Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation

Shivalika Singh[1], Angelika Romanou[2], Clémentine Fourrier[3], David I. Adelani[4], Jian Gang Ngui[5,6], Daniel Vila-Suero[3], Peerat Limkonchotiwat[5,6], Kelly Marchisio [7], Wei Qi Leong[5,6], Yosephine Susanto[5,6], Raymond Ng[5,6], Shayne Longpre[8], Sebastian Ruder[15], Wei-Yin Ko[7], Antoine Bosselut[2], Alice Oh[9], André F. T. Martins[10,11], Leshem Choshen[12], Daphne Ippolito[13], Enzo Ferrante[14], Marzieh Fadaee[1], Beyza Ermis[1], Sara Hooker[1]

[1]Cohere Labs, [2]EPFL, [3]Hugging Face, [4]Mila, McGill University & Canada CIFAR AI Chair, [5]AI Singapore, [6]National University of Singapore, [7]Cohere, [8]MIT, [9]KAIST, [10]Instituto de Telecomunicações, [11]Instituto Superior Técnico, Universidade de Lisboa, [12]MIT, MIT-IBM Watson AI Lab, [13]Carnegie Mellon University, [14]CONICET & Universidad de Buenos Aires, [15]Meta AI Research,

Correspondence: shivalikasingh, beyza, sarahooker@cohere.com

Abstract

Reliable multilingual evaluation is difficult, and culturally appropriate evaluation is even harder to achieve. A common practice to fill this gap is to machine-translate English evaluation sets, but this introduces language bias and carries over cultural assumptions, often testing knowledge irrelevant to the target audience. In this work, we highlight the extent and impact of these biases and present a multilingual evaluation framework that aims to mitigate them through improved translation and annotation practices. Through a large-scale study involving professional and community translators and annotators, we show that state-of-the-art models excel primarily by learning Western-centric concepts. Notably, we find that model rankings on full MMLU change when evaluated on a subset of questions marked as culturally sensitive. We release **Global-MMLU**, a multilingual extension of MMLU across 42 languages, with improved translation quality, expanded language coverage, and designated subsets labeled as **culturally sensitive** and **culturally agnostic** to enable a more comprehensive and equitable benchmark for evaluating language models across diverse linguistic and cultural contexts.

Global-MMLU: https://hf.co/datasets/CohereForAI/Global-MMLU

Global-MMLU Lite: https://huggingface.co/datasets/CohereForAI/Global-MMLU-Lite

1 Introduction

Despite the global reach of state-of-the-art generative AI, most evaluations still rely on English benchmarks (Zellers et al., 2019; Hendrycks et al., 2020; Suzgun et al., 2022; Zhang et al., 2023b), reflecting a primarily Western cultural perspective. This raises a pressing question: *How can we develop large language models (LLMs) that effectively serve the full spectrum of languages and cultures?*

Many widely used multilingual benchmarks rely on translations of English datasets, such as the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2020). Originally composed of English-language questions across 57 subject areas, MMLU is frequently machine-translated for multilingual assessment, forming what we term *transMMLU* (Lai et al., 2023; Üstün et al., 2024; OpenAI, 2024; Dubey et al., 2024; Bendale et al., 2024). However, such translations do not ensure cultural inclusivity and risk overrepresenting Western-centric knowledge. Many MMLU subsets, such as *US History* and *US Law*, reflect American perspectives, which may skew multilingual evaluation results. Optimizing AI models for these datasets risks reinforcing cultural biases. Additionally, machine translation introduces artifacts known as *translationese* (Bizzoni et al., 2020; Vanmassenhove et al., 2021; Koppel and Ordan, 2011), further compromising evaluation quality (Luccioni and Viviano, 2021; Kreutzer et al., 2022; Ferrara, 2023).

Our core contributions to address the above are:

Cultural Bias Analysis: We assess the cultural biases in MMLU, finding that 28% of sampled questions require Western-centric knowledge, with 84.9% of geographic references focusing on North America or Europe.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18761–18799
July 27 - August 1, 2025

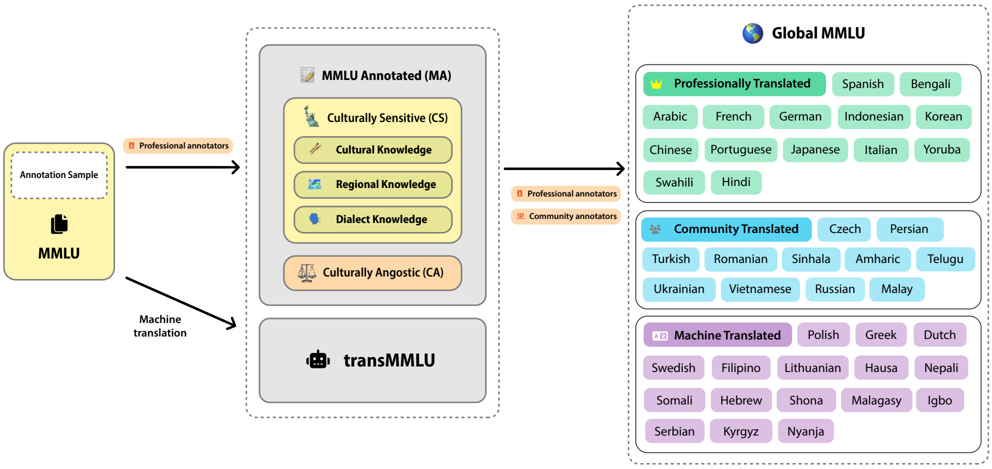

Figure 1: Overview of **Global-MMLU** preparation, incorporating professional and community annotators to refine translations and to provide rich meta-data for what questions in MMLU require *Culturally-Sensitive* (**CS**) knowledge such as 1) **Cultural**, 2) **Geographical** or 3) **Dialect** Knowledge to answer correctly.

Introducing Global-MMLU: We release a new multilingual MMLU test set spanning 42 languages, including English. **Global-MMLU**expands the original MMLU samples through a combination of professional translations with post-edits (14 languages), crowd-sourced translations (11 languages), and machine translations (16 languages). Our **cultural bias study**, offers two splits for evaluations, the **Culturally-Sensitive** (**CS**) and **Culturally-Agnostic** (**CA**) subsets. While **Global-MMLU**mitigates certain biases through translation improvements and cultural annotations that help differentiate model behavior on culturally sensitive versus culturally agnostic questions, it retains the original English-centric samples and does not introduce new culturally-specific questions.

Re-evaluation of state-of-the-art models: We evaluate the impact of cultural biases on multilingual models. Among 14 tested models, rankings on **CA** datasets shifted by an average of 3.7 positions compared to their performance on a uniform subsample of the MMLU dataset (*MMLU Annotated*). In contrast, **CS** datasets showed greater variability, with an average shift of 7.3 positions across languages.

Role of data quality improvements: Our analysis highlights notable performance differences between human-translated and machine-translated datasets for both high-resource and low-resource languages. Human-translated datasets are essential for accurately assessing model performance, especially on low-resource languages.

To improve multilingual evaluations, we recommend: (1) **Prioritizing Global-MMLU over direct MMLU translations**: **Global-MMLU** provides a more accurate and culturally inclusive benchmark. (2) **Separately reporting CA and CS performance**: Given significant ranking variations across subsets, evaluating them independently enhances transparency in multilingual model evaluations.

2 Evaluating Cultural Bias in MMLU

2.1 Data Annotation Process

We annotated a subset of the MMLU dataset to identify unintended cultural, regional, and linguistic sensitivities, referring to this annotated subset as *MMLU Annotated* (MA). In total, 200 professional and community annotators reviewed 2,850 samples from the original English MMLU dataset, comprising 50 uniformly randomly selected questions from each of the 57 exam subjects. Annotators were asked whether correctly answering each question required any of the following: (1) cultural knowledge, (2) geographic knowledge, or (3) dialect knowledge (detailed in Appendix B).

To understand the prevalence of these attributes, we labeled questions as **Culturally-Sensitive** (CS) if they required any form of cultural, geographic, or dialect knowledge. Otherwise, they were classified as **Culturally-Agnostic** (CA). This enables us to track the proportion of the dataset that requires **CS** knowledge at an aggregate level. Further details on the annotation process are provided in Appendix I.

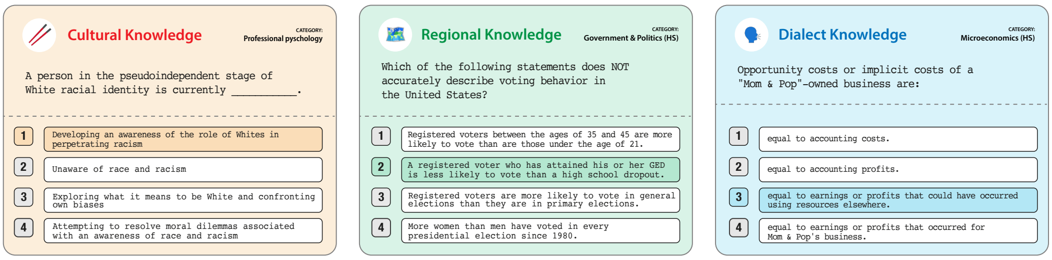

Figure 2: Examples of MMLU questions requiring cultural, regional, or dialectal knowledge.

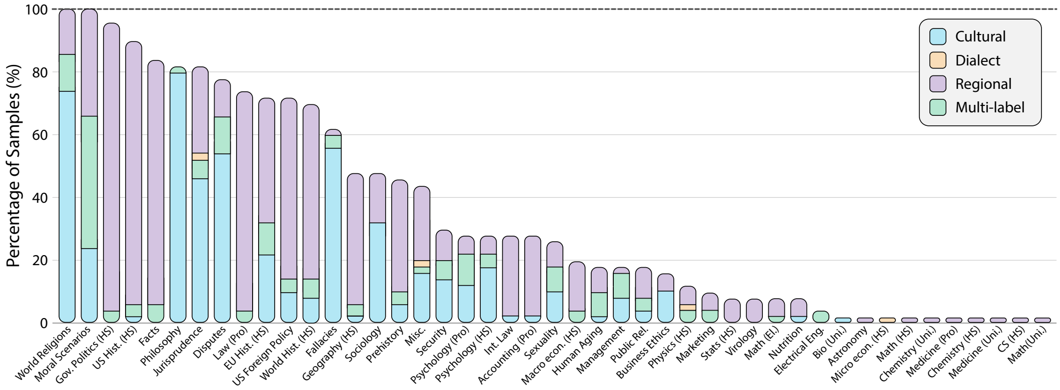

Figure 3: Proportion of samples containing cultural, regional, or dialect-specific references per subject in the MMLU dataset. All samples in *World Religions* and *Moral Scenarios* include at least one such reference. (12 subjects with no culturally sensitive samples are excluded.)

2.2 Analysis of MMLU Cultural Biases

Figure 3 summarizes the results of this extensive annotation process. Our analysis reveals that 28% of MMLU requires **CS** knowledge – defined as requiring either geographic, cultural or dialect knowledge – to be answered correctly. Geographic knowledge was the most frequently tagged bias, at 54.7%, followed by cultural (32.7%) and dialect (0.5%). 10.6% needed both cultural and geographic knowledge, and 1.5% required all three.

Western-centric culture dominates. Among the samples identified as requiring **CS**, a significant 86.5% were tagged as specific to *Western* cultural knowledge. A similar trend is observed for geographic knowledge: 64.5% of **CS** samples were tagged as needing regional knowledge of *North America*, followed by 20.4% tagged as requiring knowledge of *Europe*. This concentration indicates that progress on MMLU predominantly reflects knowledge of Western-centric cultural and regional knowledge.

Culture-specific knowledge is overrepresented for certain countries. Figure 4 shows the distribution of cultural and regional tags across countries in the **CS** dataset. Our analysis reveals that 73.9% of Western culture-related questions require knowledge of the U.S., followed by the U.K. at 8%. Similarly, 59% of Asian culture tags are tied to India, while China and Japan account for 17.9% each. Despite this, Asian cultures remain underrepresented, with only 4.0% of questions covering South Asia and 3.1% addressing East Asia. Middle Eastern culture is also underrepresented, accounting for just 2.7%. These findings underscore the dataset's heavy bias towards the U.S. For a deeper analysis of culture-region relationships and country-level breakdowns, see Appendix H.

Cultural sensitivity varies considerably across subjects. The MMLU dataset includes 57 subjects spanning four categories: *STEM, Humanities, Social Sciences*, and *Other*. We further categorized relevant *Other* subjects into *Medical* and *Business*. Figure 5 shows the distribution of the **CA** subset, revealing significant variation in cultural and regional references across subjects. *Humanities* and *Social Sciences* frequently require cultural or regional knowledge, with 68% of *Humanities* questions labeled **CS**. Some subjects, like Philosophy, Moral Scenarios, High School US History, and High School Government and Politics, exceed 80% **CS**. In contrast, *STEM* subjects showed minimal cultural bias, with only 30

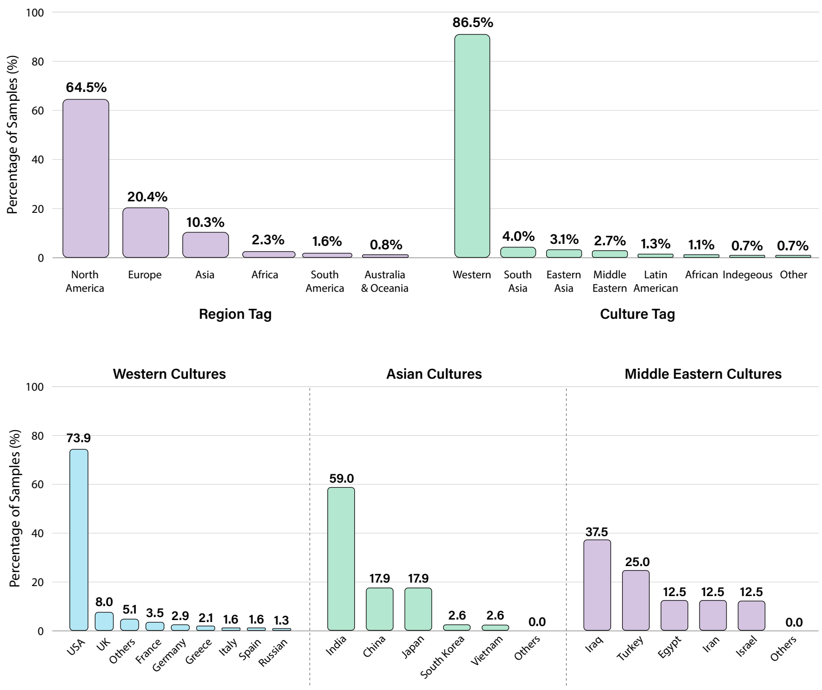

Figure 4: (Top) Region and culture distribution in the **CS** dataset, with most Region tags (64.5%) linked to North America and Culture tags (86.5%) classified as Western. (Bottom) Cultural and regional tag distribution across countries, showing each country's dataset representation. Samples without tags are excluded.

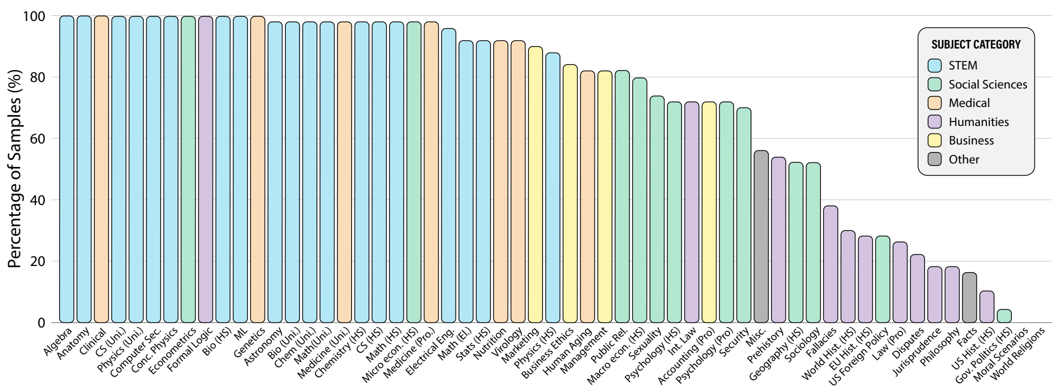

Figure 5: Proportion of samples retained per subject, after excluding those requiring cultural, geographic and dialectic knowledge (selected based on majority agreement).

of 950 samples (3.15%) classified as **CS**. Some subjects, such as Clinical Knowledge, Computer Security, and Econometrics, contained no **CS** questions. As shown in Figure 5, certain subjects inherently reflect more cultural and regional biases. Examples of **CS** and **CA** questions are provided in Appendix M.

Characteristics of CS and CA subsets. Our annotation process resulted in two aggregated annotated subsets of MMLU: **CS**, containing questions requiring dialect, cultural, or geographic knowledge, and **CA**, comprising those without such dependencies. Table 2 in Appendix F details subject and sample distributions.

Significant differences emerge in subject representation. *Social Sciences* account for 21.1% of the *MMLU Annotated*, but are over-represented in **CS** at 26.3%. Conversely, STEM, which makes up 33.3% of MMLU Annotated, is under-represented in **CS**, contributing just 2.9%. These shifts reflect how the nature of the **CS** subset emphasizes cultural and contextual knowledge over technical or scientific content. Overall, STEM, Medical, and Business categories are more prevalent in **CA** due to their globally relevant content, whereas Humanities and Social Sciences dominate **CS** due to frequent cultural or regional references. These trends are critical to model evaluations (Section 4), demonstrating how cultural biases in MMLU shape dataset composition and influence model performance.

3 Introducing Global-MMLU

Many multilingual evaluations rely on translated MMLU, with the most widely used dataset translated into 26 languages using ChatGPT (GPT-3.5) (Lai et al., 2023). We introduce **Global-MMLU**, an improved benchmark with higher-quality translations and dedicated **CS** and **CA** subsets for deeper analysis.

To enhance translation quality, we incorporate professional annotator edits and native speaker translations for 25 languages, expanding total coverage to 42 languages by including higher-quality machine translation. We incorporated professional human translations from the MMMLU dataset[1] for 14 languages. We prioritize human-verified translations to ensure reliability and reduce biases, particularly those introduced by *translationese*, which can be more pronounced in machine translation (Bizzoni et al., 2020; Vanmassenhove et al., 2021; Koppel and Ordan, 2011). Alongside these improvements, we provide **CS** and **CA** metadata to enable comprehensive subset analysis. Below, we detail our approach to improving MMLU quality, compensating human annotators for translation verification, and identifying **CS** and **CA** subsets.

3.1 Translation Process

We first translated the English MMLU dataset into 41 languages using the Google Translate API.[2] Despite its cost, we selected Google Translate due to its superior performance, as demonstrated in comprehensive evaluations spanning 102 languages (Zhu et al., 2024). It significantly outperforms alternatives like NLLB (NLLB-Team et al., 2022), GPT-4, and ChatGPT for low-resource languages (Robinson et al., 2023). While LLMs are improving in high-resource translations (Kocmi et al., 2024), they tend to favor their own generations (Panickssery et al., 2024; Shimabucoro et al., 2024). To avoid bias, we used Google Translate uniformly across all languages. A comparison with GPT-3.5-turbo (previously used for MMLU translations (Lai et al., 2023)) confirmed this choice, as Google Translate achieved higher ChrF++ scores (Popović, 2017) with lower variance across languages (see Figure 20 in Appendix J.1). After translation, native speakers reviewed and refined the outputs for accuracy and fluency. Edits were performed by *professional annotators* and *native community annotators* (details in Appendix J.2).

MMMLU Translations. As detailed in the OpenAI-o1 system card,[3] MMMLU is a professionally human-translated dataset available in 14 languages. To maximize human-translated content in **Global-MMLU**, we incorporated this dataset wherever applicable. Since MMMLU overlaps with our *Gold Set* (edited by professional annotators), we incorporated the remaining 10 languages: *Bengali, Chinese, German, Indonesian, Italian, Japanese, Korean, Portuguese, Swahili, Yoruba* – alongside *Arabic, French, Hindi, Spanish* from the *Gold Set*. Figure 19 in Appendix J illustrates edits by professionals and community contributors. Professionals edited 789 samples per language (38.5% of the Gold Set), while community members edited 362 (17.7%). With 7,565 edits in total, 36.9% of samples were reviewed. Differences in edit rates likely reflect variations in available time and resources rather than differences in translation quality across languages. Appendix J provides further analysis on translation quality and other factors.

3.2 Data Composition of Global-MMLU

Global-MMLU is our comprehensive test set including MMLU's 14K samples in 42 languages, totaling 589,764 samples. It covers human-translated, machine-translated, and original English MMLU samples. Throughout the Model Evaluations section, we report on different subsets of **Global-MMLU**, such as MMLU Annotated, Culturally-Sensitive (CS) and Culturally-Agnostic (CA) subsets. A detailed breakdown of these subsets is provided in Appendix C.

Global-MMLU Lite is a "lite" version of **Global-MMLU** covering 15 languages which are fully human translated or post-edited, along with English. It includes 200 CS and 200 CA samples per language, totaling 6,000 samples. Further details on its preparation are in Appendix E.

[1] https://huggingface.co/datasets/openai/MMMLU

[2] https://cloud.google.com/translate

[3] https://openai.com/index/openai-o1-system-card/

4 Model Evaluations

Section 2.2 highlights MMLU's strong bias toward **CS** knowledge. Here, we assess how these biases impact the evaluation of both open-weight and closed models. To do so, we analyze changes in model rankings across three subsets: ***Global-MMLU Annotated***, ***Global-MMLU Culturally-Agnostic (CA)*** and ***Global-MMLU Culturally-Sensitive (CS)***. By comparing model performance across these subsets, we aim to answer: (1) *How do models perform when culturally-sensitive samples are included?* (2) *How do models perform on culturally-agnostic samples, ensuring consistent evaluation across languages and regions?*

Experimental Setup. We evaluated 14 recent state-of-the-art language models from 9 model families, focusing on those known for their high multilingual performance. These include **small models** like Aya Expanse 8B, Gemma2 9B, SEA-LION v3 (9B), Llama 3.1 8B, Mistral Nemo 12B, and Qwen 2.5 7B; **mid-size models**, comprising Aya Expanse 32B, CommandR (34B), Gemma2 27B, and Qwen 2.5 32B; **large models**, such as Llama 3.1 70B and CommandR+; and **closed-weight models**, specifically GPT-4o and Claude Sonnet 3.5. A more detailed description of the models covered is mentioned in the Appendix K.1

We categorize the languages into two main groups for reporting the results. The first group consists of *human-translated data only*, which includes 10 languages from OpenAI's human-translated MMLU test set and 4 *Gold Set* languages from our professionally translated set. The second group includes *all our data*, combining professional, community, and machine translations. Languages are categorized as ● high-, ◐ mid-, and ○ low-resource, following Joshi et al. (2019) and Singh et al. (2024). See Table 7 in Appendix L for details.

4.1 Evaluations on Human-Translated Data

We evaluate model performance on high-quality, human-translated data, focusing on **CA** and **CS** subsets to analyze how models handle tasks with and without cultural context. Figure 6 presents results across 14 languages.

We note that the focus of this evaluation is not to compare model performances directly but to analyze their behavior on **CA** and **CS** datasets. Direct comparisons between proprietary models and open-weight models are not feasible due to significant differences in model sizes (the parameter sizes of proprietary models have not been officially disclosed) and different evaluation methods. However, the results show that proprietary models consistently outperform smaller open-source models. Interestingly, the performance gap between these models is narrower on **CS** datasets.

Additionally, we assess mid-size and large open-weight models on **Global-MMLU Lite**, a fully human-translated (or post-edited) subset evenly balanced between **CS** and **CA** samples. Unlike the full **Global-MMLU**, this balance enables clearer comparisons. Figure 7 shows that overall, models perform better on the **CA** portion.

Performance on CS is higher but more variable. On average, models achieve higher accuracy on **CS** datasets than **CA**, likely because **CS** samples are drawn primarily from Social Sciences and Humanities, where models perform well. In contrast, **CA** datasets contain more challenging categories, such as Medical and STEM (see Figure 23 in Appendix K.3.1).

However, performance on **CS** data exhibits greater variance across languages due to several factors. Culturally sensitive tasks demand deeper contextual understanding, making them more susceptible to translation quality variations. Additionally, nuanced cultural, regional, or dialectal references amplify sensitivity, as differing translations can impact performance. Many models are also trained primarily on high-resource or Western centric data, introducing biases that cause inconsistencies in less-represented contexts. On **Global-MMLU Lite**, the pattern shifts: **CS** tasks have lower average accuracies and greater variance than **CA** tasks. This highlights how cultural specificity increases performance instability, when the **CS** and **CA** samples are balanced.

4.2 Evaluations by Language Resource Availability

We analyzed model performance on **CA** and **CS** subsets across ● high-, ◐ mid-, and ○ low-resource languages (see Figure 25 in Appendix K.3. This evaluation provides insights into how models handle linguistic diversity and cultural nuances across different resource levels.

For both **CA** and **CS** datasets, ● high-resource languages consistently achieve the highest accuracy. As expected, performance declines significantly for ○ low-resource languages due to limited high-quality training data, which also in-

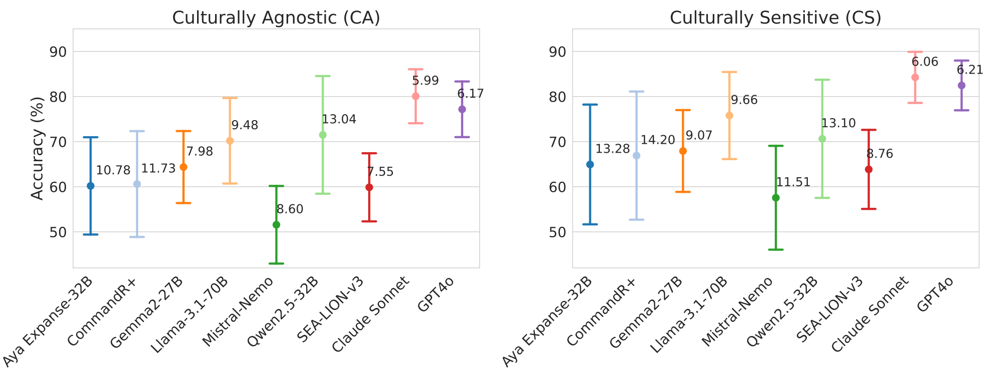

Figure 6: Model evaluations on **CA** and **CS** samples across **14 human-translated languages**. Error bars indicate standard deviation across languages.

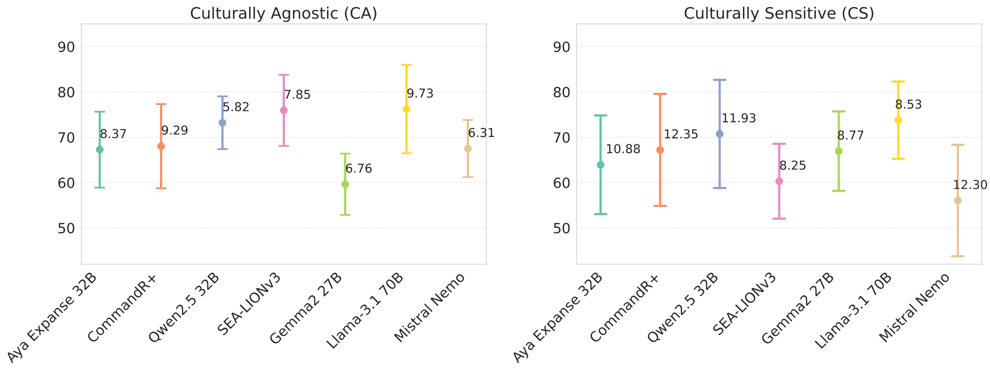

Figure 7: Model evaluations on **CA** and **CS** samples in **Global-MMLU Lite**. Error bars indicate standard deviation across languages.

creases performance variability. Standard deviation rises for mid-resource languages and even more so for low-resource languages, particularly on **CS** datasets.

The average standard deviation for high-resource languages is **3.21** on **CA** datasets and **3.86** on **CS**. For mid-resource languages, these values increase to **3.42** and **4.6**, respectively. Low-resource languages exhibit the highest variability, with averages of **6.37** on **CA** and **6.78** on **CS** – increases of 98% and 75% compared to high-resource languages. This underscores the increased sensitivity of low-resource settings, where a deeper understanding of regional and dialectal nuances is essential.

4.3 Model Rank Changes

We analyze how model rankings shift between **CA** and **CS** datasets relative to MA across all languages. Table 1 shows how model rankings shift for **human-translated** languages. Organized by resource level, it reveals the impact of dataset type, resources, and model size. For more details, including rankings for all languages, see Appendix K.3.3. The rank changes reveal three key findings:

1) Models perform differently on CA and CS datasets, with greater variation in CS. CA datasets show minimal ranking changes, with an average of 3.4 rank and 3.7 position changes. CS datasets, however, exhibit greater volatility, with an average of 5.7 rank and 7.3 position changes. Chinese, Hindi, French, German, Italian, Japanese, and Portuguese are particularly sensitive to CS knowledge. Notably, models from Aya Expanse and CommandR families tend to show positive trends on **CS** datasets, particularly for these languages.

2) Performance differences between CA and CS datasets are smaller in low-resource languages. High-resource languages demonstrate relatively stable rankings on **CA** datasets, with an average of *3.3 rank changes* and a maximum shift of *3 positions*. However, on **CS** datasets, these rise to *6.8 rank changes* and *9.1 position shifts*. Mid-resource languages show moderate variation, with rank changes averaging *3.7* on **CA** and *4.7* on **CS**, with corresponding position

Language	Dataset	Aya Exp. 8B	Aya Exp. 32B	CommandR	CommandR+	Gemma2 9B	Gemma2 27B	Llama-3.1 8B	Llama-3.1 70B	Mistral Nemo	Qwen2.5 7B	Qwen2.5 32B	SEA-LION-v3	GPT4o	Claude Sonnet
● Arabic	CA	-	-	-	-	-	-	-	-	-	↑1	-	↓1	-	-
	CS	-	↑1	-	-	-	↓1	-	↑1	-	-	↓1	-	-	-
● Chinese	CA	-	-	↓1	-	↑1	-	-	-	-	-	↑1	-	↓1	-
	CS	↑1	↑1	↑1	↑2	↑1	-	↓1	↑1	-	↓3	↓1	↓2	↑1	↓1
● English	CA	-	-	-	-	-	↓1	-	-	-	↑1	↑1	-	↓1	-
	CS	-	↑1	-	-	-	-	-	↑1	-	↓1	↓1	-	-	-
● French	CA	-	↑1	-	-	-	-	-	-	-	↓1	-	-	-	-
	CS	-	↑2	↑2	↑1	-	↓2	-	↑1	-	↓3	↓1	↑1	-	-
● German	CA	-	↓1	-	↓1	-	↑1	-	-	-	↑1	-	-	-	-
	CS	-	-	↓1	-	↑2	-	-	↑1	-	↓3	↓1	↑2	-	-
● Hindi	CA	-	↑1	↓2	↓1	↑1	-	-	-	-	-	-	↑1	-	-
	CS	↑1	↓1	↑1	↑2	-	↓1	↑1	-	↑1	↓3	↓1	-	↑1	↓1
● Italian	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	CS	-	-	↑1	↑1	-	↓1	-	↑1	-	↓2	↓1	↑1	-	-
● Japanese	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	CS	-	↑1	↑1	↑1	↑1	↓2	-	↑1	-	↓1	↓1	↓1	-	-
● Portuguese	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	CS	-	↑1	↑1	↑1	↑1	↓1	-	↑1	-	↓2	↓1	↓1	-	-
● Spanish	CA	-	↓1	-	↓1	-	↑1	-	-	-	↑1	-	-	-	-
	CS	-	-	↑1	-	↑2	-	-	↑1	-	↓3	↓1	-	-	-
◐ Bengali	CA	-	↑1	-	-	-	-	-	↓1	↓1	-	-	-	-	-
	CS	-	-	-	-	-	-	-	-	↑1	↓1	-	-	-	-
◐ Indonesian	CA	-	-	↓1	↓1	↓1	↑1	-	-	-	↑2	-	-	-	-
	CS	-	-	↑1	-	-	-	↓1	↑1	↑1	-	↓1	↓1	-	-
◐ Korean	CA	↓1	↓1	↓1	-	-	↑1	↑1	-	-	↑1	-	-	-	-
	CS	-	↑1	↑1	↓1	-	↓1	-	↑1	-	-	↓1	-	-	-
○ Sinhala	CA	-	↑1	-	-	-	-	↓3	-	-	↑2	-	-	-	-
	CS	-	↓1	↑1	↑1	-	-	-	-	-	↓1	-	-	-	-
○ Swahili	CA	-	↓1	-	-	-	-	↑1	-	-	-	-	-	-	-
	CS	-	-	↑1	-	-	-	↓1	-	-	-	-	-	↓1	↑1
○ Yoruba	CA	-	↑1	↓2	-	↓1	-	-	-	-	↑2	↑1	↓1	-	-
	CS	-	↓1	↑1	↑1	↑1	-	-	-	-	-	↓2	-	-	-

Table 1: Changes in model rankings on **CA** and **CS** datasets, based on MA, across **human-translated** languages, including English. Languages are categorized as ●high-, ◐mid-, and ○low-resource. Color-coded boxes indicate increases (↑) and decreases (↓) in rank.

changes of *4.7* and *4.9*. Among all groups, ◐ mid-resource languages show the smallest difference between **CA** and **CS** performance. ○ Low-resource languages see a larger gap between **CA** and **CS** datasets. Rank changes average *3.3* on **CA** and *3.7* on **CS**, with position changes rising to *5.7* on **CA** and *7.9* on **CS**. This group also sees the largest rank fluctuations. Table 3 highlights significant shifts, including *up to 5 positions* for Malagasy, and *13 ranking changes* for Ukrainian, underscoring how resource levels amplify variability, even in **CA** datasets.

3) Model size affects performance variability. We analyzed performance variations across three model groups, as defined in the *Model* section (excluding closed-weight models due to unknown sizes). *Large models* demonstrate higher consistency across datasets and resource levels, with *0.21* for **CA** and *0.67* for **CS** average rank changes. Their maximum *position shift is 3*, compared to *5* for *small-models*. *Mid-size models* show much bigger variability. Their average *rank changes are 0.33* for **CA** and *1.97* for **CS**, particularly in culture dependent **CS** tasks. *Small models* show minimal rank change differences between **CA** and **CS** (0.35 and 0.45, respectively), but perform worse on both datasets. Their average accuracy is 51.3% on **CA** and 54.8% on **CS**, while mid-size models achieve 59.1% and 61.7%, and large models perform at 61.6% and 66.8% on **CA** and **CS**, respectively. Model performance remains highly influenced by dataset characteristics, especially in **CS** tasks requiring cultural knowledge. A similar trend appears in **Global-MMLU Lite**, where despite being smaller and balanced, performance volatility is still higher on **CS** datasets, particularly for low-resource languages (see Table 4 in Appendix K.3). Additionally, we compare models on Human-Translated (HT) and Machine-Translated (MT) **CS** datasets, with results provided in Appendix K.3.2.

5 Conclusion

We evaluate cultural biases in MMLU and find that 28% of questions require culturally sensitive knowledge, with a strong Western bias – regional questions predominantly focus on North America and Europe. This bias persists in translated MMLU variants, limiting their effectiveness as global benchmarks. To address this, we introduce **Global-MMLU** and **Global-MMLU Lite**, multilingual multi-domain datasets that distinguish between culturally-sensitive (**CS**) and culturally-agnostic (**CA**) knowledge. By incorporating professional and crowd-sourced annotations, these subsets enable rigorous multilingual model evaluation. Our evaluation reveals that model rankings shift depending on whether evaluation focuses on **CS** or **CA** knowledge, highlighting that progress on translated MMLU is insufficient as an indicator of performance. We recommend evaluating multilingual LLMs on culturally-sensitive and agnostic subsets of **Global-MMLU** to comprehensively assess their capabilities.

6 Limitations

Uneven distribution of contributions Beyond the gold standard languages where we engaged with compensated annotators, community annotator participation was uneven across languages, potentially leading to skewed dataset distributions and limited annotator diversity in some languages.

Language and dialect coverage We focus on 42 languages for **Global-MMLU**. However, this is still only a tiny fraction of the world's linguistic diversity. Future work should improve and expand evaluations beyond the 42 languages and address how technology serves different dialects. Geo-cultural variation within a language often leads to new dialects or creoles (Zampieri et al., 2020; Wolfram, 1997), which are crucial in establishing and maintaining cultural identity (Falck et al., 2012).

Toxic or offensive speech Global-MMLU may contain some potentially harmful content, as our annotation interface didn't allow for flagging toxic or offensive speech. However, we believe the risk is low due to the dataset's focus on examination material.

Region Category Assignment: For annotating geographically sensitive questions, we initially classified regions into six regions (Africa, Asia, Europe, North America, Oceania, and South America)[4] but recommend adopting World Bank's more granular taxonomy, which includes Central America and Sub-Saharan Africa, for future annotations.[5]

Identifying cultural sensitivity does not guarantee cultural inclusion. While initiatives like **Global-MMLU** highlight cultural biases in datasets, they don't fully solve the problem. Future work must prioritize the integration of diverse culturally grounded knowledge to achieve true inclusivity and fairness in multilingual AI evaluation.

7 Ethics Statement

This work was carried out as an open science initiative by volunteer participants as well as with help of paid professional annotators. All datasets used in this work have permissive licensing. We publicly release the datasets under Apache 2.0 license.

8 Acknowledgments

We would like to thank members of the Cohere Labs community who championed this initiative and helped with annotations for the project. In particular, we recognize Ashay Srivastava, Aurélien-Morgan Claudon, Bevnm SaiAsrit, Danylo Boiko, Hanna Yukhymenko, Sai Vineetha Baddepudi Venkata Naga Sri, Sangyeon Kim, Tadesse Destaw Belay, Alperen Ünlü, Mohammed Hamdy, Muhammad Rafi Sudrajat, Olusanya Joy Naomi, Vu Trong Ki, Yiyang Nan, Abdelmoneim Shahd, Arwa ALaya, Bimasena Putra, Emad Alghamdi, Fabian Farestam, Mridul Sharma, Sayuru Bopitiya, Surya Abhinai who contributed a significant amount to each of their languages. A special thank you to Claire Cheng and Trisha Starostina for helping to coordinate the Cohere professional annotators who contributed to this project. We thank all these compensated experts who provided their language knowledge to comprehensively improve quality over our gold languages.

[4] https://www.pewresearch.org/global/2013/06/04/regional-categorization/

[5] https://ourworldindata.org/world-region-map-definitions

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Zhuang Yun Jian, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing K. Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo KABENAMUALU, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Bridget Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *ArXiv*, abs/2406.03368.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`.

Abhijit Bendale, Michael Sapienza, Steven Ripplinger, Simon Gibbs, Jaewon Lee, and Pranav Mistry. 2024. Sutra: Scalable multilingual language model architecture. *Preprint*, arXiv:2405.06694.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. ACM.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? *Preprint*, arXiv:2406.12822.

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? *Preprint*, arXiv:2406.14267.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. 2018. Deep learning for classical japanese literature. *Preprint*, arXiv:1812.01718.

Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and public participation in ai. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Xuan-Quy Dao, Ngoc-Bich Le, The-Duy Vo, Xuan-Dung Phan, Bac-Bien Ngo, Van-Tien Nguyen, Thi-My-Thanh Nguyen, and Hong-Phuoc Nguyen. 2023. Vnhsge: Vietnamese high school graduation examination dataset for large language models. *Preprint*, arXiv:2305.12199.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Oliver Falck, Stephan Heblich, Alfred Lameli, and Jens Südekum. 2012. Dialects, cultural identity, and economic exchange. *Journal of urban economics*, 72(2-3):225–239.

Allan M. Feldman. 1980. *Majority Voting*, pages 161–177. Springer US, Boston, MA.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2024. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *Preprint*, arXiv:2404.06644.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing Latin American Spanish for low-resource text-to-speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6504–6513, Marseille, France. European Language Resources Association.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

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*.

Carlos Daniel Hernandez Mena and Ivan Vladimir Meza Ruiz. 2022. Creating Mexican Spanish language resources through the social service program. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pages 20–24, Marseille, France. European Language Resources Association.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mika Hämäläinen. 2021. *Endangered Languages are not Low-Resourced!*, page 1–11. University of Helsinki.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Naomi Kipuri. 2009. Chapter ii: culture. In *UN, Department of Economic and Social Affairs, Division for Social Policy and Development, Secretariat of the Permanent Forum on Indigenous Issues (ed.), State of the world's indigenous peoples: ST/ESA/328, New York: United Nations publication*, pages 51–81.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Findings of the wmt24 general machine translation shared task: the llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu. *Preprint*, arXiv:2310.04928.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *Preprint*, arXiv:2402.12840.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

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, Singapore. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Huihan Li, Liwei Jiang, Jena D. Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *Preprint*, arXiv:2404.10199.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Foutse Yuehgoh, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Elizaveta Semenova, et al. 2024. You are what you eat? feeding foundation models a regionally diverse food dataset of world wide dishes. *arXiv preprint arXiv:2406.09496*.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *Preprint*, arXiv:2309.12342.

Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, William Chandra Tjhi, and Alham Fikri Aji. 2024. Kalahi: A handcrafted, grassroots cultural llm evaluation suite for filipino. *Preprint*, arXiv:2409.15380.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. *arXiv preprint arXiv:2406.11661*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

NLLB-Team, 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. *Preprint*, arXiv:2207.04672.

OpenAI. 2024. GPT-4 Technical Report.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. *Preprint*, arXiv:2404.13076.

Wannaphong Phatthiyaphaibun, Surapon Nonesung, Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Chompakorn Chaksangchaichot, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Wangchanlion and wangchanx mrc eval. *arXiv preprint arXiv:2403.16127*.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *Preprint*, arXiv:2312.13951.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A framework for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages. *Preprint*, arXiv:2309.07423.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

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. *38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks*.

Sheikh Shafayat, H Hasan, Minhajur Mahim, Rifki Putri, James Thorne, and Alice Oh. 2024. Benqa: A question answering benchmark for bengali and english. In *ACL Findings*.

Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. LLM see, LLM do: Leveraging active inheritance to target non-differentiable objectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9243–9267, Miami, Florida, USA. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024a. Kmmlu: Measuring massive multitask language understanding in korean. *Preprint*, arXiv:2402.11548.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2024b. Hae-rae bench: Evaluation of korean knowledge in language models. *Preprint*, arXiv:2309.02706.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Saeid Asgari Taghanaki, Aliasgahr Khani, and Amir Khasahmadi. 2024. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms. *Preprint*, arXiv:2409.02257.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Amirudin, Muhammad Ridzuan, Daniya Kareem, Ketan More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Baliah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Xavier, Amit Bhatkal, Hawau Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, and Fahad Khan. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2024. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *Preprint*, arXiv:2310.01929.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Walt Wolfram. 1997. Issues in dialect obsolescence: An introduction. *American speech*, 72(1):3–11.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea:

A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Şenel, Anna Korhonen, and Hinrich Schütze. 2024. Turkishmmlu: Measuring massive multitask language understanding in turkish. *Preprint*, arXiv:2407.12402.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Related Work

A.1 Multilingual Knowledge Evaluation

As MMLU becomes a standard for LLM evaluation (Beeching et al., 2023; OpenAI, 2024; Dubey et al., 2024; Üstün et al., 2024; Aryabumi et al., 2024), addressing its limitations and enhancing its scope is essential. For English, MMLU-redux re-annotates 3K questions across 30 subjects to refine quality (Gema et al., 2024), while MMLU-Pro expands question complexity and answer choices (Wang et al., 2024). MMLU-Pro+ further extends this by incorporating multiple correct answers and testing higher-order reasoning (Taghanaki et al., 2024). Despite these advancements, all remain English-only.

Language-specific MMLU variants typically focus on a single language, including ArabicMMLU (Koto et al., 2024), CMMLU (Li et al., 2024a), IndoMMLU (Koto et al., 2023), ThaiExam (Pipatanakul et al., 2023), TurkishMMLU (Yüksel et al., 2024), AfriMMLU (Adelani et al., 2024), Khayyam Challenge (Ghahroodi et al., 2024), KMMLU (Son et al., 2024a), HAE-RAE (Son et al., 2024b), and VNHSGE (Dao et al., 2023), covering Arabic, Chinese, Indonesian, Thai, Turkish, Persian, Korean, and Vietnamese, respectively.

Multilingual evaluation datasets include AGIEval (English/Chinese) (Zhong et al., 2023), BEnQ (English/Bengali) (Shafayat et al., 2024), EXAMS (16 languages) (Hardalov et al., 2020), and M3EXAMS (9 languages, multimodal) (Zhang et al., 2023a). While these benchmarks assess LLMs across different languages, they often lack a standardized cross-language comparison. An exception is INCLUDE, which compiles local exams from 44 languages (Romanou et al., 2024).

To broaden multilingual evaluation, MMLU has also been translated. ChatGPT-translated MMLU spans 26 languages (Lai et al., 2023), but translation quality varies across languages (Robinson et al., 2023). More recently, OpenAI released MMMLU, a professional human-translated version in 14 languages, which we incorporate into our benchmark.

A.2 Culturally-aware Evaluation

Recent research has increasingly examined LLMs' cultural alignment. Studies such as Arora et al. (2022) and Cao et al. (2023) explore LLMs' ability to understand cross-cultural differences in values and beliefs. SEA-HELM (formerly BHASA (Leong et al., 2023))[6] is an evaluation suite emphasizing Southeast Asian languages, featuring handcrafted linguistic diagnostics and manually validated SEA-IFEval and SEA-MTBench tasks. Research by Wang et al.

[6] https://leaderboard.sea-lion.ai

(2023) and Masoud et al. (2024) shows LLMs often reflect Western-centric values, even across multiple languages.

Several benchmarks assess cultural biases in LLMs, including Naous et al. (2024) and Rao et al. (2024), while Ventura et al. (2024) examines cultural biases in text-to-image diffusion models. Aakanksha et al. (2024) investigates aligning LLMs to balance linguistic and cultural diversity while minimizing harms. Additionally, studies such as Myung et al. (2024), Magomere et al. (2024), and Montalan et al. (2024) evaluate LLMs' understanding of everyday cultural knowledge across regions.

Multilingual Visual Language Model (VLM) evaluations have also gained attention. PangeaBench assesses 47 languages using 14 preexisting datasets (Yue et al., 2024), while CVQA introduces a culturally diverse Visual Question Answering benchmark covering 30 countries and 31 languages (Romero et al., 2024). Vayani et al. (2024) further extends this with a multimodal benchmark featuring culturally diverse images and text across 100 languages.

Pretraining data significantly influences cultural biases in LLMs. Chen et al. (2024) found models fine-tuned on native instructions outperform those trained on translated data. Choenni et al. (2024) highlights the reliability of machine translation versus human translation in multilingual evaluations. Aya 101, introduced by Üstün et al. (2024), employs in-language prompting and human-written data across 114 languages to reflect local cultures (Singh et al., 2024).

Efforts to enhance cultural alignment in LLMs include cost-effective fine-tuning strategies (Li et al., 2024b) and Anthropological Prompting, a novel approach that applies anthropological reasoning to improve cultural representation (AlKhamissi et al., 2024).

A.3 Participatory Open Science Projects

Participatory research empowers diverse communities to actively contribute to the research process, ensuring inclusivity, contextual relevance, and real-world impact. While most past efforts have focused on specific regions or tasks like translation, character recognition, and audio segmentation, several projects have advanced culturally diverse data collection.

For instance, Clanuwat et al. (2018) tackled the challenge of reading Kuzushiji, a historical Japanese script. MaRVL (Liu et al., 2021) collected culturally representative images from native speakers of Indonesian, Swahili, Tamil, Turkish, and Mandarin Chinese, with linguists providing captions. However, MaRVL's dataset remains limited (<8,000 samples) and is primarily for evaluation. Similarly, Hernandez Mena and Meza Ruiz (2022) developed eight open-access datasets for Mexican and Latin American Spanish via student-led contributions to tasks like audio segmentation and transcription. Other efforts, such as Cañete et al. (2020) and Guevara-Rukoz et al. (2020), have addressed resource scarcity by building datasets for Latin American Spanish.

Masakhane applied a participatory research framework to curate NLP datasets and train models for underrepresented African languages (∀ et al., 2020; Adelani et al., 2021, 2023). Similarly, Project SEALD[7], a collaboration between AI Singapore and Google Research, pioneered multilingual data collection for Southeast Asian LLMs. This initiative supports open-source multilingual models like SEA-LION[8] and its derivatives WangchanLion (Phatthiyaphaibun et al., 2024) and Sahabat-AI[9].

Other large-scale participatory projects include NusaCrowd (Cahyawijaya et al., 2023), which aggregated and standardized data for Indonesian languages, and SEACrowd[10], which extends these efforts to all Southeast Asian languages (Lovenia et al., 2024). The Aya Initiative (Singh et al., 2024; Üstün et al., 2024), with contributions from over 3,000 global participants, collected instruction data in 114 languages, fostering linguistic diversity and inclusivity to create one of the largest multilingual datasets for advancing state-of-the-art LLMs.

B Global-MMLU Knowledge Categories

Annotators were asked to identify MMLU questions where correctly answering depended upon 1) cultural knowledge, 2) geographic knowledge or 3) dialect knowledge.

Cultural Knowledge. Annotators evaluated whether answering a question required culture-specific knowledge. If so, they selected the rele-

[7] https://aisingapore.org/aiproducts/southeast-asian-languages-in-one-network-data-\seald/

[8] https://sea-lion.ai

[9] https://sahabat-ai.com

[10] https://github.com/SEACrowd

vant culture from a drop-down menu with options: Western Culture, Eastern Asian Culture, Middle Eastern Culture, South Asian Culture, African Culture, Latin American Culture, or Other. Cultural knowledge encompasses recognizing and appreciating the beliefs, values, customs, and artistic expressions of a particular group, shaped by shared traditions and heritage (Kipuri, 2009; Liu et al., 2024; Mukherjee et al., 2024).

Geographical or Regional Knowledge. Geographical knowledge refers to understanding characteristics tied to specific regions, such as natural landmarks or environmental features. Annotators determined whether answering correctly required region-specific knowledge. If applicable, they identified the relevant region from a drop-down menu with the following options: North America, South America, Europe, Asia, Africa, Australia and Oceania, and Antarctica.

Dialect Knowledge. This category involves recognizing distinctive language variations or speech patterns used by people from specific regions or communities in English. It includes slang terms, idiomatic expressions, and pronunciation differences that distinguish regional speech from standardized forms of language. Notably, this assessment was conducted on the original English sentences. Therefore, it specifically addresses variations in English dialects or regional vocabulary, rather than any nuances that might arise during the translation process.

C Global-MMLU subsets

Global-MMLU consists of the following smaller annotated subsets:

MMLU Annotated. This subset consists of 2,850 question-answer pairs sampled at uniform from the MMLU dataset (50 questions per subject), representing 20% of the original data and serving as a representative random sample. These samples are annotated in English to determine whether answering requires cultural, geographic, dialectal, or temporal knowledge. The annotations are then applied to corresponding samples in 41 other languages, resulting in a total of 119,700 samples.

Culturally-Sensitive (CS). This subset contains samples identified as requiring dialect knowledge, cultural knowledge or geographic knowledge to answer correctly. It includes 792 annotated samples in English based on majority voting by annotators. These annotations are extended to 41 additional languages, creating a dataset with 33,264 entries. This subset is particularly useful for evaluating model performance on culturally contextual tasks.

Culturally-Agnostic (CA). This subset includes samples that do not contain cultural, regional, or dialectal references. It serves as a baseline for evaluating models on tasks that do not require specific contextual knowledge. The subset consists of 2,058 annotated samples in English, which are extended to 41 languages for a total of 86,436 entries.

D Global-MMLU Subject Categories

Global-MMLU covers six diverse subject categories: STEM, Humanities, Social Sciences, Medical, Business, and Other. For a consistent approach, we adopt the classification proposed by (Hendrycks et al., 2020) for the MMLU dataset to categorize subjects as STEM, Humanities, and Social Sciences. However, we further refine the 'Other' category from the original MMLU dataset by breaking it down into two distinct categories: Medical and Business. Within the 'Other' category, subjects such as clinical knowledge, college medicine, human aging, medical genetics, nutrition, professional medicine, and virology are classified under the Medical category. Meanwhile, business ethics, management, marketing, and professional accounting fall under the Business category. It's worth noting that the 'Other' category in **Global-MMLU**, sometimes referred to as 'General Knowledge', includes the remaining two subjects from the original MMLU 'Other' category: global facts and miscellaneous.

E Global-MMLU Lite

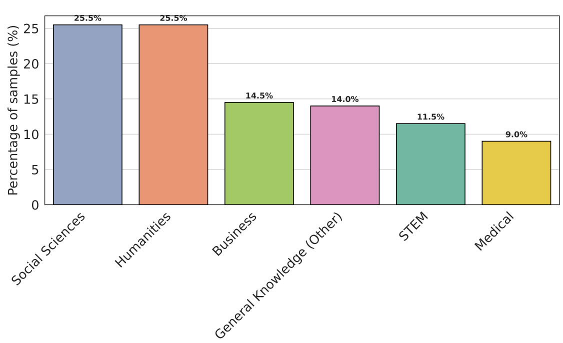

Figure 8: Distribution of samples across subject categories in **Global-MMLU Lite**

As mentioned in section 3.2, **Global-MMLU Lite** is a lighter version of **Global-MMLU** containing 200 **CS** and 200 **CA** samples per language for 15 human-translated or post-edited languages, including English.

For preparing **Global-MMLU Lite**, we took the MA subset of **Global-MMLU** containing 50 samples per subject and looked at proportion of CS and CA samples available per subject. Subjects exclusively tagged as **CS** or **CA** (14 in total) were excluded to ensure both categories were represented within each subject. Consequently, Social Sciences and Humanities subjects are more prevalent in **Global-MMLU Lite**, as shown in Figure 8.

However, we aimed for a balanced distribution across subject categories. Social Science subjects like High School Geography and Sociology had higher proportion of CS samples whereas STEM subjects like Abstract Algebra had higher number of CA samples. To maintain balance, we sampled five **CS** and five **CA** samples per subject where available. Few subjects like Anatomy or High School Mathematics had only one CS sample available, so for such subjects, only one CS and one CA sample was taken. Samples from few subjects of Business and Medical categories were slightly upsampled to ensure adequate representation.

The General Knowledge category, comprising only Miscellaneous and Global Facts, was also upsampled, with 22 samples from Miscellaneous and 8 from Global Facts per category. This adjustment ensures sufficient coverage for evaluating general knowledge capabilities. The overall goal with **Global-MMLU Lite** is to have a balanced dataset for efficient multilingual evaluation across multiple languages.

F Global-MMLU Data Statistics

Table 2 provides a detailed breakdown of the number of subjects and samples in the **CS** and **CA** subsets.

G Temporal Knowledge

As part of the annotation process, annotators were also asked to label samples for temporal or time-sensitive knowledge. This applies to questions where the correct answer may change over time due to factors such as current political leaders or economic statistics. Figure 9 shows the distribution of time sensitive samples in **MMLU Annotated**. Overall it is observed that only 2.4% of the dataset is tagged as time-sensitive and majority of these samples fall under Social Sciences, Humanities, Medical and Other categories. STEM is the only category with no time sensitive samples at all.

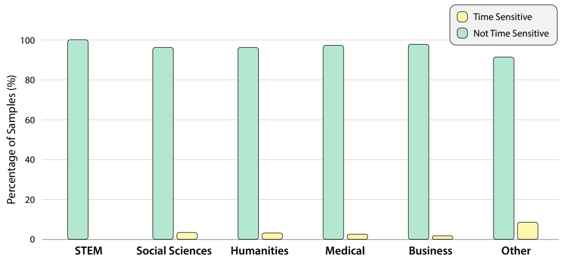

Figure 9: Distribution of time-sensitive samples across subject categories. Note that STEM subjects do not include any temporal knowledge.

H Relationship between cultural and geographical tags

H.1 Culture–Region Relations

We analyzed the samples in the **CS** dataset. Figure 10 illustrates the relationship between Western and Asian cultures and their associated regions. Among the samples labeled with a Western culture tag, 73.3% are also tagged with North America, followed by 25.5% with Europe. Similarly, 97.2% of samples labeled with Asian cultures are associated with the Asia region.

H.2 Culture Country Relations

Figure 11 shows relationship between culture and country. For the Latin American culture, the distribution is balanced, with Bolivia and Mexico comprising 33.3% each of the tags, followed by Hondurus and Peru sharing 16.7% of the tags each. For Indigenous culture, the tags are shared between two countries with USA at top with 66.7% followed by Micronesia at 33.3%. The *Other* culture category was added for representing cultures that did not fall under other pre-existing categories. We find that all samples Other category fall under Russia.

H.3 Region Country Relations

Figure 12 and 13 present country-specific information for each region: *North America*, *Europe*, and *Africa*. The United States accounts for the largest proportion of regional tags, representing

	Number of Subjects			Number of Samples			Data Proportion		
Categories	MA	CS	CA	MA	CS	CA	MA	CS	CA
STEM	19	11	19	950	23	**927**	33.3%	2.9% ↓	45.0% ↑
Humanities	13	12	11	650	**442**	208	22.8%	55.8% ↑	10.1% ↓
Social Sciences	12	11	12	600	208	**392**	21.1%	26.3% ↑	19.1% ↓
Medical	7	5	7	350	19	**331**	12.3%	2.4% ↓	16.1% ↑
Business	4	4	4	200	36	**164**	7.0%	4.5% ↓	8.0% ↑
Other	2	2	2	100	**64**	36	3.5%	8.1% ↑	1.8% ↓

Table 2: Statistics for **MA**, **CS**, and **CA** datasets. The left column displays the number of subjects included in each dataset, the middle column shows the total number of samples per category, and the right column illustrates changes in subject category distributions relative to **MA**, with arrows indicating increases or decreases in representation.

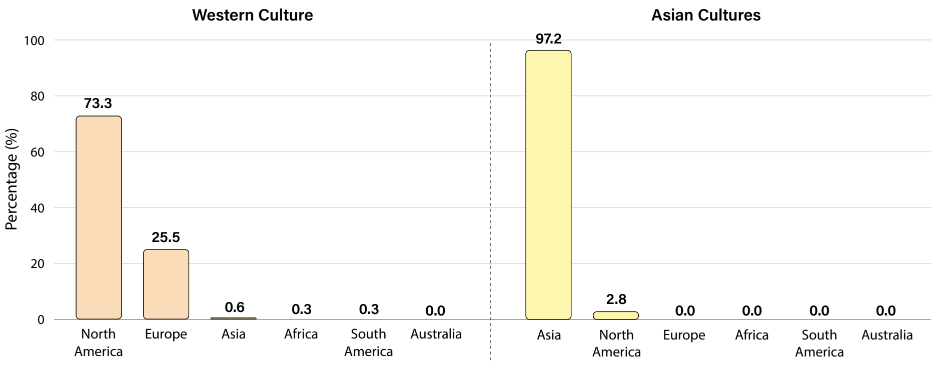

Figure 10: Relationship between Western and Asia cultures and region tags.

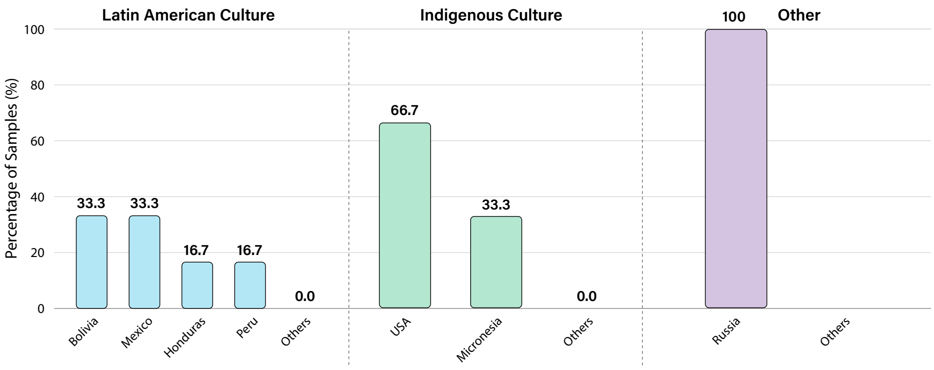

Figure 11: Relationship between culture and country tags, focusing on Latin American and Indigeneous cultures.

89.6% of the tags for the North America region, followed by Canada and the United Kingdom, each with only 0.8% of the tags. For the Europe region, the distribution is more balanced, with the United Kingdom comprising 20.1% of the tags, followed by France at 10.1%. In the Africa region, the distribution is even more balanced, with Egypt and South Africa sharing the top position at 33.3% of the tags each.

I Annotation Process

Communication. For both annotation tasks, annotators were briefed by one of the authors in a virtual introduction session and were able to ask questions and raise issues throughout the annotation task in a Discord channel. For both tasks, they were also encouraged to share frequent error patterns or artifacts that they observed throughout the tasks with the authors and capture difficult decisions and their rationales in comments for individual ratings. Similarly, they discussed ambigu-

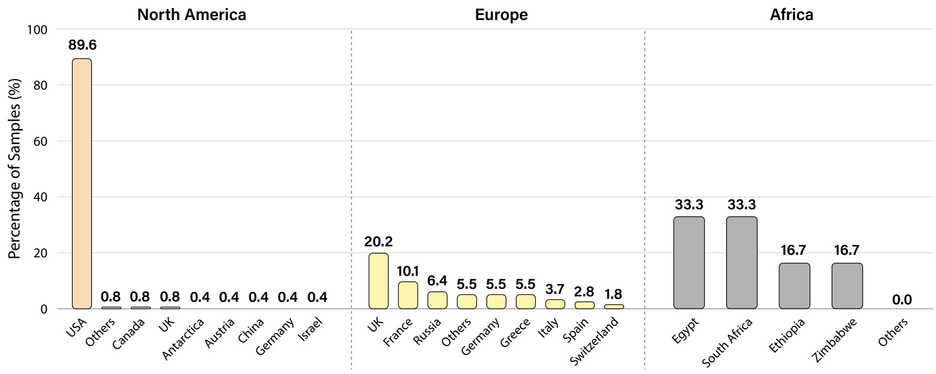

Figure 12: Relationship between region and country tags, focusing on North America, Europe and Africa regions.

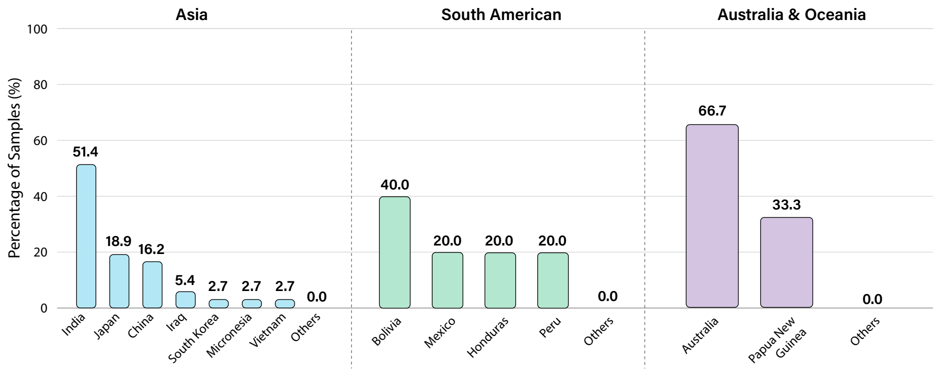

Figure 13: Relationship between region and country tags, focusing on Asia, South America and Australia.

ous cases and questions. This helped calibrate annotations across annotators and languages.

Schedule. Each of the annotation tasks was conducted as 2–3 week long sprints in collaboration with contributors from the community. There was no fixed time schedule for the annotations, and annotators contributed varying hours, depending on their availability and speed.

For the cultural sensitivity evaluation task, 100% of the selected samples were labeled whereas for the translation quality evaluation task, 37% of the provided samples were fully reviewed 12.3% of the samples were edited in total.

Interface. The annotation interface for both tasks was built using Argilla.[11] Argilla is an open-source tool that can be used for data labeling. Using Argilla's Python SDK, it was quick and easy to set up an annotation interface that could be deployed on Hugging Face Spaces. We also set up SSO so annotators could log in and easily access the UI using their Hugging Face accounts.

For cultural sensitivity evaluation, annotators were shown questions one by one from each of the 57 MMLU subjects and were asked to analyze and label the questions for presence of cultural, geographic, dialect or regional knowledge as explained in 2.1. Figure 14 in Appendix I illustrates the annotation interface used during this process. Annotators were presented with questions one at a time from each of the 57 MMLU subjects and had to analyze and label them for the presence of cultural, geographic, and dialect knowledge. Each data point was reviewed by at least three annotators, and some data points had a maximum of 10 annotators. 96.4% of all data points were reviewed by more than 3 human annotators. We classify each question as presenting cultural, geographic and dialect sensitivity according to majority vote among annotators who reviewed each data point (Feldman, 1980). If half or more of the annotators apply the same tag to a question, it is categorized under that tag. Detailed information regarding the annotators and the annotation process is available in Appendix I.

We also asked annotators to annotate for temporal knowledge to determine if answers for questions change with time. We find that only 2.4% of annotated samples depend on temporal knowledge. We provide more details about this analysis

[11] https://argilla.io/

in Appendix G.

As shown in Figure 15, for translation quality evaluation, annotators were shown the translated question and corresponding options in their chosen language on the UI. Annotators were also shown the original question and answer options in English for reference. If the translation was good in quality and correctly represented the original English text then the annotators could mark it as acceptable in quality and proceed to next question otherwise they could edit the provided translation to improve its quality.

I.1 Compensated Annotator Pool for Gold Standard Languages

Annotator Selection. The primary demographic make-up of the participants in the evaluations was recruited based on their proficiency in the language groups. The proficiency was self-reported, and the primary requirement was native or professional proficiency in the specific languages needed for the project.

Socio-Demographics. The annotator pool is comprised of people from diverse backgrounds, and this spans across socioeconomic backgrounds, careers, levels of education, and self-reported gender and sexual identities. We do not ask any annotators to share or report any of these statistical pieces of information in a formal way; any insights into this are gathered organically and through self-reporting by the annotators.

Quality Considerations. We do not believe that any socio-demographic characteristics have led to any impact on the data that has been annotated. Through every part of the project, we have reiterated the importance of this work and the fact that it is helping support a global-scale research project. We are confident in the trust we have built with the annotators in this project, and they care greatly about the overall outcome and, therefore, have been diligent in completing the task with a high degree of accuracy. Where possible, we have done our best to have annotators work on this project and be representatives of the communities that the project aims to support.

I.2 Agreement between Annotators

Inter-annotator agreement. Each data point was reviewed by at least three annotators, and some datapoints had a maximum of 10 annotators. 96.4% of all data points were reviewed by more than 3 human annotators. Given this rich set of feedback on each data point, we analyze the agreement between ratings from different annotators using *Krippendorff's Alpha* scores (Krippendorff, 2004). We observed high inter-annotator agreement across most subjects, with a unanimous cultural sensitivity agreement in the *Anatomy* subject. Six subjects showed disagreement including High-school US History, while Moral Scenarios showed the most disagreement. Detailed results are presented in Figure 17 and 18 in Appendix I.2.

For the first phase of annotations to identify culturally sensitive samples, we ensured that each sample was annotated by at least 3 annotators. We used the ratings for each sample from different annotators and aggregated it per subject to analyze the agreement among annotators. We report the corresponding Krippendorff's Alpha scores depicting annotator agreement in Figure 17 and 18. Krippendorff's Alpha values range between -1 and 1 where 1 denotes that all annotators agree unanimously and -1 denotes that the annotators are making opposite ratings. We observe reasonable disagreement among samples for *moral scenarios* for both cultural sensitivity as well as time-sensitivity annotations. 12 subjects have complete unanimous agreement regarding time-sensitivity annotations between annotators.

J Translation Analysis

J.1 Translation Quality

Figure 20 shows the translation quality comparison for Google Translate which is used to translate **Global-MMLU** and GPT-3.5-turbo which is used for translating multilingual MMLU released by (Lai et al., 2023). We see that Google Translate is significantly better across different MMLU subject categories. For this analysis, we considered samples from MMMLU dataset[12] as the human reference and only considered languages which overlapped between the two machine translated sets and human translated MMMLU.

J.2 Translation Annotators

Professional Annotators. We hired compensated professional annotators for four languages: *Arabic*, *French*, *Hindi*, and *Spanish*. These annotators reviewed the machine translations to ensure fluency and cultural appropriateness, making edits

[12] https://openai.com/index/openai-o1-system-card/

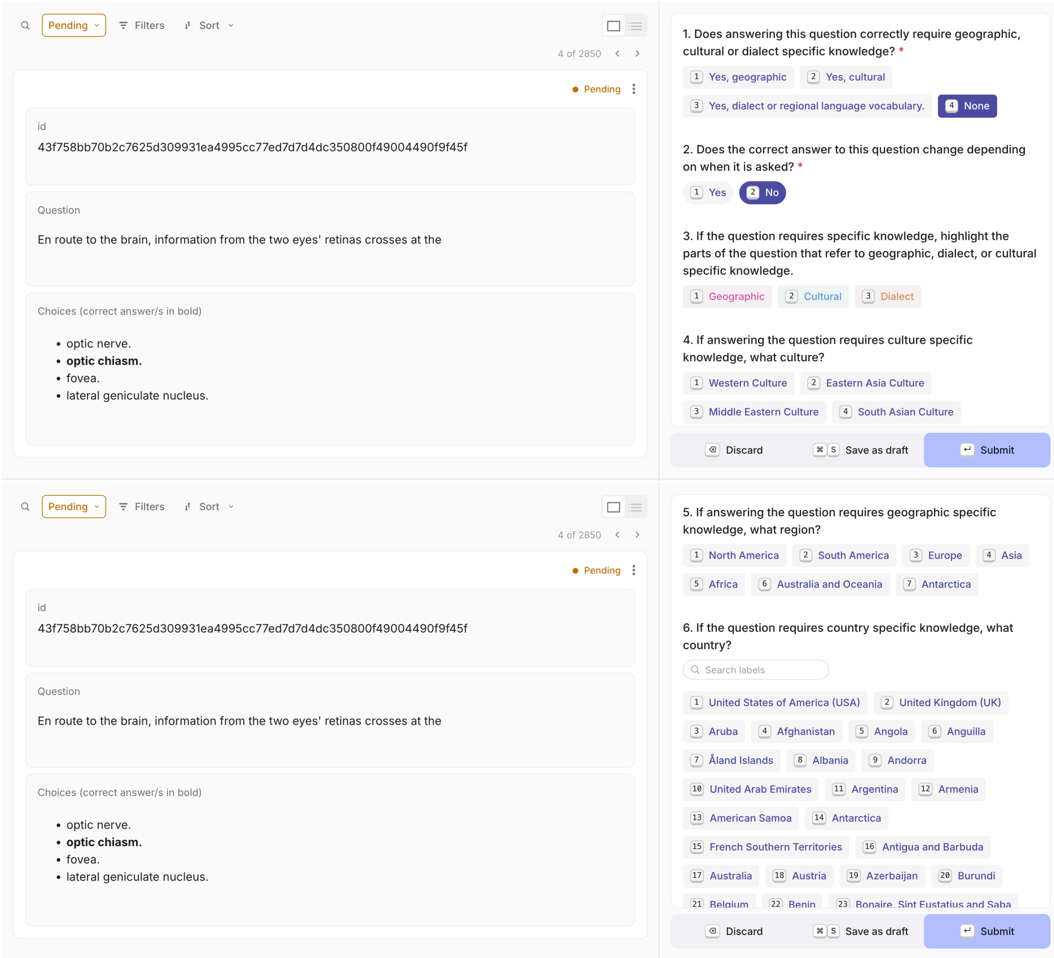

Figure 14: Cultural Sensitivity evaluation annotation interface.

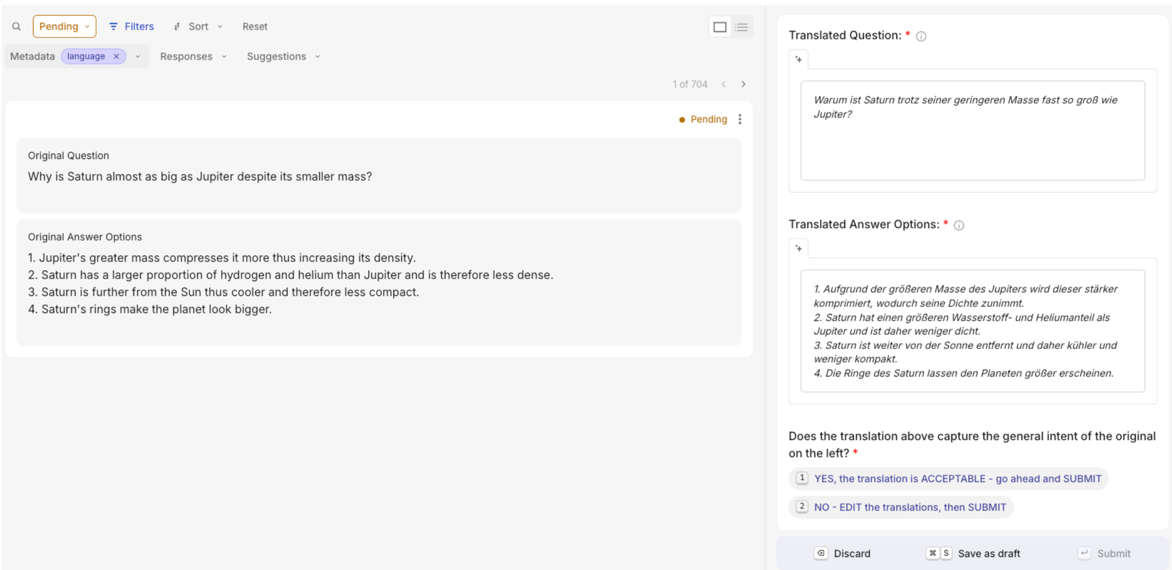

Figure 15: Translation evaluation annotation interface.

where necessary. We refer to this set of translation as our *"Gold Set"*. We include more details about compensated annotation process in section I.1.

Community Annotators. In addition to professional annotations for a subset of languages, we also facilitated community contributions to verify translation quality across a broader range

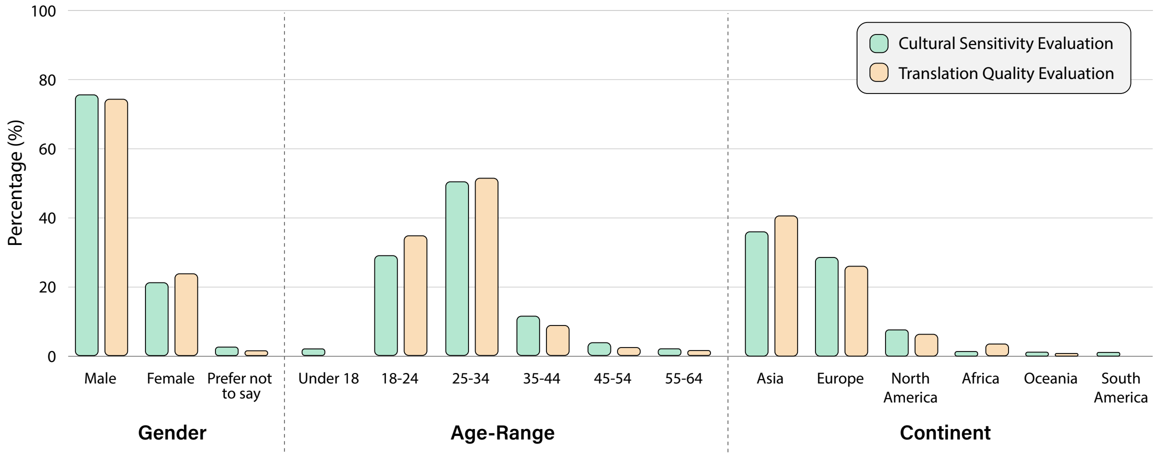

Figure 16: Demographics of annotators who registered using our annotation interface for cultural sensitivity as well as translation quality evaluation.

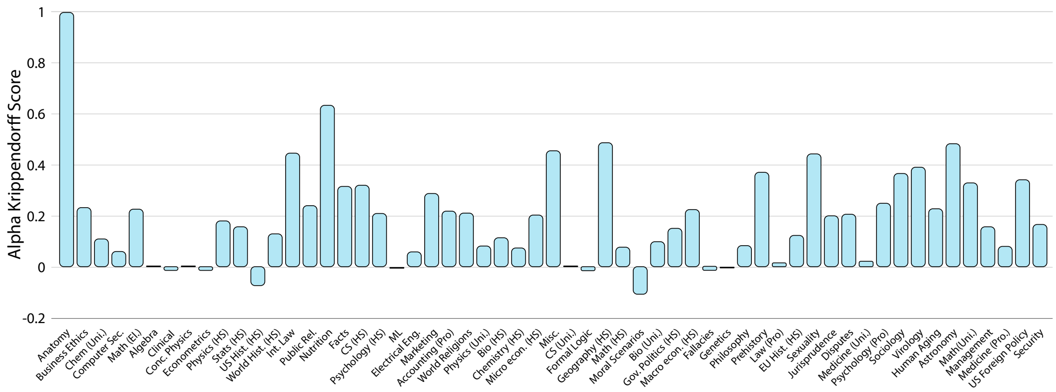

Figure 17: Krippendorff's Alpha Scores for checking annotator agreement regarding the presence of cultural or regional knowledge of samples.

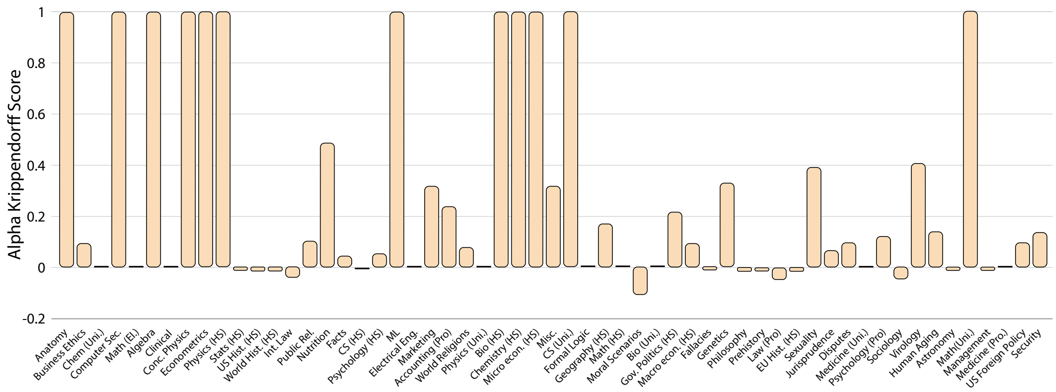

Figure 18: Krippendorff's Alpha Scores for checking annotator agreement regarding the presence of the time-sensitive nature of samples.

of languages, focusing on fluency edits and correcting poor translations. This participatory research approach (Birhane et al., 2022; Corbett et al., 2023; Delgado et al., 2023; Singh et al., 2024; Üstün et al., 2024) involved collaboration across multiple institutions globally. Such cross-sectional efforts are crucial for gathering linguistic data at scale and fostering community engagement—both essential for developing inclusive language technologies (Joshi et al., 2019; Nekoto et al., 2020; Singh et al., 2024; Romanou et al., 2024). We established a criterion requiring a min-

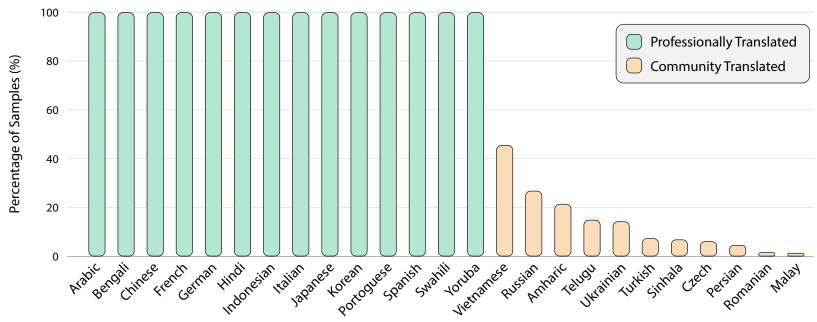

Figure 19: Percentage of Human-Translated Samples in MMLU Annotated.

imum of 50 human-translated samples for each language before its inclusion in **Global-MMLU**. This threshold was met by eleven languages: *Amharic, Czech, Malay, Persian, Romanian, Russian, Sinhala, Telugu, Turkish, Ukrainian*, and *Vietnamese*. In the following sections, we refer to this set of languages as *"Community Translated"*.

The participation of native speakers from diverse regions introduced logistical challenges in both data selection and quality control. To overcome these, we adopted Argilla[13] as our primary annotation platform. In line with our community-based approach, Argilla's collaborative features and customizable workflows enabled us to efficiently manage contributions from various regions while maintaining consistency in translation quality. Annotators were presented with both the original and machine-translated questions and answers, and were asked to edit any translations that did not accurately capture the intent of the original text. The translation interface is shown in Figure 15 in Appendix J.

J.3 Translation Edits

Figure 21 illustrates the *edit distance*, averaged over all samples within each subject category, for edits made by professional and community annotators. The edit distance, calculated using the "Levenshtein Distance" (Levenshtein, 1966), measures the differences between two strings. In this analysis, the machine translations were compared to their edited versions to compute the scores.

The results reveal that the *Humanities* category exhibits the largest edit distances, with higher values observed for questions compared to answers.

Given that longer text may inherently require more edits, we hypothesized that the observed large edit distances could be influenced by the length of the questions and answers. To account for this, we analyzed the length of each question-answer pair and computed the *Normalized Edit Distance* (NED), where the edit distance is divided by the text length, shown in Figure 22.

The analysis reveals that questions in the *Humanities* category have the greatest average length, whereas answers in the *STEM* category exhibit the highest NED. These findings suggest that while raw edit distances are influenced by text length, normalized measures provide additional insights into the complexity of edits across categories.

K Model Evaluations

K.1 Models Covered

We evaluated 14 recent state-of-the-art language models from 9 model families, focusing on those known for their high multilingual performance. These include both small and large open weight models as well as closed models. Details of each model are mentioned below:

Aya Expanse[14] is a family of models include 8B[15] and 32B[16] parameter models. Aya Expanse models support 23 languages including Arabic, Chinese (simplified & traditional), Czech, Dutch, English, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian, and Vietnamese. Aya Expanse builds on the Aya initiative which includes multilingual first releases like Aya 101 (Üstün et al., 2024), Aya 23 (Aryabumi et al., 2024) and extensive multilingual datasets such as Aya collection (Singh et al., 2024).

[13] https://github.com/argilla-io/argilla

[14] https://hf.co/blog/aya-expanse

[15] https://hf.co/CohereForAI/aya-expanse-8b

[16] https://hf.co/CohereForAI/aya-expanse-32b

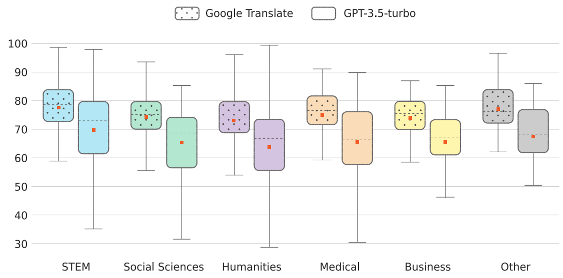

Figure 20: ChrF++ scores for Google Translate and GPT-3.5-Turbo

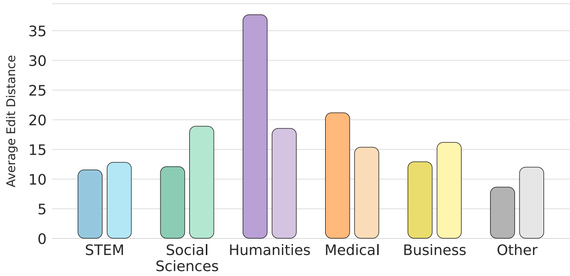

Figure 21: Average edit distance across different subject categories in MMLU. Each sample comprises a question-and-answer pair, with the left column showing edit distances for questions and the right column for answers.

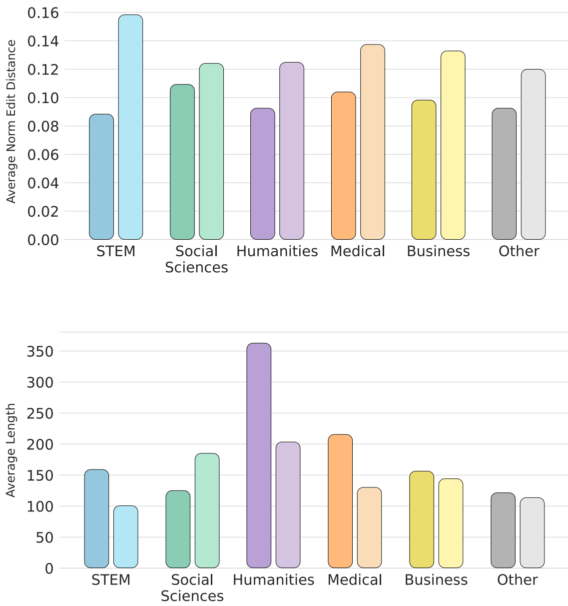

Figure 22: (Top) Average normalized edit distance and (Bottom) average question and answer lengths across different subject categories. The left column represents questions, while the right column represents answers.

Command R and R+ are open-weight models of size 34B[17] and 104B[18] respectively which both support 10 languages: *English, French, Spanish, Italian, German, Brazilian Portuguese, Japanese, Korean, Arabic, Simplified Chinese*. We use Command-R 08-2024 and Command-R+ 08-2024 for evaluation.

Gemma2 (Gemma Team et al., 2024) is part of the Gemma model family. The languages targeted are not explicitly reported. We evaluate the instruct-tuned 9B (gemma-2-9b-it) and 27B (gemma-2-27b-it) variants.

Gemma2-9B-CPT-SEA-LIONv3[19] is part of the SEA-LION[20,21] collection of models trained for Southeast Asian (SEA) languages, including Burmese, Chinese, English, Filipino, Indonesian, Javanese, Khmer, Lao, Malay, Sundanese, Tamil, Thai, and Vietnamese. We use Gemma2-9B-CPT-SEA-LIONv3-Instruct for evaluation.

Llama 3.1 (Dubey et al., 2024) is a series of open LLM models that come in three sizes: 8B, 70B, and 405B parameters. All variants support 8 languages, including English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. We use Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct for evaluation.

Mistral Nemo[22] is a 12B model which supports 11 languages including English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi.

Qwen 2.5[23] model supports up to 29 languages, including Chinese, English, French, Spanish, and Portuguese. We evaluate Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct variants of Qwen 2.5.

GPT-4o (Hurst et al., 2024) is a multilingual, multimodal closed-model and is part of the GPT-4 family. The languages targeted are not explicitly reported.

Claude Sonnet 3.5 is also a multilingual, multimodal closed-model from the Claude 3.5 family.

[17] https://hf.co/CohereForAI/c4ai-command-r-08-2024

[18] https://hf.co/CohereForAI/c4ai-command-r-plus-08-2024

[19] https://hf.co/aisingapore/gemma2-9b-cpt-sea-lionv3-instruct

[20] An acronym for **S**outh**e**ast **A**sian **L**anguages **in** **O**ne **N**etwork.

[21] https://github.com/aisingapore/sealion

[22] https://hf.co/mistralai/Mistral-Nemo-Instruct-2407

[23] https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e

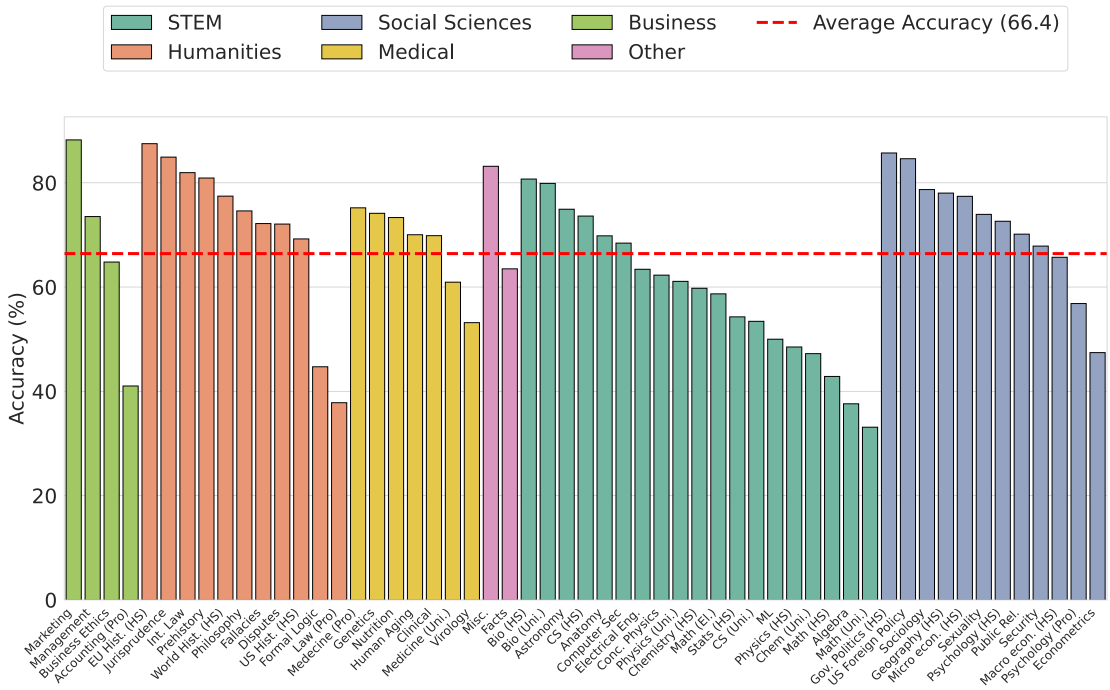

Figure 23: Aya Expanse 32B performance on each subjects.

The languages supported by this model are also unknown.

We note that all these models do not claim to support the same set of languages, and none claim to support the full set of languages we cover.

K.2 Evaluation Setup

We use *lm-evaluation-harness* (Gao et al., 2024) to evaluate the open multilingual models in a 5-shot setting. For closed models, we also do 5-shot evaluation. However, since log probabilities are not accessible via API for closed models, we send the 5-shot prompt via API and get the corresponding generation from the model. We use a system preamble to make the model respond with only the correct answer option and extract the answer from the output generation. For prompting, we follow the same approach as specified in (Hendrycks et al., 2020) and use prompt instructions in the same language as the sample.

K.3 Evaluation Results

K.3.1 Subject-level Performance

Figure 23 illustrates the performance of the Aya Expanse 32 model across various subjects, with an average accuracy of 66.4%. Notably, most *STEM* subjects fall below this average, whereas the majority of *Social Sciences* and *Humanities* subjects exceed it.

K.3.2 Human Translated vs Machine Translated

We compared models on Human-Translated (HT) and Machine-Translated (MT) **CS** datasets to gain deeper insights into model behavior. Figure 24 illustrates the model performances for one ●high-resource language (French), one ◐mid-resource language (Korean), one ○low-resource language (Yoruba).

The key finding is that models generally perform better on human-translated data for ●high-resource languages. This is likely because these languages benefit from extensive in-language training data. However, this trend shifts for ◐mid-resource languages. The figure reveals that the performance gap between HT and MT narrows for models such as Claude Sonnet and Qwen2.5 32B. Conversely, models like CommandR+ and Aya Expanse 32B continue to perform better on HT data. Notably, these two models have strong Korean language support, which can be attributed to a substantial amount of in-language training data.

For ○low-resource languages, a distinct pattern emerges. As shown in the figure, models such as Claude Sonnet and GPT-4o perform significantly better on MT data than on

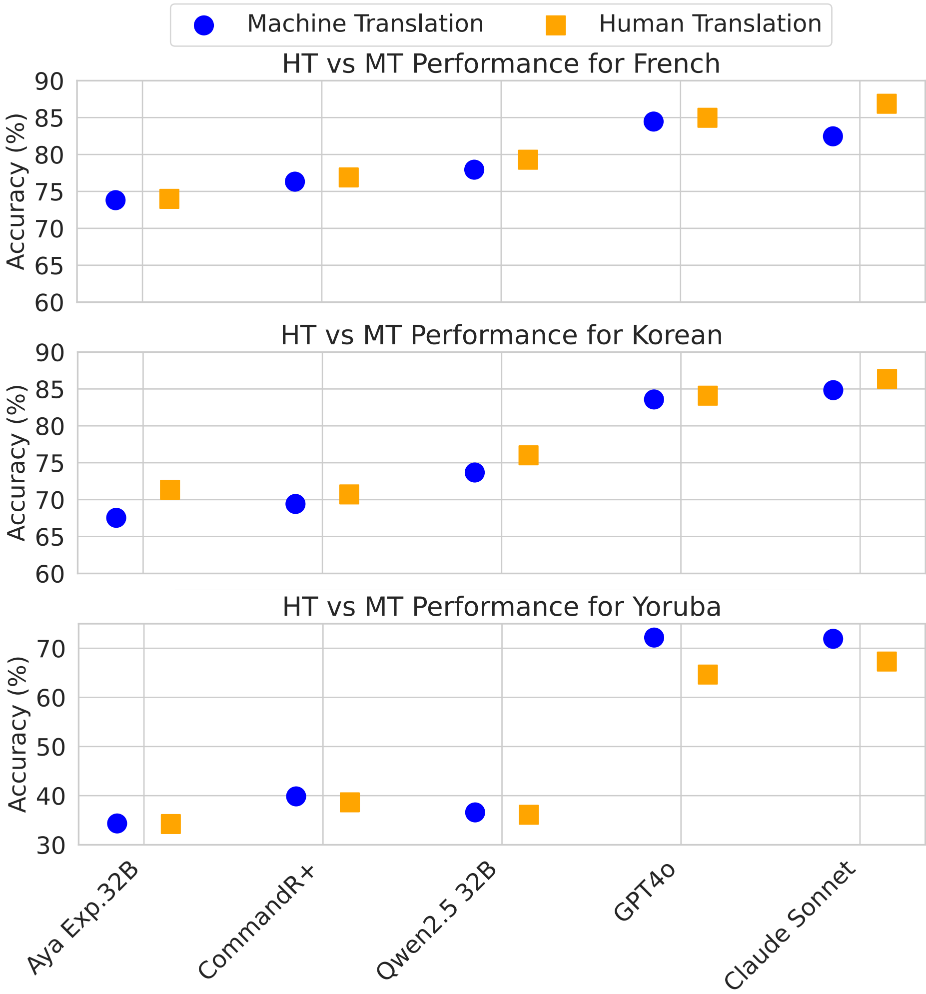

Figure 24: Comparison of model performance on *human-translated* and *machine-translated* **CS** in French, Korean, and Yoruba.

HT data. Similarly, CommandR+ and Qwen2.5 32B also show improved performance on MT data, albeit with less pronounced differences. This behavior is likely because these models primarily rely on machine-translated data for low-resource languages during training, and the distribution of the machine-translated test set aligns more closely with their training data. Notably, the only model demonstrating consistent performance across both HT and MT datasets is Aya Expanse 32B, which can be attributed to its broad coverage and strong support for low-resource languages.

These results underscore the importance of in-language or human-translated datasets for evaluating low-resource languages. The **Global-MMLU** dataset provides a valuable tool for assessing the in-language performance of large language models (LLMs) on low-resource languages, offering insights into their capabilities and limitations in such contexts.

K.3.3 Model Rank Changes

Table 5 presents the rank changes and corresponding position shifts (indicated next to the arrows) for high-resource languages, while Table 6 provides similar data for mid- and low-resource languages. The rightmost columns in each table summarize the total number of models that changed ranks (*Total Rank Change*) and the total number of position shifts in the rankings (*Total Position Change*). A detailed analysis of these results is provided in Section 4.3.

Language	Dataset	Aya Exp. 8B	Aya Exp. 32B	CommandR	CommandR+	Gemma2 9B	Gemma2 27B	Llama-3.1 8B	Llama-3.1 70B	Mistral Nemo	Qwen2.5 7B	Qwen2.5 32B	SEA-LION-v3	GPT4o	Claude Sonnet
◐Greek	CA	↓1	↓1	-	-	-	↑1	-	-	↓1	↑2	-	-	-	-
	CS	-	-	↑2	↑3	-	↓1	↑1	-	-	↓1	↓4	-	-	-
◐Ukrainian	CA	-	↑1	-	↓1	↓1	-	-	-	-	↑1	-	-	-	-
	CS	-	↑1	-	↑1	-	↓2	-	↑1	↑1	↓1	↓1	-	↑1	↓1
○Malagasy	CA	-	↓1	-	-	-	-	-	-	-	↑1	-	-	-	-
	CS	-	↑1	↑4	↑1	-	-	↓1	-	↑1	↓1	↓5	-	-	-
○Shona	CA	-	-	-	-	↓1	-	-	-	-	-	↑1	-	↓1	↑1
	CS	↑2	-	↑1	↑1	-	-	↑1	-	-	↓4	↓1	-	-	-

Table 3: Changes in model rankings on **CA** and **CS** datasets, based on MA on Greek, Ukrainian, Malagasy, and Shona.

Language	Dataset	Aya Exp. 32B	CommandR+	Gemma2 27B	Llama-3.1 70B	Mistral Nemo	Qwen2.5 32B	SEA-LION-v3
● Arabic	CA	-	↓1	↑1	-	-	-	-
	CS	↑1	-	↓1	-	-	-	-
● Chinese	CA	↑1	↓1	-	-	-	-	-
	CS	-	↑1	↓1	-	-	-	-
● English	CA	↓1	↓1	↑1	↓1	-	↑1	↑1
	CS	↑1	-	↓1	-	↑1	-	↓1
● French	CA	↑1	↓1	-	-	-	-	-
	CS	↓1	↑1	↓1	↑1	↑2	↓1	↓1
● German	CA	-	↓1	-	↓1	-	↑2	-
	CS	-	↑1	-	↓1	-	-	-
● Hindi	CA	↓1	-	-	-	-	↓2	↑3
	CS	-	-	-	-	-	-	-
● Italian	CA	↑2	↓3	-	-	-	-	↑1
	CS	-	-	-	-	↑1	-	↓1
● Japanese	CA	↑1	↓1	-	-	-	-	-
	CS	-	-	-	-	-	-	-

Language	Dataset	Aya Exp. 32B	CommandR+	Gemma2 27B	Llama-3.1 70B	Mistral Nemo	Qwen2.5 32B	SEA-LION-v3
● Portuguese	CA	↓1	↓2	↑1	↓1	-	↑1	↑2
	CS	↑1	-	↓1	-	-	-	-
● Spanish	CA	-	-	-	-	-	-	-
	CS	-	-	-	↑1	-	↓1	-
◐ Bengali	CA	↑1	-	-	-	↓1	-	-
	CS	-	-	-	-	-	-	-
◐ Indonesian	CA	-	-	-	-	-	-	-
	CS	↑1	↑1	↓2	-	-	-	-
◐ Korean	CA	↓1	↑1	-	-	-	-	-
	CS	-	-	-	-	-	-	-
○ Swahili	CA	↓1	↑1	↑1	↓1	↑1	↓1	-
	CS	↑1	↓1	-	-	-	-	-
○ Yoruba	CA	-	↓2	-	↓2	-	↑1	↑3
	CS	↑3	↑1	↓4	↑1	-	-	↓1

Table 4: Changes in model rankings on **CA** and **CS** datasets, based on total accuracy on **Global-MMLU Lite**. Languages are categorized as ●high-, ◐mid-, and ○low-resource. Color-coded boxes indicate increases (↑) and decreases (↓) in rank.

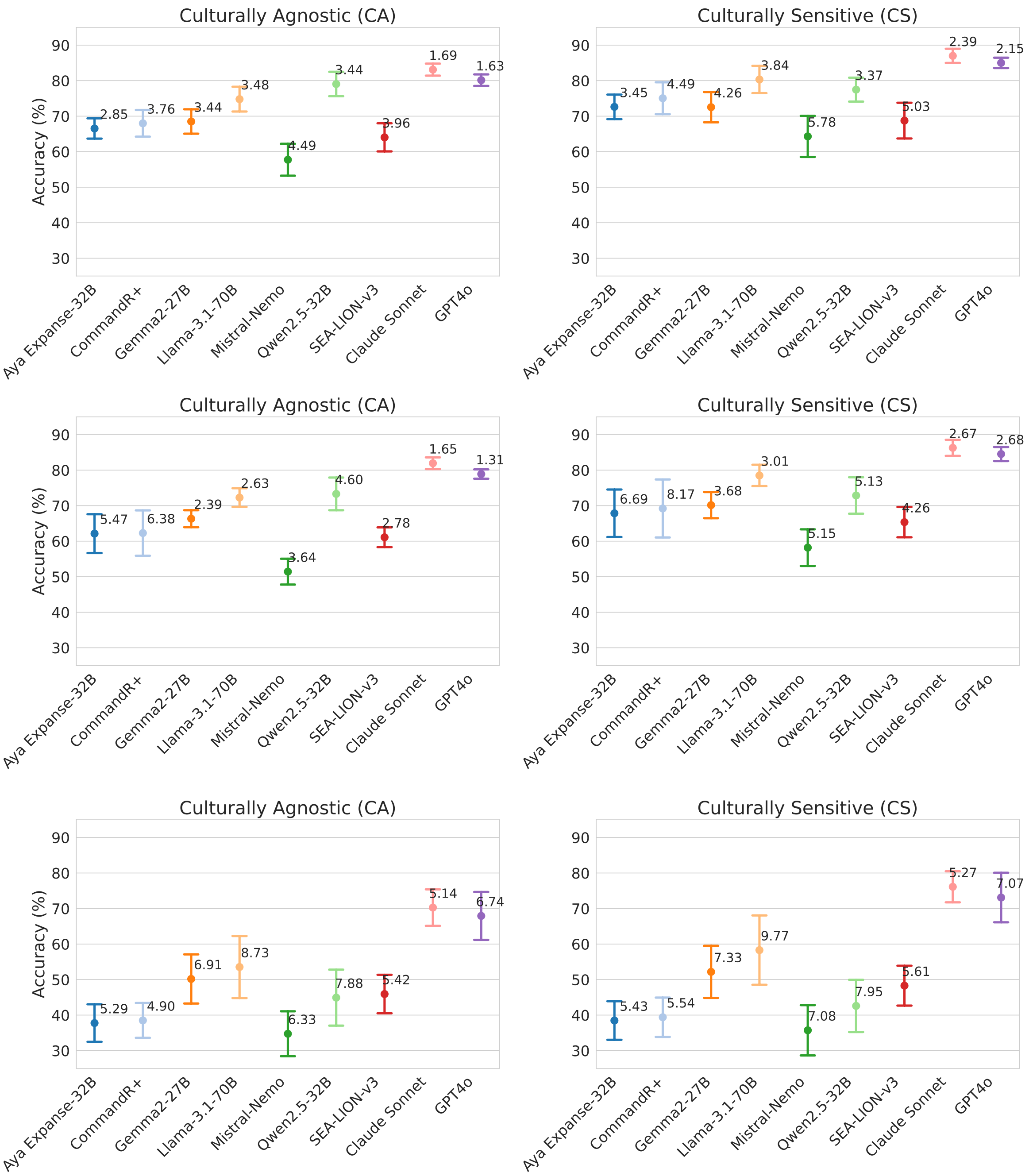

Figure 25: Model evaluations on (Top) high-resource , (Mid) mid-resource and (Bottom) low resource data samples for **CA** and **CS** subsets.

Language	Dataset	Aya Exp. 8B	Aya Exp. 32B	CommandR	CommandR+	Gemma2 9B	Gemma2 27B	Llama-3.1 8B	Llama-3.1 70B	Mistral Nemo	Qwen2.5 7B	Qwen2.5 32B	SEA-LION-v3	GPT4o	Claude Sonnet	**Total rank change**	**Total position change**
Arabic	CA	-	-	-	-	-	-	-	-	-	↑1	-	↓1	-	-	2	2
	CS	-	↑1	-	-	-	↓1	-	↑1	-	-	↓1	-	-	-	4	4
Chinese	CA	-	-	↓1	-	↑1	-	-	-	-	-	↑1	-	↓1	-	4	4
	CS	↑1	↑1	↑1	↑2	↑1	-	↓1	↑1	-	↓3	↓1	↓2	↑1	↓1	12	16
Czech	CA	-	-	-	-	-	-	-	↓1	-	-	↑1	-	-	-	2	2
	CS	↑2	↓1	-	↑3	-	↓1	↓2	-	-	-	↓1	-	-	-	6	10
Dutch	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	CS	-	-	-	↑1	↑2	↓1	-	↑1	-	↓2	↓1	-	-	-	6	8
English	CA	-	-	-	-	-	↓1	-	-	-	↑1	↑1	-	↓1	-	4	4
	CS	-	↑1	-	-	-	-	-	↑1	-	↓1	↓1	-	-	-	4	4
French	CA	-	↑1	-	-	-	-	-	-	-	↓1	-	-	-	-	2	2
	CS	-	↑2	↑2	↑1	-	↓2	-	↑1	-	↓3	↓1	↑1	-	-	8	13
German	CA	-	↓1	-	↓1	-	↑1	-	-	-	↑1	-	-	-	-	4	4
	CS	-	-	↓1	-	↑2	-	-	↑1	-	↓3	↓1	↑2	-	-	6	10
Hindi	CA	-	↑1	↓2	↓1	↑1	-	-	-	-	-	-	↑1	-	-	5	6
	CS	↑1	↓1	↑1	↑2	-	↓1	↑1	-	↑1	↓3	↓1	-	↑1	↓1	11	14
Italian	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	CS	-	-	↑1	↑1	-	↓1	-	↑1	-	↓2	↓1	↑1	-	-	7	8
Japanese	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	CS	-	↑1	↑1	↑1	↑1	↓2	-	↑1	-	↓1	↓1	↓1	-	-	9	10
Persian	CA	↑1	↑1	-	↓1	-	-	↑1	-	↓2	-	-	-	-	-	5	6
	CS	-	-	-	↑2	↑1	↓2	-	-	↑1	↑1	↓1	↑1	-	-	7	9
Polish	CA	↑2	↑1	↑2	↓1	↓1	-	↓1	-	↓1	↑2	-	↓1	-	-	9	12
	CS	-	-	↑2	↑2	-	↓1	-	↑1	↑1	↓1	↓1	-	-	-	7	9
Portuguese	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	CS	-	↑1	↑1	↑1	↑1	↓1	-	↑1	-	↓2	↓1	↓1	-	-	9	10
Russian	CA	-	↓1	↓1	↓1	↑1	-	-	-	-	↑2	-	-	-	-	5	6
	CS	↑1	-	-	↑2	↓1	↓1	↑1	-	-	↓2	↓1	↑3	-	-	8	12
Serbian	CA	-	↓1	-	↑1	-	-	-	↓1	-	↓1	↑1	-	-	-	5	5
	CS	-	↑2	↑1	↓1	↑1	-	-	-	-	-	-	-	-	-	4	5
Spanish	CA	-	↓1	-	↓1	-	↑1	-	-	-	↑1	-	-	-	-	4	4
	CS	-	-	↑1	-	↑2	-	-	↑1	-	↓3	↓1	-	-	-	5	8
Swedish	CA	-	↓1	-	↓1	-	↑1	-	-	-	↓1	-	-	-	-	4	4
	CS	-	-	↑1	-	↑2	-	-	↑1	-	↓3	↓1	-	-	-	5	8
Turkish	CA	-	-	-	↓1	↑1	-	↓1	-	-	↑1	-	-	-		4	4
	CS	-	↑2	↓1	↑1	-	↓1	-	-	-	-	↓2	-	-	-	5	7
Vietnamese	CA	-	-	-	↓1	-	↑1	↓1	-	- ↑1	-	-	-	-	-	4	4
	CS	-	↓1	↑3	-	-	↓1	↓1	-	↑1	↓1	-	-	-	-	6	8

Table 5: Model rankings with MA rank as the reference for high-resource languages (●). First row indicates changes in CA ranks, while second row shows the changes in CS ranks relative to MA. Color-coded boxes highlight increases (↑) and decreases (↓).

Language	Dataset	Aya Exp. 8B	Aya Exp. 32B	CommandR	CommandR+	Gemma2 9B	Gemma2 27B	Llama-3.1 8B	Llama-3.1 70B	Mistral Nemo	Qwen2.5 7B	Qwen2.5 32B	SEA-LION-v3	GPT4o	Claude Sonnet	**Total rank change**	**Total position change**
◐Bengali	CA	-	↑1	-	-	-	-	-	↓1	↓1	-	-	-	-	-	3	3
	CS	-	-	-	-	-	-	-	-	↑1	↓1	-	-	-	-	2	2
◐Filipino	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	CS	-	-	-	-	-	↑1	↑1	-	↓1	-	↓1	-	↓1	↑1	6	6
◐Greek	CA	↓1	↓1	-	-	-	↑1	-	-	↓1	↑2	-	-	-	-	5	6
	CS	-	-	↑2	↑3	-	↓1	↑1	-	-	↓1	↓4	-	-	-	6	12
◐Hebrew	CA	↓1	↑1	-	↓1	-	-	-	-	↑1	-	-	-	-	-	4	4
	CS	-	↑2	-	↑2	-	↓2	-	-	-	-	↓2	-	-	-	4	8
◐Indonesian	CA	-	-	↓1	↓1	↓1	↑1	-	-	-	↑2	-	-	-	-	5	6
	CS	-	-	↑1	-	-	-	↓1	↑1	↑1	-	↓1	↓1	-	-	6	6
◐Korean	CA	↓1	↓1	↓1	-	-	↑1	↑1	-	-	↑1	-	-	-	-	6	6
	CS	-	↑1	↑1	↓1	-	↓1	-	↑1	-	-	↓1	-	-	-	6	6
◐Malay	CA	-	-	-	-	↓1	-	-	↓1	-	↑1	↑1	-	-	-	4	4
	CS	-	↑1	↑1	↓1	-	-	-	-	-	↓1	-	-	-	-	4	4
◐Lithuanian	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	CS	-	-	-	↑2	-	-	-	-	-	-	-	↓2	-	-	2	4
◐Romanian	CA	-	↑1	-	↓1	-	-	↑1	↓1	↓1	-	↑1	-	-	-	6	6
	CS	-	-	-	↑2	-	↓1	-	-	-	-	-	-	-	-	2	3
◐Ukrainian	CA	-	↑1	-	↓1	↓1	-	-	-	-	↑1	-	-	-	-	4	4
	CS	-	↑1	-	↑1	-	↓2	-	↑1	↑1	↓1	↓1	-	↑1	↓1	9	10
◯Amharic	CA	-	-	↓1	↑1	↓1	-	-	-	-	-	↑1	-	-	-	4	4
	CS	↓1	↑2	↑2	↓1	-	-	-	-	↑1	↓3	-	-	-	-	6	10
◯Hausa	CA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	CS	↑1	↓1	↑3	↓1	-	-	↓1	-	↓1	↓1	↑1	-	-	-	8	10
◯Igbo	CA	-	-	↓1	-	-	-	↓1	-	-	↑1	↑1	-	-	-	4	4
	CS	-	↑1	-	-	↑1	-	-	-	↑2	↓3	-	↓1	-	-	5	8
◯Kyrgyz	CA	-	-	-	-	-	↓1	-	-	-	-	↑1	-	-	-	2	2
	CS	-	↓1	↑1	↑1	-	-	↑1	-	-	↓2	-	-	-	-	5	6
◯Malagasy	CA	-	↓1	-	-	-	-	-	-	-	↑1	-	-	-	-	2	2
	CS	-	↑1	↑4	↑1	-	-	↓1	-	↑1	↓1	↓5	-	-	-	7	14
◯Nepali	CA	-	-	-	-	-	-	-	-	↓1	↑1	-	-	-	-	2	2
	CS	-	-	-	-	-	↑1	↓1	-	↑1	-	↓1	-	↑1	↓1	6	6
◯Nyanja	CA	-	-	-	↓1	↓1	-	-	-	-	-	↑2	-	↑1	↓1	5	6
	CS	-	↓1	↑1	-	-	-	-	-	-	-	-	-	-	-	2	2
◯Shona	CA	-	-	-	-	↓1	-	-	-	-	-	↑1	-	↓1	↑1	4	4
	CS	↑2	-	↑1	↑1	-	-	↑1	-	-	↓4	↓1	-	-	-	6	10
◯Sinhala	CA	-	↑1	-	-	-	-	↓3	-	-	↑2	-	-	-	-	3	6
	CS	-	↓1	↑1	↑1	-	-	-	-	-	↓1	-	-	-	-	4	4
◯Somali	CA	-	↓2	-	↑1	-	-	-	-	-	↑1	-	-	-	-	3	4
	CS	-	↑1	↑2	↓2	-	-	↑2	-	-	↓2	↓1	-	↓1	↑1	8	12
◯Swahili	CA	-	↓1	-	-	-	-	↑1	-	-	-	-	-	-	-	2	2
	CS	-	-	↑1	-	-	-	↓1	-	-	-	-	-	↓1	↑1	4	4
◯Telugu	CA	-	↓1	-	-	-	-	-	-	-	↑1	↑1	↓1	-	-	4	4
	CS	-	↓1	↑2	↑1	↑1	-	↑1	-	↓1	↓2	↓1	-	-	-	8	10
◯Yoruba	CA	-	↑1	↓2	-	↓1	-	-	-	-	↑2	↑1	↓1	-	-	6	8
	CS	-	↓1	↑1	↑1	↑1	-	-	-	-	-	↓2	-	-	-	5	6

Table 6: Model rankings with MA rank as the reference for mid (◐) and low (◯) resource languages. First row indicates changes in CAranks, while second row shows the changes in CS ranks relative to MA. Color-coded boxes highlight increases (↑) and decreases (↓).

L Global-MMLU Languages

In this work we will refer to groups of languages to be "lower-", "mid-" or "higher"-resourced according to their recorded, written, and catalogued NLP resources (Joshi et al., 2020). We group these 5 distinct clusters following the groupings in (Singh et al., 2024) into a rough taxonomy of **lower-resourced (LR)**, **mid-resourced (MR)** and **higher-resourced (HR)**. We note that this grouping is inevitably imperfect; languages and their varieties cannot absolutely nor universally be classified based on this single dimension (Hämäläinen, 2021; Bird, 2022). The categorization in our case serves the purpose of aggregation in our analysis of the data distribution.

ISO Code	Language	Script	Resource	Type
am	Amharic	Ge'ez	Low	♢ ♤
ar	Arabic	Arabic	High	♤
bn	Bengali	Bengali	Mid	♤
cs	Czech	Latin	High	♢ ♤
de	German	Latin	High	♤
el	Greek	Greek	Mid	♢
en	English	Latin	High	♢ ♤
fil	Filipino	Latin	Mid	♢
fr	French	Latin	High	♤
ha	Hausa	Latin	Low	♢
he	Hebrew	Hebrew	Mid	♢
hi	Hindi	Devanagari	High	♤
ig	Igbo	Latin	Low	♢
id	Indonesian	Latin	Mid	♤
it	Italian	Latin	High	♤
ja	Japanese	Japanese	High	♤
ky	Kyrgyz	Cyrillic	Low	♢
ko	Korean	Hangul	Mid	♤
lt	Lithuanian	Latin	Mid	♢
mg	Malagasy	Latin	Low	♢
ms	Malay	Latin	Mid	♢ ♤
ne	Nepali	Devanagari	Low	♢
nl	Dutch	Latin	High	♢
ny	Nyanja	Latin	Low	♢
fa	Persian	Arabic	High	♢ ♤
pl	Polish	Latin	High	♢
pt	Portuguese	Latin	High	♤
ro	Romanian	Latin	Mid	♢ ♤
ru	Russian	Cyrillic	High	♢ ♤
sin	Sinhala	Sinhala	Low	♢ ♤
sn	Shona	Latin	Low	♢
som	Somali	Latin	Low	♢
es	Spanish	Latin	High	♤
sr	Serbian	Cyrillic	High	♢
sw	Swahili	Latin	Low	♤
sv	Swedish	Latin	High	♢
te	Telugu	Telugu	Low	♢ ♤
tr	Turkish	Latin	High	♢ ♤
uk	Ukrainian	Cyrillic	Mid	♢ ♤
vi	Vietnamese	Latin	High	♢ ♤
yo	Yorùbá	Latin	Low	♤
zh	Chinese	Hans	High	♤

Table 7: 42 languages in **Global-MMLU**, along with each language's script and resource category. We followed (Singh et al., 2024) and categorized languages as low, mid and high resource based on language classes proposed by (Joshi et al., 2020) (low: [0, 1, 2], mid: [3], high: [4, 5]). In *Global-MMLU*, the language is either fully machine translated ♢, fully human translated ♤, or contains both machine and human translated data ♢♤.

M MMLU Annotated Examples

Dataset	Subject	Question	Choices
CS	**US Hist. (HS)**	This question refers to the following information: "Some men look at constitutions with sanctimonious reverence, and deem them like the ark of the covenant, too sacred to be touched. They ascribe to the men of the preceding age a wisdom more than human, and suppose what they did to be beyond amendment . . . . But I know also, that laws and institutions must go hand in hand with the progress of the human mind. As that becomes more developed, more enlightened, as new discoveries are made, new truths disclosed, and manners and opinions change with the change of circumstances, institutions must advance also, and keep pace with the times." —Thomas Jefferson, 1816 Which of the following best describes a contributing factor in the crafting of the United States Constitution?	(A) Individual state constitutions written at the time of the Revolution tended to cede too much power to the federal government, leading to a call for reform on the part of Anti-Federalists. (B) The weaknesses of the Articles of Confederation led James Madison to question their efficacy and prompted a formation of the Constitutional Congress in 1787. (C) Difficulties over trade and foreign relations led to a repeal of overly restrictive tariffs required by the Articles of Confederation. (D) Washington's embarrassing failure at the Whiskey Rebellion led to Federalist demands for a new framework for federal power.
CS	**Accounting (Pro)**	Under the Sales Article of the UCC, which of the following circumstances best describes how the implied warranty of fitness for a particular purpose arises in a sale of goods transaction?	(A) The buyer is purchasing the goods for a particular purpose and is relying on the seller's skill or judgment to select suitable goods. (B) The buyer is purchasing the goods for a particular purpose and the seller is a merchant in such goods. (C) The seller knows the particular purpose for which the buyer will use the goods and knows the buyer is relying on the seller's skill or judgment to select suitable goods. (D) The seller knows the particular purpose for which the buyer will use the goods and the seller is a merchant in such goods.
CS	**Jurisprudence**	Which of the following criticisms of Llewellyn's distinction between the grand and formal styles of legal reasoning is the most compelling?	(A) There is no distinction between the two forms of legal reasoning. (B) Judges are appointed to interpret the law, not to make it. (C) It is misleading to pigeon-hole judges in this way. (D) Judicial reasoning is always formal.
CS	**Prehistory**	What is the name of the lithic technology seen in the Arctic and consisting of wedge-shaped cores, micro-blades, bifacial knives, and burins?	(A) Clovis Complex (B) Denali Complex (C) Folsom Complex (D) Nenana Complex
CS	**US Foreign Policy**	What was the key difference between US expansion pre- and post- 1865?	(A) US expansion was based on territory rather than markets post-1865 (B) US expansion was based on markets rather than territory post-1865 (C) US expansion was limited to Latin America post-1865 (D) US expansion ended after 1865

CA	**Econometrics**	Which of the following statements will be true if the number of replications used in a Monte Carlo study is small? i) The statistic of interest may be estimated imprecisely ii) The results may be affected by unrepresentative combinations of random draws iii) The standard errors on the estimated quantities may be unacceptably large iv) Variance reduction techniques can be used to reduce the standard errors	(A) (ii) and (iv) only (B) (i) and (iii) only (C) (i), (ii), and (iv) only (D) (i), (ii), (iii), and (iv)
CA	**Stats (HS)**	An assembly line machine is supposed to turn out ball bearings with a diameter of 1.25 centimeters. Each morning the first 30 bearings produced are pulled and measured. If their mean diameter is under 1.23 centimeters or over 1.27 centimeters, the machinery is stopped and an engineer is called to make adjustments before production is resumed. The quality control procedure may be viewed as a hypothesis test with the null hypothesis $H_0 : \mu = 1.25$ and the alternative hypothesis $H_a : \mu \neq 1.25$. The engineer is asked to make adjustments when the null hypothesis is rejected. In test terminology, what would a Type II error result in?	(A) A warranted halt in production to adjust the machinery (B) An unnecessary stoppage of the production process (C) Continued production of wrong size ball bearings (D) Continued production of proper size ball bearings
CA	**Formal Logic**	Construct a complete truth table for the following argument. Then, using the truth table, determine whether the argument is valid or invalid. If the argument is invalid, choose an option which presents a counterexample. (There may be other counterexamples as well.) $M \vee N \neg M \wedge \frac{O}{N}$	(A) Valid (B) Invalid. Counterexample when M and O are true and N is false (C) Invalid. Counterexample when M is true and O and N are false (D) Invalid. Counterexample when O is true and M and N are false
CA	**Geography (HS)**	Which of the following is MOST likely to experience population pressure?	(A) An industrial society with abundant natural resources and large imports of food (B) A society with a highly mechanized agricultural sector (C) A non-ecumene (D) A slash-and-burn agricultural society
CA	**Nutrition**	Why might some biochemical (eg plasma or serum) indices of micronutrient status give misleading results in people with infections or inflammatory states?	(A) Because people who are sick often alter their diets, and may eat less food. (B) Because the accuracy of some laboratory assays may be compromised in samples from people who are sick. (C) Because some metabolic pathways are altered in sick people, which changes their micronutrient requirements. (D) Because an acute phase reaction results in changes in inter-tissue distributions of certain micro-nutrients.

N Examples of Cultural, Geographical and Dialect Knowledge

This section lists some examples of cultural, geographical (or regional) and dialect knowledge that was shared with the annotators to guide them during the annotation process.

Knowledge	Applicable Examples	Non-Applicable Examples

Cultural	(A) Understanding religious customs: For instance, the significance of colored powder during Holi in Hindu culture. (B) Awareness of traditional arts: For instance, the unique styles and techniques of Indigenous Australian art, often featuring dot painting and storytelling. (C) References to liberal/conservative attitudes: We can't assume the notion of liberal is specific to a certain culture or region but it inevitably involves social values and culture. (D) References to philosophy and philosophical concepts, including philosophy of law: Some familiar philosophical concepts fall within critical cultural contexts. Hume's conception of practical reason is a familiar philosophical concept in western culture. Logical fallacies also fall under this category.	(A) Universal scientific principles: Knowledge of gravity or evolution is not exclusive to any particular culture. (B) Principles from the social sciences: The principle of social exchange, that posits that social behavior is the result of an exchange process, is used worldwide. (C) Standardized international sports: The rules and practices of soccer (football) are consistent worldwide. (D) Math questions which do not rely on local references: For example, the formula for the radius of a circle.

Geographical	(A) Natural Landmark Identification: Recognizing and knowing the significance of regional natural wonders like the Grand Canyon in the Southwestern United States or the Great Barrier Reef in Australia. (B) Environmental Awareness: Understanding the impact and importance of regional weather patterns, such as the monsoons in South Asian regions or the hurricanes in the Caribbean. (C) Historical Event Memory: Knowledge of region-specific historical occurrences, such as the Gold Rush in California during the 1850s, which transformed the region's economy and demographics. (D) Awareness of a region-specific natural phenomenon: The Northern Lights, visible in the night skies of Alaska and northern regions. (E) Systems of measurement that are specific to a geographic area: Imperial units are used to measure distance (eg. miles), volume (eg. gallons) and weight (eg. pounds) (F) Laws and regulations: A programmer uses code published online under a Creative Commons Attribution (CCBY) license in a commercial product. This license is specific to the regional geographic area it was created in. (G) Behaviors and preferences of groups in specified areas: These can be noted as both "cultural" and "geographic", as in the exam "Which of the following statements does NOT accurately describe voting behavior in the United States?" voting practices are cultural, and the US is specified as a geographic area.	(A) Global Climate Patterns: Understanding El Niño and La Niña weather phenomena, which occur worldwide and are not specific to any single region. (B) Universal Celestial Bodies: The Sun and the Moon are visible worldwide and do not possess regional specificity. (C) Standardized Geography Terms: Understanding the definition of a peninsula or archipelago is applicable to geographic features globally, not tied to regional knowledge.

Dialect	(A) Regional slang: Using the word “wicked” to mean “very good” in parts of New England, USA. Using the phrase “boot of the car” to mean “trunk” in the UK. (B) Unique idiomatic expressions: The phrase “Bob’s your uncle” in British English, meaning “there you have it” or “that’s all there is to it.” (C) Knowledge of social greetings: The customary handshake and verbal greeting of “Konnichiwa” when meeting someone in Japanese culture. (D) Words or phrases from other languages that are brought into English: as in the sentence “he has that je ne sais quoi” in which je ne sais quoi is borrowed from French	(A) Standardized technical jargon: Medical or legal terminology used internationally within professional fields. (B) Formal literary language: The writings of Shakespeare or Dickens utilize sophisticated language but are not tied to specific dialects. (C) Global brand names: Companies like Nike or Adidas use consistent branding worldwide, avoiding regional vocabulary.

O MMLU Subject Name Mapping

Original Name	Short Name
abstract_algebra	Algebra
anatomy	Anatomy
astronomy	Astronomy
business_ethics	Business Ethics
clinical_knowledge	Clinical
college_biology	Bio (Uni.)
college_chemistry	Chem (Uni.)
college_computer_science	CS (Uni.)
college_mathematics	Math (Uni.)
college_medicine	Medicine (Uni.)
college_physics	Physics (Uni.)
computer_security	Computer Sec
conceptual_physics	Conc. Physics
econometrics	Econometrics
electrical_engineering	Electrical Eng.
elementary_mathematics	Math (El.)
formal_logic	Formal Logic
global_facts	Facts
high_school_biology	Bio (HS)
high_school_chemistry	Chemistry (HS)
high_school_computer_science	CS (HS)
high_school_european_history	EU Hist. (HS)
high_school_geography	Geography (HS)
high_school_government_and_politics	Gov. Politics (HS)
high_school_macroeconomics	Macro econ. (HS)
high_school_mathematics	Math (HS)
high_school_microeconomics	Micro econ. (HS)
high_school_physics	Physics (HS)
high_school_psychology	Psychology (HS)
high_school_statistics	Stats (HS)
high_school_us_history	US Hist. (HS)
high_school_world_history	World Hist. (HS)
human_aging	Human Aging
human_sexuality	Sexuality
international_law	Int. Law
jurisprudence	Jurisprudence
logical_fallacies	Fallacies
machine_learning	ML
management	Management
marketing	Marketing
medical_genetics	Genetics
miscellaneous	Misc.
moral_disputes	Disputes
moral_scenarios	Moral Scenarios

nutrition	Nutrition
philosophy	Philosophy
prehistory	Prehistory
professional_accounting	Accounting (Pro)
professional_law	Law (Pro)
professional_medicine	Medicine (Pro)
professional_psychology	Psychology (Pro)
public_relations	Public Rel.
security_studies	Security
sociology	Sociology
us_foreign_policy	US Foreign Policy
virology	Virology
world_religions	World Religions

Table 10: This table shows the short names assigned to MMLU subjects proposed by (Hendrycks et al., 2020) in Figures 3, 5, 17, 18.