**Automated Evaluation of Language Model Performance – Machine Translation from English to Montenegrin Language**

Table of Content

[Introduction 2](#_Toc151073705)

[Methodology 3](#_Toc151073706)

[Language models selection 3](#_Toc151073707)

[Selection of metrics 5](#_Toc151073708)

[Thai language 7](#_Toc151073709)

[Reference corpora 7](#_Toc151073710)

[BLEU scores 9](#_Toc151073711)

[CER Scores 9](#_Toc151073712)

[WER Scores 10](#_Toc151073713)

[Discussion 10](#_Toc151073714)

[Conclusion 12](#_Toc151073715)

[Literatura 13](#_Toc151073716)

[Table 1: Methodology of BLEU score calculation 6](#_Toc151073717)

[Table 2: Methodology of CER and WER score calculation 7](#_Toc151073718)

[Table 3: BLEU scores for three language models on 4-gram and 3-gram models 9](#_Toc151073719)

[Table 4: CER Scores of three languageg models 9](#_Toc151073720)

[Table 5: WER Scores of three language models 10](#_Toc151073721)

# Introduction

Linguistic models represent a fundamental component of Natural Language Processing (NLP) technologies. NLP has been in the spotlight in the post-COVID-19 pandemic period, attracting attention from both experts and the general public. This is due to significant advancements in the performance of language models and the applications they enable. The reasons for this progress can be found in the increased availability of data, computational power, and innovations in the field of machine learning.

Modern language models are transforming the relationship between computers and language, innovating processes and procedures, pushing the boundaries of what to expect in terms of computer performance. Applications based on NLP or utilizing it in some segment of the process undeniably open up new possibilities, but at the same time, questions about the evaluation of language model performance arise.

The question of assessing the quality of a language model is not new, nor has the scientific community recently engaged in it. However, it is certain that the methodology that worked adequately with statistical or expert language models is undergoing scrutiny when working with modern language models created using deep learning methods.

It is undisputed that the evaluation of language model performance must be carried out through human assessment. However, over the years, numerous metric evaluation mechanisms have been developed to assist primarily in the context of the time required for a quality evaluation. The community's need for new performance metrics is primarily reflected in the complexity and multilayered nature of the problems that language models solve.

More precisely, it is not a novelty that different users of the same language model may be interested in completely different metrics, depending on the language model application for which it is used. For this reason, in practice, metrics specialized for specific tasks in the field of NLP can be found, as well as general metrics that deal with technical assumptions, regardless of the context and task of the language model. Such metrics, unburdened by context and the role of the language model in the application, will be used in this paper to evaluate the performance of modern language models in generating text in the Montenegrin language.

The paper focuses on the specificity of the Montenegrin and Thai language and the ability of language models to comprehend, process, or generate it. In conclusion, findings and recommendations are presented regarding the selection of language models for specific tasks and the automatic evaluation of performance.

# Methodology

## Language models selection

The selection of language models was carried out based on their availability and popularity in the NLP community. It is important to emphasize that each of the chosen language models is multilingual, trained on general-purpose data:GPT-3.5 Turbo[[1]](#footnote-1)

* Solar-0-70b[[2]](#footnote-2)
* LaMDA[[3]](#footnote-3)
* GPT 4[[4]](#footnote-4)
* LLAMA-2-70b[[5]](#footnote-5)

Models from the GPT series (GPT-3.5-turbo and GPT-4) represent the most popular models, primarily due to the application of ChatGPT, which became a global topic at the end of 2022 and the beginning of 2023, demonstrating the potential of this type of technology. The company OpenAI, which introduced these models to the world through a simple application, leveraged the original success of the GPT-3.5-turbo language model-driven application and quickly introduced GPT-4.

GPT-3.5-turbo was trained with data collected until September 2021. This model, along with GPT-4, is still in development. The basic differences between the two models are manifested in:

* Number of parameters - GPT-3.5-turbo has 100 billions, GPT-4 175 trillions of parameters[[6]](#footnote-6)
* the size of the context that can be analyzed in one go - GPT-3.5-turbo can contextualize 9 times more information during analysis
* multimodality - GPT-4 is capable of processing images and code, while GPT-3.5-turbo can only process textual data
* performance- - GPT-4 is signigicantly faster and more precise
* price – GPT-3.5-turbo is free, GPT-4 is not

The global popularity of the ChatGPT application not only drew public attention to generative language models but also inspired the integration of language models into existing services to enhance their quality. Consequently, GPT has been integrated into services such as Grammarly, AutoBot Builder, and nearly all Microsoft services, for which Microsoft acquired special rights from OpenAI to use the language model and integrate it into almost all its services.

This trend did not go unnoticed by other technology giants. Soon, Google and Meta reacted with their models – LaMDA and Llama, respectively.

Google introduced LaMDA as a language model primarily designed for conversational applications like chatbots. LaMDA is used through an application similar to ChatGPT called Bard, promising a more natural and engaging language compared to ChatGPT. LaMDA has slightly more parameters than GPT-3.5-turbo (137 billion) but significantly fewer than GPT-4. Similar to GPT-4, this model can process images, a capability not present in GPT-3.5-turbo. Like the GPT series, LaMDA quickly found applications across a wide range of Google-promoted applications and, along with Google's new model, PaLM, seeks to permanently change the way Google services are used.

Given its vast amount of natural language data, Meta logically launched its series of language models - Llama. However, it's worth mentioning that Llama has significantly fewer parameters than both competitors, reduced contextual analysis compared to GPT and LaMDA, no multimodality, and noticeably diminished creativity and informativeness in comparison. On the other hand, Llama is significantly more efficient in terms of resource requirements, more adaptable, and more accessible, making it very popular, especially for quick solutions.

The emphasized flexibility has led to a large number of models created through additional customization and adaptation of the basic Llama model. One such model is Solar-0-70b, tested in this work. The notable advantage over the base version is the breadth of context the model can analyze when generating results. In August 2023, this model was the top-ranked publicly available language model according to the popular HuggingFace community list, making it a candidate for analysis in this study.

Since these are solutions with varying levels of availability and developmental stages, and there have been no prior attempts to empirically test them for tasks in the Montenegrin language, each of the listed models has undergone testing in this work, and the results will continue to be discussed.

## Selection of metrics

Automatic performance testing is fundamentally based on the logic of comparing machine translation with a reference translation, typically produced by a human translator or linguist. Several different parameters can be used for this purpose, among which the following stand out:

* BLEU - Bilingual Evaluation Understudy
* CER – Character Error Rate
* WER - Word Error Rate

BLEU was introduced in a paper[[7]](#footnote-7) in 2002 with the aim of automatically evaluating machine translation results. The methodology for calculating BLEU results is quite straightforward. It counts the number of identical n-grams within pairs of texts for which the BLEU score is calculated. By default, the BLEU score calculates matches for individual words, 2-grams, 3-grams, and 4-grams. The final BLEU score is the product of all n-grams divided by the total number of n-grams. The geometric mean is used for the aggregate BLEU score of all sentence pairs. The following example illustrates the calculation of BLEU scores for two sentences:

|  |
| --- |
| Sentence #1: *Otac mora uvijek da trazi nacine da oprosti svojoj djeci ali Henrijevi zlocini protiv nas i protiv nase vjere su previse duboki da bi bili oprosteni* |
| Sentence #2: *Otac uvijek mora traziti nacine da oprosti svojoj djeci ali Henryjevi zlocini prema nama i prema nasoj vjeri su suvise duboki da bi se oprostili* |
| **Step #1**: Dividing sentence to n-grams:  1-grams: "Otac," "uvijek," "mora," "da," "trazi," "nacine," itd.  2-grams: "Otac uvijek," "mora traziti," "traziti nacine," itd.  3-grams: "Otac uvijek mora," "uvijek mora traziti," itd.  4-grams: " Otac uvijek mora traziti," "uvijek mora traziti nacine," itd. |
| **Step #2:** Comparing n-grams |
| **Step #3:** calculating precisions  1-grams: 16/17  2-grams: 12/16  3-grams: 7/15  4-grams: 1/14 |
| **Step #4:** Calculating BLEU score: |

Table 1: Methodology of BLEU score calculation

BLEU score, although developed in a time before modern language models, has been a dominant metric in fundamental tests of language model capabilities for over a decade. Bleu score, although developed in a time before modern language models[[8]](#footnote-8), has been a dominant metric in fundamental tests of lanuage model capabilitiesfor over a decade, although several authors promote paradigm shift and switch to sentence level over level of n-grams[[9]](#footnote-9).

CER (Character Error Rate) and WER (Word Error Rate) represent a similar methodology, relatively independent of usage and purpose, very simple to implement, but highly sensitive to the smallest differences. This sensitivity will be further analyzed during the interpretation of the results of this research. Both methodologies calculate the difference between the text generated by the language model and the reference model. However, WER considers differences in words, while CER is slightly more precise calculating difference in characters[[10]](#footnote-10):

Equation 1: CER score

Equation 2: WER score

Errors in the generated text, represent the substitution, addition, or deletion of letters in the case of CER results, or words in the case of WER results. Consequently, the WER result is significantly more sensitive than the CER result, as illustrated in the example shown in Table 2:

|  |
| --- |
| Sentence #1: „*Automatska evaluacija modlea“* |
| Sentence #2: „*Automatska evaluacija modela“* |
|  |
|  |

Table 2: Methodology of CER and WER score calculation

As both results range from 0 to 1, where 0 represents a perfect match and 1 indicates complete differentiation, it is clear that in the illustrative example, the sensitivity of the WER methodology significantly compromises the result. However, this sensitivity is not as alarming for longer texts, making WER one of the more popular methodologies for evaluating the performance of generative language models. The advantages and drawbacks of all three evaluation methodologies will be discussed in the third part of the paper.

# Thai language

The chapter about corpora

## Reference corpora

Since all three performance metrics heavily depend on the reference text used for comparison, it was essential to ensure representative translations from English to Montenegrin. For this purpose, a corpora of translations of television content broadcast on national television frequencies was used. This corpora was created by Assoc. Prof. Dr. Petar Božović, who was simultaneously responsible for translating the content. The corpus has been published under the name MontenegrinSubs-1.0[[11]](#footnote-11). The corpora consists of parallel texts of translations of television series from English, totaling 853,165 words. It was primarily developed for the study of linguistics[[12]](#footnote-12).

For the purposes of this research, 125 pairs of sentences were isolated from the corpus, each ranging from 100 to 150 characters, consisting of the original text and its accompanying translation. Considering the methodology of calculating BLEU, CER, and WER results, it was clear that the data needed preprocessing, primarily by removing diacritical marks (e.g., č-c, š-s) and then eliminating punctuation marks.

After preprocessing, the original English sentences were translated by language models and paired again with the reference translations. The proposed language models reacted differently to the task of translating from English to Montenegrin and performed the task with varying degrees of success. However, even at this stage of development, some models had to be abandoned such as LLAMA-2-70b[[13]](#footnote-13) because, in the first requested translation into Montenegrin, the model generated information: „*I'm not able to provide a translation of that sentence into Montenegrin as it is not a widely spoken language and I don't have access to a reliable dictionary or grammar guide for it.*“. Such algorithmic output emphasizes the importance of further investment in the field of NLP in the Montenegrin language and research towards the development of this branch of science. The translation was successfully completed using the models: GPT-3.5-turbo, Solar-0-70b, and LaMDA, while partial translation was achieved using the PaLM and GPT-4 models. The results will be discussed separately as "advanced" versions of their predecessors (LaMDA and GPT-3.5-turbo, respectively). In the final part of the research, all translations from language models will be compared with reference translations to obtain BLEU, CER, and WER results, which will be discussed in the next section of the paper.

# BLEU scores

The preliminary implementation of the evaluation using the BLEU methodology yielded expected results, along with anticipated challenges and issues described in the previous chapter. Due to the specific function for calculating BLEU results, in cases where there is no match for any combination of 4 words (4-gram) in sentence pairs, the cumulative result is 0, even though the sentences are translated quite similarly. Regardless of the number of matches on smaller n-gram instances (1-gram, 2-gram, 3-gram) in the product with zero matches at the 4-gram level, the cumulative BLEU score is zero. Additionally, such results prevent the calculation of the aggregate BLEU score for all pairs in the corpus because its calculation uses the geometric mean, which is sensitive to zero results.

There is an idea to calculate BLEU results with a lower degree of n-grams, usually three, which, in the case of three language models, was tested for the purposes of this study and gave results shown in Table 3:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Language models** | **Scores higher than 0 on 4-gram level** | **Scores higher than 0 on 3-gram level** | **Cumulative BLEU score for scores higher than 0 (4-gram level)** | **Cumulative BLEU score for scores higher than 0 (3-gram level)** |
| GPT-3.5-Turbo | 61/125 | 98/125 | **0.2352** | **0.0141** |
| Solar-0-70b | 45/125 | 75/125 | **0.2550** | **0.0136** |
| LaMDA | 64/125 | 97/125 | **0.2316** | **0.0157** |

Table 3: BLEU scores for three language models on 4-gram and 3-gram models

# CER Scores

Unlike BLEU results, there were no technical challenges with the CER methodology. However, the pronounced sensitivity of the methodology resulted in a significant number of extreme values that negatively affected the aggregated result shown in Table 4:

|  |  |
| --- | --- |
| **Language model** | **Average CER score** |
| GPT-3.5-Turbo | 44.48 |
| Solar-0-70b | 47.83 |
| LaMDA | 43.49 |

Table 4: CER Scores of three languageg models

Although using the BLEU methodology, Solar-0-70b achieved the best results at the 4-gram level, with the CER methodology, it had the highest error rate in characters compared to the reference value. On the other hand, LaMDA had the lowest error rate, even though its BLEU score was the lowest.

# WER Scores

Similar to the CER methodology, testing was conducted without significant technical issues, and the results are provided in Table 5:

|  |  |
| --- | --- |
| **Language model** | **Average WER score** |
| GPT-3.5-Turbo | 0.6792 |
| Solar-0-70b | 0.7318 |
| LaMDA | 0.6767 |

Table 5: WER Scores of three language models

WER results provided an even clearer picture of the performance of these three language models, with a pronounced deviation of the Solar-0-70b model and a slight dominance of the LaMDA language model. This points to a correlation between CER and WER results but also highlights the difference of these methodologies from the BLEU methodology.

# Discussion

Automatic evaluation can be a good indicator but fails to highlight certain fundamental shortcomings in the performance of language models. First and foremost, it is important to emphasize that the model with the best BLEU result does not use the standard ijekavian variant within the Montenegrin language. However, the fact that GPT-3.5-turbo and BARD generate text in the ijekavian variant does not necessarily mean that their performance is better; on the contrary, it suggests a higher number of words translated into Croatian due to the lack of information about the Montenegrin language or the knowledge of appropriate words in Montenegrin. Examples of such behavior are numerous: "sveučilište" (university), "uhićeni" (arrested), "izravno" (directly), "opće pobune" (general rebellion), "životopis" (CV/resume), "listopad" (October), etc. It can be assumed that Solar-0-70b, being trained on data owned by Meta, associates the Serbian dialect with the Montenegrin language, as a product of the overlap of user interests on social media platforms that Meta manages. Therefore, the translation corpus of this language model predominantly contains the ekavian variant. On the other hand, BARD and GPT associate the ijekavian variant with the Croatian language, which has a much larger user base.

Regarding GPT-3.5-turbo, it is worth mentioning that manual reading of the translations suggests that the context is not significantly missed, and there are no sentences where the translation is absolutely unclear, as is the case with the Solar-0-70b model. For example: "What fancy what folly has led and seduced you to make this most shameful rebellion against our most noble and righteous king and sovereign?" / "sta fantazija sta ludilo vodi i vabiti tebe da ucinis ovu najstidniju pobunu protiv naseg najplemenitijeg i pravednog kralja i vladara?" / "Koja fantazija koja ludost vas je vodila i zavela da izvrsite ovaj najljigaviji ustanak protiv naseg najplemenitijeg i najpravednijeg kralja i suverena?".

Additionally, by using the more advanced GPT-4 model on a limited set of 20 pairs, the BLEU score improved by 9.5% to 0.2714. This indicates a significant improvement in the performance of this series when it comes to translating from English to Montenegrin. However, it must be taken into account that GPT-3.5-turbo is free, while GPT-4 involves financial costs.

Solar-0-70b showed the best results through the BLEU methodology but weaker results in the CER and WER methodologies, which may reveal the specificity of the operation of this language model. Manual reading of the translations from this language model suggests that the context of the translation is often lost, especially in grammatical and spelling irregularities (e.g., "Njihovo ubojstvo je udario cijeli kršćanski svijet/Njihovo ubistvo je sokiralo cijelu hrišćansku zajednicu"). However, this model stands out for its speed and availability, so it should not be ignored in the context of adapting to specific tasks.

The Google language model LaMDA showed the highest precision according to automatic metrics CER and WER, but its BLEU result was lower than the other two language models. Nevertheless, a manual review of the translation results clearly indicates that LaMDA has almost perfect contextualization of translations, which is not surprising given that it was conceived and developed as a conversational model. LaMDA achieved a better BLEU result than the newer Google language model, PaLM, in the test environment; the average result of the LaMDA model was 2.2% better than PaLM.

By manually reading the translations, a subjective feeling is gained that the LaMDA model's translations seem the most natural. However, for such an assessment, the opinion of an expert in linguistics would be necessary. Automated tests assigned the greatest similarity to reference translations to the GPT-4 language model, but LaMDA closely follows. In any case, this can be a good recommendation for a linguistics novice when choosing a language model for a specific task and justifies the use of these metrics with a significant degree of caution.

Unfortunately, none of the tested language models demonstrated a sufficient level of understanding of the specifics of the Montenegrin language. On the contrary, when translating from English to Montenegrin, expressions from one of the related languages were dominantly used. This justifies research of this nature and investment in this field of science.

# Conclusion

The research aimed to present the strengths and weaknesses of automatic metrics in evaluating the performance of language models in English to Montenegrin translation tasks. While automatic metrics can serve as a good indicator when choosing an appropriate language model, the absolute recommendation is to include manual inspection and testing.

Empirical research revealed that BLEU results effectively assess adequacy and fluency in translation but poorly recognize grammatical and spelling irregularities. On the other hand, CER and WER results emphasize grammatical and spelling correctness but overlook the contextualization of the text. The findings suggest that automatic metrics can be used to evaluate language model performance in translation tasks, but it's crucial to employ a combination of metrics and manual evaluation to obtain a complete picture of the model's success.

A significant conclusion from this research is that the most popular and largest language models have very little or no information about the Montenegrin language. In attempting tasks in this language, they resort to dictionaries and dialects of related languages. This highlights the need for significant investment in this field and intensified research efforts specifically focused on the Montenegrin language.

# Literatura

1. Hirschman, “The evolution of evaluation: Lessons from the message understanding conferences”, Computer Speech and Language, 12:283–285. 1998.
2. J. Libovicky, T. Brovelli, B. Cantoni, „Machine Translation Evaluation beyond the Sentence Level“, Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 179–188, 2018.
3. K. Leung, „Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER)“, „Towards Data Science, dostupno na: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510 , 2021.
4. K. Papineni, et.al. “BLEU: a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318., 2002.
5. M. Post, „A Call for Clarity in Reporting BLEU Scores“, arXiv:1804.08771v2 [cs.CL] 12 Sep 2018
6. O. Caglayan, P. Madhyastha, L. Specia, „Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale“, arXiv:2010.13588
7. P. Božović, et.al., „Opus-MontenegrinSubs 1.0: First electronic corpus of the Montenegrin language“, Conference on Language Technologies & Digital Humanities Ljubljana, 2018.
8. P. Božović, T. Erjavec, J. Tiedemann, M. Ljubešić, V, Gorjanc, “English-Montenegrin parallel corpus of subtitles Opus-MontenegrinSubs 1.0, Slovenian language resource repository”, CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1176. 2018.
9. AutoBot Builder, https://startup.outlookindia.com/sector/saas/gupshup-launches-auto-bot-builder-powered-by-gpt-3-news-7219
10. Google, „LaMDA: our breakthrough conversation technology“, dostpuno na: https://blog.google/technology/ai/lamda/
11. Google, Bard, https://bard.google.com/
12. Google, Palm 2, https://blog.google/technology/ai/google-palm-2-ai-large-language-model/
13. Google, PaLM2, https://ai.google/discover/palm2/
14. Grammarly, https://mashable.com/article/grammarly-chatgpt-style-ai-tool-writing-editing-announcement
15. https://news.microsoft.com/source/features/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/
16. HuggingFace, Open LLM Leaderboard, https://huggingface.co/spaces/HuggingFaceH4/open\_llm\_leaderboard
17. Meta, Llama 2, https://ai.meta.com/llama/
18. Meta, Llama2, https://www.llama2.ai pristupljeno Oktobar 2023.
19. OpenAI, ChatGPT, https://chat.openai.com/
20. OpenAI, GPT-3.5-turbo, https://platform.openai.com/docs/models/gpt-3-5
21. OpenAI, GPT4, https://openai.com/gpt-4
22. Upstage, Solar-0-70b, https://huggingface.co/upstage/SOLAR-0-70b-16bit

1. OpenAI, GPT-3.5-turbo, <https://platform.openai.com/docs/models/gpt-3-5> [↑](#footnote-ref-1)
2. Upstage, Solar-0-70b, <https://huggingface.co/upstage/SOLAR-0-70b-16bit> [↑](#footnote-ref-2)
3. Google, „LaMDA: our breakthrough conversation technology“, dostpuno na: <https://blog.google/technology/ai/lamda/> [↑](#footnote-ref-3)
4. OpenAI, GPT4, <https://openai.com/gpt-4> [↑](#footnote-ref-4)
5. Meta, Llama 2, <https://ai.meta.com/llama/> [↑](#footnote-ref-5)
6. Isto kao 6. [↑](#footnote-ref-6)
7. K. Papineni, et.al. “BLEU: a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318., 2002. [↑](#footnote-ref-7)
8. M. Post, „A Call for Clarity in Reporting BLEU Scores“, arXiv:1804.08771v2 [cs.CL] 12 Sep 2018 [↑](#footnote-ref-8)
9. J. Libovicky, T. Brovelli, B. Cantoni, „Machine Translation Evaluation beyond the Sentence Level“, Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 179–188, 2018. [↑](#footnote-ref-9)
10. K. Leung, „Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER)“, „Towards Data Science, dostupno na: <https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510> , 2021. [↑](#footnote-ref-10)
11. P. Božović, T. Erjavec, J. Tiedemann, M. Ljubešić, V, Gorjanc, “English-Montenegrin parallel corpus of subtitles Opus-MontenegrinSubs 1.0, Slovenian language resource repository”, CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1176>. 2018. [↑](#footnote-ref-11)
12. P. Božović, et.al., „Opus-MontenegrinSubs 1.0: First electronic corpus of the Montenegrin language“, Conference on Language Technologies & Digital Humanities Ljubljana, 2018. [↑](#footnote-ref-12)
13. Meta, Llama2, <https://www.llama2.ai> pristupljeno Oktobar 2023. [↑](#footnote-ref-13)