First Steps in Benchmarking Latvian in Large Language Models

Inguna Skadiņa[1, 2], Bruno Bakanovs[2] and Roberts Dargis[1]
Institute of Mathematics and Computer Science[1], University of Latvia[2]
{roberts.dargis, inguna.skadina}@lumii.lv

The aim of this study was to conduct an initial assessment of natural language understanding (NLU) and reasoning skills of different large language models (LLMs) for the Latvian language.

Three different experiments performed

- The **BERT family LLMs** were evaluated using Choice of Plausible Alternatives (**COPA**) dataset.
- NLU evaluation of **two commercial LLMs** (ChatGPT-3.5 Turbo and Google Gemini 1.0) using Measuring Massive Multitask Language Understanding (**MMLU**) dataset.
- Evaluation of the impact of machine translation on the performance of different **open-source LLMs** using a multilingual **Belebele** dataset.

Evaluation of BERT Family Models

- Three models that include Latvian have been selected: multilingual BERT model (**mBERT**, 104 languages), **LVBERT** (Latvian only) and **LitLat BERT**(Latvian, Lithuanian and English).
- The **COPA dataset** consisting of 1000 common-sense causal reasoning samples were selected for evaluation. The task is to select the alternative that more plausibly has a causal relation with the premise.

```
Premise: The girl made wish.
Choice1: She saw a shooting star.
Choice2: She saw a black cat.
```

- For fine-tuning and evaluation COPA was machine translated with Tilde MT into Latvian and 280 samples of test set were post-edited.

Results (accuracy)

Model	Machine translated	Post-edited
mBert	54.62%	**55.00%**
LVBERT	60.38%	**61.54%**
LitLat BERT	58.46%	**62.69%**

Evaluation of Commercial LLMs

- Two cost-effective AI models that support Latvian and are available via a public API were selected: **GPT-3.5 Turbo** and **Google Gemini 1.0 Pro**.
- **MMLU** benchmark consists of various multiple-choice questions across 57 different subjects, grouped in four categories: human sciences, social sciences, STEM and miscellaneous (finance, accounting, global facts, etc.).

```
Which of the following was
responsible for the most death in
the 20th century?

["Earthquakes", "Volcanic
activity", "Wildfires", "Floods"]
```

- For evaluation MMLU was machine translated into Latvian with DeepL and sociology domain data were post-edited.

Results (accuracy)

	Machine translated	Post-edited
ChatGPT 3.5-Turbo	78.79%	**81.82%**
Gemini 1.0 Pro	81.82%	**90.90%**

Category	ChatGPT 3.5-Turbo	Gemini 1.0 Pro
Humanities	56.35	**60.84**
Other	55.91	**60.44**
Social Sciences	59.63	**68.89**
STEM	43.25	**49.52**
Average	52.58	**58.67**

Subject	ChatGPT 3.5-Turbo	Gemini 1.0 Pro
Humanities		
Formal Logic	31.43	**37.14**
Philosophy	60.11	**67.00**
Social Sciences		
Security Studies	53.09	**56.25**
Sociology	78.78	**80.60**
STEM		
Anatomy	**46.60**	44.19
Conceptual Physics	29.41	**52.00**
Other		
College Medicine	57.89	**58.93**
Human Aging	55.40	**64.38**

Evaluation of Open LLMs

- The most popular open LLM families were selected: **Gemma2, Llama3, Mistral-large** and **Qwen.** We also included OpenAI's **GPT-4o** and **GPT-4o-mini** models for reference as the most popular closed commercial models.
- **Belebele** is a multiple-choice machine reading comprehension dataset, consisting of 900 questions. Each question is based on a short passage from the FLORES-200 dataset and has four multiple choice answers. The dataset was **created without the use of machine translation technology**, relying solely on experts fluent in English and the target language.

Direction	Section	BLEU		chrF	
		DeepL	GPT	DeepL	GPT
Eng → Lav	passages	0.36	0.28	65.8	60.6
Eng → Lav	questions	0.29	0.18	64.7	53.1
Eng → Lav	answers	0.32	0.22	64.4	58.2
Lav → Eng	passages	0.43	0.38	69.3	67.3
Lav → Eng	questions	0.48	0.34	69.7	61.4
Lav → Eng	answers	0.34	0.26	63.7	62.0

Flores passage: The Plitvice Lakes national park is heavily forested, mainly with beech, spruce, and fir trees, and features a mixture of Alpine and Mediterranean vegetation. It has a notably wide variety of plant communities, due to its range of microclimates, differing soils and varying levels of altitude. The area is also home to an extremely wide variety of animal and bird species. Rare fauna such as the European brown bear, wolf, eagle, owl, lynx, wild cat and capercaillie can be found there, along with many more common species

Question: Which of the following rare fauna can be found in the Plitvice Lakes national park?

Answers: Turkey, Honey Badger, Cassowary, Capercaillie

	English			Latvian		
Model	**Belebele**	**DeepL**	**GPT**	**Belebele**	**DeepL**	**GPT**
gemma2:27b	94%	85%	87%	91%	90%	87%
gemma2:9b	94%	82%	85%	88%	87%	85%
gemma2:2b	83%	69%	73%	58%	55%	54%
gpt-4o	95%	**87%**	88%	**94%**	**93%**	**90%**
gpt-4o-mini	94%	83%	86%	88%	88%	85%
llama3.1:405b	**96%**	**87%**	**89%**	92%	91%	**90%**
llama3.1:70b	94%	84%	87%	87%	87%	85%
llama3.1:8b	87%	71%	74%	63%	62%	59%
mistral-large:123b	**96%**	**87%**	88%	85%	86%	80%
qwen2:72b	94%	85%	87%	87%	87%	84%
qwen2:7b	89%	79%	79%	67%	63%	61%
qwen2.5:72b	**96%**	85%	87%	91%	89%	87%
qwen2.5:32b	95%	86%	**89%**	91%	88%	86%
qwen2.5:14b	94%	83%	85%	78%	76%	73%
Average	**93%**	**82%**	**85%**	**83%**	**82%**	**79%**

Conclusion

- For the **low-resource** language Latvian, the top-performing LLMs can achieve **similar results** on both the human-created and machine-translated datasets.
- Machine translation proved **less effective for high-resource language** benchmarks, such as English, where it significantly impacted model accuracy.
- The benchmarking of open-source LLMs against proprietary systems reveals a **narrowing performance gap**.
- The choice of translation method and the inherent **properties of the language models significantly influence the outcomes** of benchmarking exercises.
- Future work should focus on creating robust, **high-quality datasets specifically for low-resource languages** and exploring novel architectures that can better generalize across linguistic diversity.
- We **provide the datasets** used in our experiments to facilitate further benchmarking of Latvian: https://github.com/LUMII-AILab/VTI-Data

This work was funded by the EU Recovery and Resilience Facility project *Language Technology Initiative* (2.3.1.1.i.0/1/22/I/CFLA/002)