Semantic Fidelity Versus Literary Quality: A Construct Validity Study of Neural Machine Translation Metrics

Dmytro Chaplynskyi[1,2,3] **Maria Shvedova**[4,5] **Ivan Kulynych**[6] **Lesia Ivashkevych**[7]

[1]Ukrainian Catholic University [2]I. Krypiakevych Institute of Ukrainian Studies [3]lang-uk
[4]NTU "Kharkiv Polytechnic Institute" [5]University of Jena
[6]Grammarly [7]NTUU "Igor Sikorsky Kyiv Polytechnic Institute"

chaplynskyi.dmytro@ucu.edu.ua, corpus.textiv@gmail.com,
ivankulynych4@gmail.com, lesia.ivashkevych@gmail.com

Abstract

Automatic machine translation metrics are the de facto standard for evaluating translation quality. Yet, it remains unclear what they actually measure. We investigate this question using a unique multilingual corpus: seven human Ukrainian translations of George Orwell's *Animal Farm*, alongside three architecturally distinct AI systems (GPT-5.2, DeepL, and Lapa, a Ukrainian-tuned LLM). Across seven neural metrics, four reference-free and three reference-based, all three AI translations rank at the top. However, stylometric analysis exposes that these same AI translations are not as lexically rich as human ones (-18% MTLD), underuse Ukrainian particles (up to $2\times$ fewer) and diminutive morphology ($2.6\times$ fewer), and converge on near-identical outputs (LaBSE pairwise similarity 0.941 vs. 0.711 for human pairs). A controlled LLM-as-a-judge experiment demonstrates a clear preference reversal: when the English source is visible, AI ranks first; when it is hidden and the judge evaluates literary quality alone, humans rise to the top and AI falls to the lower ranks. Human evaluation (1,034 pairwise judgments) is balanced across both patterns. We argue that current MT metrics reward semantic fidelity and surface fluency — properties optimized by AI systems — while failing to capture the lexical richness, cultural adaptation, and stylistic voice that characterize skilled literary translation.

1 Introduction

Neural MT metrics — COMET, COMETKiwi, XCOMET, MetricX — are trained on human judgments from news and general-domain settings. They reward closeness to the source and surface fluency. Whether this transfers to literary translation, where voice, style, and cultural adaptation matter, is an open question since literary translation differs structurally from metric training regimes. The quality of literary translation depends largely on how the text reads in the target language — on lexical richness, pragmatic nuance, rhythm, and culturally specific expression. A translation can be semantically faithful yet stylistically thin.

To test construct validity, we derive four falsifiable predictions from the hypothesis that neural metrics primarily reward semantic fidelity:

1. **Metric dominance.** Systems optimized for fidelity — including modern AI systems — should rank highest across neural metrics.

2. **Convergence.** If metrics reward proximity to the source, metric-preferred systems should be both closer to the English original and closer to one another in semantic space.

3. **Stylistic compression.** Features associated with Ukrainian literary expressiveness — lexical diversity, particles, diminutive morphology, and stylistic dispersion — should not correlate with metric rankings and may be reduced in metric-preferred systems.

4. **Preference reversal.** If fidelity drives metric scores, then removing access to the source during evaluation should alter system rankings. When evaluators judge only the Ukrainian text as literary prose, systems optimized for fidelity should lose their advantage.

We test these predictions using seven neural MT metrics (four reference-free and three reference-based), cross-lingual embedding similarity (LaBSE), multi-dimensional stylometric analysis, and two controlled LLM-as-a-judge experiments that differ only in source visibility. We further compare these results to human pairwise evaluation aggregated with TrueSkill.

2 Related Work

Automatic MT evaluation has shifted decisively from lexical-overlap metrics such as BLEU toward

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 67–79

neural approaches — including COMET (Rei et al., 2020), COMETKiwi (Rei et al., 2022), XCOMET (Guerreiro et al., 2024), and MetricX-24 (Juraska et al., 2024) — which correlate much more strongly with human adequacy judgments (Freitag et al., 2022). This progress, however, raises a largely overlooked question: what exactly do these metrics measure, and is correlation with adequacy the right criterion for evaluating literary translation? Zouhar et al. (2024) show that COMET is biased toward adequacy and penalizes valid paraphrases, while Läubli et al. (2018) demonstrate that human-parity claims dissolve at document-level evaluation.

For example, in general-domain English–Ukrainian parallel data, Chaplynskyi and Zakharov (2025) found that an ensemble of six Quality Estimation (QE) models explains only ∼60% of the variance in human quality judgments, with a nonlinear relationship between metric scores and human perception, suggesting that the gap widens further for literary text.

Prior research shows that neural systems produce fluent output but struggle with stylistic consistency, figurative language, and cultural nuance (Toral and Way, 2018; Wang et al., 2023; Karpinska and Iyyer, 2023) — precisely the dimensions that define literary quality. Vanmassenhove et al. (2019) show that NMT reduces lexical and morphological richness compared to human translation, and Zhang et al. (2025) find that automatic metrics consistently prefer machine-generated literary translations over professional translators. Existing work on literary MT either remains qualitative or applies general-domain metrics without questioning their validity for literary texts.

Stylometry offers a long-standing tradition of analyzing these dimensions. Features such as function-word frequencies, lexical diversity, and morphological patterns have been used to identify translator voice and stylistic distinctiveness (Burrows, 2002; Rybicki, 2012; Baker, 2000). Yet this tradition has not been meaningfully integrated into MT evaluation research. Zheng et al. (2023) show that GPT-4 matches both controlled and crowdsourced human preferences, and LLMs have been established as state-of-the-art MT evaluators (Kocmi and Federmann, 2023). However, Huang et al. (2024) find that source information can be counterproductive for LLM-based evaluation — a finding our source-visibility experiment directly extends. We bring these strands into dialogue.

This paper sits at the intersection of MT evaluation, literary translation, and stylometric analysis. Unlike prior work that evaluates literary MT using standard metrics, we ask whether those metrics validly measure literary quality. Unlike qualitative critiques of MT in literary contexts, we provide quantitative evidence across multiple independent dimensions.

ID	Year	Translator / System	Type
T1	1947	Ivan Cherniatynskyi	Human
T2	1984	Iryna Dybko	Human*
T3	1991	Oleksii Drozdovskyi	Human
T4	1991	Yurii Shevchuk	Human
T5	1992	Natalia Okolitenko	Human†
T6	2020	Bohdana Nosenok	Human
T7	2021	Viacheslav Stelmakh	Human
T8	—	Lapa (v0.1.2-instruct)	AI (tuned LLM)
T9	—	GPT-5.2	AI (general LLM)
T10	—	DeepL	AI (commercial NMT)

Table 1: List of Ukrainian translations of George Orwell's *Animal Farm* included in the corpus. *Free cultural adaptation. †Translated from Russian, not English.

3 Corpus

To test whether neural MT metrics capture literary translation quality, we need multiple translations of the same source, enabling comparison of translational strategies independent of source variation. Our corpus comprises ten Ukrainian translations of George Orwell's *Animal Farm* (seven human, three AI) aligned into 1,367 sentence-level segments.[1] The seven human translations are drawn from the ParaFarm corpus (Maslij (Kalashnyk) and Shvedova, 2025; Kalashnyk, 2025). All translate the identical English source, so observed differences reflect strategy rather than content.

The human translations span 1947–2021, representing diaspora (Cherniatynskyi, Dybko), early-independence (Drozdovskyi, Shevchuk, Okolitenko), and contemporary professional (Nosenok, Stelmakh) contexts, which provides natural variation in norms and conventions.

We include Dybko (1984) as a deliberate test. Because it departs substantially from the English source, we expect fidelity-oriented metrics to penalize it. We therefore retain it in full metric analyses but exclude it from certain group-level comparisons between human and AI systems.

We generate three AI translations representing distinct architectures and training regimes:

[1] Sentence boundaries do not always coincide across translations; alignment follows the source segmentation.

- **GPT-5.2**[2] — a general-purpose large language model prompted for translation (see Appendix D).
- **DeepL**[3] — a commercial neural machine translation system (used via API without additional prompting).
- **Lapa (v0.1.2-instruct)** (Paniv et al., 2025) — a 12B-parameter LLM (Gemma-3) adapted for Ukrainian, state-of-the-art on English–Ukrainian translation benchmarks, and fine-tuned on multimodal data including the filtered English–Ukrainian parallel corpus from Chaplynskyi and Zakharov (2025)

The three AI systems differ in architecture and training, so convergent stylistic patterns would point to properties of AI translation in general rather than to any single model.

The corpus has three methodological advantages: one controlled source, seven human baselines spanning 74 years, and natural variation in translation norms. The retranslation literature (Deane-Cox, 2014; Stasiuk, 2019; Lepokhin, 2023) suggests that successive translators and translation editors may deliberately diverge from predecessors, which would further increase human–human variation independently of AI convergence. The result is a rare setting for disentangling semantic fidelity, stylistic range, and metric bias.

4 Methodology

Our experimental design tests the four predictions outlined in Section 1, with each analytical component targeting a distinct aspect of the construct-validity hypothesis.

To evaluate metric dominance (Prediction 1), we assess ten Ukrainian translations using seven state-of-the-art neural MT metrics spanning reference-free and reference-based paradigms. The reference-free metrics (COMETKiwi-22, COMETKiwi-XL, XCOMET-XXL, and MetricX-24 QE) estimate translation quality using only the source sentence and candidate translation. The reference-based metrics (COMET-22, XCOMET, and MetricX-24) are implemented in a round-robin design in which each translation is scored against the other nine as pseudo-references, yielding a 10×10 pairwise matrix. For each system, we report the mean score across all nine references. Higher scores indicate better quality for COMET-family metrics, whereas lower scores indicate better quality for MetricX.

If neural metrics operationalize semantic fidelity and fluency, AI systems optimized for fidelity should rank highest.

To test convergence and source proximity (Prediction 2), we compute cross-lingual semantic similarity using LaBSE (Feng et al., 2022). For each pair of systems, aligned segments are embedded, and cosine similarity is computed per segment, with mean similarity reported. We calculate system–system similarity (AI–AI, human–human, AI–human) as well as system–source similarity between each translation and the English original.

If neural metrics reward source proximity, AI systems should exhibit higher similarity to the source, tighter clustering in semantic space, and greater separation from human translations.

To assess stylistic compression (Prediction 3), we conduct a multi-dimensional stylometric analysis capturing features of Ukrainian literary expressiveness not reducible to semantic fidelity. Lexical diversity is measured using Measure of Textual Lexical Diversity (MTLD), Moving Average Type Token Ratio (MATTR), and hapax ratio (via lexical richness and pymorphy3[4]). We select two morphosyntactic features identified by Ukrainian linguists as salient markers of literary expressiveness. Discourse particles (e.g., ж ‘indeed’, таки ‘after all’, ось ‘here/just’, бо ‘because’, аж ‘even’, ну ‘well’, мов/наче ‘as if’) have no direct English equivalents and must be actively introduced by the translator; Shevelov (1963) describes them as “the microorganisms of language, which lend it colour and flavour,” making their density a marker of literary voice. Diminutive morphology — suffixes such as -еньк-, -очк-, -ик, -оньк-, -ечк- signaling affection, intimacy, or attenuation — encodes evaluative, emotional, and pragmatic functions beyond literal smallness (Ruda, 2021), requiring the translator to perceive the emotional register of the scene. We quantify particle frequency and diversity through lemma matching and detect diminutive morphology using regex over morphologically analyzed tokens. Surface overlap between translations is measured with pairwise chrF and BLEU scores computed with sacrebleu,[5] and stylistic distance is estimated using Cosine Delta, a variant of Burrows’

[2] https://openai.com/index/gpt-5/
[3] https://www.deepl.com/
[4] https://github.com/no-plagiarism/pymorphy3
[5] https://github.com/mjpost/sacrebleu

Delta (Evert et al., 2017) based on z-score normalized function-word frequencies. We also compute the standard deviation of per-segment Ukrainian-to-English word-count ratios to assess consistency in expansion and compression. If neural metrics fail to capture stylistic richness, top-ranked systems may exhibit reduced lexical diversity, lower particle and diminutive usage, and lower stylistic dispersion.

To test preference reversal under controlled source visibility (Prediction 4), we conduct two LLM-as-a-judge experiments using GPT-5.2 with identical sampling and aggregation procedures. Note that GPT-5.2 serves a dual role: it is both one of the evaluated translation systems (T9) and the judge. We discuss the resulting self-preference risk in the Limitations section.

- *Translation-focused:* the model evaluates pairs with access to the English source, mirroring the human evaluation setup.

- *Literary-focused:* the model evaluates the same pairs without access to the source and is instructed to judge solely on literary naturalness, expressiveness, and stylistic richness in Ukrainian.

Human evaluation is conducted as a pairwise preference tournament following Romanyshyn et al. (2024). Four professional English–Ukrainian translators serve as annotators. They are shown the English source and two anonymized Ukrainian translations and select the better translation or indicate a tie. Judgments emphasize meaning preservation alongside fluency and literary quality. Because evaluators have access to the source text, this procedure reflects adequacy-oriented judgment under source visibility, comparable to the conditions under which neural metrics are trained.

All three evaluations — both LLM-as-a-judge experiments and the human evaluation — draw from the same pool: 1,128 valid segments $\times$ 45 system pairs $=$ 50,760 items, globally shuffled with a fixed seed. This ensures that even early subsets cover all 45 system pairs approximately uniformly. Left/right presentation order is randomized per pair to control for position bias. TrueSkill ratings (Herbrich et al., 2006) are computed with default parameters ($\mu_0 = 25$, $\sigma_0 = 25/3$).

If neural metrics primarily reward fidelity to the source, removing source visibility should systematically alter system rankings, providing causal evidence of construct misalignment.

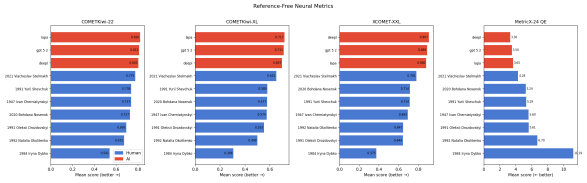

Figure 1: Comparison of reference-free neural metrics across the systems.

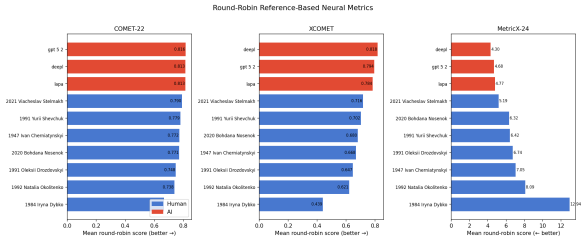

Figure 2: Round-robin comparison across the systems.

5 Results

We present results corresponding to the four predictions made in Section 1.

5.1 Prediction 1: Metric Dominance

Across every MT metric, a consistent pattern emerges: all three AI systems rank in the top three. No metric ranks any human translator above any AI system. The highest-ranked human translation (Stelmakh, 2021) consistently places fourth.

For example:

- On COMETKiwi-22, AI systems average 0.812 versus 0.724 for human translators (excluding Dybko), a gap of +0.088.

- On MetricX-24 QE (lower is better), AI systems average 3.48 compared to 5.45 for humans.

The pattern holds across COMETKiwi-XL, XCOMET-XXL, COMET-22, and round-robin MetricX-24. Figure 1 (reference-free metrics) and Figure 2 (round-robin metrics) show consistent separation between AI and humans.

As expected, Dybko's free cultural adaptation ranks last on every metric, confirming that the metrics are highly sensitive to source deviation. Notably, Lapa — a 12B-parameter model adapted from Gemma-3 — tops the COMETKiwi rankings, matching or exceeding GPT-5.2, a much larger general-purpose model. On metric scores, a

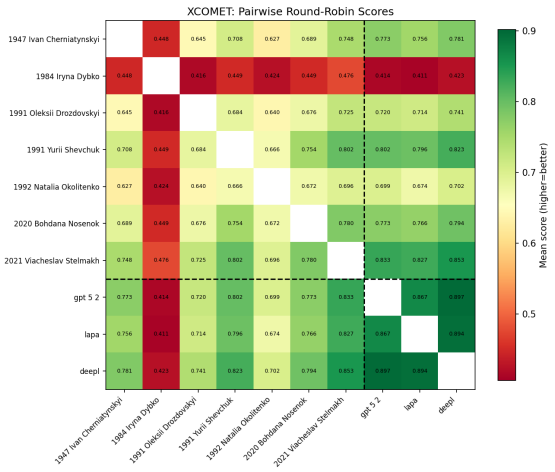

Figure 3: Heatmap of pairwise XCOMET round-robin scores (note the bottom-right AI block).

Measure	H–H	AI–AI	Gap
LaBSE cosine sim.	0.711	**0.941**	+0.230
XCOMET round-robin	0.627	**0.886**	+0.259
COMET-22 round-robin	0.750	**0.881**	+0.131
MetricX-24 round-robin	7.61	**3.38**	−4.23
chrF (surface overlap)	33.6	**43.1**	+9.5

Table 2: AI–AI vs. Human–Human average pairwise scores. For MetricX-24, lower values indicate greater similarity. Bold signifies the more similar group.

domain-tuned small model can reach parity with a system orders of magnitude larger.

These results confirm Prediction 1: neural MT metrics systematically favor AI translations in this literary corpus. Not only that, the round-robin heatmaps reveal a uniformly high-scoring 3×3 AI–AI block, while the 7×7 human–human block shows much more variation.

5.2 Prediction 2: Convergence

We next examine pairwise similarity among translations. The AI–AI LaBSE similarity averages 0.941, compared to 0.711 for human–human pairs — a +0.230 gap.

This agreement is consistent across independent measures:

The individual AI–AI LaBSE pairs — Lapa–DeepL (0.952), GPT-5.2–DeepL (0.940), Lapa–GPT-5.2 (0.932) — represent near-identical translations. The convergence holds on every measure, from neural embeddings to raw character n-grams. Three architecturally distinct systems have arrived at the same output.

AI systems are also measurably closer to the

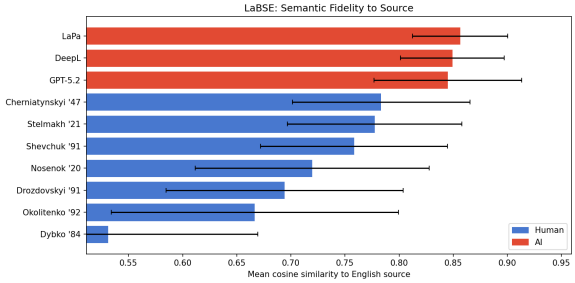

Figure 4: LaBSE semantic similarity to English source.

English source than any human translator in cross-lingual embedding space. On LaBSE, all three AI systems exceed 0.845 similarity to the source; the closest human (Cherniatynskyi, 1947) reaches 0.783. The human average (excluding Dybko) is 0.733. This gap may reflect AI systems selecting the most frequent translation equivalents, while human translators employ richer or less default lexical choices that increase semantic distance from the source.

Surface overlap (chrF/BLEU) confirms this trend. AI translations are 15–20% more similar to each other at the surface level (chrF) than any pair of human translations. The three AI systems form a tight cluster; humans spread across a much wider range — a gap that may be amplified by a deliberate "repulsion effect," as later translators often read and consciously diverge from their predecessors (Deane-Cox, 2014).

Despite architectural differences, a general-purpose LLM (GPT-5.2), a commercial NMT system (DeepL), and a domain-tuned LLM (Lapa) produce near-identical translations. The convergence persists from neural metrics to surface-level chrF scores.

This confirms Prediction 2: systems that neural metrics favor are both closest to the source and highly clustered in semantic space.

5.3 Prediction 3: Stylistic Compression

At the same time, AI translations systematically lack Ukrainian literary expressiveness. AI systems average an MTLD of 311 vs. 377 for humans (excluding Dybko) — an 18% gap. The hapax ratio (words used exactly once) tells the same story: AI 0.182 vs. human 0.215, a 15% deficit. AI translations cycle through a narrower vocabulary, concentrating 39.4% of their text within the 100 most common words (vs. 37.7% for humans).

Ukrainian particles (ж 'indeed', таки 'after all',

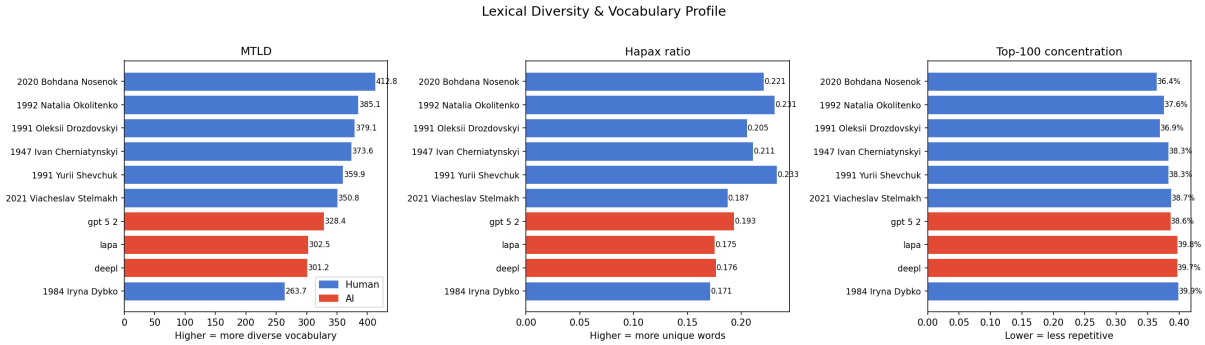

Figure 5: Lexical diversity (MTLD, Hapax, Top-100).

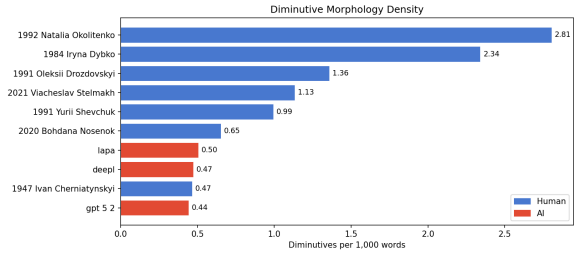

Figure 6: Diminutive morphology across translators and AI systems.

ось ‘here’, бо ‘because’, аж ‘even’, ну ‘well’, мов/наче ‘as if’) encode pragmatic nuances — emphasis, surprise, hedging — with no direct English equivalents. They must be *added* by the translator. GPT-5.2 and DeepL produce approximately 2× fewer particles than the human average.

Diminutive suffixes (-еньк-, -очк-, -ик, -оньк-; e.g., мама ‘mom’ → мамочка ‘mommy’) are a core expressive device in Ukrainian, signaling affection, irony, or intimacy. Linguists studying Ukrainian diminutives note that the category often encodes much more than literal smallness, including evaluative, emotional, or pragmatic functions in discourse (Ruda, 2021).

AI systems average 0.47 diminutives per thousand tokens; humans average 1.23 — a 2.6× gap. All three AI systems rank at the bottom, below every human translator. Lapa, despite literary fine-tuning, is indistinguishable from GPT-5.2 and DeepL on this measure.

Using function-word lemma frequencies — content-independent markers of translatorial voice — Cosine Delta measures how stylistically distinct each system is from the others. AI systems have the smallest mean pairwise distance (DeepL: 1.058, GPT-5.2: 1.059, Lapa: 1.068), placing them closest to each other and to the corpus centroid. Human translators range from 1.095 (Nosenok) to 1.160 (Dybko). By this measure, AI systems occupy the same narrow corner of the stylistic space.

Lapa ($\sigma = 0.127$) and DeepL ($\sigma = 0.132$) are the two most uniform systems, maintaining near-constant word count ratios across segments. Human translators vary more — they expand descriptive passages and compress dialogue, adapting to content. AI produces uniformly adequate output without the peaks and valleys that characterize human stylistic choices.

These results confirm Prediction 3: the systems that neural metrics rank highest exhibit reduced lexical diversity, fewer discourse particles and diminutives, lower stylistic dispersion, and more uniform expansion ratios — a systematic pattern of stylistic compression.

5.4 Prediction 4: Preference Reversal

In order to validate these results, we ran two LLM-as-a-judge experiments (∼750 pairwise comparisons each, with rankings stable between the 500- and 750-pair checkpoints) and a human evaluation (1,034 judgments) with an identical setup except for one variable: the presence of the English source.

- **LLM-as-a-judge Translation (Source Visible):** Judge sees the English source + two Ukrainian translations.
- **LLM-as-a-judge Literary (Source Hidden):** Judge sees only two Ukrainian sentences.
- **Human eval:** same setup as experiment 1.

We compute TrueSkill ratings for each experiment and compare them with metric rankings and human evaluations (1,034 matches). The results are in Table 3; the ranking shifts for AI systems are in Table 4.

System	Metrics	LLM: Translation	LLM: Literary	Human Eval
Lapa	**#1**	#4 (27.4)	#9 (22.4)	#7 (24.9)
GPT-5.2	**#2**	**#1** (30.3)	#6 (24.4)	#5 (25.3)
DeepL	**#3**	**#3** (29.2)	#8 (23.8)	**#2** (26.4)
Stelmakh 2021	#4	**#2** (29.8)	**#2** (28.6)	**#1** (27.1)
Shevchuk 1991	#5	#5 (26.7)	**#3** (26.6)	**#3** (26.3)
Cherniatynskyi 1947	#6	#6 (25.0)	#7 (24.3)	#8 (24.1)
Nosenok 2020	#7	#8 (23.5)	#5 (24.7)	#4 (25.3)
Drozdovskyi 1991	#8	#7 (24.0)	**#1** (29.5)	#6 (25.2)
Okolitenko 1992	#9	#9 (21.4)	#4 (25.2)	#9 (23.3)
Dybko 1984	#10	#10 (13.9)	#10 (21.4)	#10 (18.1)

Table 3: System rankings across four evaluation paradigms, ordered by metric rank. TrueSkill μ in parentheses. Bold = top 3 in each column.

AI system	Metric	Transl.	Literary	Shift
GPT-5.2	#2	#1	#6	↓5
DeepL	#3	#3	#8	↓5
Lapa	#1	#4	#9	↓5

Table 4: AI rank shifts when the English source is removed from evaluation.

This reversal is uniform across all AI systems and persists even when the judging model is held constant. The main experimental manipulation here is the presence or absence of the source sentence. **The top five positions in the literary ranking are all human translators.**

The most dramatic individual reversal is that of Drozdovskyi (1991). This translation ranks #8 on COMETKiwi-22 (0.693) and #7 on the translation judge — firmly in the bottom half. But on the literary judge, it leaps to #1 ($\mu = 29.5$), surpassing every system, including Stelmakh. The translation is rich in particles (10.1/1k — the highest in the corpus) and diminutives (1.36/1k). These are exactly the features metrics penalize and literary judgment rewards.

Okolitenko (1992), who translated from Russian rather than English, shows a similar pattern: #9 on metrics and human evaluation, but #4 on the source-free literary judge. The low fidelity scores are expected given the different source language, yet the Ukrainian prose is judged favorably when evaluated on its own terms.

Stelmakh (2021) is the only translation to rank in the top two across all non-metric rankings: #1 in human eval (27.1), #2 in literary judge (28.6), #2 in translation judge (29.8). On metrics, it ranks #4 — the best human. Stelmakh represents the rare translator whose work satisfies both fidelity-oriented and literary-oriented evaluation: semantically faithful enough for metrics, stylistically rich enough for readers.

The 1,034-match human evaluation places Stelmakh clearly first ($\mu = 27.1$), but positions #2–#7 form a compressed cluster ($\mu = 24.9$–26.4) with overlapping confidence intervals. DeepL ranks #2 and Shevchuk #3 — both above GPT-5.2 (#5) and Lapa (#7). Humans value fidelity enough to keep DeepL near the top, but not enough to replicate the metric-based ranking in which all three AI systems dominate.

This suggests that human judges — even when given access to the source — weigh stylistic and expressive qualities more heavily than neural metrics do. In other words, human assessment appears to balance semantic fidelity with literary naturalness, whereas neural metrics disproportionately reward the former. The human results thus reinforce the central claim of construct misalignment: the metrics that identify what is "best" do not fully correspond to what readers perceive as high-quality literary translation.

These results confirm Prediction 4: evaluation criteria shift when the source is removed, demonstrating that metric-aligned fidelity is not equivalent to literary quality.

6 Discussion

Our results suggest a clear answer: neural MT metrics primarily measure semantic fidelity to the source and surface fluency in the target language. These are the properties AI systems maximize (via RLHF, parallel training data, and instruction tuning), and these are the properties on which AI systems achieve the highest scores.

This is not a flaw in the metrics per se — semantic fidelity and fluency are genuine dimensions of translation quality. The problem is one of construct validity: the metrics claim to measure "translation

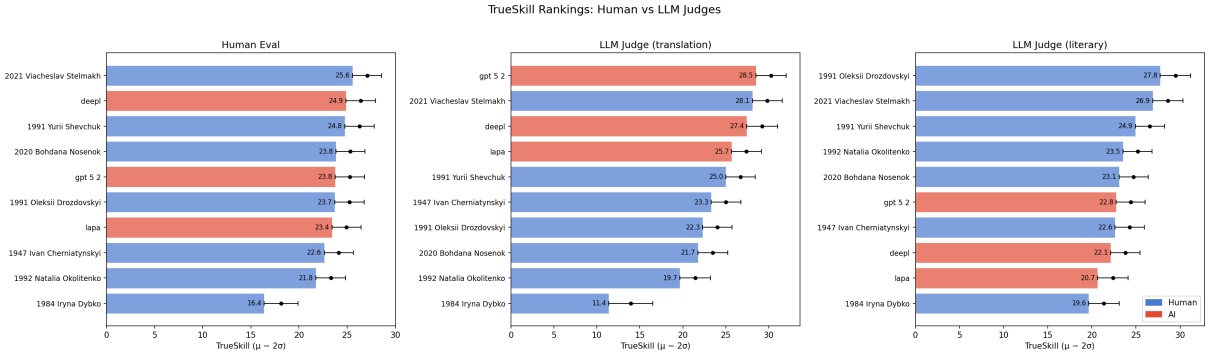

Figure 7: TrueSkill rankings across three evaluation paradigms.

quality" but actually measure a subset of quality that is systematically correlated with AI output. In the literary domain, this subset is insufficient.

The preference reversal is the strongest evidence for this claim. When the source is visible, the judge sees the same information the metrics do and produces the same ranking. When the source is hidden, the judge must evaluate the Ukrainian text on its own merits — and the ranking inverts. **The quality that survives without source access is a different quality from what metrics capture.**

Perhaps the most striking finding is the convergence of three architecturally distinct AI systems. GPT-5.2, DeepL, and Lapa produce translations with pairwise LaBSE similarity of 0.941. By every measure (neural metrics, surface overlap, cross-lingual embeddings, cosine delta), they are closer to each other than any pair of human translators is to each other.

This has implications beyond evaluation. If AI translation converges on a single point in the space of possible translations, then:

1. Switching between AI systems may not always increase diversity. While individual differences exist, the overall stylometric profiles of the three AI systems we tested are far more similar to each other than to any human translator. Whether this convergence generalizes to a broader population of MT systems is an open question requiring larger-scale replication.

2. Domain tuning closes the metric gap but not the stylistic one. Lapa (12B parameters) matches the much larger GPT-5.2 on metric scores — a strong result for a compact, domain-adapted model. Yet on stylometric measures, Lapa is indistinguishable from GPT-5.2 and DeepL: fine-tuning achieves metric parity without producing a distinct literary voice.

The stylometric deficits we document — in particles, diminutives, and lexical diversity — are not arbitrary choices but features that Ukrainian linguists identify as markers of skilled prose (see Section 4). Their absence in AI translations is not a matter of personal preference but of cultural competence: AI systems translate *what is said* but not *how it is said*.

These findings have practical consequences. For MT evaluation research, current metrics should not be applied to literary translation without explicit caveats. A literary-specific evaluation metric is needed — one that rewards vocabulary richness, cultural perceptiveness, and stylistic identity alongside semantic fidelity. For AI-assisted translation, AI output may serve as a useful starting point — human evaluators ranked DeepL second overall — but the stylistic deficits documented here suggest that post-editing for lexical and cultural richness remains necessary; across the three systems we tested, switching between AI providers did not yield greater stylistic diversity, though wider replication is needed before generalizing.

6.1 Future Directions

One avenue for improvement is prompt design. The LLM-based translations (GPT-5.2 and Lapa) were generated with the minimal translation prompt used by the Lapa team in their evaluation pipeline (Paniv et al., 2025) — adopted here for direct comparability with their reported benchmark scores — without specific instructions targeting stylistic features such as particles or diminutives. Because large language models are sensitive to instruction framing, prompts explicitly emphasizing lexical variation, pragmatic particles, and affective nuance may reduce stylistic compression without requiring fine-

tuning. Controlled comparisons between general and feature-targeted prompts would clarify this possibility.

A second direction concerns evaluation. Literary-aware metrics could incorporate stylometric dispersion, lexical richness, and discourse-pragmatic features alongside semantic fidelity. Hybrid systems combining adequacy metrics with source-free literary judgment may better reflect the multidimensional nature of translation quality.

The particles and diminutives analyzed here are salient examples that Ukrainian linguists have already highlighted, but the space of expressive markers is likely much larger. Preliminary analysis of augmentative and pejorative suffixes (-ищ(е), -иськ(о), -юг(а), -юр(а)) showed zero occurrences in all AI translations versus sparse but non-zero usage by humans — a pattern consistent with our findings, though the signal in this corpus is too weak for statistical claims. Systematic identification of further morphosyntactic markers of literary expressiveness is a promising direction for future work.

Finally, replication across genres and language pairs, particularly those with different morphological and expressive resources, would determine whether stylistic compression is language-specific or structural to current AI systems.

7 Conclusion

Across seven neural metrics, three AI translations consistently rank at the top. Yet these same translations are not as lexically rich as human ones, lack common signs of Ukrainian literary voice, and converge on near-identical output. When an LLM-as-a-judge evaluates the translations without access to the source — judging only whether the Ukrainian text reads as skilled literary prose — the AI advantage disappears, and human translators take the top five positions.

Although the metrics accurately measure what they were trained to measure — semantic fidelity to the source and surface fluency in the target — literary translation requires more than fidelity and fluency. It requires voice, cultural adaptation, and expressive richness — qualities that current metrics cannot detect and that AI systems do not produce.

Limitations

The study is limited to a single novella and a single language pair (English→Ukrainian), restricting generalization. The number of systems (seven human, three AI) constrains statistical breadth, though TrueSkill mitigates uncertainty. The seven human translations span 74 years (1947–2021) and reflect shifts in Ukrainian orthographic and stylistic norms across diaspora, Soviet-era, and post-independence periods; this temporal range is part of what makes the corpus interesting, but it also means within-human variation conflates translator voice with diachronic norm change. We note, however, that within-human pairwise LaBSE similarity (0.711) remains substantially lower than within-AI similarity (0.941), so temporal and norm-driven variation in the human set does not approach the magnitude of the AI–human gap that drives our central findings. The 1,034-match human evaluation is directionally stable, but the #2–#7 cluster is tight with overlapping confidence intervals. GPT-5.2 serves as both a translation system and the LLM judge, creating a potential self-preference bias; however, GPT-5.2 ranks only #6 in the literary condition and #5 in human evaluation, suggesting that any self-preference does not dominate the results. Replication with an architecturally distinct judge model (e.g., Claude or Gemini) would further strengthen this conclusion and is a clear next step. Our stylometric analysis targets specific morphosyntactic features (particles, diminutives, lexical diversity) but does not measure figurative language or cultural adaptation, which are also central to literary quality. Finally, sentence-level pairwise judgments cannot capture long-range narrative qualities such as sustained voice or rhythm, and our human evaluation was conducted only under source-visible conditions; a source-hidden human pairwise evaluation, mirroring the LLM literary judge, is the natural next experiment to fully decouple the source-removal effect from the human/LLM-judge contrast.

Ethical Considerations

This study uses published literary translations and publicly available AI systems. Human annotators were professional translators who participated in the study on a voluntary basis. No personally identifiable information was collected. The AI translations were generated for research purposes. AI writing assistance (Claude, Anthropic) was used

for editing and formatting the manuscript. We acknowledge that our findings about AI translation limitations should not be used to devalue human translators' work but rather to highlight the irreplaceable qualities they bring to literary translation.

Acknowledgments

We thank Vladislav Demyanov, Kateryna Buchina, and Serhiy Snihur for serving as expert translators in the human evaluation, and Taras Yaroshko for his contributions during the early stages of this research.

References

Mona Baker. 2000. Towards a methodology for investigating the style of a literary translator. *Target*, 12(2):241–266.

John Burrows. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Dmytro Chaplynskyi and Kyrylo Zakharov. 2025. A framework for large-scale parallel corpus evaluation: Ensemble quality estimation models versus human assessment. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 73–85. ACL.

Sharon Deane-Cox. 2014. *Retranslation: Translation, Literature and Reinterpretation*. Bloomsbury Academic.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of ACL 2022*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU — neural metrics are better and more robust. In *Proceedings of WMT 2022*, pages 46–68.

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

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems 19*.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of ACL 2024*, pages 3546–3562.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 492–504.

Viktoriia Kalashnyk. 2025. Creation of a multi-variant parallel corpus of Ukrainian translations of George Orwell's "Animal Farm" and its use for studying variability of the Ukrainian language. In *Language Space of the Modern World: Proceedings of the IX All-Ukrainian Scientific Conference*, pages 113–119. NaUKMA.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of WMT 2023*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of EAMT 2023*, pages 193–203.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of EMNLP 2018*, pages 4791–4796.

Yevhenii Lepokhin. 2023. Vasyl Stefanyk's short story "All alone" as interpreted in translations by Olha Kobylianska, Mary Skrypnyk and Danylo Struk in English and German. *Slavia: Journal for Slavic Philology*, 92(3):322–353.

Viktoriia Maslij (Kalashnyk) and Mariia Shvedova. 2025. ParaFarm: English-Ukrainian multiple-translation corpus (1.1). Data set.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP 2020*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of WMT 2022*.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74.

Nataliia Ruda. 2021. Diminutive contronyms in Ukrainian. *Studia Slavica Academiae Scientiarum Hungaricae*, 65(2):341–350.

Jan Rybicki. 2012. The great mystery of the (almost) invisible translator: Stylometry in translation. In *Quantitative Methods in Corpus-Based Translation Studies*, pages 231–248. John Benjamins.

George Y. Shevelov. 1963. *The Syntax of Modern Literary Ukrainian: The Simple Sentence*. Mouton.

Bohdan Stasiuk. 2019. Conflict of editorial versions of old translations and the problem of their republishing. *Zapysky Naukovoho tovarystva imeni Shevchenka*, 272:496–514.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of EMNLP 2023*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of NAACL 2025*, pages 10961–10988.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of WMT 2024*, pages 1272–1288.

A Appendix: Additional Metric Heatmaps

This appendix reports the full pairwise round-robin score matrices for the reference-based metrics, complementing the XCOMET heatmap in Figure 3.

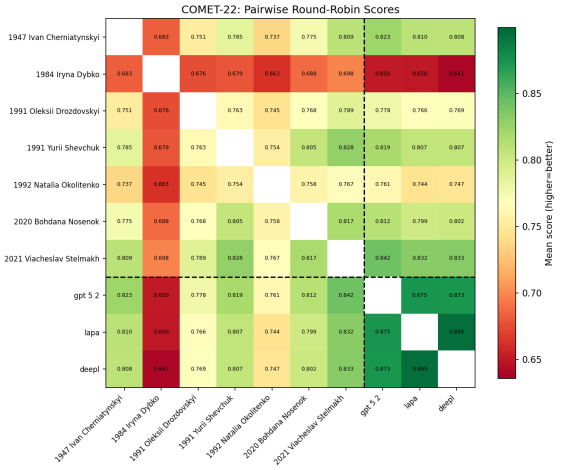

Figure 8: Heatmap of pairwise COMET-22 round-robin scores.

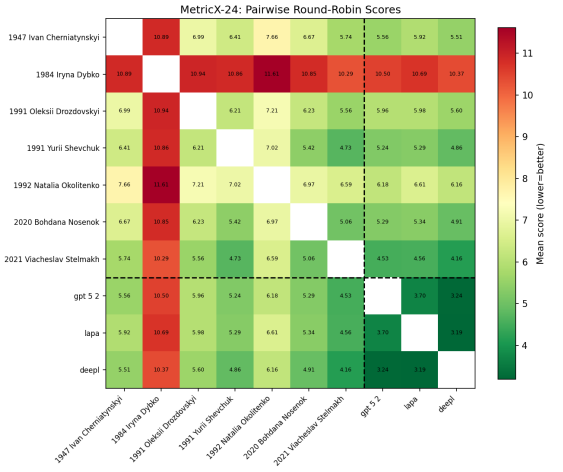

Figure 9: Heatmap of pairwise MetricX-24 round-robin scores.

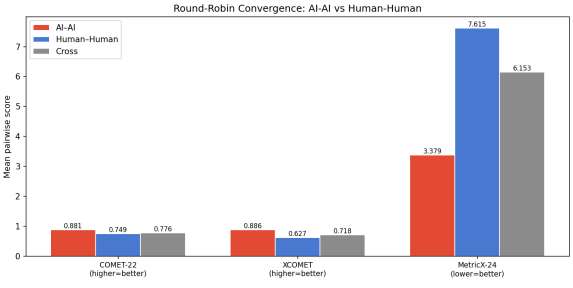

Figure 10: Round-robin convergence across MT metrics.

B Appendix: Additional LaBSE Analyses

This appendix presents additional cross-lingual similarity analyses based on LaBSE sentence em-

beddings.

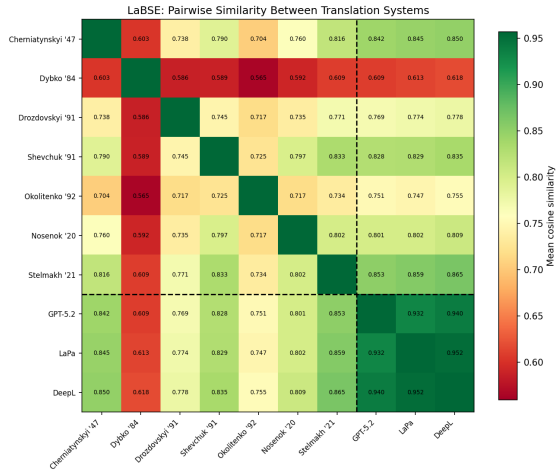

Figure 11: LaBSE pairwise similarity heatmap.

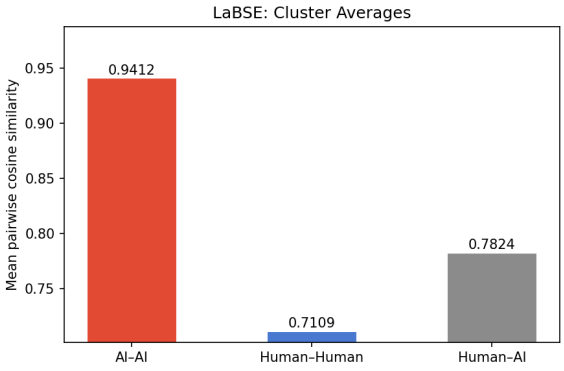

Figure 12: LaBSE cluster averages (AI–AI, Human–AI, Human–Human).

C Appendix: Additional Stylometric Analyses

This appendix collects the remaining stylometric comparisons across translators and AI systems.

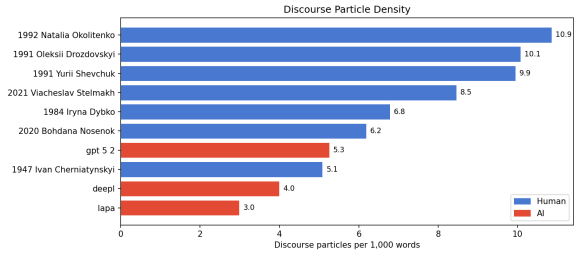

Figure 13: Discourse particle frequency across translators and AI systems.

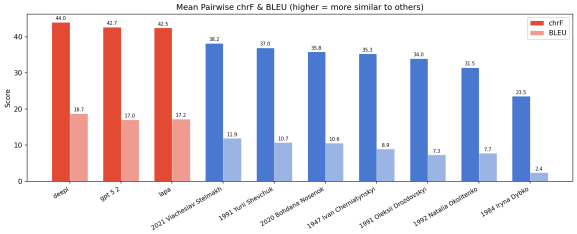

Figure 14: Mean pairwise chrF and BLEU scores.

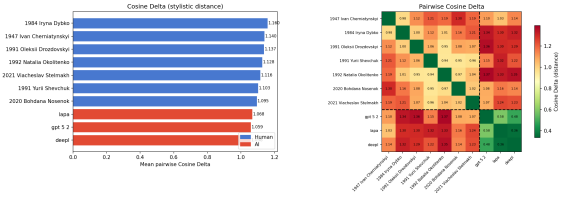

Figure 15: Cosine Delta heatmap showing stylistic distances between translators.

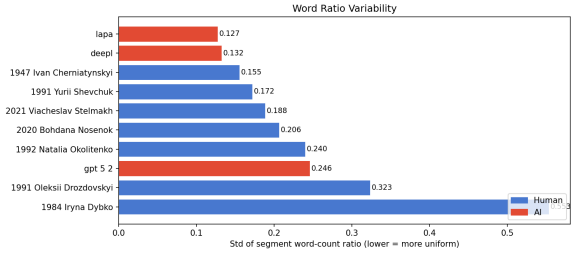

Figure 16: Word ratio uniformity across translators and AI systems.

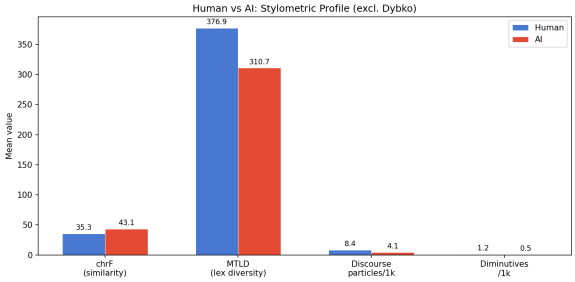

Figure 17: Stylometric convergence summary across all measures.

D Appendix: Prompts

AI translation prompt (Lapa, GPT-5.2).

> Translate the following English text to Ukrainian. Output only the translated text without any additional words or formatting, start with the translated text:

LLM judge: translation (source visible). System prompt:

> You are an expert literary Ukrainian translator evaluating translation quality. You will be given one English sentence and two Ukrainian translations. Choose the better translation.
>
> How to decide: 1. Meaning preservation — Does the translation convey the core meaning and intent of the English sentence? 2. Fluency and literary quality — Which translation reads more natural, expressive, and appropriate for literary Ukrainian?
>
> Rules: Prefer the translation that best balances intent and natural literary expression. Use "tie" only if you genuinely cannot decide. Judge each sentence independently.
>
> Respond with EXACTLY one of: `system1`, `system2`, `tie`.

User template: `English: {english} | system1: {system1} | system2: {system2}`

LLM judge: literary (source hidden). System prompt:

> You are an expert in Ukrainian literature. You will be given two Ukrainian sentences. Choose the one that sounds more literary — as if written by a skilled Ukrainian author for a published book.
>
> Judge only literary quality: naturalness, expressiveness, and stylistic richness of the Ukrainian language.
>
> Rules: Use "tie" only if you genuinely cannot decide. Judge each sentence independently.
>
> Respond with EXACTLY one of: `system1`, `system2`, `tie`.

User template: `system1: {system1} | system2: {system2}`

Both experiments use GPT-5.2 with temperature 0.0 and max 16 output tokens.

E Appendix: Reproducibility

All computational analyses are reproducible from the accompanying repositories: the translation metrics and stylometric pipeline at https://github.com/vanneqq/translation-metrics, and the human evaluation tournament implemented in Vulyk Arena at https://github.com/lang-uk/vulyk-arena. The three AI translations will be submitted as an update to the ParaFarm corpus to enable replication and further research.