Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams

Roberts Dargis, Guntis Bārzdiņš, Inguna Skadiņa, Normunds Grūzītis, Baiba Saulīte
Institute of Mathematics and Computer Science, University of Latvia
Raina bulv. 29, Riga, LV-1459, Latvia
{roberts.dargis, guntis.barzdins, inguna.skadina, normunds.gruzitis, baiba.saulite}@lumii.lv

Abstract

The latest large language models (LLM) have significantly advanced natural language processing (NLP) capabilities across various tasks. However, their performance in low-resource languages, such as Latvian with 1.5 million native speakers, remains substantially underexplored due to both limited training data and the absence of comprehensive evaluation benchmarks. This study addresses this gap by conducting a systematic assessment of prominent open-source LLMs on natural language understanding (NLU) and natural language generation (NLG) tasks in Latvian. We utilize standardized high school centralized graduation exams as a benchmark dataset, offering relatable and diverse evaluation scenarios that encompass multiple-choice questions and complex text analysis tasks.

Our experimental setup involves testing models from the leading LLM families, including Llama, Qwen, Gemma, and Mistral, with OpenAI's GPT-4 serving as a performance reference. The results reveal that certain open-source models demonstrate competitive performance in NLU tasks, narrowing the gap with GPT-4. However, all models exhibit notable deficiencies in NLG tasks, specifically in generating coherent and contextually appropriate text analyses, highlighting persistent challenges in NLG for low-resource languages.

These findings contribute to efforts to develop robust multilingual benchmarks and to improve LLM performance in diverse linguistic contexts.

1 Introduction

The dream that artificial intelligence (AI) can perform many tasks in a similar manner to humans became closer with the release of ChatGPT by OpenAI in November 2022. Today, several large language models (LLM) have been made available by global companies and are widely used by society and industry for various text generation tasks, such as question answering, text summarization, translation, etc. However, LLMs have shown considerably less reliable results for low-resource languages (Lai et al., 2023; Ahuja et al., 2024). The reason for this is the fact that most of the language data used for training LLMs is in English and few other widely spoken languages, while low-resource languages are represented by very small portions of data.

Benchmarking is a crucial step in evaluating LLM performance and capabilities across various tasks. It involves setting standardized tests or tasks to measure the LLMs' performance. A lack of benchmarks that enable comprehensive multilingual evaluation is one of the reasons why research on LLMs and machine learning models for NLP is still mostly focused on English and some other widely spoken languages.

The aim of this paper is to conduct an initial evaluation of open-weights LLM capabilities in Latvian, both in natural language understanding (NLU) and in natural language generation (NLG). The evaluation was performed using high school centralized graduation exams, overseen by the National Centre for Education. High school exams serve as an excellent benchmark dataset because they offer a relatable point of reference, allowing for comparison not only between different models but also between the performance of LLMs and the expected achievements of high school graduates.

2 Related Work

Recent advances on LLMs have led to impressive gains on NLU benchmarks, starting from GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) with 10 tasks related to different NLU problems, followed by MMLU (Hendrycks et al., 2021) which covers nearly 60 subjects (including STEM: science, technology, engineering and mathematics), and Bigbench (Srivastava et al., 2023) with more

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 289–293
November 16, 2024.

than 200 tasks, as well as many other benchmarks.

Again, many well-known benchmarks are available only in English and other widely spoken languages. Google researchers addressed the need for a highly multilingual benchmark when the first transformer-based LLMs appeared by introducing the Cross-lingual Transfer Evaluation of Multilingual Encoders (XTREME) benchmark (Hu et al., 2020) which is used to evaluate cross-lingual generalization capabilities of multilingual representations. Although XTREME covers 40 typologically diverse languages, spanning 12 language families, Baltic languages are not included in this benchmark. Similarly, the dataset for the evaluation of multilingual LLMs developed by Okapi (Lai et al., 2023), in which the English part was translated with the help of ChatGPT, covers 26 languages except the Baltic languages (the "smallest" language is Danish with 6 million speakers, followed by Slovak with 7 million speakers).

The development of test sets for specific languages involves significant human resources. Therefore a widely used strategy is to apply machine translation, with or without manual post-editing. Recently, this approach was chosen to translate the MMLU and COPA (Gordon et al., 2012) datasets into Latvian.[1] Evaluation of OpenAI ChatGPT 3.5 Turbo and Google Gemini 1.0 Pro on the machine translated MMLU dataset shows that performance of these LLMs for Latvian is worse when compared to English (Bakanovs, 2024). It should be noted, that this dataset is not manually post-edited, and machine translation most likely has introduced some errors which can result in generating wrong answers. Bakanovs (2024) experiment on a small subset of the social science domain shows that post-editing improves results by 3 percentage points for ChatGPT-3.5 Turbo and by 9 percentage points for Gemini 1.0 Pro.

Finally, GPT-4 has been evaluated by OpenAI on several benchmarks (OpenAI et al., 2024), such as MMLU, HellaSwag, AI2 Reasoning Challenge, WinoGrande, HumanEval, and DROP. When comparing GPT-4's 3-shot accuracy on MMLU across different languages, English reaches 85.5% (only 70.1% for GPT 3.5), while Latvian – 80.9%. With respect to educational tests and exams, OpenAI has reported that "GPT-4 exhibits human-level performance on the majority of professional and academic exams" (OpenAI et al., 2024).

[1]The Latvian versions of these datasets are available at https://github.com/LUMII-AILab/VTI-Data

3 Test Setup

All tests were run using the Ollama toolkit on a computer with 8x interconnected Nvidia A100 80GB GPUs.

The most popular open-source LLM families were chosen to be tested: Llama, Qwen, Gemma, and Mistral. A non-quantized *instruct-fp16* version was chosen for each model, except for Llama3.1 405B because the model was too large, therefore its 5-bit K-quantized version was used instead.

The emphasis in this article is on open-weights models. OpenAI's GPT-4o model is added just for a reference as the most popular closed-source commercial model. In the GPT-4 technical report (OpenAI et al., 2024), Latvian is classified as a low-resource language. Although it could be argued that Latvian is not as low-resource as many other languages, especially w.r.t. to the number of native speakers, it is considered as low-resource also by European researchers (Ali and Pyysalo, 2024).

4 Centralized High School Exams

In Latvia, centralized exams are a crucial component of the educational system, designed to standardize knowledge assessment across the country and to ensure that high school graduates meet national academic standards. These exams are taken at the end of the 11th or 12th grade and are required to obtain a high school diploma.

Students in Latvia must take a certain number of centralized exams, though they have some flexibility in choosing which subjects to be examined in, depending on their future academic and career aspirations. It is not expected for a student to be able to pass the exams in all subjects. The mandatory exams include Latvian language and literature, mathematics, and a foreign language of choice (usually English, but alternatives such as German, Russian, or French are available). Beyond these core subjects, students can opt to take additional exams in subjects like biology, chemistry, physics, history, geography, or informatics.

Higher education institutions in Latvia typically use these scores as part of their admission criteria, often alongside other considerations such as entrance exams or interviews. This makes the performance on centralized exams a significant factor in a student's educational trajectory.

These exams are designed and administered by

Model	Val.	Con.	Corr.
gpt-4o	1.00	0.88	**0.82**
gpt-4o-mini	1.00	0.86	0.78
llama3.1 : 405b	0.99	0.75	0.72
qwen2 : 72b	1.00	0.89	0.72
llama3 : 70b	1.00	0.88	0.71
gemma2 : 9b	1.00	0.89	0.68
gemma2 : 27b	0.97	**0.90**	0.67
llama3.1 : 70b	1.00	0.72	0.64
mistral-large : 123b	0.99	0.71	0.63
gemma2 : 2b	0.97	0.71	0.40
qwen2 : 7b	0.97	0.64	0.40
llama3 : 8b	0.92	0.43	0.31
llama3.1 : 8b	0.93	0.32	0.26
mistral-nemo : 12b	0.10	0.00	0.00

Table 1: LLM performance on MCQ tests in Latvian. Val. – validity; Con. – consistency; Corr. – correctness.

the National Centre for Education.[2] The exams are intended to assess not just rote memorization, but also critical thinking, problem-solving abilities, and application of knowledge. The structure of the exam and the types of tasks vary from year to year.

Exams of 2023 were chosen for the initial version of this benchmark, since a lot of the exercises contained multiple-choice questions (MCQ). In addition, models were also tested on text analysis task from the Latvian language and literature exam.

5 Multiple-choice Questions

A set of 72 Latvian MCQs was created, covering physics, geography, chemistry, biology, Latvian language and literature exams. Questions containing pictures and complex formulas were omitted.

The models were tested with the zero-shot learning approach. The prompt started with a question, followed by answer options and concluded with the instruction: "Atbildi formātā 'Atbilde ir X', kur X ir pareizās atbildes burts." ("Answer in the form 'Answer is X' where X is the letter of the correct answer."). The results are shown in Table 1.

The first criterion evaluated was validity – how many of the generated answers matched the expected format. Many models achieved 100% validity, indicating that instructions were understood and the zero-shot approach works well for this kind of task. For an answer to be valid it must contain the phrase "Atbilde ir" followed by a letter A–Z. There can be any number of whitespaces and asterisks (used by some models to indicate bold text in the markdown syntax) between the phrase and letter. The upper/lower case of letters is ignored. The answer may also contain extra text (usually, an explanation) before or after the phrase.

The second criterion evaluates consistency. Each prompt was sent to each model 10 times with a different seed value each time to reduce the chance of a lucky guess. The model must choose the same option for the same question every time. This criterion was evaluated on a per-question basis. To count an answer to a question as consistent, all answers to the same question must be valid, and the chosen option must be the same in every attempt.

The final measure binds it all together. For a question to be counted as correctly answered, all responses must be valid, consistent, and correct. Such a strict requirement was used to measure the true expected correctness rate. The questions in biology and geography had higher correctness scores overall. The chemistry scores were lower because some of the questions contained chemical reaction equations, and some physics questions required not only reasoning, but also calculations.

The non-quantized *fp16* models had very similar correctness compared to 5-bit K-quantized models. For such tasks, the quantized models would be more appropriate due to their significantly smaller memory and compute footprint.

6 Text Analysis and Writing Skills

One of the tasks in the Latvian language and literature exam in 2023 was to read two texts (each about 600 words) and write a text analysis (500–600 words) comparing both texts, following the principles of text composition and including the specified content components:

- Topic, relevance, and issues.
- Cultural facts, signs, or symbols in the interpretation of the cultural-historical context.
- Connection with other cultural facts beyond the provided texts.
- Text composition, genre characteristics.
- Language tools typical to the author's style in the analyzed texts.

The same task was given to the largest model from each of the LLM families. The result was

[2]Past exams are available at https://www.visc.gov.lv/lv/valsts-parbaudes-darbi

Model	Understanding (4-16)	Argumentation (3-12)	Language (0-16)	Creativity (4-16)	Total (11-60)
gpt-4o : 2024-08-06	12	9	14	11	46
gemma2 : 27b	12	9	11	10	42
llama3.1 : 405b	9	7	13	9	38
mistral-large : 123b	11	9	7	7	34
qwen2 : 72b	4	3	5	4	16

Table 2: Human expert evaluation of LLM text analysis and writing skills.

evaluated by an expert in linguistics using the same guidelines and criteria as students were evaluated on the exam. The results are evaluated according to 15 criteria divided into four categories: knowledge and understanding, argumentation, language quality, and creativity. For each criterion, students can get 1 to 4 points, except for language quality for which 0 points can be assigned as well. The overall score can range from 11 to 60 points. In our evaluation, the same scoring method was used to strictly comply with the official guidelines.

The results of the evaluation of the text analysis and writing skills are shown in Table 2. GPT-4o and Mistral are the only models that generated text within the requested length (500–600 words). Gemma2 generated almost 500 words, while Qwen2 and Llama3.1 generated about 250 words. The language quality of Llama3.1 and Gemma2 was very similar. According to the guidelines, language quality is based on absolute number of errors, thus, comparing two texts of similar relative quality, the longer one typically will have more errors and therefore a lower score.

Similarly, knowledge and understanding was based on the number of facts mentioned in the text, therefore shorter analysis had a disadvantage in this category.

The text generated by Qwen2 was very difficult to understand with many illogical sentences, which led to a low score in other categories. The text generated by Mistral had many agreement errors, such as subject-verb, noun-adjective, tense, gender, and singular/plural disagreement.

Demonstrating author's individuality was one of the conditions to get top scores in originality (part of creativity), which was lacking in all of the analysis. There was also a lack of comparison to nowadays, which was a condition to get top scores in the knowledge category.

All of the models analyzed the two texts mostly separately, using the specified content component subsections. A cohesive, fluent analysis with introduction, discussion, and conclusion was expected instead. In this task, zero-shot learning did not work well. For such tasks, examples or more detailed instructions provided in the prompt would probably lead to better results.

Experiments were also conducted using LLMs as evaluators. Each model was asked to assess the text analysis generated by every other model ten times. The results were not promising. The scores varied a lot between the runs, and the average scores by any model did not correlate with human evaluation. The worst category is language quality assessment. Most of the errors found by the models were not actually errors, and many actual errors were missed.

7 Conclusion and Further Work

The experiments validated the use of Latvian high school centralized graduation exams as a source for natural language understanding (NLU) tasks. This gives us motivation to continue the work on expanding the size of the data set. The performance gap for Latvian between the best open-weights LLMs and GPT-4o is minor. The biggest surprise was the performance of the relatively small Gemma2 27B model. The quantized version is small enough to be run on a consumer grade GPU, making it perfect for large-scale NLU tasks, such as classifying or tagging documents even without fine-tuning. This opens up huge possibilities for NLP in Latvian, especially in digital humanities.

The performance in natural language understanding did not correlate with the performance in natural language generation (NLG) in text analysis tasks, showing the importance of evaluating both tasks separately. Despite NLG shortcomings discussed, in human evaluation the best open-source LLMs achieved score above 66% (40 out of max 60 points) compared to average 56% score reported for the actual human graduation exam. Unfortu-

nately, NLG tasks are hard to evaluate, since human evaluation requires a lot of resources. Even the best models showed no correlation between their assessment of other LLM's on the NLG tasks, and the human evaluation. This makes the current generation of LLMs not well suited for NLP tasks like error detection and correction, text normalization and data denoising in Latvian.

The high out-of-vocabulary (OOV) word density score, measured against the large Latvian Thesaurus database (Grasmanis et al., 2023) was a good indicator of poor language quality, but a low number of OOV words is not an indicator of high NLG score, because most errors were grammatical errors. Finding a good automatic NLG evaluation methodology is still an open research question.

The dataset created and used in this evaluation is available as open data via a GitHub repository.[3]

Overall, the open-weights models show promising performance on Latvian, suggesting that fine-tuning such models for low-resource languages might achieve competitive results with much lower costs compared to training language-specific LLMs from scratch.

Acknowledgments

This work was supported by the EU Recovery and Resilience Facility project Language Technology Initiative (2.3.1.1.i.0/1/22/I/CFLA/002) in synergy with the Latvian Council of Science grant Common Writing Errors in Latvian (lzp-2023/1-0481).

References

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *Preprint*, arXiv:2311.07463.

Wazir Ali and Sampo Pyysalo. 2024. A Survey of Large Language Models for European Languages. *Preprint*, arXiv:2408.15040.

Bruno Bakanovs. 2024. Large Language Model Evaluation and Improvements for the Latvian Language. Master's thesis, University of Latvia.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 394–398. Association for Computational Linguistics.

Mikus Grasmanis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma, Laine Strankale, Arturs Znotins, and Normunds Gruzitis. 2023. Tēzaurs.lv – The Experience of Building a Multifunctional Lexical Resource. In *Electronic lexicography in the 21st century (eLex): Invisible Lexicography*, pages 400–418.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, et al. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

[3] https://github.com/LUMII-AILab/VTI-Data