Can Large Language Models substitute Human Annotators in Machine Translation Usability Assessment?

Matúš Kleštinec, Martin Chotár, and Daša Munková

Constantine the Philosopher University Nitra
Trieda Andreja Hlinku 1
949 01 Nitra, Slovakia
{matus.klestinec,martin.chotar,dmunkova}@ukf.sk

Abstract. This study investigates the potential of GPT-based large language models (LLMs) to evaluate the usability of English–Slovak machine translations. Usability annotations were conducted by three GPT models (GPT4o, GPT-4.1, and GPT-5) with varying temperature settings, alongside seven human annotators. The earlier GPT models (GPT-4o and GPT-4.1) produced significantly higher usability scores than human evaluators, suggesting a tendency to overestimate translation quality. In contrast, the latest model, GPT-5, exhibited evaluation patterns that were statistically comparable to those of human annotators. These findings indicate that the GPT-5 model demonstrates improved agreement with human judgment and may serve as a reliable automated annotator for assessing the usability of machine translation systems.

Keywords: LLM; ChatGPT; Machine Translation; Slovak.

1 Introduction

Evaluating the usability of machine translation (MT) outputs remains a crucial step in assessing the overall quality and practical applicability of translation systems. Traditional human based evaluation, while reliable, is time consuming, costly, and often subject to variability among annotators. Recent advances in large language models (LLMs) [10], such as the GPT models, have opened new possibilities for automating evaluation tasks traditionally performed by humans. Previous studies [3] have shown that LLMs can assess translation quality with promising consistency. However, little attention has been given to usability - a broader and more subjective criterion reflecting user satisfaction and perceived translation quality. This study aims to examine whether GPT based models can serve as reliable annotators for evaluating the usability of English-Slovak machine translation. By comparing annotations from multiple GPT versions and human evaluators, we analyze the degree of agreement between human and LLM judgments and discuss the potential of modern LLMs to complement or

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 37–43, 2025.

partially replace human annotators in translation usability assessment. According to ISO 9241-11 standard [1], usability measures the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a defined context of use. We address the following research question: To what extent can large language models GPT-4o, GPT-4.1, and GPT-5, with varying temperature settings, evaluate the usability of English-Slovak MT outputs comparably to human evaluators? The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the methodology and dataset used in the study. Section 4 presents the results and their analysis. Section 5 concludes the paper and outlines directions for future research.

2 Related Work

Jiang at al. [2] examined translation evaluation using ChatGPT versions 3.5 and 4o in the Chinese-Portuguese language pair based on two criteria: information completeness and correctness. They found that the older GPT-3.5 model exhibited relatively low similarity in human MT evaluation, whereas GPT-4 showed higher similarity. However, these results varied depending on the domain of translation. High correlations between human annotators, GPT-3.5, and GPT-4 were observed for political texts, but significant discrepancies appeared in the evaluation of the literary text, where GPT-3.5 diverged substantially from human evaluation.

Alghamdi et al. [3] examined the extent to which GPT-4o can reliably assess the quality of legal translations from English into Arabic compared to human evaluators. They analyzed ten students' translations from a university legal translation course, which were first assessed by a teacher using a predefined rubric and then evaluated by GPT-4o using the same criteria. The system was prompted to detect the translation errors, score the exam, followed by manual segment to segment comparison between GPT-4o and human evaluations. Statistical analysis revealed a high level of agreement between human and GPT-4o with no statistically significant differences. GPT-4o provided detailed feedback and correction suggestions, indicating its potential as a complementary tool for translation assessment in educational contexts.

Wang et al. [4] investigated how well automatic metrics that do not rely on reference translations correspond with human judgments in simultaneous English-Spanish translation. They analyzed 12 lecture recordings translated by three professional interpreters and the KUDO AI Speech Translator. Human evaluation focused on accuracy and intelligibility, while LLM evaluation which was performed by GPT-3.5, GPT-Ada, UMSE and All-MiniLM-L6-v2 models through semantic similarity leveraging sentence embeddings. Results showed that GPT-3.5 achieved the highest correlation with human scores, though overall agreement among human annotators was low.

Bavaresco et al. [5] compared the results of 11 LLMs with human judgments and found that agreement strongly depends on the model, dataset, and task

type. LLMs generally correlated better with human assessments on human-written texts but showed inconsistencies when evaluating machine-generated content. While some models achieved near human parity in instruction following or reasoning tasks, none consistently matched human evaluations across all contexts. They concluded that LLMs can assist but not fully replace human judges without prior calibration and task specific validation.

3 Methodology

The dataset comprised 100 segments (sentences, chapter or article titles – each treated as a segment when present in the analyzed texts). For each text, one MT output was obtained from online Google Translate service and aligned with the original English source text and its human Slovak translation. These 100 segments were manually annotated by seven human annotators (H01, H02, H03, H04, H05, H06, and H07) based on the usability criterion.

Table 1: Selection of 3 rows from dataset

Source	Translation	usability-c
I grew up in between.	Vyrastal som medzi tým.	poor (most meaning)
In 1995, he was awarded the Nobel Prize in Literature.	V roku 1995 mu bola udelená Nobelova cena za literatúru.	good (likely use it)
There was a horse in the stable.	V stajni bol kôň.	good (likely use it)

Usability was measured on a five-point scale, where 1 indicated an unacceptable (unusable) translation and 5 indicated an excellent translation. A higher value represents greater usability. These values corresponded to the following verbal categories:

1. unacceptable (non useful),
2. poor (most segments),
3. acceptable (some segments),
4. good (likely use it),
5. excellent (no editing).

To use ChatGPT to annotate the usability of translations, we imported the OpenAI class from the Python library openai. That class is necessary to sending requests to OpenAI API to annotate all sentences. We requested ChatGPT to perform the same evaluation based on a prompt we defined:
You are a professional evaluator of Slovak translations. Evaluate the following translation from English to Slovak using: Usability - how suitable the translation would be

for real use or usability measures the extent to which a translation can be used effectively, efficiently, and with satisfaction in a specific context of use (ISO, 1998). Choose one of: [unacceptable (non useful), poor (most segments), acceptable (some segments), good (likely use it), excellent (no editing)] Then output your evaluation in this format: Usability: <category> .

We applied this prompt to three models: GPT-4.1 [6], GPT-4o [7], and GPT-5 [8]. We also experimented with the Temperature parameter, which controls the randomness of the model's output. A lower value makes the model more consistent, producing consistent and predictable responses and higher value increases randomness, making the output more creative and varied [9]. For the GPT-4.1 and GPT-4o models, we used temperature settings of 0.2 and 1. For the GPT-5 model, we only used *temperature* = 1 because it was not possible to change value. In total, we performed annotation using GPT models 5 times. Annotation was performed using columns Source and Translation from Table 1 .

4 Results

The first step of the experiment involved determining the inter-annotation agreement (IAA) among evaluators. Based on the non-parametric Kendall's coefficient of concordance (used to measure agreement among multiple annotators) and the non-parametric Kendall's Tau coefficient (used for pairwise evaluator comparisons), the results indicated a low to moderate degree of agreement (*Kendall Coeff. of Concordance* = 0.2272) among evaluators when evaluating translation usability. Although the evaluators demonstrated some consistency in their usability judgments, notable deviation was observed for evaluator H01, who adopted a more stringent approach and, on average, rated translations as poorly usable (*Average Rank* = 2.27). This evaluation differed significantly from those of the other annotators [11].

The next stage examined whether the variability of the human annotations and the LLM-generated annotations was comparable. Results from Mauchly's test of sphericity ($W = 0.002363$, *Chi-Sqr.* = 575.3824, $df = 65$, $p = 0.0001$) indicated a violation of sphericity assumption ($p < 0.001$). Consequently, adjusted tests for repeated measures were applied.

The adjusted univariate test for repeated measures (*G-G Epsilon* = 0.357284, *G-G Adj. df1* = 3.930121, *G-G Adj. df2* = 389.0819, *G-G Adj. p* = 0.0001) rejected the null hypothesis that usability scores are independent of whether they are produced by human evaluators or LLMs ($p < 0.001$). This finding indicates statistically significant differences between usability evalution produced by human evaluators and those generated by LLMs. To identify specific differences between evaluators, the Tukey HSD test for multiple comparisons was applied Table 2 .

Human or LLM annotations with no statistically significant differences in mean usability ratings were grouped into the same homogenous cluster (one homogeneous group). The GPT-4o and GPT-4.1, across varying temperature settings, were classified into a single homogeneous group, suggesting they produced very similar evaluations regardless of temperature Figure 1. Moreover,

Table 2: Homogenous groups - usability

α = .05000., Error: Within MS = 67131, df = 1089.0							
Cell No.	**U**	**Mean**	**1**	**2**	**3**	**4**	**5**
6	usability_H01	2.72					****
7	usability_H02	3.28			****		
8	usability_H03	3.51	****		****		
12	usability_H07	3.54	****	****	****		
9	usability_H04	3.56	****	****	****		
5	usability_T1GPT5	3.75	****	****			
11	usability_H06	3.83	****	****			
10	usability_H05	3.89		****			
4	usability_T1GPT4o	4.35				****	
1	usability_T0.2GPT4.1	4.39				****	
2	usability_T0.2GPT4o	4.43				****	
3	usability_T1GPT4.1	4.47				****	

both models evaluated usability substantially higher than human annotators, implying that earlier GPT models may overestimate the usability of machine translation from English into Slovak. In contrast, the state of the art GPT-5 model produced usability scores that were, on average, lower than those of previous models but closer to most human annotators Figure 1. GPT-5 model appeared in two homogeneous groups (Group 1 and 2) alongside several human annotators. Specifically, it forms a homogenous group with human annotators H03, H07, H04, and H06 in one cluster and another with human annotators H07, H04, H06, and H07 in the second cluster. This confirms that the scores produced by T1GPT5 are comparable to human judgments. Conversely, evaluator H01 differed significantly not only from all LLMs but also from other human evaluators. As a result, H01 constituted a single-element homogeneous group Table 2.

5 Conclusion

The results indicate that while earlier GPT models (GPT-4o and GPT-4.1) tend to overestimate the usability of English–Slovak translations, the GPT-5 model shows a closer alignment with human annotators. Its inclusion in the same homogeneous groups as several human evaluators suggests improved calibration and a more human like assessment of translation usability. This study demonstrates that the latest generation of GPT models may serve as a more reliable substitute or complement to human annotators in usability evaluation tasks. The analysis also reveals model specific tendencies that can inform the development of automated evaluation process. By integrating LLM based annotation into usability assessment, the process can be significantly faster and more available while maintaining reliability close to human judgment. Future research could explore additional factors influencing LLM evaluations, such as adequacy or fluency, and meaning preservation to assess whether LLMs perform consistently

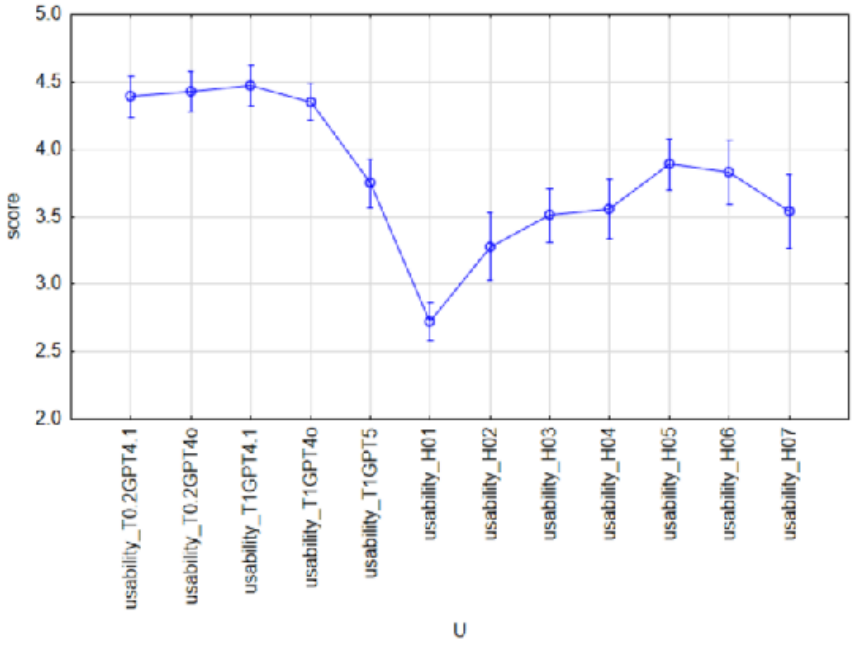

Fig. 1: Visualization of the mean usability scores produced by GPT models and human annotators

across different quality metrics. Moreover, comparative analyses between GPT-5 and other state-of-the-art models, such as Deepseek [12], Claude [13] or Gemini [14], could provide deeper insights into model-specific strengths and weaknesses. Finally, experimenting with hybrid human - AI evaluation frameworks may lead to more scalable and robust methodologies for translation quality assessment.

Acknowledgements. This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-23-0554.

Disclosure of Interests. The author have no competing interests to declare that are relevant to the content of this article.

References

1. International Standard ISO 9241-11, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11 : Guidance on usability `http://ndl.ethernet.edu.et/bitstream/123456789/87925/29/ISO9241part11.pdf`, last accessed 2025/10/20
2. Jiang, L. Jiang, Y. Han, L.: The potential of ChatGPT in translation evaluation: A case study of the Chinese-Portuguese machine translation. In: Cadernos de Tradução, 44(1) (2024) `https://doi.org/10.5007/2175-7968.2024.e98613`

3. Alghamdi, F. Alotaibi, H.: Using AI in Translation Quality Assessment: A Case Study of ChatGPT and Legal Translation Texts. In: Electronics 14(19):3893 (2025) https://doi.org/10.3390/electronics14193893
4. Wang, X. Fantinuoli, C.: Exploring the Correlation between Human and Machine Evaluation of Simultaneous Speech Translation. (2024) https://doi.org/10.48550/arXiv.2406.10091
5. Bavaresco, A. Bernardi, R. Bertolazzi, L. Elliott, D. Fernández, R. Gatt, A. Ghaleb, E. Giulianelli, M. Hanna, M. Koller, A. Martins, A. Mondorf, P. Neplenbroek, V. Pezzelle, S. Plank, B. Schlangen, D. Suglia, A. Surikuchi, A. K. Takmaz, E. Testoni, A.: LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. (2025) https://doi.org/10.48550/arXiv.2406.18403
6. OpenAI FAQ, https://openai.com/index/gpt-4-1/, last accessed 2025/10/20
7. Hurst, A. Lerer, A. Goucher, A. Perelman, A. Ramesh, A. et al.: GPT-4o System Card. (2024) https://doi.org/10.48550/arXiv.2410.21276
8. OpenAI, GPT-5 System Card, https://cdn.openai.com/gpt-5-system-card.pdf, last accessed 2025/10/20
9. OpenAI, Introducing GPT-4.1 in the API https://platform.openai.com/docs/faq/how-should-i-set-the-temperature-parameter, last accessed 2025/10/19
10. Vaswani, A. Shazeer, N. Parmar, N. Uszkoreit, J. Jones, L. Gomez, A. Kaiser, L. Polosukhin, I.: Attention Is All You Need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010. (2017) https://doi.org/10.48550/arXiv.1706.03762
11. Forgáč, F.: Quality estimation for machine translation. Dissertation thesis. Constantine the Philosopher University in Nitra. 141 p. https://opac.crzp.sk/?fn=detailBiblioFormChildK2O36&sid=46E5F073F744ACC4FE7F4E767BF7&seo=CRZP-detail-kniha, last accessed 2025/11/13
12. Guo, D. Yang, D. Zhang, H. Song, J. Zhang, R. et al.: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. (2025) https://doi.org/10.48550/arXiv.2501.12948
13. Claude API https://www.claude.com/platform/api, last accessed 2025/10/20
14. Comanici, G. Bieber, E. Schaekermann, M. Pasupat, I. Sachdeva, N. et al.: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. (2025) https://doi.org/10.48550/arXiv.2507.06261