eISSN 2783-6851

VILNIUS TECH
Vilnius Gediminas Technical University

NEW TRENDS in COMPUTER SCIENCES

2026
Volume 4
Issue 1
Pages 60–92
https://doi.org/10.3846/ntcs.2026.26227

DISTINGUISHING LEXICAL DEVIATION FROM SEMANTIC RESILIENCE IN LLM-BASED ROUND-TRIP TRANSLATION

Kalin KOPANOV ✉, Tatiana ATANASOVA

Department of Modelling and Optimization, Institute of Information and Communication Technologies – Bulgarian Academy of Sciences, Sofia, Bulgaria

Article History:
- received 26 February 2026
- accepted 3 July 2026

Abstract. Round-trip translation (RTT) is widely used for data augmentation and paraphrasing, yet its impact on meaning preservation remains poorly understood. We evaluate RTT systematically across two phases totaling 80,000 translation operations: 200 synthetic English passages are routed through single- and multi-hop pipelines spanning Bulgarian, Russian, and Chinese, translated by four systems ranging from a classical web translator (*googletrans*) to recent large language models (*4o-mini*, *Claude Haiku 4.5*, *GPT-5-nano*), with every passage undergoing ten RTT passes scored by BLEU, COMET-Kiwi, and xCOMET, joined by SBERT in the second phase. The results show that RTT outputs remain fluent while rapidly losing meaning: even the best configuration reaches only 0.646 xCOMET, and the four systems follow three distinct degradation regimes (front-loaded damage, continuous drift, and deterministic fixed-point convergence). High surface overlap masks this loss, creating a surface fidelity trap, and genre-level analysis suggests that creative and news texts are most vulnerable. Sample-level inspection further reveals that xCOMET over-penalizes legitimate paraphrastic variation while under-penalizing domain-critical substitutions in legal text, exposing a mismatch between translation-oriented metrics and paraphrase evaluation. These findings argue against RTT in meaning-sensitive workflows and call for paraphrase-specific evaluation metrics.

Keywords: round-trip translation (RTT), machine translation evaluation, semantic preservation, paraphrasing, data augmentation, synthetic data, large language models (LLM), xCOMET.

✉Corresponding author. E-mail: ***kalin.kopanov@iict.bas.bg***

1. Introduction

Text manipulation plays a key role in various Natural Language Processing (NLP) applications, including data augmentation, plagiarism detection, and text simplification. A simple and increasingly popular method for creating new text similar to the original is round-trip translation (RTT). In this method, a sentence is translated into one or more intermediate languages and then back into the source language. Each detour typically alters lexical choices and surface syntax while, in theory, preserving the underlying semantics. Modern Machine Translation (MT) Application Programming Interfaces (APIs) make RTT easy to implement, so researchers often use it as a quick data-augmentation method without custom models or heavy compute.

The same simplicity that makes RTT appealing can also mask a serious flaw. Translation systems are designed to produce smooth target-language output, not to ensure that meaning survives a full round-trip. Small changes at each step accumulate, and after just a few iterations, the final text can differ significantly from the original intent. Despite this risk, most RTT papers still report only improvements in downstream tasks. Questions about how much

meaning remains, how quickly quality declines, and which language sequences are safest remain largely unanswered.

This investigation extends our previous comparative analysis of Qwen 2.5 and Gemma 3 model attribution (Kopanov & Atanasova, 2025), where we demonstrated that these LLMs exhibit distinct lexical and stylistic patterns that enable high-accuracy automated attribution. While our initial work characterized those patterns, the present study examines how RTT affects the semantic preservation of AI-generated content. This research also complements our study of the degradation of stylometric attribution accuracy across various text-editing workflows (Kopanov & Atanasova, 2026), in which RTT is one of several transformation methods that can obscure model-specific linguistic signatures. Here, we specifically examine the RTT perspective, focusing on the fundamental question of semantic preservation rather than attribution robustness.

We address the gap with a controlled, two-phase evaluation of RTT across multiple language pipelines, translation engines, and text genres. The resulting findings challenge the common assumption that RTT produces "safe" paraphrases and highlight the need for methods that balance lexical diversity with reliable semantic preservation, as well as paraphrase-specific evaluation metrics. By providing a detailed, metric-aligned map of where and how meaning erodes, our study establishes a quantitative baseline for future paraphrasing systems to benchmark genuine semantic fidelity.

2. Related work

2.1. Round-trip translation for text augmentation

Early work on translation-based paraphrasing dates back to Barzilay and McKeown (2001) and Quirk et al. (2004). However, it was Bannard and Callison-Burch (2005) who first introduced the bilingual-pivot method, demonstrating that two English phrases that align with the same foreign-language phrase in a parallel corpus are strong candidates for paraphrase.

With Neural Machine Translation (NMT), the approach gained fluency and diversity. Mallinson et al. (2017) reported that neural pivoting produces paraphrases that are more grammatical and varied than those of statistical systems, and that a typologically distant pivot, such as Czech, requires greater surface reorganization while preserving meaning. Wieting and Gimpel (2018) scaled the idea up by releasing ParaNMT-50M, a fifty-million-sentence Czech-pivot corpus that clearly improves sentence-embedding quality.

The use of data augmentation via translation owes much to back-translation. Sennrich et al. (2016) showed that pairing target-side monolingual data with automatic translations back into the source language yields synthetic parallel corpora that substantially improve NMT quality (+2.8–3.7 BLEU on WMT 15 English↔German). Edunov et al. (2018) scaled the technique to hundreds of millions of sentences and found that sampled or noised back-translations provide a stronger training signal than beam-search outputs, an early indication that translation-based augmentation benefits from lexical diversity rather than maximal fidelity, the trade-off our study examines for iterated RTT.

RTT is now a standard data-augmentation tool. Thompson and Post (2020) treated paraphrase generation as a zero-shot multilingual translation task, thereby advancing low-resource

MT. Corbeil and Abdi Ghavidel (2021) introduced a semantic-similarity filter that retains only meaning-preserving back-translations for paraphrase identification. A survey by Mondal et al. (2023) confirms that RTT remains a low-cost strategy for expanding corpora used in translation, summarization, and dialogue.

RTT can also serve as a diagnostic probe. Crone et al. (2021) proposed using the semantic distance between a sentence and its round-trip version for reference-free quality estimation, although the correlations with human adequacy judgments remained weak for English-German and only moderate for English-Chinese. SemMT treats RTT as an automated testing workflow, using regex/DFA-based semantic checks to detect mistranslations and reporting sizable gains over textual/syntactic heuristics (Cao et al., 2022).

2.2. Evaluation metrics for meaning preservation

Early automatic evaluation relied on overlap-based metrics. BLEU (Papineni et al., 2002) measures n-gram precision but penalizes lexical change. METEOR (Denkowski & Lavie, 2014) enhances recall through stemming and synonymy, yet it still relies on explicit matches. Two RTT hops can still produce very high BLEU, whereas extended chains eventually drive BLEU down to low values, masking the cumulative semantic drift.

Embedding-based metrics relax this dependence on surface overlap. BERTScore (Zhang et al., 2020) computes token-level cosine similarity over contextual embeddings rather than exact n-gram matches, correlating more strongly with human judgments and rewarding meaning-preserving lexical variation that BLEU penalizes. It remains, however, a reference-based similarity measure with no explicit notion of adequacy or error severity, which limits its diagnostic value for locating where meaning is lost.

Neural reference-free metrics further reduce this bias. COMET-Kiwi (Rei et al., 2023) embeds sentences and predicts both adequacy and fluency from multilingual human ratings, so it does not punish legitimate variation. xCOMET (Guerreiro et al., 2024b) adds token-level error spans that reveal exactly where meaning breaks down.

Explainability is becoming essential. Amara et al. (2024) list open challenges, such as subword masking and perturbation tests, that must be met before evaluation can be considered trustworthy. Their taxonomy motivates our choice to combine COMET-Kiwi adequacy and fluency with xCOMET span diagnostics, as discussed in the following sections.

Separately, Zhuo et al. (2023) revisit RTT as a reference-free evaluation signal, showing that once copying effects (notably in Statistical Machine Translation) are controlled, RTT scores correlate with forward-translation (FT) quality, can predict FT scores and improve sentence-level quality estimation, and can flag adversarial systems via X-Check.

2.3. Translation quality and language-pair effects

RTT effectiveness depends on the quality of the forward and backward models for each language pair. Commercial systems report BLEU scores above eighty for English-German, while the same engines perform far worse for distant pairs such as English-Chinese, as summarized in Mondal et al. (2023). High-fidelity pairs tend to preserve source structure and limit variation, whereas low-fidelity pairs often inject noise instead of useful reformulation.

Typology also matters. Languages are typologically distant when they differ significantly in word order, morphology, or syntax. Mallinson et al. (2017) found that French and German, which are closely related to English, produce less diverse paraphrases than Czech; yet, Czech sometimes drifts semantically, indicating an optimal mid-range distance. Crone et al. (2021) observed that round-trip similarity tracks human adequacy poorly for English-German yet more closely for English-Chinese, where it outperformed a baseline quality estimator, reinforcing the interaction between distance and meaning retention.

Large-scale evidence from misinformation research echoes this view. Quelle et al. (2025) analyzed 264,487 fact-checks in 95 languages. They showed that about one-third of repeated claims cross language boundaries and become more semantically dissimilar after each language switch.

New training paradigms aim to strike a balance between fidelity and novelty, even for low-resource pairs. Zou et al. (2025) introduced TRANS-ZERO, a self-play framework that enables an LLM to translate across multiple languages, score candidates based on multilingual consistency, and improve itself through preference optimization, relying only on monolingual data.

2.4. Research gap

Although most RTT evaluations still rely on surface-level metrics such as BLEU, this approach is particularly insufficient for modern LLM-based translation systems, which function less as strict translators and more as creative paraphrasers, producing outputs that are fluent, grammatically polished, and confidently wrong. This generative fluency creates an evaluation blind spot: LLMs can produce what might be called fluent hallucinations, passages that read naturally yet have drifted factually, and overlap-based metrics cannot distinguish such outputs from faithful paraphrases. Large-scale evidence of meaning drift in multilingual settings (Quelle et al., 2025) and recent self-improving translation frameworks such as TRANS-ZERO (Zou et al., 2025) reinforce this concern but have not been applied to RTT as a data-augmentation technique. Beyond the metric problem, to our knowledge, the literature lacks a multidimensional account of RTT stability that distinguishes lexical drift, fluency preservation, and semantic integrity across extended recursive chains and tracks how each dimension evolves across multiple hops and across typologically diverse language pairs. Nor has prior work examined how this stability varies across discourse types: whether genres that tolerate creative rephrasing (e.g., storytelling, entertainment) are more susceptible to undetected semantic failures than genres with rigid terminology (e.g., legal, technical writing).

Our study addresses these gaps. We analyze both single-hop chains of the form English → pivot → English and multi-hop chains such as English → Lang 1 → Lang 2 → English, selecting pivots that span near, medium, and distant language families. We expand the translator pool from two to four systems across two experimental phases and assess semantic preservation using BLEU, COMET-Kiwi, xCOMET, and SBERT. A genre-level analysis across ten content categories tests whether certain text types are more resilient to RTT degradation, and a sample-level qualitative inspection probes the alignment between metric scores and genuine semantic shifts. This design quantifies the trade-off between fidelity and novelty, reveals the linguistic conditions that maximize paraphrastic utility, and shows limitations of current evaluation metrics when applied to paraphrase rather than translation outputs.

3. Experimental design

3.1. Dataset and models

We evaluate RTT using the same 200 synthetic passages (100 per model) that served as the held-out test set in our previous study of Qwen 2.5 and Gemma 3 model attribution (Kopanov & Atanasova, 2025). These passages span diverse domains to ensure topical variety in our semantic-preservation analysis. The synthetic data covers ten categories, as defined in Kopanov and Atanasova (2025): creative storytelling, technical explanations, data analysis reports, casual conversation, marketing copy, news reporting, instructional tutorials, philosophical discourse, legal and contractual writing, and entertainment reviews. Each passage contains approximately 200–250 words, providing sufficient context for meaningful semantic evaluation while remaining computationally tractable for multiple translation iterations. For reproducibility, the passages are reused verbatim and unmodified: the same 200 texts (ten per category for each source model, generated with Qwen 2.5 32B and Gemma 3 27B as described in Kopanov and Atanasova (2025)) are fed to every translation system in both phases, so that all system and pipeline comparisons are paired at the passage level. The synthetic source passages, the evaluation outputs underlying Tables 1–5, and A1–A3, and the analysis scripts are available from the corresponding author upon reasonable request.

3.2. Translation pipelines and iterative methodology

We implement six RTT pipelines that differ in typological complexity and hop count, with each pipeline undergoing 10 complete RTT cycles to track both incremental degradation patterns and cumulative semantic loss. Our single-hop pipelines include EN → BG → EN (Bulgarian), chosen as a lower-resource Slavic language with Cyrillic script that receives less training attention in commercial systems; EN → RU → EN (Russian), selected as a major Slavic language with extensive translation resources; and EN → ZH → EN (Simplified Chinese), included for maximum typological distance with its logographic writing system and Sino-Tibetan family membership.

The multi-hop pipelines extend this design by routing through multiple intermediate languages: EN → BG → RU → EN combines related Slavic languages to test whether stacking typologically similar pivots curbs cumulative noise and preserves meaning across multiple steps; EN → BG → ZH → EN mixes Slavic and Sino-Tibetan families to examine cross-family semantic drift; and EN → RU → ZH → EN provides sequential transfer through two major non-English languages with different writing systems. Each complete cycle follows the pattern Original → Pass 1 → Pass 2 → … → Pass 10, allowing us to track how semantic fidelity degrades with repeated translation.

Across all 200 passages, six pipelines, two systems, and ten passes, this results in 12,000 single-hop round-trip iterations and 12,000 multi-hop iterations. In total, that comprises 60,000 translation operations (where one operation is a single forward translation of one passage through one language hop), made up of 24,000 single-hop steps and 36,000 multi-hop steps.

The translation phase lasted 6 days. Processing was strictly sequential, with no batching or parallel threads, because reliability had priority over speed. Day 1 was allocated to the

googletrans runs: all six pipelines finished in about 9 to 14 hours, with a deliberate randomized pause of one to two seconds after every translation request to avoid rate limiting. The *4o-mini* runs occupied the remaining 5 days; each API call was followed by a fixed 0.5-second pause to respect rate limits. The entire translation schedule was completed without any interruptions or debugging, confirming that our infrastructure can handle a 10-pass round-trip translation evaluation smoothly, even under fully sequential conditions.

3.3. Extended system comparison

Having established the general phenomenon of semantic degradation across six pipeline configurations with two translation systems (Phase 1), we conduct a focused follow-up evaluation designed to probe the effect of the translation engine itself. Phase 2 retains the same 200 passages and the same 10-pass iterative methodology, but narrows the pipeline set to the two Bulgarian-centered routes, EN → BG → EN (single-hop) and EN → BG → RU → EN (multi-hop), while expanding the translator pool to four systems: *googletrans*, *4o-mini*, *Claude Haiku 4.5*, and *GPT-5-nano*.

The rationale for this design is threefold. First, Phase 1 demonstrated that the Bulgarian pivot produces the highest lexical fidelity among the single-hop configurations and that the EN → BG → RU → EN chain provides a controlled multi-hop comparison within the same Slavic language family. These two routes, therefore, provide a stable experimental baseline against which differences between translators can be isolated. Second, adding two LLM translators from diverse vendors and model generations turns the comparison from a binary web-NMT-versus-LLM contrast into a four-point spectrum that captures architectural and generational variation. Third, Phase 2 introduces a genre-level analysis across the ten content categories present in the dataset, enabling us to investigate whether certain text types are more resilient to RTT degradation than others.

The Phase 2 translation runs were completed over approximately three days. The *GPT-5-nano* runs took roughly two days: about 10 hours for EN → BG → EN and 17 hours for EN → BG → RU → EN, with the multi-hop chain, as expected, taking longer due to the additional translation step per pass. The *Haiku 4.5* runs were completed in approximately one day: about 8 hours for EN → BG → EN and 11 hours for EN → BG → RU → EN. As in Phase 1, all processing was strictly sequential to ensure reliability. Metric computation (xCOMET, COMET-Kiwi, SBERT, and BLEU across all passes) was performed on SLURM CPU nodes, which added approximately 1 additional day of wall-clock time.

Across all 200 passages, two pipelines, two new systems, and ten passes, Phase 2 adds 20,000 translation operations (8,000 single-hop steps and 12,000 multi-hop steps); the *googletrans* and *4o-mini* translation outputs on these two routes are carried over from Phase 1, with all metrics recomputed under the Phase 2 evaluation configuration (Section 3.5). Combined with Phase 1, the study covers 80,000 translation operations evaluated with four complementary metrics.

3.4. Translation systems

The initial broad evaluation (Phase 1) compares two translation approaches representing different technological paradigms. *Googletrans* serves as a classical NMT baseline throughout

both phases, allowing us to measure how LLM-based translators compare against a conventional neural machine translation system on the same data and pipelines. We access Google Translate through the unofficial yet widely used Python library *googletrans* (version 4.0.2) (Han, 2025). While this approach represents the current standard for accessible NMT in research settings, it carries stability limitations due to its unofficial nature. Our second system is OpenAI's *4o-mini* model (gpt-4o-mini-2024-07-18) accessed via API, which offers instruction-following capabilities with multilingual support.

The extended evaluation (Phase 2) adds two further LLM-based translators to enable a broader comparison across the current generation of language models. Anthropic's *Claude Haiku 4.5* (claude-haiku-4-5-20251001) is a lightweight yet capable model from the Claude family, selected as a representative of a competing LLM architecture with a distinct training methodology. It is called via the standard messages API, with a 4,096 max-token limit at the default temperature, and up to 5 retries with exponential backoff (2 s to 60 s). OpenAI's *GPT-5-nano* (gpt-5-nano-2025-08-07) is the smallest member of the GPT-5 series, configured with low reasoning effort and the model's default temperature, chosen to test whether a newer-generation model of similar size can outperform the older *4o-mini* on translation fidelity.

All systems receive explicit language specifications; the LLM-based translators are additionally instructed, in every request, with the prompt "You are a professional translator. Translate the following text from {source language} to {target language}. Output only the translated text without any explanations, notes, or additional text." All API access uses retry logic and exponential backoff to ensure stability. In summary, the decoding configurations were: *googletrans*, deterministic decoding with no user-controllable sampling parameters; *4o-mini*, temperature 0.3 with all other parameters at their API defaults; *Haiku 4.5*, the provider's default temperature with a 4,096 max-token limit; and *GPT-5-nano*, the provider's default temperature with low reasoning effort. These settings reflect each system's typical production configuration rather than a harmonized sampling regime, which has implications for cross-system comparisons: a lower temperature plausibly favors conservative, higher-fidelity rewrites, whereas default-temperature sampling introduces fresh variation at each pass. The system ranking reported in Section 4.6 should therefore be read as a comparison of systems as deployed, not of architectures under identical decoding conditions; a controlled study of decoding parameters is identified as future work (see Future research).

3.5. Evaluation framework

Our evaluation strategy combines multiple metrics to capture different aspects of semantic preservation and surface variation. We compute BLEU scores between consecutive iterations to measure surface-level stability and track the rate of lexical change throughout the RTT process. BLEU is computed with a custom BLEU-4 implementation: text is lowercased and tokenized with punctuation as separate tokens, modified n-gram precisions up to order four use add-epsilon smoothing for zero counts, and the standard brevity penalty is applied. While BLEU's limitations for text augmentation evaluation are well documented, it nevertheless reveals incremental drift patterns across translation passes.

To assess the standalone quality of final outputs, we employ COMET-Kiwi (Unbabel/wmt22-cometkiwi-da), a reference-free quality-estimation metric that scores each output

conditioned on a configurable source field, without requiring a reference translation (Rei et al., 2022). This shows whether RTT maintains surface readability despite potential semantic degradation. The Phase 1 and Phase 2 COMET-Kiwi evaluations used different source-text configurations; absolute COMET-Kiwi values are therefore not directly comparable across phases, although within-phase comparisons are unaffected. Specifically, the Phase 2 evaluation used the generation prompt as the COMET-Kiwi source field, whereas Phase 1 used the original passage; because the original passage is lexically closer to the round-trip output than the prompt, the Phase 1 scores are systematically higher. For semantic preservation analysis, we use xCOMET (Unbabel/XCOMET-XXL; 10.7B parameters) to measure fidelity between the original texts and the output of each pass, with Pass-10 scores as the primary endpoint (Guerreiro et al., 2024a). xCOMET is run in reference-only mode: the original passage occupies the reference field and the round-trip output the hypothesis field, with no source segment supplied. xCOMET provides segment-level scores and error span detection, enabling detailed analysis of where and how semantic degradation occurs.

In the extended evaluation (Phase 2), we add Sentence-BERT (SBERT) as a fourth metric to provide an independent embedding-based perspective on semantic similarity (Reimers & Gurevych, 2019). SBERT was not part of the original Phase 1 design; it was added in Phase 2 because of the Phase 1 results. We use the all-mpnet-base-v2 model, a later sentence-transformers checkpoint built on the SBERT framework, to compute cosine similarity between the original and paraphrased texts at each pass. SBERT operates in a different representational space from xCOMET: it encodes entire sentences into dense vectors optimized for semantic textual similarity, whereas xCOMET uses a translation-quality framework trained on human adequacy judgments. Including both allows us to test whether the semantic degradation observed via xCOMET is confirmed by a metric from a different methodological tradition, and to identify cases where the two metrics diverge.

Finally, we compute the Sørensen-Dice coefficient on word bigrams to assess meaningful paraphrastic variation, excluding pairs with similarity exceeding 0.9 from the semantic preservation analysis, as they represent insufficient transformation. The Dice score is calculated on the same tokenization that xCOMET produces, so no additional input-reference preprocessing is required. This filtering ensures we analyze only cases where RTT produces real textual variation rather than near-identical copies.

The four metrics span a wide range of computational costs, which bear on their practical applicability. On CPU-only SLURM nodes processing 100 passages × 2 source models × 10 passes, xCOMET (XCOMET-XXL, 10.7B parameters) requires approximately 145 minutes per evaluation run (~14.5 minutes per pass), whereas COMET-Kiwi completes the same workload in approximately 3.4 minutes (~20 seconds per pass), roughly 40 times faster. SBERT falls between the two at approximately 7 minutes per run, and BLEU is effectively instantaneous. These cost differences suggest a tiered evaluation strategy for production settings: BLEU and Sørensen-Dice can serve as real-time surface-change monitors; COMET-Kiwi provides a fast fluency and adequacy check; and xCOMET should be reserved for definitive semantic fidelity assessment, where its discriminative power justifies the computational overhead.

For the statistical analysis, note that BLEU in Tables 1 and 2 is computed at the corpus level, whereas Table 5 reports the mean of per-sample BLEU scores; the two aggregations

differ slightly for the systems carried over from Phase 1. All Pass-10 means are reported with 95% confidence intervals obtained by bias-corrected and accelerated (BCa) bootstrap resampling (10,000 resamples) over the 100 passages per source model. Because every translation system and pipeline was applied to the same set of passages, system comparisons use paired tests: a Friedman omnibus test across the four systems within each pipeline × source-model combination, followed by pairwise Wilcoxon signed-rank tests with Holm correction for the six comparisons per family, with matched-pairs rank-biserial correlations (r) reported as effect sizes. The same paired procedure is used to compare single-hop and multi-hop routes within each system, and to compare Qwen 2.5 and Gemma 3 source texts (paired by generation prompt) within each configuration; for these two analyses, Holm correction is applied over all contrasts of each analysis (the 24 Phase 1 and 8 Phase 2 hop contrasts as separate families, and the 20 source-model contrasts as one family). Non-parametric tests were chosen because per-sample xCOMET scores are bounded and skewed; significance is assessed at $\alpha = 0.05$. Complete confidence intervals for all four metrics and the full pairwise comparisons are reported in Appendix A.

4. Results

4.1. BLEU score evolution across iterations

Analysis of BLEU scores between consecutive iterations reveals distinct degradation patterns across translation systems and pipeline configurations. Inter-pass BLEU (the similarity between successive passes such as Pass 1 versus Pass 2) starts high, from 0.64 to 0.86 for *googletrans* and from 0.78 to 0.93 for *4o-mini*, yet it declines along different trajectories for each pipeline.

Figures 1–4 trace corpus-level BLEU across ten round-trip passes for every pipeline and translation engine. Red and orange curves depict Qwen 2.5 results, whereas blue and violet curves depict Gemma 3. Within each color family, line style identifies the language path: solid (EN → BG → EN; EN → BG → RU → EN), dashed (EN → RU → EN; EN → BG → ZH → EN), and dotted (EN → ZH → EN; EN → RU → ZH → EN). In the *googletrans* plots, single-hop loops degrade gradually, with the Chinese route dropping most steeply (Figure 1), while multi-hop loops fall far faster, especially the dotted chain, highlighting the impact of typological distance (Figure 2). The OpenAI *4o-mini* plots begin at higher BLEU values: single-hop loops remain comparatively stable over the entire sequence (Figure 3), and even the multi-hop loops decline more gently than their *googletrans* counterparts (Figure 4).

The BLEU trajectories in Figures 1–4 are captured numerically in Tables 1 and 2. At the first round-trip pass, *googletrans* outputs already diverge far more from the source (BLEU ≈ 0.27–0.48 across models; 0.300–0.480 for Qwen 2.5, 0.268–0.419 for Gemma 3) than do *4o-mini* outputs (≈ 0.42–0.70 across models; 0.527–0.703 for Qwen 2.5, 0.421–0.616 for Gemma 3), confirming that the web engine applies a more aggressive lexical rewrite while the LLM-based system is comparatively conservative. Degradation is sharpest whenever the chain traverses Simplified Chinese: the multi-hop EN → RU → ZH → EN route loses 25–30% of its n-gram overlap, and the single-hop EN → ZH → EN loop records the largest single-pass loss among the one-hop pipelines in both models. To quantify convergence, we adopt a 1-BLEU-point

stabilization threshold (Δ < 0.01): the earliest pass after which every subsequent change stays below one BLEU point. Under this criterion:

- OpenAI's *4o-mini* flattens quickly: all six pipelines are stable by Pass 5 (median Pass 4 for both Qwen 2.5 and Gemma 3).
- *Googletrans* stabilizes earlier for five of the six pipelines (Pass 3–6), but the EN → ZH → EN loop continues to drift in one-point increments until Pass 10 for Qwen 2.5 (Gemma 3 stabilizes at Pass 4), extending the overall stabilization window across pipelines from Pass 3 to Pass 10.

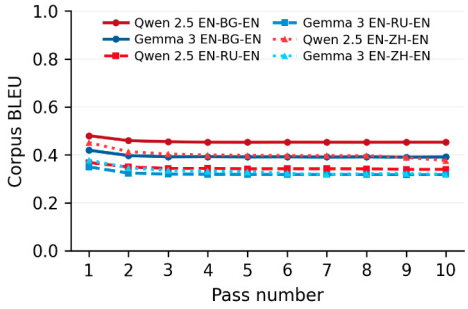

Figure 1. *Googletrans* single-hop RTT

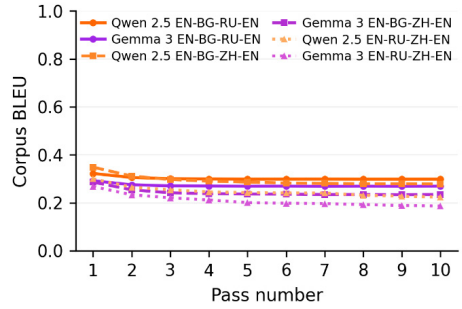

Figure 2. *Googletrans* multi-hop RTT

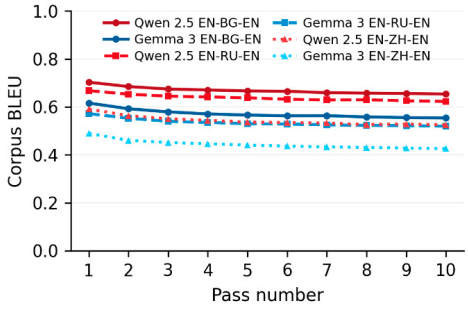

Figure 3. *4o-mini* single-hop RTT

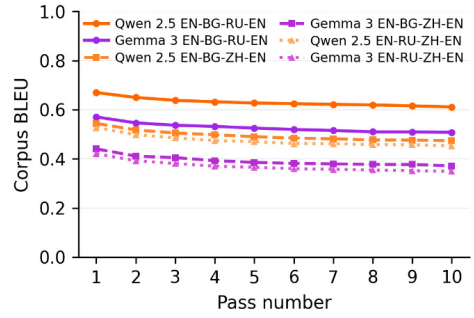

Figure 4. *4o-mini* multi-hop RTT

Table 1. BLEU per model (*googletrans*, Δ0.01)

Pipeline	Qwen 2.5 Pass 1	Gemma 3 Pass 1	Qwen 2.5 Pass 10	Gemma 3 Pass 10	Qwen 2.5 Decay %	Gemma 3 Decay %	Qwen 2.5 Stab. pass	Gemma 3 Stab. pass
EN-BG-EN	0.480	0.419	0.453	0.392	−5.63	−6.44	Pass 3	Pass 3
EN-RU-EN	0.368	0.350	0.340	0.318	−7.61	−9.14	Pass 3	Pass 3
EN-ZH-EN	0.451	0.379	0.376	0.318	−16.63	−16.09	Pass 10	Pass 4
EN-BG-RU-EN	0.323	0.292	0.299	0.270	−7.43	−7.53	Pass 3	Pass 3
EN-BG-ZH-EN	0.349	0.286	0.279	0.235	−20.06	−17.83	Pass 4	Pass 4
EN-RU-ZH-EN	0.300	0.268	0.224	0.187	−25.33	−30.22	Pass 4	Pass 6

Table 2. BLEU per model (*4o-mini*, Δ0.01)

Pipeline	Qwen 2.5 Pass 1	Gemma 3 Pass 1	Qwen 2.5 Pass 10	Gemma 3 Pass 10	Qwen 2.5 Decay %	Gemma 3 Decay %	Qwen 2.5 Stab. pass	Gemma 3 Stab. pass
EN-BG-EN	0.703	0.616	0.653	0.554	−7.11	−10.06	Pass 4	Pass 4
EN-RU-EN	0.668	0.572	0.623	0.520	−6.74	−9.09	Pass 3	Pass 4
EN-ZH-EN	0.591	0.490	0.525	0.426	−11.17	−13.06	Pass 4	Pass 3
EN-BG-RU-EN	0.670	0.570	0.611	0.508	−8.81	−10.88	Pass 4	Pass 3
EN-BG-ZH-EN	0.544	0.441	0.474	0.372	−12.87	−15.65	Pass 4	Pass 5
EN-RU-ZH-EN	0.527	0.421	0.453	0.350	−14.04	−16.86	Pass 5	Pass 5

After the stabilization point, residual BLEU fluctuations are well within the metric's typical test-retest noise, indicating that the curves have genuinely plateaued and that further iterations introduce no systematic lexical change.

4.2. Semantic preservation analysis (xCOMET)

While BLEU scores stabilize after several iterations, xCOMET analysis reveals that semantic preservation continues to deteriorate severely, with no pipeline achieving "Good" (≥ 0.70) or even "Moderate" (≥ 0.55) meaning retention by Pass 10 (averaged across Qwen 2.5 and Gemma 3). *4o-mini* achieves higher semantic fidelity than *googletrans*, averaging xCOMET scores of 0.526 across Qwen 2.5 and Gemma 3 outputs, still within the "Poor" band. By contrast, *googletrans* averages only 0.222, placing every configuration in the "Very Poor" category (Table 3).

Multi-hop chains consistently underperform their single-hop counterparts across both translation systems, with the most severe degradation observed in *googletrans* configurations involving Simplified Chinese (EN → RU → ZH → EN: 0.132). Paired Wilcoxon tests qualify this picture: the multi-hop deficit is large and uniformly significant for *googletrans* (all twelve single- versus multi-hop contrasts, rank-biserial $r \geq 0.50$, Holm-adjusted $p < 0.001$), whereas for *4o-mini* it is small (0.007–0.047 xCOMET) and significant in seven of twelve contrasts. Model-specific patterns persist throughout the analysis, with Qwen 2.5 consistently outperforming Gemma 3 in semantic preservation across all pipeline configurations (paired Wilcoxon by prompt; Holm-adjusted $p < 0.001$ and mean gap ≥ 0.072 xCOMET in every configuration). The error analysis reveals an 87–100% error detection rate with 5.8–22.6 average error spans per sample, indicating pervasive semantic degradation rather than isolated translation failures.

At the individual-sample level, several *googletrans* configurations produce negative xCOMET scores, meaning that the RTT output is rated as worse than a completely unrelated text. The minimum per-sample scores reach −0.040 (EN → BG → RU → EN, Qwen 2.5, Pass 2) and −0.031 (EN → BG → EN, Gemma 3, Pass 3). These negative values indicate semantic failure: the passage has been so distorted that xCOMET considers it actively misleading rather than merely irrelevant. *4o-mini* produces no negative xCOMET values on any configuration; across the entire study, the only LLM-generated negatives are three marginal *GPT-5-nano* samples

in Phase 2 (minimum −0.039), so crossing zero remains essentially a property of the weakest translation engine on its most challenging routes. Manual inspection of these negative-scoring outputs confirms that the damage is real (e.g., "the operative clause of an easement agreement reduced to unintelligible repetition"); however, the zero crossing itself should not be read as a threshold, since near-zero positive scores on LLM outputs frequently correspond to largely faithful paraphrases whose penalties stem from surface artifacts (Section 5.2).

4.3. Fluency and adequacy assessment (COMET-Kiwi)

Fluency and adequacy assessments reveal a paradox: despite severe semantic degradation observed in xCOMET analysis, COMET-Kiwi scores for final outputs (Pass 10) remain high across all configurations (Table 4). This disconnect between semantic preservation and fluency/adequacy metrics highlights the fundamental difference between meaning retention and linguistic fluency in iterative translation processes.

4.4. Pipeline-specific characteristics

A cross-metric inspection of Tables 1–4 shows that each loop has a recognizable profile once surface similarity (BLEU), semantic preservation (xCOMET), and fluency/adequacy (COMET-Kiwi) are considered together. The single-hop Bulgarian pivot (EN → BG → EN) retains the highest lexical overlap, with final-pass BLEU of 0.653 (*4o-mini*) and 0.453 (*googletrans*) for Qwen 2.5; Gemma 3 produces 0.554 and 0.392, respectively. The best two-hop configuration, EN → BG → RU → EN, averages 0.560 (*4o-mini*) and 0.285 (*googletrans*) across Qwen 2.5 and Gemma 3, confirming that every additional hop erodes surface form.

Table 3. Semantic preservation scores comparing final outputs (Pass 10) to original texts

Pipeline	Translation System	Qwen 2.5 xCOMET	Gemma 3 xCOMET	Quality Category
EN-BG-EN	googletrans	0.331 [0.305, 0.357]	0.199 [0.177, 0.222]	Very Poor
EN-RU-EN	googletrans	0.290 [0.266, 0.313]	0.167 [0.147, 0.187]	Very Poor
EN-ZH-EN	googletrans	0.390 [0.358, 0.419]	0.217 [0.191, 0.243]	Very Poor
EN-BG-RU-EN	googletrans	0.235 [0.213, 0.257]	0.135 [0.120, 0.150]	Very Poor
EN-BG-ZH-EN	googletrans	0.224 [0.202, 0.247]	0.141 [0.124, 0.160]	Very Poor
EN-RU-ZH-EN	googletrans	0.205 [0.184, 0.228]	0.132 [0.118, 0.147]	Very Poor
EN-BG-EN	4o-mini	0.640 [0.607, 0.673]	0.448 [0.413, 0.482]	Poor
EN-RU-EN	4o-mini	0.646 [0.611, 0.679]	0.445 [0.409, 0.481]	Poor
EN-ZH-EN	4o-mini	0.623 [0.593, 0.652]	0.425 [0.389, 0.460]	Poor
EN-BG-RU-EN	4o-mini	0.627 [0.594, 0.659]	0.425 [0.389, 0.459]	Poor
EN-BG-ZH-EN	4o-mini	0.615 [0.582, 0.648]	0.403 [0.368, 0.438]	Poor
EN-RU-ZH-EN	4o-mini	0.616 [0.582, 0.649]	0.398 [0.361, 0.435]	Poor

Note: Quality categories are based on the following xCOMET score ranges: Excellent (≥ 0.85), Good (0.70 to < 0.85), Moderate (0.55 to < 0.70), Poor (0.40 to < 0.55), Very Poor (< 0.40). The same thresholds are applied to COMET-Kiwi scores in Table 4. Values in brackets are 95% BCa bootstrap confidence intervals (10,000 resamples; n = 100 passages per source model). These bands (Quality Category) are our own operational thresholds, introduced for interpretability; they are not part of the xCOMET specification (Guerreiro et al., 2024b). Quality categories are assigned to the mean of the two source models.

Table 4. COMET-Kiwi quality scores for final outputs (Pass 10)

Pipeline	Translation System	Qwen 2.5 COMET-Kiwi	Qwen 2.5 COMET-Kiwi std	Gemma 3 COMET-Kiwi	Gemma 3 COMET-Kiwi std	Quality Category
EN-BG-EN	googletrans	0.776	0.050	0.737	0.067	Good
EN-RU-EN	googletrans	0.732	0.082	0.691	0.077	Good
EN-ZH-EN	googletrans	0.794	0.085	0.750	0.095	Good
EN-BG-RU-EN	googletrans	0.706	0.059	0.659	0.060	Moderate
EN-BG-ZH-EN	googletrans	0.723	0.079	0.680	0.075	Good
EN-RU-ZH-EN	googletrans	0.692	0.099	0.648	0.090	Moderate
EN-BG-EN	4o-mini	0.884	0.018	0.878	0.021	Excellent
EN-RU-EN	4o-mini	0.884	0.017	0.879	0.021	Excellent
EN-ZH-EN	4o-mini	0.877	0.020	0.875	0.022	Excellent
EN-BG-RU-EN	4o-mini	0.880	0.027	0.876	0.022	Excellent
EN-BG-ZH-EN	4o-mini	0.873	0.020	0.869	0.025	Excellent
EN-RU-ZH-EN	4o-mini	0.873	0.021	0.868	0.025	Excellent

Semantic scores tell a complementary story. In *4o-mini*, the two Slavic single-hops (EN → BG → EN and EN → RU → EN) score highest on xCOMET: 0.640–0.646 for Qwen 2.5 and 0.445–0.448 for Gemma 3 (all "Poor"). By contrast, under *googletrans* the strongest loop is the Chinese single-hop EN → ZH → EN, averaging 0.304 across models (still "Very Poor"). The largest meaning loss occurs in the EN → RU → ZH → EN chain, which drops to 0.616/0.398 in *4o-mini* (avg. 0.507, "Poor") and a severe 0.205/0.132 in *googletrans* (avg. 0.169, "Very Poor"), illustrating the compounding effect of typological distance.

Fluency remains high in the stronger engine: in *4o-mini*, COMET-Kiwi places EN → BG → EN and EN → RU → EN at 0.884/0.878 and 0.884/0.879, respectively (avg. 0.881, both "Excellent"), with EN → ZH → EN only slightly lower at 0.877/0.875 (avg. 0.876, also "Excellent"). Under *googletrans*, the ordering changes and the Chinese loop becomes the smoothest at 0.794/0.750 (avg. 0.772, "Good"), ahead of Bulgarian 0.776/0.737 (avg. 0.757, "Good") and Russian 0.732/0.691 (avg. 0.712, "Good").

Taken together, these figures indicate that Bulgarian pivots maximize lexical fidelity, Chinese pivots in the weaker engine maximize sentence-level fluency, Slavic single-hops in *4o-mini* minimize semantic loss, and the EN → RU → ZH → EN multi-hop performs worst or near-worst on every metric. This triangulation shows that no single pivot dominates all dimensions, highlighting trade-offs between fidelity, semantics, and fluency.

4.5. Correlation analysis between surface quality and semantic preservation

At the pipeline level, COMET-Kiwi and xCOMET rise together, with Spearman $\rho = 0.99$ ($n = 12$, $p < 0.001$). When Qwen 2.5 and Gemma 3 are considered separately, the correlation remains very strong at $\rho = 0.95$ ($n = 24$, $p < 0.001$). In both cases $\rho \geq 0.95$, confirming a strong monotonic link: pipelines that deliver higher fluency also retain more meaning, although their absolute scores diverge sharply (the configuration means share the same underlying

passages, so these correlations are descriptive rather than inferential). *4o-mini* clusters COMET-Kiwi around 0.876 with xCOMET 0.526, whereas *googletrans* sits lower at COMET-Kiwi 0.716 with xCOMET 0.222. This contrast shows that even the "better" translator lifts semantic fidelity only from "Very Poor" to "Poor," while the web-based MT system sounds polished yet discards most of the message.

Model-specific patterns persist across this correlation space: Qwen 2.5 samples consistently achieve higher xCOMET scores than Gemma 3 counterparts within identical pipeline configurations, while their surface-quality scores remain comparable. This divergence suggests that source-text characteristics are strongly associated with semantic preservation during RTT processing, independent of surface-level fluency outcomes.

Per-sample dispersion (reflected in the confidence-interval widths of Table 3 and the standard-deviation (σ) columns of Table 4) adds detail to the mean-level picture. On the semantic side, *4o-mini* pipelines show much higher dispersion (xCOMET $\sigma \approx 0.17$–0.19) than their *googletrans* counterparts ($\sigma \approx 0.10$–0.13), signaling that the LLM translator can produce near-faithful renditions for some passages yet still fails for others. By contrast, fluency is both higher and more homogeneous. Under *4o-mini*, COMET-Kiwi σ does not exceed 0.03, whereas *googletrans* ranges from 0.05 to 0.10. This variability gap explains why the Spearman ρ remains strong but not perfect: pipelines cluster along the COMET-Kiwi–xCOMET trend line, while individual passages scatter around it.

4.6. Extended system comparison

Having established the general pattern of semantic degradation across six pipelines and two systems, we now examine whether selecting a translation engine can mitigate this effect. Figure 5 presents the complete cross-comparison of all four translation systems on all four metrics across the ten passes for the two Bulgarian-centered pipelines. Table 5 summarizes the Pass-10 scores.

The four translation systems form a clear ranking in terms of semantic fidelity (xCOMET): *4o-mini* > *Haiku 4.5* > *GPT-5-nano* > *googletrans*. This ordering holds consistently for both source models and both pipeline configurations. On the single-hop EN → BG → EN route, *4o-mini* achieves 0.640 xCOMET for Qwen 2.5, followed by *Haiku 4.5* at 0.591 (−8%), *GPT-5-nano* at 0.559 (−13%), and *googletrans* at 0.331 (−48%). The same ranking and approximate gaps persist for Gemma 3 and for the multi-hop route.

This ranking is statistically robust. Friedman tests reject equality of the four systems in every pipeline × source-model combination ($\chi^2(3) \geq 177.9$, $n = 100$, all $p < 0.001$), and all 24 pairwise Wilcoxon signed-rank comparisons remain significant after Holm correction (adjusted $p \leq 0.008$). Effect sizes scale with the ranking gaps: comparisons involving *googletrans* are at or near the theoretical maximum (rank-biserial $r \geq 0.94$), the contrasts between *4o-mini* or *Haiku 4.5* and *GPT-5-nano* are large ($r \approx 0.36$–0.77), and the smallest contrast, *4o-mini* versus *Haiku 4.5*, remains significant in all four configurations ($r = 0.31$–0.59). The four-system ordering is therefore not attributable to sampling noise (full pairwise results in Table A3, Appendix A).

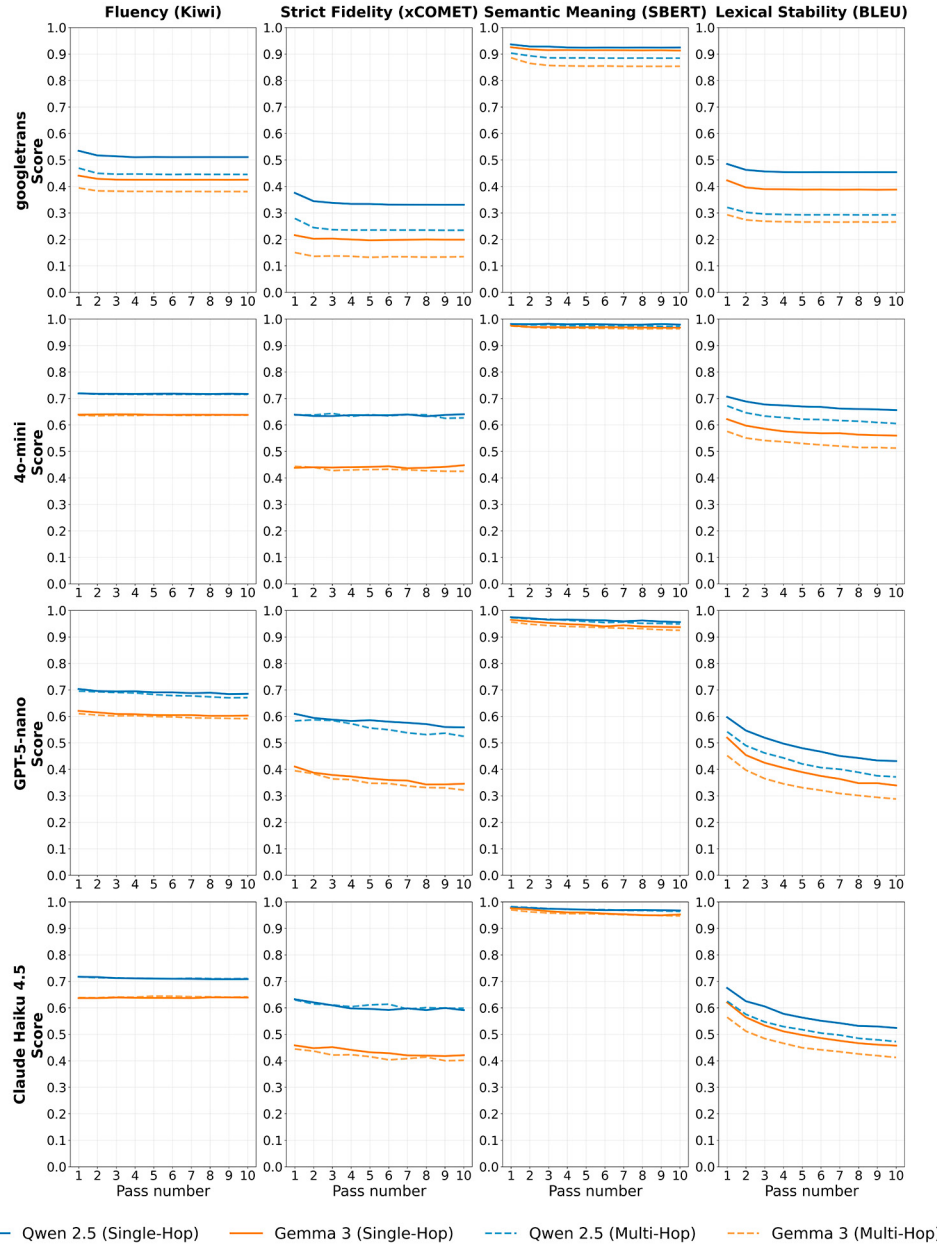

Figure 5. Comprehensive model comparison across four translation systems and four metrics. Rows represent translation systems (*googletrans*, *4o-mini*, *GPT-5-nano*, *Claude Haiku 4.5*); columns represent evaluation metrics (COMET-Kiwi, xCOMET, SBERT, BLEU). Solid lines denote single-hop (EN → BG → EN), dashed lines denote multi-hop (EN → BG → RU → EN). Blue curves show Qwen 2.5 results, orange curves show Gemma 3

Table 5. Pass-10 scores across four translation systems and two pipelines

Pipeline	Translation System	Qwen 2.5 xCOMET	Gemma 3 xCOMET	Qwen 2.5 COMET-Kiwi	Gemma 3 COMET-Kiwi	Qwen 2.5 SBERT	Gemma 3 SBERT	Qwen 2.5 BLEU	Gemma 3 BLEU
EN-BG-EN	4o-mini	0.640 [0.607, 0.673]	0.448 [0.413, 0.482]	0.717	0.638	0.979	0.968	0.656	0.560
EN-BG-RU-EN	4o-mini	0.627 [0.594, 0.659]	0.425 [0.389, 0.459]	0.715	0.637	0.972	0.963	0.605	0.512
EN-BG-EN	Haiku 4.5	0.591 [0.556, 0.623]	0.421 [0.385, 0.456]	0.709	0.639	0.968	0.952	0.524	0.457
EN-BG-RU-EN	Haiku 4.5	0.599 [0.563, 0.632]	0.401 [0.366, 0.436]	0.711	0.641	0.964	0.947	0.473	0.413
EN-BG-EN	GPT-5-nano	0.559 [0.523, 0.593]	0.346 [0.310, 0.381]	0.685	0.603	0.956	0.937	0.431	0.339
EN-BG-RU-EN	GPT-5-nano	0.525 [0.489, 0.559]	0.322 [0.291, 0.353]	0.671	0.592	0.949	0.925	0.371	0.288
EN-BG-EN	googletrans	0.331 [0.305, 0.357]	0.199 [0.177, 0.222]	0.511	0.425	0.925	0.913	0.454	0.388
EN-BG-RU-EN	googletrans	0.235 [0.213, 0.257]	0.135 [0.120, 0.150]	0.445	0.380	0.884	0.854	0.293	0.266

Note: Values in brackets are 95% BCa bootstrap confidence intervals (10,000 resamples; n = 100 passages per source model). COMET-Kiwi values are not directly comparable to Table 4 (see Section 3.5).

Beyond absolute scores, the systems differ in their degradation dynamics (Figure 5, xCOMET column), which fall into three qualitatively distinct regimes (Table 6). The first regime, *front-loaded damage*, is exhibited by *4o-mini*: xCOMET scores remain essentially flat across all ten passes (EN → BG → EN, Qwen 2.5: 0.639 at Pass 1, 0.640 at Pass 10), indicating that all semantic loss occurs in the first translation step and subsequent iterations add no further semantic loss. The second regime, *continuous drift*, characterizes *Haiku 4.5* and *GPT-5-nano*: both exhibit a gradual downward slope, with *Haiku 4.5* losing about 5–10% of its initial xCOMET score between Pass 1 and Pass 10 and *GPT-5-nano* losing 8–19%, depending on source model and pipeline. For these systems, each additional RTT iteration compounds the semantic loss. The third regime, *deterministic fixed-point convergence*, describes *googletrans*: after roughly five passes, xCOMET scores become virtually constant (EN → BG → RU → EN, Qwen 2.5: 0.235 at Pass 5, 0.235 at Pass 6, 0.235 at Pass 8, 0.235 at Pass 10), with inter-pass variation below 0.001. This plateau indicates that the NMT system reaches a translation fixed point, a sentence form that round-trips to itself, whereas the stochastic sampling of LLM-based systems prevents such convergence.

Table 6. Summary of the three degradation regimes identified across translation systems

Regime	System(s)	Semantic trajectory (xCOMET)	Fluency trajectory (COMET-Kiwi)	Practical implication
Front-loaded damage	4o-mini	Flat after Pass 1 (EN-BG-EN, Qwen 2.5: 0.639 at Pass 1, 0.640 at Pass 10); near-zero drift in all genres	Stable and high; recovery events on 10–20% of transitions	A single pass captures the full semantic cost; further passes add no additional semantic cost
Continuous drift	*Haiku 4.5*; *GPT-5-nano*	Gradual decline: Haiku 4.5 loses 5–10% and GPT-5-nano 8–19% of the Pass-1 score by Pass 10	Oscillating; recovery events on 20–31% (Haiku 4.5) and 28–38% (GPT-5-nano) of transitions	Each additional pass compounds the loss; capping the number of passes directly reduces degradation
Deterministic fixed-point convergence	googletrans	Virtually constant after about Pass 5 (inter-pass variation below 0.001)	Locked after about Pass 5; recovery events on only 9–11% of transitions	Passes beyond the fifth add computation without producing further textual change

Note: Trajectories refer to the Bulgarian-centered pipelines of Phase 2 (Section 4.6); recovery events are pass-to-pass COMET-Kiwi increases exceeding 0.005.

The multi-hop penalty is itself regime-dependent: paired single- versus multi-hop comparisons are large and significant for *googletrans* (rank-biserial $r \geq 0.79$, Holm-adjusted $p < 0.001$) but small for the LLM translators (mean xCOMET differences of −0.007 to 0.034), reaching significance in only about half of the configurations; for *Haiku 4.5* on Qwen 2.5 passages, the multi-hop route even scores marginally, though not significantly, higher (0.599 vs. 0.591). This is consistent with the front-loaded damage regime: a system that inflicts most of its semantic loss in the first translation step leaves little additional damage for a second pivot to add.

Fluency (COMET-Kiwi) remains high across all LLM-based systems (0.59–0.72), confirming that the fluency-meaning paradox observed in Phase 1 extends to the new translators. Even *GPT-5-nano*, which drops to 0.322 xCOMET on the worst configuration (EN → BG → RU → EN, Gemma 3), still produces text rated 0.592 on fluency. The *googletrans* baseline scores lower on fluency as well (0.38–0.51), consistent with its weaker overall translation quality.

Crucially, the pass-by-pass COMET-Kiwi trajectories reveal that LLM-based systems do not degrade fluency monotonically. Instead, they exhibit an oscillating pattern in which a disfluent pass is frequently followed by a partial or full recovery. To quantify this, we classify any pass-to-pass increase in COMET-Kiwi exceeding 0.005 as a recovery event. Under this criterion, *GPT-5-nano* shows recovery on 28–38% of transitions (depending on source model and pipeline), with individual samples jumping as much as +0.161 in a single pass. *Haiku 4.5* recovers on 20–31% of transitions, and *4o-mini* on 10–20%. By contrast, *googletrans* recovers on only 9–11% of transitions, and 11–20% of its samples follow an effectively non-increasing trajectory (treating pass-to-pass increases below 0.001 as flat), compared with 0–2% for the LLMs. The *googletrans* trajectories tell a different story: after roughly five passes, fluency scores become completely stable, consistent with the fixed-point convergence described above. We return to the mechanistic explanation of this contrast in Section 5. The practical implication is that LLM-based RTT is less predictable at the individual-pass level, even though its average fluency remains high.

BLEU scores reveal an additional contrast. While *4o-mini* retains high lexical overlap (0.51–0.66 at Pass 10), *GPT-5-nano* and *Haiku 4.5* produce far more aggressive surface rewrites (BLEU 0.29–0.52), indicating that these systems introduce greater lexical variation per pass. This contrast shows what we term a *surface fidelity trap*: on the EN → BG → EN route at Pass 10, *googletrans* achieves higher BLEU than *GPT-5-nano* for Qwen 2.5 (0.454 vs. 0.431), yet its xCOMET is dramatically worse (0.331 vs. 0.559). In other words, the weaker system preserves more of the original n-grams while destroying far more meaning. The same inversion appears for Gemma 3 on the single-hop route (BLEU 0.388 vs. 0.339; xCOMET 0.199 vs. 0.346). On the multi-hop route there is no crossover (*googletrans* falls below *GPT-5-nano* on both metrics), yet the disproportion persists: a modest BLEU deficit (0.293 vs. 0.371) coincides with an xCOMET score less than half of *GPT-5-nano*'s (0.235 vs. 0.525). These crossovers demonstrate that high surface overlap can mask severe semantic loss and that BLEU alone is an unreliable indicator of RTT quality, a finding that reinforces earlier warnings in the literature but is quantified here across multiple systems for the first time.

4.7. SBERT as a complementary semantic metric

SBERT cosine similarity provides an independent validation of the degradation patterns observed through xCOMET, but with a compressed dynamic range. Across all four systems, SBERT scores at Pass 10 range from 0.854 (*googletrans*, EN → BG → RU → EN, Gemma 3) to 0.979 (*4o-mini*, EN → BG → EN, Qwen 2.5). By comparison, xCOMET spans a much wider range, from 0.135 to 0.640, across the same configurations. The system ranking remains identical under both metrics (*4o-mini* > *Haiku 4.5* > *GPT-5-nano* > *googletrans*), confirming the consistency of the finding, but SBERT's narrow range (all scores above 0.85) means it cannot clearly distinguish between "Poor" and "Very Poor" semantic preservation as xCOMET does.

This compression arises because SBERT embeddings capture broad topical and structural similarity; a passage about temple exploration will still embed near other temple-exploration passages even after substantial factual distortion. xCOMET, trained on fine-grained human translation-quality judgments, is more sensitive to the kinds of meaning shifts that accumulate during RTT: entity swaps, negation flips, and register changes. For RTT evaluation, xCOMET is therefore the more discriminating metric, while SBERT serves as a useful sanity check confirming that the observed patterns are not artifacts of a single evaluation framework.

4.8. Sørensen-Dice surface similarity as a pre-metric indicator

The xCOMET evaluation pipeline computes a Sørensen-Dice coefficient on word bigrams before scoring each sample, primarily to filter near-identical round-trips (threshold > 0.9). The per-pass Dice statistics logged during evaluation, however, reveal an independent signal about raw surface distortion that merits reporting. At Pass 1 on the EN → BG → EN route (Qwen 2.5 values), the four systems rank as follows: *4o-mini* 0.740, *Haiku 4.5* 0.713, *GPT-5-nano* 0.655, *googletrans* 0.544. This ordering is identical to the final xCOMET ranking, suggesting that word-level surface overlap after a single RTT pass is already informative about downstream semantic preservation at the system level. The Gemma 3 values are consistently 0.04–0.08 lower than their Qwen 2.5 counterparts across all systems, mirroring the model

gap observed in every other metric and confirming that source-text properties influence preservation outcomes from the very first translation step.

Tracking Dice similarity across all ten passes shows distinct system profiles. *4o-mini* declines gradually from 0.740 to 0.693 (−6.3%), whereas *Haiku 4.5* drops from 0.713 to 0.569 (−20.2%) and *GPT-5-nano* from 0.655 to 0.489 (−25.3%). *Googletrans* falls quickly from 0.544 to 0.511 but stabilizes by Pass 4, after which the coefficient remains essentially constant. Multi-hop pipelines amplify these trends: *GPT-5-nano* on EN → BG → RU → EN declines from 0.598 to 0.421 (−29.5%). The high-similarity filter (Dice > 0.9) is triggered only once across all 32 evaluation runs (8 pipeline configurations × 4 metrics), for a single Gemma 3 sample under *GPT-5-nano* at Pass 1, confirming that RTT universally introduces meaningful surface variation and that the filter plays a negligible role in practice.

Because the Dice coefficient requires no model inference and can be computed quickly, it could serve as a lightweight pre-screening tool in production RTT pipelines: a low initial Dice value would signal that the translation engine is likely to produce poor semantic preservation, enabling early intervention before expensive neural evaluation is invoked.

4.9. Genre-level semantic preservation

The genre-level analysis reveals substantial variation in semantic resilience across the ten content categories (Figure 6). This analysis is based on per-category averages at Pass 10 across all four systems and both pipelines. Each genre cell averages ten passages per source model; the genre comparisons are therefore descriptive rather than inferential.

News reporting is the most vulnerable category under every system, with creative storytelling and entertainment reviews next; at the resilient end, marketing copy ranks highest for *4o-mini* and *Haiku 4.5*, and technical explanations for *GPT-5-nano* and *googletrans*. Under *4o-mini*, for instance, marketing copy attains 0.744–0.774 xCOMET for Qwen 2.5 (within the "Good" band), whereas creative storytelling drops to 0.499–0.524 (firmly "Poor"). The gap between the most and least resilient genres is substantial for every translator (approximately 0.32–0.36 xCOMET points for the three LLM-based systems and 0.15 for *googletrans*) and is comparable in magnitude to the gap between *4o-mini* and *googletrans* on the same genre.

Part of the genre effect plausibly reflects the structural properties of each text type, although Section 5.1 shows that metric sensitivity also contributes. Marketing copy and philosophical discourse tend to employ formulaic structures, abstract vocabulary, and conventional phrasing that translation systems handle reliably. Creative storytelling, by contrast, relies on specific imagery, narrative coherence, and stylistic nuance that are harder to preserve through translation. News reporting suffers because it contains proper nouns, factual claims, and temporal references that are particularly vulnerable to entity swaps and factual distortion during iterative translation.

The drift columns in Figure 6 show that genre vulnerability is not uniform across systems. *4o-mini* exhibits only small xCOMET drift in every genre (at most 0.05 points, confirming its pass-invariant behavior), while *GPT-5-nano* and *Haiku 4.5* show genre-dependent degradation: depending on the pipeline and source model, individual genres lose up to 0.14 xCOMET points between Pass 1 and Pass 10, whereas technical explanations and instructional tutorials never lose more than 0.05 points. This interaction between the system and the genre suggests that the resilience of a given text type depends not only on its inherent translational complexity but also on how a specific translation engine handles that complexity across repeated iterations.

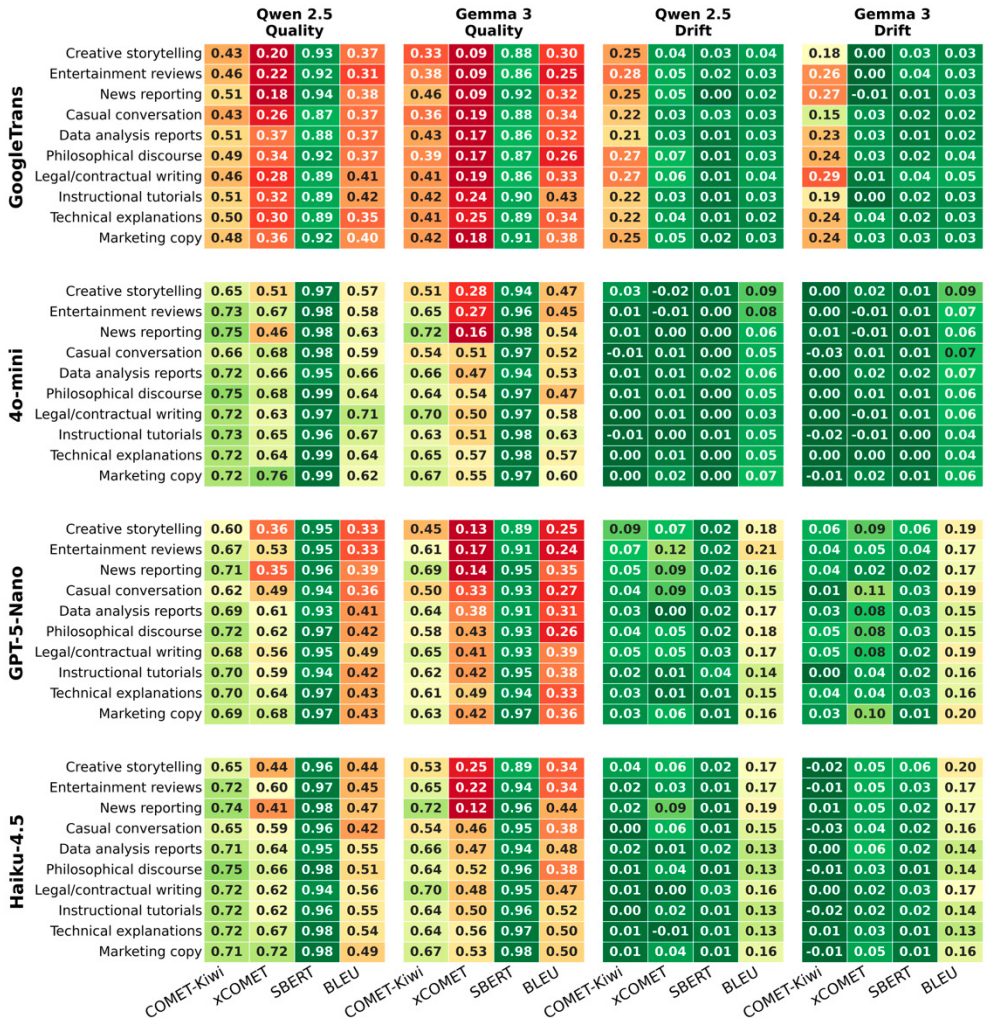

Figure 6. Genre analysis showing quality scores (Pass 10) and drift (Pass 1 to Pass 10) across all four translation systems. Each block shows Qwen 2.5 and Gemma 3 results for COMET-Kiwi, xCOMET, SBERT, and BLEU. Green indicates better preservation or lower drift; red indicates worse preservation or higher drift

5. Discussion

Our findings reveal a fundamental architectural mismatch between the objectives of translation and paraphrasing, which explains the widespread adoption of RTT despite its semantic inadequacies. Contrary to expectations, we observe a strong monotonic relationship between surface fluency and meaning retention, as evidenced by Spearman $\rho \geq 0.95$, indicating that combinations ranking higher in fluency also rank higher in semantic preservation. The sample sizes behind these correlations are modest ($n = 12$ for pipeline means, $n = 24$ for model-specific points). Therefore, at the passage level, where thousands of individual texts would be

considered, the relationship is likely to weaken, and a sizable number of high-fluency yet low-meaning outputs can still occur. Yet a substantial gap remains: in Phase 1, *4o-mini* pairs "Excellent" COMET-Kiwi scores with only "Poor" semantics, while *googletrans* achieves "Good" fluency but falls into the "Very Poor" semantic band. In other words, a superior translator narrows, but does not close, the fluency-meaning chasm: *4o-mini*'s mean semantic score (0.526) is more than double *googletrans*'s (0.222); both fall far short of acceptable preservation. This paradox arises because neural translation systems optimize for target-language naturalness rather than bidirectional semantic equivalence, resulting in an illusion of successful paraphrasing through polished but semantically diminished output.

The iterative BLEU analysis reveals important limitations in traditional RTT evaluation approaches. While BLEU scores suggest stability through high inter-pass similarity (often above 0.80), this apparent convergence obscures a substantial early loss of semantic content. RTT systems rapidly settle into simplified lexical patterns that resist further surface change, while already compromising actual meaning. This phenomenon explains why surface-level evaluation methods have supported RTT's continued use: high surface stability can coexist with significant underlying semantic degradation, rendering the approach problematic for meaning-preserving applications. A comparable surface-versus-meaning divergence appears beyond RTT: Gašpar and Knežević (2026) find that, for single-pass ChatGPT translation across four languages, high BLEU and ChrF scores can coexist with lower COMET and BLEURT scores; notably, legal text scored highest on these surface metrics yet expert evaluators rated it the least accurate domain, echoing both our surface fidelity trap and the tendency of these metrics to under-penalize legal writing.

The marked performance difference between *4o-mini* and *googletrans* shows how system architecture influences semantic preservation outcomes. *4o-mini*'s language modeling foundation provides greater semantic awareness, slowing but not preventing degradation. By contrast, *googletrans* optimizes for single-direction efficiency, producing more aggressive transformations that compound in round-trip scenarios. Even the superior system achieves only "Poor" semantic preservation, demonstrating that RTT's limitations extend beyond translation quality issues.

The three degradation regimes identified, front-loaded damage (*4o-mini*), continuous drift (*Haiku 4.5*, *GPT-5-nano*), and deterministic fixed-point convergence (*googletrans*), reflect differences in translation architecture and decoding configuration (Section 3.4). The front-loaded pattern suggests that *4o-mini* applies a consistent reformulation strategy that reaches a stable output regardless of iteration count, likely because its instruction-following behavior produces deterministic-like translations at the low temperature used (0.3). The continuous-drift pattern in newer LLM translators, particularly *GPT-5-nano* at its default temperature, indicates that each pass introduces genuinely novel variation, which is useful for diversity but harmful to fidelity. The fixed-point convergence of *googletrans* is the most mechanistically transparent: a deterministic NMT decoder maps each input to a unique output, so repeated round-trips converge to a cycle (in practice, a fixed point) within a few iterations. These regime differences mean that no single recommendation about "how many RTT passes" applies across systems; rather, the degradation dynamics must be characterized per engine before an RTT pipeline can be calibrated for a given application.

The COMET-Kiwi pass-by-pass analysis extends the regime taxonomy to a second dimension: fluency stability. While semantic scores (xCOMET) decline monotonically or plateau, the fluency trajectories of LLM-based systems oscillate. *GPT-5-nano* shows the most pronounced pattern, recovering fluency on up to 38% of transitions, versus almost never for *googletrans* (Section 4.6). This asymmetry has a straightforward mechanistic explanation: each LLM pass is a fresh stochastic generation, so a temporarily disfluent output can be "repaired" in the next pass by the model's language-modeling prior, which strongly favors fluent continuations. A deterministic NMT decoder lacks this self-correcting channel; once it converges to a fixed point, fluency is locked in along with everything else. The oscillation also deepens the fluency-meaning paradox: meaning drifts steadily downward while fluency bounces back, making it even harder for reference-free metrics like COMET-Kiwi to signal that something has gone wrong.

Across all pipelines and translation systems, passages generated by Qwen 2.5 consistently retain higher xCOMET scores than those generated by Gemma 3. Because the two sets of passages were translated under identical conditions, the gap originates in properties of the source texts themselves rather than in the translation process. A direct comparison of the two models' passages (Table A4, Appendix A) shows that they differ significantly across almost every measured dimension. Qwen 2.5 writes longer sentences (21.7 vs. 15.1 tokens on average) with more than twice the subordination, while Gemma 3 uses a richer vocabulary (MTLD 237.8 vs. 210.2; type-token ratio 0.743 vs. 0.723), shorter sentences, and pervasive markdown formatting (62% of its passages contain headers and 86% italics, against zero for Qwen 2.5). These results support a lexical rather than syntactic explanation of the preservation gap: what favors Qwen 2.5 is its more repetitive, predictable vocabulary, which translation systems reconstruct more faithfully, whereas its sentences are in fact more complex than Gemma 3's. Gemma 3's markdown is an additional, metric-specific contributor: as Section 5.2 shows, xCOMET flags markdown tokens as critical error spans, so part of the gap reflects formatting penalization rather than semantic loss. Within both models, passage length and punctuation density correlate negatively with preservation (Spearman $\rho \approx -0.6$ and $\rho \approx -0.4$, respectively; per-passage Pass-10 xCOMET on the *4o-mini* EN → BG → EN route, n = 100 per model). Passage length is balanced between the models by design and cannot account for the between-model gap; comma density, however, does differ between the models (Table A4) and forms part of the stylistic contrast described above. Fully decomposing the causal contribution of each property would require controlled manipulation of individual text features, which we leave to future work. Regardless of its cause, this finding has a practical implication: RTT-based paraphrase quality depends not only on the translation system but also on the stylistic characteristics of the input text, introducing a source-side confound that should be controlled for in future evaluations.

A related limitation concerns the nature of the source material itself. All 200 passages are LLM-generated, yet our conclusions speak to meaning-sensitive applications that typically involve human-written text. LLM-generated prose tends to be more regular than human writing (syntactically simpler, lexically more predictable, and more formulaic), properties that plausibly make it easier for translation systems to reconstruct. RTT degradation on human-written text could therefore be more severe than reported here; conversely, the redundancy and

contextual richness of human prose might buffer meaning loss. Our design cannot distinguish between these possibilities, and the absolute scores reported here should be generalized to human-written domains with caution. Replication on human-written corpora is identified as a direction for future work (see Future research). A further limitation is that each passage–pipeline–system chain was executed once: for the stochastically sampled LLM translators, each per-passage trajectory is a single draw from a distribution of possible round-trips, although every reported mean aggregates 100 such trajectories per source model.

5.1. Metric sensitivity and the translation-paraphrase mismatch

The genre-level analysis reveals a deeper methodological issue: xCOMET is a translation-quality metric, not a paraphrase-quality metric. It was trained on human judgments of translation adequacy where any deviation from the reference is potentially an error, and its error-span detection is designed to flag where the hypothesis diverges from the reference. This design creates an asymmetric sensitivity pattern when applied to RTT outputs, which are better understood as paraphrases than as translations.

Sample-level inspection of the output data clearly reveals this asymmetry. In creative storytelling, synonym substitution (e.g., "hacked through" becoming "slicing through," "swallowed by strangler figs" becoming "engulfed by intertwined fig trees") and narrative restructuring are legitimate paraphrastic variations; the meaning is preserved even though the surface form changes substantially. Yet xCOMET has no mechanism to distinguish a creative synonym that preserves intent from a mistranslation that distorts meaning. In contrast, in news reporting, xCOMET flags markdown formatting characters and proper nouns as critical error spans. One Gemma 3 news sample under *4o-mini* receives an xCOMET score of just 0.014 at Pass 1, despite the actual text being barely changed ("The findings indicate" becoming "The results show," "some dating back over a century" becoming "some of which date back over a century"). The near-zero score is driven almost entirely by formatting artifacts rather than semantic failures.

The opposite distortion appears in legal and technical texts. A legal clause translated by *4o-mini* scores 0.817 at Pass 10, yet close reading reveals that "affiliates" has been systematically replaced by "subsidiaries," terms with distinct legal definitions (a subsidiary is a type of affiliate, but not vice versa). In a contractual context, these changes could narrow the scope of liability protection. xCOMET rewards this output because the formulaic sentence structure and standardized vocabulary are well preserved, and its span-level detection does not weight domain-specific semantic criticality. Meanwhile, *googletrans* on the same passage produces genuine errors: "punitive, or exemplary damages" becoming "penal or excerpts from the bonds," and "agents, agents, agents" as a triple repetition, which xCOMET correctly identifies.

The SBERT data independently support this interpretation. Because SBERT encodes sentences into an embedding space where synonyms and paraphrases map to nearby vectors, it should show a comparable genre gap if the xCOMET gap were entirely about meaning loss, a comparison whose force is tempered by SBERT's compressed dynamic range (Section 4.7). Instead, SBERT scores are far more uniform across genres (all above 0.85), suggesting that the xCOMET genre gap is at least partly a measurement artifact of applying a translation metric to a paraphrasing task. The pattern is therefore that xCOMET over-penalizes genres

that tolerate lexical variation (creative storytelling, entertainment reviews, news reporting) while under-penalizing genres where small changes carry disproportionate semantic weight (legal writing, technical documentation). This asymmetry does not invalidate xCOMET as an evaluation tool (it remains the most discriminating metric in our battery), but it does mean that absolute scores should be interpreted with awareness of the genre being evaluated, and that the genre rankings in our analysis likely exaggerate the true semantic gap between creative and formulaic text types.

5.2. Manual validation of score-meaning alignment

To test whether these sample-level observations are systematic rather than anecdotal, we conducted a structured manual validation check on 54 original-output pairs, stratified by translation system (all four), pipeline (both Bulgarian-centered routes), genre, and score band. The sample includes same-passage panels evaluated across all four systems; representative examples are collected in Appendix B. The inspection confirmed both directions of the asymmetry and revealed its system dependence. At the low end of the scale, scores carry little information for LLM outputs: on one news passage, a near-verbatim *4o-mini* round-trip (0.012), a *Haiku 4.5* output whose only substantive error is a single direction flip ("underestimate" → "overestimate," 0.012), and a *googletrans* output with genuine semantic destruction ("a crucial system regulating global climate" → "a crucial global air-conditioning system," −0.014) are statistically indistinguishable, with the corresponding error spans concentrated on markdown tokens, proper nouns, and sub-word fragments. At the high end, the asymmetry inverts: ten of thirteen legal outputs scoring above 0.72 carried legally consequential substitutions, including the addition of "direct" to an exclusion-of-damages list (0.724), "harassment" → "violence" and the deletion of "including but not limited to" (0.895), and "eligible non-profit organizations" → "any organization working in the interests of society" (0.888); the silent omission of "officers" from a liability enumeration recurred in four of four LLM round-trips of the same clause. The check also confirmed that low scores are not uniformly spurious: every inspected negative-scoring *googletrans* output was semantically incoherent, and several LLM outputs contained real entity and stance errors correctly reflected in low scores. We therefore characterize xCOMET's failure mode, across the cases we inspected, not as random noise but as a consistent conflation of benign surface variation with meaning-bearing change: compressed by surface artifacts at the bottom of the scale and blind to domain-critical substitutions at the top. Importantly, averaged over passages, the metric's system ranking agrees with manual judgment, so the corpus-level comparisons in this study remain valid; it is per-sample and per-genre interpretation that requires caution. A full-scale human evaluation with independent annotators remains outside the scope of this study and is identified as future work.

These findings point to a broader challenge for RTT evaluation: no existing metric fully captures what matters for paraphrase quality. BLEU misses semantic drift, SBERT compresses the range, COMET-Kiwi does not measure fidelity to the original, and xCOMET applies a translation-adequacy lens that penalizes legitimate variation. The fragility of COMET-family metrics under degenerate inputs is not specific to paraphrase evaluation: Zan et al. (2025) report that COMET-Kiwi and xCOMET can assign high scores to copied or wrong-language

outputs and therefore exclude such cases before computing these metrics, independent evidence that they do not reliably penalize adequacy failures at the extremes. A purpose-built paraphrase-fidelity metric that tolerates synonym substitution and structural reorganization while remaining sensitive to factual, entity-level, and domain-specific distortions remains an open research gap.

6. Conclusions

This evaluation shows that RTT consistently fails to preserve meaning while producing fluent surface variations. Across all pipeline configurations, semantic-preservation scores remain low: *4o-mini* 0.526 and *googletrans* 0.222 (xCOMET). In contrast, surface-quality metrics such as COMET-Kiwi reach deceptively high values (≥ 0.648 in Phase 1, and 0.59–0.72 for the LLM-based systems in Phase 2), creating an illusion of success that has fueled the widespread yet unwarranted adoption of RTT.

The degradation is system-dependent: the four translators follow three distinct regimes (front-loaded damage, continuous drift, and deterministic fixed-point convergence), each with different implications for the number of RTT passes that can be tolerated before meaning is irretrievably lost. Crucially, fluency does not track meaning loss; LLM-based systems can recover surface quality between passes even as semantic content continues to erode, while the deterministic NMT baseline locks into a stable but depleted output. These dynamics, modulated by multi-hop configurations and source-text variability, mean that no single iteration count or system choice is universally safe.

The surface fidelity trap identified carries a direct practical warning: *googletrans* can retain more n-gram overlap with the original (higher BLEU) than *GPT-5-nano* while destroying far more meaning (lower xCOMET). The genre-level analysis and sample-level inspection reveal a complementary limitation of the evaluation framework itself. xCOMET, as a translation-quality metric, over-penalizes genres that tolerate lexical variation, flagging creative synonyms and even markdown formatting as critical errors, while potentially under-penalizing genres such as legal writing, where subtle term substitutions can alter contractual meaning despite preserving formulaic structure. The more uniform SBERT scores across genres are consistent with the xCOMET genre gap partly reflecting metric sensitivity rather than true semantic loss. This finding does not diminish the overall conclusion that RTT degrades meaning, but it does caution against interpreting xCOMET scores as genre-comparable absolute measures of fidelity.

Given these findings, RTT should be avoided in workflows that require semantic accuracy, including academic paraphrasing, legal documentation, medical text adaptation, technical writing, and related domains. Convenience and accessibility cannot outweigh the empirically demonstrated failure to maintain meaning at professional standards. The fact that *4o-mini* still achieves only poor semantic preservation shows that RTT limitations arise from a structural mismatch between translation objectives and meaning retention, rather than from temporary deficiencies that a larger model could easily fix.

The risk in these domains is not visible degradation that a proofreader would catch, but precisely the failure mode documented in Sections 5.1 and 5.2: fluent output in which a single substituted term silently changes scope or obligation. In legal text, replacing “affiliates” with

"subsidiaries" or "without recourse" with "without the right to appeal" narrows contractual protection while leaving the clause grammatically impeccable; in medical or academic text, the analogous shifts (a dropped qualifier, a removed hedge, an entity swap) can alter clinical meaning or misattribute claims. Because fluency metrics and casual reading both fail to flag such changes, deploying RTT in these domains would require term-level expert review of every output, which negates the very convenience that motivates RTT in the first place.

For researchers and practitioners who nonetheless employ RTT-based pipelines, our findings translate into five concrete recommendations. (1) Do not use RTT for meaning-critical text; for other applications, treat it as a lossy transformation whose damage must be measured rather than assumed. (2) Characterize the translation engine's degradation regime before deployment: for front-loaded systems such as *4o-mini*, a single pass already captures the full semantic cost; for continuously drifting systems such as *Haiku 4.5* and *GPT-5-nano*, each additional pass compounds the loss; and for deterministic NMT systems, passes beyond the fixed point add computation without textual change. (3) Never assess RTT quality with surface metrics alone: the surface fidelity trap means high BLEU can coexist with severe meaning loss, so any surface metric should be paired with a semantic metric such as xCOMET. (4) Exploit cheap early signals: in our experiments, a low Sørensen-Dice coefficient after the first pass tracked the system-level preservation ranking and can serve as an inexpensive early warning before expensive neural evaluation, though this rests on system-level rank agreement rather than passage-level correlation. (5) Interpret per-sample and per-genre xCOMET scores with the caution documented in Section 5.2, relying on them primarily for corpus-level comparisons.

7. Future research

Several directions follow from these findings. On the evaluation side, a paraphrase-specific metric that tolerates synonym substitution while remaining sensitive to factual and domain-specific distortions would address the asymmetry we observed in xCOMET's genre-level behavior. Such a metric could build on xCOMET's error-span architecture but incorporate domain-aware severity weighting. Complementing metric development, a human evaluation with independent annotators would validate the score-meaning misalignments identified in Section 5.2, and replication on human-written corpora would establish whether our findings generalize beyond LLM-generated source texts. On the systems side, a larger genre-stratified benchmark, additional translation engines (e.g., DeepL), larger or state-of-the-art models with higher parameter counts, and a controlled study of decoding parameters, particularly temperature, would clarify whether the degradation regimes identified here generalize beyond the systems and configurations tested.

More broadly, dual-objective architectures that jointly optimize for meaning preservation and stylistic control could move RTT from an unreliable heuristic toward a principled paraphrasing tool. A related direction is the interaction between RTT and AI-content detectability: as shown in our concurrent work on stylometric attribution (Kopanov & Atanasova, 2026), RTT can obscure model-specific linguistic signatures, which carries implications for academic integrity and content verification that warrant dedicated investigation.

Acknowledgements

This research is supported by the National Science Fund of Bulgaria under Contract No. КП-06-Н85/16.

Author contributions

Tatiana Atanasova conceptualized the study, established the theoretical framework, supervised the research process, and reviewed and enhanced the manuscript. Kalin Kopanov designed the methodology, developed the experimental framework and all supporting software (translation pipelines, evaluation scripts, and metric computation), conducted all experiments, performed the statistical analysis, created the visualizations, and drafted the initial manuscript. The synthetic dataset used in this study was generated in previous collaborative work by both authors (Kopanov & Atanasova, 2025). Both authors jointly interpreted the results and formulated the conclusions. All authors have read and agreed to the published version of the manuscript.

Disclosure statement

The authors declare that they have no competing financial, professional, or personal interests from other parties.

References

Amara, K., Sevastjanova, R., & El-Assady, M. (2024). Challenges and opportunities in text generation explainability. In L. Longo, S. Lapuschkin, & C. Seifert (Eds.), *Communications in computer and information science*: *Vol. 2153. Explainable artificial intelligence* (*xAI 2024*) (pp. 244–264). Springer. https://doi.org/10.1007/978-3-031-63787-2_13

Bannard, C., & Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 597–604). Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219914

Barzilay, R., & McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 50–57). Association for Computational Linguistics. https://doi.org/10.3115/1073012.1073020

Cao, J., Li, M., Li, Y., Wen, M., Cheung, S.-C., & Chen, H. (2022). SemMT: A semantic-based testing approach for machine translation systems. *ACM Transactions on Software Engineering and Methodology*, *31*(2), 1–36. https://doi.org/10.1145/3490488

Corbeil, J.-P., & Abdi Ghavidel, H. (2021). Assessing the eligibility of backtranslated samples based on semantic similarity for the paraphrase identification task. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (*RANLP 2021*) (pp. 301–308). INCOMA Ltd. https://doi.org/10.26615/978-954-452-072-4_035

Crone, N., Power, A., & Weldon, J. (2021). *Quality estimation using round-trip translation with sentence embeddings*. arXiv. https://doi.org/10.48550/arXiv.2111.00554

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, & L. Specia (Eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation* (pp. 376–380). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3348

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 489–500). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1045

Gašpar, A., & Knežević, M. (2026). Cross-lingual and cross-domain evaluation of ChatGPT's translation performance. In D. Vasić & E. Brajković (Eds.), *Communications in computer and information science*: *Vol. 2609*. *Digital transformation in education and artificial intelligence application*: *MoStart 2025* (pp. 196–211). Springer. https://doi.org/10.1007/978-3-032-02801-3_13

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. T. (2024a). *XCOMET-XXL* [Computer software]. Hugging Face. https://huggingface.co/Unbabel/XCOMET-XXL

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. T. (2024b). xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, *12*, 979–995. https://doi.org/10.1162/tacl_a_00683

Han, S. (2025). *googletrans* (Version 4.0.2) [Computer software]. GitHub. https://github.com/ssut/py-googletrans

Kopanov, K., & Atanasova, T. (2025). A comparative pattern analysis of Qwen 2.5 and Gemma 3 text generation. *WSEAS Transactions on Information Science and Applications*, *22*, 604–615. https://doi.org/10.37394/23209.2025.22.50

Kopanov, K., & Atanasova, T. (2026). Degradation of stylometric attribution accuracy for AI-generated text. *Engineering Proceedings, 150*(1), Article 98. https://doi.org/10.3390/engproc2026150098

Mallinson, J., Sennrich, R., & Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (vol. 1, pp. 881–893). Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1083

Mondal, S. K., Zhang, H., Kabir, H. M. D., Ni, K., & Dai, H.-N. (2023). Machine translation and its evaluation: A study. *Artificial Intelligence Review*, *56*, 10137–10226. https://doi.org/10.1007/s10462-023-10423-5

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Quelle, D., Cheng, C. Y., Bovet, A., & Hale, S. A. (2025). Lost in translation: Using global fact-checks to measure multilingual misinformation prevalence, spread, and evolution. *EPJ Data Science*, *14*, Article 22. https://doi.org/10.1140/epjds/s13688-025-00520-6

Quirk, C., Brockett, C., & Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In D. Lin & D. Wu (Eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 142–149). Association for Computational Linguistics.

Rei, R., Guerreiro, N. M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., de Souza, J. G. C., & Martins, A. F. T. (2023). Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation* (*WMT 2023*), (pp. 841–848). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.73

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J. G. C., Glushkova, T., Alves, D., Coheur, L., Lavie, A., & Martins, A. F. T. (2022). *wmt22-cometkiwi-da* [Computer software]. Hugging Face. https://huggingface.co/Unbabel/wmt22-cometkiwi-da

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*) (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (vol. 1, pp. 86–96). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1009

Thompson, B., & Post, M. (2020). Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation* (pp. 561–570). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.wmt-1.67

Wieting, J., & Gimpel, K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (vol. 1, pp. 451–462). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1042

Zan, C., Ding, L., Shen, L., Zhan, Y., Yang, X., & Liu, W. (2025). Building accurate translation-tailored large language models with language-aware instruction tuning. *Frontiers of Information Technology & Electronic Engineering*, *26*(8), 1341–1355. https://doi.org/10.1631/FITEE.2400458

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations* (*ICLR 2020*).

Zhuo, T. Y., Xu, Q., He, X., & Cohn, T. (2023). Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics*: *ACL 2023* (pp. 319–337). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.22

Zou, W., Yang, S., Bao, Y., Huang, S., Chen, J., & Cheng, S. (2025). Trans-zero: Self-play incentivizes large language models for multilingual translation without parallel data. *Findings of the Association for Computational Linguistics*: *ACL 2025* (pp. 12337–12347). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.637

APPENDIX

A. Statistical detail

Tables A1–A4 report bootstrap confidence intervals for all Pass-10 means, the complete pairwise system comparisons summarized in Section 4.6, and the source-text feature analysis referenced in Section 5.

Table A1. Phase 1 Pass-10 xCOMET and COMET-Kiwi means with 95% confidence intervals

Metric	System	Pipeline	Qwen 2.5	Gemma 3
xCOMET	googletrans	EN-BG-EN	0.331 [0.305, 0.357]	0.199 [0.177, 0.222]
xCOMET	googletrans	EN-RU-EN	0.290 [0.266, 0.313]	0.167 [0.147, 0.187]
xCOMET	googletrans	EN-ZH-EN	0.390 [0.358, 0.419]	0.217 [0.191, 0.243]
xCOMET	googletrans	EN-BG-RU-EN	0.235 [0.213, 0.257]	0.135 [0.120, 0.150]
xCOMET	googletrans	EN-BG-ZH-EN	0.224 [0.202, 0.247]	0.141 [0.124, 0.160]
xCOMET	googletrans	EN-RU-ZH-EN	0.205 [0.184, 0.228]	0.132 [0.118, 0.147]
xCOMET	4o-mini	EN-BG-EN	0.640 [0.607, 0.673]	0.448 [0.413, 0.482]
xCOMET	4o-mini	EN-RU-EN	0.646 [0.611, 0.679]	0.445 [0.409, 0.481]
xCOMET	4o-mini	EN-ZH-EN	0.623 [0.593, 0.652]	0.425 [0.389, 0.460]
xCOMET	4o-mini	EN-BG-RU-EN	0.627 [0.594, 0.659]	0.425 [0.389, 0.459]
xCOMET	4o-mini	EN-BG-ZH-EN	0.615 [0.582, 0.648]	0.403 [0.368, 0.438]
xCOMET	4o-mini	EN-RU-ZH-EN	0.616 [0.582, 0.649]	0.398 [0.361, 0.435]
COMET-Kiwi	googletrans	EN-BG-EN	0.776 [0.767, 0.786]	0.737 [0.723, 0.749]
COMET-Kiwi	googletrans	EN-RU-EN	0.732 [0.711, 0.745]	0.691 [0.673, 0.703]
COMET-Kiwi	googletrans	EN-ZH-EN	0.794 [0.773, 0.808]	0.750 [0.728, 0.766]
COMET-Kiwi	googletrans	EN-BG-RU-EN	0.706 [0.695, 0.717]	0.659 [0.648, 0.671]
COMET-Kiwi	googletrans	EN-BG-ZH-EN	0.723 [0.706, 0.737]	0.680 [0.665, 0.694]
COMET-Kiwi	googletrans	EN-RU-ZH-EN	0.692 [0.670, 0.709]	0.648 [0.628, 0.664]
COMET-Kiwi	4o-mini	EN-BG-EN	0.884 [0.880, 0.887]	0.878 [0.874, 0.883]
COMET-Kiwi	4o-mini	EN-RU-EN	0.884 [0.880, 0.887]	0.879 [0.875, 0.883]
COMET-Kiwi	4o-mini	EN-ZH-EN	0.877 [0.873, 0.881]	0.875 [0.871, 0.879]
COMET-Kiwi	4o-mini	EN-BG-RU-EN	0.880 [0.873, 0.884]	0.876 [0.872, 0.880]
COMET-Kiwi	4o-mini	EN-BG-ZH-EN	0.873 [0.869, 0.877]	0.869 [0.864, 0.874]
COMET-Kiwi	4o-mini	EN-RU-ZH-EN	0.873 [0.869, 0.877]	0.868 [0.863, 0.873]

Note: Values in brackets are 95% BCa bootstrap confidence intervals (10,000 resamples; n = 100 passages per source model).

Table A2. Phase 2 Pass-10 means with 95% confidence intervals

System	Pipeline	Metric	Qwen 2.5	Gemma 3
4o-mini	EN-BG-EN	xCOMET	0.640 [0.607, 0.673]	0.448 [0.413, 0.482]
4o-mini	EN-BG-EN	COMET-Kiwi	0.717 [0.705, 0.726]	0.638 [0.620, 0.654]
4o-mini	EN-BG-EN	SBERT	0.979 [0.971, 0.983]	0.968 [0.959, 0.973]
4o-mini	EN-BG-EN	BLEU	0.656 [0.642, 0.670]	0.560 [0.543, 0.577]
4o-mini	EN-BG-RU-EN	xCOMET	0.627 [0.594, 0.659]	0.425 [0.389, 0.459]
4o-mini	EN-BG-RU-EN	COMET-Kiwi	0.715 [0.700, 0.725]	0.637 [0.620, 0.653]
4o-mini	EN-BG-RU-EN	SBERT	0.972 [0.963, 0.977]	0.963 [0.954, 0.968]
4o-mini	EN-BG-RU-EN	BLEU	0.605 [0.585, 0.622]	0.512 [0.495, 0.531]
Haiku 4.5	EN-BG-EN	xCOMET	0.591 [0.556, 0.623]	0.421 [0.385, 0.456]
Haiku 4.5	EN-BG-EN	COMET-Kiwi	0.709 [0.698, 0.718]	0.639 [0.621, 0.654]
Haiku 4.5	EN-BG-EN	SBERT	0.968 [0.959, 0.974]	0.952 [0.942, 0.959]
Haiku 4.5	EN-BG-EN	BLEU	0.524 [0.508, 0.540]	0.457 [0.438, 0.477]
Haiku 4.5	EN-BG-RU-EN	xCOMET	0.599 [0.563, 0.632]	0.401 [0.366, 0.436]
Haiku 4.5	EN-BG-RU-EN	COMET-Kiwi	0.711 [0.700, 0.720]	0.641 [0.624, 0.656]
Haiku 4.5	EN-BG-RU-EN	SBERT	0.964 [0.956, 0.970]	0.947 [0.935, 0.955]
Haiku 4.5	EN-BG-RU-EN	BLEU	0.473 [0.454, 0.489]	0.413 [0.394, 0.431]
GPT-5-nano	EN-BG-EN	xCOMET	0.559 [0.523, 0.593]	0.346 [0.310, 0.381]
GPT-5-nano	EN-BG-EN	COMET-Kiwi	0.685 [0.673, 0.696]	0.603 [0.585, 0.620]
GPT-5-nano	EN-BG-EN	SBERT	0.956 [0.947, 0.963]	0.937 [0.927, 0.945]
GPT-5-nano	EN-BG-EN	BLEU	0.431 [0.416, 0.446]	0.339 [0.323, 0.357]
GPT-5-nano	EN-BG-RU-EN	xCOMET	0.525 [0.489, 0.559]	0.322 [0.291, 0.353]
GPT-5-nano	EN-BG-RU-EN	COMET-Kiwi	0.671 [0.658, 0.682]	0.592 [0.573, 0.610]
GPT-5-nano	EN-BG-RU-EN	SBERT	0.949 [0.937, 0.956]	0.925 [0.915, 0.933]
GPT-5-nano	EN-BG-RU-EN	BLEU	0.371 [0.357, 0.386]	0.288 [0.273, 0.303]
googletrans	EN-BG-EN	xCOMET	0.331 [0.305, 0.357]	0.199 [0.177, 0.222]
googletrans	EN-BG-EN	COMET-Kiwi	0.511 [0.497, 0.525]	0.425 [0.413, 0.439]
googletrans	EN-BG-EN	SBERT	0.925 [0.911, 0.935]	0.913 [0.900, 0.924]
googletrans	EN-BG-EN	BLEU	0.454 [0.440, 0.469]	0.388 [0.372, 0.405]
googletrans	EN-BG-RU-EN	xCOMET	0.235 [0.213, 0.257]	0.135 [0.120, 0.150]
googletrans	EN-BG-RU-EN	COMET-Kiwi	0.445 [0.434, 0.457]	0.380 [0.372, 0.389]
googletrans	EN-BG-RU-EN	SBERT	0.884 [0.872, 0.895]	0.854 [0.838, 0.867]
googletrans	EN-BG-RU-EN	BLEU	0.293 [0.282, 0.303]	0.266 [0.253, 0.280]

Note: Values in brackets are 95% BCa bootstrap confidence intervals (10,000 resamples; n = 100 passages per source model).

Table A3. Pairwise system comparisons on Pass-10 xCOMET (Wilcoxon signed-rank tests, paired by passage)

Pipeline	Source model	System A	System B	Mean diff. (A−B)	r	Adjusted p
EN-BG-EN	Qwen 2.5	4o-mini	Haiku 4.5	0.049	0.59	< 0.001
EN-BG-EN	Qwen 2.5	4o-mini	GPT-5-nano	0.082	0.73	< 0.001
EN-BG-EN	Qwen 2.5	4o-mini	googletrans	0.310	1.00	< 0.001
EN-BG-EN	Qwen 2.5	Haiku 4.5	GPT-5-nano	0.033	0.36	= 0.002
EN-BG-EN	Qwen 2.5	Haiku 4.5	googletrans	0.260	1.00	< 0.001
EN-BG-EN	Qwen 2.5	GPT-5-nano	googletrans	0.228	0.99	< 0.001
EN-BG-EN	Gemma 3	4o-mini	Haiku 4.5	0.027	0.43	< 0.001
EN-BG-EN	Gemma 3	4o-mini	GPT-5-nano	0.102	0.77	< 0.001
EN-BG-EN	Gemma 3	4o-mini	googletrans	0.249	1.00	< 0.001
EN-BG-EN	Gemma 3	Haiku 4.5	GPT-5-nano	0.075	0.65	< 0.001
EN-BG-EN	Gemma 3	Haiku 4.5	googletrans	0.222	0.99	< 0.001
EN-BG-EN	Gemma 3	GPT-5-nano	googletrans	0.147	0.95	< 0.001
EN-BG-RU-EN	Qwen 2.5	4o-mini	Haiku 4.5	0.028	0.31	= 0.008
EN-BG-RU-EN	Qwen 2.5	4o-mini	GPT-5-nano	0.102	0.77	< 0.001
EN-BG-RU-EN	Qwen 2.5	4o-mini	googletrans	0.392	1.00	< 0.001
EN-BG-RU-EN	Qwen 2.5	Haiku 4.5	GPT-5-nano	0.074	0.64	< 0.001
EN-BG-RU-EN	Qwen 2.5	Haiku 4.5	googletrans	0.364	1.00	< 0.001
EN-BG-RU-EN	Qwen 2.5	GPT-5-nano	googletrans	0.290	1.00	< 0.001
EN-BG-RU-EN	Gemma 3	4o-mini	Haiku 4.5	0.024	0.33	= 0.004
EN-BG-RU-EN	Gemma 3	4o-mini	GPT-5-nano	0.103	0.77	< 0.001
EN-BG-RU-EN	Gemma 3	4o-mini	googletrans	0.290	0.98	< 0.001
EN-BG-RU-EN	Gemma 3	Haiku 4.5	GPT-5-nano	0.080	0.67	< 0.001
EN-BG-RU-EN	Gemma 3	Haiku 4.5	googletrans	0.267	0.97	< 0.001
EN-BG-RU-EN	Gemma 3	GPT-5-nano	googletrans	0.187	0.96	< 0.001

Note: r is the matched-pairs rank-biserial correlation. p-values are adjusted with the Holm–Bonferroni procedure within each pipeline × source-model family of six comparisons (n = 100 passages per source model). Positive mean differences indicate higher xCOMET for System A.

Table A4. Lexical, syntactic, and formatting features of the Qwen 2.5 and Gemma 3 source passages

Feature	Qwen 2.5	Gemma 3	p
Passage length (tokens)	225.3	225.0	= 0.739
Mean sentence length (tokens)	21.7	15.1	< 0.001
Mean word length (characters)	5.44	5.53	= 0.005
Type-token ratio	0.723	0.743	< 0.001
Lexical diversity (MTLD)	210.2	237.8	< 0.001
Subordinating conjunctions per sentence	0.65	0.27	< 0.001
Commas per sentence	1.42	1.11	< 0.001

End of Table A4

Feature	Qwen 2.5	Gemma 3	p
Passages with markdown headers (%)	0	62	< 0.001
Passages with markdown bold (%)	9	38	< 0.001
Passages with markdown italics (%)	0	86	< 0.001

Note: Means over 100 passages per model; p-values from Wilcoxon signed-rank tests paired by generation prompt. Lexical diversity is measured with MTLD (factor threshold 0.72); subordination counts a fixed list of subordinating conjunctions and relativizers per sentence; markdown features are detected before sentence segmentation. p-values are unadjusted for multiple comparisons.

B. Annotated examples from the manual validation check

Table B1 presents representative original → output pairs from the 54-pair manual validation check described in Section 5.2, grouped by the three patterns it revealed: xCOMET over-penalizes faithful paraphrase, under-penalizes meaning-changing edits in legal text, and correctly penalizes genuine degradation.

Table B1. Representative examples of score–meaning misalignment, grouped by pattern

Pattern	Genre / system / xCOMET	Decisive change (original → Pass-10 output)	Why the score is wrong (or right)
Over-penalized	News 4o-mini, EN-BG-EN 0.012	Near-verbatim output; every "critical" error span falls on markdown tokens, proper nouns, and the transliterated name "Óskarsdóttir" → "Oscarsdottir"	Meaning preserved; the near-zero score is driven by formatting, not semantics
Over-penalized	News Haiku 4.5, EN-BG-EN 0.012	"underestimate" → "overestimate"; "trap warmer waters" → "suppress warm water"; rest near-verbatim	One genuine direction flip in otherwise faithful text, indistinguishable in score from total destruction
Over-penalized	News Haiku 4.5, EN-BG-EN 0.157	"El-Khoury" → "El-Houri"; "suggest" → "testifying to"	Meaning preserved; a transliterated name and one strengthened hedge drive a near-total penalty
Over-penalized	Creative 4o-mini, EN-BG-EN 0.403	"shimmering orb" → "shimmering sphere"; "echoed off the cliffs" → "echoed against the rocks"	Legitimate synonym substitution penalized as if it were error
Under-penalized	Legal 4o-mini, EN-BG-EN 0.870	"proprietary business data" → "trade secrets"; "without recourse" → "without the right to appeal"	Contractual scope narrowed while the score stays near "Good"
Under-penalized	Legal Haiku 4.5, EN-BG-EN 0.724	"direct" added to the excluded-damages list; "officers" dropped from "affiliates, officers, employees, agents"	Liability exclusion broadened and a protected party removed
Under-penalized	Legal Haiku 4.5, EN-BG-EN 0.895	"harassment" → "violence"; "including but not limited to" deleted	Conduct clause and open-ended enumeration altered at near-"Excellent"

End of Table B1

Pattern	Genre / system / xCOMET	Decisive change (original → Pass-10 output)	Why the score is wrong (or right)
Under-penalized	Legal Haiku 4.5, EN-BG-RU-EN 0.888	"eligible non-profit organizations" → "any organization working in the interests of society"	Eligibility constraint dissolved
Under-penalized	Legal GPT-5-nano, EN-BG-EN 0.764	"officers" dropped from "affiliates, officers, employees, agents"; "willful misconduct" → "intentional acts"	A protected party removed at a high score; the same omission recurs in all four LLM round-trips of this clause
Correctly penalized	Entertainment 4o-mini, EN-BG-EN 0.157	"ostensibly set in 1888" → "clearly set … in 1888"	Reviewer's skeptical stance inverted to a confident one; a real change, though the penalty is severe for a single flip
Correctly penalized	News GPT-5-nano, EN-BG-RU-EN 0.020	"engage targets" → "identify targets"; "loitering munitions" → "illumination rounds"; "General Evelyn Hayes" → "Evan Hayes"	Domain-critical military meaning altered fluently
Correctly penalized	Creative googletrans, EN-BG-RU-EN −0.019	"dense canopy" → "the godfather"; "The temple loomed" → "the temple was a problem"	Genuine semantic destruction; negative score warranted
Correctly penalized	News googletrans, EN-BG-EN 0.059	"the cargo ship Ever Given ran aground" → "a cargo ship was given, Ran was awarded"; "delays" → "entertainment"	Entities and meaning destroyed

Note: Excerpts are abbreviated to the decisive fragments; values are per-sample Pass-10 xCOMET scores. "Over-penalized" – meaning preserved but scored low; "under-penalized" – meaning changed but scored high; "correctly penalized" – meaning genuinely altered or destroyed. The over- and under-penalized rows are the metric failures; the correctly-penalized rows show that low scores are not uniformly spurious.