CONSTANTINE
THE PHILOSOPHER FACULTY
UNIVERSITY OF NATURAL SCIENCES
IN NITRA AND INFORMATICS

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

RASLAN 2025 – Kouty nad Desnou, Czech Republic
The Nineteenth Workshop on Recent Advances in Slavonic Natural Language Processing

DEPARTMENT OF INFORMATICS
Faculty of Natural Sciences and Informatics CPU in Nitra

Martin Chotár, Matúš Kleštinec and Daša Munková

Introduction

- Evaluating MT usability is crucial for assessing translation quality and practical applicability. Traditional human evaluation, while reliable is:
  - Time-consuming
  - Costly
  - Subject to variability among annotators
- LLMs enable automation of evaluation tasks previously done by humans.
- Recent advances expand possibilities for MT assessment

Definition of Usability

- According to ISO 9241-11 standard, 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.

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?

Methodology

- Dataset comprised 100 segments
  - Sentences, chapter or article titles – each treated as a segment when present in the analyzed text
- 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)

Methodology - usability

- Usability was measured on a five-point scale:
  - 1. unacceptable (non useful),
  - 2. poor (most segments),
  - 3. acceptable (some segments),
  - 4. good (likely use it),
  - 5. excellent (no editing).

Methodology - Temperature

- Temperature parameter, which controls the randomness of the model's output.
- A lower value makes the model more consistent, producing consistent and predictable responses.
- Higher value increases randomness, making the output more creative and varied.

Methodology – LLM

- OpenAI API
- GPT models with temperatures:
  - GPT-4o (Temperature: 0.2 and 1),
  - GPT-4.1 (Temperature: 0.2 and 1),
  - GPT-5 (Temperature: 1)

  - In case of GPT-5 was not possible to change temperature.

Methodology - prompt

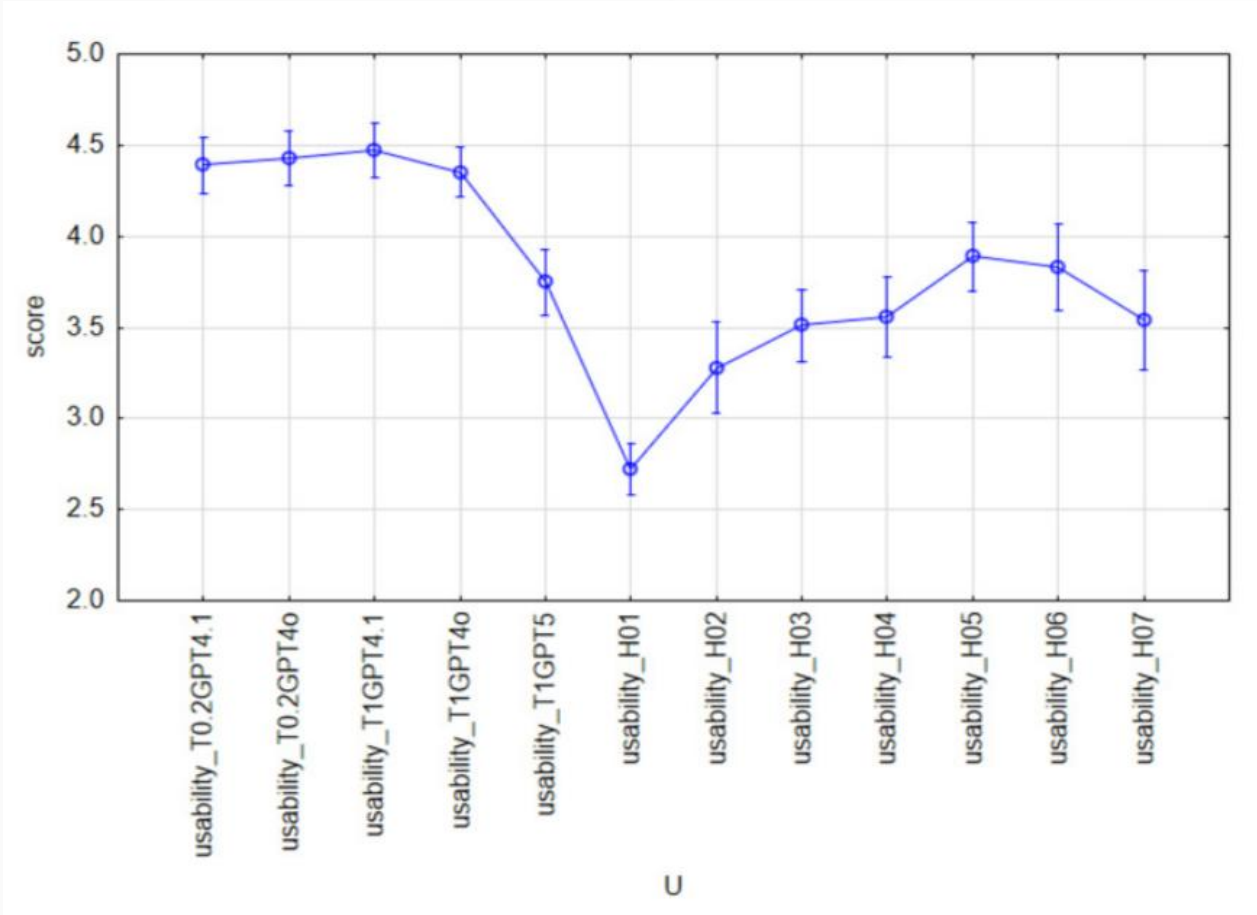
- *You are a professional evaluator of Slovak translations. Evaluate the following translation from English to Slovak using:*
- *Usability - how suitable the translation would be for real use or usability measures the extent to which a translation can be used effectively, efficiently, and with satisfaction in a specific context of use (ISO, 1998).*
- *Choose one of: [unacceptable (non useful), poor (most segments), 4 acceptable (some segments), good (likely use it), excellent (no editing)] Then output your evaluation in this format: Usability: <category> .*

Results – Homogenous groups - usability

- Tukey HSD test for multiple comparisons was applied
- Five homogenous groups
- GPT models 4o and 4.1 are in separate group
- GPT-5 is comparable with most of annotators

$\alpha = .05000.$, 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				****	

Results - mean usability scores produced by GPT models and human annotators



Conclusion

- Earlier GPT models (GPT-4o and GPT-4.1) overestimate the usability of English-Slovak translation
- GPT-5 model shows a closer alignment with human annotators.

- Integrating LLM based annotation in usability assessment = significally faster and more available, while maintaining reliability close to human judgment.

Future work

- Exploring additional factors influencing LLM evaluation: adequacy, fluency and meaning preservation to assess LLM capabilities across different quality metrics,
- Comparative analyses between GPT-5 and other state-of-the-art models, such as Deepseek, Claude, Gemini etc.

Thank you for your attention

- Martin Chotár: martin.chotar@ukf.sk
- Matúš Kleštinec: matus.klestinec@ukf.sk
- Daša Munková: dmunkova@ukf.sk