Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures

v0.1

Tyler A. Chang[1*], **Catherine Arnett**[2*], and
Authors at the 5th Multilingual Representation Learning (MRL) Workshop[†]

[1]UC San Diego, [2]EleutherAI
[†]For full authorship list, see §A.
[*]Equal contribution

Abstract

To date, there exist almost no culturally-specific evaluation benchmarks for large language models (LLMs) that cover a large number of languages and cultures. In this paper, we present **Global PIQA**, a participatory commonsense reasoning benchmark for over 100 languages, constructed by hand by 320 researchers from 65 countries around the world. The 116 language varieties in Global PIQA cover five continents, 14 language families, and 23 writing systems. In the non-parallel split of Global PIQA, over 50% of examples reference local foods, customs, traditions, or other culturally-specific elements. We find that state-of-the-art LLMs perform well on Global PIQA in aggregate, but they exhibit weaker performance in lower-resource languages (up to a 41% accuracy gap, despite random chance at 50%). Open models generally perform significantly worse than proprietary models. Global PIQA highlights that in many languages and cultures, everyday knowledge remains an area for improvement, alongside more widely-discussed capabilities such as complex reasoning and expert knowledge. Beyond its uses for LLM evaluation, we hope that Global PIQA provides a glimpse into the wide diversity of cultures in which human language is embedded.

`https://mrlbenchmarks.github.io/`

`Global PIQA` `mrlbenchmarks`

1 Introduction

Nearly all prominent multilingual benchmarks for large language models (LLMs) translate existing English datasets into other languages (e.g. XNLI: Conneau et al., 2018, XCOPA: Ponti et al., 2020, Belebele: Bandarkar et al., 2024, XStoryCloze: Lin et al., 2022, MGSM: Shi et al., 2023, Global MMLU: Singh et al., 2025, etc). As a result, the vast majority of the world's languages lack culturally-specific evaluation datasets that cover local customs, traditions, and everyday life for speakers of those languages. The culturally-specific datasets that do exist generally still rely heavily on translation or are limited to a relatively small number of languages (Citations; §B).

This lack of culturally-specific datasets is particularly relevant in the domain of commonsense reasoning, where LLMs are evaluated for physical, social, and world knowledge that is broadly known by the majority of people in a community. Commonsense reasoning capabilities have long been a desirable property of LLM-based systems, evaluated through popular benchmarks such as HellaSwag (Zellers et al., 2019) and PIQA (Bisk et al., 2020). Because commonsense reasoning focuses on everyday physical and social activities, and it has its basis in community knowledge, it differs greatly across languages and cultures. This variation across communities is particularly noticeable

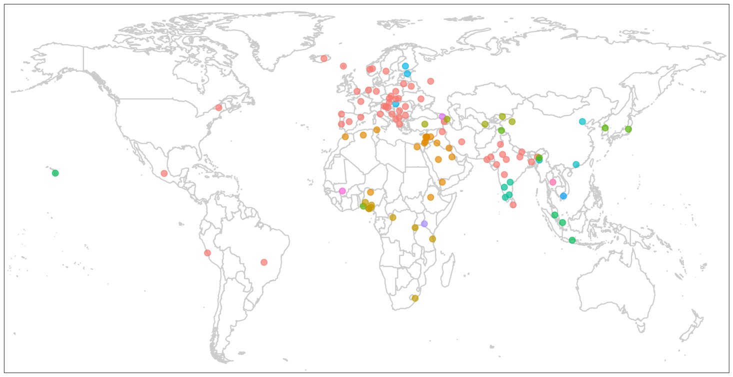

Figure 1: Map of the 116 language varieties represented in Global PIQA, colored according to language families from Glottolog (Hammarström et al., 2023).

when compared to the relative uniformity of more abstract capabilities, such as mathematical or logical reasoning, which have been the focus of many recent LLM evaluation benchmarks (Citations). Unfortunately, culturally-specific commonsense reasoning evaluation datasets do not exist for the vast majority of the world's languages.

To fill this gap, we present **Global PIQA**, a culturally-specific physical commonsense reasoning benchmark created by native speakers of over 100 language varieties across the globe. In contrast to previous multilingual benchmarks, examples in the non-parallel split of Global PIQA are written directly in each language, largely by NLP researchers who speak the language, involving very little translation. Authors were given flexibility to determine the topics and domains for their examples, in order to develop "target-language original prompts" (Kreutzer et al., 2025) that are appropriate for each linguistic and cultural context. All contributors to the datasets were offered authorship on this paper, to reflect the significance of these intellectual contributions to the project.

We then evaluate state-of-the-art LLMs on Global PIQA. We find that proprietary models perform well in aggregate, with the best performing model achieving an accuracy of 91.7%. In some ways, this is expected, as Global PIQA is designed to evaluate *commonsense* knowledge that is widely known in each cultural and linguistic community. However, Global PIQA highlights disparities between high- and low-resource languages; for example, the best performing model for Sub-Saharan African languages reaches an accuracy of only 80.2% for those languages, compared to 95.6% for European languages (with chance at 50%). Open weight models generally perform worse (best performing model: 82.4% across all languages) than proprietary models. We hope that Global PIQA will enable researchers to measure and ultimately close the multilingual performance gap both across languages and between open and proprietary models.[1]

2 Background and Related Work

Multilingual evaluation datasets. Most multilingual evaluations for standard LLM tasks (e.g. question answering and mathematical reasoning) are the product of translation from English (e.g. EU20: Thellmann et al., 2024, mArenaHard: Dang et al., 2024, Okapi: Lai et al., 2023, MMLU-ProX: Xuan et al., 2025, MGSM: Shi et al., 2023, etc). In some cases, the translations are automatic without any human verification, which can lead to unnatural examples and low-quality datasets due to artifacts from machine translation (Singh et al., 2025). In other cases, benchmarks are professionally translated or use human-verified translations (e.g. Belebele: Bandarkar et al., 2024, MMMLU: OpenAI, 2024, IrokoBench: Adelani et al., 2025, GlobalMMLU: Singh et al., 2025, XQUAD: Artetxe et al., 2020,

[1]We release Global PIQA under a CC-BY-SA 4.0 license. Global PIQA is intended only for evaluation. We do not allow training of AI systems on Global PIQA, or on AI-generated data that uses Global PIQA as a seed. We release the raw materials and unbalanced dataset at `https://github.com/mrlbenchmarks/global-piqa`, along with further details about the dataset and evaluation.

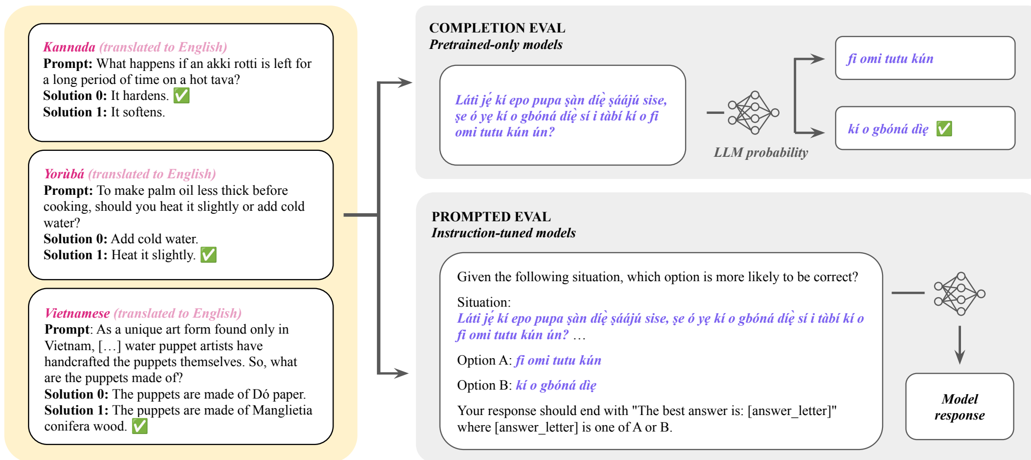

Figure 2: The format of Global PIQA examples. Each example can be used either in a completion setting (to evaluate pretrained-only models) or a prompted setting (to evaluate instruction-tuned models). Evaluation method details are in §5.

etc). These benchmarks are less likely to suffer from quality issues related to machine translation, but they are still not necessarily culturally relevant for the target languages. Benchmarks translated from English have been found to propagate Anglocentric perspectives and values (Singh et al., 2025; Kreutzer et al., 2025).

Culturally-specific evaluation. Culturally-specific evaluation is critical for designing models that align with values other than those from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) cultures (Henrich et al., 2010). Culturally-specific benchmarks have been constructed for a variety of languages (e.g. INCLUDE: Romanou et al., 2025, TyDi QA: Clark et al., 2020, CulturalBench: Chiu et al., 2025, MultiLoKo: Hupkes and Bogoychev, 2025, DOSA: Seth et al., 2024, BLEnD: Myung et al., 2024, etc), and datasets such as MMLU (Hendrycks et al., 2021) have been localized to other languages (e.g. CMMLU: Li et al., 2024, KMMLU: Son et al., 2025, ArabicMMLU: Koto et al., 2024, TurkishMMLU: Yüksel et al., 2024, and IndoMMLU: Koto et al., 2023). Results from these localized benchmarks tend to correlate more strongly with human judgments of model quality than results from translated or non-localized benchmarks (Wu et al., 2025). Still, these datasets often focus on challenging knowledge questions in localized topics, rather than commonsense cultural knowledge which is often widely known in the community but not documented on the web.

Physical Interaction: Question Answering (PIQA). To define the task format and scope for our benchmark, we take inspiration from PIQA (Bisk et al., 2020). PIQA aims to measure physical commonsense reasoning, which we note in §1 is likely to vary substantially across languages and cultures. In Global PIQA, we define physical commonsense reasoning as a broad collection of related tasks relying on knowledge of physical properties, affordances (types of actions an agent can perform with an object; Gibson, 1979; Jones et al., 2022), and physical and temporal relations. Each example in PIQA consists of a "goal" (or prompt) and two possible solutions, one correct and one incorrect (e.g. Figure 2). Prompt-solution pairs can consist of sentence beginnings and completions, questions and answers, or goals (e.g. making specific food dishes) and solutions. Even five years after its initial release, PIQA is still being used in evaluations, e.g. reported in technical reports for releases such as Gemma 3 (Team Gemma et al., 2025) and Llama 3 (Meta AI, 2024). Despite its broad usage as a benchmark for English, PIQA has not been translated or broadly adapted as a multilingual benchmark, much less extended to massively multilingual and culturally-specific settings.[2]

[2]Üstün et al. (2024) machine-translate PIQA into 93 languages to train the Aya model, but these translations are not human verified. Translations also exist on Hugging Face for Catalan and Basque.

3 Global PIQA: Non-Parallel Split

Thus, we construct Global PIQA, a physical commonsense reasoning benchmark for 116 language varieties. The primary split of Global PIQA is non-parallel (i.e. not translated across languages) to allow authors to write culturally-specific examples for their languages. Following the PIQA dataset (Bisk et al., 2020; §2), each example consists of a prompt and two candidate solutions. These can be used to evaluate either a pretrained-only model (Figure 2, center) or an instruction-tuned model (Figure 2, right). In every example, determining the correct solution is designed to require physical commonsense reasoning, although we allow for fairly flexible definitions of physical commonsense (§D.2).

3.1 Organizing a Global and Participatory Benchmark

For the non-parallel split of Global PIQA, authors contributed datasets following the task format described above (details in §3.2). Authors provided their datasets with short dataset descriptions,[3] and all authors of included datasets were offered co-authorship on this paper. To date, the Global PIQA project has involved 320 contributors across 65 countries and 165 university or company affiliations. Our authors range from early career undergraduate researchers to professors at major global universities. Here, we describe key decisions that made the collaboration a success.

- **Researchers and authorship.** One major reason for this benchmark's success was that we recruited NLP researchers themselves to construct the datasets. In this setup, researchers benefit from co-authorship on a large benchmark paper, and they have both the domain expertise and motivation to write high quality examples. Participation is entirely voluntary. This contrasts with benchmarks where external annotators are paid to create datasets, with little incentive to create high quality examples. Our collaborative approach involving other NLP researchers is less exploitative, and because many of our authors develop technologies for their language(s), authors also benefit from having a high quality benchmark in their language(s).
- **Recruiting.** We were able to recruit a diverse group of contributors through large online communities, low-resource NLP community organizations, social media, and personal connections.[4] We also identified NLP researchers with experience constructing benchmarks and language models for specific languages or language families, and we contacted them directly to broaden our reach. We maintained a spreadsheet of interested volunteers (with contact information and languages spoken) to keep volunteers informed throughout the process.
- **Early feedback.** We allowed authors to send initial examples and preliminary versions of their datasets for feedback well before the dataset submission deadline. This contrasts with traditional shared tasks at NLP conferences, where participants have minimal interaction with the organizers prior to submitting. Furthermore, we held FAQ meetings one month before the deadline, held at multiple times to accommodate different time zones, and we maintained a consistently-updated set of slides with instructions and FAQs for creating the Global PIQA datasets.
- **Timeline.** The shared task was announced in the last week of June 2025, with a submission deadline of September 15. This allowed almost three months to recruit contributors and for groups to develop datasets. The timeline was not so long, however, that momentum was lost. The bulk of feedback to participants and recruiting happened in the second half of the three-month period.
- **Data quantity.** We required a minimum of 100 items per language for each submitted dataset. We found that this quantity was doable so as not to discourage researchers from participating, but large enough to ensure that researchers put significant thought into creating their datasets.
- **Flexible deadlines and acceptances.** After the dataset submission deadline, we continued to allow submissions for languages and dialects that were still missing from the benchmark. We individually reached out to volunteers who had signed up for specific missing languages, and in many cases, we were able to work out later deadlines that were more amenable to those authors. In cases where an initial dataset submission did not meet quality checks (§3.3), the dataset was not simply rejected; instead, we worked with the authors to make improvements for the dataset to be accepted.

[3]Dataset descriptions ranged from single paragraphs to full length papers. Individual dataset descriptions that the individual authors have decided to publicly release are at X.

[4]We publicized the Global PIQA task through announcements on the Eleuther AI Discord, the LINGUIST List, Masakhane, X/Twitter, BlueSky, and LinkedIn.

3.2 Dataset Construction Methods

We asked authors to construct at least 100 examples in their language, all manually checked by a native speaker of the language. Translated examples directly from the English PIQA dataset are not included in the non-parallel split of Global PIQA. Authors were asked to construct examples (*prompt*, *solution0*, *solution1*) where (1) the correct solution relates to physical properties of one or more objects, and (2) an average person who speaks the language natively would likely know the answer. We encouraged authors to include culturally-specific examples that might not be easily translatable into English, or that might require regional or cultural commonsense knowledge. Specifically, in the guidelines sent to all authors, we encouraged examples based on "local foods, places, everyday objects, customs, traditions, religions, literature, folklore, or art forms". We asked authors to vary the length of their examples (e.g. to include some examples greater than 25 words long), make the two candidate solutions as similar as possible (while still having one be unambiguously correct and the other unambiguously incorrect), and avoid having the incorrect solution be "so absurd that it is extremely obvious". Full guidelines sent to authors are in §X.

Aside from these guidelines, authors were provided substantial flexibility in creating the datasets for their languages. This is a benefit of having researchers construct their own datasets; as native speakers and researchers working in each language, they themselves are experts who can ensure the quality of their respective datasets. This flexibility also allowed each author to construct a dataset that was culturally specific to their language and dialect, in the way that they believed was best. Method descriptions for individual languages are in §H.

Diverse methods. Indeed, authors used a wide variety of methods to brainstorm and construct examples. We encouraged authors to manually write examples, and Nout of Ngroups wrote their examples manually. Some authors (Ngroups) wrote examples motivated by content on websites or other resources in their language, such as recipe blogs, DIY pages, question forums, or how-to books. Many groups (Ngroups) brainstormed examples based on specific topic categories, such as food, home, clothing, transportation, hobbies, or religion. The vast majority of groups (Nout of Ngroups) explicitly reported making their datasets at least partially culturally-specific, covering local foods, clothing, traditions, everyday life, and/or customs. Examples of hand-picked culturally-specific examples from Global PIQA are shown in Figure X.

The majority of authors (Nout of Ngroups) also reported writing examples based on everyday situations. For example, one group spent one month adapting examples from naturally-occurring sentences spoken by family and friends, and another group read examples aloud to their parents and grandparents to verify "colloquial [language] usage, cultural appropriateness, and everyday realism". All groups had examples written or checked by at least one native speaker, and many groups (Ngroups) had multiple native speakers check each example. Brief method details for individual groups are in §H, and we highly encourage readers to explore these individual dataset descriptions.

A small number of groups (N out of Ngroups) used LLMs to generate topic ideas, but not to generate examples themselves. An even smaller number of groups used LLMs to initially generate examples, before filtering, editing, and manual verification by the authors (Ngroups). In these cases, LLMs had to be prompted carefully so as not to generate easy and generic examples; for example, one group reported that "our preliminary attempts involved using state-of-the-art Large Language Models (LLMs) to generate question candidates. However, we found these outputs to be consistently inadequate" (for Tamil). Another group reported that LLMs "produced poor quality samples; no such items were included in the final dataset" (for Azerbaijani). Of the N groups that still used LLMs to generate initial examples, the authors reported needing to filter the resulting datasets heavily for quality (e.g. keeping only 14.6% and 22.0% of examples in the two independent groups who reported the proportions of examples kept).

3.3 Compiling the Dataset

The next step in constructing the Global PIQA non-parallel split was to run quality checks and compile the dataset for each language. For each language, we standardized column names, added unique example IDs, and normalized language codes to use ISO 639-3 individual language codes (e.g. `cmn` for Mandarin Chinese, c.f. macrolanguage codes; language code details in §C) with ISO 15924 script codes (e.g. `latn` for Latin script). In cases where a dataset used a specific dialect within an individual language code, we appended an optional four-letter region code; for example, the Global

PIQA language code for Brazilian Portuguese is `por_latn_braz`. Finally, to inspect the data more easily, we generated machine translations into English using Gemini 2.5 Pro (October 2025). The translation prompt used is in §D.1.

Additional manual annotation and cultural specificity. Based on these LLM-generated English translations, we dropped examples that did not fit the task description (e.g. we dropped several abstract logic puzzles and complex mathematical reasoning questions). We also dropped examples that seemed trivially easy based on the English translations. Finally, we annotated examples as "culturally-specific" if they met at least one of three criteria: (1) the example requires knowledge of a word that does not translate well into English, e.g. specific food dishes or local brands, (2) the example describes specific holidays, folklore, traditions, or sayings, or (3) the correct solution likely varies by region, e.g. involving local norms, laws, or customs. Annotation details, along with motivations for our operationalization of cultural specificity, are in §D.2. In cases where all examples were quite non-culturally-specific, or where dropping trivial and off-task examples led to a dataset with under 100 examples in the language, we worked with the authors for that language to reach the 100 example threshold and to increase the number of culturally-specific examples in their dataset.

Subsampling. After cleaning but before any subsampling, the full dataset consists of 28K examples covering 116 language codes (§C). Because this full dataset is highly skewed across languages and often overwhelmed by non-culturally-specific examples or repeated examples about similar topics, we subsample to an *official non-parallel split* of exactly 100 examples per language to use for model evaluations. Subsampling details for the official non-parallel split are in §D.3, and we provide an overview here. First, where possible (i.e. when this does not reduce our sample size to less than 100 examples for a given language), we filter out examples where the two candidate solutions differ in length by more than 25 bytes, when normalized to English byte equivalents. We also filter out examples whose non-stopword tokens overlap by more than 50% with another example in the dataset, using the per-language tokenizers from Goldfish (Chang et al., 2024).[5] This aims to ensure diversity across topics for the official Global PIQA dataset for each language.

Finally, we sample 100 examples from this filtered subset for each language. We sample culturally-specific examples before non-culturally specific examples (as annotated in §3.3), and within each of these categories, we first sample examples that did not use any LLMs in the creation process. In the resulting official non-parallel split, 58% of examples are annotated as culturally-specific, and only 4.0% of examples are written with the help of LLMs. All examples have been manually validated by at least one native speaker of the respective language, and N%have been validated by multiple native speakers.

3.4 Official Non-Parallel Split

The resulting official non-parallel split of Global PIQA contains 100 examples per language for 116 language codes. When excluding region codes, the Global PIQA non-parallel split covers N language-script combinations and 101 unique ISO 639-3 language codes. These languages cover five continents, 14 language families, and 23 scripts (writing systems). The full list of languages is in §C. Importantly, the dataset contains 58% culturally-specific examples, as annotated in §3.3, enabling evaluations across a wide variety of global cultures. Across languages, the mean prompt length is N English character equivalents (§D.3), with mean correct solution length of Xand mean incorrect solution length of X.

4 Global PIQA: Parallel Split

We are in the process of developing a parallel split of Global PIQA. This dataset will consist of PIQA-style items originally written in English, which we will machine translate and send to the authors of the non-parallel split for correction and validation. The vast majority of Global PIQA authors have professional working proficiency in English on top of their native language(s). As discussed in Section 2, parallel (translated) datasets are inherently biased towards the source language.

[5] Due to the lack of available resources for many low-resource languages in our dataset, we define stopword tokens as tokens that appear in at least 25% of examples in the Global PIQA dataset for that language. Details for token overlap filtering are in §D.3.

Therefore, we will aim to make the English parallel split as non-culturally-specific as possible. While parallel evaluation datasets do not allow for culturally-specific evaluations, they allow researchers to make more direct comparisons across languages; for example, in Global PIQA, we hope that the parallel split will allow us to determine whether performance differences across languages are due to (1) differences in models' physical commonsense reasoning capabilities in different languages, vs. (2) differences in how well the models perform in different cultural contexts, as evaluated in the non-parallel split.

5 Results for State-of-the-Art LLMs

Finally, we evaluate existing LLMs on Global PIQA. We find that proprietary models perform well when averaged across all languages, but performance is substantially worse for some languages and regions. Open-weight models under-perform relative to closed models, both in aggregate and for each individual language.

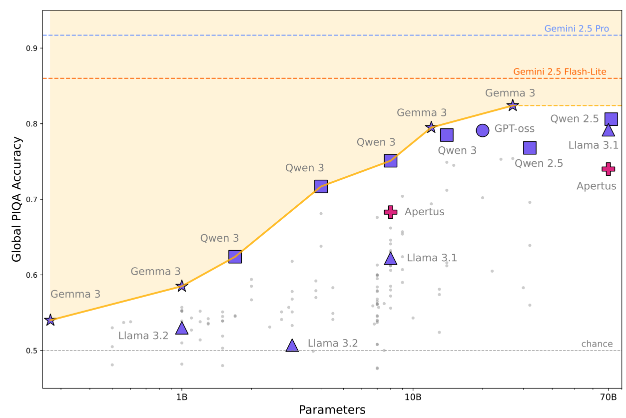

Figure 3: Accuracy averaged across all languages vs. parameter count for open-weight models. We highlight top-performing models. Shape indicates model family, and color indicates openness (open-weight vs. open-data models). All other models are plotted as gray dots. Chance (50%; gray) performance and performance for Gemini 2.5 Pro and Gemini 2.5 Flash-Lite are plotted as dashed lines.

5.1 Evaluation Format

We evaluate models in one of two formats (Figure 2): completion or prompted. All examples in Global PIQA are amenable to either format. Both versions are implemented in the LM Evaluation Harness (Gao et al., 2024).[6]

Completion evaluation: For models that are not tuned to follow instructions (i.e. pretrained-only or "base" models), we compute the log-probability from the LLM for each candidate solution given the prompt, normalized by the length of each solution in bytes: $P(solution \mid prompt) / \, \mathrm{len}(solution)$. If the correct solution has a higher normalized probability than the incorrect solution, then we mark the model correct for that example.

Prompted evaluation: For models that are tuned to follow instructions (e.g. the vast majority of proprietary models, and instruction-tuned and RL-tuned open models), we prompt the LLM with the prompt template in Figure 2. We sample a maximum of 1024 generated response tokens (allowing another 1024 "thinking" tokens for thinking models), then we use exact string matching to mark answers as correct or incorrect. Evaluation method details are in §F.

For smaller models (e.g. up through the 2–4B "weight class"; Michaelov et al., 2024), we find that base models evaluated with the completion format perform better than instruction-tuned (IT) models evaluated with the prompted format. This is consistent with the claim that instruction-following imposes auxiliary task demands that may obscure capabilities in smaller models (Hu and Frank, 2024). In the main text here, we only report results for models with 7B+ parameters, and thus all results in this section use the prompted evaluation format. In Appendix G.1, we report results using the completion format. For all models, we report accuracy, where chance accuracy is 50%.

5.2 Models

We evaluate a large range of open, open-weight, and proprietary (closed) models on Global PIQA. We evaluate pretrained-only models exclusively with the completion format and instruction-tuned

[6]Link to tasks.

models exclusively with the prompted format (§5.1). Evaluated models include BLOOM (Workshop et al., 2022), Apertus (Hernández-Cano et al., 2025), Qwen 2.5 and 3 (Yang et al., 2024a; Team, 2025), Llama 3.1 and 3.2 (Meta AI, 2024), Gemma 2 and 3 (Team Gemma et al., 2024, 2025), XGLM (Lin et al., 2021), Aya and Command R (Dang et al., 2024; Cohere et al., 2025), GPT-5 (full, nano, and mini; OpenAI, 2025), Sonnet 4.5 (Anthropic, 2025), and Gemini 2.5 (pro, flash, and flash-lite; Google DeepMind, 2025c,b,a). Proprietary models (GPT-5, Claude, and Gemini) are evaluated with thinking on, with details in §F. The open-weight models range from 300M to 70B parameters.

We also evaluate open-weight models that are trained to focus on one language or region, including Kanana (Bak et al., 2025; Korean), PhoGPT (Nguyen et al., 2023a; Vietnamese), SeaLLM (Zhang et al., 2025; Southeast Asian languages), Salamandra (Gonzalez-Agirre et al., 2025a; European languages), EuroLLM (Martins et al., 2025; European languages), Poro 2 (Zosa et al., 2025; Finnish), Cheetah (Adebara et al., 2024; African languages), Sailor2 (Dou et al., 2025; Southeast Asian languages), and Jais (Sengupta et al., 2023; Inception, 2024; Arabic). We prioritize models that were requested by the authors of the datasets, and we prioritize models pretrained from scratch over adapted and fine-tuned models. See §F.1 for the full list of models we evaluate.

5.3 Results

In Table 1, we report accuracies averaged across languages per region in the Global PIQA non-parallel split, along with the overall accuracy per model. Because each language has exactly 100 examples in the official non-parallel split, average performance across all languages is equivalent to the macroaverage. The best-performing model overall is Gemini 2.5 Pro, with an average score of 91.7%. Gemini 2.5 Pro achieves the highest score of any model for seven of the ten regions in Table 1. The best open-weight model overall is Gemma 3 27B (average score of 82.4%), outperforming open-weight models even at the 70–72B parameter scale. Gemma 3 27B performs best out of the open-weight models for languages in Eastern Europe, the Middle East, North Africa, Sub-Saharan Africa, Central Asia, and South Asia. Overall, open-weight model performance steadily increases as parameter counts increase, but there remains a gap between the top proprietary models and the

Model	Western Europe	Eastern Europe	Middle East	North Africa	Subsaharan Africa	Central Asia	South Asia	SE Asia	East Asia	Americas	Avg.
7-10B Weight Class											
Qwen3 (8B)	**80.6**	**79.1**	**74.2**	**66.8**	**56.3**	**70.3**	**76**	**83**	**82.4**	**94.8**	**75.1**
Gemma 2 (9B)	78.1	76.1	70.5	64.8	43.7	65	71.1	79.5	75	93.2	70.4
Apertus (9B)	72.6	73.3	64.3	62	55.3	66	69.1	70.2	67.4	88.2	68.3
Aya Expanse (8B)	64.8	67.1	69.7	61.6	**56.3**	52.5	60.8	65	71	79.7	64.1
Llama 3.1 (8B)	66.6	64	62	55.6	50.6	55.7	61.5	67.5	68.4	81.8	62.2
Command R (7B)	60	59.2	64.2	60.2	50.9	50.8	59.3	60.8	68.6	72.8	59.5
12-20B Weight Class											
Gemma 3 (12B)	83.6	82.6	**79.8**	**78**	65.5	**78.5**	**80.9**	82.8	77.8	92.5	**79.5**
GPT-oss (20B)	**84.6**	81	79.6	73.8	**65.9**	75.5	79.3	86.3	81.2	**94.8**	79.1
Qwen 3 (14B)	84	**83.2**	76.6	71.8	57.6	75.8	80	**86.7**	**85.8**	**94.8**	78.5
Phi-4 (14B)	81.9	78.8	72.7	66	58	64.7	76	78.7	77	**94.8**	74.5
27-32B Weight Class											
Gemma 3 (27B)	**86.1**	**86.5**	**82.9**	**80.2**	**67.2**	**80.7**	**82.6**	**87.3**	82.2	95.8	**82.4**
Qwen 2.5 (32B)	84.6	78.9	78.1	72.2	60.2	65	75.1	86.5	**85**	**96**	76.8
Aya Exapanse (32B)	80.6	77.8	79.5	72.6	58.2	61.5	72.8	80	80.6	94.8	74.7
Command R (32B)	72.7	73.6	75.2	68.4	55.9	52.5	67.2	74.2	76	88	69.6
70-72B Weight Class											
Qwen 2.5 (72B)	**88.7**	**84.6**	**82**	**76**	61.5	**76**	77.7	**88.7**	**88.2**	**97.8**	**80.6**
Llama 3.1 (70B)	83.7	82.1	79.2	74.8	**66.2**	75.8	**79.8**	83.7	79.6	93.5	79.2
Apertus (70B)	77.7	78.2	73.8	70.4	61.7	70.5	73.1	77.2	74	91.2	74
Closed Models											
Gemini 2.5 Pro	**95.6**	**95.2**	**92.4**	**93.8**	**80.2**	**93.2**	**90**	92.3	91	97.5	**91.7**
Gemini 2.5 Flash	94.1	93.7	90.2	90.4	76.3	92.2	88.1	91.7	90.2	**97.8**	89.8
Claude Sonnet 4.5	94.3	93.4	88.7	88.4	76.4	90	88.3	**94.2**	71	97	89.1
GPT-5	94.7	93.9	89.2	89.6	70.4	**93.2**	83.4	93.5	**91.4**	**97.8**	88.3
GPT-5 mini	93.1	92.5	85.7	83.4	73.6	90.7	85.4	92.7	87.2	97.3	88.1
Gemini Flash-Lite	91.5	90	85.5	86.4	70.3	88	85.3	92	68.2	96	86
GPT-5 nano	81.6	80.2	75.2	70.6	52.2	75.5	72.4	86.2	78.6	95	75.7

Table 1: Aggregated accuracies across all regions. Results for all models are in §G.1. All results here are for instruction-tuned models evaluated with the prompted evaluation format (§5.1).

strongest open-weight models (Figure 3). We hope that Global PIQA will help direct progress towards closing the gap between open and proprietary LLMs.

Global PIQA also highlights languages for which state-of-the-art LLMs underperform. There are 18 languages for which the best-performing LLM achieves less than 90% accuracy; in ad hoc human evaluations for 12 languages, average human performance was 95.1% (§E). In fact, for seven languages, the top score is less than 80%: Burushaski (`bsk_arab`: 66%), Chakavian (`ckm_latn`: 74%), Ekpeye (`ekp_latn`: 56%), Idoma (`idu_latn`: 71%), Lingala (`lin_latn`: 68%), Manipuri (`mni_mtei`: 56%), and Urhobo (`urh_latn`: 62%). In Sub-Saharan African languages, even the best performing model only reaches an accuracy of 80.2% (Table 1). Notably, we find that in low-resource languages, there are systematically higher refusal rates (examples where the LLM refuses to answer or returns a null response) from some proprietary models, particularly GPT-5. In the seven languages where the best model accuracy was less than 80%, refusal rates from GPT-5, Sonnet 4.5, and Gemini 2.5 Pro were 43.9%, 0.0%, and 1.4% respectively. Refusal rate details are in §F.2, and we release the full list of best models per language on GitHub also in appendix?.

6 Discussion and Conclusion

In this paper, we present Global PIQA, a physical commonsense reasoning benchmark covering 116 language varieties. Unlike previous benchmarks, Global PIQA is a participatory benchmark, constructed by hand by 320 researchers across 65 countries; this enables the construction of a culturally-specific non-parallel split, where 59% of examples reference local foods, clothing, customs, traditions, or other culturally-specific elements. We find that proprietary LLMs perform well on Global PIQA, but there are still significant disparities for some languages and regions. Open weight models generally have lower accuracies than proprietary models, but Global PIQA allows researchers to clearly quantify the gap between open and proprietary models in multilingual settings. Notably, Global PIQA measures culturally-specific everyday knowledge, demonstrating that in many languages, areas for improvement can be as simple as everyday reasoning, rather than exclusively complex reasoning and expert knowledge.

Limitations. Of course, Global PIQA has several limitations. First, the sample size per language is only 100 examples; in the future, we hope that our participatory approach to benchmark construction will facilitate the construction of larger datasets. Second, we note that while Global PIQA contains culturally-specific examples, these examples are snapshots specific to our authors and researchers, not necessarily representative of entire cultures. Cultural stereotypes may be present in the dataset, although all examples are verified by native speakers of the languages. Finally, we emphasize that more languages is not necessarily better when constructing multilingual benchmarks; researchers should work with communities themselves to determine if and how they want their languages included. In Global PIQA, we have sought to work together with native speakers as authors, giving authors flexibility and ownership over how they construct their datasets.

Global PIQA v1. This paper currently describes Global PIQA v0.1; for Global PIQA v1, we plan for significant additions in the coming months. First, as discussed in §4, we are developing a parallel split of Global PIQA, which will double the size of the dataset. In addition, we are still looking for contributors to expand the language coverage of both the non-parallel and parallel splits of Global PIQA. Prospective contributors can register their interest through this form or visit the project website to get involved. We particularly welcome contributions for less-resourced languages and language varieties.

We close by noting that the scale of participation in this project far exceeded the organizers' expectations. The result is a manually curated, culturally-specific evaluation dataset with unprecedented language coverage. We are excited to continue developing community-led open-source multilingual evaluations, and we believe that this is an extremely promising avenue for addressing the critical lack of benchmarks for the vast majority of the world's languages.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. (2024a). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. (2024b). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Adebara, I., Elmadany, A., and Abdul-Mageed, M. (2024). Cheetah: Natural language generation for 517 African languages. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12798–12823, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Adelani, D. I., Ojo, J., Azime, I. A., Zhuang, J. Y., Alabi, J. O., He, X., Ochieng, M., Hooker, S., Bukula, A., Lee, E.-S. A., Chukwuneke, C. I., Buzaaba, H., Sibanda, B. K., Kalipe, G. K., Mukiibi, J., Kabongo Kabenamualu, S., Yuehgoh, F., Setaka, M., Ndolela, L., Odu, N., Mabuya, R., Osei, S., Muhammad, S. H., Samb, S., Guge, T. K., Sherman, T. V., and Stenetorp, P. (2025). IrokoBench: A new benchmark for African languages in the age of large language models. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. (2025). gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. (2023). Falcon-40B: an open large language model with state-of-the-art performance.

Alves, D. M., Pombal, J., Guerreiro, N. M., Martins, P. H., Alves, J., Farajian, A., Peters, B., Rei, R., Fernandes, P., Agrawal, S., Colombo, P., de Souza, J. G. C., and Martins, A. F. T. (2024). Tower: An open multilingual large language model for translation-related tasks.

An, S., Bae, K., Choi, E., Choi, K., Jungkyu Choi, S., Hong, S., Hwang, J., Jeon, H., Jeongwon Jo, G., Jo, H., et al. (2024). Exaone 3.5: Series of large language models for real-world use cases. *arXiv e-prints*, pages arXiv–2412.

Anthropic (2025). Claude sonnet 4.5 system card. Technical report, Anthropic. Version date as published (system card).

Arnett, C., Chang, T. A., and Bergen, B. K. (2024). A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the Annual Meeting of the Special Interest Group on Under-Resourced Languages*.

Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Bae, K., Choi, E., Choi, K., Jungkyu Choi, S., Choi, Y., Han, K., Hong, S., Hwang, J., Hwang, T., Jang, J., et al. (2025). Exaone 4.0: Unified large language models integrating non-reasoning and reasoning modes. *arXiv e-prints*, pages arXiv–2507.

Bak, Y., Lee, H., Ryu, M., Ham, J., Jung, S., Nam, D. W., Eo, T., Lee, D., Jung, D., Kim, B., et al. (2025). Kanana: Compute-efficient bilingual language models. *arXiv preprint arXiv:2502.18934*.

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., and Khabsa, M. (2024). The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Bisk, Y., Zellers, R., Le bras, R., Gao, J., and Choi, Y. (2020). PIQA: Reasoning about Physical Commonsense in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Chang, T. A., Arnett, C., Tu, Z., and Bergen, B. K. (2024). Goldfish: Monolingual language models for 350 languages. *Preprint*.

Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., and Choi, Y. (2025). CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Cohere, T., Ahmadian, A., Ahmed, M., Alammar, J., Alizadeh, M., Alnumay, Y., Althammer, S., Arkhangorodsky, A., Aryabumi, V., Aumiller, D., et al. (2025). Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2024). Tucano: Advancing neural text generation for portuguese. *arXiv preprint arXiv:2411.07854*.

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., et al. (2024). Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

DeepSeek-AI (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Dou, L., Liu, Q., Zhou, F., Chen, C., Wang, Z., Jin, Z., Liu, Z., Zhu, T., Du, C., Yang, P., Wang, H., Liu, J., Zhao, Y., Feng, X., Mao, X., Yeung, M. T., Pipatanakul, K., Koto, F., Thu, M. S., Kydlíček, H., Liu, Z., Lin, Q., Sripaisarnmongkol, S., Sae-Khow, K., Thongchim, N., Konkaew, T., Borijindargoon, N., Dao, A., Maneegard, M., Artkaew, P., Yong, Z.-X., Nguyen, Q., Phatthiyaphaibun, W., Tran, H. H., Zhang, M., Chen, S., Pang, T., Du, C., Wan, X., Lu, W., and Lin, M. (2025). Sailor2: Sailing in south-east asia with inclusive multilingual llm. *arXiv preprint arXiv:2502.12982*.

Ekgren, A., Cuba Gyllensten, A., Stollenwerk, F., Öhman, J., Isbister, T., Gogoulou, E., Carlsson, F., Casademont, J., and Sahlgren, M. (2024). GPT-SW3: An autoregressive language model for the Scandinavian languages. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Faysse, M., Fernandes, P., Guerreiro, N. M., Loison, A., Alves, D. M., Corro, C., Boizard, N., Alves, J., Rei, R., Martins, P. H., Casademunt, A. B., Yvon, F., Martins, A., Viaud, G., HUDELOT, C., and Colombo, P. (2025). CroissantLLM: A truly bilingual french-english language model. *Transactions on Machine Learning Research*.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. (2024). The language model evaluation harness.

Gen2B (2025). Hygpt 1.0: Technical report. Tech. report, Gen2B. Version 1.0, May 9, 2025.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Psychology Press.

Gonzalez-Agirre, A., Pàmies, M., Llop, J., Baucells, I., Da Dalt, S., Tamayo, D., Saiz, J. J., Espuña, F., Prats, J., Aula-Blasco, J., et al. (2025a). Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Gonzalez-Agirre, A., Pàmies, M., Llop, J., Baucells, I., Dalt, S. D., Tamayo, D., Saiz, J. J., Espuña, F., Prats, J., Aula-Blasco, J., Mina, M., Rubio, A., Shvets, A., Sallés, A., Lacunza, I., Pikabea, I., Palomar, J., Falcão, J., Tormo, L., Vasquez-Reina, L., Marimon, M., Ruíz-Fernández, V., and Villegas, M. (2025b). Salamandra technical report.

Google DeepMind (2025a). Gemini 2.5 flash-lite model card. Technical report, Google DeepMind. Model card, published via DeepMind media server.

Google DeepMind (2025b). Gemini 2.5 flash model card. Technical report, Google DeepMind. Model card, stable release.

Google DeepMind (2025c). Gemini 2.5 pro model card. Technical report, Google DeepMind. Last updated June 27, 2025.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2023). *Glottolog 4.8*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., Hakimi, I., et al. (2025). Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*.

Hu, J. and Frank, M. (2024). Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Hupkes, D. and Bogoychev, N. (2025). MultiLoKo: a multilingual local knowledge benchmark for llms spanning 31 languages. *arXiv preprint arXiv:2504.10356*.

Inception (2024). Jais family model card.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*. Version v1.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., and Bergen, B. (2022). Distrubutional Semantics Still Can't Account for Affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Koto, F., Aisyah, N., Li, H., and Baldwin, T. (2023). Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., and Baldwin, T. (2024). ArabicMMLU: Assessing massive multitask language understanding in Arabic. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Kreutzer, J., Briakou, E., Agrawal, S., Fadaee, M., and Kocmi, T. (2025). Déjà vu: Multilingual LLM evaluation through the lens of machine translation evaluation. In *Second Conference on Language Modeling*.

Lai, V., Nguyen, C., Ngo, N., Nguyen, T., Dernoncourt, F., Rossi, R., and Nguyen, T. (2023). Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In Feng, Y. and Lefever, E., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., and Baldwin, T. (2024). CMMLU: Measuring massive multitask language understanding in Chinese. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. (2021). Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., Stoyanov, V., and Li, X. (2022). Few-shot learning with multilingual generative language models. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Martins, P. H., Alves, J., Fernandes, P., Guerreiro, N. M., Rei, R., Farajian, A., Klimaszewski, M., Alves, D. M., Pombal, J., Boizard, N., et al. (2025). Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Meta AI (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Michaelov, J., Arnett, C., and Bergen, B. (2024). Revenge of the fallen? recurrent models match transformers at predicting human language comprehension metrics. In *First Conference on Language Modeling*.

Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R. A., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., Gutiérrez-Basulto, V., Ibáñez García, Y., Lee, H., Muhammad, S. H., Park, K., Rzayev, A. S., White, N., Yimam, S. M., Pilehvar, M. T., Ousidhoum, N., Camacho-Collados, J., and Oh, A. (2024). Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Ng, R., Nguyen, T. N., Huang, Y., Tai, N. C., Leong, W. Y., Leong, W. Q., Yong, X., Ngui, J. G., Susanto, Y., Cheng, N., et al. (2025). SEA-LION: Southeast asian languages in one network. *arXiv preprint arXiv:2504.05747*.

Nguyen, D. Q., Nguyen, L. T., Tran, C., Nguyen, D. N., Phung, D., and Bui, H. (2023a). Phogpt: Generative pre-training for vietnamese. *arXiv preprint arXiv:2311.02945*.

Nguyen, Q., Pham, H., and Dao, D. (2023b). Vinallama: Llama-based vietnamese foundation model. *arXiv preprint arXiv:2312.11011*.

Ociepa, K. and Azurro Team (2024). Introducing apt3-1b-base: Polish language model.

Ociepa, K., Łukasz Flis, Kinas, R., Wróbel, K., and Gwoździej, A. (2025). Bielik v3 small: Technical report.

OpenAI (2024). Multilingual massive multitask language understanding (mmmlu).

OpenAI (2025). Gpt-5 system card. Technical report, OpenAI. Version dated August 7, 2025.

Owen, L., Tripathi, V., Kumar, A., and Ahmed, B. (2024). Komodo: a linguistic expedition into indonesia's regional languages. *arXiv preprint arXiv:2403.09362*.

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., and Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Romanou, A., Foroutan, N., Sotnikova, A., Nelaturu, S. H., Singh, S., Maheshwary, R., Altomare, M., Chen, Z., Haggag, M. A., A, S., Amayuelas, A., Amirudin, A. H., Boiko, D., Chang, M., Chim, J., Cohen, G., Dalmia, A. K., Diress, A., Duwal, S., Dzenhaliou, D., Florez, D. F. E., Farestam, F., Imperial, J. M., Islam, S. B., Isotalo, P., Jabbarishiviari, M., Karlsson, B. F., Khalilov, E., Klamm, C., Koto, F., Krzemiński, D., de Melo, G. A., Montariol, S., Nan, Y., Niklaus, J., Novikova, J., Ceron, J. S. O., Paul, D., Ploeger, E., Purbey, J., Rajwal, S., Ravi, S. S., Rydell, S., Santhosh, R., Sharma, D., Skenduli, M. P., Moakhar, A. S., soltani moakhar, B., Tarun, A. K., Wasi, A. T., Weerasinghe, T. O., Yilmaz, S., Zhang, M., Schlag, I., Fadaee, M., Hooker, S., and Bosselut, A. (2025). INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Rostami, P., Salemi, A., and Dousti, M. J. (2024). PersianMind: A Cross-Lingual Persian-English Large Language Model.

Roussis, D., Voukoutis, L., Paraskevopoulos, G., Sofianopoulos, S., Prokopidis, P., Papavasileiou, V., Katsamanis, A., Piperidis, S., and Katsouros, V. (2025). Krikri: Advancing open large language models for greek.

Sarvam AI (2025). Sarvam-m: Explorations in post training and inferencing optimizations for a hybrid indic llm. `https://www.sarvam.ai/blogs/sarvam-m`.

Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., Afzal, O. M., Kamboj, S., Pandit, O., Pal, R., Pradhan, L., Mujahid, Z. M., Baali, M., Han, X., Bsharat, S. M., Aji, A. F., Shen, Z., Liu, Z., Vassilieva, N., Hestness, J., Hock, A., Feldman, A., Lee, J., Jackson, A., Ren, H. X., Nakov, P., Baldwin, T., and Xing, E. (2023). Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Seth, A., Ahuja, S., Bali, K., and Sitaram, S. (2024). DOSA: A dataset of social artifacts from different Indian geographical subcultures. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., and Wei, J. (2023). Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ruder, S., Ko, W.-Y., Bosselut, A., Oh, A., Martins, A., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., and Hooker, S. (2025). Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Son, G., Lee, H., Kim, S., Kim, S., Muennighoff, N., Choi, T., Park, C., Yoo, K. M., and Biderman, S. (2025). KMMLU: Measuring massive multitask language understanding in Korean. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Team, Q. (2025). Qwen3 technical report.

Team Gemma, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Team Gemma, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Thellmann, K., Stadler, B., Fromm, M., Buschhoff, J. S., Jude, A., Barth, F., Leveling, J., Flores-Herr, N., Köhler, J., Jäkel, R., et al. (2024). Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

Üstün, A., Aryabumi, V., Yong, Z., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., and Hooker, S. (2024). Aya model: An instruction finetuned open-access multilingual language model. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Voukoutis, L., Roussis, D., Paraskevopoulos, G., Sofianopoulos, S., Prokopidis, P., Papavasileiou, V., Katsamanis, A., Piperidis, S., and Katsouros, V. (2024). Meltemi: The first open large language model for greek.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Wu, M., Wang, W., Liu, S., Yin, H., Wang, X., Zhao, Y., Lyu, C., Wang, L., Luo, W., and Zhang, K. (2025). The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. *arXiv preprint arXiv:2504.15521*.

Xuan, W., Yang, R., Qi, H., Zeng, Q., Xiao, Y., Feng, A., Liu, D., Xing, Y., Wang, J., Gao, F., et al. (2025). Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. (2024a). Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*. Version v2 (revised 3 January 2025).

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. (2024b). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yoo, K. M., Han, J., In, S., Jeon, H., Jeong, J., Kang, J., Kim, H., Kim, K.-M., Kim, M., Kim, S., et al. (2024). Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Yüksel, A., Köksal, A., Senel, L. K., Korhonen, A., and Schuetze, H. (2024). TurkishMMLU: Measuring massive multitask language understanding in Turkish. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zhang, W., Chan, H. P., Zhao, Y., Aljunied, M., Wang, J., Liu, C., Deng, Y., Hu, Z., Xu, W., Chia, Y. K., Li, X., and Bing, L. (2025). SeaLLMs 3: Open foundation and chat multilingual large language models for Southeast Asian languages. In Dziri, N., Ren, S. X., and Diao, S., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 96–105, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhao, Y., Liu, C., Deng, Y., Ying, J., Aljunied, M., Li, Z., Bing, L., Chan, H. P., Rong, Y., Zhao, D., et al. (2025). Babel: Open multilingual large language models serving over 90% of global speakers. *arXiv preprint arXiv:2503.00865*.

Zosa, E., Luoma, J., Hakala, K., Virtanen, A., Koistinen, M., Luukkonen, R., Reunamo, A., Pyysalo, S., and Burdge, J. (2025). Poro 2: Continued pretraining for language acquisition. LumiOpen.

A Author Contributions

Global PIQA would not be possible without the efforts of all of the authors. We note that we intentionally do not list authors by contributed language. This is to preserve privacy, as some authors would prefer not to be contacted by a large number of unaffiliated projects that require expertise in their language. Correspondence should be sent to the lead authors (`tachang@ucsd.edu` and `catherine@eleuther.ai`) or to `mrl.benchmarks@gmail.com`. Global PIQA is a community effort, and it does not necessarily reflect the opinions or views of the authors' affiliated organizations.

Leads
Tyler A. Chang*, UC San Diego
Catherine Arnett*, EleutherAI

*Joint first authorship.

Contributors[7]
Abdelrahman Eldesokey, King Abdullah University of Science and Technology (KAUST)
Abdelrahman Sadallah, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Abeer Kashar, University of Waterloo
Abolade Daud, Masakhane
Abosede Grace Olanihun, Obafemi Awolowo University
Adamu Labaran Mohammed, Independent
Adeyemi Praise, Tonative
Adhikarinayum Meerajita Sharma, Banasthali Vidyapith
Aditi Gupta, International Institute of Information Technology Hyderabad
Afitab Iyigun, Boston University
Afonso Simplício, NOVA School of Science and Technology, NOVA University Lisbon
Ahmed Essouaied, Higher School of Communication of Tunis
Aicha Chorana, University of Laghouat
Akhil Eppa, Independent
Akintunde Oladipo, The African Research Collective

[7] In alphabetical order by first name.

Akshay Ramesh, Vellore Institute of Technology - Chennai
Aleksei Dorkin, University of Tartu
Alfred Malengo Kondoro, Hanyang University
Alham Fikri Aji, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Ali Eren Çetintaş, Middle East Technical University
Allan Hanbury, TU Wien
Alou Dembele, RobotsMali
Alp Niksarli, Davidson College
Amin Bajand, Linköping University
Amol Khanna, Booz Allen Hamilton
Ana Chkhaidze, University of California San Diego
Ana Condez, NOVA School of Science and Technology, NOVA University Lisbon
Andiswa Mkhonto, Independent
Andrew Hoblitzell, Purdue University
Andrew Tran, Independent
Angelos Poulis, Boston University
Anirban Majumder, Amazon Science (work done independently, outside of their role at Amazon)
Anna Vacalopoulou, Institute for Speech and Language Processing, Athena Research Center
Annette Kuuipolani Kanahele Wong, University of Hawai'i at Mānoa
Annika Simonsen, University of Iceland
Anton Kovalev, University of Massachusetts Lowell
Ashvanth.S, Cohere Labs Community
Barkin Kinay, Robert College
Bashar Alhafni, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Benedict Cibalinda Busole, Independent
Bernard Ghanem, King Abdullah University of Science and Technology (KAUST)
Bharti Nathani, Banasthali Vidyapith
Biljana Stojanovska, University of Rijeka
Bola Agbonile, Zabbot LLC
Bragi Bergsson, Independent
Bruce Torres Fischer, University of Hawai'i at Hilo
Burak Tutar, Middle East Technical University
Burcu Alakuş Çınar, Middle East Technical University
Cade J. Kanoniakapueo Kane, University of Hawai'i at Mānoa
Can Udomcharoenchaikit, Vidyasirimedhi Institute of Science and Technology
Catherine Arnett, Eleuther AI
Chadi Helwe, King Abdullah University of Science and Technology (KAUST)
Chaithra Reddy Nerella, International Institute of Information Technology Hyderabad
Chen Cecilia Liu, Independent
Chiamaka Glory Nwokolo, University of Ibadan
Cristina España-Bonet, German Research Center for Artificial Intelligence (DFKI), Barcelona Supercomputing Center (BSC)
Cynthia Amol, Maseno University
DaeYeop Lee, Pohang University of Science and Technology, Modulabs
Dana Arad, Technion – Israel Institute of Technology
Daniil Dzenhaliou, École Polytechnique Fédérale de Lausanne (EPFL)
Daria Pugacheva, Artificial Intelligence Research Institute (AIRI)
Dasol Choi, Yonsei University, AIM Intelligence
Daud Abolade, University of Lagos
David Liu, Boston University
David Semedo, NOVA School of Science and Technology, NOVA University Lisbon
David Stap, NXAI
Deborah Popoola, Tonative
Deividas Mataciunas, AQ22
Delphine Nyaboke, Independent
Dhyuthy Krishna Kumar, Independent
Diogo Glória-Silva, NOVA School of Science and Technology, NOVA University Lisbon
Diogo Tavares, NOVA School of Science and Technology, NOVA University Lisbon
Divyanshu Goyal, Independent
DongGeon Lee, Pohang University of Science and Technology
Duygu Ataman, Middle East Technical University
Ebele Nwamaka Anajemba, Nnamdi Azikiwe University, Awka
Egonu Ngozi Grace, Alvan Ikoku Federal College of Education, Owerri
Elena Mickel, Independent
Elena Tutubalina, Artificial Intelligence Research Institute (AIRI)
Elias Herranen, Independent
Emile Anand, Cognition AI
Emmanuel Habumuremyi, Rwanda Journalists Association
Emuobonuvie Maria Ajiboye, Delta State University, Abraka
Eryawan Presma Yulianrifat, Universitas Indonesia
Esther Adenuga, The African Research Collective
Ewa Rudnicka, Wrocław University of Science and Technology
Fabian Schmidt, University of Würzburg
Faith Olabisi Itiola, University of Ibadan
Faran Taimoor Butt, Moscow Institute of Physics and Technology (MIPT)
Fathima Thekkekara, Independent
Fatima Haouari, University of Sheffield
Filbert Aurelian Tjiaranata, Universitas Indonesia

Firas Laakom, King Abdullah University of Science and Technology (KAUST)
Francesca Grasso, University of Turin
Francesco Orabona, King Abdullah University of Science and Technology (KAUST)
Francesco Periti, KU Leuven, Flanders Make
Gbenga Kayode Solomon, Adekunle Ajasin University
Gia Nghia Ngo, True North International School
Gloria Udhehdhe-oze, University of Port Harcourt
Gonçalo Martins, NOVA School of Science and Technology, NOVA University Lisbon
Gopi Naga Sai Ram Challagolla, Independent
Guijin Son, OneLineAI
Gulnaz Abdykadyrova, Independent
Hafsteinn Einarsson, University of Iceland
Hai Hu, City University of Hong Kong
Hamidreza Saffari, Polytechnic University of Milan
Hamza Zaidi, University of Waterloo
Haopeng Zhang, University of Hawai'i at Mānoa
Harethah Abu Shairah, King Abdullah University of Science and Technology (KAUST)
Harry Vuong, Independent
Hele-Andra Kuulmets, University of Tartu
Houda Bouamor, Carnegie Mellon University in Qatar
Hwanjo Yu, Pohang University of Science and Technology
Iben Nyholm Debess, University of the Faroe Islands
Ikhlasul Akmal Hanif, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Ikhyun Cho, University of Illinois Urbana-Champaign
Inês Calvo, NOVA School of Science and Technology, NOVA University Lisbon
Inês Vieira, NOVA School of Science and Technology, NOVA University Lisbon
Isaac Manzi, Independent
Ismail Daud, University of Ibadan
Itay Itzhak, Technion – Israel Institute of Technology
Iuliia (Julia) Alekseenko, IHU Strasbourg, University of Strasbourg, CNRS, INSERM
Ivan Belashkin, Independent
Ivan Spada, University of Turin
Ivan Zhelyazkov, Independent
Jafar Isbarov, Virginia Tech
Jaka Čibej, University of Ljubljana
Jake Brinton, Boston University
Jan Kocoń, Wrocław University of Science and Technology
Jan Čuhel, Independent
Jauza Akbar Krito, Universitas Gadjah Mada
Jebish Purbey, Cohere Labs Community
Jennifer Mickel, EleutherAI Community
Jennifer Za Nzambi, Independent
Jenny Kunz, Linköping University
Jihae Jeong, Pohang University of Science and Technology
Jimena Tena Dávalos, Universidad Michoacana de San Nicolás de Hidalgo
Jinu Lee, University of Illinois Urbana-Champaign
John Yi, Boston University
Jongin Kim, Boston University
Joseph Chataignon, University of Bern
Joseph Marvin Imperial, University of Bath, National University Philippines
João Magalhães, NOVA School of Science and Technology, NOVA University Lisbon
Jubeerathan Thevakumar, University of Moratuwa
Judith Land, Independent
Junchen Jiang, Shanghai Jiao Tong University
Jungwhan Kim, NAVER Cloud
Kairit Sirts, University of Tartu
Kamesh R, Sathyabama Institute of Science and Technology
Kamesh V, Sathyabama Institute of Science and Technology
Kanda Patrick Tshinu, Tshwane University of Technology
Kaustubh Ponkshe, École Polytechnique Fédérale de Lausanne (EPFL)
Kavsar Huseynova, Baku Higher Oil School
Ke He, Shanghai Jiao Tong University
Kelly Buchanan, Stanford University
Kengatharaiyer Sarveswaran, University of Jaffna
Kerem Zaman, University of North Carolina at Chapel Hill
Khalil Mrini, Oracle
Kian Kyars, Independent
Krister Kruusmaa, Tallinn University
Kusum Chouhan, Banasthali Vidyapith
Kätriin Kukk, Linköping University
Lainitha Krishnakumar, University of Moratuwa
Lana Ayodeji Joseph, Masakhane
Laura Castro Sánchez, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
Laura Porrino Moscoso, Universidad Alfonso X El Sabio
Leshem Choshen, MIT, MIT-IBM Watson AI Lab
Levent Sencan, Boston University
Lilja Øvrelid, University of Oslo
Lisa Alazraki, Imperial College London
Lovina Ehimen-Ugbede, University of Alicante
Luheerathan Thevakumar, Independent
Luxshan Thavarasa, University of Moratuwa
Mahnoor Malik, NED University of Engineering and Technology

Mamadou K. Keita, Rochester Institute of Technology
Mansi Jangid, Banasthali Vidyapith
Marco De Santis, University of Udine
Marcos García, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
Marek Suppa, Cisco Systems
Mariam D’Ciofalo, Independent
Marii Ojastu, University of Tartu
Maryam Sikander, Independent
Mausami Narayan, Independent
Maximos Skandalis, Laboratoire d’Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), CNRS, University of Montpellier
Mehak Mehak, Independent
Mehmet İlteriş Bozkurt, Middle East Technical University
Melaku Bayu Workie, Addis Ababa University
Menan Velayuthan, University of Jaffna
Michael Leventhal, RobotsMali
Michał Marcińczuk, CodeNLP (Gdańsk, Poland)
Mirna Potočnjak, Independent
Mohammadamin Shafiei, University of Milan
Mridul Sharma, Institute for Research and Innovation in Intelligent Systems (IRIIS)
Mrityunjaya Indoria, Banasthali Vidyapith
Muhammad Ravi Shulthan Habibi, Universitas Indonesia
Murat Kolić, Independent
Nada Galant, Čakavski sabor
Naphat Permpredanun, Independent
Narada Maugin, Paris Cité University
Nicholas Kluge Corrêa, University of Bonn
Nikola Ljubešić, Jožef Stefan Institute
Nirmal Thomas, Pratham International
Nisansa de Silva, University of Moratuwa
Nisheeth Joshi, Banasthali Vidyapith
Nitish Ponkshe, University of Minnesota Twin Cities
Nizar Habash, NYU Abu Dhabi
Nneoma C. Udeze, Northwestern University
Noel Thomas, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Nouhoum Coulibaly, RobotsMali
Noémi Ligeti-Nagy, Eötvös Loránd University (ELTE), Research Centre for Linguistics
Nsengiyumva Faustin, University of Rwanda
Odunayo Kareemat Buliaminu, University of Benin
Odunayo Ogundepo, The African Research Collective
Oghojafor Godswill Fejiro, Delta State University, Abraka
Ogundipe Blessing Funmilola, University of Ibadan
Okechukwu God’spraise, Tonative
Olanrewaju Samuel, Stonybrook University
Olaoye Deborah Oluwaseun, University of Ilorin
Olasoji Akindejoye, University of Ibadan
Olga Popov, Artificial Intelligence Research Institute (AIRI)
Olga Snissarenko, Kazakhstan Branch of Lomonosov Moscow State University
Onyinye Anulika Chiemezie, Nnamdi Azikiwe University, Awka
Orkun Kinay, University of Edinburgh
Osman Tursun, Queensland University of Technology
Owoeye Tobiloba Moses, University of Ibadan
Oyelade Oluwafemi Joshua, University of Ilorin
Oyesanmi Fiyinfoluwa, University of Johannesburg
Pablo Gamallo, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
Pablo Rodríguez Fernández, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
Palak Arora, DIT Univeristy
Pedro Valente, NOVA School of Science and Technology, NOVA University Lisbon
Peter Rupnik, Jožef Stefan Institute
Philip Oghenesuowho Ekiugbo, National Institute for Nigerian Languages, Aba
Pramit Sahoo, Independent
Prokopis Prokopidis, Institute for Speech and Language Processing, Athena Research Center
Pua Niau-Puhipau, University of Hawai’i at Mānoa
Quadri Yahya, University of Abuja
Rachele Mignone, University of Turin
Raghav Singhal, École Polytechnique Fédérale de Lausanne (EPFL)
Ram Mohan Rao Kadiyala, Cohere Labs Community
Raphael Merx, The University of Melbourne
Rapheal Afolayan, University of Ilorin
Ratnavel Rajalakshmi, Vellore Institute of Technology - Chennai
Rishav Ghosh, Independent
Romina Oji, Linköping University
Ron Kekeha Solis, University of Hawai’i at Mānoa
Rui Guerra, NOVA School of Science and Technology, NOVA University Lisbon
Rushikesh Zawar, Independent
Sa’ad Nasir Bashir, Bayero University Kano
Saeed Alzaabi, NYU Abu Dhabi
Sahil Sandeep, Vellore Institute of Technology - Chennai
Sai Pavan Batchu, Independent
SaiSandeep Kantareddy, Independent
Salsabila Zahirah Pranida, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Sam Buchanan, University of California Berkeley

Sander Land, Writer Inc.
Sarah Sulollari, University of Vienna
Sardar Ali, Independent
Saroj Sapkota, Institute for Research and Innovation in Intelligent Systems (IRIIS)
Saulius Tautvaisas, Independent
Sayambhu Sen, Amazon Alexa
Sayantani Banerjee, IIT Madras
Sebastien Diarra, RobotsMali
SenthilNathan.M, Independent
Sewoong Lee, University of Illinois Urbana-Champaign
Shaan Shah, University of California San Diego
Shankar Venkitachalam, Independent
Sharifa Djurabaeva, Dennis-Yarmouth High School
Sharon Ibejih, Tonative
Shivanya Shomir Dutta, Vellore Institute of Technology - Chennai
Siddhant Gupta, IIT Roorkee
Silvia Paniagua Suárez, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
Sina Ahmadi, University of Zurich
Sivasuthan Sukumar, University of Moratuwa
Siyuan Song, University of Texas at Austin
Snegha A., IIT Bombay
Sokratis Sofianopoulos, Institute for Speech and Language Processing, Athena Research Center
Sona Elza Simon, IIT Bombay
Sonja Benčina, Parafraza
Sophie Gvasalia, Lightcast
Sphurti Kirit More, Independent
Spyros Dragazis, Boston University
Stephan Kaufhold, University of California San Diego
Suba.S, Independent
Sultan AlRashed, King Abdullah University of Science and Technology (KAUST)
Surangika Ranathunga, Massey University
Taiga Someya, The University of Tokyo
Taja Kuzman Pungeršek, Jožef Stefan Institute
Tal Haklay, Technion – Israel Institute of Technology
Tasi'u Jibril, Bayero University Kano
Tatsuya Aoyama, Georgetown University
Tea Abashidze, Independent
Terenz Jomar Dela Cruz, Independent
Terra Blevins, Northeastern University
Themistoklis Nikas, Boston University
Theresa Dora Idoko, Benue State University
Thu Mai Do, The Dewey Schools Hanoi
Tilek Chubakov, Independent Researcher
Tommaso Gargiani, Independent
Uma Rathore, Banasthali Vidyapith
Uni Johannesen, University of the Faroe Islands
Uwuma Doris Ugwu, Ignatius Ajuru University of Education
Vallerie Alexandra Putra, Bina Nusantara University
Vanya Bannihatti Kumar, Independent
Varsha Jeyarajalingam, University of Jaffna
Varvara Arzt, TU Wien
Vasudevan Nedumpozhimana, Trinity College Dublin Ireland
Viktoria Ondrejova, Comenius University Bratislava
Viktoryia Horbik, Independent
Vishnu Vardhan Reddy Kummitha, Independent
Vuk Dinić, Independent
Walelign Tewabe Sewunetie, African Institute of Mathematical Sciences (AIMS) Research and Innovation Centre (RIC)
Winston Wu, University of Hawai'i at Hilo
Xiaojing Zhao, Hong Kong Polytechnic University
Yacouba Diarra, RobotsMali
Yaniv Nikankin, Technion – Israel Institute of Technology
Yash Mathur, Independent
Yixi Chen, Zhejiang University
Yiyuan Li, University of North Carolina at Chapel Hill
Yolanda Xavier, Linguistics Research Centre of NOVA University Lisbon
Yonatan Belinkov, Technion – Israel Institute of Technology, Kempner Institute, Harvard University
Yusuf Ismail Abayomi, Obafemi Awolowo University
Zaid Alyafeai, King Abdullah University of Science and Technology (KAUST)
Zhengyang Shan, Boston University
Zhi Rui Tam, National Taiwan University
Zilu Tang, Boston University
Zuzana Nadova, Universidad del País Vasco
Álvaro Arroyo, University of Oxford
İbrahim Ethem Deveci, Middle East Technical University

Evaluation Infrastructure

Baber Abbasi, EleutherAI
Stella Biderman, EleutherAI

Workshop Organizers

Catherine Arnett, EleutherAI
David Stap, University of Amsterdam
Duygu Ataman, New York University
Fabian Schmidt, University of Würzburg
Hila Gonen, University of Washington
Jiayi Wang, University College London
Tyler A. Chang, UC San Diego
David Ifeoluwa Adelani, McGill University and Mila

Acknowledgements: We also thank several anonymous contributors who preferred not to be authors on this paper. The research of Yolanda Xavier is supported by Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UID/3213/2025 – Linguistics Research Centre of NOVA University Lisbon (CLUNL) and by the Doctoral Grant (FCT PhD grant) number 2022.13977.BD from the same funder. Group 0133 would like to thank the MbazaNLP community, including students from the University of Rwanda, School of Art and Languages.

B Comparison to Existing Benchmarks

C Language Codes and Included Languages

We normalize all language codes in Global PIQA to use ISO 639-3 individual language codes (three letters), ISO 15924 script codes (four letters), and an optional custom four-letter region code for dialects within an individual language code. For example, the code for Mexican Spanish is `spa_latn_mexi`, and the code for Peninsular Spanish (as spoken in Spain) is `spa_latn_spai`. When the language code was unclear for an individual dataset based on the description from the authors, we worked with authors to identify the specific ISO 639-3 and ISO 15924 codes that would best reflect their dataset.

For clarity, we note:

- ISO 639-3 macrolanguage codes are often used in other work for some languages. We use individual language codes for more precision, and here, we show mappings from commonly-used macrolanguage codes to the ISO 639-3 individual codes used in Global PIQA:
  - Mandarin Chinese: `zho` → `cmn`
    Cantonese Chinese: `zho` → `yue`
    Standard Estonian: `est` → `ekk`
    Norwegian Bokmål: `nor` → `nob`
    Norwegian Nynorsk: `nor` → `nno`
    Nepali: `nep` → `npi`
    Iranian Persian (Farsi): `fas` → `pes`
    Swahili (Kiswahili): `swa` → `swh`
    Northern Uzbek: `uzb` → `uzn`
    Standard Malay: `msa` → `zsm`
    Central Kurdish: `kur` → `ckb`
- Dialects of Arabic are often separate individual language codes. In Global PIQA, we have: list here.
- A Filipino dataset (language code `fil`) separate from the Tagalog dataset (language code `tgl`) was not included, despite the two being considered separate individual language codes in ISO 639-3. This is because native speakers of Tagalog often refer to the two languages interchangeably; Filipino is the standardized national language of the Philippines, but it draws influence primarily from Tagalog.

Using these language, script, and optional region codes, Global PIQA contains 116 unique language varieties. This includes N unique ISO language-script combinations, N unique ISO 639-3 language codes, and 23 unique ISO 15924 script codes. Language counts per region, family, and resource level are shown in Tables X, X, and X respectively. Language families use ... Glottolog? Regions use... ? Resource levels from Joshi et al.

A full table of languages will go here!

D Dataset Cleaning, Compilation, and Sampling Details

D.1 Non-Parallel Split: Cleaning and Compilation

As described in §3, authors contributed datasets to the Global PIQA non-parallel split for their own language(s). At minimum, each dataset contributed to Global PIQA contained a `prompt`, `solution0`,

Region	# Langs
Europe	40
South Asia	26
Subsaharan Africa	14
Middle East	13
Southeast Asia	6
East Asia	5
Central Asia	4
North Africa	4
North America	2
South America	2
Oceania	1

Table 2: Table 1

language_family	language
Indo-European	60
Afro-Asiatic	16
Atlantic-Congo	10
Austronesian	6
Turkic	6
Sino-Tibetan	5
Dravidian	4
Uralic	3
Japonic	1
Kartvelian	1
Koreanic	1
Mande	1
Nilotic	1
Tai-Kadai	1
isolate	1

Table 3: Table 2

Resource Level	language
0	11
1	32
2	8
3	24
4	19
5	23

Table 4: Joshi et al. (2020) resource levels

Table 5: re-generate this, based on old list of languages

solution1, and label column. For each dataset, we first removed exact duplicate examples and invalid examples where the two solutions were identical. We normalized column names, moved supplemental information (e.g. "topic" fields or other columns added by individual groups) to a supplement column, and we converted all text fields to use UTF-8 text encoding. For transparency, we annotated any examples that used LLMs to initially generate the example; this is a relatively small number of examples (N%before subsampling, then N%in the official non-parallel split), and all examples are human validated (see method descriptions in §H). For several datasets, we found that sentence completion examples (i.e. examples where the prompt is an incomplete sentence, and the candidate solutions complete the sentence) contained prompts ending with ellipses ("...") or underscores ("___", i.e. fill-in-the-blank). We removed these ending ellipses and underscores, as the completions are concatenated directly onto the prompts when fed into LLMs in the completion setup (§F).

As a preliminary check, we used Google Translate to translate a random subset of ∼20 examples per dataset, to identify any egregious errors (e.g. all examples far too easy, not following the task format, or large numbers of repetitive examples). Based on this preliminary check, if any datasets were clearly not culturally specific (see annotation guidelines in §D.2), we asked the dataset authors for optional revisions to add more culturally-specific examples. In these cases, we asked authors to modify or add examples to include words that are unlikely to translate well into other languages, such as food words, words for types of clothing, or local brand names.

After this initial cleaning and revision, to better inspect the data, we used Gemini 2.5 Pro to translate each prompt+solution into English for all datasets. The translation prompt used is in Figure X, and we accessed Gemini 2.5 Pro (October 2025) through Google's Gemini API using a paid API key. The resulting machine translations are available in the publicly-released Global PIQA dataset. Of course, we note that translation quality into English is likely worse for lower-resource languages; for example, the English machine translations for Burushaski (bsk_arab) examples appear consistently poor. Even in high-resource languages, the machine translations sometimes correct the incorrect solution when translating a prompt with the incorrect solution. Still, based on the translations, for many languages we were able to spot-check labels in the datasets for examples that had clear correct answers. From this cursory verification, we found two datasets with systematic errors where the annotated labels were often flipped to be incorrect; we worked with the authors of these datasets to correct and revalidate the labels.

Finally, we combined all datasets per language, and we added unique example IDs including group (i.e. dataset) number, example index, and language code. For groups that submitted parallel datasets in multiple languages (e.g. Group 0065 for eight dialects of Arabic, or Group 0042 for Catalan and Peninsular Spanish), the parallel examples have the same group number and example index, only differing in language code. This allows the small number of parallel examples in the non-parallel split of the Global PIQA dataset to still be found.

D.2 Non-Parallel Split: English Annotations of Task Adherence and Cultural Specificity

Next, we used the machine-generated English translations of all examples to annotate cultural specificity and loose adherence to the task description. Annotations were completed by one of the primary authors, who is a native English speaker.

Task adherence. We use the following guidelines for adherence to the task description:

1. Drop examples that consist of a complex or abstract logical problem, as these do not fit the task description of physical commonsense reasoning. For example, we drop complex logic puzzles and computer programming questions.
2. Drop examples that appear both generic and extremely easy based on the English translation. For example, we drop examples such as "*When you heat water, it becomes [hot/cold]*".
3. Keep examples that query common knowledge about locations (e.g. locations of cities or famous monuments, or common events to observe in particular cities). We count these under a loose definition of physical commonsense.
4. Keep examples that query social or cultural knowledge. These examples often describe regional customs, norms, and traditions, which we would like to remain present in Global PIQA. Additionally, these examples may arguably still be considered physical commonsense due to the embodied nature of everyday human interactions.
5. Where possible, drop examples that query obscure historical factoids. In some languages, there are too few total examples to drop all such examples, so a small number of historical knowledge questions may still be present in the dataset. These examples are generally apparent from the machine translations to English.

Based on these guidelines, we dropped approximately N out of Nexamples in the submitted datasets, before subsampling for the official non-parallel split. In cases where this filtering caused the number of examples in a language to drop below 100 examples, we worked directly with authors to reach the 100 example minimum. We note that given our fairly flexible definitions of physical commonsense reasoning, we do not guarantee that the entire Global PIQA dataset evaluates physical commonsense reasoning in a strict sense; some examples may be better categorized as social commonsense, cultural knowledge, or common knowledge.

Cultural specificity. Because cultural specificity is fairly subjective and perspective-dependent, we attempt to provide clear guidelines for when we annotated an example as "culturally-specific". Our definition of culturally-specific covers both culturally-*specific* examples, i.e. examples that are only relevant in a specific region or language, and culturally-*sensitive* examples, i.e. examples whose solution varies across regions or languages (Citations?). When we use the term "culturally specific", we refer to this broad definition. We formulated the guidelines here in an attempt to reduce potential bias and the presence of stereotypes in our annotations of cultural specificity. We annotate examples for cultural specificity using these guidelines:

1. Some datasets have some examples marked as culturally specific by the dataset authors. We annotate these examples as culturally specific; this defers to the authors (members of the cultural communities) to choose examples that they believe reflect their culture, giving more ownership back to the communities themselves.
2. We annotate examples as culturally specific if they describe specific holidays, folklore, traditions, sayings, or aphorisms in the language.
3. We annotate an example as culturally specific if its solution likely varies by region. For example, traffic rules and social norms are likely to vary across regions.
4. If an example contains a word that does not translate well into English, then we annotate it as culturally specific. This can include words for local food dishes, traditional objects or articles of clothing, or local brands. We do *not* count city names (or person names), as many examples that simply mention a city are not actually specific to that city. We acknowledge that some words are ambiguously "English" vs. borrowed from another language; in these cases, we use our best judgment based on how commonly the word is used in English.

5. We do *not* count the presence of local ingredients or objects if they have widely used English words, such as corn, rice, beans, or many fruits and vegetables, even if these items vary in popularity across regions. In other words, we do not annotate an example as culturally specific solely based on the presence of these items. This guideline aims to reduce bias where some examples might otherwise be annotated as culturally specific based on stereotypical associations between specific foods and corresponding regions or cultures.

6. In cases where the English machine translation appears to be extremely low quality, such that the topic of the example is not clear, we use our best judgment based on the previous guidelines. We lean towards annotating cultural specificity in borderline cases, because we expect that machine translation systems are more likely to perform poorly in culturally-specific scenarios (Citation?).

Through these annotations, we primarily aim to have a coarse filter for cultural specificity, such that we can up-sample culturally-specific examples in the following section. In the full non-parallel dataset (i.e. before subsampling to the official split), N%of examples are annotated as culturally specific. We note that even when marked as culturally specific, many examples do not actually require knowledge of the referenced culturally-specific item or tradition to correctly answer the prompt; in many cases, the culturally-specific element is referenced, but the correct answer can be inferred naively from the rest of the context.

D.3 Non-Parallel Split: Subsampling to the Official Split

Before subsampling, the Global PIQA non-parallel split is highly skewed across languages. For example, before subsampling, Hindi contains Nexamples and Yoruba contains Nexamples, while many other languages have close to the minimum dataset submission requirement of 100 examples. The full dataset before subsampling is available at `githuburl`. Due to the imbalance across languages, we select a subsample of 100 diverse and maximally culturally-specific examples in each language as the official non-parallel split of Global PIQA. This enables efficient evaluations of state-of-the-art LLMs across all languages in Global PIQA.

When filtering, we apply the following stages per language; we continue to the next stage unless that stage would cause the dataset for the language to fall below 100 examples. This allows us to maximize the quality and diversity of the examples for each language while still maintaining at least 100 examples per language. We note that the extremely low quality examples and off-task examples were already filtered out by the cleaning and annotations in §D.1 and §D.2. We apply the following filtering stages in order (or until reaching 100 examples in the language):

1. We remove any duplicate prompts, i.e. examples that have the same prompt but different pairs of solutions. This is generally a very small number of examples (e.g. one or two examples), and zero examples for most datasets. This filtering step drops a total of N examples across all languages. Note that exact duplicate examples (i.e. same prompt and same solutions) were already removed in §D.1.

2. We filter out examples where the two candidate solutions differ in length by more than 25 English byte equivalents. We compute English byte equivalents by computing the solution lengths first in raw UTF-8 bytes, then dividing by the language's byte premium (Arnett et al., 2024), which is the estimated number of bytes used to encode text in the language compared to content-matched (parallel) text in English. We perform this filtering step to attempt to minimize any length biases in the dataset, where longer solutions might be assigned systematically lower probabilities than shorter solutions by pretrained-only models, leading to a bias towards shorter solutions for those models. This filtering step drops a total of N examples across all languages.

3. We filter our examples whose non-stopword tokens overlap by more than 50% with another example in the dataset. Specifically, we tokenize all examples using the Goldfish tokenizer for the language (Chang et al., 2024). For the N Global PIQA languages not covered by the 350 languages in Goldfish, we use a simple space-based tokenizer after removing common punctuation symbols; all Global PIQA languages without a Goldfish tokenizer use scripts that separate words with spaces. Upon tokenizing all examples, we define stopword tokens as tokens that appear in at least 25% of examples for the language. Then, we sort examples by length (in order to give longer examples priority), and we loop through all examples, dropping any examples in which greater than 50% of its non-stopword tokens are contained in another previously-encountered example. This filtering step aims to increase the diversity of examples in the official Global PIQA

Language	Acc.	Language	Acc.
Slovenian (slv_latn)	97%	Croatian (hrv_latn)	100%
Serbian (srp_latn)	97%	Macedonian (mkd_cyrl)	92%
Catalan (cat_latn)	94%, 95%, 98%	Estonian (ekk_latn)	95%
Tamil (tam_taml)	95%	European Portuguese (por_latn_port)	91%, 95%
Algerian Arabic (arq_arab)	95%	Moroccan Arabic (ary_arab)	95%
Mandarin Chinese (cmn_hans)	95%	Mandarin Chinese (cmn_hant)	93%

Table 6: Ad hoc human evaluations, showing accuracies for individual human annotators for various languages, on individual dataset contributions to the Global PIQA non-parallel split. Details in §E.

non-parallel split, particularly for languages with large numbers of examples covering similar topics. This filtering step drops a total of N examples across all languages.

Finally, we sample 100 examples from the filtered subset for each language. We sample culturally-specific examples before non-culturally specific examples (as annotated in §D.2), and within each of these categories, we first sample examples that did not use any LLMs in the creation process. We shuffle the correct and incorrect solutions to balance 0 and 1 labels.

E Ad Hoc Human Evaluations

We do not explicitly perform a human evaluation study due to the substantial resources that it would take to run a study for the large number of languages involved in Global PIQA. However, several groups reported human evaluations on their dataset contributions to the non-parallel split, where a native speaker was asked to choose correct solutions without access to the "ground truth" labels, or inter-annotator agreement percentages were reported (from which we can compute an analogy to human "accuracy" by treating the other annotator's labels as the "ground truth"). On top of this, we conducted ad hoc human evaluation with one author (a native speaker of Mandarin Chinese) for the Mandarin Chinese datasets (simplified and traditional Chinese characters, `cmn_hans` and `cmn_hant`) in the Global PIQA official non-parallel split, after observing somewhat low scores in the language for some models (e.g. GPT-5 with less than 90% accuracy, given that Mandarin Chinese is a high-resource language). Accuracies for individual human annotators for the 12 language varieties with available human results are shown in Table 6.

In these ad hoc human evaluations, mean human annotator accuracy was 95.1%, and none of the fifteen individual annotators had accuracy below 91%. Of course, we note that there is likely some sampling bias, where dataset authors who chose to run human evaluations were also more likely to construct high quality datasets in the first place. That said, we even observe high accuracies for Mandarin Chinese (95% and 93%), in which we ran our ad hoc human evaluation after dataset submissions and compilation, independent of the dataset authors. These results suggest that human accuracy on the Global PIQA non-parallel split is likely to be at least 90%, and potentially as high as 95%. After running these ad hoc evaluations, examples were updated based on disagreeing labels.

F Evaluation Details

F.1 Full List of Models

We evaluate Global PIQA on 146 models, including 7 closed models and 139 open-weight models: GPT-SW3 1.3B (Ekgren et al., 2024); APT3 1B (Ociepa and Azurro Team, 2024); Salamandra 2B and 7B (Gonzalez-Agirre et al., 2025b); Aya Expanse (Dang et al., 2024); Command R 7B and 32B (Cohere et al., 2025); Ganda Gemma [8] and Swahili Gemma[9], HyGPT 10B (Gen2B, 2025); EXAONE 3.5 7.8B and 32B (An et al., 2024) and EXAONE 4 1.2B and 32B (Bae et al., 2025), Poro 2 8B (Zosa et al., 2025), Viking 7B[10] and 13B[11]; Qwen 2.5 (500M 1.5B, 3B, 7B 14B, and 32B, 72B; Yang

[8] `https://huggingface.co/CraneAILabs/ganda-gemma-1b`

[9] `https://huggingface.co/CraneAILabs/swahili-gemma-1b`

[10] `https://huggingface.co/LumiOpen/Viking-7B`

[11] `https://huggingface.co/LumiOpen/Viking-13B`

et al., 2024b), and Qwen 3 (600M, 1.7B, 4B, 8B, 14B, 32B; Team, 2025); SeaLLMs v3 1.5B and 7B (Zhang et al., 2025); Babel 9B (Zhao et al., 2025); Tucano 1.1B and 2.4B (Corrêa et al., 2024); Cheetah (Adebara et al., 2024); TowerBase and TowerInstruct v0.1 7B and 13B (Alves et al., 2024), Komodo 7B (Owen et al., 2024); Gemma SEA-LION v3 9B and Llama SEA-LION 8B (Ng et al., 2025); Gromenauer 7B[12]; BLOOM (560M, 1.1B, 1.7B, 3B, 7.1B; Workshop et al., 2022); Croissant LLM v0.1 1B (Faysse et al., 2025); DeepSeek R1 Distill Qwen (1.5B, 7B 14B; DeepSeek-AI, 2025); XGLM (1.7b, 2.9, 4.5B, 7.5B; Lin et al., 2021; Gemini 2.5 Pro (Google DeepMind, 2025c), Flash (Google DeepMind, 2025b), Flash-Lite (Google DeepMind, 2025a); Gemma 2 (2B, 9B, 27B; Team Gemma et al., 2024) and Gemma 3 (270M, 1B, 4B, 12B, 27B; Team Gemma et al., 2025); GPT-5 (full size, nano, and mini; OpenAI, 2025); Llama Krikri 8B (Roussis et al., 2025); Meltemi v1.5 7B (Voukoutis et al., 2024); Jais (1.3B, 2.7B, 6.7B, 30B; Sengupta et al., 2023; Inception, 2024); Kanana 1.5 (2.1B and 8B; Bak et al., 2025); Llama 3.1 (8B base and instruct, 70B base and instruct) and 3.2 (1B base and instruct, 3B base and instruct; Meta AI, 2024); Phi-3 (medium and mini instructAbdin et al., 2024a), Phi-3.5 mini instruct, and Phi-4 (full and mini instruct; Abdin et al., 2024b); Mistral v0.1 7B, Mistral v0.3 7B, Mistral Small, and Mixtral v0.1 (Jiang et al., 2023, 2024); HyperCLOVAX (500M and 1.5B; Yoo et al., 2024); GPT-oss 20B (Agarwal et al., 2025); Sailor2 (1B, 8B, 20B; Dou et al., 2025); Minerva[13] (1B, 3B, and 7B); Sarvam-m (Sarvam AI, 2025); Claude Sonnet 4.5 (Anthropic, 2025); Bielik v3 (1.5B and 4.5B; Ociepa et al., 2025); Apertus (8B and 70B; (Hernández-Cano et al., 2025)); Falcon (7B; (Almazrouei et al., 2023); PersianMind v1.0 (Rostami et al., 2024); EuroLLM (9B; Martins et al., 2025); vinaLlama (2.7B and 7B; Nguyen et al., 2023b); and PhoGPT (7.5B; Nguyen et al., 2023a).

F.2 Refusals from Proprietary Models

G Additional Results

fix north african/subsaharan african sections

G.1 By Region (full)

[12] https://huggingface.co/bertin-project/Gromenauer-7B

[13] https://huggingface.co/collections/sapienzanlp/minerva-llms-661e6011828fe67de4fe7961

Model	Western Europe	Eastern Europe	Middle East	North Africa	Subsaharan Africa	Central Asia	South Asia	Southeast Asia	East Asia	Americas	Avg.
Sub-1B Weight Class (LL)											
google/gemma-3-270m	53.3	53.3	49.7	53	63	55.5	51.4	55	55.8	59	54
bigscience/bloom-560m	52.6	53.2	50.5	51	61.1	54	52.1	53.8	53	64.7	53.7
Qwen/Qwen2.5-0.5B	52	51.7	50.5	51.6	62.4	53.2	50.3	53.2	53.8	61.7	53
1B Weight Class (LL)											
google/gemma-3-1b-pt	59.8	59.1	55.5	52.6	62.8	55.7	54.6	59.8	61.8	72.3	58.5
CraneAILabs/swahili-gemma-1b	56.7	55.3	52.1	51.6	61.1	53.7	53.4	58.7	57.4	67.7	55.7
CraneAILabs/ganda-gemma-1b	53.7	54.6	52.1	52.8	62	54.5	53	58	56.6	69.3	55.3
meta-llama/Llama-3.2-1B	56.7	56.6	49.2	53.2	64.4	51.5	52.2	53.7	52.4	65.3	55.2
facebook/xglm-1.7B	57.3	53.2	51.5	53.8	57	50.8	52.6	58	54.8	67.3	54.6
Qwen/Qwen2.5-1.5B	54.2	53.8	49	53.6	63.2	53.7	50	57.2	57.6	72.7	54.5
bigscience/bloom-1b7	54.2	52.6	52.8	53.6	60.2	53	52.4	57.2	55.2	70.3	54.5
bigscience/bloom-1b1	53.7	52.3	51.2	52.2	61	53.2	52.7	58.7	57	67	54.3
inceptionai/jais-family-1p3b	52.4	52.4	56.9	58.6	62.6	54.2	50.7	52.3	53	57.3	54.2
speakleash/Bielik-1.5B-v3	55.5	53.1	52.7	53.2	62.3	52.5	50.3	49.3	53.6	66.3	53.9
SeaLLMs/SeaLLMs-v3-1.5B	54.2	53.6	47.7	53.8	61.3	53.8	50.3	54.7	57.8	70.7	53.9
sail/Sailor2-1B	50.3	52.9	52.7	50.8	61.4	53	51.9	62.7	55.4	54	53.9
kakaocorp/kanana-1.5-2.1b-base	53.6	53.3	50.8	49.4	62.7	54.7	52.1	48.7	57.8	60	53.8
AI-Sweden-Models/gpt-sw3-1.3b	57.3	53	52.3	52	60.7	54.5	50.6	50.2	53.2	54.3	53.6
croissantllm/CroissantLLMChat-v0.1	54.9	54.1	49.7	53.8	58.3	54	50.6	51.5	50.8	63.3	53.5
sapienzanlp/Minerva-1B-base-v1.0	53.7	52.1	48.8	50.6	61.4	50.5	51.3	51.7	57.4	53	52.8
TucanoBR/Tucano-1b1	50.4	51.9	51	52.6	59.7	51.2	51.7	49.8	54.2	64.7	52.7
Azurro/APT3-1B-Base	47.6	51.7	51.4	48.8	56.9	52	49.2	51.5	52.2	49	51
2-3B Weight Class (LL)											
google/gemma-3-4b-pt	70.6	65.1	59.4	53.8	65.2	65.7	58	68.7	60.6	83.3	63.8
google/gemma-2-2b	61.8	59.2	53	53.6	65	55	52.9	61.5	61	75.7	58.5
facebook/xglm-4.5B	61.5	60.9	52.3	52.6	62.1	57	51.2	59.8	60.8	76	58.3
meta-llama/Llama-3.2-3B	61.3	58.4	53.4	50.2	65	59.7	54.1	56.7	55.6	74	57.8
Qwen/Qwen2.5-3B	60.8	55.8	51.5	50.8	65.1	55.2	50.2	59.8	60.6	78	56.8
BSC-LT/salamandra-2b	62.1	59.2	50.5	53.4	63	52.2	52.2	52.2	53.8	72.7	56.7
facebook/xglm-2.9B	57.5	55.5	50.7	53	59.1	53.2	52.5	58.3	58	73.3	55.7
bigscience/bloom-3b	55.3	53.2	51.9	53.4	60.3	51.8	52.9	58.5	58.4	71.7	55.1
inceptionai/jais-family-2p7b	56.4	53.2	56.9	55.6	63.9	54.2	51.5	49	54.4	64.7	55.1
speakleash/Bielik-4.5B-v3	57.6	54	53.6	52.6	62.7	53.7	51.2	52.8	51.6	69.3	54.9
vilm/vinallama-2.7b	52.3	54.6	53.3	53.6	60.2	53.2	51.4	55	53.8	58.3	54.1
sapienzanlp/Minerva-3B-base-v1.0	55	52.9	50.8	54.4	60.9	50	51	49.8	55.2	57	53.3
TucanoBR/Tucano-2b4	51.9	52.5	49.7	53.6	61.2	50	52.4	49.7	52.6	65.7	53.1
UBC-NLP/cheetah-base	49.9	49.3	49.2	49.4	51.7	44	52.2	48.7	50.6	49.3	49.9
2-3B Weight Class (Gen)											
Qwen/Qwen3-4B	77.3	74.8	70.2	63.6	56.9	66.2	71.7	77.8	78.6	93.8	71.7
google/gemma-3-4b-it	69.2	69.6	67.6	67.6	57	61.5	70.3	71.5	67.8	87.5	68.1
Qwen/Qwen2.5-3B-Instruct	65.7	63.6	62.2	58.8	55.7	51.5	56.2	70.3	66.2	87.8	61.8
google/gemma-2-2b-it	61.7	61.6	56.5	56.2	52.3	46.8	59	65.5	61.4	82	59.4
microsoft/Phi-4-mini-instruct	60.5	61.6	55.8	52	56.6	52.8	58	61.3	59	81.8	59.2
microsoft/Phi-3.5-mini-instruct	63.6	57.7	57	55.6	55.7	53.7	52.2	58.8	62.4	87	57.9
microsoft/Phi-3-mini-4k-instruct	63.1	55.3	52.4	57.4	55.1	53	53.5	60.2	57	87.5	57
speakleash/Bielik-4.5B-v3.0-Instruct	60.9	56.1	49.7	44.4	49.6	49.5	51.3	54.2	53.6	81.8	54.1
meta-llama/Llama-3.2-3B-Instruct	53	53.7	37.6	29.4	44.8	52.7	54.2	58.8	56.6	69.5	50.7
BSC-LT/salamandra-2b-instruct	14.9	11.8	11.5	9.4	11.4	13.5	13.2	10.2	10.8	10.3	12.1
TucanoBR/Tucano-2b4-Instruct	1.43	2.68	6.42	5	2.14	2.75	3.75	2.67	4.6	1.5	3.22
7-10B Weight Class (LL)											
swiss-ai/Apertus-8B-2509	74.5	70.8	60.5	58.6	63	70.3	57.8	71.7	59.6	83.7	66.2
aisingapore/Gemma-SEA-LION-v3-9B-IT	72.5	68.7	60.5	57.6	66.4	60.7	59.1	72	64	83.7	65.7
google/gemma-2-9b	73.6	68.7	60.4	56.8	65.4	62.5	57.8	69.8	61.8	85.7	65.4
utter-project/EuroLLM-9B	74.3	68	56.3	58.2	63.8	51.2	53.4	52.7	61.6	87.3	62.4
aisingapore/Llama-SEA-LION-v3-8B-IT	68.5	61.3	55.5	54	64	62	56.9	65.2	57.6	77.3	61.2
meta-llama/Llama-3.1-8B	67.9	62.5	56.1	55.8	64.2	60.8	55.9	63.7	56.6	78.3	61.2
BSC-LT/salamandra-7b	71.2	65.3	50.9	52.6	65.2	55.5	50.1	53.8	56.2	82	59.9
Tower-Babel/Babel-9B	61.1	58.1	55.3	55.2	63.1	56.2	54.1	63.8	61.6	84.3	59
LumiOpen/Llama-Poro-2-8B-base	65.8	58.8	52.6	52	66	54.7	52.8	60	56.4	75.3	58.5
Qwen/Qwen2.5-7B	62.8	57.4	52.9	55.8	62.4	54.5	51.7	64.8	63.4	83.7	58.4
sail/Sailor2-8B	58.7	55.6	53.1	55	66	52.2	52.1	73.8	63.2	75	58
SeaLLMs/SeaLLMs-v3-7B	61.5	56.7	52.8	54	62.4	55.2	52.7	61.8	66.8	77	57.9
mistralai/Mistral-7B-v0.1	64.3	59.9	51.1	51	63.2	54.2	51.2	57.8	57.6	76.7	57.8
ilsp/Llama-Krikri-8B-Base	61.1	58	51.3	50.4	64.1	53.2	53.2	58.5	54.4	74.7	57.2
Unbabel/TowerBase-7B-v0.1	61.9	57.4	51.5	52.2	62.6	54.7	50.8	55.8	57	77.3	56.8
bertin-project/Gromenauer-7B	59.9	57.4	51.2	53.2	63	55.5	51.6	56	55.8	73.3	56.5
facebook/xglm-7.5B	58.3	55.6	53.5	51.8	61	52.5	52.4	59.7	56.2	75.7	56.2
Yellow-AI-NLP/komodo-7b-base	57.9	55.7	51.8	52.4	62.9	55	52.1	59.5	55.8	69.7	56.1
LumiOpen/Viking-7B	66.3	55	50.6	56.4	60.9	54.7	50.3	52.3	56.2	68.3	56.1
inceptionai/jais-family-6p7b	57.9	53.7	60.5	56.6	61.4	54.5	49.7	54.7	56.6	68	55.9
bigscience/bloom-7b1	54.7	52.7	53	54.8	61.1	54.2	53.6	58.5	59.8	79.7	55.7
kakaocorp/kanana-1.5-8b-base	56.9	54.2	52.2	54.2	62.7	53.5	52.9	56	59.6	68	55.6
ilsp/Meltemi-7B-v1.5	57	55.4	52.1	53	61.2	53.2	49.2	53.5	53.4	66.7	54.6
tiiuae/falcon-7b	57.8	53.1	48.2	49.8	65.2	53.2	50.6	51.3	56.4	75.3	54.5
sapienzanlp/Minerva-7B-base-v1.0	56.3	53.6	50.8	48.8	62.7	54	51.7	51.7	52.6	63.7	54
vilm/vinallama-7b	52.1	53.6	52.6	53.8	60.1	52.8	49.7	56.7	54.8	61	53.7
universitytehran/PersianMind-v1.0	53.8	52.2	50	53.2	63.4	51.5	53.1	52.2	52.8	55.7	53.5
vinai/PhoGPT-7B5	48.7	49.2	47.2	50	50.8	49.7	52.1	51.2	54.4	49	50

Table 7: Aggregated results across all regions.

Model	Western Europe	Eastern Europe	Middle East	North Africa	Subsaharan Africa	Central Asia	South Asia	Southeast Asia	East Asia	Americas	Avg.
7-10B Weight Class (Gen)											
Qwen/Qwen3-8B	80.6	79.1	74.2	66.8	56.3	70.3	76	83	82.4	94.8	75.1
google/gemma-2-9b-it	78.1	76.1	70.5	64.8	43.7	65	71.1	79.5	75	93.2	70.4
swiss-ai/Apertus-8B-Instruct-2509	72.6	73.3	64.3	62	55.3	66	69.1	70.2	67.4	88.2	68.3
Qwen/Qwen2.5-7B-Instruct	72.4	69	69.8	59.8	57.5	59	64.2	76.8	74.4	90.5	67.6
LumiOpen/Llama-Poro-2-8B-Instruct	70.6	68	62.2	56.2	55.9	57	63	69.3	63.8	85	64.6
CohereLabs/aya-expanse-8b	64.8	67.1	69.7	61.6	56.3	52.5	60.8	65	71	79.7	64.1
ilsp/Llama-Krikri-8B-Instruct	67.9	66.6	60.7	58.8	55.9	52.2	60.6	67.3	65.6	84.5	63.2
sail/Sailor2-8B-Chat	65.1	62.4	62.8	59.6	56.5	51.5	61.8	79.2	62.4	86	63
meta-llama/Llama-3.1-8B-Instruct	66.6	64	62	55.6	50.6	55.7	61.5	67.5	68.4	81.8	62.2
LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct	64.4	62.1	61	54	54.8	53.7	62.2	61.5	66.4	87.3	61.7
mistralai/Mixtral-8x7B-Instruct-v0.1	74	66.1	57.7	53.6	50	44	55.7	67.7	64	87.8	61.3
utter-project/EuroLLM-9B-Instruct	66.2	67.3	62.8	60.2	52.3	48	52.8	57.7	65.4	85.3	60.7
kakaocorp/kanana-1.5-8b-instruct-2505	64.3	61.6	58.6	53.8	54.4	52	58.5	65.7	70.6	81.2	60.6
SeaLLMs/SeaLLMs-v3-7B-Chat	65.4	60.5	59.1	51.6	52.8	52.2	56	68.2	69.2	87	60
mistralai/Mistral-7B-Instruct-v0.3	61.5	65.2	59.2	54.6	58.6	52	54.3	60.5	59.8	81.2	60
CohereLabs/c4ai-command-r7b-12-2024	60	59.2	64.2	60.2	50.9	50.8	59.3	60.8	68.6	72.8	59.5
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	60.1	59.8	52.9	53	48.8	46	54.3	56.2	60.8	78.7	56.2
mistralai/Mistral-7B-Instruct-v0.1	55.4	57.4	51.9	51.6	51.1	48.8	50	54.2	52.4	76.5	53.9
vilm/vinallama-7b-chat	52.7	52.4	47.4	47.2	48.3	51.5	48.1	53.8	54.6	64	50.8
Unbabel/TowerInstruct-7B-v0.1	52.5	52.5	48.9	50.8	52.1	50.8	47	49.3	52.2	60	50.8
sapienzanlp/Minerva-7B-instruct-v1.0	48.8	49.3	46.6	47	47.7	48.2	46.8	47.8	44.8	46.5	47.7
ilsp/Meltemi-7B-Instruct-v1.5	49.3	48.5	49.6	44.6	49.1	41.3	43.8	48.7	43.4	62.7	47.6
BSC-LT/salamandra-7b-instruct	18.1	19.1	19.8	23.2	20.1	18.8	19.3	18.5	13.6	19.3	19.1
12-20B Weight Class (Gen)											
google/gemma-3-12b-it	83.6	82.6	79.8	78	65.5	78.5	80.9	82.8	77.8	92.5	79.5
openai/gpt-oss-20b	84.6	81	79.6	73.8	65.9	75.5	79.3	86.3	81.2	94.8	79.1
Qwen/Qwen3-14B	84	83.2	76.6	71.8	57.6	75.8	80	86.7	85.8	94.8	78.5
Qwen/Qwen2.5-14B-Instruct	80.9	77.1	76.7	72.6	60.4	62.7	72.8	81.7	84.2	95.5	74.9
microsoft/phi-4	81.9	78.8	72.7	66	58	64.7	76	78.7	77	94.8	74.5
deepseek-ai/DeepSeek-R1-Distill-Qwen-14B	79.1	73.6	74.4	66.8	57.1	52.2	66.9	81.8	80	93.8	71.2
sail/Sailor2-20B-Chat	75.3	71.6	69.9	65.6	59.2	54.2	68.6	79.2	80.2	93.5	70.2
microsoft/Phi-3-medium-4k-instruct	68.7	62.9	57.5	56.6	55.5	52.2	55.1	65.8	67.8	94.3	61.2
Unbabel/TowerInstruct-13B-v0.1	54.4	52	51.5	50.8	51.4	50.8	51.2	52.8	53.4	64.2	52.4
27-32B Weight Class (Gen)											
google/gemma-3-27b-it	86.1	86.5	82.9	80.2	67.2	80.7	82.6	87.3	82.2	95.8	82.4
Qwen/Qwen2.5-32B-Instruct	84.6	78.9	78.1	72.2	60.2	65	75.1	86.5	85	96	76.8
google/gemma-2-27b-it	85	81.6	77.5	78.4	41.8	70.3	76.3	84.7	79.4	94.5	75.4
sarvamai/sarvam-m	83.7	79.2	73.2	64.8	54.1	64.7	79.5	79.8	79.8	94.3	75.3
CohereLabs/aya-expanse-32b	80.6	77.8	79.5	72.6	58.2	61.5	72.8	80	80.6	94.8	74.7
CohereLabs/c4ai-command-r-08-2024	72.7	73.6	75.2	68.4	55.9	52.5	67.2	74.2	76	88	69.6
LGAI-EXAONE/EXAONE-3.5-32B-Instruct	68.9	67.5	59.8	54.8	57.2	51	61	67.3	72.4	86.7	63.9
mistralai/Mistral-Small-Instruct-2409	66.6	65.9	56.5	56.2	53.9	48.5	55.2	65.5	60.4	83.3	60.5
inceptionai/jais-family-30b-8k-chat	60.6	57	71.3	66	53.4	50.2	51.4	58.5	57.6	79.2	58.3
LGAI-EXAONE/EXAONE-4.0-32B	70.2	58	61.1	44.4	44.3	21.7	47	65.5	71.2	93	56
70-72B Weight Class (Gen)											
Qwen/Qwen2.5-72B-Instruct	88.7	84.6	82	76	61.5	76	77.7	88.7	88.2	97.8	80.6
meta-llama/Llama-3.1-70B-Instruct	83.7	82.1	79.2	74.8	66.2	75.8	79.8	83.7	79.6	93.5	79.2
swiss-ai/Apertus-70B-Instruct-2509	77.7	78.2	73.8	70.4	61.7	70.5	73.1	77.2	74	91.2	74

Table 8: Aggregated results across all regions. make model names prettier, add links?

G.2 European Languages

Model	cat_latn	eng_latn	fao_latn	fin_latn	fra_latn_fran	glg_latn	isl_latn	ita_latn	nld_latn	nno_latn	nob_latn	swe_latn	deu_latn	spa_latn_spai	Avg.
Sub-1B Weight Class (LL)															
google/gemma-3-270m	0.59	0.69	0.61	0.48	0.46	0.54	0.45	0.57	0.54	0.44	0.53	0.5	nan	nan	0.533
bigscience/bloom-560m	0.64	0.59	0.53	0.47	0.51	0.55	0.46	0.51	0.53	0.51	0.58	0.43	nan	nan	0.526
Qwen/Qwen2.5-0.5B	0.62	0.62	0.57	0.44	0.53	0.58	0.43	0.55	0.46	0.52	0.5	0.42	nan	nan	0.52
1B Weight Class (LL)															
google/gemma-3-1b-pt	0.63	0.72	0.56	0.58	0.57	0.58	0.5	0.74	0.6	0.48	0.52	0.69	nan	nan	0.597
AI-Sweden-Models/gpt-sw3-1.3b	0.53	0.62	0.64	0.5	0.48	0.52	0.57	0.53	0.52	0.58	0.58	0.81	nan	nan	0.573
facebook/xglm-1.7B	0.73	0.69	0.58	0.66	0.58	0.61	0.47	0.69	0.44	0.43	0.51	0.48	nan	nan	0.572
meta-llama/Llama-3.2-1B	0.68	0.73	0.61	0.54	0.5	0.66	0.47	0.59	0.56	0.41	0.51	0.54	nan	nan	0.567
2-3B Weight Class (LL)															
google/gemma-3-4b-pt	0.72	0.77	0.66	0.75	0.68	0.77	0.54	0.8	0.77	0.58	0.65	0.78	nan	nan	0.706
BSC-LT/salamandra-2b	0.74	0.72	0.58	0.66	0.56	0.72	0.38	0.7	0.57	0.54	0.52	0.76	nan	nan	0.621
google/gemma-2-2b	0.7	0.78	0.6	0.57	0.6	0.64	0.44	0.76	0.6	0.48	0.55	0.69	nan	nan	0.618
facebook/xglm-4.5B	0.7	0.7	0.54	0.66	0.61	0.62	0.5	0.71	0.59	0.48	0.56	0.71	nan	nan	0.615
2-3B Weight Class (Gen)															
Qwen/Qwen3-4B	0.78	0.82	0.64	0.74	0.74	0.87	0.65	0.84	0.75	0.79	0.77	0.83	0.74	0.86	0.773
google/gemma-3-4b-it	0.64	0.76	0.63	0.68	0.74	0.7	0.55	0.63	0.68	0.71	0.72	0.79	0.68	0.78	0.692
Qwen/Qwen2.5-3B-Instruct	0.6	0.77	0.56	0.59	0.72	0.67	0.58	0.65	0.72	0.6	0.58	0.69	0.7	0.77	0.657
microsoft/Phi-3.5-mini-instruct	0.58	0.8	0.5	0.69	0.73	0.72	0.46	0.7	0.69	0.58	0.45	0.63	0.58	0.79	0.636
7-10B Weight Class (LL)															
swiss-ai/Apertus-8B-2509	0.85	0.85	0.72	0.79	0.72	0.81	0.57	0.79	0.75	0.64	0.64	0.81	nan	nan	0.745
utter-project/EuroLLM-9B	0.81	0.77	0.69	0.75	0.7	0.83	0.5	0.85	0.79	0.63	0.73	0.87	nan	nan	0.743
google/gemma-2-9b	0.8	0.84	0.65	0.73	0.69	0.83	0.52	0.83	0.82	0.63	0.68	0.81	nan	nan	0.736
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.74	0.85	0.64	0.74	0.7	0.78	0.52	0.82	0.8	0.57	0.69	0.85	nan	nan	0.725
7-10B Weight Class (Gen)															
Qwen/Qwen3-8B	0.8	0.9	0.67	0.85	0.76	0.89	0.72	0.86	0.78	0.75	0.73	0.9	0.82	0.86	0.806
google/gemma-2-9b-it	0.71	0.87	0.64	0.78	0.79	0.79	0.7	0.8	0.79	0.75	0.74	0.9	0.82	0.86	0.781
mistralai/Mixtral-8x7B-Instruct-v0.1	0.7	0.84	0.52	0.72	0.79	0.76	0.62	0.73	0.78	0.73	0.73	0.8	0.81	0.83	0.74
swiss-ai/Apertus-8B-Instruct-2509	0.67	0.75	0.71	0.74	0.77	0.66	0.63	0.72	0.72	0.74	0.76	0.79	0.74	0.76	0.726
12-20B Weight Class (Gen)															
openai/gpt-oss-20b	0.79	0.91	0.74	0.88	0.82	0.88	0.81	0.85	0.85	0.89	0.8	0.89	0.88	0.85	0.846
Qwen/Qwen3-14B	0.87	0.92	0.7	0.87	0.86	0.88	0.71	0.87	0.82	0.88	0.79	0.88	0.84	0.87	0.84
google/gemma-3-12b-it	0.84	0.85	0.73	0.84	0.86	0.87	0.82	0.79	0.81	0.84	0.84	0.9	0.84	0.87	0.836
microsoft/phi-4	0.81	0.89	0.65	0.79	0.78	0.82	0.73	0.85	0.87	0.83	0.78	0.88	0.9	0.88	0.819
27-32B Weight Class (Gen)															
google/gemma-3-27b-it	0.92	0.89	0.79	0.81	0.82	0.85	0.86	0.82	0.77	0.91	0.86	0.92	0.94	0.9	0.861
google/gemma-2-27b-it	0.81	0.82	0.72	0.87	0.88	0.86	0.78	0.84	0.87	0.84	0.9	0.89	0.91	0.91	0.85
Qwen/Qwen2.5-32B-Instruct	0.87	0.91	0.73	0.81	0.89	0.88	0.79	0.9	0.82	0.81	0.74	0.88	0.91	0.9	0.846
sarvamai/sarvam-m	0.86	0.87	0.67	0.81	0.84	0.91	0.72	0.87	0.89	0.87	0.83	0.9	0.84	0.84	0.837
70-72B Weight Class (Gen)															
Qwen/Qwen2.5-72B-Instruct	0.92	0.92	0.75	0.87	0.86	0.92	0.8	0.92	0.88	0.87	0.87	0.97	0.93	0.94	0.887
meta-llama/Llama-3.1-70B-Instruct	0.79	0.9	0.75	0.8	0.82	0.77	0.8	0.86	0.8	0.87	0.84	0.93	0.91	0.88	0.837
swiss-ai/Apertus-70B-Instruct-2509	0.71	0.77	0.68	0.83	0.74	0.8	0.77	0.76	0.75	0.76	0.78	0.82	0.86	0.85	0.777
Closed Models (Gen)															
gemini-pro	1.0	0.92	0.96	0.95	0.96	0.99	0.93	0.94	0.91	0.93	0.94	0.99	0.98	0.98	0.956
gpt-5	0.98	0.96	0.9	0.96	0.96	0.94	0.92	0.96	0.88	0.93	0.92	0.99	0.98	0.98	0.947
sonnet-4-5	0.99	0.97	0.93	0.91	0.96	0.94	0.91	0.97	0.9	0.93	0.93	0.97	nan	nan	0.943
flash	0.98	0.94	0.95	0.94	0.92	0.97	0.91	0.92	0.91	0.91	0.94	0.97	0.96	0.96	0.941

Table 9: Western European

Model	als_latn	bel_cyrl	bul_cyrl	ces_latn	ckm_latn	ell_grek	hrv_latn	hye_armn	lit_latn	mkd_cyrl	Avg.
Sub-1B Weight Class (LL)											
bigscience/bloom-560m	0.54	0.6	0.49	0.58	0.53	0.52	0.49	0.45	0.61	0.51	0.532
Qwen/Qwen2.5-0.5B	0.52	0.58	0.51	0.48	0.43	0.54	0.52	0.54	0.57	0.53	0.522
google/gemma-3-270m	0.53	0.57	0.55	0.48	0.44	0.52	0.61	0.46	0.48	0.5	0.514
1B Weight Class (LL)											
google/gemma-3-1b-pt	0.52	0.67	0.53	0.61	0.51	0.56	0.64	0.53	0.56	0.61	0.574
meta-llama/Llama-3.2-1B	0.56	0.61	0.5	0.61	0.45	0.59	0.62	0.52	0.55	0.55	0.556
croissantllm/CroissantLLMChat-v0.1	0.55	0.57	0.56	0.5	0.58	0.56	0.58	0.5	0.58	0.48	0.546
CraneAILabs/swahili-gemma-1b	0.54	0.67	0.53	0.53	0.43	0.55	0.49	0.49	0.59	0.57	0.539
2-3B Weight Class (LL)											
google/gemma-3-4b-pt	0.61	0.73	0.71	0.63	0.46	0.62	0.69	0.61	0.6	0.68	0.634
facebook/xglm-4.5B	0.59	0.64	0.62	0.64	0.49	0.55	0.69	0.58	0.61	0.6	0.601
meta-llama/Llama-3.2-3B	0.58	0.64	0.66	0.63	0.5	0.59	0.56	0.49	0.55	0.56	0.576
google/gemma-2-2b	0.5	0.65	0.61	0.61	0.51	0.58	0.58	0.57	0.55	0.58	0.574
2-3B Weight Class (Gen)											
Qwen/Qwen3-4B	0.64	0.9	0.94	0.76	0.56	0.54	0.83	0.61	0.75	0.81	0.734
google/gemma-3-4b-it	0.68	0.8	0.81	0.62	0.51	0.6	0.79	0.6	0.75	0.87	0.703
microsoft/Phi-4-mini-instruct	0.54	0.61	0.75	0.62	0.56	0.57	0.63	0.52	0.62	0.66	0.608
google/gemma-2-2b-it	0.5	0.76	0.75	0.68	0.39	0.4	0.64	0.47	0.65	0.71	0.595
7-10B Weight Class (LL)											
swiss-ai/Apertus-8B-2509	0.63	0.8	0.75	0.8	0.47	0.59	0.79	0.72	0.78	0.79	0.712
google/gemma-2-9b	0.53	0.72	0.72	0.78	0.49	0.65	0.77	0.69	0.76	0.7	0.681
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.57	0.77	0.75	0.77	0.51	0.64	0.76	0.62	0.73	0.68	0.68
utter-project/EuroLLM-9B	0.52	0.71	0.76	0.77	0.48	0.65	0.74	0.54	0.78	0.63	0.658
7-10B Weight Class (Gen)											
Qwen/Qwen3-8B	0.72	0.93	0.92	0.81	0.54	0.65	0.9	0.78	0.86	0.89	0.8
google/gemma-2-9b-it	0.68	0.89	0.94	0.76	0.46	0.65	0.91	0.63	0.85	0.88	0.765
swiss-ai/Apertus-8B-Instruct-2509	0.76	0.81	0.82	0.77	0.52	0.61	0.83	0.64	0.86	0.8	0.742
Qwen/Qwen2.5-7B-Instruct	0.57	0.74	0.83	0.77	0.52	0.56	0.77	0.47	0.65	0.87	0.675
12-20B Weight Class (Gen)											
google/gemma-3-12b-it	0.82	0.91	0.92	0.79	0.49	0.68	0.95	0.78	0.96	0.93	0.823
Qwen/Qwen3-14B	0.74	0.94	0.95	0.81	0.51	0.72	0.96	0.81	0.9	0.87	0.821
microsoft/phi-4	0.73	0.92	0.92	0.84	0.58	0.64	0.91	0.67	0.83	0.94	0.798
openai/gpt-oss-20b	0.78	0.92	0.95	0.75	0.53	0.67	0.9	0.74	0.81	0.89	0.794
27-32B Weight Class (Gen)											
google/gemma-3-27b-it	0.88	0.98	0.95	0.86	0.53	0.73	0.97	0.81	0.98	0.95	0.864
google/gemma-2-27b-it	0.76	0.95	0.92	0.82	0.47	0.75	0.94	0.73	0.91	0.92	0.817
sarvamai/sarvam-m	0.67	0.92	0.91	0.85	0.46	0.71	0.86	0.78	0.77	0.87	0.78
CohereLabs/aya-expanse-32b	0.68	0.9	0.86	0.82	0.51	0.73	0.88	0.71	0.77	0.9	0.776
70-72B Weight Class (Gen)											
Qwen/Qwen2.5-72B-Instruct	0.7	0.96	0.95	0.9	0.57	0.8	0.95	0.72	0.88	0.94	0.837
meta-llama/Llama-3.1-70B-Instruct	0.79	0.95	0.95	0.8	0.54	0.72	0.87	0.82	0.93	0.93	0.83
swiss-ai/Apertus-70B-Instruct-2509	0.82	0.85	0.94	0.79	0.5	0.63	0.9	0.67	0.89	0.89	0.788
Closed Models (Gen)											
gemini-pro	0.97	0.98	0.95	0.97	0.74	0.89	1.0	0.96	0.98	1.0	0.944
flash	0.95	0.98	0.96	0.96	0.69	0.84	1.0	0.95	0.98	1.0	0.931
sonnet-4-5	0.95	0.99	0.97	0.94	0.65	0.84	1.0	0.94	0.99	0.99	0.926
gpt-5-mini	0.94	0.97	0.97	0.93	0.67	0.85	0.98	0.92	0.97	0.98	0.918

Table 10: Eastern European Indo-European pt 1

Model	pol_latn	por_latn_port	ron_latn	slk_latn_sari	slv_latn_cerk	srp_cyrl	srp_latn	ukr_cyrl	bos_latn	rus_cyrl	slk_latn	slv_latn	Avg.
Sub-1B Weight Class (LL)													
google/gemma-3-270m	0.49	0.54	0.59	0.54	0.57	0.58	0.69	0.68	nan	nan	nan	nan	0.585
bigscience/bloom-560m	0.53	0.54	0.68	0.56	0.5	0.48	0.53	0.58	nan	nan	nan	nan	0.55
Qwen/Qwen2.5-0.5B	0.54	0.51	0.6	0.52	0.46	0.5	0.62	0.57	nan	nan	nan	nan	0.54
1B Weight Class (LL)													
google/gemma-3-1b-pt	0.62	0.63	0.64	0.58	0.54	0.71	0.74	0.78	nan	nan	nan	nan	0.655
meta-llama/Llama-3.2-1B	0.51	0.61	0.58	0.59	0.54	0.74	0.69	0.69	nan	nan	nan	nan	0.619
CraneAILabs/swahili-gemma-1b	0.5	0.57	0.73	0.52	0.53	0.6	0.63	0.69	nan	nan	nan	nan	0.596
CraneAILabs/ganda-gemma-1b	0.59	0.6	0.66	0.51	0.5	0.61	0.67	0.6	nan	nan	nan	nan	0.593
2-3B Weight Class (LL)													
google/gemma-3-4b-pt	0.67	0.69	0.69	0.62	0.53	0.85	0.86	0.84	nan	nan	nan	nan	0.719
BSC-LT/salamandra-2b	0.64	0.63	0.67	0.56	0.59	0.78	0.81	0.79	nan	nan	nan	nan	0.684
facebook/xglm-4.5B	0.62	0.57	0.58	0.6	0.57	0.77	0.8	0.77	nan	nan	nan	nan	0.66
google/gemma-2-2b	0.64	0.61	0.65	0.59	0.55	0.7	0.72	0.81	nan	nan	nan	nan	0.659
2-3B Weight Class (Gen)													
Qwen/Qwen3-4B	0.74	0.81	0.95	0.62	0.5	0.8	0.83	0.81	0.96	0.85	nan	nan	0.787
google/gemma-3-4b-it	0.68	0.71	0.9	0.53	0.51	0.62	0.72	0.77	0.95	0.73	nan	nan	0.712
Qwen/Qwen2.5-3B-Instruct	0.69	0.72	0.87	0.54	0.46	0.71	0.64	0.71	0.89	0.76	nan	nan	0.699
google/gemma-2-2b-it	0.63	0.67	0.85	0.52	0.5	0.59	0.57	0.61	0.86	0.72	nan	nan	0.652
7-10B Weight Class (LL)													
swiss-ai/Apertus-8B-2509	0.75	0.71	0.72	0.64	0.5	0.88	0.88	0.9	nan	nan	nan	nan	0.748
google/gemma-2-9b	0.75	0.72	0.74	0.64	0.55	0.82	0.88	0.86	nan	nan	nan	nan	0.745
utter-project/EuroLLM-9B	0.78	0.74	0.71	0.7	0.56	0.74	0.8	0.91	nan	nan	nan	nan	0.742
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.77	0.73	0.74	0.6	0.55	0.83	0.84	0.87	nan	nan	nan	nan	0.741
7-10B Weight Class (Gen)													
Qwen/Qwen3-8B	0.82	0.87	0.98	0.66	0.6	0.82	0.82	0.84	0.96	0.82	nan	nan	0.819
google/gemma-2-9b-it	0.79	0.88	0.98	0.63	0.36	0.74	0.78	0.87	0.98	0.83	nan	nan	0.784
swiss-ai/Apertus-8B-Instruct-2509	0.74	0.8	0.97	0.52	0.45	0.75	0.78	0.77	0.98	0.82	nan	nan	0.758
mistralai/Mixtral-8x7B-Instruct-v0.1	0.78	0.86	0.94	0.54	0.45	0.78	0.69	0.68	0.94	0.77	nan	nan	0.743
12-20B Weight Class (Gen)													
Qwen/Qwen3-14B	0.88	0.91	1.0	0.71	0.65	0.91	0.9	0.92	0.98	0.86	nan	nan	0.872
google/gemma-3-12b-it	0.88	0.85	0.95	0.7	0.68	0.89	0.89	0.83	0.99	0.83	nan	nan	0.849
openai/gpt-oss-20b	0.82	0.9	0.97	0.66	0.52	0.86	0.9	0.9	0.98	0.88	nan	nan	0.839
microsoft/phi-4	0.77	0.91	0.99	0.66	0.56	0.87	0.82	0.84	0.99	0.88	nan	nan	0.829
27-32B Weight Class (Gen)													
google/gemma-3-27b-it	0.89	0.92	0.98	0.78	0.68	0.89	0.95	0.88	0.98	0.89	nan	nan	0.884
google/gemma-2-27b-it	0.83	0.91	0.99	0.74	0.57	0.84	0.82	0.87	1.0	0.83	nan	nan	0.84
Qwen/Qwen2.5-32B-Instruct	0.86	0.94	0.98	0.69	0.51	0.83	0.84	0.89	0.99	0.85	nan	nan	0.838
CohereLabs/aya-expanse-32b	0.87	0.92	0.99	0.75	0.5	0.77	0.84	0.81	0.92	0.83	nan	nan	0.82
70-72B Weight Class (Gen)													
Qwen/Qwen2.5-72B-Instruct	0.94	0.93	0.99	0.76	0.61	0.92	0.87	0.94	1.0	0.88	nan	nan	0.884
meta-llama/Llama-3.1-70B-Instruct	0.86	0.9	1.0	0.68	0.57	0.8	0.86	0.85	0.99	0.89	nan	nan	0.84
swiss-ai/Apertus-70B-Instruct-2509	0.78	0.89	0.94	0.7	0.56	0.82	0.79	0.82	0.93	0.84	nan	nan	0.807
Closed Models (Gen)													
gemini-pro	1.0	0.93	0.99	0.91	0.87	0.96	0.96	0.95	1.0	0.95	1.0	0.98	0.958
gpt-5	0.98	0.95	1.0	0.93	0.77	0.97	0.98	0.97	1.0	0.96	1.0	0.98	0.958
sonnet-4-5	0.98	0.93	1.0	0.92	0.78	0.97	0.96	0.96	nan	nan	0.97	0.99	0.946
flash	0.97	0.94	0.99	0.94	0.81	0.96	0.95	0.9	0.99	0.93	0.98	0.98	0.945

Table 11: Eastern European Indo-European pt 2

Model	azj_latn	est_latn	hun_latn	kat_geor	tur_latn	Avg.
Sub-1B Weight Class (LL)						
bigscience/bloom-560m	0.5	0.57	0.53	0.49	0.43	0.504
google/gemma-3-270m	0.46	0.53	0.48	0.47	0.49	0.486
Qwen/Qwen2.5-0.5B	0.47	0.47	0.48	0.46	0.46	0.468
1B Weight Class (LL)						
Azurro/APT3-1B-Base	0.56	0.59	0.55	0.51	0.49	0.54
google/gemma-3-1b-pt	0.48	0.52	0.5	0.57	0.55	0.524
facebook/xglm-1.7B	0.48	0.58	0.47	0.45	0.59	0.514
CraneAILabs/swahili-gemma-1b	0.45	0.55	0.44	0.51	0.6	0.51
2-3B Weight Class (LL)						
google/gemma-3-4b-pt	0.53	0.53	0.63	0.5	0.69	0.576
facebook/xglm-2.9B	0.56	0.57	0.51	0.5	0.63	0.554
facebook/xglm-4.5B	0.49	0.59	0.5	0.46	0.68	0.544
meta-llama/Llama-3.2-3B	0.49	0.5	0.52	0.56	0.58	0.53
2-3B Weight Class (Gen)						
Qwen/Qwen3-4B	0.72	0.55	0.91	0.51	0.79	0.696
google/gemma-3-4b-it	0.71	0.51	0.76	0.52	0.76	0.652
Qwen/Qwen2.5-3B-Instruct	0.54	0.57	0.66	0.53	0.73	0.606
microsoft/Phi-4-mini-instruct	0.53	0.49	0.71	0.49	0.71	0.586
7-10B Weight Class (LL)						
swiss-ai/Apertus-8B-2509	0.63	0.68	0.6	0.54	0.74	0.638
utter-project/EuroLLM-9B	0.54	0.72	0.7	0.49	0.66	0.622
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.56	0.56	0.65	0.52	0.79	0.616
google/gemma-2-9b	0.56	0.58	0.67	0.5	0.73	0.608
7-10B Weight Class (Gen)						
Qwen/Qwen3-8B	0.81	0.46	0.93	0.55	0.83	0.716
google/gemma-2-9b-it	0.72	0.61	0.89	0.51	0.81	0.708
swiss-ai/Apertus-8B-Instruct-2509	0.7	0.56	0.79	0.48	0.8	0.666
ilsp/Llama-Krikri-8B-Instruct	0.5	0.58	0.8	0.54	0.75	0.634
12-20B Weight Class (Gen)						
google/gemma-3-12b-it	0.83	0.67	0.9	0.7	0.82	0.784
openai/gpt-oss-20b	0.8	0.72	0.94	0.62	0.83	0.782
Qwen/Qwen3-14B	0.84	0.63	0.95	0.62	0.82	0.772
microsoft/phi-4	0.72	0.47	0.92	0.5	0.81	0.684
27-32B Weight Class (Gen)						
google/gemma-3-27b-it	0.85	0.77	0.93	0.72	0.88	0.83
google/gemma-2-27b-it	0.84	0.62	0.91	0.54	0.91	0.764
sarvamai/sarvam-m	0.79	0.68	0.95	0.55	0.82	0.758
Qwen/Qwen2.5-32B-Instruct	0.8	0.52	0.92	0.52	0.87	0.726
70-72B Weight Class (Gen)						
Qwen/Qwen2.5-72B-Instruct	0.86	0.64	0.95	0.6	0.9	0.79
meta-llama/Llama-3.1-70B-Instruct	0.83	0.64	0.87	0.61	0.88	0.766
swiss-ai/Apertus-70B-Instruct-2509	0.82	0.71	0.73	0.54	0.81	0.722
Closed Models (Gen)						
gemini-pro	0.95	0.96	0.96	0.94	0.95	0.952
gpt-5	0.95	0.97	0.95	0.89	0.96	0.944
flash	0.96	0.9	0.98	0.89	0.93	0.932
sonnet-4-5	0.95	0.91	0.96	0.89	0.93	0.928

Table 12: Eastern European non-Indo-European

G.3 North African languages

Model	aeb_arab	amh_ethi	arq_arab	ary_arab	arz_arab	Avg.
Sub-1B Weight Class (LL)						
google/gemma-3-270m	0.53	0.48	0.52	0.56	0.56	0.53
Qwen/Qwen2.5-0.5B	0.53	0.48	0.51	0.55	0.51	0.516
bigscience/bloom-560m	0.52	0.43	0.53	0.52	0.55	0.51
1B Weight Class (LL)						
inceptionai/jais-family-1p3b	0.62	0.49	0.6	0.53	0.69	0.586
croissantllm/CroissantLLMChat-v0.1	0.57	0.57	0.51	0.49	0.55	0.538
SeaLLMs/SeaLLMs-v3-1.5B	0.58	0.51	0.55	0.5	0.55	0.538
facebook/xglm-1.7B	0.58	0.46	0.59	0.54	0.52	0.538
2-3B Weight Class (LL)						
inceptionai/jais-family-2p7b	0.56	0.45	0.61	0.53	0.63	0.556
sapienzanlp/Minerva-3B-base-v1.0	0.5	0.46	0.57	0.64	0.55	0.544
google/gemma-3-4b-pt	0.59	0.43	0.58	0.52	0.57	0.538
TucanoBR/Tucano-2b4	0.53	0.58	0.54	0.54	0.49	0.536
2-3B Weight Class (Gen)						
google/gemma-3-4b-it	0.73	0.74	0.7	0.62	0.59	0.676
Qwen/Qwen3-4B	0.71	0.6	0.61	0.67	0.59	0.636
Qwen/Qwen2.5-3B-Instruct	0.65	0.52	0.61	0.59	0.57	0.588
microsoft/Phi-3-mini-4k-instruct	0.63	0.64	0.51	0.53	0.56	0.574
7-10B Weight Class (LL)						
swiss-ai/Apertus-8B-2509	0.64	0.51	0.59	0.53	0.66	0.586
utter-project/EuroLLM-9B	0.62	0.51	0.58	0.56	0.64	0.582
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.54	0.55	0.59	0.58	0.62	0.576
google/gemma-2-9b	0.56	0.53	0.59	0.56	0.6	0.568
7-10B Weight Class (Gen)						
Qwen/Qwen3-8B	0.69	0.68	0.68	0.63	0.66	0.668
google/gemma-2-9b-it	0.69	0.66	0.59	0.66	0.64	0.648
swiss-ai/Apertus-8B-Instruct-2509	0.63	0.67	0.57	0.58	0.65	0.62
CohereLabs/aya-expanse-8b	0.68	0.52	0.58	0.63	0.67	0.616
12-20B Weight Class (Gen)						
google/gemma-3-12b-it	0.83	0.78	0.76	0.74	0.79	0.78
openai/gpt-oss-20b	0.79	0.73	0.69	0.79	0.69	0.738
Qwen/Qwen2.5-14B-Instruct	0.79	0.69	0.71	0.71	0.73	0.726
Qwen/Qwen3-14B	0.79	0.65	0.72	0.68	0.75	0.718
27-32B Weight Class (Gen)						
google/gemma-3-27b-it	0.8	0.82	0.81	0.79	0.79	0.802
google/gemma-2-27b-it	0.79	0.79	0.74	0.8	0.8	0.784
CohereLabs/aya-expanse-32b	0.77	0.55	0.77	0.78	0.76	0.726
Qwen/Qwen2.5-32B-Instruct	0.8	0.65	0.7	0.73	0.73	0.722
70-72B Weight Class (Gen)						
Qwen/Qwen2.5-72B-Instruct	0.82	0.69	0.74	0.73	0.82	0.76
meta-llama/Llama-3.1-70B-Instruct	0.8	0.78	0.72	0.73	0.71	0.748
swiss-ai/Apertus-70B-Instruct-2509	0.74	0.67	0.68	0.69	0.74	0.704
Closed Models (Gen)						
gemini-pro	0.96	0.89	0.95	0.94	0.95	0.938
flash	0.9	0.86	0.92	0.92	0.92	0.904
gpt-5	0.94	0.77	0.93	0.89	0.95	0.896
sonnet-4-5	0.92	0.88	0.87	0.87	0.88	0.884

Table 13: N Africa

G.4 Middle East Languages

Model	acq_arab	afb_arab	apc_arab_jord	apc_arab_leba	apc_arab_pale	apc_arab_syri	arb_arab	ars_arab	ckb_arab	heb_hebr	pes_arab	acm_arab	Avg.
Sub-1B Weight Class (LL)													
Qwen/Qwen2.5-0.5B	0.44	0.59	0.44	0.56	0.54	0.47	0.49	0.53	0.59	0.42	0.49	nan	0.505
bigscience/bloom-560m	0.51	0.53	0.42	0.55	0.47	0.47	0.61	0.58	0.45	0.5	0.46	nan	0.505
google/gemma-3-270m	0.46	0.57	0.51	0.51	0.45	0.46	0.52	0.57	0.5	0.47	0.45	nan	0.497
1B Weight Class (LL)													
inceptionai/jais-family-1p3b	0.54	0.57	0.66	0.52	0.54	0.59	0.65	0.74	0.51	0.46	0.48	nan	0.569
google/gemma-3-1b-pt	0.53	0.6	0.56	0.56	0.49	0.5	0.59	0.71	0.48	0.54	0.55	nan	0.555
bigscience/bloom-1b7	0.56	0.52	0.51	0.59	0.51	0.45	0.54	0.59	0.47	0.55	0.52	nan	0.528
speakleash/Bielik-1.5B-v3	0.56	0.53	0.44	0.6	0.49	0.5	0.62	0.61	0.45	0.49	0.51	nan	0.527
2-3B Weight Class (LL)													
google/gemma-3-4b-pt	0.55	0.59	0.61	0.5	0.54	0.5	0.69	0.77	0.46	0.54	0.78	nan	0.594
inceptionai/jais-family-2p7b	0.61	0.63	0.63	0.52	0.55	0.56	0.66	0.73	0.48	0.45	0.44	nan	0.569
speakleash/Bielik-4.5B-v3	0.55	0.6	0.35	0.54	0.58	0.53	0.66	0.56	0.53	0.54	0.46	nan	0.536
meta-llama/Llama-3.2-3B	0.48	0.48	0.5	0.56	0.49	0.52	0.57	0.65	0.47	0.54	0.61	nan	0.534
2-3B Weight Class (Gen)													
Qwen/Qwen3-4B	0.71	0.63	0.74	0.73	0.67	0.65	0.75	0.74	0.49	0.68	0.73	0.9	0.702
google/gemma-3-4b-it	0.64	0.64	0.63	0.72	0.72	0.65	0.66	0.64	0.59	0.66	0.7	0.86	0.676
Qwen/Qwen2.5-3B-Instruct	0.63	0.62	0.6	0.68	0.49	0.59	0.66	0.68	0.56	0.55	0.54	0.86	0.622
microsoft/Phi-3.5-mini-instruct	0.56	0.52	0.5	0.63	0.57	0.62	0.51	0.6	0.5	0.51	0.53	0.79	0.57
7-10B Weight Class (LL)													
swiss-ai/Apertus-8B-2509	0.6	0.6	0.6	0.58	0.54	0.61	0.58	0.67	0.49	0.61	0.78	nan	0.605
inceptionai/jais-family-6p7b	0.64	0.62	0.72	0.55	0.61	0.63	0.7	0.8	0.44	0.46	0.48	nan	0.605
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.6	0.53	0.61	0.6	0.53	0.51	0.64	0.7	0.47	0.61	0.85	nan	0.605
google/gemma-2-9b	0.58	0.54	0.65	0.61	0.48	0.53	0.61	0.69	0.46	0.69	0.8	nan	0.604
7-10B Weight Class (Gen)													
Qwen/Qwen3-8B	0.69	0.73	0.81	0.84	0.74	0.7	0.72	0.82	0.46	0.64	0.8	0.95	0.742
google/gemma-2-9b-it	0.66	0.66	0.67	0.78	0.67	0.66	0.74	0.77	0.46	0.62	0.86	0.91	0.705
Qwen/Qwen2.5-7B-Instruct	0.72	0.66	0.71	0.74	0.8	0.61	0.68	0.74	0.52	0.61	0.69	0.9	0.698
CohereLabs/aya-expanse-8b	0.64	0.67	0.73	0.73	0.68	0.73	0.7	0.71	0.5	0.62	0.76	0.89	0.697
12-20B Weight Class (Gen)													
google/gemma-3-12b-it	0.81	0.79	0.82	0.85	0.79	0.79	0.77	0.83	0.62	0.76	0.83	0.92	0.798
openai/gpt-oss-20b	0.83	0.73	0.87	0.85	0.77	0.7	0.77	0.86	0.56	0.78	0.89	0.94	0.796
Qwen/Qwen2.5-14B-Instruct	0.72	0.73	0.8	0.88	0.72	0.8	0.8	0.78	0.54	0.69	0.79	0.95	0.767
Qwen/Qwen3-14B	0.79	0.74	0.85	0.81	0.75	0.75	0.73	0.84	0.49	0.72	0.82	0.9	0.766
27-32B Weight Class (Gen)													
google/gemma-3-27b-it	0.79	0.79	0.86	0.9	0.82	0.84	0.83	0.87	0.65	0.79	0.89	0.92	0.829
CohereLabs/aya-expanse-32b	0.8	0.82	0.82	0.84	0.78	0.76	0.77	0.81	0.6	0.76	0.89	0.89	0.795
Qwen/Qwen2.5-32B-Instruct	0.83	0.75	0.85	0.83	0.79	0.85	0.79	0.82	0.42	0.73	0.8	0.91	0.781
google/gemma-2-27b-it	0.78	0.73	0.81	0.78	0.82	0.74	0.79	0.79	0.56	0.71	0.88	0.91	0.775
70-72B Weight Class (Gen)													
Qwen/Qwen2.5-72B-Instruct	0.81	0.75	0.91	0.9	0.83	0.85	0.83	0.87	0.5	0.8	0.88	0.91	0.82
meta-llama/Llama-3.1-70B-Instruct	0.8	0.71	0.76	0.85	0.8	0.73	0.82	0.82	0.61	0.85	0.85	0.9	0.792
swiss-ai/Apertus-70B-Instruct-2509	0.77	0.69	0.75	0.82	0.67	0.73	0.78	0.67	0.56	0.73	0.84	0.85	0.738
Closed Models (Gen)													
gemini-pro	0.94	0.89	0.95	0.95	0.91	0.91	0.92	0.9	0.89	0.95	0.93	0.95	0.924
flash	0.89	0.85	0.96	0.94	0.86	0.9	0.9	0.87	0.81	0.95	0.94	0.95	0.902
gpt-5	0.95	0.88	0.97	0.94	0.93	0.9	0.95	0.89	0.49	0.89	0.95	0.97	0.893
sonnet-4-5	0.91	0.9	0.96	0.93	0.89	0.9	0.89	0.89	0.67	0.9	0.92	nan	0.887

Table 14: Mid East

G.5 Subsaharan African Languages

Model	bam_latn	hau_latn	ibo_latn	idu_latn	iso_latn	pcm_latn	urh_latn	yor_latn	zul_latn	ekp_latn	kin_latn	lin_latn	luo_latn	swh_latn	Avg.
Sub-1B Weight Class (LL)															
google/gemma-3-270m	0.53	0.58	0.68	0.66	0.68	0.7	0.54	0.67	0.63	nan	nan	nan	nan	nan	0.63
Qwen/Qwen2.5-0.5B	0.53	0.67	0.7	0.62	0.7	0.67	0.53	0.58	0.62	nan	nan	nan	nan	nan	0.624
bigscience/bloom-560m	0.58	0.6	0.62	0.62	0.62	0.64	0.59	0.64	0.59	nan	nan	nan	nan	nan	0.611
1B Weight Class (LL)															
meta-llama/Llama-3.2-1B	0.52	0.66	0.68	0.61	0.74	0.72	0.54	0.66	0.67	nan	nan	nan	nan	nan	0.644
Qwen/Qwen2.5-1.5B	0.49	0.64	0.71	0.67	0.69	0.72	0.54	0.64	0.59	nan	nan	nan	nan	nan	0.632
google/gemma-3-1b-pt	0.48	0.65	0.65	0.63	0.67	0.72	0.54	0.64	0.67	nan	nan	nan	nan	nan	0.628
kakaocorp/kanana-1.5-2.1b-base	0.49	0.63	0.66	0.69	0.68	0.7	0.53	0.65	0.61	nan	nan	nan	nan	nan	0.627
2-3B Weight Class (LL)															
google/gemma-3-4b-pt	0.49	0.69	0.71	0.69	0.68	0.79	0.54	0.55	0.73	nan	nan	nan	nan	nan	0.652
Qwen/Qwen2.5-3B	0.53	0.65	0.67	0.65	0.77	0.77	0.58	0.62	0.62	nan	nan	nan	nan	nan	0.651
google/gemma-2-2b	0.55	0.67	0.66	0.67	0.68	0.8	0.55	0.6	0.67	nan	nan	nan	nan	nan	0.65
meta-llama/Llama-3.2-3B	0.5	0.65	0.68	0.65	0.68	0.76	0.59	0.67	0.67	nan	nan	nan	nan	nan	0.65
2-3B Weight Class (Gen)															
google/gemma-3-4b-it	0.48	0.54	0.57	0.64	0.51	0.84	0.48	0.41	0.69	0.56	0.55	0.49	0.46	0.76	0.57
Qwen/Qwen3-4B	0.5	0.63	0.55	0.59	0.62	0.9	0.51	0.52	0.49	0.54	0.46	0.51	0.55	0.59	0.569
microsoft/Phi-4-mini-instruct	0.6	0.51	0.51	0.55	0.56	0.85	0.55	0.48	0.56	0.55	0.53	0.55	0.5	0.63	0.566
Qwen/Qwen2.5-3B-Instruct	0.54	0.49	0.58	0.48	0.56	0.85	0.51	0.6	0.53	0.55	0.53	0.5	0.51	0.57	0.557
7-10B Weight Class (LL)															
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.51	0.76	0.68	0.71	0.66	0.85	0.49	0.63	0.69	nan	nan	nan	nan	nan	0.664
sail/Sailor2-8B	0.56	0.79	0.66	0.61	0.71	0.77	0.56	0.58	0.7	nan	nan	nan	nan	nan	0.66
LumiOpen/Llama-Poro-2-8B-base	0.54	0.68	0.72	0.66	0.7	0.77	0.55	0.64	0.68	nan	nan	nan	nan	nan	0.66
google/gemma-2-9b	0.47	0.74	0.7	0.68	0.69	0.81	0.54	0.61	0.65	nan	nan	nan	nan	nan	0.654
7-10B Weight Class (Gen)															
mistralai/Mistral-7B-Instruct-v0.3	0.57	0.55	0.55	0.59	0.59	0.82	0.53	0.59	0.7	0.5	0.51	0.49	0.55	0.67	0.586
Qwen/Qwen2.5-7B-Instruct	0.54	0.56	0.54	0.64	0.51	0.86	0.52	0.56	0.61	0.5	0.54	0.54	0.48	0.65	0.575
sail/Sailor2-8B-Chat	0.5	0.75	0.55	0.43	0.49	0.78	0.52	0.58	0.69	0.44	0.58	0.61	0.43	0.56	0.565
CohereLabs/aya-expanse-8b	0.54	0.48	0.56	0.52	0.51	0.86	0.52	0.53	0.6	0.5	0.53	0.51	0.58	0.64	0.563
12-20B Weight Class (Gen)															
openai/gpt-oss-20b	0.6	0.8	0.68	0.56	0.62	0.88	0.49	0.66	0.8	0.5	0.81	0.5	0.56	0.77	0.659
google/gemma-3-12b-it	0.53	0.83	0.66	0.45	0.55	0.91	0.64	0.59	0.82	0.56	0.73	0.54	0.52	0.84	0.655
Qwen/Qwen2.5-14B-Instruct	0.55	0.66	0.49	0.61	0.58	0.92	0.51	0.46	0.57	0.51	0.66	0.57	0.58	0.78	0.604
sail/Sailor2-20B-Chat	0.52	0.7	0.6	0.53	0.47	0.88	0.47	0.53	0.77	0.48	0.62	0.55	0.51	0.66	0.592
27-32B Weight Class (Gen)															
google/gemma-3-27b-it	0.48	0.74	0.67	0.55	0.58	0.91	0.61	0.63	0.88	0.57	0.8	0.5	0.61	0.88	0.672
Qwen/Qwen2.5-32B-Instruct	0.6	0.48	0.62	0.57	0.61	0.9	0.47	0.59	0.55	0.55	0.55	0.61	0.61	0.72	0.602
CohereLabs/aya-expanse-32b	0.5	0.59	0.64	0.46	0.58	0.9	0.46	0.52	0.65	0.47	0.53	0.55	0.53	0.77	0.582
LGAI-EXAONE/EXAONE-3.5-32B-Instruct	0.5	0.62	0.57	0.55	0.59	0.88	0.49	0.52	0.69	0.47	0.45	0.54	0.45	0.69	0.572
70-72B Weight Class (Gen)															
meta-llama/Llama-3.1-70B-Instruct	0.56	0.77	0.65	0.59	0.53	0.92	0.6	0.54	0.76	0.57	0.72	0.63	0.53	0.9	0.662
swiss-ai/Apertus-70B-Instruct-2509	0.59	0.55	0.56	0.58	0.59	0.85	0.46	0.61	0.74	0.41	0.67	0.63	0.57	0.83	0.617
Qwen/Qwen2.5-72B-Instruct	0.55	0.6	0.65	0.52	0.59	0.92	0.55	0.65	0.68	0.5	0.57	0.59	0.44	0.8	0.615
Closed Models (Gen)															
gemini-pro	0.81	0.96	0.86	0.6	0.77	0.95	0.6	0.87	0.89	0.52	0.94	0.68	0.9	0.88	0.802
sonnet-4-5	0.65	0.92	0.89	0.63	0.63	0.95	0.52	0.8	0.89	nan	nan	nan	nan	nan	0.764
flash	0.79	0.96	0.88	0.48	0.77	0.96	0.53	0.79	0.9	0.38	0.93	0.66	0.76	0.89	0.763
gpt-5-mini	0.52	0.94	0.88	0.58	0.64	0.94	0.55	0.71	0.86	nan	nan	nan	nan	nan	0.736

Table 15: Subsaharan Africa

G.6 Southeast Asian Languages

Model	ind_latn	jav_latn	tgl_latn	tha_thai	vie_latn	zsm_latn	Avg.
Sub-1B Weight Class (LL)							
google/gemma-3-270m	0.6	0.55	0.5	0.58	0.57	0.5	0.55
bigscience/bloom-560m	0.52	0.56	0.49	0.51	0.68	0.47	0.538
Qwen/Qwen2.5-0.5B	0.47	0.48	0.55	0.54	0.63	0.52	0.532
1B Weight Class (LL)							
sail/Sailor2-1B	0.73	0.49	0.6	0.67	0.71	0.56	0.627
google/gemma-3-1b-pt	0.68	0.56	0.6	0.6	0.65	0.5	0.598
CraneAILabs/swahili-gemma-1b	0.64	0.56	0.58	0.57	0.63	0.54	0.587
bigscience/bloom-1b1	0.66	0.51	0.56	0.53	0.73	0.53	0.587
2-3B Weight Class (LL)							
google/gemma-3-4b-pt	0.73	0.58	0.74	0.67	0.82	0.58	0.687
google/gemma-2-2b	0.7	0.44	0.6	0.63	0.73	0.59	0.615
Qwen/Qwen2.5-3B	0.67	0.52	0.56	0.63	0.71	0.5	0.598
facebook/xglm-4.5B	0.7	0.48	0.56	0.61	0.67	0.57	0.598
2-3B Weight Class (Gen)							
Qwen/Qwen3-4B	0.91	0.68	0.71	0.81	0.75	0.81	0.778
google/gemma-3-4b-it	0.86	0.61	0.7	0.72	0.59	0.81	0.715
Qwen/Qwen2.5-3B-Instruct	0.82	0.57	0.61	0.76	0.76	0.7	0.703
google/gemma-2-2b-it	0.79	0.6	0.62	0.57	0.54	0.81	0.655
7-10B Weight Class (LL)							
sail/Sailor2-8B	0.79	0.62	0.79	0.7	0.82	0.71	0.738
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.78	0.6	0.71	0.73	0.79	0.71	0.72
swiss-ai/Apertus-8B-2509	0.76	0.57	0.75	0.7	0.83	0.69	0.717
google/gemma-2-9b	0.76	0.55	0.75	0.68	0.79	0.66	0.698
7-10B Weight Class (Gen)							
Qwen/Qwen3-8B	0.91	0.75	0.85	0.87	0.79	0.81	0.83
google/gemma-2-9b-it	0.95	0.69	0.82	0.79	0.69	0.83	0.795
sail/Sailor2-8B-Chat	0.89	0.76	0.79	0.8	0.73	0.78	0.792
Qwen/Qwen2.5-7B-Instruct	0.91	0.6	0.71	0.82	0.77	0.8	0.768
12-20B Weight Class (Gen)							
Qwen/Qwen3-14B	0.95	0.81	0.92	0.85	0.83	0.84	0.867
openai/gpt-oss-20b	0.94	0.78	0.94	0.86	0.77	0.89	0.863
google/gemma-3-12b-it	0.94	0.69	0.87	0.81	0.78	0.88	0.828
deepseek-ai/DeepSeek-R1-Distill-Qwen-14B	0.91	0.61	0.82	0.86	0.85	0.86	0.818
27-32B Weight Class (Gen)							
google/gemma-3-27b-it	0.96	0.85	0.88	0.86	0.81	0.88	0.873
Qwen/Qwen2.5-32B-Instruct	0.95	0.76	0.87	0.89	0.87	0.85	0.865
google/gemma-2-27b-it	0.93	0.73	0.87	0.87	0.83	0.85	0.847
CohereLabs/aya-expanse-32b	0.93	0.79	0.81	0.67	0.79	0.81	0.8
70-72B Weight Class (Gen)							
Qwen/Qwen2.5-72B-Instruct	0.98	0.79	0.9	0.9	0.9	0.85	0.887
meta-llama/Llama-3.1-70B-Instruct	0.96	0.67	0.9	0.81	0.81	0.87	0.837
swiss-ai/Apertus-70B-Instruct-2509	0.93	0.69	0.66	0.76	0.75	0.84	0.772
Closed Models (Gen)							
sonnet-4-5	0.96	0.94	0.94	0.93	0.91	0.97	0.942
gpt-5	0.94	0.92	0.96	0.92	0.93	0.94	0.935
gpt-5-mini	0.96	0.89	0.93	0.93	0.89	0.96	0.927
gemini-pro	0.96	0.92	0.91	0.92	0.9	0.93	0.923

Table 16: Southeast Asia

G.7 South Asian Languages

Model	ben_latn	bho_deva	dhd_deva	guj_gujr	mar_deva	nag_latn	npi_deva	rwr_deva	sin_sinh	snd_arab	snd_deva	urd_arab	urd_latn	asm_beng	ben_beng	hin_deva	pan_guru	Avg.
Sub-1B Weight Class (LL)																		
bigscience/bloom-560m	0.56	0.57	0.62	0.5	0.51	0.5	0.48	0.53	0.51	0.51	0.57	0.6	0.48	nan	nan	nan	nan	0.534
google/gemma-3-270m	0.52	0.47	0.59	0.53	0.45	0.55	0.53	0.52	0.56	0.52	0.55	0.59	0.45	nan	nan	nan	nan	0.525
Qwen/Qwen2.5-0.5B	0.47	0.46	0.62	0.45	0.48	0.55	0.46	0.4	0.5	0.57	0.59	0.48	0.49	nan	nan	nan	nan	0.502
1B Weight Class (LL)																		
google/gemma-3-1b-pt	0.4	0.5	0.58	0.58	0.45	0.53	0.56	0.54	0.58	0.61	0.61	0.71	0.55	nan	nan	nan	nan	0.554
CraneAILabs/swahili-gemma-1b	0.56	0.47	0.63	0.63	0.41	0.52	0.5	0.55	0.5	0.57	0.56	0.68	0.48	nan	nan	nan	nan	0.543
CraneAILabs/ganda-gemma-1b	0.56	0.48	0.62	0.58	0.47	0.58	0.49	0.54	0.47	0.55	0.55	0.64	0.5	nan	nan	nan	nan	0.541
facebook/xglm-1.7B	0.47	0.48	0.57	0.6	0.48	0.49	0.55	0.49	0.59	0.47	0.57	0.73	0.49	nan	nan	nan	nan	0.537
2-3B Weight Class (LL)																		
google/gemma-3-4b-pt	0.44	0.55	0.58	0.64	0.62	0.56	0.65	0.56	0.53	0.62	0.56	0.77	0.66	nan	nan	nan	nan	0.595
meta-llama/Llama-3.2-3B	0.44	0.54	0.62	0.57	0.46	0.56	0.51	0.5	0.55	0.55	0.56	0.63	0.56	nan	nan	nan	nan	0.542
bigscience/bloom-3b	0.46	0.53	0.6	0.54	0.47	0.52	0.51	0.55	0.57	0.56	0.6	0.64	0.44	nan	nan	nan	nan	0.538
facebook/xglm-2.9B	0.49	0.53	0.57	0.56	0.48	0.55	0.53	0.48	0.47	0.48	0.59	0.73	0.47	nan	nan	nan	nan	0.533
2-3B Weight Class (Gen)																		
Qwen/Qwen3-4B	0.54	0.71	0.74	0.81	0.82	0.72	0.82	0.81	0.55	0.94	0.68	0.86	0.66	0.87	0.74	0.82	0.86	0.762
google/gemma-3-4b-it	0.6	0.69	0.57	0.78	0.83	0.54	0.78	0.77	0.54	0.93	0.63	0.83	0.75	0.8	0.82	0.83	0.75	0.732
google/gemma-2-2b-it	0.46	0.56	0.65	0.65	0.61	0.55	0.67	0.71	0.54	0.74	0.56	0.76	0.55	0.63	0.58	0.71	0.64	0.622
microsoft/Phi-4-mini-instruct	0.41	0.56	0.53	0.61	0.58	0.51	0.61	0.6	0.55	0.85	0.53	0.65	0.55	0.61	0.54	0.73	0.63	0.591
7-10B Weight Class (LL)																		
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.46	0.57	0.65	0.63	0.56	0.64	0.63	0.57	0.53	0.68	0.57	0.78	0.7	nan	nan	nan	nan	0.613
swiss-ai/Apertus-8B-2509	0.42	0.54	0.62	0.65	0.59	0.54	0.69	0.54	0.55	0.62	0.58	0.77	0.68	nan	nan	nan	nan	0.599
google/gemma-2-9b	0.41	0.53	0.6	0.61	0.55	0.6	0.58	0.5	0.61	0.56	0.59	0.79	0.73	nan	nan	nan	nan	0.589
aisingapore/Llama-SEA-LION-v3-8B-IT	0.44	0.52	0.61	0.67	0.48	0.55	0.59	0.53	0.55	0.67	0.56	0.75	0.61	nan	nan	nan	nan	0.579
7-10B Weight Class (Gen)																		
Qwen/Qwen3-8B	0.6	0.79	0.73	0.86	0.83	0.63	0.84	0.85	0.55	0.99	0.74	0.95	0.83	0.92	0.86	0.89	0.87	0.808
google/gemma-2-9b-it	0.69	0.76	0.75	0.76	0.87	0.66	0.87	0.88	0.43	0.93	0.71	0.92	0.86	0.87	0.84	0.93	0.85	0.799
swiss-ai/Apertus-8B-Instruct-2509	0.5	0.6	0.67	0.8	0.79	0.57	0.77	0.74	0.53	0.97	0.61	0.81	0.65	0.86	0.79	0.83	0.76	0.721
Qwen/Qwen2.5-7B-Instruct	0.51	0.63	0.65	0.73	0.68	0.62	0.75	0.73	0.52	0.83	0.65	0.71	0.69	0.67	0.79	0.81	0.68	0.685
12-20B Weight Class (Gen)																		
google/gemma-3-12b-it	0.78	0.86	0.8	0.91	0.91	0.73	0.9	0.93	0.64	1.0	0.67	0.93	0.92	0.87	0.88	0.9	0.88	0.854
openai/gpt-oss-20b	0.78	0.82	0.81	0.9	0.88	0.8	0.9	0.93	0.5	1.0	0.74	0.9	0.84	0.88	0.87	0.95	0.89	0.846
Qwen/Qwen3-14B	0.72	0.82	0.81	0.91	0.86	0.66	0.87	0.92	0.56	0.99	0.71	0.95	0.84	0.91	0.88	0.92	0.92	0.838
microsoft/phi-4	0.63	0.84	0.83	0.79	0.88	0.65	0.83	0.93	0.58	0.93	0.73	0.84	0.71	0.82	0.81	0.92	0.8	0.795
27-32B Weight Class (Gen)																		
google/gemma-3-27b-it	0.84	0.85	0.81	0.9	0.92	0.75	0.96	0.93	0.63	1.0	0.79	0.98	0.93	0.95	0.91	0.9	0.9	0.879
google/gemma-2-27b-it	0.71	0.82	0.75	0.9	0.87	0.72	0.92	0.9	0.56	0.98	0.71	0.9	0.91	0.86	0.91	0.91	0.87	0.835
sarvamai/sarvam-m	0.84	0.84	0.78	0.89	0.88	0.76	0.87	0.88	0.38	0.88	0.77	0.93	0.87	0.93	0.87	0.88	0.9	0.832
Qwen/Qwen2.5-32B-Instruct	0.75	0.79	0.79	0.82	0.8	0.73	0.85	0.88	0.55	0.94	0.77	0.91	0.81	0.88	0.85	0.94	0.86	0.819
70-72B Weight Class (Gen)																		
meta-llama/Llama-3.1-70B-Instruct	0.69	0.88	0.83	0.9	0.89	0.66	0.84	0.94	0.6	1.0	0.67	0.96	0.87	0.91	0.89	0.92	0.88	0.843
Qwen/Qwen2.5-72B-Instruct	0.77	0.83	0.78	0.9	0.87	0.7	0.9	0.91	0.49	0.96	0.7	0.96	0.88	0.87	0.93	0.92	0.93	0.841
swiss-ai/Apertus-70B-Instruct-2509	0.47	0.76	0.69	0.81	0.9	0.6	0.87	0.85	0.51	0.98	0.7	0.91	0.82	0.94	0.86	0.87	0.72	0.78
Closed Models (Gen)																		
gemini-pro	0.95	0.91	0.91	0.91	0.94	0.87	0.98	0.95	0.9	1.0	0.95	0.98	1.0	0.95	0.94	0.95	0.93	0.942
flash	0.96	0.91	0.89	0.92	0.96	0.78	0.99	0.96	0.87	1.0	0.91	0.96	0.98	0.94	0.94	0.93	0.95	0.932
sonnet-4-5	0.96	0.92	0.9	0.94	0.94	0.88	1.0	0.95	0.81	1.0	0.86	0.97	0.99	nan	nan	nan	nan	0.932
gpt-5	0.94	0.88	0.92	0.95	0.9	0.86	0.97	0.94	0.78	1.0	0.88	0.97	1.0	0.95	0.96	0.96	0.95	0.93

Table 17: South Asian, Indo-European

Model	bsk_arab	mni_beng	tam_taml	kan_knda	mal_mlym	mni_mtei	tel_telu	Avg.
Sub-1B Weight Class (LL)								
Qwen/Qwen2.5-0.5B	0.52	0.47	0.54	nan	nan	nan	nan	0.51
bigscience/bloom-560m	0.47	0.43	0.5	nan	nan	nan	nan	0.467
google/gemma-3-270m	0.47	0.45	0.48	nan	nan	nan	nan	0.467
1B Weight Class (LL)								
SeaLLMs/SeaLLMs-v3-1.5B	0.56	0.49	0.53	nan	nan	nan	nan	0.527
kakaocorp/kanana-1.5-2.1b-base	0.56	0.47	0.54	nan	nan	nan	nan	0.523
meta-llama/Llama-3.2-1B	0.54	0.47	0.54	nan	nan	nan	nan	0.517
google/gemma-3-1b-pt	0.54	0.44	0.55	nan	nan	nan	nan	0.51
2-3B Weight Class (LL)								
meta-llama/Llama-3.2-3B	0.51	0.51	0.58	nan	nan	nan	nan	0.533
TucanoBR/Tucano-2b4	0.51	0.56	0.5	nan	nan	nan	nan	0.523
google/gemma-2-2b	0.56	0.45	0.55	nan	nan	nan	nan	0.52
google/gemma-3-4b-pt	0.51	0.46	0.57	nan	nan	nan	nan	0.513
2-3B Weight Class (Gen)								
google/gemma-3-4b-it	0.47	0.48	0.78	0.76	0.57	0.61	0.77	0.634
Qwen/Qwen3-4B	0.48	0.52	0.64	0.77	0.6	0.47	0.78	0.609
microsoft/Phi-4-mini-instruct	0.5	0.53	0.68	0.52	0.54	0.53	0.58	0.554
speakleash/Bielik-4.5B-v3.0-Instruct	0.56	0.43	0.56	0.51	0.52	0.53	0.55	0.523
7-10B Weight Class (LL)								
ilsp/Llama-Krikri-8B-Base	0.53	0.53	0.55	nan	nan	nan	nan	0.537
kakaocorp/kanana-1.5-8b-base	0.58	0.46	0.57	nan	nan	nan	nan	0.537
sail/Sailor2-8B	0.53	0.43	0.64	nan	nan	nan	nan	0.533
bertin-project/Gromenauer-7B	0.53	0.47	0.58	nan	nan	nan	nan	0.527
7-10B Weight Class (Gen)								
Qwen/Qwen3-8B	0.45	0.52	0.69	0.85	0.64	0.46	0.91	0.646
swiss-ai/Apertus-8B-Instruct-2509	0.41	0.5	0.74	0.78	0.6	0.5	0.8	0.619
CohereLabs/c4ai-command-r7b-12-2024	0.55	0.57	0.61	0.67	0.52	0.53	0.48	0.561
LumiOpen/Llama-Poro-2-8B-Instruct	0.45	0.51	0.66	0.57	0.55	0.51	0.62	0.553
12-20B Weight Class (Gen)								
Qwen/Qwen3-14B	0.55	0.59	0.78	0.85	0.71	0.54	0.92	0.706
google/gemma-3-12b-it	0.59	0.56	0.82	0.85	0.71	0.49	0.88	0.7
microsoft/phi-4	0.66	0.54	0.77	0.8	0.58	0.54	0.82	0.673
openai/gpt-oss-20b	0.51	0.46	0.71	0.94	0.73	0.37	0.93	0.664
27-32B Weight Class (Gen)								
sarvamai/sarvam-m	0.49	0.54	0.78	0.92	0.71	0.57	0.91	0.703
google/gemma-3-27b-it	0.45	0.51	0.83	0.83	0.77	0.63	0.86	0.697
CohereLabs/aya-expanse-32b	0.48	0.49	0.78	0.7	0.7	0.56	0.57	0.611
Qwen/Qwen2.5-32B-Instruct	0.47	0.44	0.74	0.74	0.58	0.43	0.71	0.587
70-72B Weight Class (Gen)								
meta-llama/Llama-3.1-70B-Instruct	0.47	0.62	0.82	0.88	0.66	0.48	0.89	0.689
Qwen/Qwen2.5-72B-Instruct	0.45	0.59	0.73	0.67	0.65	0.56	0.7	0.621
swiss-ai/Apertus-70B-Instruct-2509	0.39	0.55	0.72	0.72	0.64	0.42	0.85	0.613
Closed Models (Gen)								
gemini-pro	0.54	0.88	0.9	0.94	0.86	0.52	0.94	0.797
flash	0.47	0.85	0.88	0.93	0.81	0.45	0.9	0.756
gpt-5-mini	0.46	0.57	0.82	nan	nan	nan	nan	0.617
gpt-5	0.09	0.45	0.87	0.94	0.82	0.1	0.94	0.601
Closed Models								
gemini-pro	0.54	0.88	0.9	0.94	0.86	0.52	0.94	0.797
flash	0.47	0.85	0.88	0.93	0.81	0.45	0.9	0.756
gpt-5-mini	0.46	0.57	0.82	nan	nan	nan	nan	0.617
gpt-5	0.09	0.45	0.87	0.94	0.82	0.1	0.94	0.601
sonnet-4-5	0.48	0.64	nan	nan	nan	nan	nan	0.56
flash-lite	0.41	0.62	nan	nan	nan	nan	nan	0.515
gpt-5-nano	0.28	0.37	0.68	nan	nan	nan	nan	0.443

Table 18: South Asian, non-Indo-European Languages

G.8 East Asian Languages

Model	cmn_hans	cmn_hant	jpn_jpan	kor_hang	yue_hant	Avg.
Sub-1B Weight Class (LL)						
google/gemma-3-270m	0.53	0.52	0.64	0.54	0.56	0.558
Qwen/Qwen2.5-0.5B	0.49	0.47	0.69	0.47	0.57	0.538
bigscience/bloom-560m	0.47	0.43	0.61	0.56	0.58	0.53
1B Weight Class (LL)						
google/gemma-3-1b-pt	0.56	0.56	0.74	0.55	0.68	0.618
kakaocorp/kanana-1.5-2.1b-base	0.53	0.53	0.65	0.61	0.57	0.578
SeaLLMs/SeaLLMs-v3-1.5B	0.56	0.46	0.71	0.52	0.64	0.578
Qwen/Qwen2.5-1.5B	0.51	0.51	0.7	0.52	0.64	0.576
2-3B Weight Class (LL)						
google/gemma-2-2b	0.55	0.54	0.74	0.48	0.74	0.61
facebook/xglm-4.5B	0.52	0.5	0.75	0.58	0.69	0.608
Qwen/Qwen2.5-3B	0.52	0.52	0.76	0.51	0.72	0.606
google/gemma-3-4b-pt	0.53	0.47	0.79	0.53	0.71	0.606
2-3B Weight Class (Gen)						
Qwen/Qwen3-4B	0.8	0.76	0.87	0.68	0.82	0.786
google/gemma-3-4b-it	0.64	0.61	0.85	0.6	0.69	0.678
Qwen/Qwen2.5-3B-Instruct	0.67	0.6	0.8	0.57	0.67	0.662
microsoft/Phi-3.5-mini-instruct	0.62	0.56	0.82	0.55	0.57	0.624
7-10B Weight Class (LL)						
SeaLLMs/SeaLLMs-v3-7B	0.65	0.57	0.75	0.59	0.78	0.668
aisingapore/Gemma-SEA-LION-v3-9B-IT	0.57	0.54	0.79	0.6	0.7	0.64
Qwen/Qwen2.5-7B	0.6	0.53	0.76	0.57	0.71	0.634
sail/Sailor2-8B	0.62	0.57	0.81	0.5	0.66	0.632
7-10B Weight Class (Gen)						
Qwen/Qwen3-8B	0.82	0.81	0.89	0.77	0.83	0.824
google/gemma-2-9b-it	0.7	0.67	0.86	0.71	0.81	0.75
Qwen/Qwen2.5-7B-Instruct	0.77	0.69	0.92	0.62	0.72	0.744
CohereLabs/aya-expanse-8b	0.69	0.66	0.89	0.62	0.69	0.71
12-20B Weight Class (Gen)						
Qwen/Qwen3-14B	0.88	0.82	0.91	0.85	0.83	0.858
Qwen/Qwen2.5-14B-Instruct	0.88	0.79	0.94	0.73	0.87	0.842
openai/gpt-oss-20b	0.81	0.81	0.94	0.69	0.81	0.812
sail/Sailor2-20B-Chat	0.81	0.77	0.92	0.71	0.8	0.802
27-32B Weight Class (Gen)						
Qwen/Qwen2.5-32B-Instruct	0.85	0.82	0.92	0.76	0.9	0.85
google/gemma-3-27b-it	0.82	0.72	0.93	0.8	0.84	0.822
CohereLabs/aya-expanse-32b	0.73	0.72	0.93	0.81	0.84	0.806
sarvamai/sarvam-m	0.81	0.71	0.9	0.79	0.78	0.798
70-72B Weight Class (Gen)						
Qwen/Qwen2.5-72B-Instruct	0.88	0.86	0.93	0.82	0.92	0.882
meta-llama/Llama-3.1-70B-Instruct	0.73	0.71	0.91	0.8	0.83	0.796
swiss-ai/Apertus-70B-Instruct-2509	0.73	0.71	0.9	0.62	0.74	0.74
Closed Models (Gen)						
gpt-5	0.86	0.9	0.96	0.92	0.93	0.914
gemini-pro	0.87	0.9	0.94	0.92	0.92	0.91
flash	0.9	0.88	0.97	0.91	0.85	0.902
gpt-5-mini	0.85	0.78	0.96	0.87	0.9	0.872

Table 19: East Asia

Model	kaz_cyrl	kir_cyrl	uig_arab	uzn_latn	Avg.
Sub-1B Weight Class (LL)					
google/gemma-3-270m	0.59	0.52	0.53	0.58	0.555
bigscience/bloom-560m	0.54	0.56	0.48	0.58	0.54
Qwen/Qwen2.5-0.5B	0.55	0.46	0.51	0.61	0.532
1B Weight Class (LL)					
google/gemma-3-1b-pt	0.57	0.52	0.5	0.64	0.557
kakaocorp/kanana-1.5-2.1b-base	0.54	0.5	0.56	0.59	0.547
AI-Sweden-Models/gpt-sw3-1.3b	0.6	0.53	0.52	0.53	0.545
CraneAILabs/ganda-gemma-1b	0.59	0.48	0.53	0.58	0.545
2-3B Weight Class (LL)					
google/gemma-3-4b-pt	0.76	0.64	0.55	0.68	0.657
meta-llama/Llama-3.2-3B	0.69	0.55	0.46	0.69	0.597
facebook/xglm-4.5B	0.72	0.51	0.5	0.55	0.57
Qwen/Qwen2.5-3B	0.57	0.52	0.54	0.58	0.552
2-3B Weight Class (Gen)					
Qwen/Qwen3-4B	0.71	0.69	0.55	0.7	0.662
google/gemma-3-4b-it	0.61	0.68	0.51	0.66	0.615
microsoft/Phi-3.5-mini-instruct	0.61	0.5	0.48	0.56	0.537
microsoft/Phi-3-mini-4k-instruct	0.62	0.47	0.48	0.55	0.53
7-10B Weight Class (LL)					
swiss-ai/Apertus-8B-2509	0.79	0.68	0.54	0.8	0.703
google/gemma-2-9b	0.79	0.6	0.43	0.68	0.625
aisingapore/Llama-SEA-LION-v3-8B-IT	0.74	0.61	0.5	0.63	0.62
meta-llama/Llama-3.1-8B	0.74	0.55	0.5	0.64	0.608
7-10B Weight Class (Gen)					
Qwen/Qwen3-8B	0.72	0.7	0.6	0.79	0.703
swiss-ai/Apertus-8B-Instruct-2509	0.67	0.74	0.63	0.6	0.66
google/gemma-2-9b-it	0.7	0.66	0.56	0.68	0.65
Qwen/Qwen2.5-7B-Instruct	0.6	0.6	0.52	0.64	0.59
12-20B Weight Class (Gen)					
google/gemma-3-12b-it	0.74	0.89	0.74	0.77	0.785
Qwen/Qwen3-14B	0.78	0.85	0.63	0.77	0.757
openai/gpt-oss-20b	0.77	0.81	0.67	0.77	0.755
microsoft/phi-4	0.66	0.71	0.65	0.57	0.647
27-32B Weight Class (Gen)					
google/gemma-3-27b-it	0.79	0.88	0.72	0.84	0.807
google/gemma-2-27b-it	0.72	0.86	0.53	0.7	0.703
Qwen/Qwen2.5-32B-Instruct	0.61	0.69	0.59	0.71	0.65
sarvamai/sarvam-m	0.65	0.77	0.46	0.71	0.647
70-72B Weight Class (Gen)					
Qwen/Qwen2.5-72B-Instruct	0.76	0.83	0.64	0.81	0.76
meta-llama/Llama-3.1-70B-Instruct	0.79	0.83	0.65	0.76	0.758
swiss-ai/Apertus-70B-Instruct-2509	0.65	0.78	0.63	0.76	0.705
Closed Models (Gen)					
gemini-pro	0.9	0.98	0.94	0.91	0.932
gpt-5	0.92	0.98	0.93	0.9	0.932
flash	0.88	0.99	0.9	0.92	0.922
gpt-5-mini	0.87	0.98	0.85	0.93	0.908

Table 20: Central Asia

Model	fra_latn_cana	haw_latn	por_latn_braz	spa_latn_peru	spa_latn_mexi	Avg.
Sub-1B Weight Class (LL)						
bigscience/bloom-560m	0.7	0.58	0.55	0.69	nan	0.63
Qwen/Qwen2.5-0.5B	0.7	0.54	0.44	0.71	nan	0.597
google/gemma-3-270m	0.66	0.54	0.42	0.69	nan	0.578
1B Weight Class (LL)						
google/gemma-3-1b-pt	0.75	0.59	0.66	0.76	nan	0.69
Qwen/Qwen2.5-1.5B	0.8	0.53	0.62	0.76	nan	0.677
bigscience/bloom-1b7	0.75	0.5	0.64	0.72	nan	0.653
facebook/xglm-1.7B	0.67	0.58	0.57	0.78	nan	0.65
2-3B Weight Class (LL)						
google/gemma-3-4b-pt	0.86	0.53	0.79	0.85	nan	0.758
facebook/xglm-4.5B	0.78	0.59	0.68	0.82	nan	0.718
Qwen/Qwen2.5-3B	0.82	0.52	0.73	0.79	nan	0.715
google/gemma-2-2b	0.83	0.51	0.66	0.78	nan	0.695
2-3B Weight Class (Gen)						
Qwen/Qwen3-4B	0.93	0.47	0.92	0.96	0.94	0.844
Qwen/Qwen2.5-3B-Instruct	0.89	0.51	0.8	0.92	0.9	0.804
google/gemma-3-4b-it	0.86	0.52	0.82	0.96	0.86	0.804
microsoft/Phi-3-mini-4k-instruct	0.91	0.51	0.78	0.97	0.84	0.802
7-10B Weight Class (LL)						
google/gemma-2-9b	0.89	0.52	0.84	0.84	nan	0.772
utter-project/EuroLLM-9B	0.92	0.43	0.85	0.85	nan	0.763
Qwen/Qwen2.5-7B	0.88	0.54	0.81	0.82	nan	0.762
Tower-Babel/Babel-9B	0.89	0.51	0.81	0.83	nan	0.76
7-10B Weight Class (Gen)						
Qwen/Qwen3-8B	0.96	0.52	0.91	0.98	0.94	0.862
google/gemma-2-9b-it	0.94	0.53	0.87	0.96	0.96	0.852
Qwen/Qwen2.5-7B-Instruct	0.94	0.55	0.83	0.97	0.88	0.834
LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct	0.88	0.56	0.8	0.94	0.87	0.81
12-20B Weight Class (Gen)						
Qwen/Qwen3-14B	0.97	0.62	0.88	0.99	0.95	0.882
deepseek-ai/DeepSeek-R1-Distill-Qwen-14B	0.93	0.65	0.92	0.95	0.95	0.88
Qwen/Qwen2.5-14B-Instruct	0.97	0.56	0.92	0.97	0.96	0.876
openai/gpt-oss-20b	0.96	0.55	0.91	0.98	0.94	0.868
27-32B Weight Class (Gen)						
google/gemma-3-27b-it	0.97	0.66	0.96	0.96	0.94	0.898
Qwen/Qwen2.5-32B-Instruct	0.97	0.63	0.93	0.98	0.96	0.894
google/gemma-2-27b-it	0.94	0.65	0.9	0.98	0.96	0.886
sarvamai/sarvam-m	0.97	0.53	0.92	0.93	0.95	0.86
70-72B Weight Class (Gen)						
Qwen/Qwen2.5-72B-Instruct	0.96	0.68	0.97	0.99	0.99	0.918
meta-llama/Llama-3.1-70B-Instruct	0.91	0.56	0.9	0.98	0.95	0.86
swiss-ai/Apertus-70B-Instruct-2509	0.89	0.58	0.92	0.98	0.86	0.846
Closed Models (Gen)						
gemini-pro	0.98	0.95	0.95	0.98	0.99	0.97
flash	0.98	0.91	0.97	0.98	0.98	0.964
gpt-5	0.98	0.91	0.98	0.98	0.97	0.964
flash-lite	0.97	0.87	0.94	0.97	nan	0.938

Table 21: Americas and Oceania

G.9 Central Asian Languages

G.10 Americas and Oceania

G.11 Base vs. Instruction versions of open-weight models

Instruct is better for 7B and up, sometimes better for 3-4B, almost always worse for less than 3B

H Individual Dataset Descriptions

Here, we provide brief descriptions of the methods that individual groups used to construct their contributions to the non-parallel split of Global PIQA (§3). Longer dataset description papers that authors consented to release are at `todo`. Authors were recruited and organized as described in §3.1, and all contributors were offered authorship. The vast majority have chosen to be authors on this paper. This project would not be possible without the efforts of all authors.

We note that we intentionally do not list authors with their groups and languages. This is to preserve privacy, as some authors would prefer not to be contacted by a large number of unaffiliated projects that require expertise in their language.

Group 0000: Hindi (`hin_deva`: N examples)
Manually written in English by a native Hindi speaker, machine-translated into Hindi using Google Translate, then checked, corrected, and refined by the dataset author. Approximately 25% of examples are designed to be culturally-grounded, with references to specific Indian culinary items, musical instruments, common fauna, and social traditions, such as customs within a wedding ceremony.

Group 0001: Telugu (`tel_telu`: N examples)
Manually written by a native Telugu speaker, with examples crafted to reflect realistic scenarios encountered in Telugu households, agriculture, cooking, transportation, and daily problem-solving. Each question was double-checked, and edge cases and ambiguous situations were discarded to ensure high quality.

Group 0002: French (Canadian) (`fra_latn_cana`: N examples)
Topic ideas were brainstormed using LLMs, but examples were all written manually. All examples were checked or written by a native speaker.

Group 0003: Yoruba (`yor_latn`: N examples)
Examples from English PIQA were translated and culturally adapted to Yoruba by a native Yoruba speaker. Care was taken to preserve Yoruba idiomatic forms, and for culturally unique contexts, questions were created directly in Yoruba rather than translated. Culturally-specific domains include cooking, clothing, farming, weather, transportation, religion, household practices, and festivals.

Group 0004: French (`fra_latn_fran`: N examples)
Manually written by a native French speaker, with examples crafted by observing daily life and social interactions, and by browsing French websites for topics such as furniture, home goods, sports, and news. Many examples were designed to be specific to French culture, e.g. including French food and social norms, or how to take the metro in Paris.

Group 0005: Finnish (`fin_latn`: N examples)
Manually written by a native Finnish speaker, with many examples covering Finnish culture and everyday life. Topics include traditional foods, household chores, log cabin terms, saunas, winter activities, reindeer-related terms, and Finnish sports and traditions.

Group 0006: Hungarian, Romanian (`hun_latn`, `ron_latn`: N examples)
Examples were written in English, translated into Hungarian and Romanian (by native speakers of those languages), and reviewed by another translator. All translators and editors were offered authorship.

Group 0007: Ukrainian (`ukr_cyrl`: N examples)

Manually written by a native Ukrainian speaker, and checked by another native speaker, both from Western Ukraine. Topics were inspired by Ukrainian websites and blogs, as well as personal knowledge, covering Ukrainian cuisine, traditions, superstitions, and local Ukrainian festivities.

Group 0008: Mandarin (`cmn_hans`, `cmn_hant`: N examples)
Manually written by a native Mandarin speaker and verified by another native speaker. Examples were balanced across culturally-specific food, clothing and materials, musical instruments, and other objects. Examples were written using Chinese simplified characters, but also translated into traditional characters using Google Translate with human verification.

Group 0009: Hebrew (`heb_hebr`: N examples)
Manually written by a native Hebrew speaker, with examples covering specific Hebrew linguistic constructions, along with Israeli cultural knowledge, such as places, food, climate, and Jewish religion and culture. By design, some items may resist direct translation into other languages, and in some cases, translation may alter the validity of the designated correct answer.

Group 0010: Indonesian (`ind_latn`: N examples)
Examples were generated with the assistance of ChatGPT (GPT-5) using carefully guided prompts to produce PIQA-style examples. All examples were manually reviewed, corrected, and finalized by a native speaker of Indonesian to ensure quality, correctness, and cultural relevance. Because the original LLM-generated examples were often fairly generic, at least 50 examples were manually edited to reflect uniquely Indonesian contexts (e.g. local foods, household practices, and traditional objects). The dataset was written in Standard Indonesian (Bahasa Indonesia).

Group 0011: Italian (`ita_latn`: N examples)
Manually written by a native Italian speaker. ChatGPT was occasionally used to correct typos or to find appropriate words that did not immediately come to mind, but never to generate examples themselves. All final versions of examples were human verified. To include examples reflecting Italian culture, some examples were motivated by online recipes and websites in Italian.

Group 0012: Hausa (`hau_latn`: N examples)
Manually written by a native Hausa speaker, using culturally-relevant themes to motivate example creation. Themes included traveling, food, school, exams, driving, and health.

Group 0013: Portuguese (Brazilian) (`por_latn_braz`: N examples)
Manually written by a native Brazilian Portuguese speaker, covering food, traditions, regional objects, daily activities, and environmental contexts that are common to Brazil, particularly southern Brazil.

Group 0014: Dutch (`nld_latn`: N examples)
Manually written by a native Dutch speaker, using specific culturally-relevant topics to motivate example creation. Topics include bicycle maintenance techniques, preparation of traditional Dutch foods, managing Dutch rainfall, and navigating Amsterdam's narrow spaces. All examples were verified by another native speaker.

Group 0015: Tagalog / Filipino (`tgl_latn`: N examples)
Manually written by a native Tagalog speaker. A separate Filipino dataset was not included, as many native speakers of Tagalog do not draw a strong distinction between the two. Examples in this dataset were written to be culturally-specific to the Philippines, covering three main topics: (1) cooking and baking, (2) crafts and construction of cultural objects, and (3) art, dances, and literature. The author cross-checked information using

websites such as Philippine Wikipedia, Philippine government blogs on culture, and informal verification from fellow native speakers living in the Philippines.

Group 0016: Vietnamese (`vie_latn`: N examples)
Manually written by a native Vietnamese speaker, and examples contain Vietnamese cultural contexts such as everyday objects, weather, clothing, routines, safety, school, simple social norms, and holidays.

Group 0017: Russian, Iraqi Arabic (Gelet) (`rus_cyrl`, `acm_arab`: N examples)
Manually written by native Russian and Iraqi Arabic (Gelet) speakers, covering everyday topics such as weather, transportation, home safety, work, hobbies, nature, sports, school, and technology. For a more culturally-specific subset, approximately 20 examples for Iraqi Arabic were translated from the Modern Standard Arabic dataset from Group 0065; a native speaker of Iraqi Arabic selected examples that were culturally relevant to their region.

Group 0018: Korean (`kor_hang`: N examples)
Manually written and verified by three native Korean speakers. Examples were written to cover popular Korean games, food, and mandatory military service.

Group 0019: Mandarin (`cmn_hans`: N examples)
Manually written by a native Mandarin speaker, covering traditional Chinese culture, food, objects, everyday life, customs, and computer use. Some examples were motivated by reading guidebooks on transportation, cooking, or safety operations. Some examples were also designed to cover recently-developed technologies from within the past five to ten years.

Group 0020: Kannada (`kan_knda`: N examples)
Manually written by a native Kannada speaker, and verified by another native speaker. Examples reflect cultural aspects of Karnataka (an Indian state where Kannada is widely spoken), as well as everyday scenarios.

Group 0021: Yoruba (`yor_latn`: N examples)
Manually written by a native Yoruba speaker, and verified by another native speaker. Examples are written to be relevant to the Yoruba land, including festivals, traditions, foods, and clothing.

Group 0022: Slovenian, Croatian, Serbian, Macedonian, Slovenian Cerkno, Chakavian (`slv_latn`, `hrv_latn`, `srp_latn`, `srp_cyrl`, `mkd_cyrl`, `slv_latn_cerk`, `ckm_latn`: N examples)
Manually written by native speakers of Slovenian, Croatian, Serbian, Macedonian, and two dialects: Slovenian Cerkno and Croatian Chakavian. Authors attempted to include culturally-relevant examples for their language(s). Examples were motivated by everyday objects, life hacks, recipes, and/or assembly manuals in each language. For each dataset, another co-author with significant understanding of the language or dialect solved the task without access to labels. Human accuracies were 97%, 100%, 97%, and 92%, excluding the two low-resource dialects. Labels were adjusted based on disagreements from this cross-check.

Group 0023: Tagalog (`tgl_latn`: N examples)
Manually written by a native Tagalog speaker, using both common spoken Tagalog (Northern and Manila dialects) and the Filipino dialect. Writing style varies between street-spoken Tagalog and formal Tagalog, and topics focus on daily life in the agricultural town of Talavera, Nueva Ecija (e.g. fishing and cooking). Some examples were inspired by Instructables posts, adapted to be culturally-relevant.

Group 0024: French (`fra_latn_fran`: N examples)

Manually written and reviewed by native French speakers, using French as spoken in mainland France. Examples were written by observing everyday actions, with distracting information added to some prompts to make the examples more challenging.

Group 0025: Polish (`pol_latn`: N examples)
Manually written and reviewed by native Polish speakers. Authors drew upon their knowledge of Polish history, culture, customs, and everyday habits.

Group 0026: Norwegian Bokmål, Norwegian Nynorsk (`nob_latn`, `nno_latn`: N examples)
Manually written in Norwegian Bokmål by native Norwegian speakers, including examples covering local foods, activities, traditions, folklore, and indigenous culture. Text embedding similarity search and then manual verification were used to ensure that examples were not direct translations of English PIQA. Examples were translated into Norwegian Nynorsk using the Nynorsk dictionary from LEXIN OsloMet, and checked by a Norwegian speaker who used Norwegian Nynorsk in school.

Group 0027: Malay (`zsm_latn`: N examples)
Manually written by a native Malay speaker, using Standard Malay (Bahasa Melayu). Examples were designed to cover local commonsense, social norms, food and drink, religious life, and everyday routines. Examples were written with natural Malay phrasing and colloquial register where appropriate.

Group 0028: Faroese (`fao_latn`: N examples)
Manually written and reviewed by native Faroese speakers. Approximately 35 examples were written to be specific to the Faroe Islands, focusing on Faroese food preparation and preservation techniques, weather patterns, traditional clothing, wool and knitting, and geography.

Group 0029: Urdu (`urd_arab`: N examples)
This dataset was written by native Urdu speakers, using Gemini 2.5 Flash and Claude Sonnet 4 for example clarification and refinement. Local websites such as UrduPoint were used to motivate examples, and examples were designed to reflect everyday life in Pakistan, including Pakistani food preparation, household practices, social customs, and traditional crafts. The dataset is written in Standard Pakistani Urdu, with every example checked by at least two native speakers.

Group 0030: Uzbek (`uzn_latn`: N examples)
Manually written by a native Northern Uzbek speaker, drawing from real-life experiences and commonly-used expressions in Uzbek. Colloquial phrases are used where appropriate. The dataset is written using Latin script, although Cyrillic script is also widely used in Uzbekistan.

Group 0031: Icelandic (`isl_latn`: N examples)
Manually written by native Icelandic speakers, covering culturally-specific topics such as food and cooking, holidays and traditions, civics and culture, folklore, geography, history, and agriculture. Some examples were inspired by browsing the Icelandic science web (`https://www.visindavefur.is/`).

Group 0032: Bengali (`ben_beng`: N examples)
Manually written by a native Bengali speaker, with culturally grounded examples reflecting daily life in Bangladesh and West Bengal, India. Examples were written to reflect everyday topics such as household chores, seasonal weather, agriculture, cooking, storage, and material interactions.

Group 0033: Tunisian Arabic (`aeb_arab`: N examples)
This dataset was created using a mix of manual writing and LLM generation, with all examples verified by two native speakers of Tunisian Arabic. The examples are written to reflect everday life in Tunisia, including

cooking practices, traditional music and instruments, household activities, local customs, and everyday objects. Because Tunisian Arabic is primarily a spoken dialect with no standardized orthography, some linguistic variation may appear across examples.

Group 0034: Marathi (`mar_deva`: N examples)
Manually written by native Marathi speakers, using Marathi as spoken in Pune City, Maharashtra, India (i.e. Puneri dialect). Examples were written to cover culturally-specific everyday topics such as education and exams, cooking and household activities, sports and games, and shopping and technology.

Group 0035: Japanese (`jpn_jpan`: N examples)
One subset of this dataset was created by native Japanese speakers using ChatGPT to translate English PIQA examples and to replace lexical elements with Japanese-specific counterparts. Another subset prompted ChatGPT to generate novel Japanese examples that required knowledge of Japanese cultural norms and conventions. Of the translated subset, 35 out of 145 passed quality checks by the native speakers, and of the novel generations, 66 out of 300 generated examples passed quality checks. All examples were verified by two native Japanese speakers.

Group 0036: Italian (`ita_latn`: N examples)
Manually written by native Italian speakers, covering household, cuisine, and entertainment domains, focusing on everyday scenarios reflecting local Italian practices. All examples were validated for fluency, correctness, and adherence to the task description by another native speaker.

Group 0037: Indonesian (`ind_latn`: N examples)
Manually written and verified by native Indonesian speakers, with examples motivated by the authors' general knowledge, past experiences, and daily life activities. By design, some prompts incorporated culturally specific Indonesian elements, such as food and traditional musical instruments. All examples were checked by at least two native speakers.

Group 0038: Vietnamese (`vie_latn`: N examples)
Manually written and verified by native Vietnamese speakers, highlighting both Kinh Vietnamese culture and minority ethnic culture (e.g. from the 50+ ethnic minority groups in present-day Vietnam). Examples cover culturally-specific knowledge such as cooking and farming methods, folklore, traditions, well-known cultural events, and minority ethnic culture. All examples were checked by at least two native speakers.

Group 0039: Korean (`kor_hang`: N examples)
Korean questions were collected from Naver Knowledge iN1, a popular Korean Q&A platform, covering diverse everyday scenarios where Korean users seek practical advice on physical tasks and problem-solving. Qwen3-4B, Qwen3-32B, and HCX-14B were used to identify PIQA-style questions, keeping only questions where all three models unanimously agreed that the question fit the task description (less than 1% of the originally collected examples). Then, GPT-4o was used to refine questions and generate incorrect solutions. Two native Korean speakers independently validated each question, improving question clarity, calibrating difficulty levels, and verifying cultural appropriateness. KoSentenceBERT was used to removed near-duplicate questions. Of the final dataset, approximately 85 questions contain elements specific to Korean culture such as traditional foods and cooking methods, clothing care, housing systems, specialized appliances, and cultural practices.

Group 0040: Urdu (`urd_arab`, `urd_latn`: N examples)
Manually written by a native Urdu speaker using Latin script, in line with the way many Pakistanis communicate on social media platforms. Examples were transliterated into Urdu script using Gemini 2.5 Flash and then manually verified.

Group 0041: Hebrew (`heb_hebr`: N examples)
Manually written by native Hebrew speakers, with each example verified by another native speaker. Approximately 55 examples cover everyday Israeli life or Jewish religious practices, including recipes, household cleaning techniques, cultural traditions, and religious customs. For some examples, motivation for topics came from Wikipedia articles or from lists of everyday objects obtained by prompting LLMs.

Group 0042: Catalan, Peninsular Spanish (`cat_latn`, `spa_latn_spai`: N examples)
Manually written in Catalan by a native Catalan and Spanish speaker, covering everyday topics such as clothing, festivity, folklore, food, literature, music, and sports. Many examples include concepts and situations that are specific to Catalan-speaking communities, and some examples do not translate well into other languages. The Catalan dataset underwent human evaluation by three native speakers, who achieved accuracies of 94%, 95%, and 98% respectively; examples were then adjusted based on this cross-checking. The dataset was translated into Spanish using Google Translate, then human verified, keeping examples for Spanish only if they remained valid after translation.

Group 0043: Polish (`pol_latn`: N examples)
Manually written by a native Polish speaker based on physics topics, including fundamental laws of physics, material properties, and principles governing interactions between materials. Online materials describing at-home basic experiments were used to motivate some examples, and several Polish-specific words (e.g. cooking and food items) were used.

Group 0045: Belarusian (`bel_cyrl`: N examples)
Manually written in conversational Belarusian by native Belarusian speakers, inspired by household situations, local customs, and guides on Belarusian life. LLMs were then used for paraphrasing, lengthening examples, and normalizing style, and then all examples were checked again by two native speakers.

Group 0046: Swedish (`swe_latn`: N examples)
Manually written by a native Swedish speaker, and checked by another native speaker. Roughly half of examples include Swedish slang, traditions, or foods, or hard-to-translate Swedish words.

Group 0047: Bulgarian (`bul_cyrl`: N examples)
Manually written by a native Bulgarian speaker, and checked by another native speaker. Examples are designed to test specific types of physical commonsense reasoning, with distractors (incorrect solutions) that are still semantically related to the prompts. Examples are interwoven with Bulgarian cultural elements and require knowledge of Bulgarian morphological cues (e.g. word inflections).

Group 0048: Mandarin, Cantonese (`cmn_hans`, `yue_hant`: N examples)
Manually written and reviewed by native Mandarin and Cantonese speakers, based on online encyclopedias and guidebooks in Mandarin and Cantonese. Example domains include activities (e.g. sports), food, geography, and art.

Group 0049: Yoruba, Igbo, Naija (Nigerian Pidgin), Hausa, Isoko, Urhobo, Idoma (`yor_latn`, `ibo_latn`, `pcm_latn`, `hau_latn`, `iso_latn`, `urh_latn`, `idu_latn`: N examples)
Manually written by native speakers of Yoruba, Hausa, Igbo, Idoma, Urhobo, Naija (Nigerian Pidgin English), and Isoko, as part of a community effort by the Linguistics Island community of linguists. Examples cover specific linguistic structures, and topics include food, culture, education, and technology.

Group 0050: Bengali, Mandarin, Greek, Korean, Turkish (`ben_beng`, `cmn_hans`, `cmn_hant`, `ell_grek`, `kor_hang`, `tur_latn`: N examples)

Manually written by native speakers of Bengali, Mandarin (Taiwanese using traditional characters, mainland using simplified characters), Greek, Korean, and Turkish. All examples were checked by another native speaker of the language. Many examples were written by first thinking of a culturally-specific item, then brainstorming physical properties of that item that could be incorporated into a PIQA-style example.

Group 0051: Uyghur (`uig_arab`: N examples)
Manually written by a native speaker of Uyghur, with each example proofread by five native speakers and using a Uyghur spell-checker. Examples were inspired by Uyghur literary materials, including cultural and traditional texts, proverbs and sayings, folklore collections, and instructional manuals.

Group 0052: Urdu (`urd_arab`: N examples)
Manually written by a native speaker of Urdu, covering domains such as cooking, religion, weather, science, and household activities. Examples were designed to cover regional cuisine, local household items, and local daily practices. LLMs were used to brainstorm ideas, but not to generate final examples.

Group 0053: Bengali (`ben_latn`: N examples)
Manually written by a native Bengali speaker using "Banglish", or Bengali language written in Latin script, often used by Bengali speakers in online settings and informal communication. Examples cover culturally-specific topics such as Bengali religious festivals and practices, traditional foods and cooking, household objects and tools, traditional games and activities, seasonal practices and nature, and folk traditions and customs. ChatGPT was used to brainstorm additional cultural topics, but not to generate examples.

Group 0055: Estonian (`est_latn`: N examples)
Manually written by native Estonian speakers, covering culturally relevant elements such as traditional Estonian foods, local materials, and region-specific practices. Inspiration for some examples was drawn from the "Maybe I'm Lucky" feature of Sõnaveeb, the language portal maintained by the Institute of the Estonian Language, generating randomly-selected Estonian words. Examples were each tested on six randomly-selected LLMs, and examples that all models got correct were dropped or edited. For human evaluation, another native speaker achieved an accuracy of 95%; examples were then adjusted based on this cross-checking.

Group 0056: Dutch (`nld_latn`: N examples)
Manually written by a native Dutch speaker, and reviewed by another native speaker. It includes culturally-relevant topics such as chocolate sprinkles on bread, ice skating, dikes, local sports, and specific dishes. LLMs, including GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4, were used in drafting samples, suggesting topics, and proofreading, but overall, their performance was found to be severely lacking in understanding the task and generating suitable examples.

Group 0057: Estonian, Persian (Farsi), Swedish (`est_latn`, `pes_arab`, `swe_latn`: N examples)
The Estonian part of this dataset was manually written by a native Estonian speaker, and reviewed by another native speaker. Topics include Estonian food, companies, places, cultural events and holidays, and typical activities and phenomena during different seasons of the year. The Farsi part of this dataset was manually written and reviewed by native Farsi speakers, covering six thematic categories: cooking and food, housekeeping and cleaning, daily life and social customs, driving and travel, health and safety, and life hacks and tools. The dataset emphasizes cultural and contextual knowledge, and inspiration was drawn from online articles in Farsi. The Swedish part of this dataset was manually written by a native Swedish speaker, and reviewed by another native speaker, drawing inspiration from online sources that cover everyday physical activities (e.g. sports, gardening, household life, traditional festivities, and traffic-related scenarios).

Group 0058: Hindi, Sindhi, Punjabi, Manipuri, Bengali, Gujarati, Marathi, Nepali, Bhojpuri, Marwari, Dhundhari, Nagamese (`hin_deva`, `snd_deva`, `pan_guru`, `mni_beng`, `bho_deva`, `guj_gujr`, `mar_deva`, `npi_deva`, `ben_beng`, `rwr_deva`, `dhd_deva`, `nag_latn`: N examples)
Examples in this dataset were primarily adapted from reasoning textbooks in English and Hindi that are widely used for preparation for competitive exams. Examples were written to reflect India-specific cultural contexts. Each example was manually or semi-automatically (i.e. machine-translated with human verification) translated into the 12 target languages, with careful preservation of meaning, cultural familiarity, and syntactic naturalness. All examples were independently labeled by two native speakers to ensure validity.

Group 0059: Lingala (`lin_latn`: N examples)
Manually written by a native Lingala speaker, covering culturally-specific everyday contexts and daily life.

Group 0060: Greek (`ell_grek`: N examples)
Manually written and reviewed by native Greek speakers. Some prompts are adapted from a variety of online material, including government and non-governmental organization (NGO) publications, academic theses, course presentations, commercial product brochures, and Wikipedia. Approximately 40% of the final examples are annotated by the authors as culturally specific.

Group 0061: Sindhi (`snd_arab`: N examples)
Manually written by a native Sindhi speaker, using Standard Sindhi (Vicholi Sindhi) in the Perso-Arabic script. Examples are culturally grounded in folklore, history, literature, foods, festivals, traditions, and everyday life in Sindh, Pakistan.

Group 0062: Swahili, Dhuluo, Lingala (`swh_latn`, `luo_latn`, `lin_latn`: N examples)
The dataset was manually written and reviewed by native speakers of Swahili, Dholuo, and Lingala, covering topics such as food, agriculture, transportation, and household practices. The Swahili examples are split between Kenyan and Tanzanian Swahili; these two varieties are structurally similar, but Tanzanian contributions emphasize domestic and rural practices, while Kenyan contributions highlight more urban contexts. The Lingala examples focus on rural life in Central Africa, including cassava preparation, termite cooking, fishing, river transport, market trading, and home construction.

Group 0063: Albanian (`als_latn`: N examples)
Manually written by a linguist specializing in Albanian and a native speaker of Albanian. Topics cover domains such as cooking, cleaning, object construction, Albanian traditional activities (e.g. music, dances, weddings), cultural practices, and agricultural tasks. The authors note that both dataset creators primarily reside outside the main Albanian-speaking continuum, potentially affecting the representativeness of the selected topics.

Group 0064: Indonesian (`ind_latn`: N examples)
This dataset was created by native Indonesian speakers using GPT-4o with careful prompting to generate culturally-specific examples. Topics include agriculture, art, daily activities, family relationships, fisheries and trade, food, religious holidays, traditional games, and wedding traditions. Examples were filtered for fluency, correctness, and adherence to the task format, and SentenceBERT was used to filter out near-duplicate examples. All examples were reviewed and edited by two native Indonesian speakers, using Standard Indonesian (Bahasa Indonesia). The filtering stages (including filtering for ambiguous solutions) resulted in removing 85.4% of the original LLM-generated examples.

Group 0065: Modern Standard Arabic, Syrian Arabic, Emirati Arabic, Tunisian Arabic, Algerian Arabic, Moroccan Arabic, Egyptian Arabic, Palestinian Arabic (`arb_arab`, `apc_arab_syri`, `afb_arab`, `aeb_arab`, `arq_arab`, `ary_arab`, `arz_arab`, `apc_arab_pale`: N examples)

Manually written by native speakers of eight Arabic dialects (including Modern Standard Arabic). Examples were written by all of the authors to be balanced across locales, and the resulting dataset was translated into each Arabic dialect by the respective native speaker. Domains covered include household, clothing, cooking, hospitality, events, and religion.

Group 0066: Galician (`glg_latn`: N examples)
Manually written and reviewed by native Galician speakers. Approximately half of the dataset covers Galician traditions and seasonal festivities, local customs and folklore, or traditional instruments. Galician websites (e.g. Galician Wikipedia, or local websites) were used to motivate some examples, but none of the content on these sites was used directly.

Group 0067: Malayalam (`mal_mlym`: N examples)
Manually written and reviewed by native Malayalam speakers from different regions of Kerala: one from Muvattupuzha (Idukki and Kottayam dialects), and one from Ottappalam (Palakkad and Thrissur dialects). Examples were written to cover topics specific to Kerala, such as local weather, traditional food recipes, regional flora and fauna, cultural flair, and religious traditions.

Group 0068: Persian (Farsi) (`pes_arab`: N examples)
This dataset was created by native Farsi speakers using a hybrid LLM and manual approach. LLMs were prompted to propose high-level categories and illustrative examples, spanning both everyday knowledge and culturally-specific practices. Based on these examples, the authors either created new samples from scratch inspired by the proposed categories or edited the LLM-generated examples. All examples were reviewed and edited by two native speakers.

Group 0069: Hindi, Telugu (`hin_deva`, `tel_telu`: N examples)
This dataset was created by native Hindi and Telugu speakers, using a hybrid LLM and manual approach. First, native speakers wrote a small set of seed examples which were used to prompt Gemini to expand the dataset. Each generated example was reviewed and edited by native speakers. The Hindi portion of the dataset uses Standard Hindi, which is widely understood across Northern India, with many prompts inspired by cultural practices such as food preparation, household activities, and regional crafts. The Telugu portion is based on Standard Telugu, spoken in Telangana and Andhra Pradesh, and it reflects daily life in those regions, from traditional agricultural practices to the handling of clay utensils.

Group 0070: Yemeni Arabic, Egyptian Arabic, Tunisian Arabic, Saudi Arabic, Jordanian Arabic, Lebanese Arabic (`acq_arab`, `arz_arab`, `aeb_arab`, `ars_arab`, `apc_arab_jord`, `apc_arab_leba`: N examples)
Manually written by native speakers of six Arabic dialects. Examples cover culturally-specific topics such as food, locations, religion, art, games, cultural items, and clothing.

Group 0071: Gujarati (`guj_gujr`: N examples)
This dataset was created by a native Gujarati speaker, using a hybrid LLM and manual approach. ChatGPT was prompted to generate examples, and a native Gujarati speaker manually filtered and edited all examples. Topics include household activities, local festivals, food, school settings, kitchen tools, farm life, animals, seasons, games, common objects, and geography, all reflective of Gujarati customs and environments.

Group 0072: Norwegian Bokmål (`nob_latn`: N examples)
Manually written by a native Norwegian speaker, using Norwegian Bokmål. The dataset covers Norwegian-specific activities, such as the preparation of traditional food dishes and the use of traditional objects.

Group 0073: Nepali (`npi_deva`: N examples)

Manually written and reviewed by native speakers of Nepali, based on topics including household tasks, personal care, outdoor activities, crafts, sports, and recreational pursuits. Another split of this dataset was generated with LLMs and human-verified, but only the human-written examples are included in Global PIQA.

Group 0074: Tamil (`tam_taml`: N examples)
Manually written by native Tamil speakers, focusing on Tamil cooking, including traditional Indian food preparation, ingredients, and terminology.

Group 0075: Tamil (`tam_taml`: N examples)
Manually written and reviewed by native Tamil speakers. Examples cover cultural and traditional dimensions of Sri Lankan life, including food practices, health and safety, religious traditions, rituals and customs, literature and arts, and traditional dress and identity.

Group 0076: Malayalam (`mal_mlym`: N examples)
Manually written by a native Malayalam speaker, and checked by other native speakers. Topics include local culture, cuisine, etiquette, superstitions, religion, and life hacks. Motivation for examples was often drawn from everyday objects in the author's household. Several prompts intentionally illustrate linguistic features unique to Malayalam.

Group 0077: Russian (`rus_cyrl`: N examples)
Manually written by a native Russian speaker, covering topics such as cooking, safety measures, basic physics, and basic computer use. Some questions are designed to be based on Russian culture.

Group 0078: Marathi (`mar_deva`: N examples)
This dataset was created by native Marathi speakers, using a hybrid LLM and manual approach. ChatGPT was prompted to generate examples, and native Marathi speakers manually filtered and edited all examples. Topics include household activities, local festivals, food, school settings, kitchen tools, farm life, animals, seasons, games, common objects, and geography, all reflective of Marathi customs and environments.

Group 0079: Bengali, Hindi, Kannada, Tamil, Malayalam (`ben_beng`, `hin_deva`, `kan_knda`, `tam_taml`, `mal_mlym`: N examples)
This dataset was created using LLM generation with human verification by native speakers of Bengali, Hindi, Kannada, Tamil, and Malayalam. LLMs (Gemini 2.5 Pro and Qwen 3) and translation models (MADLAD-400) were used in a multi-stage pipeline to identify topic clusters in English PIQA, to generate localized examples in English (localized to specific Indian states where the respective languages are widely spoken), to translate examples to the respective languages, then to correct any errors in the translations. After this pipeline, native speakers validated all examples.

Group 0080: Russian (`rus_cyrl`: N examples)
Examples in this dataset were generated by prompting GPT 5, GPT 4.1, and o4-mini with information from Russian school textbooks. All examples were manually edited and verified by native Russian speakers.

Group 0081: Telugu (`tel_telu`: N examples)
Manually written and reviewed by native Telugu speakers, using occasional Godavari regional slang. Topics include household activities, food preparation, natural phenomena, and cultural practices.

Group 0082: Telugu, Nepali, Hindi (`tel_telu`, `npi_deva`, `hin_deva`: N examples)
Manually written and reviewed by native Telugu, Nepali, and Hindi speakers. Embeddings of English translations were used to ensure that no examples were duplicates of English PIQA examples, and Gemini

2.5 Flash was used to verify the correctness of some examples. Posthoc, some examples were modified to incorporate more culturally-specific elements.

Group 0083: Hindi (`hin_deva`: N examples)
Manually written and reviewed by native Hindi speakers, focusing on everyday scenarios. Topics include food and cooking, household chores, health and safety, festivals and traditions, travel, technology and gadgets, environment and hygiene, personal care, and emergency situations.

Group 0085: Hindi, Kannada, Telugu, Malayalam (`hin_deva`, `kan_knda`, `tel_telu`, `mal_mlym`: N examples)
Manually written and reviewed by native speakers of Hindi, Kannada, Telugu, and Malayalam. Examples were written to be relevant to speakers of the respective language, covering topics such as food, clothing, household items, everyday life, festivals, and traditions. GPT-4 was used initially to generate examples for inspiration, but all examples in the final dataset are manually written.

Group 0086: Greek (`ell_grek`: N examples)
This dataset was manually constructed by a native Greek speaker, by navigating Greek websites on the internet, searching for sentences about a given topic, then adapting the sentences for the task. Topics include puzzles and riddles, household, cooking and recipes, driving, gardening, DIY, sports, construction, vacation, spatiotemporal orientation, and dance.

Group 0087: Turkish (`tur_latn`: N examples)
Manually written by native Turkish speakers, motivated by Turkish content such as food blogs, household advice websites, and health institution pages. All examples were manually verified by several Turkish speakers.

Group 0088: Yoruba, Nigerian Pidgin (Naijá) (`yor_latn`, `pcm_latn`: N examples)
Manually written and reviewed by native Yoruba and Nigerian Pidgin speakers. First, the authors compiled a list of everyday physical items relevant to both cultures, inspired by online videos, language dictionaries, and social media. Then, realistic scenarios were manually written for different items, and these prompts were used as the basis for examples.

Group 0089: Marwari, Marathi (`mar_deva`, `rwr_deva`: N examples)
Manually written and reviewed by native Marathi and Marwari speakers, covering culturally-specific topics such as home, cooking, farming and rural contexts, weather, and desert travel.

Group 0090: Telugu (`tel_telu`: N examples)
Manually written and reviewed by native Telugu speakers, using Kosta Andhra Telugu, a dialect spoken in coastal Andhra Pradesh, India. Examples in the dataset cover local festivals and traditional foods.

Group 0091: Tamil (`tam_taml`: N examples)
Manually written and reviewed by native Tamil speakers, after an initial attempt to use LLMs produced examples that were often generic, obvious, or culturally inaccurate. In the final dataset, all examples are either entirely manually written or substantially rewritten and refined from a primitive LLM-generated example. Culturally-specific topics include traditional rituals, literature and history, agrarian and folk wisdom, and art.

Group 0092: Bengali (`ben_beng`: N examples)
Manually written by a native Bengali speaker, and reviewed by other native speakers. The dataset uses standard colloquial Bengali as commonly spoken in Kolkata, India, and it includes references to local customs, food, holidays and traditions, and household objects.

Group 0093: Slovak, Šariš Slovak (`slk_latn`, `slk_latn_sari`: N examples)
Manually written by native speakers of Slovak and the Šariš dialect of Slovak. Examples were inspired by content on DIY and home improvement sites in Slovak, but no content was copied directly.

Group 0094: Assamese, Bengali, Hindi, Malayalam, Manipuri (`asm_beng`, `ben_beng`, `hin_deva`, `mai_deva`, `mal_mlym`, `mni_mtei`, `ory_orya`, `tel_telu`: N examples)
Manually written and reviewed by native speakers of Assamese, Bengali, Hindi, Malayalam, and Manipuri, covering everyday topics such as food, rituals, tools, climate, and household practices. Additional manual verification is in progress for Maithili, Orya, and Telugu datasets.

Group 0095: Italian (`ita_latn`: N examples)
Manually written and reviewed by native Italian speakers, covering culturally-specific topics such as local foods, artisanal products, domestic practices, and folklore.

Group 0096: Thai (`tha_thai`: N examples)
Manually written by a native Thai speaker. Inspired by browsing the internet in Thai, some examples cover local landmarks, art, cooking, and customs that are unique to Thailand.

Group 0097: Hindi, Marathi, Tamil (`hin_deva`, `mar_deva`, `tam_taml`: N examples)
Manually written and reviewed by native speakers of Hindi, Marathi, and Tamil, covering culturally-relevant everyday scenarios in Indic contexts, such as food preparation, household chores, and electronic device usage. Examples underwent extensive validation and rewriting, including reading examples aloud to parents, grandparents, and younger relatives.

Group 0098: Hindi (`hin_deva`: N examples)
Manually written by a native Hindi speaker, and reviewed by another native speaker. Examples were drawn from diverse domains such as traditional Indian games, handicrafts, festivals, musical instruments, and everyday life.

Group 0099: Czech (`ces_latn`: N examples)
Manually written and reviewed by native Czech speakers, covering domains such as everyday activities, cooking, household tasks, and activities related to traditional Czech customs or sayings. Some examples use Moravian and Silesian dialects, or contemporary Gen Z and Gen Alpha slang (e.g. “skibidi” and “6-7”). For examples using slang or dialects, the authors consulted external collaborators from those demographic groups to ensure correct usage. Examples were passed into GPT-5 and Claude Opus 4.1 for edits, and a small number of examples were generated directly by the LLMs themselves; all examples underwent human validation by multiple native speakers.

Group 0100: Thai (`tha_thai`: N examples)
Manually written by a native Thai speaker, using the central Thai dialect. Examples cover specific Thai knowledge, such as Muay Thai movements.

Group 0101: Sinhala (`sin_sinh`: N examples)
Manually written and reviewed by native Sinhala speakers, covering domains such as literature, religion, mythology, sports, food, and history, primarily in a Sri Lankan context.

Group 0102: Turkish, Azerbeijani, Kyrgyz (`tur_latn`, `azj_latn`, `kir_cyrl`: N examples)
This dataset was written and reviewed by native speakers of Turkish, Azerbaijani, and Kyrgyz. Topics include household routines, cooking, driving, and seasonal conditions, along with everyday and culturally-specific items. Some examples in Turkish were initially generated using GPT-5, but many Turkish examples are fully

original, and all examples were verified by native speakers. LLMs were not used for Azerbaijani or Kyrgyz; for example, for Azerbaijani, trials with GPT–5 and Gemini 2.5 Pro produced poor quality samples.

Group 0103: Tamil (`tam_taml`: N examples)
Manually written and reviewed by native speakers of Tamil, using Sri Lankan Tamil and covering domains such as domestic chores, culinary practices, agriculture, and traditional artifacts. Examples were deduplicated with n-grams and SBERT embeddings. When evaluated by humans, four native speakers agreed unanimously on the label for 95% of examples.

Group 0104: Korean (`kor_hang`: N examples)
This dataset was constructed by native Korean speakers using a hybrid LLM and manual approach. Using a multi-stage pipeline, LLMs were given Korean-specific seed scenarios to (1) generate examples, (2) validate the questions, (3) validate the solutions, (4) generate distractor solutions, and (5) validate distractors. Finally, examples were deduplicated, and biased answers (e.g. examples that could be solved with simple heuristics) were removed. All final examples were validated by a native Korean speaker.

Group 0105: Kinyarwanda (`kin_latn`: N examples)
Manually written by a native Kinyarwanda speaker, and reviewed by another native speaker, using the standard dialect spoken in education and media. Examples cover everyday scenarios such as household activities, tools and objects, food, transportation, and weather.

Group 0106: Swahili (`swh_latn`: N examples)
Manually written by a native Swahili speaker, covering a variety of everyday contexts.

Group 0107: Central Kurdish (`ckb_arab`: N examples)
Manually written by a native Kurdish speaker, using Central Kurdish (also known as Sorani). Examples focus on village life and traditional practices (e.g. cooking, handicrafts, agriculture, animal husbandry, and customs), domains where Kurdish possesses a rich and nuanced vocabulary.

Group 0108: Hungarian (`hun_latn`: N examples)
Manually written and reviewed by native Hungarian speakers, covering a variety of physical phenomena and incorporating Hungarian cultural context.

Group 0109: Turkish (`tur_latn`: N examples)
Manually written by a native Turkish speaker, with some sentences adapted from online food recipes.

Group 0110: Russian (`rus_cyrl`: N examples)
Manually written and reviewed by two native Russian speakers, covering everyday scenarios. Some examples cover culturally-specific holidays or foods.

Group 0112: Javanese (`jav_latn`: N examples)
One native Javanese speaker contracted five other annotators through Prolific at a rate of 8 GBP per hour, which is significantly above the minimum hourly wage in Indonesia. Many examples were written to be culturally-specific, covering local music, food, nature, and daily life. Generally, this dataset uses the Ngoko register, or casual language in Javanese. Although a standardized writing guideline exists for Javanese, it is not universally followed, and there is substantial variation in orthography and spelling. Annotators were allowed to write in the form they naturally used, to better capture authentic language use. The final examples were reviewed by the primary author of this dataset.

Group 0113: Georgian (`kat_geor`: N examples)

Manually written and reviewed by native Georgian speakers, covering everyday knowledge and activities. Some examples drew inspiration from the Georgian book, "Imagination and Skillful Hands" by Neli Okropiridze, which offers tips and tricks for a range of DIY projects and was once widely used in the Georgian community.

Group 0114: Burushaski (`bsk_arab`: N examples)
Manually written by a native Burushaski speaker, using the Yasin dialect. All examples were checked for grammatical correctness, cultural relevance, and physical commonsense validity.

Group 0115: Peruvian Spanish (`spa_latn_peru`: N examples)
This dataset was manually compiled by native Spanish speakers. Sentences were adapted from naturally occurring speech among the dataset authors' family and friends. Some examples were drawn from public-interest topics in Lima, Peru, including local traditions or the conduct of public officials. Any names, addresses, or direct identifiers were removed and replaced with more generic placeholders. To capture authentic language usage, tense and punctuation were not standardized but instead left reflective of colloquial speech.

Group 0116: Russian (`rus_cyrl`: N examples)
This small dataset was manually written and reviewed by native Russian speakers from the South Ural Mountains region of Russia. Several examples are designed to test local commonsense knowledge.

Group 0117: Hawaiian ('Ōlelo Hawai'i) (`haw_latn`: N examples)
Manually written by second-language 'ōlelo Hawai'i speakers, and verified by native speakers. Examples cover a wide range of scenarios, including contexts specific to Hawai'i, the Hawaiian language, and Hawaiian culture, as well as everyday situations. All Hawaiian text was written in modern orthography, including both the 'okina and kahakō. Relevant to anyone using this dataset, the dataset authors note the distinction between no'ono'o Hawai'i (Hawaiian ways of thinking) and no'ono'o Haole (foreign ways of thinking) as applied to NLP, where "data representation choices risk importing external frameworks. Preserving no'ono'o Hawai'i ensures that datasets and computational models reflect culturally grounded perspectives, maintaining authenticity and integrity in the development of Hawaiian language technologies".

Group 0118: Portuguese (European) (`por_latn_port`: N examples)
Manually written and reviewed by native European Portuguese speakers, with many examples covering Portuguese culture (e.g. references to festivities, holidays, and the preparation of traditional dishes). Two native speakers evaluated the dataset without access to labels, achieving accuracies of 90.7% and 95.4% respectively.

Group 0119: Algerian Arabic, Moroccan Arabic (`arq_arab`, `ary_arab`: N examples)
This dataset was crowdsourced from native Algerian and Moroccan (Darija) Arabic speakers. All examples were checked by other native speakers for naturalness, correctness, and cultural relevance. Contributors and annotators participated voluntarily without monetary compensation. Recruitment occurred via open community channels; participants gave informed consent, could withdraw at any time, and were not subject to coercion or undue influence. No personally identifiable information was collected. Across three annotators, average pairwise agreement on labels was over 95% (Cohen's kappa > 0.90 for all pairs).

Group 0120: Amharic (`amh_ethi`: N examples)
Approximately half of this dataset was manually written by a native Amharic speaker; the other half was generated by using Gemini 2.5 to expand the size of the dataset. All examples were then verified by multiple native speakers. Examples focus on the topics of sports, culture, history, politics, and education.

Group 0121: German (`deu_latn`: N examples)
Manually written by a native German speaker, covering culturally-specific topics such as food and customs that might not be well known outside of Germany. ChatGPT was used to help double-check grammar and spelling, but not to generate examples.

Group 0122: German (`deu_latn`: N examples)
Manually written by a native German speaker, covering topics such as sports, household, gardening, and entertainment.

Group 0123: English (USA and UK) (`eng_latn`: N examples)
This dataset was obtained by filtering the English PIQA test set to approximately 100 high-quality examples. Examples were excluded if they contained typos or nonsensical answer choices; some examples were modified to correct these errors. Many examples were selected based on cultural relevance to English-speaking contexts in the United States of America or the United Kingdom (e.g. US Thanksgiving, or American football). The resulting dataset was validated by another native English speaker.

Group 0124: Amharic (`amh_ethi`: N examples)
Manually written by a native Amharic speaker, and validated by other native speakers. Examples cover everyday contexts in Ethiopian society, including traditions, customs, food, history, and proverbs.

Group 0125: Bambara (`bam_latn`: N examples)
This dataset was compiled by native Bambara speakers. Some examples were based on content from French quizzes on technical knowledge, translated into Bambara by professional translators. Other examples were written to be culturally-specific to Bambara-speaking contexts. All examples were refined and validated by native Bambara speakers.

Group 0126: Peninsular Spanish (`spa_latn_spai`: N examples)
Manually written by a native Spanish speaker, using central-northern Peninsular Spanish (e.g. as spoken in Madrid and the interior of Castilla y León). Examples cover culturally-specific foods, customs, and domestic practices.

Group 0127: Eastern Armenian (`hye_armn`: N examples)
Manually written by an Armenian speaker, and checked by a native speaker. Prompts were first outlined in English then translated to Eastern Armenian. Topics include cutlery and tableware, fabrics and clothing, laundry, and cooking. A small number of examples are specific to Armenian culture.

Group 0128: Lithuanian (`lit_latn`: N examples)
Manually written by a native speaker of Lithuanian, with examples constructed using a mix of domain expertise and simple Lithuania-related questions. GPT-5 was used to brainstorm ideas, but not to generate examples.

Group 0129: Lithuanian (`lit_latn`: N examples)
Examples in this dataset were generated based on Wikipedia articles using GPT-5, then manually rephrased and checked by two native speakers of Lithuanian. Topics include traditional Lithuanian food, traditions, places, and literature.

Group 0130: Zulu (`zul_latn`: N examples)
Manually written by a native speaker of isiZulu, with examples written to reflect everyday scenarios and local cultural practices.

Group 0131: Kazakh (`kaz_cyrl`: N examples)
Manually written by a native speaker of Kazakh, using the Northeastern Kazakh dialect, and including some specific words that are commonly used in Karaganda city. Examples cover culturally-specific topics, including food, drinks, music, customs, animals, games, history, architecture and monuments, weather, nature, clothing, and jewelry.

Group 0132: Bosnian (`bos_latn`: N examples)
Manually written by a native Bosnian speaker, using the Ijekavian standard. The dataset covers regionally salient vocabulary and scenarios, including cooking, household tasks, nature, and religious and social customs.

Group 0133: Kinyarwanda (`kin_latn`: N examples)
Manually written and reviewed by native Kinyarwanda speakers. Examples cover everyday domains such as everyday objects, weather, folklore, and literature. The dataset is written in standard Kinyarwanda, without dialectal variations such as those spoken in the northern and southern provinces of Rwanda.

Group 0134: Peninsular Spanish, Mexican Spanish (`spa_latn_spai`, `spa_latn_mexi`: N examples)
Manually written by native Spanish speakers, covering a variety of subtypes of physical commonsense reasoning. Examples reference local foods, places, traditions, architecture, and everyday objects and tasks in Spain and Mexico (for Peninsular and Mexican Spanish respectively). The Peninsular and Mexican Spanish datasets differ at the topic, lexical, and syntactic levels, to reflect differences between the two dialects. All examples in the two datasets were verified and edited by a native Spanish speaker living in Spain or Mexico respectively.

Group 0135: Ekpeye (`ekp_latn`: N examples)
Manually written by a native Ekpeye speaker, with topics covering everyday life, local Nigerian foods, and local customs.