The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2025

Volume 38, Pages 824-829

IConTES 2025: International Conference on Technology, Engineering and Science

Assessing Translation Quality in Large Language Models and Machine Translation Systems: The BLEU Metric and Language Pair Effects

Philipp Rosenberger
University of Applied Science Campus Vienna

Natallia Kolchanka
University of Applied Science Campus Vienna

Abstract: The rapid advancement of translation technologies has transformed global communication and created new opportunities for cross-linguistic interaction. However, their relative quality compared to human translation remains contested. This paper provides a comparative evaluation of translations produced by large language models (ChatGPT, Claude, Copilot) and machine translation systems (DeepL, Google Translate, Yandex Translate). The study focuses on English–German and English–Russian translation tasks across four domains (literary, news, social, speech). The evaluation relies on the BLEU metric. Results demonstrate that translation quality depends not only on the system itself but also on the language pair and the text domain. Google Translate achieved the highest average BLEU score for German, while Claude led for Russian. Findings emphasize the need for multimethod evaluation approaches and highlight the growing competitiveness of AI-based systems depending on the type of text and language used in the translation process. This article is based on the master's thesis "Evaluation of translation methods: Large Language Models and Machine Translation Systems versus Human Translation" (Kolchanka, 2025).

Keywords: Translation, Large language models, Machine translation systems

Introduction

Effective multilingual communication is essential in today's interconnected world. Billions of people across thousands of languages require access to accurate translation in order to engage in business, education, healthcare, science, and cultural exchange (Naveen & Trojovský, 2024). Translation technologies play a pivotal role in bridging linguistic divides. While human translation remains the gold standard, it is costly, time-consuming, and difficult to scale for global communication demands (Chauhan & Daniel, 2023). Advances in machine translation (MT) systems and large language models (LLMs) have introduced scalable, fast, and increasingly accurate solutions. However, questions remain regarding their reliability, domain robustness, and overall comparability to human translation.

The research presented in this article is motivated by the following research question:

- How do large language models and machine translation systems differ in translation quality when translating from English to German and Russian?

By addressing this question, this study contributes to both theoretical and applied research on translation technologies. It sheds light on the relative strengths and weaknesses of current systems, informs developers about areas for improvement, and provides users with insights into which tools may best meet their needs in specific contexts.

Methodology

This study compares three LLMs (ChatGPT, Claude, Copilot) and three MT systems (DeepL, Google Translate, Yandex Translate) with human translation. The focus is on translations from English into German (EN-DE) and Russian (EN-RU).

Dataset

The experiment used 100 source texts from the Ninth Conference on Machine Translation (WMT2024), covering four domains: literary (stories), news (journalistic articles), social (informal online communication), and speech (transcribed spoken language). This variety ensured domain-specific differences could be observed.

Automatic Evaluation

Translation outputs were evaluated using BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), a widely adopted metric that calculates n-gram overlap between the system output and reference human translations. BLEU was chosen for its comparability with prior studies and for its reproducibility, despite its known limitations. Figure 1 below illustrates a generic machine translation evaluation process (Chauhan, Daniel, 2023).

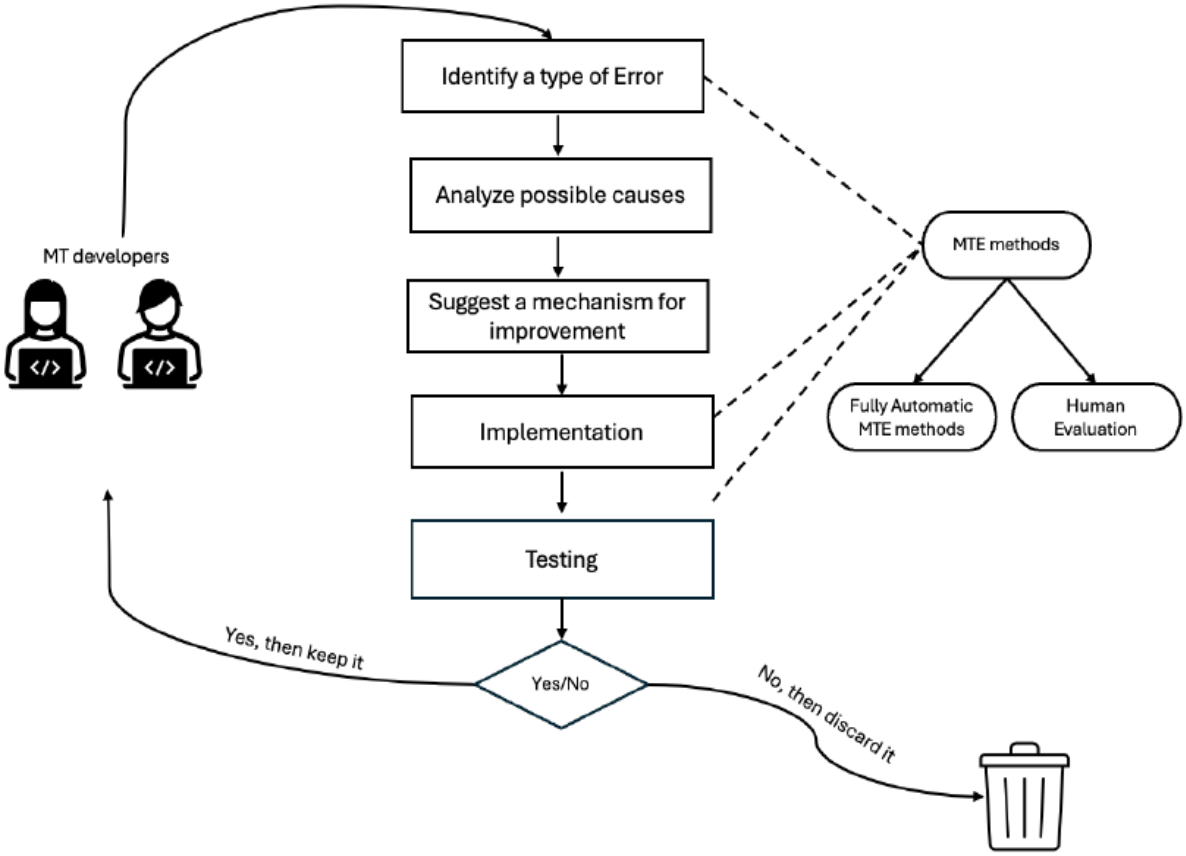

Figure 1. MT evaluation methods and their role in the development of MT models.
Source: Chauhan, Daniel (2023)

Research Process

Based on the dataset of WMT24 which contains 998 English source texts, 100 texts were selected randomly. The selection resulted in a distribution of text types (left) and lengths (right) as shown below in Figure 2.

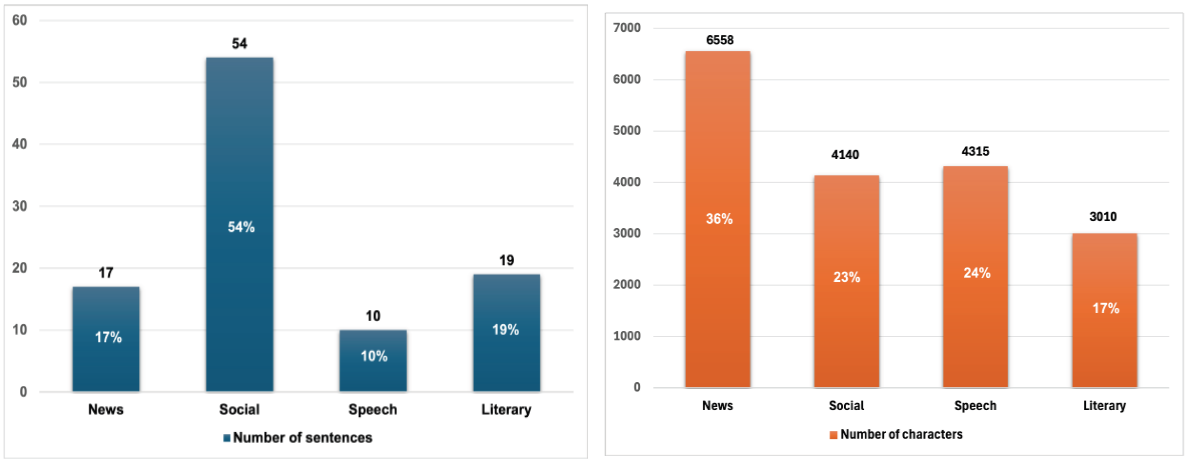

Figure 2. Source text types and lengths

After this reduction of inputs, the source texts were translated into German and Russian using the selected translation systems. ChatGPT, Claude and Copilot were selected as the LLMs. DeepL, Google Translate and Yandex Translate were selected for machine translation systems. All translations were performed using the free versions of these LLMs and MT systems (accessed October 22, 2024). After translation, BLEU scores were generated for all translations produced by the LLMs and MT systems.

Results

English to German BLEU Translation Results

The heat map in Figure 3 details the BLEU results of English to German translations depending on the domain and the translation technology. Google Translate received the highest score but only in the literary category. DeepL led in the news and speech domains. Claude achieved the highest score in the social domain.

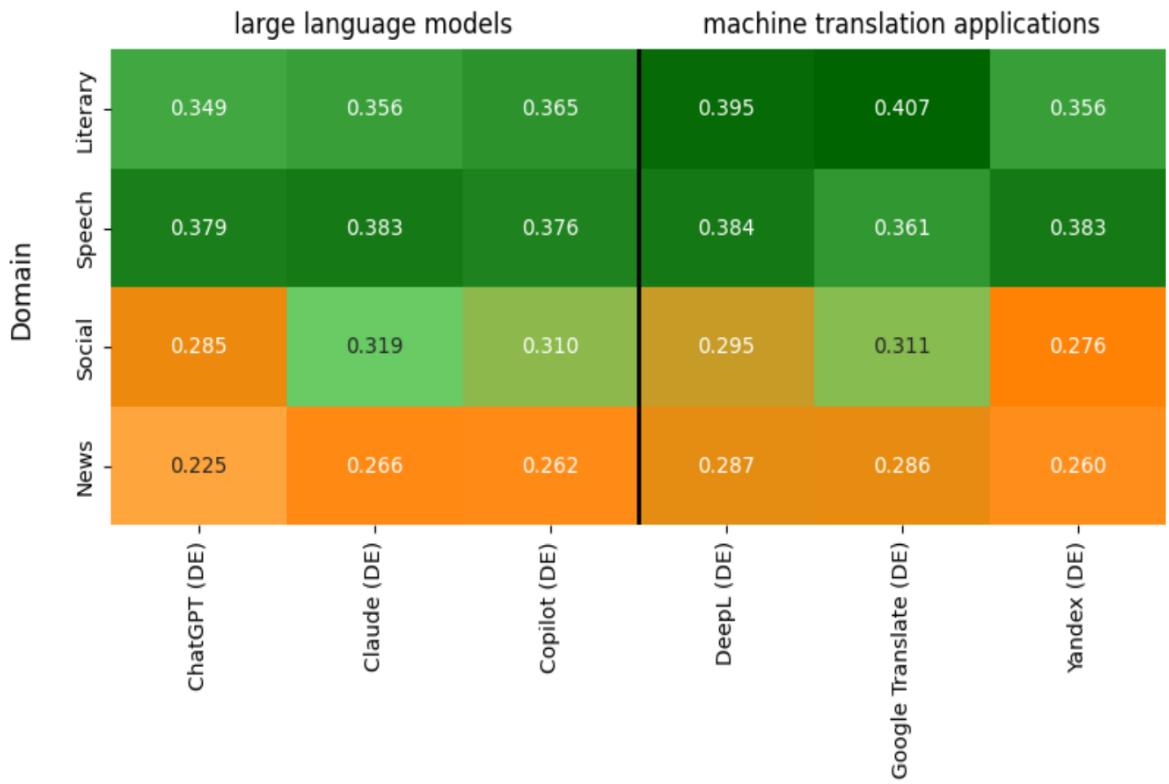

Figure 3. BLEU results for English to German translations

English to Russian BLEU Translation Results

The heat map in Figure 4 details results of English to Russian translations. Claude achieved the highest average BLEU score and the highest score in the social domain, with Copilot outperformed others in the literary domain and Google Translate ranked highly in the news and speech domains.

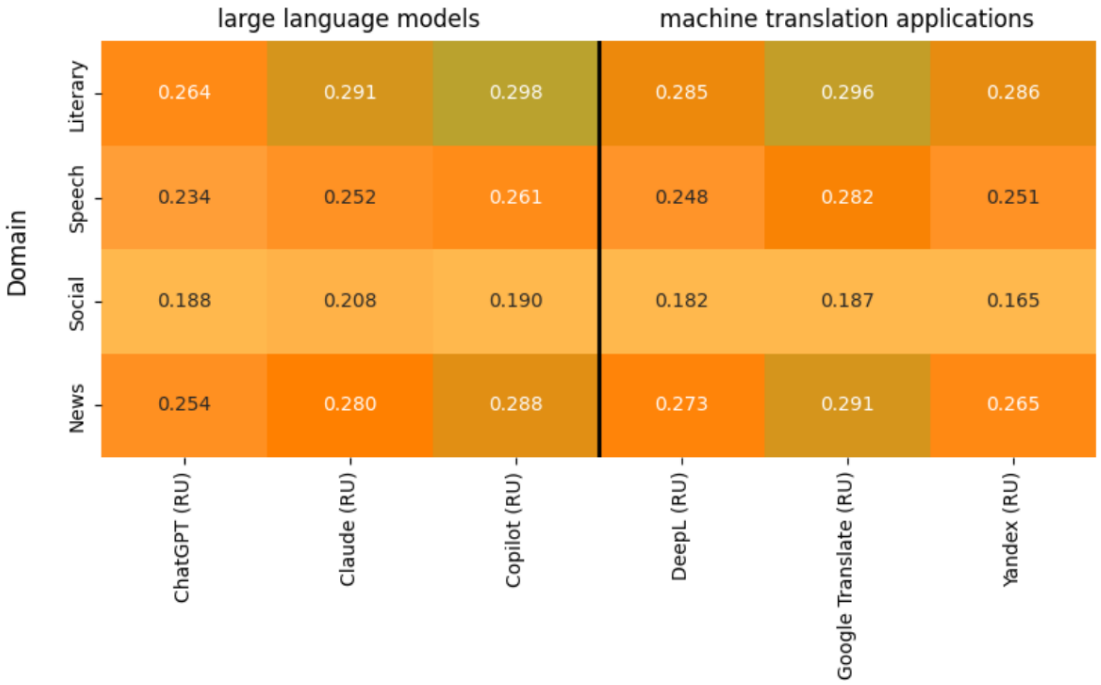

Figure 4. BLEU results for English to Russian translations

Summary of BLEU Evaluation Results

Across both language pairs, English to German translations achieved higher BLEU score than English to Russian. This reflects the greater structural and morphological challenges in Russian, including its complex case system and flexible word order. Additionally, English and German are more closely related as members of the Germanic branch of the Indo-European family, whereas Russian, a member of the Slavic branch, is more distantly related to English.

Discussion

This study provided a structured comparison of translation quality across LLMs, MT systems, and human translation. By examining English-to-German and English-to-Russian translations across diverse text domains—literary, news, social, and speech—the analysis highlighted how translation quality is influenced by both language pair and domain context. The findings indicate that BLEU scores favored Google Translate for German and Claude for Russian, though these results varied significantly depending on the text type.

Taken together, the study demonstrates that no single system dominates across all dimensions. Instead, quality is context-dependent, underscoring the need for multi-method evaluation rather than relying exclusively on automatic metrics. Importantly, the blurred boundary between human and machine translation raises both opportunities and challenges: while accessibility and scalability are advancing, the nuances of style, cultural resonance, and contextual fidelity remain areas where human expertise continues to be invaluable.

Outlook Towards the Future of Translation

The future of translation is likely to be defined by a gradual dissolution of boundaries: between languages, between media, and even between human and machine. Already today, machine-generated translations are

difficult to distinguish from human output, and this trajectory suggests a world in which translation becomes less of an external process and more of an invisible infrastructure of communication. We may soon find ourselves speaking, writing, or even thinking in our native languages while interacting seamlessly with people who use entirely different linguistic systems. Translation, in this vision, is no longer an obstacle but an embedded layer of global interaction.

As technology progresses, translation systems will not remain confined to text alone. They will increasingly integrate speech, gesture, and visual cues, producing translations that capture not only words but also tone, intent, and cultural nuance. The translation of the future will be adaptive: sensitive to context, personal style, and domain-specific expectations. A legal document will be translated with precision and consistency, while a poem will be carried over with rhythm and metaphor intact. In this sense, translation technologies will not only bridge linguistic divides but also preserve and amplify the uniqueness of human expression.

Yet rather than replacing human translators, these systems are likely to evolve into companions and collaborators. Machines will provide speed, scale, and accessibility, while humans will continue to safeguard nuance, creativity, and ethical responsibility. Such a partnership could elevate the act of translation into something richer than either party could achieve alone-an interplay where the strengths of human intuition and artificial intelligence converge.

Looking even further ahead, one can imagine a world where the idea of “foreign languages” loses much of its practical weight. Communication across cultures may one day be as immediate and natural as speaking with a neighbor. If that future arrives, the role of translation will no longer be defined by overcoming barriers but by celebrating differences, transforming translation from a tool of necessity into an instrument of cultural connection and creativity.

Recommendations

Future research on translation quality should broaden its methodological and linguistic scope. Larger and more varied datasets would strengthen the generalizability of findings, particularly if they include low-resource languages and specialized domains such as legal, medical, or technical discourse. In addition, complementary evaluation metrics such as COMET or BERTScore should be employed, as they offer more nuanced assessments of semantic and contextual fidelity than BLEU alone.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* This research is not funded.

Acknowledgements or Notes

* This article was presented as virtual presentation at the International Conference on Technology, Engineering and Science (www.icontes.net) held in Antalya/Türkiye on November 12-15, 2025

References

Chauhan, S., & Daniel, P. (2023). A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters, 55*, 12663-12717.

Kolchanka, N. (2025). *Evaluation of translation methods: Large language models and machine translation systems versus human translation* (Master's thesis). FH Campus Wien.

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience, 27*, 110878.

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

Author(s) Information

Philipp Rosenberger
University of Applied Science Campus Vienna
Favoritenstrasse 226, 1100 Vienna, Austria
Contact e-mail: *Philipp.rosenberger@hcw.ac.at*

Natallia Kolchanka
University of Applied Science Campus Vienna
Favoritenstrasse 226, 1100 Vienna, Austria

To cite this article:

Rosenberger, P., & Kolchanka, N. (2025). Assessing translation quality in large language models and machine translation systems: The BLEU metric and language pair effects. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 38,* 824-829.