UDC 81'322 **DOI: 10.18413/2313-8912-2026-12-1-0-5**

Tatiana A. Litvinova[1]
Galina A. Zavarzina[2]

Metrics for Cultural Semantic Integrity in LLMs: A Low-Resource Perspective

[1] Voronezh State Pedagogical University,
86 Lenin St., Voronezh, 394043, Russia
E-mail: centr_rus_yaz@mail.ru
ORCID: 0000-0002-6019-3700

[2] Voronezh State Pedagogical University,
86 Lenin St., Voronezh, 394043, Russia
E-mail: zga1311@mail.ru
ORCID: 0000-0002-9129-9591

Received 04 February 2026; accepted 15 March 2026; published 30 March 2026

Abstract: Multilingual large language models (LLMs) are predominantly trained and evaluated within English-centric pipelines. However, the semantic consequences of English-language mediation at the level of textual representations remain poorly understood beyond surface-level similarity measures. This paper puts forward a metric-based approach to evaluating the cultural and semantic integrity of texts produced using multilingual large language models (LLMs), with a specific focus on low-resource languages.

We set forth a set of complementary embedding-based metrics designed to diagnose how English mediation reshapes textual semantic representations at multiple levels. Using English-mediated back-translation via an LLM as a controlled diagnostic probe, we compare a high-resource language (Russian) with a low-resource language (Lingala). Texts are embedded into a shared semantic space, and semantic integrity is assessed using three metrics: Semantic Self-Similarity (SSI), capturing local semantic recognizability; Neighborhood Preservation Score (NPS), measuring the stability of local semantic relations; and axis-based drift, quantifying directional semantic bias along an interpretable semantic opposition.

The results reveal a pronounced cross-linguistic asymmetry. Russian texts maintain high semantic self-similarity, indicating strong surface-level semantic preservation, but display only moderate neighborhood preservation, reflecting nontrivial structural reorganization. In contrast, Lingala texts show severe degradation in both semantic self-similarity and neighborhood preservation, indicating a collapse of relational semantic structure under English mediation. Additionally, Lingala – but not Russian – exhibits a small yet systematic directional drift along the examined semantic axis. What is of importance is that this directional bias is independent of structural instability, which is indicative of multiple, distinct mechanisms of English-centric effect.

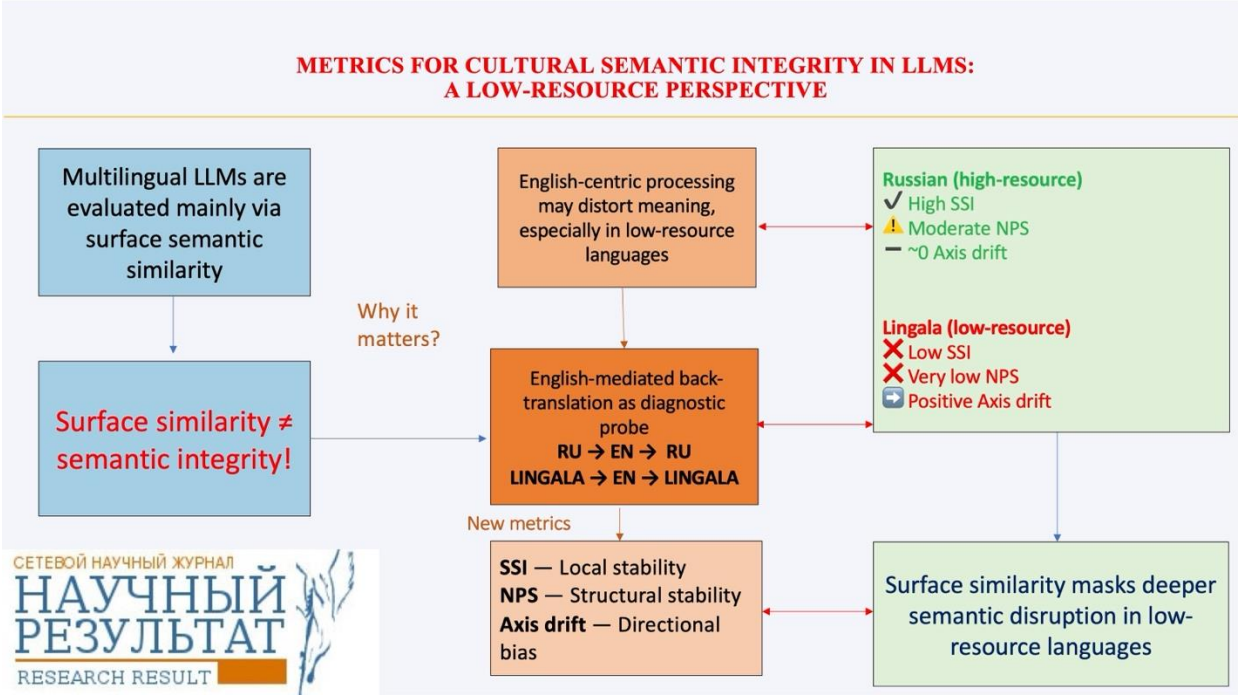
These findings indicate that surface similarity metrics considerably underestimate semantic disruption, particularly for low-resource languages. The suggested framework provides a scalable diagnostic toolkit for assessing semantic integrity in multilingual LLM representations and is directly applicable to the analysis and evaluation of LLM-generated texts beyond translation-based scenarios. Although we are validating the

framework using Russian and Lingala, the proposed metrics are intended for use with other low-resource languages and in multilingual settings.

Keywords: Cultural Semantic Integrity; Large Language Models; Low-Resource Languages; Semantic Drift; Multilingual Embeddings; English-Centric Bias.

Acknowledgements: The study is supported by the Ministry of Education of the Russian Federation within the framework of the state task in the field of science (topic number QRPK-2025-0013).

How to cite: Litvinova, T. A., Zavarzina, G. A. (2026). Metrics for Cultural Semantic Integrity in LLMs: A Low-Resource Perspective, *Research Result. Theoretical and Applied Linguistics*, 12 (1), 123–145. DOI: 10.18413/2313-8912-2026-12-1-0-5



УДК 81’322 **DOI: 10.18413/2313-8912-2026-12-1-0-5**

Т. А. Литвинова[1]
Г. А. Заварзина[2]

Метрики культурно-семантической эквивалентности для больших языковых моделей: взгляд со стороны малоресурсных языков

[1] Воронежский государственный педагогический университет,
394043, Россия, Воронеж, ул. Ленина, 86
E-mail: centr_rus_yaz@mail.ru
ORCID: 0000-0002-6019-3700

[2] Воронежский государственный педагогический университет,
394043, Россия, Воронеж, ул. Ленина, 86
E-mail: zga1311@mail.ru
ORCID: 0000-0002-9129-9591

*Статья поступила 04 февраля 2026 г.; принята 15 марта 2026 г.;
опубликована 30 марта 2026 г.*

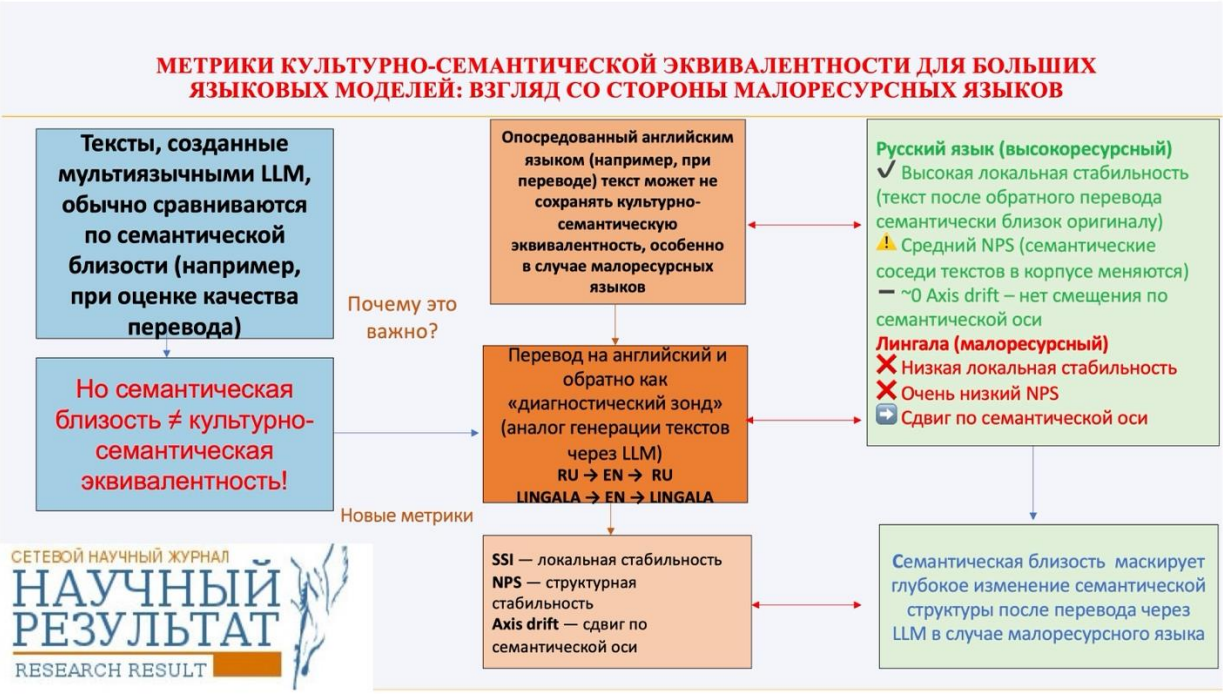
Аннотация: Мультиязычные большие языковые модели (LLM) преимущественно обучаются и оцениваются на англоязычном материале как доминирующем в датасете. Однако семантические последствия опосредования английским языком их выдачи (например, в результате перевода либо при генерации текстов) на уровне текстовых представлений до сих пор остаются недостаточно изученными за пределами поверхностных мер семантического сходства. В данной работе предлагается метрико-ориентированная методология оценки **культурно-семантической целостности** текстов, созданных с использованием многоязычных LLM, с особым акцентом на малоресурсные языки.

Мы предлагаем набор взаимодополняющих метрик на основе мультиязычных эмбеддингов, предназначенных для исследования эффекта влияния английского посредничества на нескольких уровнях. Используя англо-опосредованный обратный перевод с помощью LLM в качестве диагностического инструментария, мы сравниваем такое влияние на тексты на высокоресурсном (русский) и малоресурсном (лингала) языках. Тексты (оригиналы и обратные переводы) проецируются в общее семантическое пространство, а их семантическая эквивалентность оценивается с помощью трёх метрик: 1) метрики ***семантического автосходства*** **(Semantic Self-Similarity, SSI)**, отражающее семантическую близость текста-оригинала и текста после обратного перевода; 2) метрики ***сходства семантических соседей*** **(Neighborhood Preservation Score, NPS)**, измеряющей стабильность локальных семантических связей; 3) метрики ***сдвига по семантической оси*** **(axis-based drift)**, количественно описывающей семантическое смещение вдоль интерпретируемой семантической оппозиции.

Результаты исследования показывают выраженную межъязыковую асимметрию. Русские тексты сохраняют высокую семантическую схожесть до и после перевода, что указывает на сохранение поверхностной семантики, однако демонстрируют лишь умеренную сохранность семантических соседей, отражающую нетривиальную структурную реорганизацию датасета. В противоположность этому тексты на языке лингала показывают резкое ухудшение как семантической схожести, так и сохранности семантических окрестностей, что свидетельствует о коллапсе реляционной семантической структуры под влиянием английского языка как промежуточного канала. Кроме того, лингала — в отличие от русского — демонстрирует небольшое, но систематическое направленное смещение вдоль рассматриваемой семантической оси. Принципиально важно, что данное направленное смещение не зависит от структурной нестабильности, что указывает на наличие нескольких различных механизмов влияния английского языка при посредничестве LLM.

Полученные результаты демонстрируют, что метрики поверхностного семантического сходства существенно недооценивают масштабы семантических искажений при обратном переводе, особенно в случае малоресурсного языка. Предложенная методика представляет собой масштабируемый диагностический инструментарий для оценки семантической эквивалентности текстов, созданных с помощью многоязычных LLM.

Ключевые слова: Культурно-семантическая эквивалентность; большие языковые модели; малоресурсные языки; семантический сдвиг; многоязычные эмбеддинги; англоцентричное смещение.
Финансирование: Исследование выполнено при поддержке Министерства просвещения Российской Федерации в рамках государственного задания в сфере науки (тема № QRPK-2025-0013).
Информация для цитирования: Литвинова Т. А., Заварзина Г. А. Метрики культурно-семантической эквивалентности для больших языковых моделей: взгляд со стороны малоресурсных языков // Научный результат. Вопросы теоретической и прикладной лингвистики. 2026. Т. 12. № 1. С. 123–145. DOI: 10.18413/2313-8912-2026-12-1-0-5



1. INTRODUCTION

Large Language Models (LLMs) have rapidly become a fundamental infrastructure for natural language processing, enabling high-quality text generation, summarization, and translation across a broad range of languages. Their apparent multilingual competence resulted in widespread adoption of LLM-based pipelines for non-English content, frequently with no explicit consideration of how linguistic and cultural meanings might be structurally reorganized during internal representation and text generation.

A growing body of research has indicated that LLMs exhibit systematic biases, including demographic, social, and ideological biases, which might affect both generated content and downstream decisions. Comprehensive surveys have shown that these biases are not isolated anomalies but rather structural properties of large-scale language models trained on heterogeneous and unevenly distributed data sources (Gallegos et al., 2024). These findings have given rise to extensive efforts in bias detection, evaluation, and mitigation within LLM outputs (Xu et al., 2025). However, most bias-focused evaluations concentrate on content-level distortions or decision outcomes, rather than in the underlying semantic structure of text collections.

Beyond well-studied social biases, recent research increasingly has been highlighting cultural and linguistic biases as a distinct and insufficiently explored

dimensions of LLM behavior (Narayan et al., 2025). Even multilingual models, which are explicitly trained to handle multiple languages, have been shown to disproportionately reflect English-centric semantic norms and evaluative patterns. Empirical evidence suggests that multilingual language models are not necessarily multicultural; they might encode moral and evaluative meanings in ways that align more closely with dominant English-language distributions than with language-specific conceptual systems (Havaldar et al., 2023). For a comprehensive review of over 300 papers related to benchmarks and methods used for cultural inclusion in multimodal language models, see Pawar et al. (2025).

In this study, we use cultural semantics in an operational, non-essentialist sense: we refer to language-specific patterns of meaning organization that reflect shared conventions, norms, and evaluative distinctions in usage, without equating a language with a single homogeneous culture. Accordingly, cultural semantic integrity denotes the extent to which these language-specific meaning organizations remain stable under English-mediated processing. Importantly, the paper does not claim to 'measure culture' directly; instead, it provides diagnostics for representational semantic distortion that may disproportionately affect culturally embedded distinctions in low-resource settings.

At the representation level, modern NLP systems rely heavily on distributional semantic embeddings, which map texts from different languages into a shared vector space. Models such as LaBSE and related multilingual sentence encoders were designed in order to support cross-lingual semantic comparability by means of aligning representations across languages (Feng et al., 2022). While these models achieve impressive performance on semantic similarity and retrieval benchmarks, prior research on bias in distributional representations has shown that embedding spaces can preserve and even amplify latent structural biases present in the training data (Lauscher and Glavaš, 2019). Consequently, similarity in embedding space fails to guarantee the preservation of culturally salient semantic distinctions and, more importantly, the relational topology of semantic neighborhoods.

Recent research on multilingual and culturally aware language models has been highlighting the challenge of preserving local semantic structures amid the influence of dominant languages (Guo et al., 2025). Surveys of multilingual LLMs indicate that English frequently serves as a latent pivot language, shaping semantic organization even when models generate text directly in other languages (Guo et al., 2025; Nie et al., 2024). This suggests that non-English texts generated or transformed by LLMs might undergo a form of implicit semantic normalization, aligning them with English-centric discourse conventions.

In spite of these advances, existing evaluation frameworks predominantly concentrate on the answers provided by LLMs to culturally or socially sensitive questions or, in case of text analysis, on surface-level attributes such as fluency or translation accuracy. However, metrics commonly used to evaluate machine translation or text generation quality are not designed to detect structural semantic changes. Consequently, subtle yet systematic semantic reorganizations of texts influenced by LLMs might go unnoticed, especially in non-English and low-resource languages.

In this study, this gap is being addressed by means of introducing a framework for analyzing textual semantic integrity through LLM-based English-mediated back-translation. Rather than evaluating translation quality, we concentrate on how the textual semantic structure changes when texts pass through English as an intermediate representational space. Using back-translation through English as a controlled diagnostic probe, we make implicit semantic re-projection effects explicit and measurable. We operationalize semantic integrity by means of multilingual text-level embeddings and

quantify changes at three complementary levels: semantic self-similarity, neighborhood preservation, and the stability of culturally salient semantic oppositions. This framework is applied to two languages with distinct resource profiles: Lingala, representing a low-resource language with language-internal semantic structures, and Russian, representing a high-resource language with a well-developed non-English discourse tradition. By analyzing each language independently within a unified embedding space, we demonstrate that English-mediated processing induces systematic semantic reorganization across resource settings, with varying magnitudes and manifestations.

What is important is that, although the ongoing study makes use of back-translation as an explicit transformation, our goal is not to model real-world translation pipelines. Instead, back-translation serves as a methodological tool in order to isolate Anglocentric semantic re-projection, which also occurs implicitly during direct LLM text generation in non-English languages. Therefore, the suggested metrics are therefore applicable beyond translation scenarios, providing a foundation for analyzing semantic shifts in LLM-generated Russian texts and other non-English outputs.

This paper makes the following contributions:

1. Conceptual contribution. We introduce semantic integrity as a distinct evaluation dimension for multilingual LLM processing, emphasizing the preservation of structural semantic relationships rather than surface-level fluency or translation accuracy.

2. Methodological contribution. We set forth using back-translation through English as a controlled diagnostic probe that makes implicit Anglocentric semantic re-projection observable.

3. Metric contribution. We operationalize semantic integrity using a concise set of embedding-based metrics. Unlike standard evaluation metrics, these indicators capture changes in relational meaning and semantic topology.

4. Empirical contribution. We make use of the suggested framework to Russian and Lingala, which represent high-resource and low-resource languages, respectively.

Thus, by shifting the focus from surface adequacy to relational meaning, this study contributes to a more nuanced understanding of multilingual LLM behavior and highlights the need for evaluation frameworks that account for the preservation of cultural and semantic structures.

2. RELATED WORK

Although the ongoing study does not aim to assess translation quality of the texts, previous research on multilingual language models, machine translation, and text generation offers crucial context for understanding how semantic preservation is operationalized in evaluation practices. In most existing frameworks, semantic preservation is implicitly equated with semantic similarity, which is typically estimated either through surface-level overlap or embedding-based similarity between the system output and a reference text. The gradual transition from surface-level, n-gram overlap metrics to embedding-based similarity measures reflects significant progress in capturing meaning beyond lexical form, but also reveals limitations that motivate the current work.

Because similarity-based evaluation ultimately relies on the properties of multilingual representations learned during pretraining, it is essential to consider how these representations are formed and what biases they encode. Modern multilingual pretraining approaches, such as mBERT and related models trained using masked language modeling, have shown that one neural model can support cross-lingual transfer across many languages by means of learning shared language representations (Conneau et al., 2020). However, in practice, the distribution of training data remains heavily skewed toward high-resource languages, with English dominating both web-scale corpora and downstream benchmarks. Recent analyses

estimate that over 90% of the training data for many language models consists of English text (Brinkmann et al., 2025). As a result, multilingual representations frequently display systematic centering effects, where linguistic organization reflects dominant English-language distributions even when processing non-English inputs (Papadimitriou et al., 2023; Guo et al., 2024).

This English-dominated representational landscape is mirrored in machine translation, where English frequently functions as an explicit or implicit pivot language. Machine translation provides a particularly transparent case of English-mediated processing, as semantic representations are explicitly routed by means of English as an intermediate space, making the effects of English dominance directly observable. Parallel work in multilingual machine translation has therefore sought to reduce the English bottleneck by developing many-to-many architectures and large-scale data mining strategies designed in order to minimize reliance on English as an intermediate representation (Fan et al., 2021). While such approaches mitigate explicit English-centric pipeline dependencies, they fail to eliminate deeper representational influences arising from English-dominated pretraining.

A central methodological consequence of English-mediated processing concerns how meaning preservation has been evaluated in practice. Historically, evaluation of machine translation relied primarily on surface n-gram overlap metrics, with BLEU as the canonical example (Papineni et al., 2002). Although effective for measuring lexical correspondence, such metrics correlate imperfectly with meaning preservation, particularly for typologically distant languages, paraphrastic variation, or culturally loaded expressions. These limitations motivated the development of embedding-based or learned metrics intended in order to approximate semantic similarity beyond exact lexical overlap (Sun and Wang, 2024).

Metrics such as BERTScore (Zhang et al., 2019) operationalize semantic similarity through contextualized token representations and alignment-based matching, while learned metrics such as COMET improve correlation with human judgments by leveraging neural embeddings and contextual modeling (Rei et al., 2020). However, in spite of their use of neural representations, these approaches remain fundamentally reference-based and are optimized to assess semantic alignment with a target output. As such, they are not designed to test whether semantic identity and relational structure remain stable under transformation via translation.

A related perspective on transformation effects emerges from research on translationese (Liu et al., 2025), which has shown that translated texts display systematic surface-level and stylistic deviations from texts originally written in the target language. These deviations have been extensively studied as detectable traces of translation as a categorical phenomenon. While this line of work provides important insights into stylistic and distributional effects of translation, it primarily operates at the level of surface form and register.

We argue that English-centric processing in LLMs leaves deeper traces than those typically captured by translationese studies. Rather than producing only stylistic fingerprints, English mediation can induce structural and directional imprints in semantic space itself. From this perspective, translationese can be understood as a surface manifestation of more fundamental representational distortions. The metrics set forth in the ongoing study are designed in order to capture these deeper effects by analyzing semantic self-consistency and relational stability, rather than stylistic deviation alone.

These issues become particularly salient in culturally specific and low-resource settings. Recent advances in machine translation, especially multilingual pretrained models, have considerably improved translation quality for many low-resource

languages. Benchmarks such as Ethiobenchmark (Tonja et al., 2024) and models such as Toucan, an Afrocentric MT system supporting 156 African language pairs (Elmadany et al., 2024), demonstrate notable gains in translation accuracy across African languages.

In spite of this progress, a persistent gap remains in the effective translation and evaluation of culturally specific content due to the inherent cultural differences associated with various languages, not fully captured through MT techniques (Pawar et al., 2025). Moreover, many low-resource and mid-resource languages lack scalable evaluation benchmarks, and reliance on human annotation or community-driven data collection, while essential, is difficult to scale (Nekoto et al., 2020). Together, these limitations highlight the need for evaluation approaches that do not depend on large parallel corpora or culturally exhaustive reference sets, and that can operate under conditions of limited data availability (Rashid et al., 2025; Pawar et al., 2025).

One practical strategy for such evaluation is to make use of controlled transformations as diagnostic probes rather than as approximations of real-world pipelines. As such strategy, we apply backtranslation, a classic validation method of machine translation in which scholars compare an original text to a version of that text that has been translated from its original language to another language (in our case English) and then back again to the original (Moon et al., 2020). Recent work has explored backtranslation paired with sentence embeddings as a diagnostic strategy for assessing meaning preservation beyond surface overlap (Chew et al., 2025). Exceptional adaptability of LLMs in backtranslation (namely, GPT-4o-mini) was also indicated in recent paper (Wang and Lin, 2025).

The ongoing work extends this diagnostic perspective by means of systematically combining English-mediated back-translation with structure-sensitive embedding analysis. Unlike standard MT metrics such as BLEU or BERTScore which evaluate translation accuracy relative to a reference, our metrics are reference-free and diagnostic in nature. By concentrating on semantic self-similarity, neighborhood preservation, and directional drift in embedding space, we capture changes in semantic integrity that remain invisible to surface-level or reference-based evaluation metrics.

3. MATERIALS AND METHODS

3.1. Data Collection and Corpus Construction

3.1.1. Russian Corpus

The Russian dataset was drawn from an existing experimental corpus of written texts produced by the same group of authors under controlled conditions (Litvinova et al., 2016; Litvinova, 2021). All of the texts were originally authored by native Russian speakers and annotated with multiple metadata variables, including text type, author identifiers, and psychometric measures. For the purposes of the ongoing study, only the text content and text type annotations were made use of.

Two text types were selected: descriptive texts and narrative texts. These genres were chosen because they represent common and functionally distinct modes of written discourse while being produced by the same authors, thereby controlling for individual stylistic and lexical variation. In order to avoid sample size imbalance and genre-related confounds, an equal number of texts was randomly sampled from each genre.

Prior to sampling, texts were filtered to remove empty entries and non-textual artifacts. Document length was measured as the number of whitespace-separated tokens. After genre balancing, an empirical length window was determined based on the 10th and 90th percentiles of word counts in the balanced Russian subset. Only texts falling within this window were retained. Following this procedure, a total of 483 Russian texts were selected for further analysis. Document

lengths in the final Russian corpus ranged from 74 to 244 words.

The Russian corpus was treated as a single-language reference set; text type was not modeled as an independent experimental factor because the analysis focuses on within-text semantic stability under transformation, rather than on between-genre variation.

3.1.2. Lingala Corpus

The Lingala data were derived from a large cleaned corpus of Latin-script Lingala text containing over 500,000 textual lines (*NLLB Team*). The original corpus had previously undergone standard normalization procedures, including lowercasing, removal of punctuation and digits, whitespace normalization, and trimming of empty lines.

Because the Lingala corpus consisted of short textual units rather than document-level texts, an additional document construction step was required. Lingala documents were created by concatenating consecutive cleaned lines into fixed-size blocks. The number of lines per block was treated as a tunable parameter and optimized to produce document-level texts comparable in length to the Russian corpus. It is to be noted that while concatenation might introduce heterogeneous discourse boundaries, this procedure was applied uniformly and was intended to approximate document-level semantic structure rather than discourse coherence.

Specifically, candidate block sizes ranging from 6 to 30 lines were evaluated. For each block size, documents were constructed and their word counts computed. The block size that maximized the number of documents falling within the Russian length window (74–244 words) was selected. This procedure yielded a block size of six lines per document, producing Lingala documents with a median length of approximately 105 words.

Following the document construction, Lingala texts were filtered using the same length window as the Russian corpus. A random sample of 483 Lingala documents was then selected to exactly match the size of the Russian dataset. Final Lingala document lengths ranged from 74 to 224 words, with a mean of approximately 110 words. This length-matching strategy was adopted to reduce confounding effects of document length on embedding-based similarity measures.

3.1.3. Corpus Alignment and Final Dataset

The final dataset consisted of 966 documents in total, comprising 483 Russian and 483 Lingala texts. The two language subsets were balanced with respect to the number of documents and broadly comparable in document length distributions. Word count was retained as a tracked variable and later used for robustness checks to ensure that observed effects were not driven by length differences.

Each document was assigned a unique identifier and labeled with a language tag (RU or LN). Only the document identifier, language label, and text content were used as inputs to subsequent processing stages, making sure that all of the downstream analyses operated on fully anonymized textual data.

This corpus construction strategy was designed to satisfy three key criteria: 1) comparability across languages in terms of document length and format; 2) control over author-related variability in the Russian data; 3) scalability to larger corpora and additional languages.

What is important is that the goal of this stage was not to create linguistically homogeneous or genre-pure corpora, but to construct document-level datasets suitable for probing the stability of semantic structure under controlled transformation using embedding-based methods.

3.2. Methodology

3.2.1. English-Mediated Back-Translation as a Diagnostic Probe

Rather than treating translation as an end task, we make use of English-mediated back-translation as a controlled diagnostic probe in order to identify structural semantic

changes induced by English-centric processing. The key idea is to keep the intended meaning fixed by design while forcing the text to pass through English as an intermediate linguistic channel.

Formally, for each original text T in language L, we construct an English-mediated back-translated version T^{BT} using a two-step translation process. The text is first translated from L into English, yielding T^{EN}, and then translated back into the original language (Figure 1).

Figure 1. Scheme of English-mediated back-translation as a diagnostic probe
Рисунок 1. Англо-опосредованный обратный перевод как диагностический инструмент

$$T \xrightarrow{L \to \mathrm{EN}} T^{EN} \xrightarrow{\mathrm{EN} \to L} T^{BT}$$

This transformation makes English an explicit intermediate representation, enabling controlled analysis of semantic transformations induced by English mediation.

Both translation steps are performed by means of the same large language model (gpt-4o-mini) via the OpenAI Responses API. For each step, decoding was run with temperature = 0 and a fixed prompt template instructing the model to preserve meaning and to output only translations. Namely, we used the following prompt: **You are a precise translator. Translate the user's text to {TARGET_LANGUAGE}. Preserve meaning, register, and named entities. Avoid unnecessary rephrasing. Do not add explanations, comments, or extra text. Output only the translation. Source language: {SOURCE_LANGUAGE}.** To prevent output expansion, fixed maximum number of output tokens was imposed. Requests were cached by a hash of the input text and language to ensure reproducibility and to allow resumable processing under rate limits. This design minimizes stochastic generation effects and makes sure that observed differences between T and T^{BT} are unlikely to be driven by sampling variability.

What is of importance is that we do not interpret back-translation quality in terms of fluency or adequacy. Instead, back-translation serves as a stress test of semantic stability: if semantic representations are stable under English mediation, then T and T^{BT} should occupy similar positions in semantic embedding space. Conversely, systematic divergence indicates that English acts as a non-neutral filter capable of reorganizing semantic structure.

This probing strategy has two advantages. First, it isolates the effect of English mediation independently of downstream task objectives or application-specific evaluation criteria. Second, it applies uniformly to both high-resource and low-resource languages, enabling controlled cross-linguistic comparison. It is to be noted that although back-translation is used as an explicit transformation in this study, it is not intended to model real-world translation pipelines. Rather, it functions as a methodological tool for exposing Anglocentric semantic re-projection mechanisms that may also operate implicitly during direct non-English text generation by means of multilingual LLMs.

3.2.2. Semantic Representation and Embedding Space

In order to quantify semantic stability under English-mediated transformation, all of the texts were embedded by means of a multilingual neural embedding model (text-embedding-3-small via OpenAI Responses API). The embedding model supports multilingual inputs and produces representations for both high-resource and low-resource languages, including Lingala, within a shared vector space.

Although such models are frequently referred to as “sentence embedding” models, in the present study embeddings are computed over entire documents rather than individual sentences. Each input text – original or back-translated – is passed to the embedding model

as a single sequence, yielding one vector per document. This choice is indicative of the analytical focus of the study: semantic integrity is evaluated at the level of global textual meaning, rather than at the level of isolated lexical items or sentence fragments.

This text-level representation is essential in order to capture higher-order semantic relations, including thematic coherence, evaluative orientation, and relational structure across concepts, which might not be observable from word-level or sentence-level embeddings alone.

Because all of the texts are embedded by means of the same model and the same representational geometry, differences between vectors can be interpreted as semantic differences induced by transformation, rather than as artifacts of incompatible embedding spaces. In particular, comparing each text to its own back-translated version within the same space avoids the need for external alignment procedures or language-specific normalization.

This approach is different from traditional cross-lingual studies that rely on separate monolingual embedding spaces aligned post hoc via bilingual dictionaries or linear transformations. By means of operating directly in a unified multilingual space, we make sure that observed semantic shifts reflect changes in representation induced by English mediation, rather than alignment noise or cross-space mapping error.

What is of importance is that all of the semantic comparisons are performed within-text, comparing each document only to its own back-translated version. This design ensures that any systematic differences observed cannot be attributed to topic variation, author identity, or content mismatch, but instead are indicative of changes induced by the English-mediated transformation itself.

It is important to highlight that the embedding model is treated as a measurement instrument rather than a ground-truth representation of meaning. While the same multilingual embedding space is used for all languages, the absolute fidelity with which different languages are represented might vary. However, because all analyses rely on relative, within-text comparisons between an original text and its back-translated counterpart, transformation-induced semantic shifts can be assessed reliably even in low-resource settings.

3.2.3. Metrics

Semantic Self-Similarity (SSI)

The first evaluation metric captures local semantic stability between a text and its English-mediated reconstruction. For each text *i*, we compute **Semantic Self-Similarity (SSI)** as the cosine similarity (1) between the embedding of the original text $\boldsymbol{e_i}$ and that of its English-mediated back-translated version $\boldsymbol{e^{BT}_i}$:

$$\mathrm{SSI}_i = \cos(\mathbf{e}_i, \mathbf{e}_i^{BT}). \quad (1)$$

SSI quantifies the degree to which a text remains semantically recognizable to itself after passing through English as an intermediate representational channel. Unlike BERTScore, which evaluates semantic alignment between an output and a reference, SSI evaluates whether a text remains semantically recognizable to itself after English-mediated transformation.

In practice, SSI values observed in our experiments are in the interval [0,1], with higher values corresponding to greater semantic stability. SSI is a purely local, within-text metric: it evaluates each document in isolation and does not take into account its position relative to other texts in the embedding space. As a result, SSI alone fails to capture changes in relational or structural meaning, such as collective shifts of texts within the semantic space or reorganization of local semantic neighborhoods.

Consequently, high SSI does not imply full preservation of semantic structure at the corpus level. Texts might remain individually recognizable while undergoing systematic reconfiguration in their relations to other

texts. For this reason, SSI is complemented by structure-sensitive metrics that explicitly capture neighborhood stability and directional semantic drift.

Neighborhood Preservation Score (NPS)

In order to assess structural semantic stability, we introduce the **Neighborhood Preservation Score (NPS)**, which evaluates whether a text retains its relative position among other texts in the semantic space after English mediation. Unlike SSI, which captures semantic recognizability in isolation, NPS explicitly measures changes in relational meaning.

For each original document embedding $\boldsymbol{e_i}$, we identify its set of k nearest neighbors N_i within the corpus by means of cosine similarity. The same procedure is applied to the embedding of back-translated text $\boldsymbol{e^{BT}_i}$, yielding a corresponding neighbor set $\boldsymbol{N^{BT}_i}$. Nearest-neighbor search is performed within the same language subset to avoid cross-lingual contamination.

Neighborhood preservation (2) is quantified using the Jaccard similarity between the two neighbor sets:

$$\mathrm{NPS}_i = \frac{|\mathcal{N}_i \cap \mathcal{N}_i^{BT}|}{|\mathcal{N}_i \cup \mathcal{N}_i^{BT}|}. \quad (2)$$

In the ongoing study, we fix k = 20, balancing sensitivity to local semantic structure against robustness to noise. Jaccard similarity is used because it captures changes in neighborhood composition independently of rank order, concentrating on whether semantic affiliations are preserved rather than on fine-grained distance fluctuations.

NPS values close to 1 indicate that a text maintains its relative position among semantically similar texts, while low values signal a reorganization of semantic topology. Crucially, a text may exhibit high SSI but low NPS, indicating that while its overall meaning remains recognizable, its semantic affiliations within the corpus have shifted. This distinction is central to present analysis: NPS captures the extent to which English mediation reshapes the local structure of language-specific semantic spaces, revealing changes that remain invisible to text-level similarity measures alone.

Directional Semantic Drift along Interpretable Axes

In order to examine whether English-mediated processing induces directional semantic bias beyond structural instability, we performed an axis-based analysis by means of predefined moral–evaluative oppositions. For each language, an interpretable semantic axis was constructed from a small set of seed words representing positive and negative poles of a broadly shared moral dimension. For Russian, the positive pole was defined by the words *честность* (honesty), *достоинство* (dignity), and *справедливость* (justice), while the negative pole consisted of *ложь* (lie), *предательство* (betrayal), and *подлость* (meanness). For Lingala, the positive pole included *malamu* (good) and *bosantu* (purity/virtue), and the negative pole included *mabe* (bad) and *mbindo* (impurity).

The axis vector was defined as the difference between the mean embeddings of the positive- and negative-pole seed words, by means of the same multilingual embedding model as for document representations. Each document embedding was projected onto the corresponding language-specific axis using cosine similarity, yielding an axis projection score that indicates the document's position along the semantic dimension.

Directional semantic drift is quantified as the difference between axis projections of the back-translated and original text:

$$\Delta\mathrm{AXIS}_i = \cos(\mathbf{e}_i^{BT}, \mathbf{a}) - \cos(\mathbf{e}_i, \mathbf{a}), \quad (3)$$

where e_i and $e^{BT}{}_i$ denote the embeddings of the original and back-translated versions of document i, respectively, and **a** denotes the axis vector.

Positive drift values indicate systematic movement toward the positive pole of the semantic axis under English mediation, while negative values indicate movement toward the negative pole. Importantly, the axis is used here as a diagnostic probe rather than as an exhaustive model of moral semantics. The goal is not to fully characterize cultural meaning, but rather to test whether English mediation induces consistent directional reorientation along an interpretable semantic dimension. In order to compare drift distributions across languages, we applied a Wilcoxon rank-sum test.

3.2.4. Summary of Evaluation Framework

Together, SSI, NPS, and axis-based drift provide a three-level diagnostic framework for evaluation of semantic integrity under English mediation. SSI (local stability) assesses whether a text remains semantically recognizable in isolation. NPS (structural stability) evaluates whether a text preserves its local semantic neighborhood within the corpus, capturing changes in relational meaning and semantic topology. Axis drift (directional bias) captures systematic movement along interpretable semantic dimensions induced by English mediation.

By means of combining these complementary metrics, the framework enables us to distinguish between qualitatively different modes of semantic change: surface-level semantic preservation (high SSI), structural semantic reorganization (low NPS), directional semantic bias (non-zero axis drift), semantic breakdown (low SSI and low NPS).

Applying this framework to a low-resource language (Lingala) and a high-resource reference language (Russian) allows us to disentangle baseline effects of English-mediated processing from language-specific vulnerability. The suggested methodology is model-agnostic, scalable, and directly applicable both to translation-based diagnostic probes and to the analysis of texts generated directly by LLMs in non-English languages.

4. RESULTS

4.1. Semantic Self-Similarity under English-Mediated Back-Translation

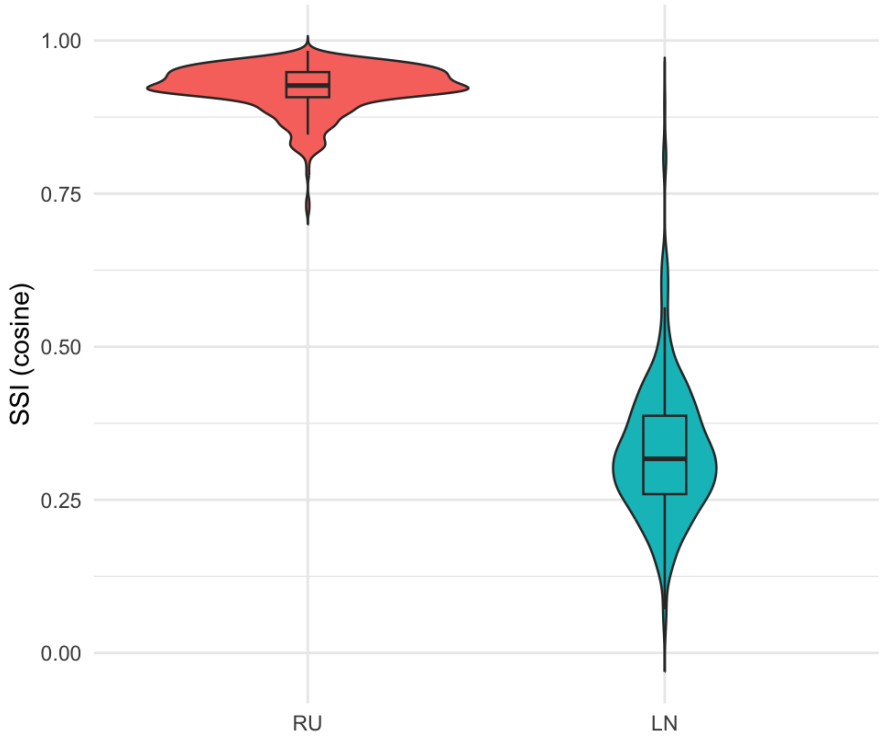
Semantic Self-Similarity (SSI) reveals a pronounced asymmetry between Russian and Lingala texts under English-mediated back-translation.

For Russian, SSI values are consistently high (mean = 0.922; median = 0.926), with a narrow interquartile range (Q1 = 0.907, Q3 = 0.948). This indicates that, at the level of pairwise semantic similarity, Russian texts largely preserve their original meaning after passing through English.

Conversely, Lingala displays considerably lower SSI (mean = 0.329; median = 0.317), with a broader dispersion (Q1 = 0.259, Q3 = 0.387). These values indicate that English-mediated processing substantially reduces semantic self-recognizability for Lingala texts even at the level of self-similarity (Figure 2).

Importantly, the Russian results demonstrate that high-resource languages may appear semantically stable when evaluated solely through local similarity metrics, potentially masking deeper representational shifts.

Figure 2. Semantic self-similarity (SSI) by language
Рисунок 2. Семантическое сходство (SSI) по языкам



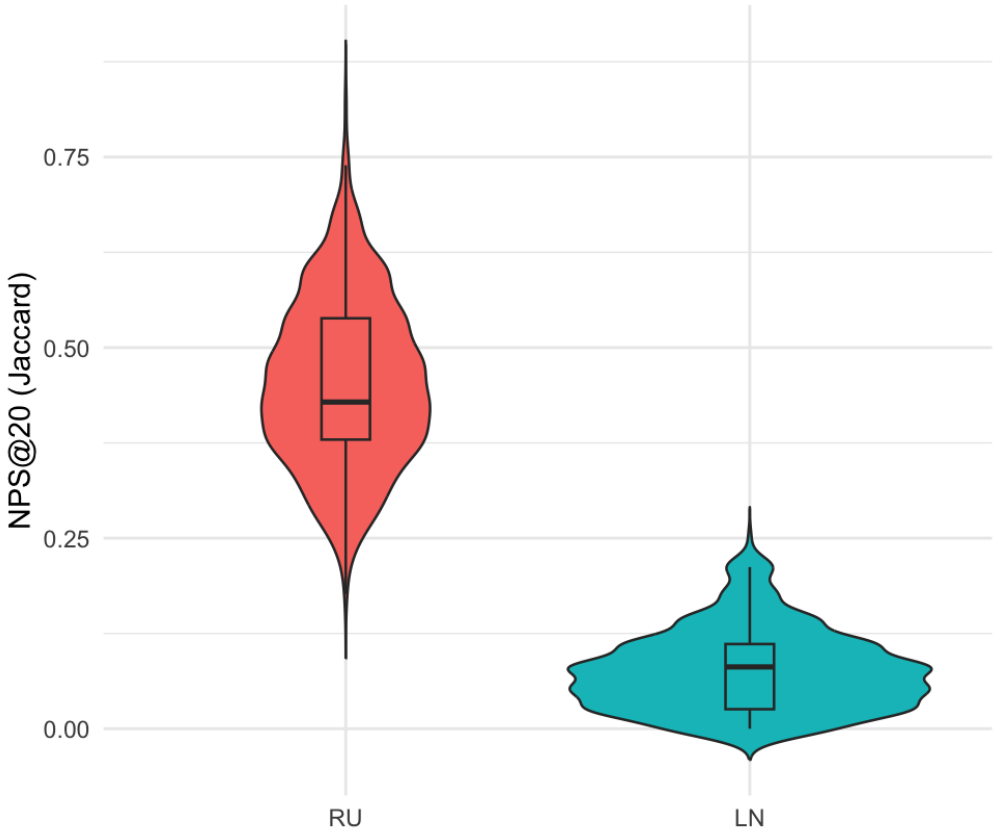
4.2. Neighborhood Preservation and Structural Semantic Stability

Neighborhood Preservation Score (NPS) provides a complementary view by means of capturing changes in relational semantic structure.

Russian texts retain a substantial portion of their semantic neighborhoods after back-translation (mean = 0.453; median = 0.429), indicating moderate structural stability. Although imperfect, this level of preservation suggests that English mediation fails to radically reorganize the Russian semantic space.

Lingala texts, conversely, display extremely low neighborhood preservation (mean = 0.075; median = 0.081). In lots of cases, the intersection between original and back-translated nearest-neighbor sets is minimal, with first-quartile values close to zero. This implies that, even when some surface meaning is retained, Lingala texts are systematically repositioned within the semantic space after English mediation (Figure 3).

Figure 3. Neighborhood Preservation Score (NPS) by language
Рисунок 3. Коэффициент сохранности семантических соседей (NPS) по языкам



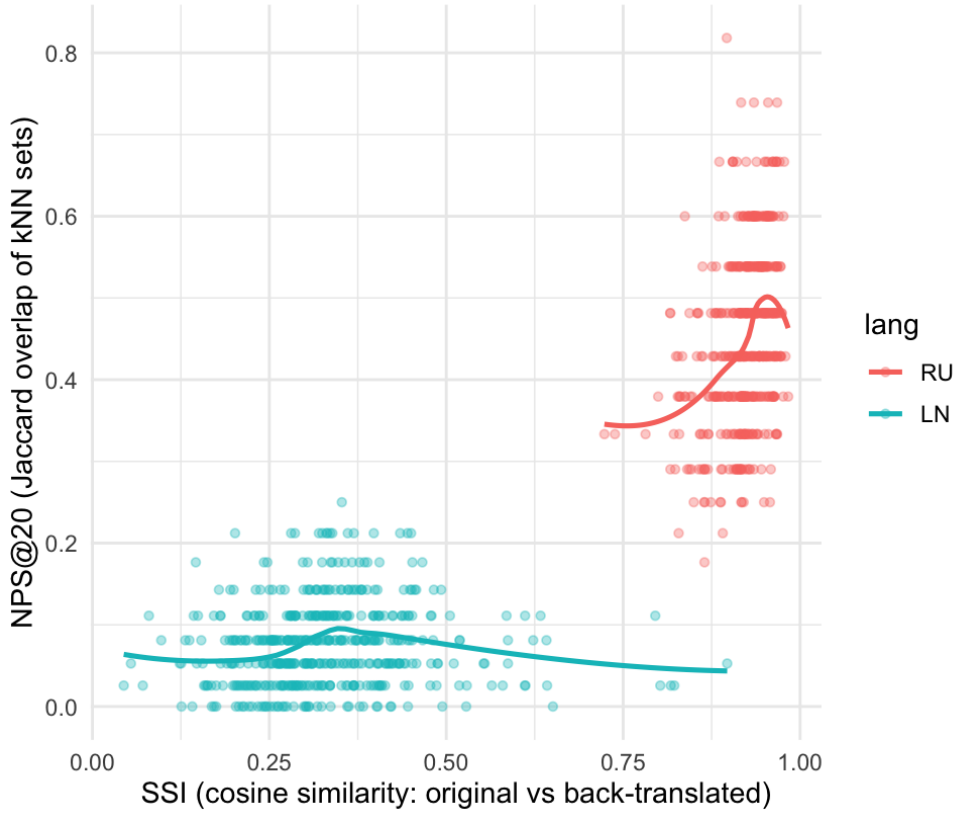
The observed effect sizes are exceptionally large (Cohen's d = 7.09 for SSI and d = 4.47 for NPS). This magnitude indicates not a marginal shift but rather a near-complete separation between the distributions of high-resource and low-resource languages under English-mediated processing.

Rank-based effect sizes (Cliff's δ=0.99) further confirm near-complete distributional separation between languages. In over 98% of random pairings, Russian texts display higher semantic stability than Lingala texts.

These results indicate that English mediation disproportionately disrupts the relational semantics of low-resource language, fragmenting culturally and contextually grounded meaning structures.

Taken together, SSI and NPS indicate a critical dissociation between local semantic preservation and structural semantic integrity (Figure 4).

Figure 4. Local vs. structural semantic stability by language
Рисунок 4. Локальная и структурная семантическая стабильность по языкам



As Figure 4 clearly shows, for Russian, high SSI coexists with moderate NPS, suggesting that while individual texts remain semantically recognizable, their relative positions within the semantic landscape shift non-trivially. This effect would remain largely invisible under surface-level evaluation metrics. For Lingala, both SSI and NPS are severely degraded, indicating not only loss of semantic content but also a breakdown of culturally coherent semantic organization. This dissociation indicates that English-mediated multilingual processing can preserve apparent meaning while simultaneously reshaping deeper semantic relations, and this effect is considerably more pronounced for low-resource languages.

English-mediated processing does not induce gradual semantic degradation. Instead, it produces a regime shift in semantic stability profiles: high-resource languages retain surface semantic identity while low-resource languages experience structural semantic collapse.

4.3. Directional Semantic Drift along an Interpretable Axis

Beyond structural instability, we examine whether English mediation induces directional semantic bias along an interpretable moral axis.

We compute axis projections for original and back-translated texts and define axis drift as the difference between these projections. Lingala texts display a small albeit systematic positive drift along this axis (median ΔAXIS $\approx$ 0.036), whereas Russian texts remain near zero (median ΔAXIS $\approx$ 0.003). A Wilcoxon rank-sum test confirms a considerable cross-linguistic difference ($p < 2.2e{-}16$).

It is to be noted that the magnitude of axis drift is not correlated with neighborhood preservation within either language,

indicating that directional semantic bias and structural semantic reorganization constitute distinct effects of English mediation.

4.4. Robustness to Text Length and Segmentation

In order to assess whether the observed effects could be driven by segmentation artifacts or fragment length, we examined the relationship between semantic stability metrics and text length.

Within both languages, correlations between text length and either SSI or NPS are weak ($|\rho| < 0.12$), indicating that variation in document length fail to considerably account for differences in semantic stability. Furthermore, restricting the analysis to the longest quartile of Lingala texts preserves the same qualitative pattern: Lingala remains markedly less stable than Russian in both SSI and NPS.

Together, these findings rule out segmentation conventions and fragment length as primary explanations for the observed cross-linguistic asymmetry.

4.5. Illustrative embedding-level response to English-mediated semantic drift (Lingala)

To improve interpretability, we include a short qualitative example illustrating how the three metrics respond to English-mediated semantic drift.

Original (Lingala):

nazongaki mboka na ndenge mibale elingi koloba ngai moko moto nakómaki lisusu epai na biso mpe nakómaki sikoyo na likoki ya komipesa na makambo ya nzambe … ata na boyangeli ya bakoministe nasengelaki kokangama na solo mpe kosakola bokonzi ya nzambe lokola elikya bobele moko mpo na banyokwami …

English-mediated back-translation (Lingala):

Mi retorné al país de dos maneras, queriendo decirme a mí mismo que una vez más había vuelto a nosotros, y ahora estoy en posición de comprometerme con los asuntos de Dios … incluso bajo el liderazgo comunista, tenía que aferrarme a la verdad y proclamar la soberanía de Dios como la única esperanza para los oprimidos …

While the back-translated version remains fluent and preserves the general narrative structure, a subtle yet systematic shift is observed in the moral framing and agency structure of the text. In particular, expressions of personal commitment and moral obligation are rendered more neutral and report-like, with reduced emphasis on internal resolve and evaluative intensity.

At the embedding level, this transformation is reflected across all three proposed diagnostics. Semantic Self-Similarity (SSI) decreases, indicating reduced within-text semantic recognizability after English mediation. Neighborhood Preservation Score (NPS) also drops, suggesting that the text is repositioned among semantically different neighbors and loses part of its original relational context. Crucially, the text exhibits a directional semantic drift along the moral–evaluative axis, moving toward a more neutral or attenuated region of the space.

This example illustrates how English-mediated processing can preserve surface plausibility while simultaneously reshaping relational meaning and evaluative orientation effects that remain largely invisible to surface-level similarity metrics but are systematically captured by the proposed framework.

5. DISCUSSION

The ongoing study provides a fine-grained analysis of how English-centric processing affects semantic integrity across languages with different resource profiles. Rather than a uniform degradation, we identify two distinct and partially independent mechanisms by means of which English mediation reshapes meaning.

First, English mediation induces structural semantic reorganization, indicated in reduced neighborhood preservation. This effect is present even for a high-resource language such as Russian, where surface semantic similarity remains high.

Consequently, local similarity metrics alone can give a misleading impression of semantic robustness, as they fail to capture changes in relational meaning.

Second, for the low-resource language Lingala, English mediation additionally produces directional semantic bias along an operationally interpretable moral–evaluative dimension. This bias manifests as a systematic axis drift that is absent or negligible in Russian. What is of importance is that this directional shift operates independently of structural instability, suggesting that English mediation can simultaneously reshape semantic topology and nudge texts along specific cultural coordinates.

Together, these findings argue against treating English mediation as a neutral or purely technical intermediate step. Instead, it functions as an active representational filter whose effects depend on the resource status and cultural embedding of the source language.

From a methodological perspective, our results demonstrate the limitations of surface-level evaluation metrics and motivate the inclusion of relational and directional diagnostics when assessing multilingual representations. From a broader perspective, the findings raise concerns about the uncritical deployment of English-centric multilingual models in contexts involving low-resource languages and culturally sensitive content.

Finally, while this study concentrates on translation-mediated probing, the suggested framework naturally extends to the analysis of LLM-generated texts, where English-centric training might induce similar forms of structural drift and directional bias without any explicit translation step.

This study has several ***limitations***. First, the use of English-mediated back-translation serves as a diagnostic probe rather than a direct model of real-world text generation. Although this design isolates the effect of English mediation, future work should apply the proposed metrics directly to LLM-generated texts in non-English languages.

Second, the analysis relies on a single interpretable semantic axis with a limited lexicon, chosen for illustrative purposes. While sufficient to demonstrate directional bias, the use of richer interpretable axes, potentially grounded in semantic primitives or mental lexicon frameworks, may strengthen the connection between directional semantic drift and culturally specific meaning structures.

Third, the study focuses on two languages representing contrasting resource profiles. Extending the framework to a broader typological and cultural range of languages would further clarify the generality of the observed mechanisms.

Finally, while embedding-based metrics capture relational semantic structure, they inherit biases from the underlying embedding models. Future work should explore model-specific variability and assess robustness across alternative representation spaces.

Statistical hypothesis testing across metrics (e.g., variance-based analyses) may provide a more fine-grained comparison of robustness profiles across models and languages.

In spite of these limitations, the present study establishes a principled and scalable framework for diagnosing semantic integrity under English-centric processing and opens multiple avenues for investigating multilingual fairness in large language models.

6. CONCLUSIONS

This work demonstrates that English-centric processing induces non-uniform and multi-level semantic effects across languages. By means of separating local semantic preservation from structural semantic stability, we show that high surface similarity fails to guarantee preservation of relational meaning.

Within this framework, cultural semantic integrity should be understood not as a property of a language or a cultural group per se, but as a representational property of multilingual LLM processing. It captures the

extent to which language-specific semantic organizations shaped by shared norms, evaluative distinctions, and usage patterns remain structurally and directionally stable under English-centric mediation. By disentangling surface-level semantic recognizability, relational semantic structure, and directional evaluative bias, the proposed metrics provide an operational and non-essentialist approach to assessing cultural semantic integrity in multilingual representations.

For Russian, English mediation preserves overall semantic recognizability while inducing measurable structural reorganization — an effect that would remain largely invisible under standard evaluation metrics. For Lingala, English mediation results in both structural semantic collapse and a systematic directional bias along a culturally interpretable axis. These effects are robust to text length and segmentation, underscoring their structural nature rather than being artifacts of corpus construction.

Crucially, our analysis reveals that structural reorganization and directional semantic bias operate as distinct mechanisms, rather than as points along a single continuum of degradation. This distinction provides a more nuanced understanding of how English-centric influence manifests in multilingual language models.

Methodologically, the study highlights the limitations of surface-level similarity metrics and motivates the inclusion of relational and directional diagnostics in multilingual evaluation. Substantively, the findings raise concerns about the uncritical use of English-mediated representations in low-resource and culturally sensitive contexts.

While the experiments make use of translation-based probing, the suggested framework naturally extends to the analysis of LLM-generated texts, where similar forms of structural drift and semantic bias may arise with no explicit translation. As such, the approach provides a foundation for future work on fairness, cultural integrity, and semantic reliability in multilingual large language models.

Declarations. We used service https://wordvice.ai for English proofreading, including spelling, grammar, and stylistic edits. This service did not generate substantive content or perform any analysis. No generative tool made interpretive or methodological decisions without human oversight, and no confidential or personally identifiable data were shared to third-party services.

References

Brinkmann, J., Wendler, C., Bartelt, C. and Mueller, A. (2025). *Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages*. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.312 *(In English)*

Chew, E., Chakraborti, M., Weisman, W. and Frey, S. (2025). Evaluating Machine Translation Solutions for Accessible Multi-Language Text Analysis: A Back-translation based Approach, *Computational Communication Research*, vol. 7, issue 1. https://doi.org/10.5117/CCR2025.1.5.CHEW *(In English)*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G. et al. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747 *(In English)*

Elmadany, A., Adebara, I. and Abdul-Mageed, M. (2024). Toucan: Many-to-Many Translation for 150 African Language Pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, 13189–13206, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.781 *(In English)*

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., et al. (2021). Beyond English-centric multilingual machine translation. J. *Mach. Learn. Res.* 22, 1, Article 107, 4839–4886. https://dl.acm.org/doi/abs/10.5555/3546258.3546365 *(In English)*

Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W. (2022). *Language-agnostic BERT Sentence Embedding*. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 878–891, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62 *(In English)*

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, Md M., Kim, S. et al. (2024). Bias and Fairness in Large Language Models: *A Survey*. *Computational Linguistics*, 50 (3): 1097–1179. https://doi.org/10.1162/coli_a_00524 *(In English)*

Guo, Y., Conia, S., Zhou, Z. Li, M. and Potdar, S. et al. (2025). *Do Large Language Models Have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs* In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3823–3838, July 27 – August 1, 2025, Association for Computational Linguistics [Online], Available at: https://aclanthology.org/2025.acl-long.193.pdf (Accessed 15.03.2026). *(In English)*

Havaldar, S., Rai, S., Singhal, B., Liu, L., Guntuku, S. et al. (2023). *Multilingual Language Models are not Multicultural: A Case Study in Emotion*. In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 202–214, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.19 *(In English)*

Lauscher, A. and Glavaš, G. (2019). Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (SEM 2019)*, 85–91, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1010 *(In English)*

Litvinova T., Litvinova, O., Zagorovskaya, O., Seredin, P., Sboev A. et al. (2016). *"Ruspersonality": A Russian corpus for authorship profiling and deception detection*, 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT), St. Petersburg, Russia, 2016, 1–7. https://doi.org/10.1109/FRUCT.2016.7584767 *(In English)*

Litvinova, T. (2021). *RusIdiolect: A New Resource for Authorship Stu*dies. In: Antipova, T. (eds) Comprehensible Science. ICCS 2020. Lecture Notes in Networks and Systems, vol 186. Springer, Cham. https://doi.org/10.1007/978-3-030-66093-2_2 *(In English)*

Liu, Y., Zhang, W., Wang, Y., Tang, J., Zhang, P. et al. (2025). Translationese-index: Using Likelihood Ratios for Graded and Generalizable Measurement of Translationese. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 12521–12538, Suzhou, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.633 *(In English)*

Moon, J., Cho, H. and Eunjeong L. P. (2020). Revisiting Round-trip Translation for Quality Estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 91–104, Lisboa, Portugal. European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.11/ *(In English)*

Narayan, M.A., Pasmore, J., Sampaio, E., Raghavan, V., Maity, S. et al. (2025). Mitigating Bias in Large Language Models Through Culturally-Relevant LLMs. In 2025 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS), Evanston, IL, USA, 2025, 1–7. https://doi.org/10.1109/ETHICS65148.2025.11098204 *(In English)*

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T. et al. (2020). *Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages*. In Findings of the Association for Computational Linguistics: EMNLP 2020, 2144–2160, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195 *(In English)*

Nie, S., Fromm, M., Welch, C., Görge, R., Karimi, A. et al. (2024). *Do Multilingual Large Language Models Mitigate Stereotype Bias*? In Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, 2024, 65–83. pages 65–83 August 16, 2024. Association for Computational Linguistics [Online], Available at: https://aclanthology.org/2024.c3nlp-1.6.pdf (Accessed 15.03.2026). *(In English)*

NLLB Team. Scaling neural machine translation to 200 languages. (2024). Nature 630, 841–846. https://doi.org/10.1038/s41586-024-07335-x *(In English)*

Papadimitriou, I., Lopez, K. and Jurafsky, D. (2023). Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, 1194–

1200, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.89 *(In English)*

Papineni, K., Roukos, S., Ward, T. and 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, 311–318. https://doi.org/10.3115/1073083.1073135 *(In English)*

Pawar, S., Park, J., Jin, J., Arora, A., Myung, J. et al. (2025). Survey of Cultural Awareness in Language Models: Text and Beyond, *Computational Linguistics* 51 (3): 907–1004. https://doi.org/10.1162/COLI.a.14 *(In English)*

Rashid, Q., Liemt, E., Shih, T., Ebinama, A., Ramos, K. et al. (2025). *Amplify Initiative: Building A Localized Data Platform for Globalized AI*, ArXiv, abs/2504.14105 https://doi.org/10.48550/arXiv.2504.14105 *(In English)*

Rei, R., Craig, S., Farinha, A.C. and Lavie, A. (2020). *COMET: A Neural Framework for MT Evaluation*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2685–2702, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213 *(In English)*

Sun, K. and Wang, R. (2024). Textual Similarity as a Key Metric in Machine Translation Quality Estimation, ArXiv, arXiv:2406.07440v1 [cs.CL]. https://doi.org/10.48550/arXiv.2406.07440 *(In English)*

Tonja A.L., Azime, I.A., Belay, T. D., Yigezu, M. G., Moges, A. Ah M. et al. (2024). EthioLLM: Multilingual Large Language Models for Ethiopian Languages with Task Evaluation. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 6341–6352, Torino, Italia. ELRA and ICCL [Online], Available at: https://aclanthology.org/2024.lrec-main.561/ (Accessed 15.03.2026). *(In English)*

Wang, Y. and Lin, Z. (2025). Revisiting Round-Trip Translation with LLMs and Agentic Translation. *2025 IEEE 5th International Conference on Software Engineering and Artificial Intelligence (SEAI)*, 172–178. https://doi.org/10.1109/SEAI65851.2025.11108752 *(In English)*

Xu, Y., Hu, L., Zhao, J., Qiu, Z., Ye, Y. and Gu, H. (2025). A Survey on Multilingual Large Language Models: Corpora, Alignment, And Bias. *Frontiers of Computer Science*, 19, 1911362. https://doi.org/10.1007/s11704-024-40579-4 *(In English)*

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y. (2019). *BERTScore: Evaluating Text Generation with BERT*. ArXiv, abs/1904.09675. https://doi.org/10.48550/arXiv.1904.09675 *(In English)*

Список литературы

Brinkmann, J., Wendler, C., Bartelt, C. and Mueller, A. (2025). *Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages* // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Pp. 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.312

Chew, E., Chakraborti, M., Weisman, W. and Frey, S. (2025). Evaluating Machine Translation Solutions for Accessible Multi-Language Text Analysis: A Back- translation based Approach // Computational Communication Research. Vol. 7, Issue 1. https://doi.org/10.5117/CCR2025.1.5.CHEW

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G. et al. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Pp. 8440–8451. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Elmadany, A., Adebara, I. and Abdul-Mageed, M. 2024. Toucan: Many-to-Many Translation for 150 African Language Pairs // Findings of the Association for Computational Linguistics: ACL 2024. Pp. 13189–13206, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.781

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., et al. (2021). Beyond English-centric multilingual machine translation // J. Mach. Learn. Res. 22, 1, Article 107. Pp. 4839–

4886. https://dl.acm.org/doi/abs/10.5555/3546258.3546365

Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W. (2022). *Language-agnostic BERT Sentence Embedding*. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Pp. 878–891. Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, Md M., Kim, S. et al. (2024). Bias and Fairness in Large Language Models: A Survey // Computational Linguistics. Vol. 50 (3). Pp. 1097–1179. https://doi.org/10.1162/coli_a_00524

Guo, Y., Conia, S., Zhou, Z. Li, M. and Potdar, S. et al. (2025). *Do Large Language Models Have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs* // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Pp. 3823–3838, July 27 – August 1, 2025, Association for Computational Linguistics [Online], Available at: https://aclanthology.org/2025.acl-long.193.pdf (Accessed 15.03.2026).

Havaldar, S., Rai, S., Singhal, B., Liu, L., Guntuku, S. et al. (2023). *Multilingual Language Models are not Multicultural: A Case Study in Emotion* // Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis. Pp. 202–214. Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.19

Lauscher, A. and Glavaš, G. (2019). Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors // *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (SEM 2019)*. Pp. 85–91, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1010

Litvinova T., Litvinova, O., Zagorovskaya, O., Seredin, P., Sboev A. et al. (2016). *"Ruspersonality": A Russian corpus for authorship profiling and deception detection* // 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT), St. Petersburg, Russia, 2016. Pp. 1–7, https://doi.org/10.1109/FRUCT.2016.7584767

Litvinova, T. (2021). *RusIdiolect: A New Resource for Authorship Studies* // Antipova, T. (eds) Comprehensible Science. ICCS 2020. Lecture Notes in Networks and Systems. Vol. 186. Springer, Cham. https://doi.org/10.1007/978-3-030-66093-2_2

Liu, Y., Zhang, W., Wang, Y., Tang, J., Zhang, P. et al. (2025). Translationese-index: Using Likelihood Ratios for Graded and Generalizable Measurement of Translationese // *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Pp. 12521–12538. Suzhou, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.633

Moon, J., Cho, H. and Eunjeong L. P. (2020). Revisiting Round-trip Translation for Quality Estimation // Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. Pp. 91–104. Lisboa, Portugal. European Association for Machine Translation [Online], Available at: https://aclanthology.org/2020.eamt-1.11/ (Accessed 15.03.2026).

Narayan, M.A., Pasmore, J., Sampaio, E., Raghavan, V., Maity, S. et al. (2025). Mitigating Bias in Large Language Models Through Culturally-Relevant LLMs // 2025 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS), Evanston, IL, USA, 2025, Pp. 1–7, https://doi.org/10.1109/ETHICS65148.2025.11098204

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T. et al. (2020). *Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages* // Findings of the Association for Computational Linguistics: EMNLP 2020, Pp. 2144–2160, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195

Nie, S., Fromm, M., Welch, C., Görge, R., Karimi, A. et al. (2024). *Do Multilingual Large Language Models Mitigate Stereotype Bias*? // Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, 2024. Pp. 65–83. Association for Computational Linguistics [Online], Available at: https://aclanthology.org/2024.c3nlp-1.6.pdf (Accessed 15.03.2026).

NLLB Team. Scaling neural machine translation to 200 languages. (2024). Nature. Vol. 630. Pp. 841–846, https://doi.org/10.1038/s41586-024-07335-x

Papadimitriou, I., Lopez, K. and Jurafsky, D. (2023). Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models // *Findings of the Association for Computational Linguistics: EACL 2023*. Pp. 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.89

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

Pawar, S., Park, J., Jin, J., Arora, A., Myung, J. et al. (2025). Survey of Cultural Awareness in Language Models: Text and Beyond // Computational Linguistics. Vol. 51 (3). Pp. 907–1004. https://doi.org/10.1162/COLI.a.14

Rashid, Q., Liemt, E., Shih, T., Ebinama, A., Ramos, K. et al. (2025). *Amplify Initiative: Building A Localized Data Platform for Globalized AI,* ArXiv, abs/2504.14105 https://doi.org/10.48550/arXiv.2504.14105 .

Rei, R., Craig, S., Farinha, A.C. and Lavie, A. (2020). *COMET: A Neural Framework for MT Evaluation* // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Pp. 2685–2702. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Sun, K. and Wang, R. (2024). Textual Similarity as a Key Metric in Machine Translation Quality Estimation // ArXiv, arXiv:2406.07440v1 [cs.CL], https://doi.org/10.48550/arXiv.2406.07440

Tonja, A.L., Azime, I.A., Belay, T. D., Yigezu, M. G., Moges, A. Ah M. et al. (2024). EthioLLM: Multilingual Large Language Models for Ethiopian Languages with Task Evaluation // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Pp. 6341–6352, Torino, Italia. ELRA and ICCL [Online], Available at: https://aclanthology.org/2024.lrec-main.561/ (Accessed 15.03.2026).

Wang, Y. and Lin, Z. (2025). Revisiting Round-Trip Translation with LLMs and Agentic Translation // *2025 IEEE 5th International Conference on Software Engineering and Artificial Intelligence (SEAI)*, Pp. 172-178. https://doi.org/10.1109/SEAI65851.2025.11108752

Xu, Y., Hu, L., Zhao, J., Qiu, Z., Ye, Y. and Gu, H. (2025). A Survey on Multilingual Large Language Models: Corpora, Alignment, And Bias // Frontiers of Computer Science. Vol. 19, 1911362. https://doi.org/10.1007/s11704-024-40579-4

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y. (2019). *BERTScore: Evaluating Text Generation with BERT* // ArXiv, abs/1904.09675. https://doi.org/10.48550/arXiv.1904.09675

The authors have read and approved the final manuscript.

Авторы прочитали и одобрили окончательный вариант рукописи.

Conflicts of interests: the authors have no conflicts of interest to declare.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Татьяна Александровна Литвинова, доктор филологических наук, профессор, Воронежский государственный педагогический университет, кафедра русского языка, современной русской и зарубежной литературы, Воронеж, Россия

Tatiana A. Litvinova, Doctor of Philological Sciences (D.Sc. in Philology), professor, Voronezh State Pedagogical University, Department of Russian Language, Modern Russian and Foreign Literature, Voronezh, Russia

Галина Анатольевна Заварзина, доктор филологических наук, заведующий кафедрой, Воронежский государственный педагогический университет, кафедра русского языка, современной русской и зарубежной литературы, Воронеж, Россия

Galina A. Zavarzina, Doctor of Philological Sciences (D.Sc. in Philology), Head of Department, Voronezh State Pedagogical University, Department of Russian Language, Modern Russian and Foreign Literature, Voronezh, Russia