LAG-MMLU: Benchmarking Frontier LLM Understanding in Latvian and Giriama

Naome A. Etori[1], **Arturs Kanepajs**[2], **Kevin Lu**[3], **and Randu Karisa**[2]
[1]University of Minnesota-Twin Cities
[2]Independent Researcher
[3]Bellarmine College Preparatory
etori001@umn.edu

Abstract

This paper evaluates the language understanding capabilities of various large language models (LLMs) through an analysis of 112 translated and human-edited questions from the Multitask Language Understanding (MMLU) dataset, focusing specifically on two underrepresented languages: Latvian and Giriama. The study compares the performance of six state-of-the-art (SOTA) models, with OpenAI's o1-preview model demonstrating superior performance across all languages, significantly outperforming non-proprietary models in Latvian and Giriama. Human editing of automated translations from English to Latvian yielded only a small, statistically insignificant improvement in performance estimates, suggesting that machine-translated benchmarks may be sufficient for comparing model performance in languages with established digital resources like Latvian. However, automated translation to Giriama proved infeasible, and model performance in Giriama remained poor, highlighting the persistent challenges LLMs face with low-resource languages. These findings underscore the need for high-quality datasets and improved machine translation capabilities for underrepresented languages, emphasizing the importance of localized benchmarks and human evaluation in addressing cultural and contextual limitations in AI models.

1 Introduction

The potential benefits of advanced artificial intelligence (AI) are vast, but to ensure these advantages are globally accessible, it's crucial that AI systems perform well across multiple languages. Previous research has highlighted a significant disparity between the performance of frontier large language models (LLMs) in English compared to other languages, particularly those with limited resources (Cohere For AI team, 2024; OpenAI, 2024; Dubey et al., 2024).

Recently, there has been growing interest in assessing the capabilities of LLMs, with studies such as HELM (Liang et al., 2022), BIG-Bench (Srivastava et al., 2022), LAMBADA (Paperno et al., 2016) evaluating various model functions. However, these evaluations mostly focus on English, leaving a gap in assessing LLMs' multilingual performance. As new language technologies based on LLMs rapidly emerge, evaluating their multilingual effectiveness is crucial (Blasi et al., 2021).

As AI models continue to evolve, it's essential to monitor how this language gap is narrowing. Users working with models in various languages could greatly benefit from comparative performance analyses across different linguistic contexts. However, evaluating model performance in non-English languages presents challenges, for example manual translation is time-consuming, and this has forced the NLP community to focus on a selection of tasks and languages only. Moreover, it has become standard practice to machine translate the training set but use human translation for test sets (Choenni et al., 2024). While automated translation of benchmarks is cost-effective, it raises concerns about quality. Conversely, human translations, though potentially more accurate, can be prohibitively expensive. Driven by these considerations this study aims to address the following key questions:

- Q1: Which LLM performs best in both Latvian and Giriama tasks?
- Q2: How do model performance levels differ between English, Latvian, and Giriama?
- Q3: How does human post-editing of translations affect benchmark quality compared to pure machine translation?

In our work, we utilize the Massive Multitask Language Understanding (MMLU) benchmark,

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 109–120
March 3-4, 2025 ©2025 University of Tartu Library

which covers 57 subjects ranging from STEM to humanities and social sciences. Our goal is to enhance the understanding of LLMs performance in low-resource languages, with a specific focus on Latvian and Giriama, and to contribute to the development of AI systems that are both linguistically and culturally inclusive.

2 Related works

2.1 Multilingual models across cultures and languages

State-of-the-art (SOTA) massively multilingual language Models (MMLMs) such as mBERT (Devlin, 2018), XLMR (Conneau, 2019), and mT5 (Xue, 2020) support 100+ languages worldwide and have shown exceptional proficiency in both understanding and generating text across diverse linguistic contexts. Additionally, generative models like GPT-4 (Achiam et al., 2023), LLaMA (Touvron et al., 2023) and BLOOM (Le Scao et al., 2023) are also gaining world recognition for their contributions to advancing natural language generation and understanding. Significant challenges remain in ensuring cultural sensitivity and language equity (Dawson et al., 2024).

Studies have shown that multilingual models perform well on high-resource languages like English, French, and German, but struggle with low-resource languages (Li et al., 2024; Hedderich et al., 2020; Ranathunga and De Silva, 2022), particularly in Africa (Adelani et al., 2021; Alabi et al., 2022; Adebara et al., 2024) and South Asia (Lahoti et al., 2022; Baruah et al., 2021), due to limited training data (Adebara et al., 2024; Magueresse et al., 2020). Challenges such as cultural nuances (Romero et al., 2024; Winata et al., 2024), dialectal variation (Faisal et al., 2024), and code-switching (Winata et al., 2021) further hinder model performance. While efforts like cross-lingual transfer learning and culturally relevant datasets have been made to address these issues (Hu et al., 2020; Winata et al., 2022; Liu et al., 2021), performance gaps persist in underrepresented languages.

2.2 Datasets, benchmarks, or libraries for evaluating multi-lingual models

Most existing multilingual NLP benchmarks such as (Hendrycks et al., 2020; Hu et al., 2020; Wang, 2018; Wang et al., 2019; Guzmán et al., 2019) are heavily skewed toward high-resource languages, particularly those in the Indo-European language family, and reflect predominantly Western cultural contexts. As a result, these benchmarks fail to capture the linguistic and cultural diversity of the global population, making them less reliable in assessing the performance of multilingual language models (MMLMs) across underrepresented languages and cultures (Bender, 2019).

Recent works have focused on expanding multilingual datasets to better reflect the linguistic and cultural diversity across the world. Projects such as (Romero et al., 2024; Winata et al., 2024; Kirby et al., 2016; Miquel-Ribé and Laniado, 2019; Moran et al., 2022; Adebara et al., 2024; Ifeoluwa Adelani et al., 2024; Costa-jussà et al., 2022) are making strides in enhancing the representation of multilingual models, leveraging community-driven initiatives to build localized datasets. These efforts have highlighted the importance of understanding the cultural context in which language is used, rather than relying solely on translation-based approaches (Tiedemann, 2020).

2.3 Human evaluation of multilingual and multicultural aspects of models

Human ability to understand language is general, flexible, and robust (Wang, 2018; Lin and Och, 2004). Hence, human evaluations are typically considered the gold standard in natural language generation to assess the effectiveness of multilingual models (Clark et al., 2021; Chiang and Lee, 2023), particularly in evaluating their ability to generate text that aligns with diverse linguistic and cultural contexts. Automatic metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) even though commonly used, often fail to capture cultural nuances, making human evaluation essential for a more comprehensive assessment (Kocmi et al., 2021).

Human evaluations are essential for assessing how well multilingual models handle grammatical, syntactical, and contextual differences, particularly in low-resource languages where machine models often struggle with culturally specific terms (Costa-jussà et al., 2022). Evaluating multicultural aspects is even more challenging due to cultural references, social norms, and context-dependent meanings. Human raters are better at identifying these nuances, using criteria such as appropriateness, bias detection, and cultural sensitivity (Choenni et al., 2024)

Language	Question and Answers (Question subject: miscellaneous)
English	According to the children's nursery rhyme what type of ocean did Columbus sail in 1492? A: calm **X**, B: blue ✓, C: windy **X**, D: really big **X**
Giriama	Kulingana na wira wa kitalu cha ahoho ni aina yani ya bahari ambayo Columbus wasafiri makathi ga 1492? A: Kuhurira **X**, B: buluu ✓, C: peho **X**, D: bomu jeri **X**
Latvian (autotranslated)	Saskaņā ar bērnu bērnudārza atskaņu, kāda veida okeānu Kolumbs kuģoja 1492. gadā? A: Mierīgs **X**, B: zils ✓, C: Vējains **X**, D: Ļoti liels **X**
Latvian (autotranslated & edited)	Saskaņā ar bērnudārza pantiņu, kāda veida okeānu Kolumbs kuģoja 1492. gadā? A: Mierīgu **X**, B: Zilu ✓, C: Vējainu **X**, D: Ļoti lielu **X**

Table 1: **Sample question translated into Giriama and Latvian (AT: autotranslated, AT+E: autotranslated and edited) with correct answers marked (✓) and incorrect answers marked (X). The correct answer "blue" in English refers to the popular children's rhyme "In 1492, Columbus sailed the ocean blue," which is a cultural reference that may not resonate in Latvian or Giriama without further explanation.**

3 Methodology

3.1 Datasets

The MMLU dataset (Hendrycks et al., 2021) includes 57 subjects spanning various disciplines such as mathematics, history, computer science, law, and more. The dataset features over 15,000 questions from publicly available sources such as practice tests for exams like the GRE and USMLE. These questions are categorized by difficulty, from elementary to advanced professional levels. The benchmark is designed to evaluate models in zero-shot and few-shot settings, aiming to assess their world knowledge and problem-solving ability across diverse subjects.

3.2 Languages covered

Our benchmarks encompass Latvian and Giriama, two languages that are quite distinct both in their geographic origins and linguistic structures:

- **Latvian (lav)**: spoken by approximately 1.75 million people primarily in Latvia, belongs to the Baltic branch of the Indo-European language family and is closely related to Lithuanian, though they are not mutually intelligible. Latvian has lower digital resources as compared to high-resource languages like English, German, or Chinese and limited representation in widely used multilingual benchmarks. The complexity of Latvian, such as its rich morphology (seven cases, gender system, and inflectional forms), further adds to the difficulty of processing it with LLMs, which often struggle with the intricate grammatical structures of low- and medium-resource languages. It remains underrepresented in many NLP applications (Darģis et al., 2024).
- **Giriama (nyf)**: Giriama, or Kigiryama, is a Bantu language spoken by around 700,000 people, primarily in Kilifi County, Kenya. It is one of the nine (9) Mijikenda languages, classified under the Northeastern Bantu subgroup of the Niger-Congo family. Like many Bantu languages, Giriama is agglutinative, using affixes to express grammatical relations, and features a complex noun class system that affects agreement with verbs and adjectives. Predominantly oral, Giriama has limited written texts, though recent efforts have promoted literacy using the Latin alphabet. Despite these efforts, Giriama remains underresourced in linguistic and digital documentation.

3.2.1 Dataset collection

We created our dataset by randomly selecting 112 questions and answers from the MMLU (Massive Multitask Language Understanding) benchmark (Hendrycks et al., 2021). The dataset preparation involved three versions:

1. The original English questions (baseline)
2. Machine translations of these questions into Latvian using MyMemory API (MyMemory, 2024)
3. Human-edited translations in both Latvian and Giriama

For Giriama, we skipped machine translation since automatic translation systems frequently misidentified the language as Swahili. This three-version approach enabled us to compare LLM performance across machine-translated and human-edited content.

3.2.2 Translations and annotation process

We recruited one language coordinator, who also doubled as a translator for the Giriama language. The translator holds a master's degree in computer science and is a native speaker of the language, with extensive experience as a translator. As a token of appreciation, we provided compensation for

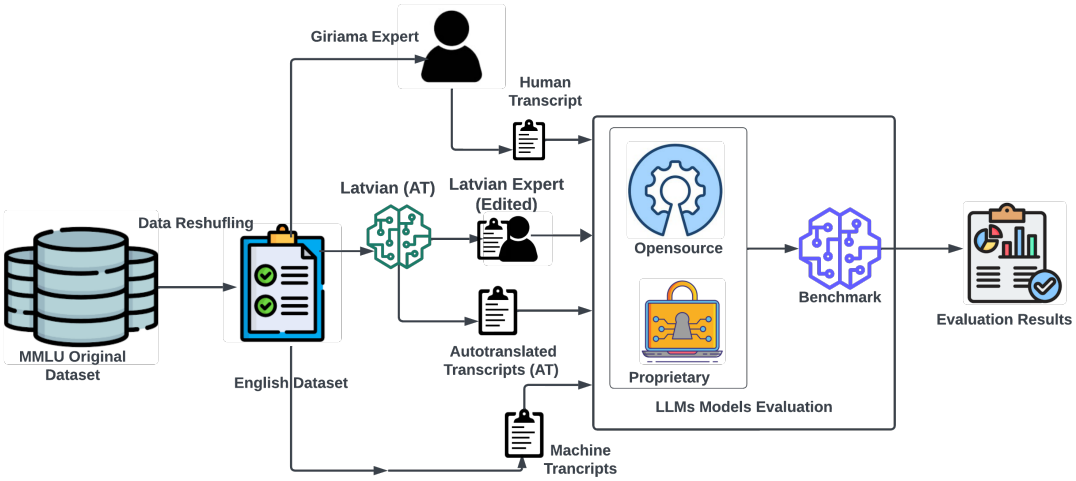

Figure 1: **Frontier LLMs in Latvian and Giriama Dataset and Benchmarking pipeline**

the work completed. For the Latvian translations, a Latvian-fluent annotator reviewed and edited the machine-translated questions. The focus was on correcting any errors that could hinder comprehension or lead to misinterpretations of the answer options. This human-edited process ensured a higher level of accuracy in both the Giriama and Latvian translations.

3.3 Task covered

Our work focuses on evaluating the multilingual understanding of LLMs by assessing their ability to process translated questions from the MMLU benchmark across three languages: English, Latvian, and Giriama. The translation tasks involve both machine-generated and human-annotated translations. Specifically, the task examines how well the models comprehend and answer 112 questions from English into Latvian and Giriama. The objective is to compare the performance of LLMs in handling machine translations versus human-annotated versions, thereby exploring the necessity and impact of human involvement in translation tasks, particularly in low-resource languages like Giriama.

4 Evaluation metrics

We evaluated the performance of six LLMs on four distinct language tasks: English, Latvian, machine-translated Latvian, and Giriama using an accuracy score. A uniform temperature setting of 0.5 was applied across all models, except for the o1-preview, for which only a fixed temperature of 1 was supported.

For each model, accuracy was computed as the proportion of correct outputs from a test set comprising 112 samples. To account for uncertainty in the performance estimates, we employed the Wilson score interval. This method provides a more accurate estimation of confidence intervals for binomial proportions p such as model accuracy by considering the sample size n and desired confidence level (typically set at $z = 1.96$ for a 95% confidence interval). The Wilson interval is preferred over traditional intervals like Wald due to its robustness, particularly with smaller sample sizes, ensuring more reliable confidence bounds around the accuracy metric.

We tested statistical significance using a two-proportion z-test, comparing each model's performance against the highest-performing model in its respective task category. This approach allowed us to ascertain whether differences in accuracy were statistically significant or occurred due to random chance. The evaluation process leveraged the UK AISI Inspect framework (AI Safety Institute, 2024), which provided a standardized structure for implementing and automating our assessment.

5 Experiment

5.1 Model choice

We employed a combination of six (6) closed and open large LLMs to evaluate their performance across English, Latvian, and Giriama translations.[1] The closed models selected for this study include

[1]Specifically: claude-3-5-sonnet-20241022, gemini-1.5-pro-002, gpt-4o-2024-08-06, Meta-Llama-3.1-405B-Instruct-Turbo, mistral-large-2407, and o1-preview-2024-09-12.

o1-preview, GPT-4o, and versions of Claude and Gemini, all of which are proprietary models known for their SOTA performance and extensive use in commercial applications. These models were chosen due to their established capabilities in handling a wide range of tasks, particularly in high-resource languages like English.

In contrast, open models such as Llama and Mistral were also included in the evaluation due to the transparency regarding their underlying architecture and training data, hence valuable for our use case. We aim to provide a comprehensive comparison of their effectiveness in low-resource languages, while also exploring the potential trade-offs between proprietary solutions and more customizable, open models.

6 Results and discussions

6.1 Model performance on languages

Table 2 presents the performance results of six LLMs across four languages—English, Latvian, machine-translated Latvian (denoted as Latvian (AT)), and Giriama. The results reflect varying degrees of proficiency across these languages, with a notable performance disparity between high-resource (English) and low-resource (Latvian and Giriama) languages.

The **o1-preview** model demonstrated superior performance across all three languages, achieving an accuracy of 87.5% in English, 84.8% in Latvian, and 82.1% in machine-translated Latvian. While the model's performance declined in Giriama, it still led the other models with an accuracy of 64.3%, suggesting relative robustness in handling lower-resource languages. The relatively small performance gap between English and Latvian shows the model's effectiveness in transferring knowledge to a non-English, medium-resource language.

Mistral showed the weakest performance across all languages, with English accuracy at 76.8% and a sharp decline in Latvian (57.1%**) and Giriama (34.8%**). This underscores the challenges of Mistral model in processing low-resource languages and its inability to maintain consistent accuracy across diverse linguistic contexts.

o1-preview model demonstrates the highest performance in Giriama, though the differences between o1, GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro are statistically insignificant. In contrast, Llama 3.1 405B and Mistral Large 2 show notably lower performance, struggling to handle Giriama and Latvian

6.2 Cross-language performance gaps

The performance disparities observed between English (Figure 2), Giriama (Figure 3) and Latvian (Figure 4) underscore the challenges faced by current LLMs in processing both medium- and low-resource languages.

The average gap between English and Latvian performance across all models is 9.3%, which is comparable to approximately two-thirds of the performance difference between GPT-3.5 and GPT-4 in English (OpenAI et al., 2024). However, this gap narrows for higher-performing models like **o1-preview**, where the difference becomes less pronounced. Large differences in this gap are primarily observed in the performance of **Mistral**.

In contrast, Giriama—a low-resource Bantu language—exhibits a much more pronounced performance gap, with average model accuracy dropping sharply to (47.6%), underscoring the limitations of cross-lingual transfer learning in handling languages with limited digital resources and complex linguistic structures.

The results reveal a consistent performance gap between more resourced languages and less resourced languages. On average, the models perform best in English (83.6%), followed by Latvian (74.3%) and machine-translated Latvian (71.3%), with the lowest performance observed in Giriama (47.6%).

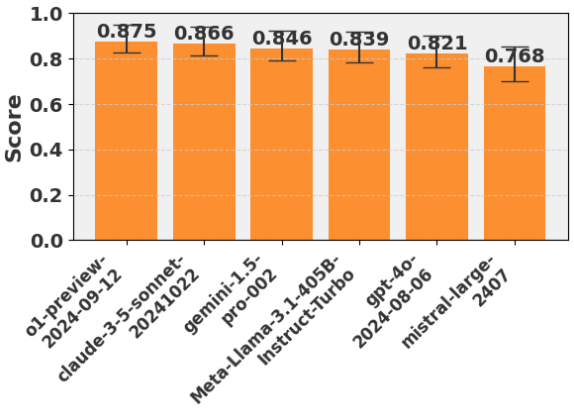

Figure 2: **Model performance in English.** Error bars represent 95% Wilson confidence intervals.

6.3 Impact of human-edited vs. machine-translated Data

For Latvian translations, human editing provided a modest improvement over machine translation, with accuracy increasing by 3.0% on average (see

Model	English	Latvian	Latvian (AT)	Giriama
o1-preview-2024-09-12	**0.875**	**0.848**	**0.821**	**0.643**
claude-3-5-sonnet-20241022	0.866	0.804	0.777	0.482*
gemini-1.5-pro-002	0.846	0.786	0.732	0.509*
Meta-Llama-3.1-405B-Instruct-Turbo	0.839	0.688**	0.643**	0.411***
gpt-4o-2024-08-06	0.821	0.759	0.723	0.464**
mistral-large-2407	0.768*	0.571***	0.580***	0.348***
AVG	**0.836**	**0.743**	**0.713**	**0.476**

Table 2: **Model performance across languages. AT: autotranslated. Each model: n=112; AVG: n=672. Boldface indicates the highest score in each column. Asterisks indicate statistically significant differences from the highest-scoring model within each language variant (*: p<0.05, **: p<0.01, ***: p<0.001), computed using two-proportion z-test.**

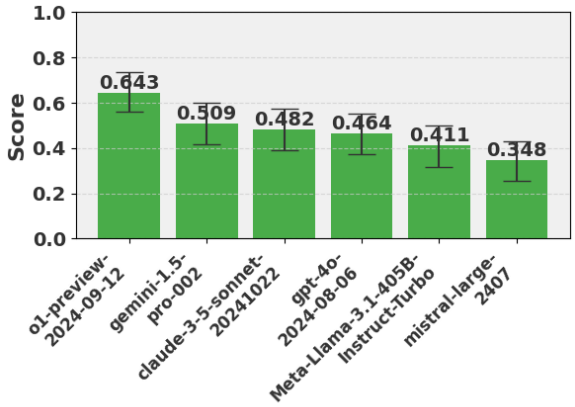

Figure 3: **Model performance in Giriama.** Error bars represent 95% Wilson confidence intervals.

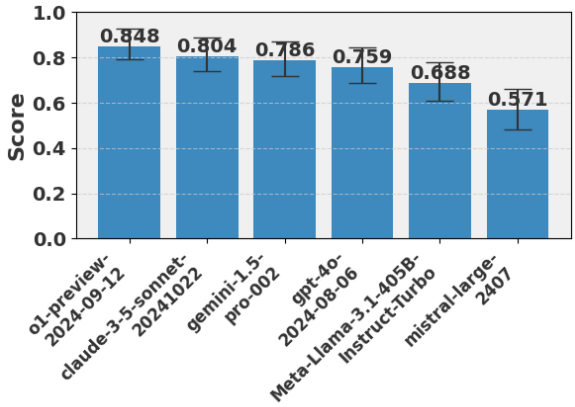

Figure 4: **Model performance in Latvian.** Error bars represent 95% Wilson confidence intervals.

Table 2). While this difference is not statistically significant, it suggests that human involvement remains valuable for languages with complex morphology and syntax. However, we note that our baseline used a free translation service - SOTA machine translation might further narrow this gap.

Giriama presented more significant challenges. The language was consistently misidentified as Swahili by translation systems, making automatic translation infeasible. This technical limitation, combined with uniformly poor model performance across all tested models, emphasizes the need for increased linguistic resources and human expertise when working with low-resource African languages.

6.4 Impact of the temperature setting

As noted previously, we used a temperature setting of 0.5 for all models except for o1-preview, which only permits temperature=1. To assess whether this non-uniform temperature setting significantly affected our results, we conducted additional tests on the edited Latvian translations. We ran the five other models at temperature=1 and compared their performance against the temperature=0.5 runs. The results were similar across models, with only Llama 3.1 405B showing slight improvement (+0.9%). On average, performance marginally declined (-0.4%), but none of the differences were statistically significant. We conclude that the non-uniform temperature settings did not materially impact our findings.

6.5 Implications for multilingual LLM development

The substantial performance drop from English to Giriama across all models reflects the broader challenges in scaling LLMs for low-resource languages. While advancements in multilingual modeling have closed some gaps for medium-resource languages like Latvian (Darġis et al., 2024), this study highlights the considerable distance yet to be covered in adequately supporting low-resource languages, particularly African languages like Giriama. These

results underscore the importance of developing more inclusive benchmarks and expanding the availability of high-quality training data to ensure that LLMs are more equitable across diverse linguistic contexts.

6.6 Bias and considerations for future Research

Anecdotal evidence showed that some of the tested models were much better at translating questions and answers than the free translation service. Future research could make use of the LLM translation capabilities. However, it is important not to bias the results in favor of one model or another: it is not inconceivable that a given model finds its own translations easier to interpret than those of other models (which is another hypothesis to explore). Alternatively, it is possible to use other translation services and human translation services together or separately.

These, as other benchmark results, may be subject to bias due to potential data contamination. (Bean et al., 2024). The English MMLU dataset is more likely to have been included in or influenced the models' training data. This could lead to an overestimation of the performance gap between languages, as models might have prior exposure to the English questions.

Cultural context introduces another potential source of bias and reduced relevance in this study. For example, Professional Law questions are based on the U.S. legal system, not Kenyan or Latvian law, which may lead to less accurate responses when questions are presented in Giriama or Latvian. This mismatch between the source material's cultural context and the target languages could affect model performance independently of linguistic factors. Future research could assess the impact of cultural context by using a larger sample size and analyzing model performance in culturally sensitive subcategories like Professional Law. However, U.S.-centric legal questions are inherently limited in evaluating legal expertise within other contexts. Adapting such questions to local contexts is crucial but may require costly specialist knowledge.

Expanding the sample size in future studies could yield more robust results. The scope of this investigation was primarily constrained by two factors: the human resources required for editing translations and the available resources for model API access.

7 Conclusion

Our evaluation of six frontier LLMs across English, Latvian, and Giriama reveals several critical insights about the current state of multilingual AI capabilities:

1. **Model-specific language gaps:** While all models showed performance degradation in non-English languages, proprietary models (particularly o1-preview with only a 2.7% English-Latvian gap) maintained relatively consistent performance compared to open-source alternatives (up to 19.7% gap). This suggests that recent advances in commercial AI systems are beginning to address historical English-centric bias, though significant gaps remain in open-source alternatives.
2. **Translation quality impact:** For Latvian, human editing of machine translations improved accuracy by only 3.0% on average, indicating that automated translations may be sufficient for benchmark creation in languages with established digital infrastructure. This finding could significantly reduce the cost and effort of developing multilingual evaluations.
3. **Low-resource language challenges:** The dramatic performance drop in Giriama (average accuracy 47.6% vs 83.6% in English) reveals fundamental limitations in current approaches to low-resource language support. The failure of machine translation for Giriama highlights how technological gaps compound the challenges of language accessibility.

These findings have immediate implications for both research and deployment. For research, they highlight the viability of using machine translation for creating benchmarks in medium-resource languages and the need for better methods to support low-resource languages. For deployment, our results suggest that while LLMs are becoming viable for medium-resource languages like Latvian, significant work remains before they can reliably serve low-resource language communities.

Future work should prioritize two key areas: (1) developing more efficient methods for extending LLM capabilities to low-resource languages without requiring extensive compute or data resources, and (2) creating evaluation frameworks that explicitly measure both linguistic accuracy and cultural appropriateness. The substantial gap in low-resource language performance emphasizes that achieving truly equitable AI requires not just tech-

nical advancement, but sustained investment in linguistic resources and community engagement.

8 Limitations

Our work presents several limitations that should be acknowledged. First, no formal quality control measures, such as inter-annotator agreement (IAA) or Cohen's Kappa, were employed to assess the consistency and reliability of the translations in our dataset. This could affect the overall validity of the translation quality.

The dataset size is relatively small, consisting of only 500 questions per language. While this dataset provides preliminary insights, the dataset size limits the generalizability of the results, and larger datasets would be necessary to draw more robust conclusions.

This study's scope was limited to six language models and two non-English languages due to API access costs. A more comprehensive evaluation would require greater financial resources to test additional models and languages.

Finally, Giriama, as a low-resource language, faces unique challenges due to limited linguistic resources, which may lead to oversimplified translations and insufficient validation, affecting the dataset's quality. Unlike Latvian, which has more established digital resources, Giriama may lack the tools for thorough quality control, increasing the risk of inaccuracies.

9 Ethical considerations

Native speakers translated the MMLU dataset into Giriama and Latvian to ensure linguistic and cultural accuracy. However, several potential ethical concerns arise in this process:

- **Cultural Relevance and Sensitivity**: While linguistic fidelity was prioritized, the dataset contains many questions grounded in Western, specifically American, cultural contexts such as historical references to Columbus or moral standards in the US. When translating such questions into Latvian or Giriama, there is a risk of imposing culturally foreign concepts onto the target audience, potentially alienating speakers or distorting meaning. For instance, some questions may have no direct equivalent in Giriama or Latvian law and moral philosophy. This can lead to mistranslation or misunderstanding, as the target audience may not relate to or fully grasp the original cultural context.
- **Linguistic Complexity and Vocabulary Gaps**: Many questions in the dataset involve highly technical and specialized terminology from subjects such as law, science, and ethics (such as "neurotransmitters," "Pauli exclusion principle"). Low-resource languages like Giriama may not have established vocabulary for such specialized terms, resulting in challenges for accurate translation. Translators must decide whether to borrow terms from English or create new ones, both of which have ethical implications that could undermine linguistic purity or lead to confusion or lack of consistency in the target language **Cultural Bias in Translation**: The MMLU dataset reflects Western-centric knowledge and perspectives, which pose ethical challenges when translating into low-resource languages like Giriama or Latvian. Without careful adaptation, cultural differences in political ideologies, social norms, or gender roles may be misrepresented, leading to misunderstandings. These biases can hinder the performance of language models by failing to accurately capture the nuances of the target cultures, reducing their effectiveness in real-world applications.

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*.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 african languages. *arXiv preprint arXiv:2401.01053*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

AI Safety Institute. 2024. Inspect - ai safety institute. `https://inspect.`

ai-safety-institute.org.uk/. Accessed: 2024-10-15.

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. *arXiv preprint arXiv:2204.06487*.

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–32.

Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low-Resource and Extinct Languages. _eprint: 2406.06196.

Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14:34.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Rochelle Choenni, Sara Rajaee, Christof Monz, and Ekaterina Shutova. 2024. On the evaluation practices in multilingual nlp: Can machine translation offer an alternative to human translations? *arXiv preprint arXiv:2406.14267*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Cohere For AI team. 2024. Policy Primer - The AI Language Gap.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Roberts Darģis, Arturs Znotins, Ilze Auziņa, Baiba Saulīte, Sanita Reinsone, Raivis Dejus, Antra Klavinska, and Normunds Gruzitis. 2024. Balsutalka. lv-boosting the common voice corpus for low-resource languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2080–2085.

Fiifi Dawson, Zainab Mosunmola, Sahil Pocker, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2024. Evaluating cultural awareness of llms for yoruba, malayalam, and english. *arXiv preprint arXiv:2410.01811*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *arXiv preprint arXiv:2403.11009*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, et al. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv e-prints*, pages arXiv–2406.

Kathryn R Kirby, Russell D Gray, Simon J Greenhill, Fiona M Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E Blasi, Carlos A Botero, Claire Bowern, Carol R Ember, et al. 2016. D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, 11(7):e0158391.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–34.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Zihan Liu, Genta I Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the importance of word order information in cross-lingual sequence labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13461–13469.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Marc Miquel-Ribé and David Laniado. 2019. Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 620–629.

Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Pelloni, and Tanja Samardzic. 2022. Teddi sample: Text data diversity sample for language comparison and multilingual nlp. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1150–1158.

MyMemory. 2024. Mymemory translation memory - api documentation.

https://mymemory.translated.net/doc/spec.php. Accessed: 2024-10-15.

OpenAI. 2024. O1 system card. Technical report, OpenAI. Accessed on October 16, 2024.

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, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report.

_eprint: 2303.08774.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Surangika Ranathunga and Nisansa De Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the nlp world. *arXiv preprint arXiv:2210.08523*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preoţiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *arXiv preprint arXiv:2103.13309*.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, et al. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *arXiv preprint arXiv:2410.12705*.

L Xue. 2020. mt5: A massively multilingual pretrained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.