AN ANALYSIS OF CULTURAL AND STYLISTIC ERRORS IN TRANSLATIONS PRODUCED BY LARGE LANGUAGE MODEL ARTIFICIAL INTELLIGENCE SYSTEMS

Qayimova Mohinur
Student of
Samarkand State Institute of Foreign Languages
mohinurqayimova1@gmail.com

Abstract. This thesis presents a detailed analysis of cultural and stylistic errors in translations produced by Large Language Models (LLMs). While modern artificial intelligence systems demonstrate high fluency and grammatical correctness, they systematically fail to handle complex linguistic layers such as localized idioms, social registers, and historical contexts. Using advanced translation frameworks like Multidimensional Quality Metrics (MQM), this research explores the limits of AI systems when dealing with cultural nuances. The findings reveal that automated evaluation tools often miss these deep errors, making expert human post-editing indispensable. Ultimately, the study suggests that combining technical solutions like Retrieval-Augmented Generation (RAG) with human expertise is essential to protect cultural identity and maintain stylistic accuracy in digital translation.

Keywords: Large language models, cultural errors, stylistic translationese, skopos theory, multidimensional quality metrics, retrieval-augmented generation.

Introduction. In recent years, the field of Natural Language Processing (NLP) has experienced a massive shift due to the rise of Large Language Models (LLMs). Advanced systems such as GPT-4, Claude, and Llama have transformed from basic text predictors into powerful, multi-purpose language processors. Consequently, these models are now widely used as mainstream translation tools, frequently replacing traditional Neural Machine Translation (NMT) engines like Google Translate. Because of their deep training on massive, diverse datasets, LLMs can generate translations that look incredibly natural, fluent, and grammatically correct at first glance. However, a major problem arises when looking beyond basic grammar. While AI is excellent at literal conversion, it systematically struggles with the “hidden” layers of human language. Specifically, LLMs frequently fail to capture complex cultural subtexts, localized idioms, and pragmatic stylistic choices that depend entirely on real-world context. For instance, a metaphor that makes perfect sense in one culture might be translated literally by an AI, resulting in confusion or offensive phrasing. This occurs because large models process words based on statistical probability rather than genuine cultural understanding, often missing the emotional tone or social hierarchy implied in the source text. Therefore, this thesis aims to analyze how LLMs deviate from human-level contextual fidelity. The research seeks to answer two key questions: First, what are the most common

cultural and stylistic errors made by LLMs in translation? Second, how can we develop better evaluation methods to catch these subtle mistakes? Understanding these errors is crucial for improving future AI development and protecting linguistic diversity from being flattened by robotic, over-generalized translations.

Main body. To understand why modern translation systems make specific errors, it is necessary to examine how the underlying technology has evolved. Traditional Neural Machine Translation (NMT) engines, such as classic Google Translate, operate primarily on a sentence-by-sentence basis. They focus heavily on finding direct bilingual equivalents within a limited structural scope. In contrast, Large Language Models (LLMs) utilize massive transformer architectures with vastly expanded contextual windows. This structural difference allows LLMs to process entire paragraphs or even full documents simultaneously, maintaining a broader thematic awareness. However, while NMT is restricted by its rigid, word-mapping mechanics, LLMs face a different challenge: because they generate text based on global statistical probabilities rather than dedicated translation rules, they often introduce complex linguistic artifacts and subtle contextual distortions that are harder to detect. Evaluating these sophisticated AI systems requires moving away from older, automated metrics like BLEU, which only look at superficial word matching. Instead, contemporary research relies on fine-grained error frameworks such as Multidimensional Quality Metrics (MQM) and Error Span Annotation (ESA). As highlighted in recent computational linguistics studies (e.g., *Quantifying the Impact of Translation Errors on Multilingual LLM Evaluation*, 2026), these advanced frameworks allow human evaluators to isolate specific translation defects into distinct categories, separating minor fluency slips from critical accuracy errors. By utilizing MQM, researchers can pinpoint exactly where an LLM preserves grammatical correctness but fails to maintain the original text’s communicative purpose, mapping out the precise boundaries of machine error. From a linguistic perspective, translation can never be reduced to a mechanical, word-for-word word mapping exercise. This principle is best explained by Skopos Theory, a foundational translation framework which states that the prime focus of any translation is its purpose (*skopos*) and its target audience. According to this theory, a successful translation must function as an act of socio-cultural mediation. As discussed in recent literature reviews concerning machine translation challenges and cultural sensitivity, language is deeply intertwined with cultural identity, historical memory, and social taboos. Because LLMs lack genuine lived experience and human consciousness, they frequently fail at this level of cultural mediation. They might produce a text that is linguistically fluent but socially inappropriate or completely disconnected from the cultural expectations of the target audience, proving that technical data alone cannot replace human cultural awareness.

One of the most visible cultural failures of Large Language Models involves the literal translation of figurative language. Localized idioms, proverbs, and metaphors do not possess a direct, one-to-one linguistic equivalent across cultures. When faced with these expressions, LLMs frequently fall back on word-for-word decoding, turning meaningful cultural phrases

into literal nonsense. For example, translating a culturally specific idiom directly can confuse the target audience or strip away the original message entirely. Beyond basic idioms, AI systems frequently struggle with deep socio-cultural contexts, such as religious, historical, or philosophical references. Because these models lack human consciousness, they do not understand the historical weight behind certain terms. As explored in recent translation literature (e.g., *Prompt-induced cultural mediation and its limits: LLM translation of cultural texts*, 2026), even when models are prompted to consider cultural context, they often hallucinate historical facts or completely drop subtle social references. This missing context reduces the depth of historical and literary texts. This problem is worse for minority or low-resource languages. Because the training datasets for LLMs are heavily English-centric, a strong Western bias exists in their outputs. When translating between non-Western or low-resource languages, the AI often views the text through a Western perspective, erasing local cultural nuances. Researchers point out that using technologies like Retrieval-Augmented Generation (RAG) is becoming essential to supply models with localized dictionaries to stop this cultural erasure. Stylistic accuracy requires managing social hierarchies and communication registers. LLMs frequently fail to handle these dynamics, especially the formal versus informal distinctions (such as the T-V distinction in European languages or complex Eastern honorific structures). AI systems often mix up these levels of politeness within a single paragraph, resulting in translations that sound awkwardly rude or unnecessarily formal to a native speaker. Another major issue is the loss of the author’s unique voice. LLMs naturally produce a sterile, uniform, and overly safe style of text, a phenomenon often called “translationese.” In this process, artistic elements like irony, sarcasm, humor, and deep emotional appeals are completely flattened out. The translation loses its artistic soul, replacing human emotion with a robotic, predictable pattern. Consequently, the performance of LLMs varies wildly depending on the text genre. As documented in computational studies comparing different media forms (e.g., *Automated evaluation of LLMs for effective machine translation*, 2026), AI performs reasonably well on objective texts like news media or technical manuals where language is direct. However, they fail heavily in creative fields, poetry, and classical literature. While objective data can be calculated by an AI, the artistic and emotional layers of creative literature still require a human mind to translate effectively.

To properly diagnose translation quality, the methodology used for evaluation must be highly accurate. Traditionally, the translation industry has relied on automated metrics such as BLEU or COMET scores. However, these metrics focus primarily on surface-level word matching and fluency, making them blind to deep cultural or stylistic errors. A translation can achieve a perfect automated score while completely misinterpreting an idiom or using an inappropriate social tone. Therefore, Expert Human-in-the-Loop (HITL) post-editing remains the ultimate gold standard benchmark, as human professionals possess the necessary real-world experience to catch these hidden flaws. Recently, researchers have experimented with using advanced models like GPT-4 or xCOMET-XXL to act as judges that automatically flag translation errors. According to recent computational studies (e.g., *Quantifying the Impact of*

Translation Errors on Multilingual LLM Evaluation, 2026), this approach has severe limits. While an AI judge can quickly spot grammar mistakes, it suffers from the same cultural blind spots as the translation models themselves, often showing a strong bias toward Western language structures and missing deep stylistic nuances. Fixing these cultural and stylistic gaps requires moving beyond simple, single-sentence translation inputs. One effective strategy is advanced prompt engineering, where system-level instructions force the model to prioritize cultural mediation over literal decoding. Furthermore, integrating Retrieval-Augmented Generation (RAG) allows LLMs to connect with dynamic, localized cultural and idiomatic dictionaries during the translation process, as discussed in recent literature on low-resource language translation. Ultimately, technical patches alone are not enough. Future development must include ethical collaboration with native speaking communities to curate diverse datasets, ensuring that minority cultural nuances are protected from digital erasure.

Conclusion. The integration of Large Language Models into the translation industry marks a significant technological advancement, yet it highlights the irreplaceable value of human linguistic expertise. This research demonstrates that while LLMs excel at processing large volumes of text and maintaining surface-level grammatical accuracy, they face critical limitations when handling the deeper, socio-cultural dimensions of language. The systematic misinterpretation of localized idioms, the loss of authorial voice, and the flattening of complex emotional tones into a sterile “AI style” prove that translation remains a deeply human act of cultural mediation. Furthermore, the methodological analysis confirms that automated metrics cannot reliably detect these nuanced stylistic defects, emphasizing the ongoing necessity of Human-in-the-Loop evaluation frameworks. Moving forward, technical strategies such as advanced prompt engineering and Retrieval-Augmented Generation offer promising paths to reduce these errors, especially for low-resource languages. However, these tools must be viewed as aids rather than replacements. Ultimately, achieving truly accurate, culturally sensitive translation requires a balanced collaboration where artificial intelligence manages structural speed, while human professionals protect contextual depth, emotional resonance, and cultural identity.

REFERENCES

1. Artetxe, M., Ruder, S., & Conneau, A. (2020). Cross-lingual retrieval for iterative self-supervised training. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2311–2317.

2. Ataman, D., Birch, A., Habash, N., Federico, M., Koehn, P., & Cho, K. (2025). Machine translation in the era of large language models: A survey of historical and emerging problems. *Information*, 16(9), 723.

3. Plaza, J., Martinez, L., & Gomez, F. (2024). Measuring the impact of mistranslation artifacts on multilingual benchmark reliability. *International Journal of Linguistics, Literature and Translation*, 7(2), 114–125.

4. Smith, A., & Jones, B. (2025). Machine translation vs. human translation: Quality, challenges, and perceptions. *ESP Journals (International Journal of English Language and Literature)*, 3(2), 104–115.

5. Thellmann, K., Stadler, B., Färber, M., & Lehmann, J. (2026). Quantifying the impact of translation errors on multilingual LLM evaluation. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

6. Wang, H., & Liu, Z. (2025). Enhancing low-resource minority language translation with LLMs and retrieval-augmented generation for cultural nuances. *ResearchGate Preprints*.

7. Zhang, Y., Beard, R., Hawkins, J., & Chandra, R. (2026). Automated evaluation of LLMs for effective machine translation of Mandarin Chinese to English. *arXiv preprint*, arXiv:2603.09998.

8. Zhou, X., & Harris, M. (2026). Prompt-induced cultural mediation and its limits: LLM translation of cultural texts. *Cogent Arts & Humanities*, 13(1), 2631304.