First Steps in Benchmarking Latvian in Large Language Models

Inguna Skadiņa[1, 2, 3], **Bruno Bakanovs**[2, 4], **Roberts Dargis**[1, 3]
[1] Institute of Mathematics and Computer Science, University of Latvia
[2] University of Latvia
[3] {inguna.skadina, roberts.dargis}@lumii.lv
[4] bakanovs26@gmail.com

Abstract

The performance of multilingual large language models (LLMs) in low-resource languages, such as Latvian, has been under-explored. In this paper, we investigate the capabilities of several open and commercial LLMs in the Latvian language understanding tasks. We evaluate these models across several well-known benchmarks, such as the Choice of Plausible Alternatives (COPA) and Measuring Massive Multitask Language Understanding (MMLU), which were adapted into Latvian using machine translation. Our results highlight significant variability in model performance, emphasizing the challenges of extending LLMs to low-resource languages. We also analyze the effect of post-editing on machine-translated datasets, observing notable improvements in model accuracy, particularly with BERT-based architectures. We also assess open-source LLMs using the Belebele dataset, showcasing competitive performance from open-weight models when compared to proprietary systems. This study reveals key insights into the limitations of current LLMs in low-resource settings and provides datasets for future benchmarking efforts.

1 Introduction

The recent progress of large language models (LLMs) has made them very popular and widely used. Being the most widely used natural language processing technique (NLP) today, LLMs differ in their performance depending on several key factors, such as, the quality and size of the training data, the model architecture, the computational resources used for training, and the specific tasks they are evaluated on.

Most of the language data used for training LLMs is in English and few other widely spoken languages, while other languages, especially less- and low-resourced, are represented by very small portions of data. For example, in recently developed EuroLLM Multilingual Language Models for Europe, English language data form 50% of training data, while low-resourced languages, such as Latvian, Lithuanian, Estonian, Finnish, and others are represented by about 1% of data (Martins et al., 2024). As a result, although many language models are multilingual and powerful in language transfer, they have generally demonstrated considerably less reliable results on low-resource languages (Lai et al., 2023; Ahuja et al., 2024).

The fast growth of LLMs in size, language coverage, and overall quality, has made benchmarking critical for assessing LLM performance and capabilities across various tasks. A wide range of benchmarks are available to evaluate different capabilities of large language models. They span multiple categories, including natural language understanding and generation, robustness, ethics, or biases of the models (Chang et al., 2024). LLMs have demonstrated impressive gains on natural language understanding (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., 2020) covering nearly 60 subjects and Bigbench (Srivastava et al., 2023) with more than 200 tasks, as well as many other benchmarks. However, many of these benchmarks focus on the English language, as well as some other widely spoken languages and only some attempts have been made to evaluate LLM performance on low-resource languages.

A recent study of LLMs for European languages (Ali and Pyysalo, 2024) has identified eight EU languages as low-resource (Croatian, Estonian, Irish, Latvian, Lithuanian, Maltese, Slovak, and

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 86–95

Slovene).

The aim of this paper is to conduct an initial assessment of natural language understanding and reasoning skills of different LLMs for the low-resource Latvian language:

- our first group of experiments aims to evaluate NLU capabilities of different BERT family (Devlin et al., 2019) LLMs using Choice of Plausible Alternatives (COPA) dataset (Section 3);
- as next, we evaluate the performance of two commercial LLMs (ChatGPT-3.5 Turbo and Google Gemini 1.0) on widely used Measuring Massive Multitask Language Understanding (MMLU) dataset (Section 4);
- finally, we use a multilingual Belebele dataset to understand the impact of machine translation on the performance of different open-source LLMs (Section 5).

We provide the datasets used in our experiments to facilitate further benchmarking of Latvian[1], ensuring that researchers have access to the resources necessary to replicate and build upon our work. By making these datasets publicly available, we aim to support the development of robust tools and methodologies for the Latvian language, as well as foster collaboration and facilitate advancements in natural language understanding for low-resourced languages.

2 Related Work

Latvian is an Indo-European language of the Baltic branch with about 1.5 million native speakers. Taking into account its size, it is rather well supported by language technologies (Skadiņa et al., 2022). However, in the context of LLMs the Latvian language is a low-resource language (Ali and Pyysalo, 2024).

Before 2024, only limited research has been conducted on the performance of BERT family language models (e.g., Znotiņš and Bārzdiņš (2020), Vīksna and Skadiņa (2020)). A widely used Latvian dataset to assess LLM performance on different natural language processing tasks (NER, POS-tagging, dependency parsing) is FullStack-LV dataset (Gruzitis et al., 2018). Comparison of several BERT family models that include Latvian (mBERT (Devlin et al., 2019), LVBERT (Znotiņš and Bārzdiņš, 2020), and LitLat BERT (Ulčar and Robnik-Šikonja, 2021)) has been performed by Ulčar and Robnik-Šikonja (2021). The evaluation showed that the LitLat BERT model has the best performance in named entity recognition, part-of-speech tagging, and word analogy tasks, whereas LVBERT demonstrated the best score for the dependency parsing task.

Until 2024, there were no datasets available to assess the natural language understanding and reasoning skills of LLMs in Latvian and compare them across different models or languages. For example, mBERT's performance has been evaluated using the XNLI dataset (Conneau et al., 2018) - an evaluation corpus for language transfer and cross-lingual sentence understanding in 15 languages, but it does not contain any Latvian samples. 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, but does not include any of the the Baltic languages (the "smallest" language is Danish with 6 million speakers, followed by Slovak with 7 million speakers).

Latvian is mentioned as one of the languages on which the GPT-4 model was evaluated with MMLU benchmark (Achiam et al., 2023). The prompts were machine-translated from English into Latvian. 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% (Achiam et al., 2023).

Different approach has been chosen by Darģis et al. (2024), who used standardized Latvian high school centralized graduation exams as a benchmark dataset. They showed that several open-source models have reached competitive performance in NLU tasks, narrowing the gap with GPT-4, while keeping notable deficiencies in natural language generation tasks (specifically in generating coherent and contextually appropriate text analyses).

Recently META has released the Belebele benchmark (Bandarkar et al., 2024). This benchmark was used to evaluate three masked language models (XLM-V, INFOXLM and XLM-R) and several LLMs (GPT3.5-TURBO, FALCON, and LLAMA). The accuracy of these models for the Latvian language varies from 37.6% for FALCON

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

40B 5-shot In-Context Learning model to 74.1% for Translate-Train-All XLM-V Large model.

Finally, European LLM leader-board that includes Latvian has been recently published on HuggingFace.[2] This leaderboard provides a comparison of more than 15 open-source multilingual LLMs across several machine-translated benchmarks – ARC, GSM8K, HellaSwag, MMLU and TruthFullQA.

3 Evaluation of BERT Family Models

While today LLMs offer broad multilingual capabilities, they may not always be the best solution for low-resourced languages, thus in some cases BERT-based models still remain relevant as a cost-effective, adaptable, and open-source alternative for research and real-world applications in under-represented languages. Although several BERT models include Latvian, their NLU capabilities have not been assessed due to the absence of necessary evaluation datasets.

3.1 COPA Dataset

In our first experiment, conducted in early spring of 2024, we evaluated several BERT models using the machine-translated [3] version of the Choice of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011).

The COPA dataset consists of 1000 commonsense casual reasoning samples. The task is to select the alternative that more plausibly has a causal relation with the premise. The dataset is split equally into two parts, one for development and the other for evaluation.

3.2 Selected Models

The following models that include Latvian have been selected: multilingual BERT model (mBERT, Devlin et al. (2019)), LVBERT (Znotiņš and Bārzdiņš, 2020), and LitLat BERT (Ulčar and Robnik-Šikonja, 2021). mBERT and LVBERT models implement the BERT reference architecture, while the LitLat BERT model is based on RoBERTa-base architecture (Liu et al., 2019). The mBERT model is pre-trained on a corpus that includes text from 104 languages, the LitLat BERT model is trained on Latvian (LV), Lithuanian (LT), and English (EN), and LVBERT is trained solely on Latvian. None of them share training datasets; however, there is some overlap between mBERT and LVBERT models, as they both contain Wikipedia datasets. Table 1 summarizes language models selected for the evaluation, their language coverage and the parameter count.

Model	Languages	Parameters (million)
mBERT	104 lang.	110
LVBERT	LV	110
LitLat BERT	LV, LT, EN	125

Table 1: Selected language models.

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

Table 2: Accuracy of BERT models on COPA dataset.

3.3 Experimental Setup

BERT models require fine-tuning of the pre-trained model for COPA task. For this, model weights were acquired from HugginFace website.[4] We added an additional linear layer and a softmax function to the pre-trained models. During the fine-tuning process for the COPA task, we experimented with different learning rates ($5e$-5, $4e$-5, $3e$-5, $2e$-5) while keeping the batch size fixed at 32 and training for 10 epochs.

We split the development dataset into 400 samples used for training and 100 samples for validation. The highest accuracy on the validation dataset on all models was achieved using $5e$-5 learning rate.

3.4 Results

The evaluation dataset consists of 500 machine-translated samples, from which 260 were post-edited by native Latvian speaker. Table 2 compares the evaluation results between 260 machine-translated and post-edited samples.

[2] https://huggingface.co/spaces/openGPT-X/european-llm-leaderboard

[3] In this experiment we used Tilde Translator https://tilde.ai/machine-translation/

[4] https://huggingface.co/google-bert/bert-base-multilingual-cased, https://huggingface.co/AiLab-IMCS-UL/lvbert, https://huggingface.co/EMBEDDIA/litlat-bert

Notably, post-edited machine translation samples bring an improvement of a few percentages. The most significant improvement has been observed for the LitLat BERT model with more than 4 percentage points. Similar gains have been noticed with BERT models in Estonian, where the post-editing lead to an improvement of a few percentages (Kuulmets et al., 2022). When compared to English BERT model (70.6%) the Latvian models perform significantly worse.

4 Evaluation of Commercial LLMs

As next, in spring 2024, we evaluated the performance of several commercial models on the Latvian language to assess their capabilities in handling low-resource languages.

4.1 MMLU Dataset

Measuring Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2020) consists of various multiple-choice questions across 57 different subjects, grouped in four categories: human sciences (philosophy, history, jurisprudence, etc.), social sciences (economics, sociology, geography, etc.), STEM (high school mathematics, college computer science, etc.), and miscellaneous (finance, accounting, global facts, etc.). The motivation for selecting MMLU benchmark comes from both its popularity and the fact that the results are available for wide-range of LLMs, including OpenAI's GPT-4 (Achiam et al., 2023), Google's Gemini family of models (GeminiTeam et al., 2024), and the recently announced NVIDIA's NVLM 1.0 (Dai et al., 2024). Similarly to COPA, MMLU was not available in Latvian, and thus was machine-translated for our experiments.

4.2 Selected Language Models

For our experiments, we selected two cost-effective AI models that support Latvian and are available via a public API: GPT-3.5 Turbo[5] and Google Gemini 1.0 Pro.[6] These models were chosen based on their balance of affordability and performance, making them suitable for conducting comprehensive tests without exceeding budget constraints.

[5] https://platform.openai.com/docs/models/gpt-3-5-turbo

[6] https://ai.google.dev/gemini-api/docs/models/gemini

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

Table 3: MMLU evaluation results (accuracy) in sociology domain with machine-translated and post-edited prompts.

4.3 Experimental Setup

The evaluation of ChatGPT-3.5 Turbo and Gemini Pro 1.0 was performed using the API provided by the developers of the models. During the evaluation of Gemini Pro 1.0, the safety filters were disabled, since with the default configuration for some prompts, no answer was provided. Both models were evaluated using 2-shot prompts, i.e., the first two multiple-choice question-answer pairs serve as examples and the model is expected to provide the correct answer for the third question.

During the evaluation, we observed that sometimes the output of models is inconsistent with the expected format. For example, if the correct answer is **D**, the model could also output variations, e.g., **(D)**, **D. 0,4**, or **(D) 0,4**. These cases were also considered as correct answers. This approach differs from Laskar et al. (2023) where the authors performed additional manual evaluation of prompts.

4.4 Results

Table 4 shows the evaluation results of machine-translated MMLU dataset per subject for both Gemini Pro 1.0 and ChatGPT-3.5 Turbo models. Overall, the accuracy of Gemini 1.0 Pro is 6.09 percentage points higher than ChatGPT-3.5 Turbo. For multiple subjects the difference of accuracy exceeds 20 percentage points. For instance, college biology, econometrics, human sexuality. However, there are also subjects, in which ChatGPT-3.5 Turbo model performed better, like computer security and public relations.

Our choice of few-shot prompts differs from those reported for English. ChatGPT-3.5 Turbo reached 67% accuracy for English using 0-shot prompts. The average accuracy of our results for Latvian across all subjects is 52.58%. The difference is significant, considering that our evaluation provided two additional examples. For English ChatGPT-3.5 Turbo accuracy of 5-shot prompts is around 70% and for Gemini 1.0 Pro accuracy is

Subject	ChatGPT-3.5 Turbo	Gemini Pro 1.0
abstract algebra	**37.500**	32.260
anatomy	**46.600**	44.190
astronomy	54.000	**72.920**
business ethics	**48.480**	40.625
clinical knowledge	**57.950**	55.290
college biology	37.500	**62.500**
college chemistry	34.375	**37.930**
college computer science	39.390	**41.940**
college mathematics	24.240	**40.000**
college medicine	57.890	**58.930**
college physics	**29.410**	28.125
computer security	**78.780**	63.630
conceptual physics	29.410	**52.000**
econometrics	31.580	**52.630**
electrical engineering	41.600	**56.520**
elementary mathematics	**43.200**	41.530
formal logic	31.430	**37.140**
global facts	33.330	**39.390**
high school biology	64.070	**77.450**
high school chemistry	**42.420**	42.370
high school computer science	63.630	**78.790**
high school European history	70.900	**77.780**
high school geography	61.530	**78.460**
high school government and politics	65.625	**79.370**
high school macroeconomics	50.000	**69.230**
high school mathematics	**34.090**	30.120
high school microeconomics	55.700	**69.620**
high school physics	40.810	**45.830**
high school psychology	65.190	**80.190**
high school statistics	40.270	**42.860**
high school US history	63.240	**76.120**
high school world history	64.100	**69.620**
human aging	55.400	**64.380**
human sexuality	58.130	**78.570**
international law	**75.000**	65.000
jurisprudence	69.400	**88.890**
logical fallacies	51.850	**53.700**
machine learning	40.540	**50.000**
management	73.530	**76.470**
marketing	78.200	**89.120**
medical genetics	**66.670**	63.640
miscellaneous	66.530	**71.150**
moral disputes	52.170	**59.650**
moral scenarios	**26.510**	24.480
nutrition	63.730	**64.700**
philosophy	60.109	**67.000**
prehistory	**59.260**	52.880
professional accounting	30.430	**47.190**
professional law	33.140	**51.970**
professional medicine	51.110	**66.670**
professional psychology	50.980	**53.000**
public relations	**72.200**	50.000
security studies	53.090	**56.250**
sociology	78.780	**80.600**
US foreign policy	72.720	**78.790**
virology	43.640	**48.150**
world religions	**75.440**	66.670
Average	52.577	**58.672**

Table 4: Comparison of Gemini Pro 1.0 and ChatGPT-3.5 Turbo on MMLU (accuracy, %).

around 71.8% .

We also verify the impact of post-editing. As the dataset is vast, post-editing was performed only for the prompts of the sociology subject. The results in Table 3 show an increase of accuracy for both models – ChatGPT-3.5 Turbo achieves a 3.03 percentage point increase, while Gemini 1.0 Pro achieves a more substantial gain of 9,08 percentage points.

5 Evaluation of Open LLMs

We continue to explore the impact of machine translation on benchmarking using the recently released multilingual Belebele dataset, which includes Latvian. We compare the performance of several popular open-weight LLM families (Gemma, Llama, Mistral, and Qwen) using both the original and machine-translated versions of the dataset.

5.1 Belebele Dataset

Belebele is a multiple-choice machine reading comprehension dataset (Bandarkar et al., 2024). The dataset was created without the use of machine translation technology, relying solely on experts fluent in English and the target language. For each language the dataset contains 900 questions. Each question is based on a short passage from the FLORES-200 dataset (NLLBTeam et al., 2022) and has four multiple choice answers.

To assess the impact of machine translation we translated English (EN) part of the Belebele dataset into Latvian (LV) and Latvian part into English using two different machine translation strategies – machine translation system DeepL[7] and GPT-4o-mini with system prompting. We used the original English and Latvian parts of this dataset as references to evaluate translations. Results of the automatic evaluation are summarized in Table 5. For both translation directions DeepL demonstrates better translation (BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) scores), when compared to GPT-4o-mini. Since Latvian is a low-resource morphologically rich free-word order language, automatic scores for English->Latvian machine translation direction are lower than for Latvian->English direction.

[7] https://www.deepl.com/en/translator

```
LV: Izlasi tekstu un atbildi uz jautājumu:
EN: Read the text and answer to the question:

{{flores_passage}}
{{question}}
A: {{option1}}
B: {{option2}}
C: {{option3}}
D: {{option4}}

LV: Atbildi formātā 'Pareizā atbilde ir X',
kur X ir pareizās atbildes burts.

EN: Answer in form 'Correct answer is X',
where X is the letter of the correct answer.
```

Figure 1: Prompt structure.

5.2 Selected Language Models

The most popular open LLM families were selected: Gemma2 (GemmaTeam et al., 2024), Llama3 (Dubey et al., 2024), Mistral-large (Jiang et al., 2023) and Qwen (Bai et al., 2023). A 5-bit K-quantized version was used for every model. We also included OpenAI's GPT-4o and GPT-4o-mini models for reference as the most popular closed commercial models.

5.3 Experimental Setup

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

The questions were asked directly in a zero-shot approach with each model's default system prompt (see Figure 1).

Some models answered with just the required phrase, some also added explanation. Therefore we used a case-insensitive regular expression:

(?:Atbilde ir|Answer is)[\s()*([A-D])*

to find the model's answer in the response.

Each question was asked three times with three different seeds to test the robustness of the models. Robustness was measured as percentage of questions to which the model chose the same answer in all three cases. The top models scored 99% robustness on human translated English data and 98% for human translated Latvian data.

5.4 Results

The evaluation results (accuracy) for different LLMs are summarized in Table 6. Each of 900 questions is considered to be answered correctly only if all three responses were equal and correct.

[8] https://github.com/ollama/ollama

Language pair	Section	BLEU		chrF	
		DeepL	GPT	DeepL	GPT
English->Latvian	passages	0.36	0.28	65.8	60.6
English->Latvian	questions	0.29	0.18	64.7	53.1
English->Latvian	answers	0.32	0.22	64.4	58.2
Latvian->English	passages	0.43	0.38	69.3	67.3
Latvian->English	questions	0.48	0.34	69.7	61.4
Latvian->English	answers	0.34	0.26	63.7	62.0

Table 5: Evaluation of DeepL and GPT-4o-mini translations (BLEU and ChrF scores).

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

Table 6: Evaluation results for different LLMs on original (Belebele column) and machine translated (DeepL and GPT columns) datasets (accuracy).

5.4.1 Original Belebele Dataset

The best result of 96% accuracy for English is achieved by several models - Qwen2.5, Mistral-large, Llama3.1, while for Latvian only gpt-4o achieved 94% accuracy, followed by several open LLMs - llama3.1:405b with 92% accuracy and gemma2:27b and qwen2.5:72b and 32b with 91% accuracy. gpt-4o also seems the most balanced model with only one percentage point difference in accuracy between Latvian and English.

In general, the model's accuracy seems to correlate with the parameter size - the smaller the model, the lower is accuracy. Although our results are not directly comparable with the results obtained by the authors of the Belebele dataset, it seems that most recent LLMs demonstrate better "understanding" of low-resource languages and the results of the best open-weight LLMs differ only by 2-3 percentage points when compared to commercial ones.

5.4.2 Machine Translated Datasets

Evaluation results in Table 6 demonstrate a decrease of accuracy in case of machine-translated datasets. For English, the accuracy for the machine-translated dataset is always below 90%, dropping by at least 5 percentage points.

In case of Latvian, most of the models demonstrate comparable performance for both original and machine-translated datasets, with only a 1-3 percentage point decrease when tested on MT-datasets.

Although the automatic evaluation of MT (see Table 5) indicated that DeepL MT outperformed GPT in terms of standard MT quality metrics, the results for English in this natural language understanding task showed a different trend. Specifically, models demonstrated better performance when using the GPT-translated dataset rather than

the DeepL-translated version.

6 Conclusion

In this study, we provided an initial assessment of several large language models' performance in Latvian across different natural language understanding tasks.

Results of our evaluation of multilingual commercial and open-source models highlights the disparities in model accuracy when applied to low-resource languages.

Our findings indicate that for the low-resource language Latvian, the top-performing LLMs can achieve similar results on both the original (human-created) and machine-translated datasets. However, machine translation proved less effective for high-resource language benchmarks, such as English, where it significantly impacted model accuracy.

While machine translation offers a feasible route to generate benchmarks for low-resource languages, it is not without its pitfalls. The choice of translation method and the inherent properties of the language models significantly influence the outcomes of benchmarking exercises.

Additionally, the benchmarking of open-source LLMs against proprietary systems reveals a narrowing performance gap. Despite these advances, significant challenges remain, including the lack of comprehensive evaluation datasets tailored to Latvian.

By introducing adapted versions of the COPA[9] and MMLU[10] datasets and evaluating models on the Belebele dataset, this paper lays the groundwork for further research in benchmarking.[11]

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.

Acknowledgments

This work was supported by the "Language Technology Initiative" project (No. 2.3.1.1.i.0/1/22/I/CFLA/002), funded by the European Union Recovery and Resilience Mechanism Investment and the National Development Plan.

[9]Datasets available at `https://github.com/LUMII-AILab/VTI-Data/tree/main/copa`

[10]Datasets available at `https://github.com/LUMII-AILab/VTI-Data/tree/main/mmlu`

[11]Datasets from all our experiments are available at `https://github.com/LUMII-AILab/VTI-Data`

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. *arXiv preprint arXiv:2311.07463*.

Wazir Ali and Sampo Pyysalo. 2024. A survey of large language models for European Languages. *arXiv preprint arXiv:2408.15040*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

Roberts Darģis, Guntis Bārzdiņš, Inguna Skadiņa, Normunds Gruzitis, and Baiba Saulite. 2024. Evaluating open-source LLMs in low-resource languages: Insights from Latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293, Miami, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

GeminiTeam, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2024. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

GemmaTeam, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, and et. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Normunds Gruzitis, Lauma Pretkalniņa, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens. 2018. Creation of a balanced state-of-the-art multilayer corpus for NLU. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hele-Andra Kuulmets, Andre Tättar, and Mark Fishel. 2022. Estonian language understanding: a case study on the copa task. *Baltic Journal of Modern Computing*, 10.

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.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for Europe. *arXiv preprint arXiv:2409.16235*.

NLLBTeam, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco

Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*.

Inguna Skadiņa, Baiba Saulīte, Ilze Auziņa, Normunds Grūzītis, Andrejs Vasiļjevs, Raivis Skadiņš, and Mārcis Pinnis. 2022. Latvian language in the digital age: The main achievements in the last decade. *Baltic Journal of Modern Computing*, 10(3):490–503.

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*.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Training dataset and dictionary sizes matter in bert models: the case of baltic languages. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 162–172. Springer.

Rinalds Vīksna and Inguna Skadiņa. 2020. Large language models for Latvian named entity recognition. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press.

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. *Advances in neural information processing systems*, 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.

Artūrs Znotiņš and Guntis Bārzdiņš. 2020. Lvbert: Transformer-based model for latvian language understanding. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press.