

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

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

<sup>1</sup>Cohere For AI, <sup>2</sup>EPFL, <sup>3</sup>Hugging Face, <sup>4</sup>Mila, McGill University & Canada CIFAR AI Chair, <sup>5</sup>AI Singapore,

<sup>6</sup>National University of Singapore, <sup>7</sup>Cohere, <sup>8</sup>MIT, <sup>9</sup>KAIST, <sup>10</sup>Instituto de Telecomunicações, <sup>11</sup>Instituto Superior Técnico, Universidade de Lisboa, <sup>12</sup>MIT, MIT-IBM Watson AI Lab, <sup>13</sup>Carnegie Mellon University,

<sup>14</sup>CONICET & Universidad de Buenos Aires, <sup>15</sup>Meta AI Research

## Abstract

Cultural biases in multilingual datasets pose significant challenges for their effectiveness as global benchmarks. These biases stem not only from differences in language but also from the cultural knowledge required to interpret questions, reducing the practical utility of translated datasets like MMLU. Furthermore, translation often introduces artifacts that can distort the meaning or clarity of questions in the target language. A common practice in multilingual evaluation is to rely on machine-translated evaluation sets, but simply translating a dataset is insufficient to address these challenges. In this work, we trace the impact of both of these issues on multilingual evaluations and ensuing model performances. Our large-scale evaluation of state-of-the-art open and proprietary models illustrates that progress on MMLU depends heavily on learning Western-centric concepts, with 28% of all questions requiring culturally sensitive knowledge. Moreover, for questions requiring geographic knowledge, an astounding 84.9% focus on either North American or European regions. Rankings of model evaluations change depending on whether they are evaluated on the full portion or the subset of questions annotated as culturally sensitive, showing the distortion to model rankings when blindly relying on translated MMLU. We release **Global-MMLU** , an improved MMLU with evaluation coverage across 42 languages – with improved overall quality by engaging with compensated professional and community annotators to verify translation quality while also rigorously evaluating cultural biases present in the original dataset. This comprehensive **Global-MMLU**  set also includes designated subsets labeled as **culturally sensitive**  and **culturally agnostic**  to allow for more holistic, complete evaluation.

**Global-MMLU** : <https://hf.co/datasets/CohereForAI/Global-MMLU>

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

<sup></sup>First author. <sup></sup>Principal senior advisors.

Corresponding authors: {shivalika, beyza, sarahooker}@cohere.com

---

## 1 Introduction

*I contain multitudes. – Walt Whitman, 1855*

Language cannot be simply reduced to a utilitarian tool, otherwise there would be no reason to have so many diverse ways for saying the same thing or referring to similar concepts. Indeed, language is also a marker of belonging and a repository of cultural knowledge (Labov, 1963; 1986; Karlik, 2023). Today, state-of-the-art generative AI is used around the world and yet evaluation of these systems is primarily conducted using English benchmarks (Zellers et al., 2019; Hendrycks et al., 2020; Suzgun et al., 2022; Zhang et al., 2023b). Where multilingual evaluations are relied upon, these are often simply machine translations of widely adopted English benchmarks (Lai et al., 2023; Üstün et al., 2024).

A pressing question arises: *how can we develop large language models (LLMs) that perform effectively and fairly across the full spectrum of languages and cultures?* The lack of comprehensive evaluation benchmarks for many languages poses a significant obstacle for researchers and practitioners striving to create truly multilingual systems. Often, a common practice is to simply translate English benchmarks into other languages. In this work, we consider the implications of this given one of the most ubiquitous examples – the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2020). Originally compiled using sources in the English language across 57 diverse subject areas such as elementary mathematics, computer science, and law, the dataset is often machine-translated into resources for multilingual assessment, which we collectively term *transMMLU* (Lai et al., 2023; Üstün et al., 2024; OpenAI, 2024; Dubey et al., 2024; Bendale et al., 2024). However, the growing adoption of automatically translated “*as-is*” *transMMLU* as a barometer of global AI progress deserves closer inspection and reflection.

While widely adopted for multilingual evaluations, the multilinguality achieved through the translation of English datasets does not guarantee multiculturality. Evaluating on blindly-translated datasets risks overemphasizing Western-centric concepts and knowledge. Cultural bias can reduce the dataset’s practical effectiveness (and conceptual relevance) as a global benchmark when translated. For example, the original English MMLU dataset contains several subsets which are US-specific, such as examinations in *US History*, *US Accounting*, and *US Law*. Such cultural bias reduces the dataset’s practical effectiveness (and conceptual relevance) as a global benchmark when translated. Furthermore, as these translated datasets become adopted for multilingual evaluation and developers optimize models for performance on *transMMLU* datasets, we risk overfitting to the datasets’ cultural biases and incidentally setting multilingual evaluation standards to be aligned with certain culture paradigms. Second, while machine translation expands language coverage, it also introduces practical evaluation challenges. Translation artifacts known as *translationese* (Bizzoni et al., 2020; Vanmassenhove et al., 2021; Koppel & Ordan, 2011) can be introduced, which causes a breakdown in evaluation quality. Automatic data curation is also known to often exacerbate common data quality issues (Luccioni & Viviano, 2021; Kreutzer et al., 2022; Ferrara, 2023; Caswell et al., 2020).

Our effort to address the above is twofold. We conduct an extensive evaluation to quantify the impact of cultural biases in MMLU on model evaluations to-date *and* contribute improvements to the overall translation quality to solve linguistic qualms. We hire professional annotators to verify translation quality and include improvements from rigorous per-question post-edits as well

---

as human translations. We release the comprehensive improved dataset **Global-MMLU**  for 42 languages:

Amharic, Arabic, Bengali, Chinese, Czech, Dutch, English, Filipino, French, German, Greek, Hausa, Hebrew, Hindi, Igbo, Indonesian, Italian, Japanese, Korean, Kyrgyz, Lithuanian, Malagasy, Malay, Nepali, Nyanja, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Sinhala, Somali, Shona, Spanish, Swahili, Swedish, Telugu, Turkish, Ukrainian, Vietnamese, and Yoruba.

To address regional and cultural biases, we systematically annotate a subset of the original English MMLU to identify questions where correctly answering requires cultural, geographical, or dialect-specific knowledge. We refer to such questions as being *Culturally-Sensitive* (**CS** )<sup>1</sup>, in contrast to questions which do not require this prior knowledge, referred to as being *Culturally-Agnostic* (**CA** ). We evaluate 14 state-of-the-art open-weight and proprietary models from 9 model families, focusing on those known for their high multilingual performance. This enables rigorous evaluation of how such models serve diverse language users and isolates how ranking may be subverted by questions which require primarily Western-centric knowledge. Through extensive evaluations, we consistently find that *cultural sensitivity* has a significant impact on model rankings. Our core contributions can be enumerated as follows:

- **Analysis of MMLU for cultural biases:** We observe that progress on MMLU depends heavily on learning Western-centric concepts. Out of the annotated sample, we found that 28% of questions require specific knowledge of Western cultures. Moreover, for questions requiring geographic knowledge, an astounding 84.9% focus on either North American or European regions.
- **Introducing Global-MMLU** : We release a new multilingual MMLU test set spanning 42 languages, including English. This dataset combines professional translations with post-edits (14 languages), crowdsourced translations (11 languages), and machine translations (16 languages). By integrating this dataset with our cultural bias study, evaluations can now report on both the **CS**  and **CA**  subsets. Additionally, we introduce **Global-MMLU Lite**  that provides a compact but high-quality alternative for multilingual evaluation.
- **Re-evaluation of state-of-the-art models:** We evaluate the impact of the re-annotated dataset on the relative performance of multilingual models. Among the 14 models tested, rankings on **CA**  datasets exhibited an average of 3.4 rank changes and 3.7 position shifts compared to their performance on a uniform subsample of the MMLU dataset (*MMLU Annotated* ). However, **CS**  datasets showed significantly greater variability, with an average of 5.7 rank changes and 7.3 position shifts across all 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, as relying solely on machine-translated data may obscure the true capabilities of models in these contexts. Without access to high-quality human-translated or in-language datasets, the evaluation of low-resource language performance remains uncertain.

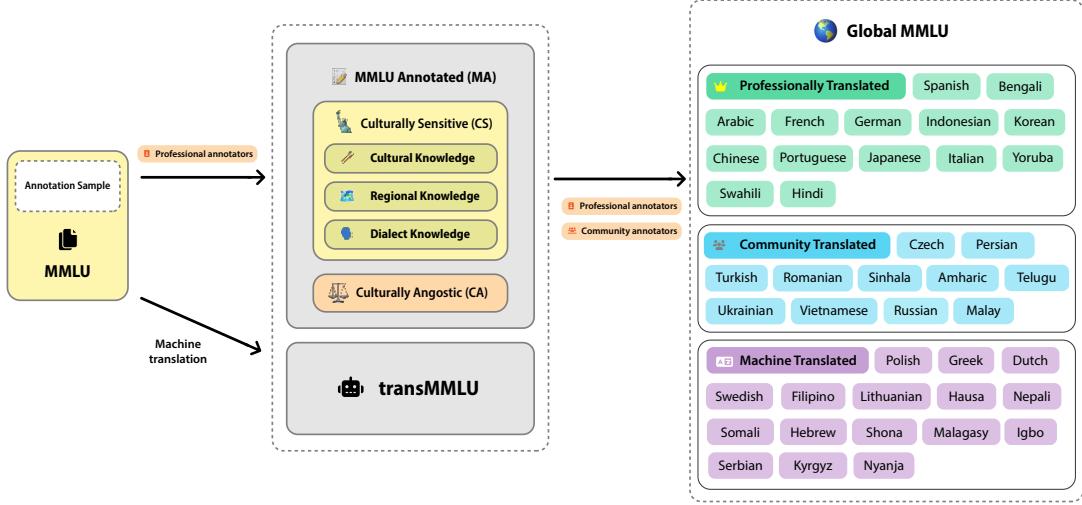


Figure 1: Overview of **Global-MMLU** preparation process. We engage with professional and community annotators to improve the quality of translated MMLU. Additionally, we engage in extensive annotation to provide rich meta-data for what questions in MMLU require *Culturally-Sensitive* (CS ) knowledge such as 1) **Cultural Knowledge** , 2) **Geographical Knowledge** or 3) **Dialect Knowledge** to answer correctly. We release this improved **Global-MMLU** alongside extensive metadata annotations.

Stemming from our comprehensive results, we make the following recommendations for multilingual evaluation of generative models:

- **Report on Global-MMLU** , instead of **translated MMLU**. We recommend prioritizing **Global-MMLU** over translated versions of MMLU for multilingual evaluation. With its extensive language coverage and improvements based on professional annotations and post-edited translations, **Global-MMLU** provides a more reliable and accurate benchmark for assessing model performance across diverse languages.
- **Report performance on culturally-sensitive and culturally-agnostic subsets separately.** Our analysis demonstrates significant variability in model rankings between **CA**  and **CS**  datasets, with **CS**  subsets showing greater variability. This variability, especially pronounced for low-resource languages and smaller models, highlights the importance of evaluating these subsets independently. We recommend reporting performance on **CA**  and **CS**  subsets separately to provide a clearer understanding of model capabilities and better address the unique challenges posed by cultural and linguistic nuances in **CS**  tasks.

## 2 Evaluating cultural bias in MMLU

### 2.1 Data Annotation Process

The goal of this work is to study how cultural biases in translated datasets influence the performance of widely-used multilingual models. To achieve this, we worked with 200 professional compensated and community annotators to review MMLU questions from the original English MMLU dataset to assess its cultural sensitivity. Annotators were presented with a

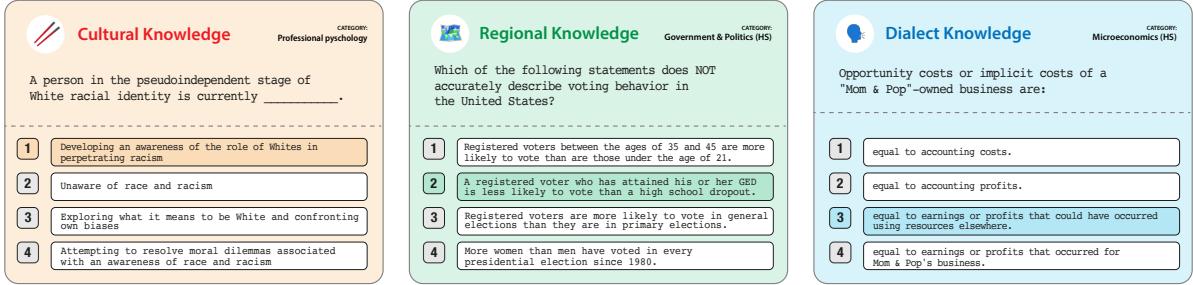


Figure 2: Examples of questions from MMLU dataset labelled as requiring cultural, regional or dialectal knowledge.

representative random sample from each of the 57 exam subjects that compose MMLU (50 per subject), totaling 2,850 samples. This annotated set is referred to as *MMLU Annotated* (MA) throughout the paper. Annotators were asked to identify questions where correctly answering depended upon 1) cultural knowledge 2) geographic knowledge or 3) dialect knowledge .

We provide more context about each of these categories below:

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

Figure 20 in Appendix H 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, 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

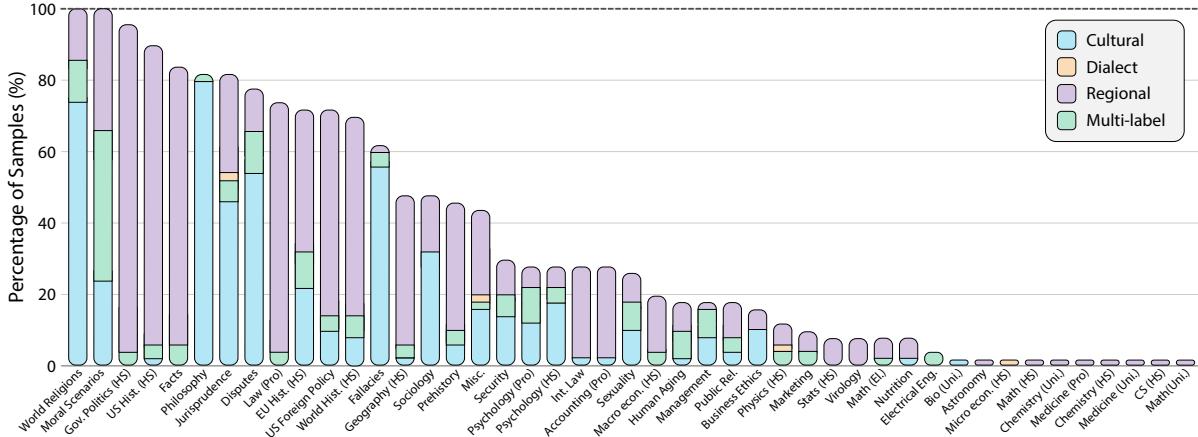


Figure 3: Proportion of samples containing cultural, regional, or dialect-specific references per subject in the MMLU dataset. Notably, all samples in the *World Religions* and *Moral Scenarios* subjects include at least one such reference. Note that 12 subjects did not contain any Culturally-Sensitive CS 🇺🇸 samples and have been excluded from the figure.

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 about the annotators and the annotation process is available in Appendix H.

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 temporal analysis in the Appendix D.

To understand the prevalence of these attributes at an aggregate level, we also assign a label of **Culturally-Sensitive** (CS 🇺🇸) if either **Dialect Knowledge** 🗣, **Cultural Knowledge** 🎨 or **Geographic Knowledge** 🌏 are positively attributed to an example. If none of these properties are present, we deem an example to be **Culturally-Agnostic** (CA 🚫). This enables us to track at an aggregate level the fraction of the entire MMLU that requires CS 🇺🇸 knowledge.

## 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 knowledge of either geographic knowledge 🌏, cultural knowledge 🎨 or dialect knowledge 🗣 – to be answered correctly. Among these, geographic knowledge 🌏 emerges as the most frequently tagged bias, representing 54.7% of all CS 🇺🇸 questions. Cultural knowledge 🎨 follows at 32.7%, while dialect-specific knowledge 🗣 accounts for a mere 0.5% of all questions. Additionally, 10.6% of questions require both cultural and geographic knowledge, and 1.5% involve a combination of all three types of nuanced knowledge.

**Western-centric culture dominates.** Among the samples identified as requiring culturally sensitive CS 🇺🇸, a significant 86.5% were tagged as specific to *Western* cultural knowledge. In contrast, the next closest category, *South Asian* cultural knowledge, accounted for only 4% of the cultural tags. As Figure 4 shows, Latin American, African and Indigenous cultures are

represented by 1.3%, 1.1% and 0.7% of the tags, respectively. This shows performing well on MMLU heavily depends on mastering Western-centric 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 regional knowledge of *Europe*. This concentration indicates that progress on MMLU predominantly reflects knowledge of Western concepts and regions.

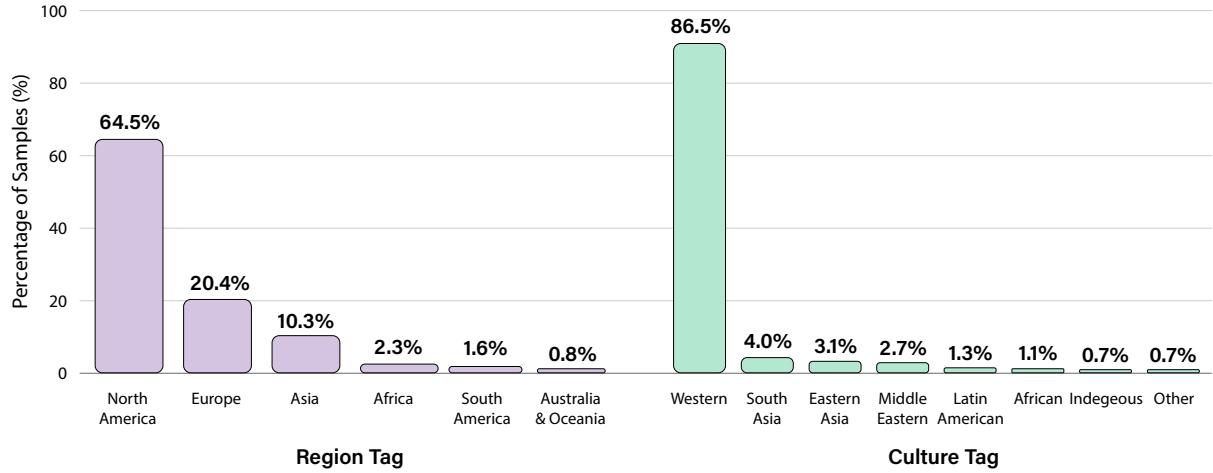


Figure 4: Distribution of region (left) and culture (right) categories found in **CS** dataset. The majority of Region tags (64.5%) correspond to North America, while the majority of Culture tags (86.5%) are classified as Western. We have excluded samples that do not contain any region or culture tags or contain multiple region or culture tags from this figure.

**Culture-specific knowledge is overfit to a few countries.** Figure 5 illustrates the distribution of cultural and regional tags across countries within the **CS** dataset. Our analysis reveals that 73.9% of questions related to Western culture require knowledge about the United States, followed by the United Kingdom at 8%, with smaller contributions from countries like France and Germany. In contrast, Asian culture tags are predominantly associated with India, accounting for 59%, while China and Japan represent only 17.9% each of the questions requiring knowledge of Asian culture. Despite this, the overall representation of Asian cultures remains limited, with only 4.0% of questions pertaining to South Asia and 3.1% to East Asia in the MMLU dataset. Similarly, Middle Eastern culture is largely represented by Iraq (37.5%) and Turkey (25%), yet its overall presence in the dataset is minimal, with just 2.7% of questions addressing Middle Eastern cultural knowledge. These findings highlight the dataset’s strong bias toward the United States, with a significant portion of cultural tags tied to the U.S. For further analysis of the culture–region relationship and detailed country-level insights, see Appendix G.

**Cultural sensitivity varies considerably across subjects.** The MMLU dataset, introduced by Hendrycks et al. (2020), includes 57 subjects spanning four categories: *STEM*, *Humanities*, *Social Sciences*, and *Other*. From the *Other* category, we selected relevant subjects and further categorized them into *Medical* (Chen et al., 2023) and *Business*. Additional details about this categorization are provided in Appendix B.

Figure 6 illustrates the data distribution for the **CA** subset, revealing significant variation

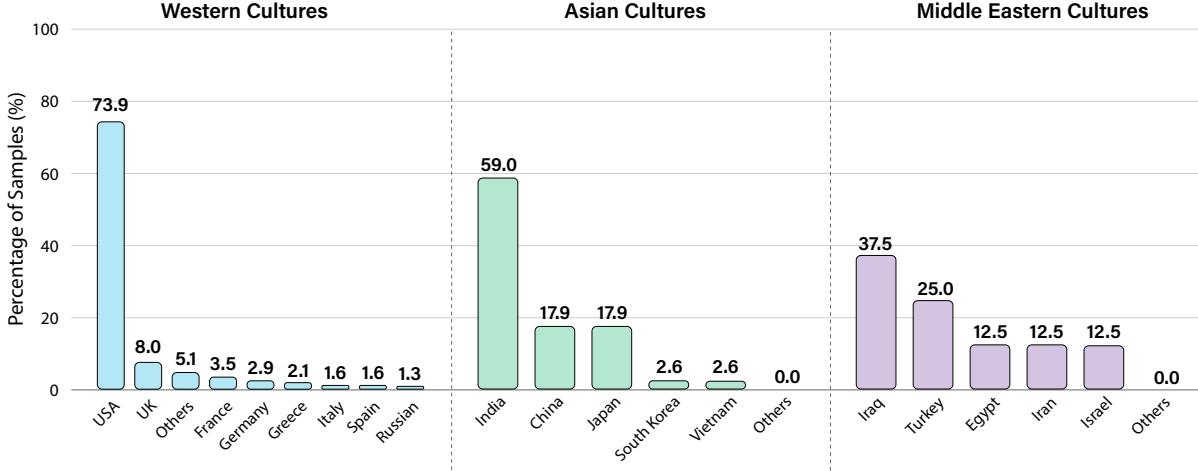


Figure 5: Distribution of cultural and regional tags across countries in the **CS** dataset. The percentages indicate the representation of each country within the dataset. We have excluded samples that do not contain any country tags or contain multiple country tags from this figure.

in cultural and regional references between different MMLU subjects and subject categories. Questions from categories in *Humanities* and *Social Sciences* frequently required cultural or regional knowledge, while those from the *STEM* and *Medical* categories generally did not. Overall for *Humanities*, 68% of all questions were tagged as **CS**. However, this bias was even more pronounced for certain subjects within *Humanities*. Notably, more than 80% of samples for subjects like Philosophy, Moral Scenarios<sup>1</sup>, High School US History, and High School Government and Politics were deemed **CS**. Within the *STEM* category, only 30 out of 950 samples (3.15%) were identified as **CS**, and for subjects such as Clinical Knowledge, Computer Security, and Econometrics all question examples were classified as **CA**. These findings, detailed in Figure 6, unsurprisingly reveal that certain subjects inherently exhibit more cultural or regional biases. We provide examples of MMLU questions annotated as **CS** (Culturally Sensitive) and **CA** (Culturally Agnostic) in the Appendix J.

**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 23 and 24 in Appendix H.2.

**Characteristics of CS versus CA subsets.** Our extensive annotation process resulted in two aggregated annotated subsets of MMLU: **CS**, which includes all questions labeled as requiring dialect knowledge 🗨, cultural knowledge 📚, or geographic knowledge 🌎 to answer

<sup>1</sup>Morals might share universal truths and moral decisions may be well-defined given an underlying belief system, but this does not seem to be the case in this scenario. That is, we observe that Moral Scenarios in MMLU are geared towards Western Culture, and therefore **CS** knowledge, as it specifies “moral standards in the US” in the instruction.

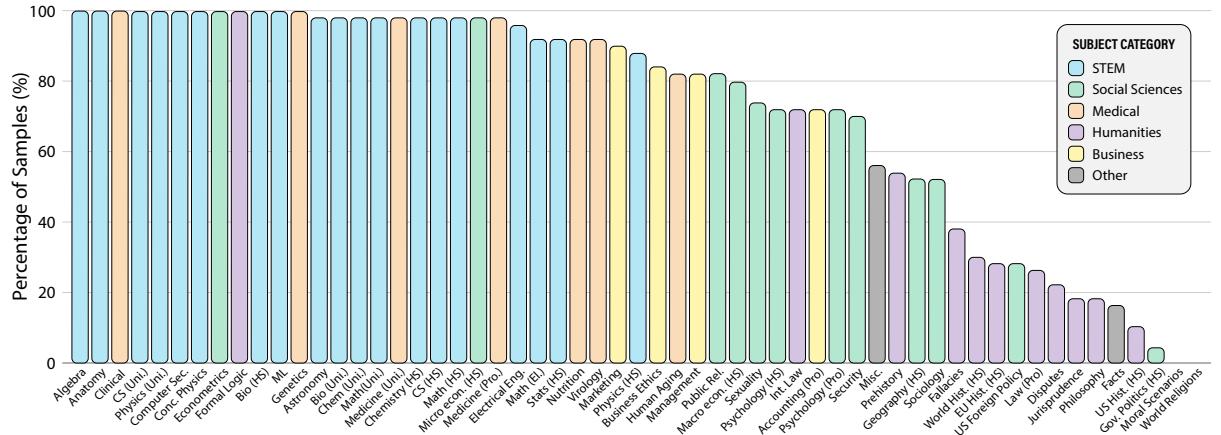


Figure 6: Proportion of samples retained per subject, after excluding those requiring cultural, geographic and dialectic knowledge (selected based on majority agreement).

	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 1: 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.

correctly, and **CA $\ddot{\text{S}}$** , comprising questions that do not require knowledge from these categories. Table 1 provides a detailed breakdown of the number of subjects and samples in the **CS $\ddot{\text{U}}$**  and **CA $\ddot{\text{S}}$**  subsets.

We observe notable differences in subject distribution between the **CA** and **CS** subsets, leading to shifts in category representation. For instance, while questions from the *Social Sciences* category make up 21.1% of the MMLU Annotated , a uniformly balanced subsample of the original MMLU, they are over-represented in **CS**, accounting for 26.3% of all questions requiring **CS** knowledge. Conversely, questions from the STEM category, which contribute 33.3% of the MMLU Annotated , are under-represented in **CS**, making up only 2.9% of all questions identified as requiring **CS** knowledge. These shifts reflect how the nature of the **CS** subset emphasizes cultural and contextual knowledge over technical or scientific content.

Overall, the proportions of STEM, Medical, and Business categories increase in the **CA** subset due to their globally relevant content. Conversely, Humanities and Social Sciences are over-represented in the **CS** subset compared to the original MMLU, as these fields frequently include cultural or regional references. These findings are critical to the model evaluations in Section 4,

---

illustrating how cultural references in MMLU influence dataset composition and, ultimately, model performance.

### 3 Introducing Global-MMLU

To date, many multilingual evaluations have relied on translated MMLU with the most widely adopted *existing multilingual MMLU translation* dataset being translated into 26 languages using ChatGPT<sup>2</sup> supported by GPT-3.5 (Lai et al., 2023). We release an improved **Global-MMLU**  benchmark which is both of higher quality and also supports analysis on both **CS**  and **CA**  subsets.

Here, we improve quality by incorporating professional edits and translations from native speakers for a subset of languages and expanding coverage to 42 languages. We achieve this through a combination of paid professional translations, community contributions, and higher-quality machine translation. This effort involved professionally compensated annotators for four gold-standard languages and a broader pool of community annotators who contributed to translations in 11 additional languages. Where available, we also included the professional human translations from the MMMLU dataset<sup>3</sup> for 14 languages. We rely as much as possible on human-verified translations to ensure that the translations are reliable and minimize the biases introduced, specifically *translationese* which might be more pronounced in Machine Translation (Bizzoni et al., 2020; Vanmassenhove et al., 2021; Koppel & Ordan, 2011). Alongside these quality improvements through human verification, we include the metadata for the **CS**  and **CA**  annotations developed in the previous sections to allow for analysis on all subsets of data. Below, we provide further details about our efforts to improve the quality of MMLU and engage compensated human annotators in translating and verifying quality as well as identifying the **CS**  and **CA**  subsets.

#### 3.1 Translation Process

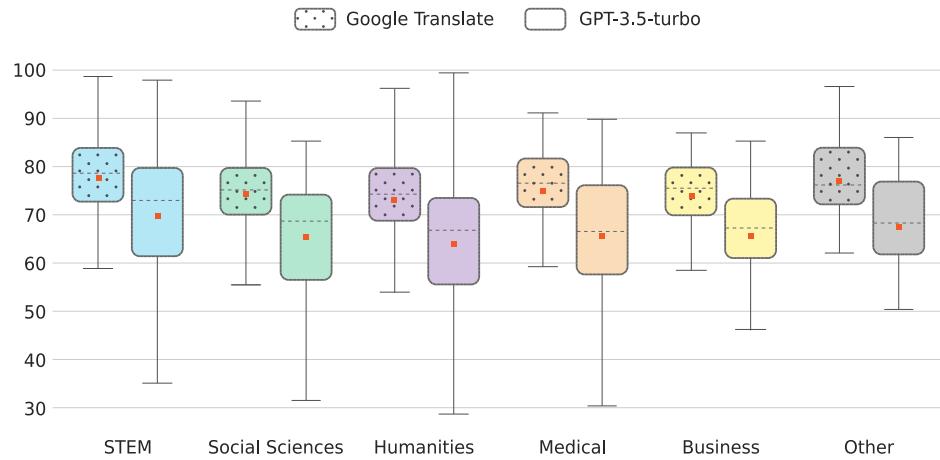


Figure 7: ChrF++ scores for Google Translate and GPT-3.5-Turbo

<sup>2</sup><https://chat.openai.com/chat>

<sup>3</sup><https://huggingface.co/datasets/openai/MMMLU>

---

We first translated the English MMLU dataset into 41 languages using the Google Translate API.<sup>4</sup> Despite its cost, we chose to use Google Translate because comprehensive evaluations spanning 102 languages (Zhu et al., 2024) demonstrate that Google Translate significantly outperforms alternatives such as NLLB (NLLB-Team et al., 2022), GPT-4, and ChatGPT, on low-resource languages (Robinson et al., 2023). Recent work (Kocmi et al., 2024) have shown that LLMs have begun to surpass popular online translation tools like Google Translate for machine translation on specific high-resource languages. However, given that there is a known tendency for models to favor their own generations (Panickssery et al., 2024; Shimabucoro et al., 2024), we decided to use Google Translate for every language in order to avoid introducing bias into model evaluations. To empirically validate this choice, we compared Google Translate’s outputs with translations performed by GPT-3.5-turbo, which had been previously used to translate the MMLU dataset (Lai et al., 2023). As shown in Figure 7, we find that Google Translate achieved higher ChrF++ scores (Popović, 2017) across all subjects and lower deviation in performance across languages, consistent with the findings of previous research (Popović, 2017) about its superiority in translation quality. Following the translation process, native speakers reviewed and edited the translations to ensure accuracy and fluency, thereby enhancing global representation. These edits were performed by two types of annotators: *professional annotators* and *native community 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 where necessary. We refer to this set of translation as our “*Gold Set*”. We include more details about compensated annotation process in the Appendix H.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 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 minimum 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<sup>5</sup> 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

---

<sup>4</sup><https://cloud.google.com/translate>

<sup>5</sup><https://github.com/argilla-io/argilla>

in Figure 21 in Appendix I.

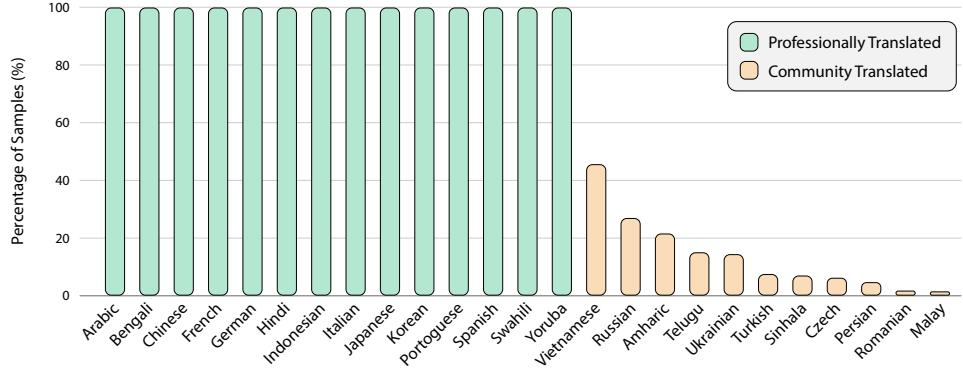


Figure 8: Percentage of Human-Translated Samples in MMLU Annotated

**MMMLU Translations.** As detailed in the OpenAI-o1 system card,<sup>6</sup> MMMLU<sup>7</sup> is a professionally human-translated dataset released by OpenAI in 14 languages. To maximize the inclusion of human-translated content in **Global-MMLU** , we incorporated this dataset wherever possible. Since MMMLU overlaps with our *Gold Set*, we utilized the remaining 10 languages: *Bengali*, *Chinese*, *German*, *Indonesian*, *Italian*, *Japanese*, *Korean*, *Portuguese*, *Swahili*, *Yoruba* from this dataset.

Figure 8 highlights the number of samples edited by professional annotators and community contributors. A total of 7,565 edits were made, accounting for 36.9% of the samples reviewed. On average, professional annotators edited 789 samples per language (38.5% of the total) in the *Gold Set*, while community contributors edited 362 samples per language (17.7% of the total). It is important to note that the differences in edit rates likely reflect variations in time and resources available to professional versus community annotators, and cannot be interpreted as differences in translation quality across languages. Additional analyses of question and answer lengths, as well as edit distances across subject categories, are presented in Appendix I.

### 3.2 Data Composition of Global-MMLU

**Global-MMLU** is our comprehensive test set encompassing all 14K samples from MMLU across 42 languages (including English), resulting in a total of 589,764 samples, created by integrating multiple data sources, including human-translated datasets, machine translations, and the original English MMLU. Throughout the Model Evaluations section, we also report on different subsets of **Global-MMLU** , described as follows:

**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.

<sup>6</sup><https://openai.com/index/openai-o1-system-card/>

<sup>7</sup><https://hf.co/datasets/openai/MMMLU>

---

**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.

**Global-MMLU Lite** ✨. This 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 C.

## 4 Model Evaluations

One of the key findings from Section 2.2 is that MMLU presents severe biases towards CS 🎨 knowledge. In this section, we seek to understand how these biases may have impacted evaluation of open-weights and closed models. To do so, we measure changes to model rankings on 3 subsets of data: *Global-MMLU Annotated* 📝, *Global-MMLU Culturally-Agnostic (CA)* ⚖️ and *Global-MMLU Culturally-Sensitive (CS)* 🎨. By comparing model performance across these three subsets, we aim to address the following questions: (1) *How do models perform on the MMLU test set when it includes culturally-sensitive samples?* and (2) *How do models perform on samples that do not require specific contextual knowledge, ensuring consistent and fair evaluations across different languages and regions?*

### 4.1 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 E. *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.*

**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 (i.e., GPT-4o and Claude-Sonnet 3.5), 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.

**Languages.** We categorize the languages into two main groups for reporting the results. The first group consists of human-translated data only, which covers 10 languages from OpenAI’s human-translated MMLU test set and 4 additional languages from our professionally translated set. The second group contains all our data (combining professional, community and machine translations), organized by language resource availability (high-resource, mid-resource, and low-resource languages as defined by Joshi et al. (2019) and categorized in (Singh et al., 2024). We report results for each of these categories. The **high-resource** languages are Arabic, Chinese, Czech, Dutch, English, French, German, Hindi, Italian, Japanese, Persian, Polish, Portuguese, Russian, Spanish, Swedish, Turkish, Vietnamese, **mid-resource** languages are Bengali, Filipino, Greek, Hebrew, Indonesian, Korean, Lithuanian, Malay, Romanian, Serbian, Ukrainian and **low-resource** languages are Amharic, Hausa, Igbo, Kyrgyz, Malagasy, Nepali, Nyanja, Shona, Sinhala, Somali, Swahili, Telugu, Yoruba.

## 4.2 Results

**Evaluations on Human-Translated Data.** To assess the performance of models on high-quality, human-translated data, we conducted evaluations using the subset of 14 languages with human-translated data. The analysis focuses on both the **CA** and **CS** subsets to explore how models handle tasks with and without cultural context.

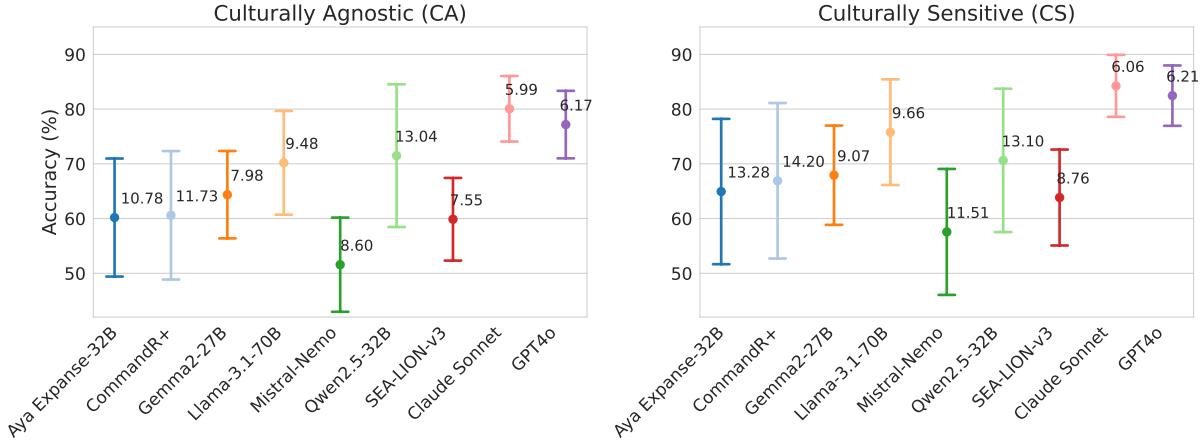


Figure 9: Model evaluations on **CA** and **CS** data samples on **human-translated 14 languages**. The error bars indicates the standard deviation across languages.

We evaluated 14 models from 9 different model families, including 2 closed-source models. Figure 9 presents the results aggregated across 14 languages. We note that the focus of this evaluation is not to compare model performances directly but to analyze their behaviors on **CA** and **CS** datasets. Direct comparisons between proprietary models and open-weight models are not feasible due to significant differences in model sizes (although we note that the parameter sizes of proprietary models have not been officially disclosed) and different evaluation methods. Nonetheless, the results show that closed-source proprietary models, such as GPT-4o and Claude 3.5 Sonnet, consistently outperform smaller open-source models. Interestingly, the performance gap between these models is narrower on **CS** datasets than on **CA** 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 10 shows that overall, models perform better on the **CA**  portion.

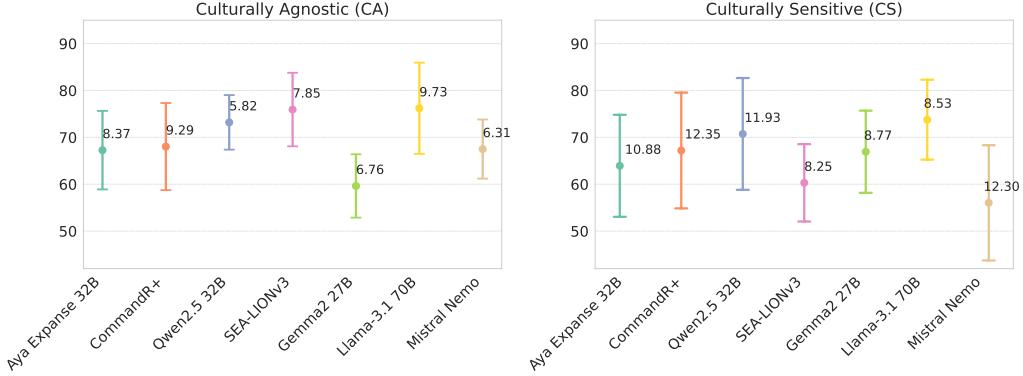


Figure 10: Model evaluations on **CA**  and **CS**  samples in **Global-MMLU Lite** . Error bars indicate standard deviation across languages.

**Performance on CS samples is higher but presents more variance** Another key observation is that the average accuracy across all models is higher on **CS**  datasets compared to **CA**  datasets. This trend can be attributed to the nature of the **CS**  samples, which are predominantly drawn from Social Sciences and Humanities domains where models generally perform better. In contrast, **CA**  datasets include more challenging categories, such as Medical and STEM, as illustrated in Figure 15.

However, the standard deviation in performance across languages is higher for **CS**  data than for **CA**  data for all models. This can be attributed to several factors: culturally sensitive tasks are inherently more challenging and require deeper contextual understanding, making them more susceptible to variations in translation quality. Nuanced cultural, regional, or dialectal references in **CS**  tasks often amplify this sensitivity, as differences in how these references are translated can affect model performance. Furthermore, many large language models are trained predominantly on data from high-resource or Western cultures, leading to biases that favor these contexts and cause inconsistencies when applied to less-represented cultures.

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.

**Evaluations Across High-, Mid-, and Low-Resource Languages.** To analyze model performance across languages with varying resource availability, we evaluated the models on **CA**  and **CS**  subsets, categorized into  high-,  mid-, and  low-resource languages. This evaluation provides insights into how models handle linguistic diversity and cultural nuances across different resource levels.

**Performance degrades on low-resource languages with higher variability** For both **CA**  and **CS**  datasets,  high-resource languages consistently achieve the highest average accuracy across all models. As expected, performance declines significantly for  low-resource languages due to the limited availability of high-quality training data, which hinders model generalization. This decline is accompanied by an increase in performance variability, with

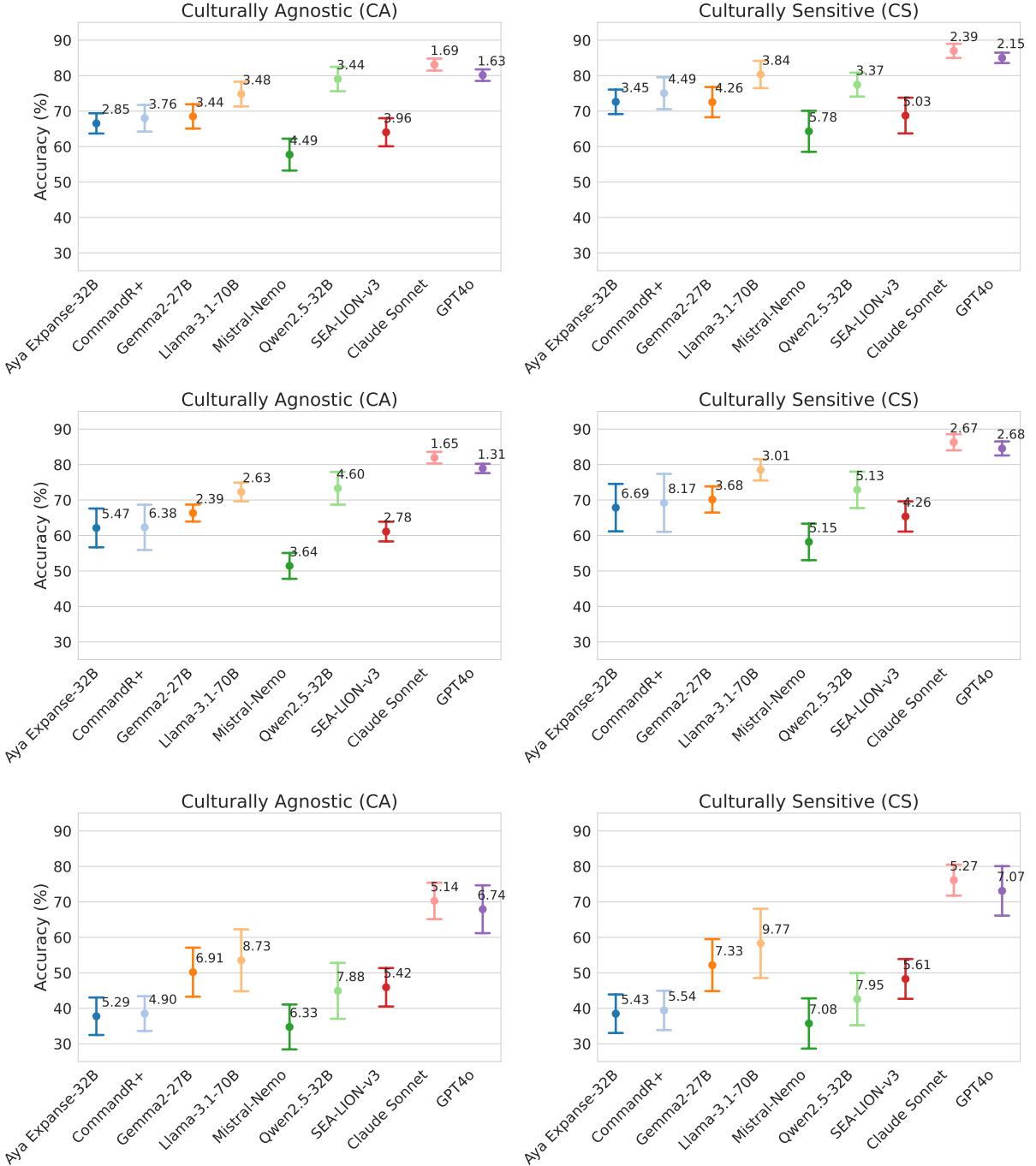


Figure 11: Model evaluations on (Top) ● high-resource , (Mid) ○ mid-resource and (Bottom) ○ low resource data samples for **CA**⚖️ and **CS**��️ subsets.

the standard deviation rising 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**��️ datasets. For ○ mid-resource languages, these values increase to **3.42** and **4.6**,

---

respectively. ○Low-resource languages exhibit significantly higher standard deviations, with averages rising to **6.37** on **CA** datasets and **6.78** on **CS** datasets. These represent increases of 98% and 75% compared to high-resource languages, highlighting the greater variability and sensitivity in low-resource settings. This increased variability in model performances highlights the challenges of culturally sensitive tasks, which demand a nuanced understanding of regional or dialectal references. Across all level of resourcefulness, performance on **CS** shows higher variability than **CA**.

**Model Rank Changes.** This section explores how model performance rankings differ between **CA** and **CS** datasets, calculated relative to their ranks on **MA**, across multiple languages. Table 2 highlights rank changes for **human-translated** languages, organized by resource level: ●high-resource, ○mid-resource, and ○low-resource. These rankings offer valuable insights into how dataset type, resource availability and model size impact model performances. Comprehensive rankings for all languages are available in Table 6 and Table 7 in Appendix F.1.

The rank changes reveal three key findings:

**1) Models perform differently across **CA** and **CS** datasets, with the latter showing greater variation.** Rankings on **CA** datasets exhibit minimal changes. For instance, Italian, Japanese, and Portuguese show no rank changes, while Arabic and French each experience only two shifts, each by one position.

On the other hand, model performance varies significantly on **CS** datasets. Chinese and Hindi emerge as the most sensitive languages to culture-specific knowledge, with models showing both increases and decreases in rankings. Similar variations are evident in French, German, Italian, Japanese, and Portuguese. Notably, models from the Aya Expanse and CommandR families tend to show positive trends on **CS** datasets, particularly for these languages. On average, across all languages, **CA** datasets see 3.4 rank changes and 3.7 position changes, whereas **CS** datasets experience markedly higher volatility, with 5.7 rank changes and 7.3 position changes.

**2) The difference between performances on **CA** and **CS** datasets are less on 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, ranking changes are more pronounced, with an average of 6.8 rank changes and 9.1 position shifts. In contrast, ○mid-resource languages display moderate variability. While *small models* face slightly greater fluctuations on **CS** datasets, their performance on **CA** datasets remains more consistent. For ○mid-resource languages, the average rank changes are 3.7 on **CA** and 4.7 on **CS**, with corresponding position changes of 4.7 and 4.9. Among the three resource groups, ○mid-resource languages show the smallest difference between **CA** and **CS** performance.

○Low-resource languages show an increase in the difference between **CA** and **CS** rank changes compared to ○mid-resource. Average rank changes are 3.3 on **CA** datasets and 3.7 on **CS**, with position changes rising to 5.7 on **CA** and 7.9 on **CS**. Notably, this group also experiences the largest rank changes. Table 3 highlights the most significant changes across all languages, including **rank shifts of up to 5 positions** for Malagasy, and **13 ranking changes** for the models on Ukrainian. These findings underscore how resource levels amplify rank changes, even within **CA** datasets.

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	-	-	-
● Chinese	CA	-	-	↓1	-	↑1	-	-	-	-	↑1	-	↓1	↑1	-
● English	CA	-	-	↑1	↑1	↑1	↑2	↑1	-	↓1	↑1	↑1	↑1	↑1	↓1
● French	CA	-	↑1	-	-	-	-	-	-	↓1	-	↓1	↓1	-	-
● German	CA	-	↓1	-	↓1	-	↑1	-	-	↑1	-	-	-	-	-
● Hindi	CA	-	↑1	↓2	↓1	↑1	-	-	-	↑1	↓3	↓1	↑1	↑1	↓1
● Italian	CA	-	-	-	-	-	-	-	-	↑1	-	-	-	-	-
● Japanese	CA	-	-	-	-	-	-	-	-	↓1	↓1	↓1	↓1	-	-
● Portuguese	CA	-	-	-	-	-	-	-	-	↑1	↓2	↓1	↓1	-	-
● Spanish	CA	-	↓1	-	↓1	-	↑1	-	-	↑1	↓3	↓1	-	-	-
○ Bengali	CS	-	↑1	-	-	-	-	-	↓1	↓1	-	-	-	-	-
○ Indonesian	CS	-	-	↓1	↓1	↓1	↑1	-	-	↑2	-	-	-	-	-
○ Korean	CS	↓1	↓1	↓1	-	-	↑1	↑1	-	-	↑1	-	-	-	-
○ Sinhala	CS	-	↑1	-	-	-	-	↓3	-	↑2	-	-	-	-	-
○ Swahili	CS	-	↓1	-	-	-	-	↑1	-	-	-	-	-	↓1	↑1
○ Yoruba	CS	-	↑1	↓2	-	↓1	-	-	-	↑2	↑1	↓1	-	-	-

Table 2: 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.

**3) Model size influences performance variations.** We analyzed performance variations across three model groups, as defined in the *Model* section (excluding closed-weight models due to unknown sizes). Our findings highlight distinct trends for large, mid-size, and small models:

*Large models* demonstrate higher consistency across datasets and resource levels. The average rank changes for large models are minimal, at 0.21 for **CA** and 0.67 for **CS**. The maximum position shift for models in this group is 3 while it can be 5 for *small-models*. This consistency reflects their robustness and higher capacity to generalize across diverse datasets.

*Mid-size models*, on the other hand, show much bigger variability. Their average rank changes are

0.33 for **CA** and 1.97 for **CS**, indicating they are more sensitive to dataset characteristics, particularly in the **CS** datasets that require cultural knowledge.

*Small models* exhibit the smallest difference in rank change between **CA** and **CS** (0.35 and 0.45, respectively). However, this apparent stability stems from their weaker overall performance across both datasets. For instance, the average accuracy for small models 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.

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	🕒	↓1	↓1	-	-	-	↑1	-	-	↓1	↑2	-	-	-	-
🕒Ukrainian	🕒	-	↑1	-	↓1	↓1	-	-	-	-	↑1	-	-	↑1	↓1
🕒Malagasy	🕒	-	↓1	-	↑4	↑1	-	-	-	-	↑1	↓1	↓5	-	-
🕒Shona	🕒	-	-	-	-	↓1	-	-	-	-	↑1	-	↓1	↑1	-
		↑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.

Overall, we can conclude that dataset characteristics significantly impact model performance across all model sizes, though the magnitude of variability differs. Across all groups, models demonstrate sensitivity to the diverse cultural and linguistic nuances present in **CS** datasets, with performance variations reflecting their capacity to adapt to dataset-specific nuances. This pattern holds consistently, regardless of model size, though the magnitude of variability differs.

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 as shown in Table 4.

**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 12 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.

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.

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 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.

## 5 Related Work

### 5.1 Multilingual Knowledge Evaluation

As the MMLU benchmark has become a standard for evaluating LLMs (Beeching et al., 2023; OpenAI, 2024; Dubey et al., 2024; Üstün et al., 2024; Aryabumi et al., 2024), addressing its limitations and introducing enhancements are essential to maintaining high evaluation standards. For English, Gema et al. (2024) manually re-annotated 3K questions across 30 MMLU subjects to identify quality or problematic questions and released it as MMLU-redux. Wang et al. (2024) introduced an extended version of this dataset, MMLU-Pro, which adds more challenging,

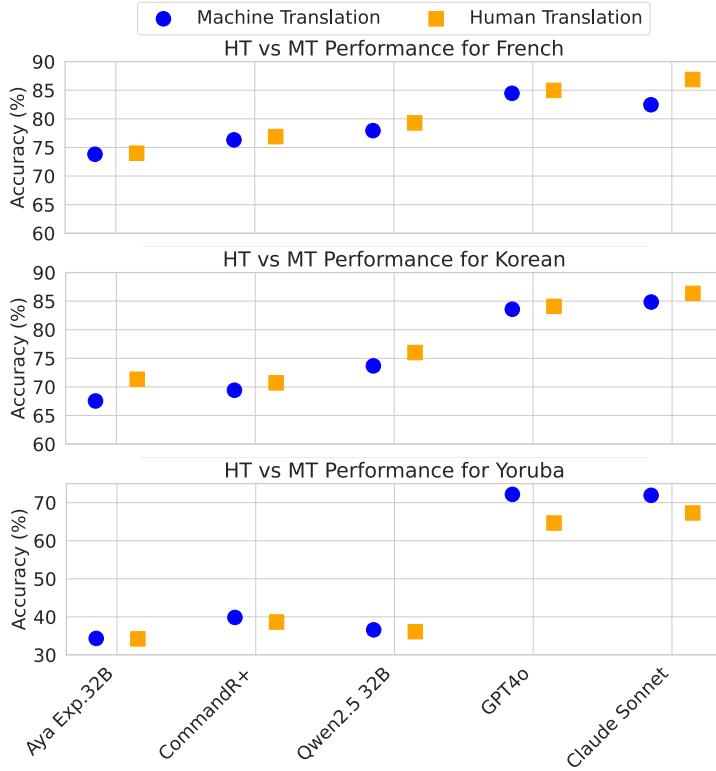


Figure 12: Comparison of model performance on *human-translated* and *machine-translated* CS in French, Korean, and Yoruba.

reasoning-focused questions and expands the answer choice set from four to ten options. MMLU-Pro+ extends the previous work by incorporating questions with multiple correct answers across diverse domains and evaluating higher-order reasoning in LLMs (Taghanaki et al., 2024). While these efforts enhance the difficulty and diversity of tasks, they remain restricted to English alone.

Language-specific variants of comprehensive multiple-choice exam benchmarks are typically centered around a single language. Examples include ArabicMMLU (Koto et al., 2024), CMMLU (Li et al., 2024a), IndoMMLU (Koto et al., 2023), ThaiExam (Pipatanakul et al., 2023), , Turkish-MMLU (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 VNHSQE (Dao et al., 2023) covering Arabic, Chinese, Indonesian, Thai, Turkish, Persian, Korean, and Vietnamese, respectively.

There have been multiple efforts to design and construct evaluation datasets that cater to multilingual settings. AGIEval is a compilation of human-centric standardized exams to assess language model performance in English and Chinese (Zhong et al., 2023). BEnQ is similar but for English and Bengali (Shafayat et al., 2024). EXAMS is a multilingual high school examination collection covering 16 languages (Hardalov et al., 2020). M3EXAMS is a multimodal multilingual benchmark supporting 9 languages with three educational levels (Zhang et al., 2023a). Both evaluation sets process exams on various topics in different countries and build per-language benchmarks. These initiatives strive to evaluate the performance of language models across various languages; however, they often support a small number of languages and lack a consistent,

---

standardized framework for direct comparison between languages. We note recent work IN-CLUDE as an exception to this as one of the most extensive evaluation benchmarks, compiled from local exams across various countries and languages, covering 44 languages (Romanou et al., 2024).

To enable evaluation across a wider range of languages, efforts have also been made to translate the MMLU dataset into multiple languages. Lai et al. (2023) use ChatGPT to translate the English MMLU dataset into 26 languages. However, the quality of translations produced by ChatGPT can vary significantly across different languages and is not always reliable (Robinson et al., 2023). More recently OpenAI released MMMLU by translating MMLU into 14 languages using professional human translators, and we incorporate this high-quality dataset into our benchmark.

## 5.2 Culturally-aware Evaluation

Recent research has increasingly focused on examining the cultural alignment of LLMs. Studies such as Arora et al. (2022) and Cao et al. (2023) have explored LLMs’ ability to understand cross-cultural differences in values and beliefs. To ensure accurate cross-cultural and cross-linguistic representation, SEA-HELM<sup>8</sup> (previously known as BHASA (Leong et al., 2023))<sup>9</sup> is an evaluation suite which emphasizes Southeast Asian languages and contains a variety of tasks, including manually handcrafted linguistic diagnostics as well as manually translated and validated SEA-IFEval and SEA-MTBench. Wang et al. (2023) and Masoud et al. (2024) demonstrate that LLMs often reflect values and opinions aligned with Western culture, a trend that persists across multiple languages. Additionally, benchmarks like those introduced by Naous et al. (2024) and Rao et al. (2024) aim to measure cultural biases in LLMs, while Ventura et al. (2024) investigates cultural biases within text-to-image diffusion models, proposing a comprehensive suite of cultural evaluation techniques. Aakanksha et al. (2024) studied aligning language models balancing dual objectives: addressing and optimizing for a non-homogeneous set of languages and cultural preferences based upon annotations from professional multilingual annotators while minimizing both global and local harms. Some studies focus on specific cultural aspects, such as Myung et al. (2024), Magomere et al. (2024), and Montalan et al. (2024), which evaluate LLMs’ understanding of everyday cultural knowledge across diverse cultures and regions.

In addition, several studies have explored evaluating multilingual visual language models (VLMs). PangeaBench is a holistic evaluation suite encompassing 14 pre-existing datasets covering 47 languages (Yue et al., 2024). Romero et al. (2024) presents CVQA, a culturally diverse multilingual Visual Question Answering benchmark that includes culturally-driven images and questions across 30 countries and 31 languages. Vayani et al. (2024) introduces a multimodal benchmark including culturally diverse images paired with text across 100 languages.

Numerous studies have also explored the role of pre-training in shaping the cultural biases present in LLMs. For example, Chen et al. (2024) examines the impact of native versus translated data on LLM instruction tuning and evaluation. Their findings reveal that models fine-tuned with native instructions typically outperform those trained using translated data. Similarly, Choenni et al. (2024) investigates the reliability of machine translation as a substitute for human translation in

---

<sup>8</sup>An acronym for SouthEast Asian Holistic Evaluation of Language Models.

<sup>9</sup><https://leaderboard.sea-lion.ai>

---

large-scale multilingual evaluations, highlighting its effectiveness across a diverse set of languages. Üstün et al. (2024) released the Aya-101 model and focused on in-language prompting and using a comprehensive dataset of human-written data for instruction tuning large language models across 114 languages to reflect local culture and preferences (Singh et al., 2024). Additionally, significant efforts have been made to incorporate knowledge from various cultures into LLMs to achieve broader cultural alignment. For instance, Li et al. (2024b) proposes a cost-effective fine-tuning strategy to embed cultural differences into LLMs, facilitating better representation and understanding of global cultural nuances. Meanwhile, AlKhamissi et al. (2024) introduces “Anthropological Prompting” a novel method that employs anthropological reasoning to enhance the cultural alignment of LLMs.

### 5.3 Participatory Open Science Projects

Participatory research empowers diverse communities to actively contribute to the research process, ensuring that outcomes are inclusive, contextually relevant, and address real-world needs. Previous participatory research efforts have primarily focused on specific regions or tasks such as translation, character recognition, audio segmentation, and transcription. For instance, Clanuwat et al. (2018) addressed the challenge of reading and understanding Kuzushiji, an old cursive style of Japanese writing no longer commonly used. Another notable example of culturally diverse data collection is MaRVL (Multicultural Reasoning over Vision and Language; Liu et al., 2021), where native speakers of five typologically, genealogically, and geographically diverse languages (*Indonesian, Swahili, Tamil, Turkish, and Mandarin Chinese*) contributed images reflecting their cultures. Professional linguists fluent in these languages then wrote captions for the images. However, MaRVL’s dataset is relatively small, with fewer than 8,000 data points, limiting its use to evaluation purposes. Similarly, Hernandez Mena & Meza Ruiz (2022) developed eight open-access resources for Mexican and Latin American Spanish by establishing a social service program where students voluntarily contributed to tasks like audio segmentation and transcription. Notably, these efforts are largely concentrated on image and speech, unlike our work, which focuses on text. Cañete et al. (2020) spearheaded the collection of a Latin American Spanish dataset to train a language model. Guevara-Rukoz et al. (2020) explored the development of a crowd-sourced corpus for Latin American Spanish dialects to address resource scarcity for these languages. Masakhane utilized a participatory research framework to curate NLP datasets and build models for several underrepresented African languages (V et al., 2020; Adelani et al., 2021; 2023). Aligned with the goals of having a participatory framework and open-access resources, Project SEALD,<sup>10,11</sup> a collaboration between AI Singapore and Google Research, pioneered multilingual data collection for Large Language Models (LLMs) in Southeast Asia (SEA). The output of this project continues to contribute to the development of open-source multilingual models in this region, namely SEA-LION<sup>12</sup> and its derivatives, such as WangchanLion (Phatthiyaphai-bun et al., 2024) and Sahabat-AI.<sup>13</sup> Similarly, the NusaCrowd initiative by Cahyawijaya et al. (2023) focused on aggregating and standardizing data sources for Indonesian languages. The ongoing SEACrowd project<sup>14</sup> represents a similar effort, aiming to standardize data resources for all Southeast Asian languages (Lovenia et al., 2024). The Aya Initiative, through a global community effort of 3,000 contributors, collected instruction data in 114 languages, fostering lin-

<sup>10</sup>An acronym for Southeast Asian Languages in One Network Data.

<sup>11</sup><https://aisingapore.org/aiproducts/southeast-asian-languages-in-one-network-data-seald/>

<sup>12</sup><https://sea-lion.ai>

<sup>13</sup><https://sahabat-ai.com>

<sup>14</sup><https://github.com/SEACrowd>

---

guistic diversity and inclusivity to create one of the largest multilingual datasets for advancing state-of-the-art language models (Singh et al., 2024; Üstün et al., 2024).

## 6 Conclusion

We evaluate the cultural biases present in MMLU and find that 28% of all questions require culturally-sensitive knowledge. In particular, progress on MMLU depends heavily on learning Western-centric concepts. For questions requiring geographic knowledge, the vast majority focus on North America and Europe. This cultural bias remains in translated variants of MMLU that are widely used for multilingual LLM evaluation, which reduces the dataset’s practical effectiveness as a global benchmark and risks over-indexing evaluations on Western-centric idioms and knowledge.

We examine the impact of translation artifacts and cultural bias on multilingual model rankings. 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.

Finally, we evaluate a large group of state-of-the-art open-weight and proprietary models to understand performance differences on both these subsets. We find that model rankings change depending on whether models are assessed on culturally-sensitive or culturally-agnostic subsets, highlighting that progress on translated MMLU is insufficient as an indicator of performance. Instead, we recommend evaluations for multilingual reports on **Global-MMLU** and both **CA** and **CS** subsets as part of the holistic evaluation of progress in multilingual LLM capabilities. As part of our commitment to the research ecosystem, we release **Global-MMLU** and **Global-MMLU Lite** under a fully permissive license for use in evaluations at <https://hf.co/datasets/CohereForAI/Global-MMLU> and <https://huggingface.co/datasets/CohereForAI/Global-MMLU-Lite>.

## 7 Limitations

**Uneven distribution of contributions** Beyond the gold standard languages where we engaged with compensated annotators, participation from community annotators was heavily skewed across languages. Despite a large volume of community annotators, there was a ‘long tail’ of annotators only contributing one or two annotations. Similarly, there is a huge gap between languages with the highest number of contributions and ones with the lowest number of contributions. Consequently, this suggests potential unevenness in dataset distributions across different languages and a lack of annotator diversity within some languages dominated by one or two frequent contributors.

**Language and dialect coverage** We focus on 42 lanugages for **Global-MMLU** . However, this is still only a tiny fraction of the world’s linguistic diversity. Of the world’s approximately 7,000 languages, only half of them are captured in any sort of written form (Adda et al., 2016). Of this half, only a few hundred are included on the internet in machine readable corpora (Adda et al., 2016). Future work is needed to continue to improve evaluations beyond these 42 languages

---

and to take into account how technology serves different dialects (a topic we do not address here). Geo-cultural variation within a language often gives rise to new dialects or creoles over time (Zampieri et al., 2020; Wolfram, 1997) and, as such, dialects can serve an important function in establishing and maintaining cultural identity(Falck et al., 2012). Many different dialects that are generally recognized as belonging to a single parent language are not represented in this evaluation dataset.

**Toxic or offensive speech** Our annotation interface does not contain specific flags for toxic, harmful, or offensive speech, so it is possible that **Global-MMLU**  contains some data that could be considered harmful. We believe this is of relatively low risk because of the nature of the original MMLU and the focus on examination material. However, we did not monitor or track this explicitly during our cultural sensitivity annotations or translation post-edits.

**Region Category Assignment:** For the annotation of geographically sensitive questions, we classified regions into six geographic regions (Africa, Asia, Europe, North America, Oceania, and South America).<sup>15</sup> However, based upon discussions we would going forward recommend switching to the taxonomy proposed by the World Bank which is more granular and includes separate designations for Central America and Sub-Saharan Africa.<sup>16</sup>

**Identifying cultural sensitivity does not guarantee cultural inclusion.** We acknowledge that efforts like the proposed Global-MMLU highlight important limitations in current datasets by identifying gaps in non-Western cultural representation. Identifying whether a dataset is culturally agnostic or not is highly relevant as mere translations may create the illusion that datasets are being more culturally inclusive and validating models in that sense, while this is not the real case. However, it must be noted that they do not fully resolve the issue. Future work must prioritize the integration of diverse culturally grounded knowledge to achieve true inclusivity and fairness in multilingual AI evaluation.

## 8 Acknowledgments

We would like to thank members of the Cohere For AI community who championed this initiative and helped with annotating samples for cultural sensitivity as well as improving translation quality across many languages. 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.

---

<sup>15</sup><https://www.pewresearch.org/global/2013/06/04/regional-categorization/>

<sup>16</sup><https://ourworldindata.org/world-region-map-definitions>

---

## References

- Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12027–12049, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.emnlp-main.671>.
- Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14, 2016. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2016.04.023>. URL <https://www.sciencedirect.com/science/article/pii/S1877050916300370>. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.
- 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. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021. doi: 10.1162/tacl\_a\_00416. URL <https://aclanthology.org/2021.tacl-1.66>.
- 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, Muhibdin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemedu Yigezu, Tajuddeen Gwadabe, Idris Abdulkumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeab 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. MasakhaNEWS: News

---

topic classification for African languages. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *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)*, pp. 144–159, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.10. URL <https://aclanthology.org/2023.ijcnlp-main.10>.

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 Yuehgooh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Bridget Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. Irokobench: A new benchmark for african languages in the age of large language models. *ArXiv*, abs/2406.03368, 2024. URL <https://api.semanticscholar.org/CorpusID:270258352>.

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

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

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. Aya 23: Open weight releases to further multilingual progress, 2024. URL <https://arxiv.org/abs/2405.15032>.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. [https://huggingface.co/spaces/open\\_llm-leaderboard-old/open\\_llm\\_leaderboard](https://huggingface.co/spaces/open_llm-leaderboard-old/open_llm_leaderboard), 2023.

Abhijit Bendale, Michael Sapienza, Steven Ripplinger, Simon Gibbs, Jaewon Lee, and Pranav Mistry. Sutra: Scalable multilingual language model architecture, 2024. URL <https://arxiv.org/abs/2405.06694>.

Steven Bird. Local languages, third spaces, and other high-resource scenarios. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7817–7829, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.539. URL <https://aclanthology.org/2022.acl-long.539>.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. ACM, October 2022. doi: 10.1145/3551624.3555290. URL <http://dx.doi.org/10.1145/3551624.3555290>.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. How human is machine translationese? comparing human and machine translations of text and speech. In Marcello Federico, Alex Waibel, Kevin Knight, Satoshi

---

Nakamura, Hermann Ney, Jan Niehues, Sebastian Stüker, Dekai Wu, Joseph Mariani, and Francois Yvon (eds.), *Proceedings of the 17th International Conference on Spoken Language Translation*, pp. 280–290, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.34. URL <https://aclanthology.org/2020.iwslt-1.34>.

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 Ghafari, 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. NusaCrowd: Open source initiative for Indonesian NLP resources. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13745–13818, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.868. URL <https://aclanthology.org/2023.findings-acl.868>.

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

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6588–6608, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.579. URL <https://aclanthology.org/2020.coling-main.579>.

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

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models?, 2024. URL <https://arxiv.org/abs/2406.12822>.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosslut. Meditron-70b: Scaling medical pretraining for large language models, 2023. URL <https://arxiv.org/abs/2311.16079>.

Rochelle Choenni, Sara Rajaei, Christof Monz, and Ekaterina Shutova. On the evaluation practices in multilingual nlp: Can machine translation offer an alternative to human translations?, 2024. URL <https://arxiv.org/abs/2406.14267>.

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

---

Eric Corbett, Emily Denton, and Sheena Erete. 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, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623228. URL <https://doi.org/10.1145/3617694.3623228>.

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. Vnhsge: Vietnamese high school graduation examination dataset for large language models, 2023. URL <https://arxiv.org/abs/2305.12199>.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 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*, 2023. URL <https://api.semanticscholar.org/CorpusID:263605822>.

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

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

Allan M. Feldman. *Majority Voting*, pp. 161–177. Springer US, Boston, MA, 1980. ISBN 978-1-4615-8141-3. doi: 10.1007/978-1-4615-8141-3\_10. URL [https://doi.org/10.1007/978-1-4615-8141-3\\_10](https://doi.org/10.1007/978-1-4615-8141-3_10).

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*, November 2023. ISSN 1396-0466. doi: 10.5210/fm.v28i11.13346. URL <http://dx.doi.org/10.5210/fm.v28i11.13346>.

∀, 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 Chinenyem Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL <https://aclanthology.org/2020.findings-emnlp.195>.

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. A framework for few-shot language model evaluation, 07 2024. URL <https://zenodo.org/records/12608602>.

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. Are we done with mmlu?, 2024. URL <https://arxiv.org/abs/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 Barra, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL <https://arxiv.org/abs/2403.08295>.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language?, 2024. URL <https://arxiv.org/abs/2404.06644>.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. Crowdsourcing Latin American Spanish for low-resource text-to-speech. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6504–6513, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.801>.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

---

*Processing (EMNLP)*, pp. 5427–5444, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.438. URL <https://aclanthology.org/2020.emnlp-main.438>.

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

Carlos Daniel Hernandez Mena and Ivan Vladimir Meza Ruiz. Creating Mexican Spanish language resources through the social service program. In Chris Callison-Burch, Christopher Cieri, James Fiumara, and Mark Liberman (eds.), *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pp. 20–24, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.nidcp-1.4>.

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

Mika Hämäläinen. *Endangered Languages are not Low-Resourced!*, pp. 1–11. University of Helsinki, March 2021. doi: 10.31885/9789515150257.1. URL <http://dx.doi.org/10.31885/9789515150257.1>.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srivivasan, Satwik Bhattacharya, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. Unsung challenges of building and deploying language technologies for low resource language communities. In Dipti Misra Sharma and Pushpak Bhattacharya (eds.), *Proceedings of the 16th International Conference on Natural Language Processing*, pp. 211–219, International Institute of Information Technology, Hyderabad, India, December 2019. NLP Association of India. URL <https://aclanthology.org/2019.icon-1.25>.

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

Meryem Karlik. Exploring the impact of culture on language learning: How understanding cultural context and values can deepen language acquisition. *International Journal of Language, Linguistics, Literature and Culture*, 2:5–11, 09 2023. doi: 10.59009/ijllc.2023.0035.

Naomi Kipuri. 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, pp. 51–81, 2009.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 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*, pp. 1–46, 2024.

---

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

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu, 2023. URL <https://arxiv.org/abs/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. Arabicmmlu: Assessing massive multitask language understanding in arabic, 2024. URL <https://arxiv.org/abs/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. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. doi: 10.1162/tacl\_a\_00447. URL <https://aclanthology.org/2022.tacl-1.4>.

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

William Labov. The social motivation of a sound change. *Word*, 19(3):273–309, 1963.

William Labov. The social stratification of (r) in new york city department stores. In *Dialect and language variation*, pp. 304–329. Elsevier, 1986.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 318–327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.28. URL <https://aclanthology.org/2023.emnlp-demo.28>.

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

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

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024a. URL <https://arxiv.org/abs/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. Culture-gen: Revealing global cultural perception in language models through natural language prompting, 2024b. URL <https://arxiv.org/abs/2404.10199>.

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

---

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.emnlp-main.818>.

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. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5155–5203, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.296. URL <https://aclanthology.org/2024.emnlp-main.296>.

Alexandra Luccioni and Joseph Viviano. What’s in the box? an analysis of undesirable content in the Common Crawl corpus. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 182–189, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.24. URL <https://aclanthology.org/2021.acl-short.24>.

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

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions, 2024. URL <https://arxiv.org/abs/2309.12342>.

Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, William Chandra Tjhi, and Alham Fikri Aji. Kalahi: A handcrafted, grassroots cultural llm evaluation suite for filipino, 2024. URL <https://arxiv.org/abs/2409.15380>.

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

---

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

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models, 2024. URL <https://arxiv.org/abs/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. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*, 2020.

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. No language left behind: Scaling human-centered machine translation, 2022.

OpenAI. GPT-4 Technical Report. 2024. URL <https://arxiv.org/abs/2303.08774>.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations, 2024. URL <https://arxiv.org/abs/2404.13076>.

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

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. Typhoon: Thai large language models, 2023. URL <https://arxiv.org/abs/2312.13951>.

Maja Popović. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL <https://aclanthology.org/W17-4770>.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A framework for measuring the cultural adaptability of large language models, 2024. URL <https://arxiv.org/abs/2404.12464>.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. Chatgpt mt: Competitive for high- (but not low-) resource languages, 2023. URL <https://arxiv.org/abs/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 Dzenhaliov, 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. Include: Evaluating multilingual language understanding with regional knowledge, 2024. URL <https://arxiv.org/abs/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. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks*, 2024.

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

Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. LLM see, LLM do: Leveraging active inheritance to target non-differentiable objectives. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9243–9267, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.521. URL <https://aclanthology.org/2024.emnlp-main.521>.

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. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL <https://aclanthology.org/2024.acl-long.620>.

Guojin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. Kmmlu: Measuring massive multitask language understanding in korean, 2024a. URL <https://arxiv.org/abs/2402.11548>.

Guojin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. Hae-rae bench: Evaluation of korean knowledge in language models, 2024b. URL <https://arxiv.org/abs/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. Challenging big-

---

bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Saeid Asgari Taghanaki, Aliasgahr Khani, and Amir Khasahmadi. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms, 2024. URL <https://arxiv.org/abs/2409.02257>.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2203–2213, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.188. URL <https://aclanthology.org/2021.eacl-main.188>.

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 Cabrerera, 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. All languages matter: Evaluating lmms on culturally diverse 100 languages, 2024. URL <https://arxiv.org/abs/2411.16508>.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models, 2024. URL <https://arxiv.org/abs/2310.01929>.

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

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 Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL <https://arxiv.org/abs/2406.01574>.

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

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyaranayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*, 2024. URL <https://arxiv.org/abs/2410.16153>.

- 
- Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Şenel, Anna Korhonen, and Hinrich Schütze. Turkishmmlu: Measuring massive multitask language understanding in turkish, 2024. URL <https://arxiv.org/abs/2407.12402>.
- Marcos Zampieri, Preslav Nakov, and Yves Scherrer. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612, 2020.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, 2023a. URL <https://arxiv.org/abs/2306.05179>.
- Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023b.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023. URL <https://arxiv.org/abs/2304.06364>.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765–2781, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.176. URL <https://aclanthology.org/2024.findings-naacl.176>.
- 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. Aya model: An instruction finetuned open-access multilingual language model, 2024. URL <https://arxiv.org/abs/2402.07827>.

## A 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 5: 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 ♦♣.

## B 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.

## C Global-MMLU Lite ✨

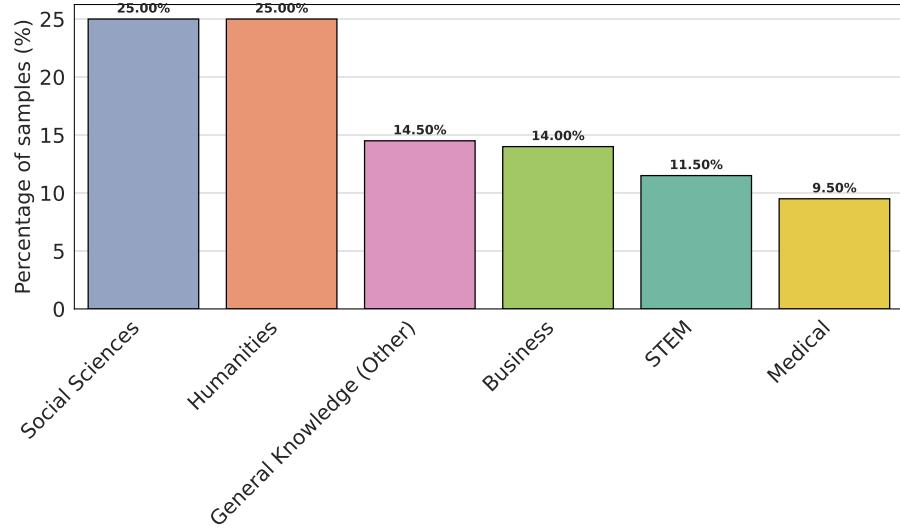


Figure 13: Distribution of samples across subject categories in **Global-MMLU Lite** ✨

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 13.

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.

## D 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 14 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.

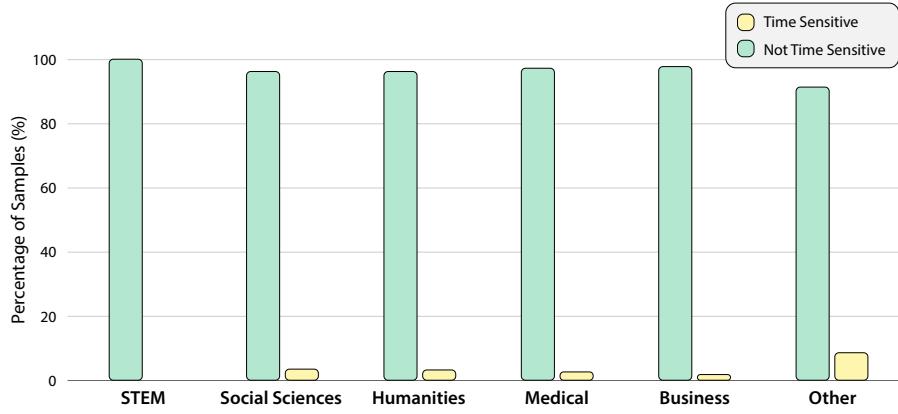


Figure 14: Distribution of time-sensitive samples across subject categories. Note that STEM subjects do not include any temporal knowledge.

## E Models Covered

Details of each model are described below:

- **Aya Expanse**<sup>17</sup> family of models include 8B<sup>18</sup> and 32B<sup>19</sup> 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](#)).
- **Command R and R+** are open-weight models of size 34B<sup>20</sup> and 104B<sup>21</sup> 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.

<sup>17</sup><https://hf.co/blog/aya-expanse>

<sup>18</sup><https://hf.co/CohereForAI/aya-expanse-8b>

<sup>19</sup><https://hf.co/CohereForAI/aya-expanse-32b>

<sup>20</sup><https://hf.co/CohereForAI/c4ai-command-r-08-2024>

<sup>21</sup><https://hf.co/CohereForAI/c4ai-command-r-plus-08-2024>

- 
- **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<sup>22</sup>** is part of the SEA-LION<sup>23,24</sup> 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) Llama 3.1 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<sup>25</sup>** is a 12B model which supports 11 languages including English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi.
  - **Qwen 2.5<sup>26</sup>** models support 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. The languages supported by this model are also unknown.

## F Additional Results

### F.1 Model Rank Changes

Table 6 presents the rank changes and corresponding position shifts (indicated next to the arrows) for high-resource languages, while Table 7 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 4.2.

---

<sup>22</sup><https://hf.co/aisingapore/gemma2-9b-cpt-sea-lionv3-instruct>

<sup>23</sup>An acronym for Southeast Asian Languages in One Network.

<sup>24</sup><https://github.com/aisingapore/sealion>

<sup>25</sup><https://hf.co/mistralai/Mistral-Nemo-Instruct-2407>

<sup>26</sup><https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e>

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	↑↑	-	-	↑1	-	-	-	↓1	-	↑1	-	-	↓1	-	-	2	2	
Chinese	↑↑	-	-	↓1	-	↑1	-	↓1	-	↑1	-	-	↑1	-	↓1	4	4	
Czech	↑↑	-	-	-	↑3	-	↓1	↓2	-	-	-	↑1	-	↓3	↓1	↓2	↑1	↓1
Dutch	↑↑	-	-	-	-	-	-	-	-	-	-	↓2	↓1	-	-	-	0	0
English	↑↑	-	-	-	-	↓1	-	-	-	↑1	-	↑1	↑1	-	↓1	-	4	4
French	↑↑	-	↑1	-	-	-	-	-	-	-	↓1	-	-	-	-	-	2	2
German	↑↑	-	↓1	-	↓1	-	↑1	-	-	↑1	-	↑1	-	-	-	-	4	4
Hindi	↑↑	-	↑1	↓2	↓1	↑1	-	↓1	↑1	-	↑1	↓3	↓1	↑2	-	-	5	6
Italian	↑↑	-	-	↑1	↑1	-	↓1	-	↑1	-	↓2	↓1	↑1	-	-	-	0	0
Japanese	↑↑	-	↑1	↑1	↑1	↑1	↓2	-	↑1	-	↓1	↓1	↓1	↑2	-	-	9	10
Persian	↑↑	↑1	↑1	-	↓1	-	-	↑1	-	↓2	-	-	-	-	-	-	5	6
Polish	↑↑	↑2	↑1	↑2	↓1	↓1	-	↓1	-	↓1	↑2	-	↓1	-	-	-	7	9
Portuguese	↑↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
Russian	↑↑	-	↓1	↓1	↓1	↑1	-	-	-	-	↑2	-	-	-	-	-	5	6
Serbian	↑↑	-	↓1	-	↑1	-	-	-	↓1	-	↓1	↑1	-	-	-	-	4	5
Spanish	↑↑	-	↓1	-	↓1	-	↑1	-	-	↑1	-	↓3	↓1	-	-	-	4	4
Swedish	↑↑	-	↓1	-	↓1	-	↑1	-	-	↓1	-	↓1	-	-	-	-	4	4
Turkish	↑↑	-	-	↓1	↑1	-	↓1	-	-	↑1	-	↑1	-	-	-	-	4	4
Vietnamese	↑↑	-	-	-	↓1	-	↑1	↓1	-	↑1	-	↑1	↓1	-	-	-	6	8

Table 6: 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	⬆️	-	↑1	-	-	-	-	-	↓1	↓1	-	-	-	-	-	3	3
● Filipino	⬆️	-	-	-	-	-	-	↑1	↑1	-	↓1	-	↓1	-	-	2	2
● Greek	⬆️	↓1	↓1	-	-	↑1	-	-	↓1	↑2	-	-	-	-	-	5	6
● Hebrew	⬆️	↓1	↑1	-	↓1	-	-	-	↑1	-	-	-	-	-	-	4	4
● Indonesian	⬆️	-	-	↓1	↓1	↓1	↑1	-	-	↑2	-	-	-	-	-	5	6
● Korean	⬆️	↓1	↓1	↓1	-	-	↑1	↑1	-	↑1	-	-	↓1	↓1	-	6	6
● Malay	⬆️	-	-	-	↓1	-	-	↓1	-	↑1	↑1	-	↓1	-	-	4	4
● Lithuanian	⬆️	-	-	-	↑2	-	-	-	-	-	-	↓2	-	-	-	0	0
● Romanian	⬆️	-	↑1	-	↓1	-	-	↑1	↓1	↓1	-	↑1	-	-	-	6	6
● Ukrainian	⬆️	-	↑1	-	↓1	↓1	-	-	-	↑1	-	-	↑1	↓1	-	4	4
○ Amharic	⬇️	-	-	↓1	↑1	↓1	-	-	-	↑1	↓3	-	-	-	-	4	4
○ Hausa	⬇️	-	-	-	-	-	-	-	-	-	↓1	↓1	↑1	-	-	6	10
○ Igbo	⬇️	-	-	↓1	-	-	↑1	-	-	↑1	↑1	-	-	-	-	0	0
○ Kyrgyz	⬇️	-	-	-	-	↓1	-	-	-	↑1	-	-	↓2	-	-	8	10
○ Malagasy	⬇️	-	↓1	-	-	-	-	-	-	↑1	-	-	↓1	-	-	2	2
○ Nepali	⬇️	-	-	-	-	-	-	↑1	↓1	↑1	-	-	↓1	↓1	-	2	2
○ Nyanja	⬇️	-	-	↓1	↓1	-	-	-	-	↑2	-	↑1	↓1	-	-	5	6
○ Shona	⬇️	-	-	-	↓1	-	-	↑1	-	↑1	↓1	↑1	↓1	↑1	-	4	4
○ Sinhala	⬇️	-	↑1	-	-	-	↓3	-	-	↑2	-	-	↓1	-	-	3	6
○ Somali	⬇️	-	↓2	-	↑1	-	-	-	-	↑1	-	-	↓2	↓1	↑1	4	4
○ Swahili	⬇️	-	↓1	-	-	-	↑1	-	-	-	-	-	↓1	↑1	-	2	2
○ Telugu	⬇️	-	↓1	-	-	-	-	↑1	-	↑1	↑1	↓1	-	-	-	4	4
○ Yoruba	⬇️	-	↑1	↓2	-	↓1	-	-	-	↑2	↑1	↓1	-	-	-	8	10
○	⬆️	-	↑1	↓2	-	↓1	-	-	-	↑2	↑1	↓1	-	-	-	6	8
○	⬇️	-	↓1	↑1	↑1	↑1	-	-	-	↓2	-	-	-	-	-	5	6

Table 7: Model rankings with MA 📈 rank as the reference for mid (●) and low (○) 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 (↓).

## F.2 Subject-level Performance

Figure 15 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.

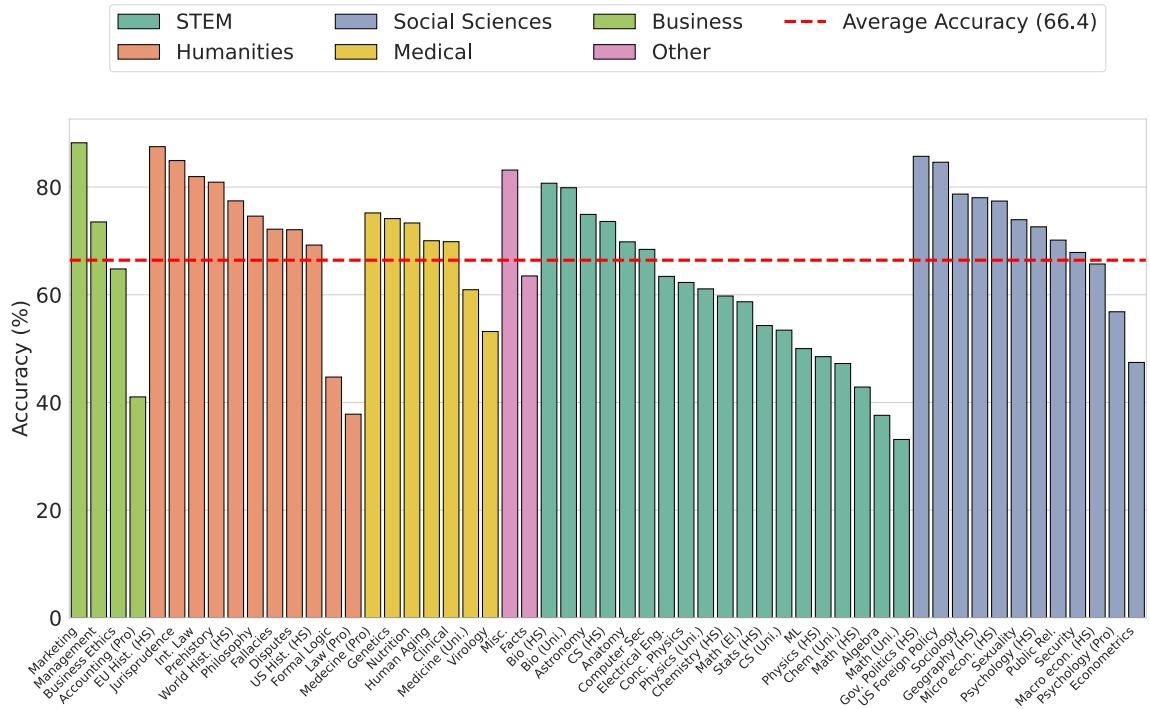


Figure 15: Aya Expanse 32B performance on each subjects.

## G Relationship between cultural and geographical tags

## G.1 Culture–Region Relations

We analyzed the samples in the **CS** dataset. Figure 16 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.

## G.2 Culture Country Relations

Figure 17 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 Honduras 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.

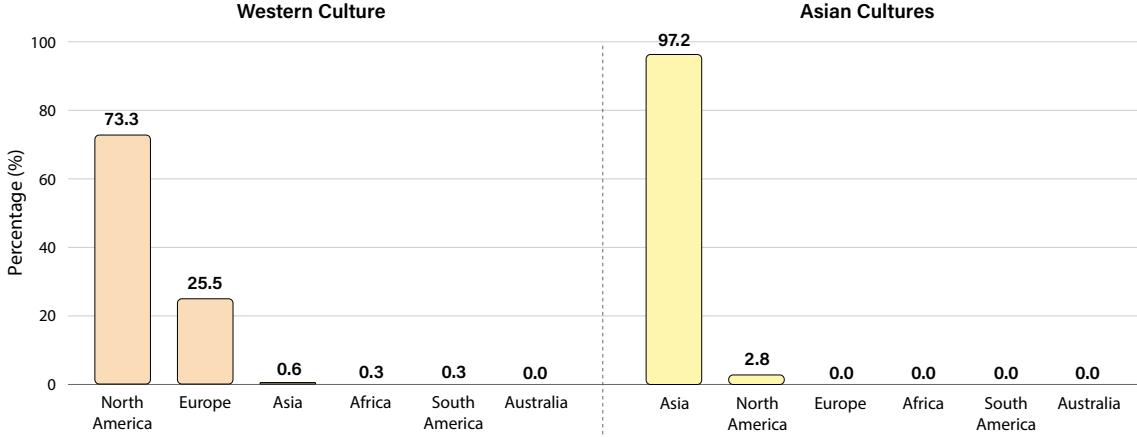


Figure 16: Relationship between Western and Asia cultures and region tags.

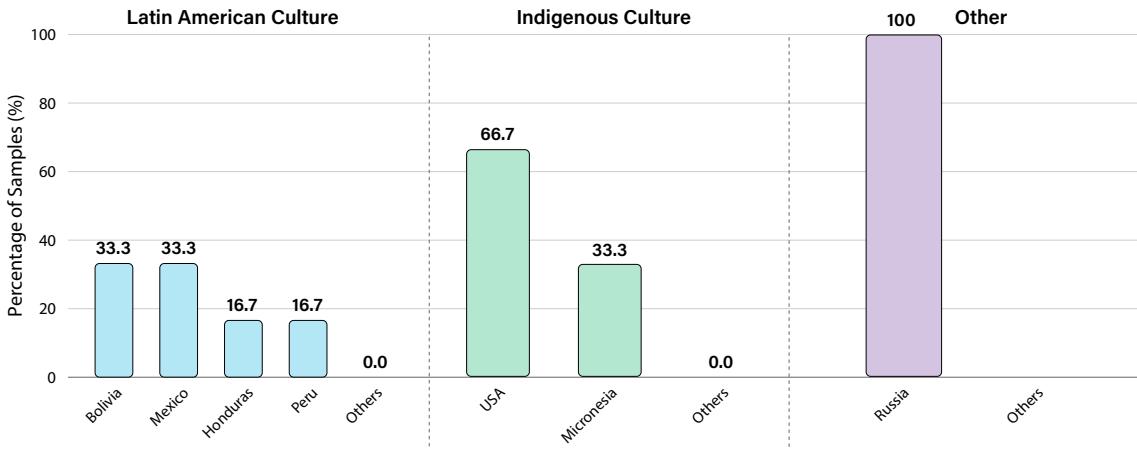


Figure 17: Relationship between culture and country tags, focusing on Latin American and Indigenous cultures.

### G.3 Region Country Relations

Figure 18 and 19 present country-specific information for each region: *North America*, *Europe*, and *Africa*. The United States accounts for the largest proportion of regional tags, representing 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.

## H 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

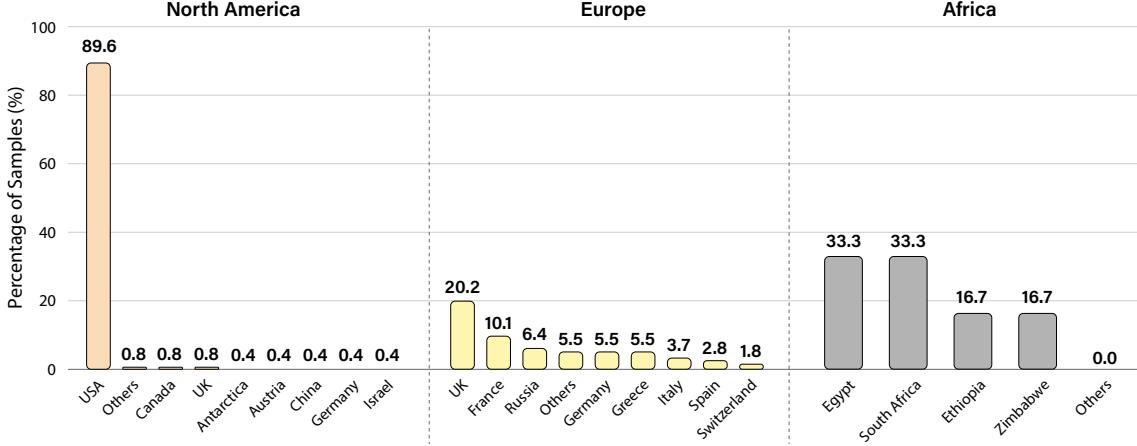


Figure 18: Relationship between region and country tags, focusing on North America, Europe and Africa regions.

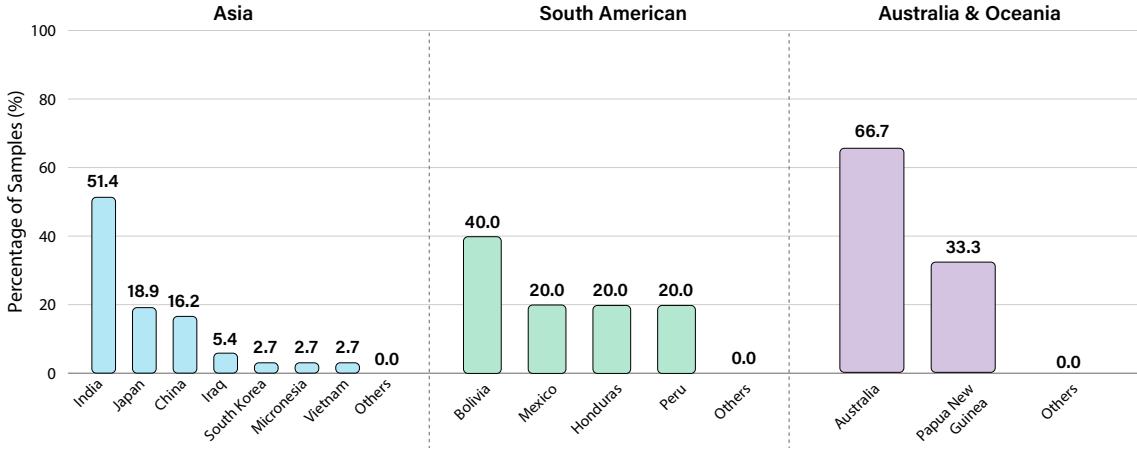


Figure 19: Relationship between region and country tags, focusing on Asia, South America and Australia.

capture difficult decisions and their rationales in comments for individual ratings. Similarly, they discussed ambiguous 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.<sup>27</sup> Argilla is an open-source tool that can be used for data labeling. Using Argilla’s Python SDK, it was quick

<sup>27</sup><https://argilla.io/>

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 and shown in Figure 20.

Figure 20: Cultural Sensitivity evaluation annotation interface.

As shown in Figure 21, 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.

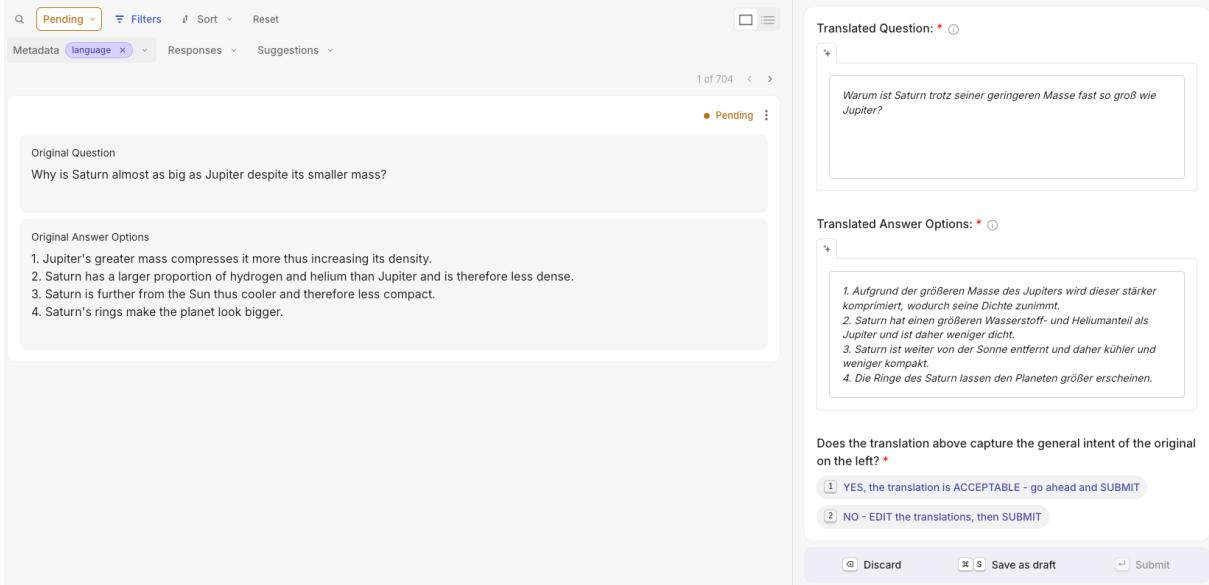


Figure 21: Translation evaluation annotation interface.

## H.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.

## H.2 Agreement between Annotators

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 23 and

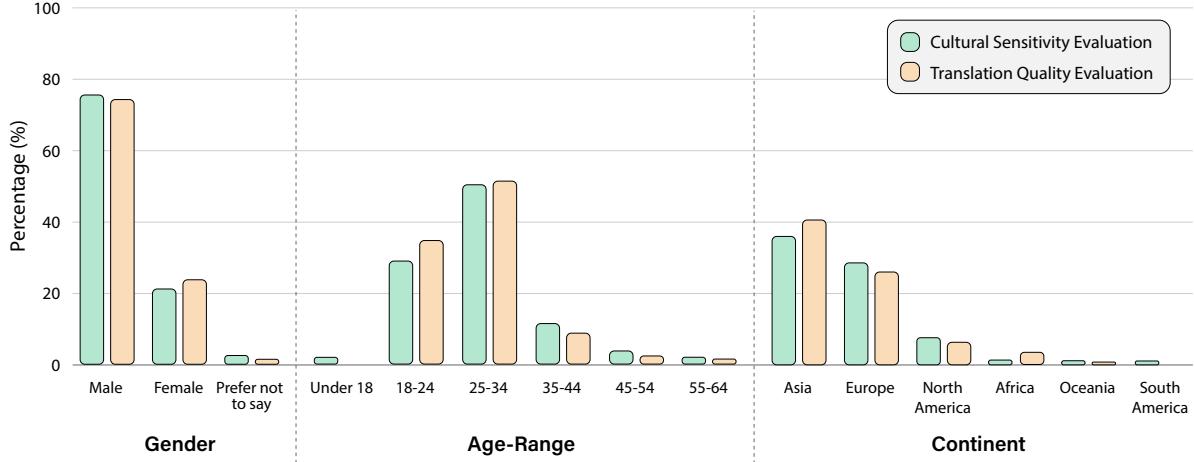


Figure 22: Demographics of annotators who registered using our annotation interface for cultural sensitivity as well as translation quality evaluation.

**24.** 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.

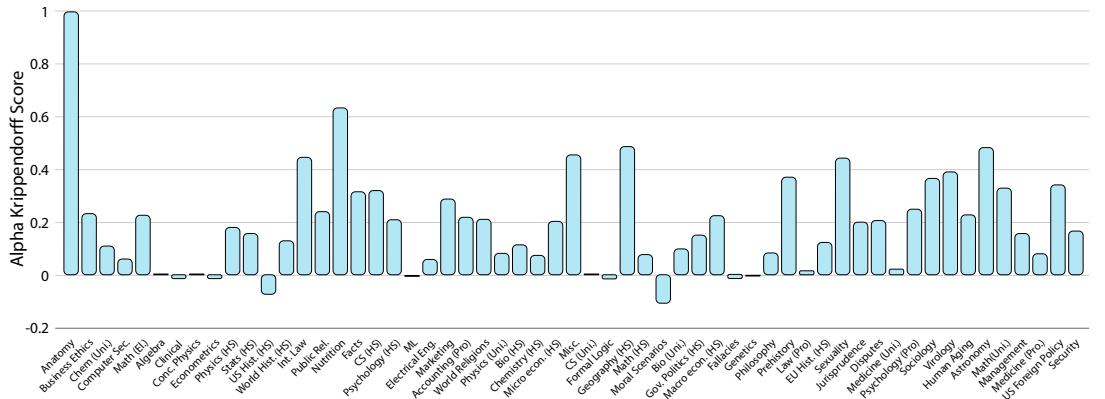


Figure 23: Krippendorff's Alpha Scores for checking annotator agreement regarding the presence of cultural or regional knowledge of samples.

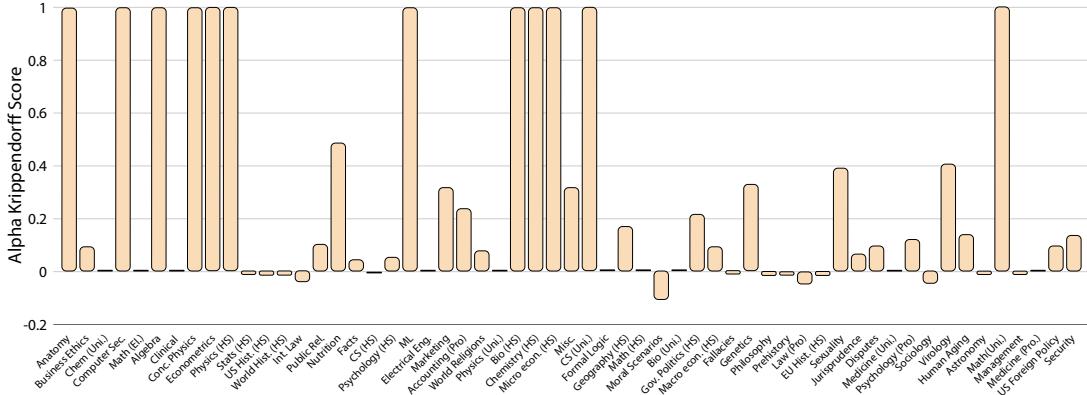


Figure 24: Krippendorff’s Alpha Scores for checking annotator agreement regarding the presence of the time-sensitive nature of samples.

## I Translation Analysis

### I.1 Translation Quality

Figure 7 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<sup>28</sup> as the human reference and only considered languages which overlapped between the two machine translated sets and human translated MMMLU.

### I.2 Translation Edits

Figure 25 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 26.

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.

<sup>28</sup><https://openai.com/index/openai-o1-system-card>

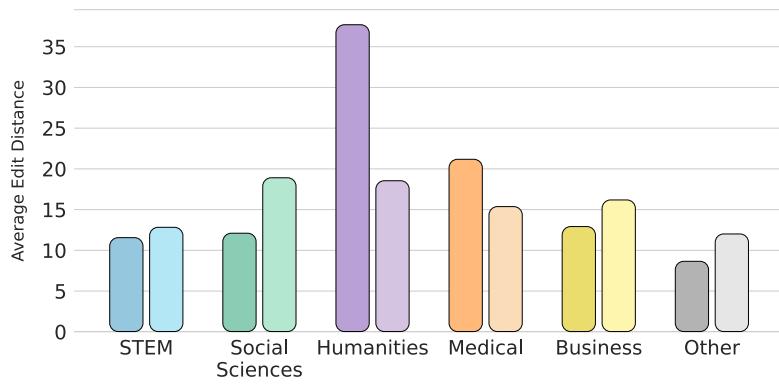


Figure 25: Average edit distance across different subject categories in MMLU. Each sample comprises a question-and-answer pair, with the left column showing edit distances for questions and the right column for answers.

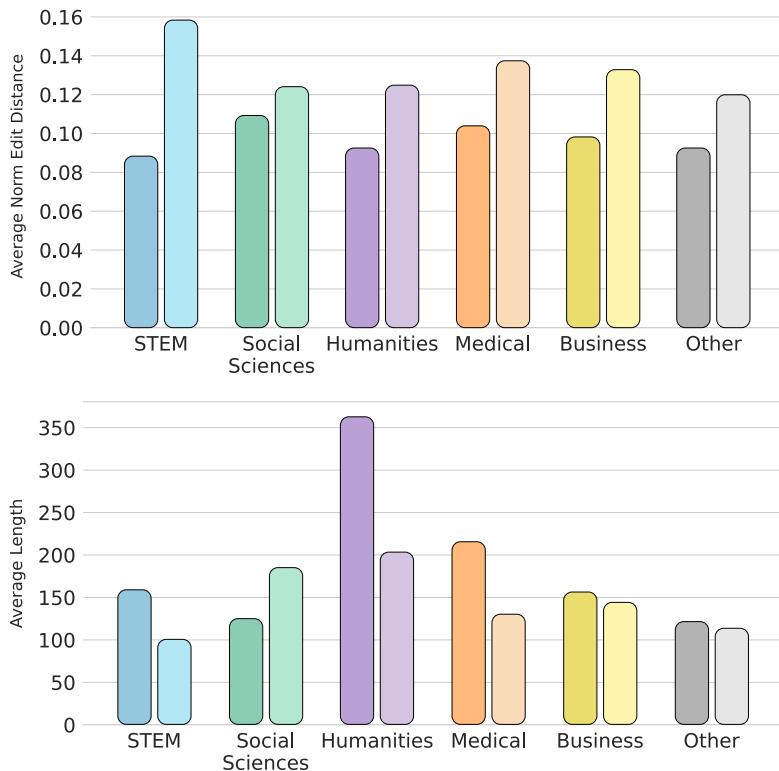


Figure 26: (Top) Average normalized edit distance and (Bottom) average question and answer lengths across different subject categories. The left column represents questions, while the right column represents answers.

## J MMLU Annotated Examples

Dataset	Subject	Question	Choices
	<b>US Hist. (HS)</b>	<p>This question refers to the following information:</p> <p>“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.”</p> <p>—Thomas Jefferson, 1816</p> <p>Which of the following best describes a contributing factor in the crafting of the United States Constitution?</p>	<p>(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.</p> <p>(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.</p> <p>(C) Difficulties over trade and foreign relations led to a repeal of overly restrictive tariffs required by the Articles of Confederation.</p> <p>(D) Washington’s embarrassing failure at the Whiskey Rebellion led to Federalist demands for a new framework for federal power.</p>
	<b>Accounting (Pro)</b>	<p>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?</p>	<p>(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.</p> <p>(B) The buyer is purchasing the goods for a particular purpose and the seller is a merchant in such goods.</p> <p>(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.</p> <p>(D) The seller knows the particular purpose for which the buyer will use the goods and the seller is a merchant in such goods.</p>
	<b>Jurisprudence</b>	<p>Which of the following criticisms of Llewellyn's distinction between the grand and formal styles of legal reasoning is the most compelling?</p>	<p>(A) There is no distinction between the two forms of legal reasoning.</p> <p>(B) Judges are appointed to interpret the law, not to make it.</p> <p>(C) It is misleading to pigeon-hole judges in this way.</p> <p>(D) Judicial reasoning is always formal.</p>

	<b>Prehistory</b>	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
	<b>US Foreign Policy</b>	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
	<b>Econometrics</b>	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)
	<b>Stats (HS)</b>	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
	<b>Formal Logic</b>	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.) $M \vee N \rightarrow M \wedge \frac{O}{N}$	(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

4	<b>Geography (HS)</b>	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
4	<b>Nutrition</b>	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.

## K 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	<p>(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.</p> <p>(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.</p> <p>(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.</p> <p>(D) Awareness of a region-specific natural phenomenon: The Northern Lights, visible in the night skies of Alaska and northern regions.</p> <p>(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)</p> <p>(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.</p> <p>(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.</p>	<p>(A) Global Climate Patterns: Understanding El Niño and La Niña weather phenomena, which occur worldwide and are not specific to any single region.</p> <p>(B) Universal Celestial Bodies: The Sun and the Moon are visible worldwide and do not possess regional specificity.</p> <p>(C) Standardized Geography Terms: Understanding the definition of a peninsula or archipelago is applicable to geographic features globally, not tied to regional knowledge.</p>
--------------	--	---

Dialect	<p>(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.</p> <p>(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.”</p> <p>(C) Knowledge of social greetings: The customary handshake and verbal greeting of “Konnichiwa” when meeting someone in Japanese culture.</p> <p>(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</p>	<p>(A) Standardized technical jargon: Medical or legal terminology used internationally within professional fields.</p> <p>(B) Formal literary language: The writings of Shakespeare or Dickens utilize sophisticated language but are not tied to specific dialects.</p> <p>(C) Global brand names: Companies like Nike or Adidas use consistent branding worldwide, avoiding regional vocabulary.</p>
---------	---	---

## L 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](#), [6](#), [23](#), [24](#).