

INCLUDE-2.0: Disentangling Knowledge Gaps in Multilingual LLM Evaluation

Angelika Romanou^{1*}, María Grandury^{1*}, Clara Meister¹, Negar Foroutan¹,
Jabez Magomere², Anna Sotnikova¹, Orestis Papakyriakopoulos³, Scott A. Hale²,
Shamsuddeen Hassan Muhammad⁴, Antoine Bosselut¹

¹EPFL, ²University of Oxford, ³Technical University of Munich, ⁴Imperial College London

Akshay Kumar Dubey, Alfonso Amayuelas, Amit Agarwal, Arion Das, Carlos Arriaga, Cheng Yu, Dalia Ali,
David Quispe-Castillo, Fenal Ashokbhai Ilasariya, Francisco-Javier Rodrigo-Ginés, Gonzalo Martínez, Gouthami
Vadithya, Guido Ivetta, Hitesh Laxmichand Patel, Javier Conde, Jorge Vallego, Juan J. Calderón G., Karan Gupta, Marcos J.
Gomez, Marvin M. Agüero-Torales, Mithil Bangera, Naira Paola Arnez Jordan, Oscar M. Cumbicus-Pineda, Priyaranjan
Pattnayak, Reewos Erwin Talla Chumpitaz, Sahil Aggarwal, Sahil Raj, Shruti Jalan, Srikant Panda, Susana Zhou, Yeshil Bangera

Abstract

Current multilingual benchmarks cover often conflate language proficiency with the ability to answer questions grounded in specific regions and communities. We introduce INCLUDE-2.0, a large-scale benchmark for region-aware multilingual factual evaluation, which covers 91 languages from 86 countries (127 total language-country pairs), comprising over 220K curated samples drawn from native regional sources. Unlike existing benchmarks that often treat language as a proxy for region, INCLUDE-2.0 distinguishes between questions whose answers are independent of context, vary by country, concern widely documented facts about a place, or depend on knowledge primarily expressed through local linguistic and cultural sources. Each item is annotated with its regional knowledge type and domain, and paired with an English translation, enabling controlled analysis of the relationship between language proficiency and possession of regional knowledge. Evaluating frontier and open-weight models on INCLUDE-2.0, we find that aggregate measures of multilingual accuracy mask substantial performance differences across regional knowledge types, languages, and regions, and that frontier models generally perform worse on locally embedded knowledge than other knowledge types. INCLUDE-2.0 is available at <https://huggingface.co/datasets/epfl-nlp/include-2.0>

1 Introduction

Large language models (LLMs) are increasingly deployed across languages, countries, and cultural contexts, making it important to evaluate whether they encode the knowledge needed to serve users

*Equal contribution.

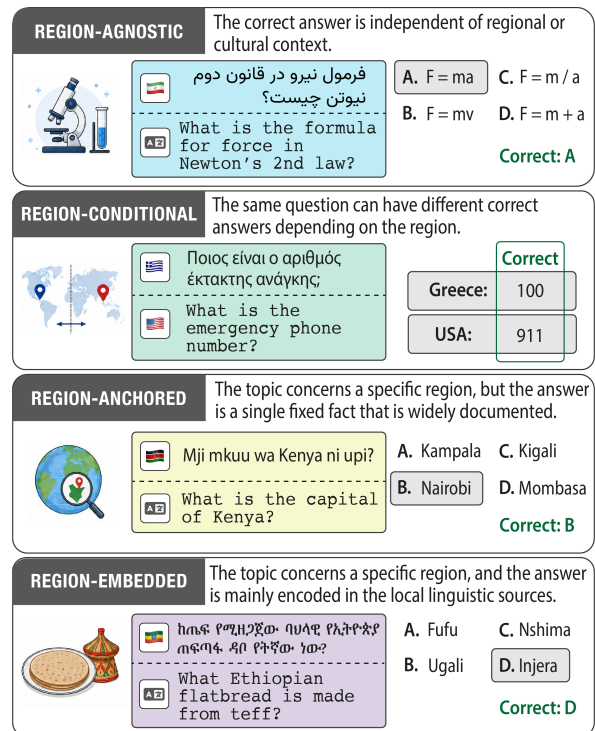


Figure 1: Regional factual knowledge taxonomy in INCLUDE-2.0. The taxonomy separates questions whose answers are region-independent, region-dependent, widely documented regional facts, or available through local linguistic and cultural sources.

in the regions where those languages are spoken. Current multilingual evaluations, however, often conflate language proficiency with regional grounding, particularly when translated from an initially-English dataset. A model may fail because it does not understand the language, because it lacks the relevant regional facts, or because it defaults to English- or Western-centric assumptions when faced with locally grounded questions. As illustrated in Figure 1, factual questions may differ

Benchmark	#Languages	#Scripts	#Countries	#Language-Country Pairs	Knowledge Type	Region Bias Analysis	Data Origin
GlobalMMLU	42	13	1	–	US-centric	✓	Human/Machine Translated
MMLU	14	8	1	–	US-centric	–	Machine-Translated
MMLU-ProX	29	11	1	–	US-centric	–	Human/Machine Translated
EXAMS	16	3	–	–	Regional & Global	–	Native
MultiLoKo	31	12	–	–	Regional	–	Native + Human/Machine Translated
BLEN	13	7	16	16	Regional	–	Native + Machine Translated
CulturalBench	1	1	45	45	Regional	–	Human + Synthetic Generation
INCLUDE	44	15	52	–	Regional & Global	–	Native
INCLUDE-2.0 (Ours)	91	22	86	127	Regional & Global	✓	Native

Table 1: Benchmark comparison. INCLUDE-2.0 offers the broadest coverage across languages, scripts, and countries, while uniquely combining regional and global knowledge, region-bias analysis, and native data collection.

substantially in how regional their answers are: some are independent of region, some require a regional context to determine the answer, some concern widely documented facts about a region, and others depend on locally embedded linguistic or cultural knowledge. Conflating these cases precludes diagnosing whether multilingual models genuinely support diverse regional contexts, or merely express globally widespread knowledge in many languages.

Recent benchmarks address this problem by moving beyond translated English datasets. Benchmarks such as INCLUDE (Romanou et al., 2024), EXAMS (Hardalov et al., 2020), BLEN (Myung et al., 2024), CulturalBench (Chiu et al., 2025), and Kaleidoscope (Salazar et al., 2025) introduce questions grounded in local educational, cultural, and regional contexts. However, important gaps remain. First, language coverage is still narrow relative to the number of languages present in modern pretraining corpora. As shown in Table 1, prior multilingual factual and regional benchmarks cover at most 44 languages, a small fraction of the 1000+ languages represented in modern pretraining corpora (Penedo et al., 2025). Second, many evaluations treat language as a proxy for region, even though the same language may be used across countries with different laws, institutions, political & social systems, and cultural practices. Third, regional knowledge is often treated as a single umbrella category, mixing questions that test general language competence, globally documented facts about a region, and knowledge that is deeply embedded in local context. As a result, existing evaluations conflate different failure modes, such as language understanding, factual recall, or region-specific grounding.

To address these limitations, we introduce INCLUDE-2.0, a multilingual benchmark for region-aware knowledge evaluation. INCLUDE-2.0

builds on INCLUDE (Romanou et al., 2024), but substantially extends it in both coverage and diagnostic structure, expanding coverage from 44 to 91 languages, covering 127 language-country pairs. Beyond scale, INCLUDE-2.0 introduces the regional taxonomy shown in Figure 1: *region-agnostic*, *region-conditional*, *region-anchored*, and *region-embedded*, disentangling linguistic competence from regional factual grounding.

We summarize our contributions as threefold:

- **Benchmark resource.** We introduce INCLUDE-2.0, covering 91 languages, 22 scripts, 86 countries, 127 language-country pairs, and over 220,000 samples. INCLUDE-2.0 specifically includes the largest factual knowledge collection for African languages to date, with 32 African languages, and one of the most diverse Indian language coverages, with 14 languages. Each item is annotated with its regional knowledge category, and all items are paired with English translations for controlled analysis.
- **Conceptual framework.** We propose a taxonomy of regional knowledge that operationalizes regionality and enables multilingual performance to be decomposed into language ability and regional knowledge.
- **Analyses.** Using INCLUDE-2.0, we analyze current LLMs’ understanding of multilingual factual knowledge, whether multilingual competence improves with model scale, how language ability interacts with regional knowledge, and whether models exhibit regional biases within pluricentric languages.

2 Operationalizing Regional Knowledge

In multilingual evaluation, *regional knowledge* refers to information, practices, conventions, and norms whose validity, prevalence, or recognition

depends on a specific geographic or cultural context (Romanou et al., 2024). Capturing this dimension is essential for benchmarks that are designed to measure not only linguistic competence but also a model’s grounding in the linguistic contexts where it is deployed. Without it, evaluation conflates language proficiency with the Western-centric defaults that dominate multilingual evaluation benchmarks. We operationalize regionality along the *type* of knowledge a question tests, and the *availability* of that knowledge across linguistic communities.

2.1 Regional Knowledge Types

We define four categories based on the relationship between a question and the region it concerns.

Region-agnostic. The correct answer is independent of any regional or cultural context. Examples include pure mathematics, physical laws, formal logic, and biochemical mechanisms. Cross-lingual performance gaps on such knowledge are therefore more likely to reflect language-related factors rather than gaps in region-specific information. This property makes region-agnostic knowledge a useful control axis for disentangling linguistic effects from regional or cultural factors.

Region-conditional. The same question can have different correct answers depending on the region. This category includes region-specific rules and institutions, such as speed limits, voting ages, prescription policies, civil procedure, and professional licensing requirements, as well as topics whose accepted answers may vary across national curricula, public institutions, or local conventions, such as historical interpretations, religious or political positions, and disputed events. In all cases, the answer is fixed within a given regional context but may change across regions. This is especially important for languages spoken across multiple countries (e.g., Spanish, Arabic, English, French, Portuguese), where the language alone is not sufficient to determine the correct answer.

Region-anchored. The topic of a query concerns a specific region, but the answer is a fixed globally documented fact. Examples include major historical events, internationally known figures, or large-scale demographic information. Such items test whether a model has acquired information *about* a region, even though the information is typically accessible through widely indexed sources, including English-language reference materials.

Region-embedded. The topic concerns a specific region and has a fixed answer, but the underlying knowledge is primarily encoded within a particular linguistic or cultural community and is sparsely represented elsewhere. Examples include traditional musical instruments, vernacular cuisine, folk practices, local historical figures, and regionally specialized terminology. Unlike region-anchored knowledge, region-embedded knowledge is typically learnable only through native-language or specialized regional sources.

2.2 Regional Knowledge Availability

Availability captures the extent to which a region-coupled fact is represented across cross-lingual discourse. We model this as a continuum operationalized by a heuristic question used during annotation: *Is this fact reliably documented in major English-language reference sources, or does retrieval typically require native-language or specialized regional sources?* This dimension primarily distinguishes region-anchored from region-embedded knowledge and has direct implications for multilingual pretraining as low-availability knowledge is less likely to be acquired from English-centric training data alone.

3 The INCLUDE-2.0 Benchmark

INCLUDE-2.0 is a multilingual benchmark covering 91 languages across 86 countries (127 language–country pairs), spanning 22 scripts and consisting of over 220,000 curated data samples. All questions in the dataset are presented in their native languages and scripts, and have a 4-option multiple-choice question (MCQ) answering format. Additionally, we provide English translations to facilitate parallel-data analysis. INCLUDE-2.0 covers 29 knowledge domains, as summarized in Table 5. The benchmark also covers a wide breadth of linguistic diversity: Ten languages appear across multiple countries, and ten countries contribute data in more than one language (Table 3 in the Appendix). This section details the data collection procedure for INCLUDE-2.0, alongside the subsequent data annotation phase that enables fine-grained model evaluation.

3.1 Data Collection

To construct INCLUDE-2.0, we collect sources of multiple-choice examinations in collaboration with native speakers and regional associations. Following a similar protocol to Romanou et al. (2024),

we collect *Academic Exams* spanning diverse subjects (e.g., Humanities & STEM) and educational levels (e.g., middle & high school & university), *Professional Certifications & Licenses*, and *Regional Licenses* such as exams administered by regional authorities to assess specific qualifications (e.g., driving, marine licenses). For high-resource languages where suitable language-specific public benchmarks already exist, we incorporate these resources directly. However, new unpublished data still account for $\sim 78\%$ of the full benchmark. During data collection, we design INCLUDE-2.0 to assess multilingual capabilities that extend beyond academic exams to include region-specific knowledge following the taxonomy introduced in Section 2. All collected data undergo native-speaker quality control to verify extraction accuracy, remove context-dependent items that cannot be represented textually, add missing reference context when needed, and assign exam-level metadata. Further details on the data collection process and contributor compensation are provided in Appendix A.1.

3.2 Data Annotation

To enable fine-grained and stratified evaluation, we augment all samples in INCLUDE-2.0 with regional metadata. Specifically, we use GPT-5.5 to annotate each question according to the regional knowledge categories introduced in Section 2.1 and the knowledge-availability framework described in Section 2.2. To assess annotation reliability, we conduct a sampled human evaluation in which annotators independently perform the same labeling task. We then measure agreement between human annotations and the LLM-generated labels, and find that agreement rates between humans and LLMs yielded raw exact agreement of 0.760 and an unweighted Cohen’s κ of 0.676, indicating substantial agreement between humans and LLM raters. Further details of the annotation protocol and agreement analysis are provided in Appendix A.1.

We also translate every multiple-choice question, including both the question text and answer options, into English using GPT-5.5. These translations allow us to disentangle errors caused by limited access to the underlying knowledge from errors arising from a model’s difficulty processing or expressing that knowledge in a particular language. We therefore use English translations as a paired reference for each original-language item, enabling controlled within-sample comparisons of

model performance across languages.

4 Experimental Setup

We benchmark INCLUDE-2.0 on commercial models: GPT-5.5, GPT-5.4-mini (Singh et al., 2025a), Claude Opus 4.7,¹ Gemini-3.5 Flash,² and on leading open-weight model families: Llama 3.1 (Grattafiori et al., 2024), Qwen 3 (Yang et al., 2025), Qwen 3.5, Gemma 3 (Gemma Team, 2025), Gemma 4,³ Mistral (Jiang et al., 2023), Apertus (Apertus et al., 2025), Aya Expanse (Dang et al., 2024), and Tiny Aya (Salamanca et al., 2026). For open-weight and commercial models on multiple-choice benchmarks, we prompt the model to output the answer letter (detailed prompt in the Appendix A.2.2).⁴ We run Claude Opus 4.7 with greedy decoding and GPT-5 variants with reasoning effort set to low and API-default sampling.

5 Results & Analysis

We use INCLUDE-2.0 to evaluate current LLMs along four axes: overall performance across languages and regional knowledge types (§5.1), whether failures reflect language ability or missing regional knowledge (§5.2), the effect of model scale (§5.3), and whether a single aggregate score represents performance at the language and country level (§5.4). As a metric, we report in-language accuracy (i.e., accuracy when a question is presented in its original, native language) macro-averaged across languages unless otherwise noted.

¹Claude Opus 4.7 system card

²Gemini-3.5 Flash model card

³Gemma 4 Model

⁴Using log-probabilities to score answer options has been an alternative standard for MCQA evaluations historically (Gao et al., 2024), but as log-probabilities are unavailable for several commercial models in our evaluation, and can introduce systematic biases tied to tokenization and length normalization (Biderman et al., 2024), we decide instead to use the model’s ability to generate the correct answer in a specified format as the evaluation target. We acknowledge this choice may conflate the evaluation of regional knowledge recall with fine-grained instruction-following learned during post-training, but the latter is also often a minimal requirement for modern LLMs to be practically useful.

Model	Region-agnostic		Region-conditional		Region-anchored		Region-embedded		All Question Types	
	In-Language	English	In-Language	English	In-Language	English	In-Language	English	In-Language	English
<i>Commercial Frontier Models</i>										
Gemini 3.5 Flash	81.1	75.5	77.7	69.7	86.4	81.7	75.2	63.3	79.7	71.2
GPT-5.5	78.7	76.0	76.2	71.6	86.1	82.5	72.5	63.4	77.5	71.3
GPT-5.4-mini	72.1	73.2	64.8	64.4	79.4	80.4	62.5	55.7	70.1	66.8
Claude Opus 4.7	77.0	76.0	75.9	67.7	83.3	82.2	67.0	60.8	74.5	70.2
<i>Open-Weight Models</i>										
Apertus-70B	41.4	50.6	39.5	41.2	45.1	55.5	32.2	32.8	38.5	43.9
Apertus-8B	39.7	48.9	36.8	41.6	43.6	54.6	32.5	33.3	38.0	42.6
Aya Expanse-32B	50.0	63.0	48.8	52.0	57.2	70.0	38.7	42.2	48.0	55.1
Aya Expanse-8B	39.8	56.3	38.2	46.1	42.8	59.3	35.2	38.8	39.4	49.3
Gemma 3-4B	49.6	58.8	43.9	47.3	50.5	59.7	36.1	39.6	45.1	50.8
Gemma 3-12B	60.3	66.5	53.7	53.0	63.3	71.2	43.3	44.2	55.6	57.9
Gemma 3-27B	65.2	69.2	57.2	57.7	68.5	73.2	46.9	45.5	59.9	60.7
Gemma 4-31B	70.9	72.1	66.5	65.2	80.0	77.9	57.0	51.0	67.3	64.7
Mistral-7B-instruct	36.4	53.6	37.8	47.3	39.6	57.4	32.4	38.3	35.7	47.8
Qwen 3-8B	60.1	69.7	52.5	55.9	61.6	69.7	38.8	42.0	54.5	59.3
Qwen 3-14B	62.5	71.2	54.4	56.4	64.0	73.9	41.9	43.9	57.3	61.0
Qwen 3-32B	64.3	71.6	54.6	58.2	62.7	72.6	43.1	45.1	58.6	62.0
Qwen 3.5-27B	64.0	68.8	54.3	56.1	66.9	73.3	43.0	41.6	58.3	59.1
Tiny Aya Earth	38.0	45.8	34.8	38.8	35.4	47.0	31.7	34.1	35.4	40.7
Tiny Aya Fire	39.6	49.6	40.8	40.9	38.5	49.5	32.7	35.9	37.0	43.5
Tiny Aya Global	36.0	46.3	33.2	40.8	35.9	46.1	30.1	33.4	34.9	40.8
Tiny Aya Water	39.2	48.5	36.3	42.3	36.7	47.8	31.8	34.8	36.6	42.6
Llama 3.1-70B_instruct	56.7	66.8	55.1	58.8	65.4	73.0	45.3	45.6	54.5	59.2
Llama 3.1-8B_instruct	37.5	53.8	36.9	43.6	38.4	57.6	29.5	35.4	34.7	47.0

Table 2: Average performance (%) across all languages in INCLUDE-2.0, both overall and stratified by different regional categories. Quantities are computed as macro-averages across languages. Specific model names/variants can be found in Table 6. In the *In-Language* setting, questions are presented in their native language; in the *English* setting, they are translated into English. Bold values refer to the best *API* and *open-weight* models per regional category.

5.1 Performance summary

Q. Which models perform best across languages and regional knowledge types?

Finding. All four frontier models have higher average in-language accuracy than every open-weight model. The best open-weight model, Gemma 4-31B, is 12.4 points below the best frontier model. Every model is less accurate on Region-embedded than on Region-anchored items, by 14.5 points on average for the frontier models and 14.4 for the open-weight models.

Overall performance. Frontier models have the highest average in-language accuracy on INCLUDE-2.0 (Table 2). The best open-weight model, Gemma 4-31B, is 12.4 points below the best frontier model. Interesting, frontier models tend to have higher in-language accuracy compared to English translation accuracy, while the open-weight models tend to show the reverse phenomenon (with the notable exception of Gemma 4-31B).

Performance by regional knowledge type. Accuracy is not uniform across regional categories (Table 2). For frontier models, in-language accuracy drops by 14.5 points on average from Region-anchored to Region-embedded items. For open-weight models the average drop is 14.4 points. The size of this drop differs by model, with more than

20 points for some open-weight systems (Table 2). The drop has the same direction for frontier and open-weight models, which shows that high accuracy on widely documented regional facts does not imply comparable accuracy on knowledge concentrated in local sources. We explore this gap in §5.2 by testing whether Region-embedded items are answered less accurately because the knowledge is harder to apply or because it is absent from the model.

5.2 Language versus knowledge

Q. When a model answers incorrectly, is the cause weak language ability or missing regional knowledge?

Finding. Both factors contribute and comparing performance across different prompt languages helps distinguish them. Some regional facts are missed by the models in both the in-language and English setting (termed “knowledge-dark” items) and concentrate in specific countries. For questions answered correctly in only one setting, frontier and open-weight models differ in which setting they perform better.

A model can fail to answer a question for many different reasons that are conflated when observing aggregate accuracy alone. For example, an answer may be more indexable from a model’s parametric knowledge through a native-language prompt, or

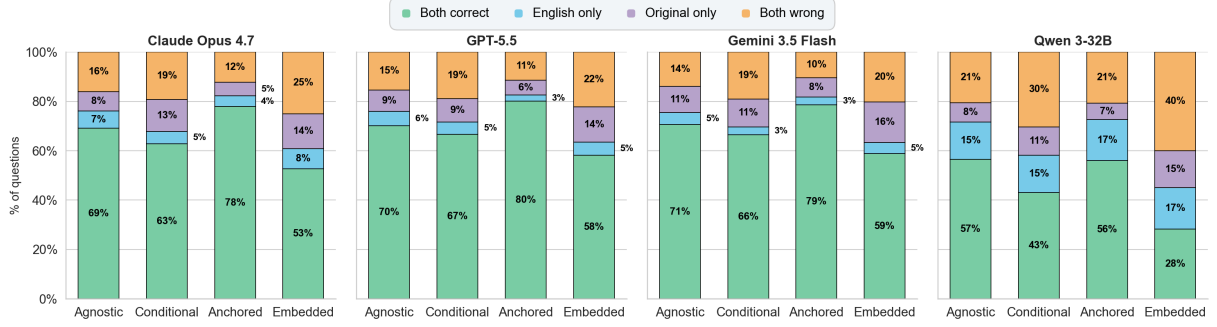


Figure 2: Language-balanced four-outcome decomposition across regional categories. For each model, language, and category, item outcomes are partitioned into both correct, English prompt correct only, Original language prompt correct only, and both wrong, then macro-averaged over languages. Region-embedded items show the largest shift from both-correct outcomes toward both-wrong outcomes, while the disagreement cells motivate the cross-lingual accessibility analysis.

it may be accessible in English but lost when processing in the original language due to the model’s struggles with basic native language understanding. To disentangle these effects, we use the parallel English-translated benchmark vs. the original in-language dataset to decompose answer outcomes into four cases: both correct, English-only correct, in-language-only correct, and both wrong. Figure 2 visualizes this decomposition across regional categories, macro-averaged over languages. The both-wrong cases identify regional facts that appear unretrievable under either language, while the disagreement cases distinguish whether knowledge is more accessible in the original language or only in English, (one of) the model’s predominant training languages.

Knowledge-dark regional items. To identify cases of completely missing regional⁵ knowledge, we define an item as *knowledge-dark* when the 23-model population performs at or below chance in both English and in-language comprehension. Requiring failure both in English and in-language controls for any cross-lingual accessibility issues (a topic we explore next). Items that meet this criterion most plausibly reflect knowledge that is absent from the current model population rather than merely inaccessible. Knowledge-dark items concentrate in particular countries rather than spreading evenly (Table 13). Across regional-native items the overall dark fraction is 13.6%, while a few countries exceed 30%. Among countries with large item

counts, the highest fractions are Ghana and Nigeria. While high fractions are also reported for Qatar, Israel, and Hungary, but they are computed on far smaller samples ($n \leq 259$) and thus have large uncertainty. The concentration in specific countries indicates that failures on regional knowledge are localized.

Cross-lingual accessibility. We also measure cross-lingual accessibility, where a model answers correctly in only one language but not another. Such cases suggest that access to the relevant knowledge may depend on the language of the question. Under this hypothesis, the in-language-only cases—where a model answers correctly in the original language but not in English—correspond to facts that are more easily accessible using the in-language formulation of the question. This pattern is consistent with a model’s recall being more successful when the query language matches the language most relevant to the fact, though other factors, such as translation quality, could also produce this pattern. The English-only cases—where the model answers correctly in English but not in the original language—correspond to facts accessible in English but not in the original language. These cases suggest either weaker linguistic abilities or accessibility issues in the original language. Our analyses of such cases are designed to separate these two possible explanations. Concretely, we summarize the English-only direction with $\Delta = D_{\text{spec}} - D_{\text{agn}}$, corresponding to the English-to-in-language accuracy drop on regional items (D_{spec}) minus the same drop on region-agnostic items (D_{agn}). Subtracting D_{agn} removes the part of the drop attributable to general difficulty with the

⁵In this study, we constrain our *regional* definition to the Region-anchored and Region-embedded categories, excluding Region-conditional items because failures on them may reflect regional conflation or defaulting to the wrong jurisdiction, rather than lack of access to a fixed regional fact.

original language, so a positive Δ isolates regional knowledge the model can produce in English but not in the original language. Appendix B.1 gives the full definition and the difference-in-differences argument for how Δ controls for general language ability.

On region-specific items, accuracy is higher in the original language than in English for all four frontier models, by 4.3 to 9.7 points, and higher in English than in the original language for eighteen of the twenty open-weight models, by up to 11.0 points (Table 14). Among region-specific items answered correctly in exactly one of the two languages, the share correct in the original language only is 61% to 73% across the frontier models and 36% to 61% across the open-weight models. Therefore, on the same region-specific items, the direction of the language advantage varies across models, potentially reflecting differences in multilingual training data. We perform a statistical significance test that supports this result in Section B.1.

To investigate the English-only cases, we compare their prevalence on region-specific and region-agnostic items. D_{spec} , the share of region-specific items answered correctly in English but incorrectly in the original language, runs from 9.3% to 57.5% across the 24 models. After subtracting the corresponding share on region-agnostic items, the remaining difference is at most 5.8 percentage points and is positive for 16 of the 20 open-weight models. The similar prevalence of English-only cases on regional and region-agnostic items suggests that general language difficulty plays a substantial role, with a smaller additional gap on regional items.

5.3 Model scale

Q. Do larger open-weight models show a smaller gap between English and original-language accuracy?

Finding. Larger open-weight models have higher accuracy on Region-agnostic items, but there is no reliable evidence that the English and in-language gap narrows when scaling up a model.

In this experiment, we examine whether increased parametric capacity correlates with increased multilingual competence. We restrict the analysis to region-agnostic items, where performance differences are less likely to reflect missing regional knowledge, and to open-weight models (frontier systems are excluded, since they report no public parameter counts). To perform a regression analysis, we model answer correctness as a

function of parameter count,⁶ evaluation condition (English vs. in-language), and an interaction term between these two variables. The interaction term captures whether larger models reduce the English vs. in-language gap (a positive coefficient means the gap narrows as size increases). We provide details of this formulation in Appendix B and show the learned coefficients in Table 11, where β_1 , β_2 and β_3 correspond to parameter count, the evaluation condition, and the interaction term.

Overall, scale improves average accuracy, but we find no reliable evidence that it reduces the English vs. in-language gap. The positive value of the *model size coefficient* β_1 shows that larger open-weight models have higher accuracy on region-agnostic items, but the *in-language coefficient* β_2 is negative (with 0 outside the credible interval⁷), meaning that performance on in-language questions is lower than on English questions at the average model size. The latter result aligns with the results in the previous section about English vs. in-language performance (§5.1). Finally, the *size $\times$ in-language interaction* term β_3 is positive but small and uncertain, with 0 inside the credible intervals for both formulations of the regression model. The positive point estimate of the β_3 coefficient is consistent with a slight narrowing of the English–in-language gap as size increases, but its credible interval includes zero, indicating this result is not statistically significant.

5.4 Aggregate versus local performance

Q. Do global model rankings generalize across languages and countries?

Finding. No. The best model overall (Gemini 3.5 Flash) achieves the highest in-language accuracy in 55 of 91 languages, but is outperformed in the remaining 36, with statistically significant differences in 8.

Global models may not be the best local models. Benchmark leaderboards usually report a single global model ranking. We test how well that ranking represents per-language performance by comparing the global leaderboard, ranked by each model’s macro-average in-language accuracy, against per-language leaderboards. We summarize the agreement between the global and per-language rankings with the Spearman rank correlation, com-

⁶We take the log of parameter count and center this variable for numeric stability.

⁷A credible interval is the Bayesian analogue of a confidence interval; we report credible intervals because the model is fit in a Bayesian framework (ADVI).

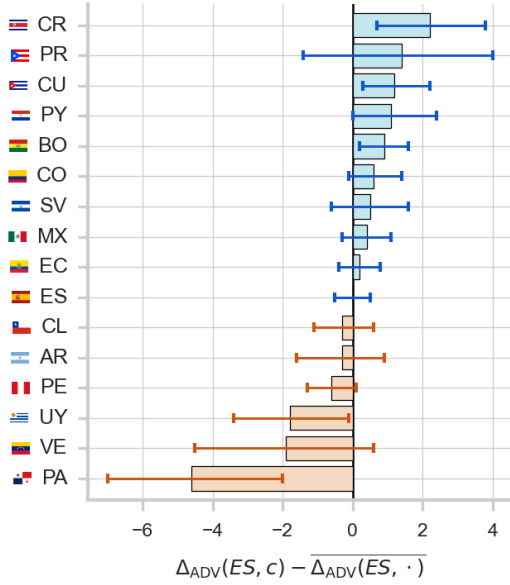


Figure 3: Per-region residual of the in-language advantage $\Delta_{\text{ADV}}(\text{ES}, c) - \Delta_{\text{ADV}}(\text{ES}, \cdot)$ across Spanish-speaking countries, macro-averaged over all 22 evaluated models with bootstrap 95% CIs. Positive values indicate that using Spanish-language questions helps more for that country than for all Spanish-speaking regions on average.

puted per language and then averaged. Separately, we count the languages whose top-ranked model differs from the global top-ranked model.⁸

Per-language rankings correlate strongly with the global ranking in both settings (Appendix Table 8), so the relative order of models is largely preserved across languages. The top-ranked model is less stable. In the in-language setting the global winner, Gemini 3.5 Flash, is not the per-language winner for 36 of 91 languages. In the English setting, the global winner, GPT-5.5, changes for 63 of 91 languages. In-language rankings correlate slightly less with the global ranking than English rankings do, indicating more movement below the top rank under in-language evaluation. The five languages with the weakest correlation are Oromo, Tangale, Maithili, Jju, and Ewe (Appendix Table 9). A global score can therefore name the wrong best model for a given language, which supports reporting per-language results with the global ranking.

Variation within pluricentric languages. We ask whether the benefit of asking questions in-language is uniform across countries that share

⁸To limit the effect of sampling noise, we resample items within each language and count a language as changed only when its best local model has higher accuracy than the global winner in a significant fraction of resamples.

the same language. For each *language–country* (ℓ, c) pair, we compute the in-language advantage $\Delta_{\text{ADV}} = \text{Acc}_{\text{IL}} - \text{Acc}_{\text{EN}}$, where positive values indicate higher accuracy in the original language than in English. We then compare each pair’s advantage to the average advantage for countries sharing the same language, yielding a gap residual: $\Delta_{\text{ADV}}(\ell, c) - \overline{\Delta_{\text{ADV}}}(\ell, \cdot)$. A positive gap residual means that in-language querying helps more for that country than for the language on average, while a negative value means it helps less. We restrict the analysis to region-specific items and estimate uncertainty by bootstrapping items within each *language–country* pair.

Using Spanish as a case study, Figure 3 shows that the in-language advantage varies across Spanish-speaking countries even with the language fixed. For Costa Rica, Cuba, and Bolivia, the in-language advantage is larger than the Spanish average by a margin whose 95% interval excludes zero. For Panama and Uruguay, it is smaller than the average by such a margin. The residuals for Puerto Rico, Paraguay, and Venezuela include zero within their 95% intervals, so they are not distinguishable from the Spanish average. Because the gap residual is defined from each country’s in-language minus English accuracy, it reflects the size of the in-language advantage rather than absolute in-language accuracy. Appendix Table 15 reports the full set of pluricentric languages, including Arabic, Chinese, and Portuguese. Countries with fewer than 30 region-specific items are excluded, which is why only four pluricentric languages appear. These results indicate that the benefit of querying the model in-language for regional knowledge need not be uniform within a shared language, and that an aggregated language-level score does not capture within-language regional variation.

6 Related work

Early Large Language Model (LLM) evaluation relied on English-centric benchmarks such as HelLaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and BigBench (Srivastava et al., 2023). These benchmarks became widely used as tools for evaluating and improving LLM pre-training, but they were all English-only benchmarks, limiting their practical utility for multilingual LLM development.

Meanwhile, early multilingual benchmarks were

typically restricted to basic linguistic and natural language understanding tasks (Conneau et al., 2018; Ponti et al., 2020; Ruder et al., 2021; Liang et al., 2020; Asai et al., 2024). While these datasets enabled multilingual evaluation, they were initially designed to test transfer from high-resource to low-resource languages in finetuning contexts, rather than to understand the complex knowledge and reasoning capabilities of LLMs across languages.

To bridge the multilingual evaluation gap, recent efforts shifted to translating resources representing complex knowledge from English sources, such as MMMLU (Hendrycks et al., 2021), Global MMLU (Singh et al., 2025b), and PangeaBench (Yue et al., 2025). However, these works evaluated multilingual systems on knowledge from Western ecosystems, as reflected in these originally-English benchmarks. Concurrently, other works filled this gap by benchmarking LLMs on crowd-sourced datasets that evaluated the cultural nuances, societal norms, and values linked to local ecosystems (Myung et al., 2024; Shi et al., 2024; Magomere et al., 2025; Rao et al., 2025). Finally, benchmarks like EXAMS (Hardalov et al., 2020), INCLUDE (Romanou et al., 2024), and Kaleidoscope (Salazar et al., 2025) sourced data from localized academic and professional examinations, explicitly introducing the notion of regionality by evaluating models using resources from specific local environments. INCLUDE-2.0 expands regional evaluation to 91 languages across 86 countries, providing the broadest coverage yet that enables a detailed analysis of distinct regional knowledge types without conflating geography with culture.

7 Conclusion

We introduced INCLUDE-2.0, a benchmark for evaluating multilingual knowledge across 91 languages, 86 countries and 127 language-country pairs. INCLUDE-2.0 comprises native-source questions, sample-level annotations, and English translations to distinguish failures in language ability from gaps in regional knowledge. Our results show important variation among categories of regionality, as frontier models struggle most with locally embedded knowledge, and larger open-weight models do not reliably become equally effective across English and original-language prompts. These findings motivate evaluation that measures multilingual coverage and also whether models can access and apply knowledge in the contexts where their lan-

guages are used.

Limitations

While INCLUDE-2.0 substantially expands multilingual and regional evaluation coverage, several limitations remain.

Taxonomy. First, our distinction between *region-anchored* and *region-embedded* knowledge is operationalized through an annotation heuristic on cross-lingual availability, *i.e.*, whether a fact is reliably documented in globally distributed reference sources, or primarily encoded in native-language and specialized regional materials. Consequently, the *region-embedded* category is definitional as items are labeled region-embedded because their supporting evidence is believed by LLM/human annotators to be concentrated outside the English-centric web. An independent quantitative proxy for availability would let future work calibrate this gap against an external signal rather than the annotation heuristic itself, and we leave this calibration to subsequent work.

Benchmark. Second, our benchmark adopts a multiple-choice question (MCQ) format exclusively. While MCQs enable scalable evaluation across a large number of languages and domains, they do not fully capture open-ended reasoning, explanation quality, or generative capabilities. Models may sometimes succeed through elimination strategies or pattern matching without demonstrating robust understanding of the underlying regional knowledge. As a result, strong performance on INCLUDE-2.0 should not be interpreted as a complete measure of grounded regional reasoning or multilingual generation ability. We leave the development of open-ended and generative regional knowledge evaluation settings to future work.

Some stages of the benchmark construction pipeline rely on LLM-based processing, including metadata annotation and translation. Although we conduct sampled human validation for the annotation process, systematic biases or inaccuracies introduced by the annotating model may still persist, particularly for low-resource languages or culturally nuanced content. These biases may influence both the categorization of regional knowledge and the downstream analyses derived from those annotations. Future work could reduce these risks through larger-scale expert annotation and multi-annotator validation pipelines across diverse lin-

guistic communities. Finally, all English translations in the benchmark are generated using GPT-5.5 without full human validation. Consequently, translation errors or subtle semantic shifts may still affect these analyses, particularly for culturally nuanced or region-specific concepts.

A discussion of the potential risks of INCLUDE-2.0 can be found in the Appendix C.

Acknowledgments

We thank Aditi Khandelwal, Deniz Bayazit and Badr Alkhamissi for reading and providing comments on drafts of this paper. We are grateful to the Masakhane and SomosNLP communities for connecting us with data contributors. We also thank the contributors to the first version of INCLUDE for their foundational work on the benchmark. We gratefully acknowledge the support of the Swiss National Science Foundation (No. 215390), the AI2050 program at Schmidt Sciences (Grant #G-25-69783), Sony Group Corporation, and the Swiss National Supercomputing Center (CSCS) in the form of an infrastructure engineering and development project. This research is with support from Google.org and the Google Cloud Research Credits program for the Gemini Academic Program. This project is also funded by the European Union (ERC, RESPECT-LM, 101222478). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project IDs a139 and a145 on Alps.

References

- Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Durech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, and 84 others. 2025. [Apertus: Democratizing open and compliant llms for global language environments](#). *Preprint*, arXiv:2509.14233.
- Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. [BUFFET: Benchmarking large language models for few-shot cross-lingual transfer](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.
- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sid Black, Jordan Clive, Anthony DiPofi, Julien Etzaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, and 10 others. 2024. [Lessons from the trenches on reproducible evaluation of language models](#). *ArXiv*, abs/2405.14782.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *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.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Alexis Conneau, Rutu Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [A framework for few-shot language model evaluation](#).

- Gemma Team. 2025. [Gemma 3 Technical Report](#). *arXiv preprint arXiv:2503.19786*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#).
- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. [EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations (ICLR)*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. [XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.
- Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Yuehgo Foutse, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Samantha Dalal, Beatrice Marchegiani, Elizaveta Semenova, Lauren Crais, and Siobhan Mackenzie Hall. 2025. [The world wide recipe: A community-centred framework for fine-grained data collection and regional bias operationalisation](#). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’25*, page 246–282, New York, NY, USA. Association for Computing Machinery.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, V  ctor Guti  rrez-Basulto, Yazm  n Ib    ez Garc  a, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. [Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.
- Guilherme Penedo, Hynek Kydl    ek, Vinko Sabol  ec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all—adapting pre-training data processing to every language](#). *arXiv preprint arXiv:2506.20920*.
- Edoardo Maria Ponti, Goran Glava  , Olga Majewska, Qianchu Liu, Ivan Vuli  , and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. [Normad: A framework for measuring the cultural adaptability of large language models](#). In *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 2373–2403.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Sneha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. [Include: Evaluating multilingual language understanding with regional knowledge](#). *arXiv preprint arXiv:2411.19799*.
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. [XTREME-R: Towards more challenging and nuanced multilingual evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alejandro R. Salamanca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajae, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. [Tiny aya: Bridging scale and multilingual depth](#). *Preprint*, arXiv:2603.11510.
- Israfel Salazar, Manuel Fern  ndez Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh,

Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, and 1 others. 2025. Kaleidoscope: In-language exams for massively multilingual vision evaluation. *arXiv preprint arXiv:2504.07072*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rog rio Abreu De Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025a. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Shivalika Singh, Angelika Romanou, Cl mentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025b. [Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). In *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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri  Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#).

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal llm for 39 languages. In *International Conference on Learning Representations*, volume 2025, pages 47758–47811.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Annotation of INCLUDE-2.0 metadata

A.1.1 Data collection

All collected data of INCLUDE-2.0 undergo quality control by native speakers. After automatic exam extraction, annotators review each question to verify that the question text and answer options were correctly extracted from the source documents. When extraction errors are identified, annotators manually correct the content using the original materials as reference. Annotators also filter out samples that depend on images or tables, and ensure that questions requiring additional context, such as reading-comprehension items, include the relevant reference text in the question field. Finally, they assign exam-level metadata to each sample, including the question language, exam topic (in both English and the original language), the academic level when applicable, and country of origin.

The data collection effort targets pluricentric languages and plurilingual countries. Table 3 highlights Spanish (20 countries) and Arabic (8) among languages, and India (14 languages) and Nigeria (13) among countries. Full counts are in Table 7.

Language	# Countries	Country	# Languages
Spanish	20	India	14
Arabic	8	Nigeria	13
French	2	Ghana	6
Portuguese	2	Ethiopia	4
Bengali	2	Kenya	3
Chinese	2	Mozambique	3
Dutch	2	Spain	3
Igbo	2	Morocco	2
Swahili	2	Pakistan	2
Urdu	2	Paraguay	2

Table 3: Languages spanning multiple countries (left) and countries with multiple languages (right).

A.1.2 Human annotator compensation

Contributors of newly collected data (*i.e.*, data not already public) choose between two forms of compensation: monetary payment or paper co-authorship. Monetary compensation depends on the number of question–answer (QA) pairs delivered, with bonuses for parsing effort and domain diversity. Table 4 summarizes the scheme; all amounts are in USD.

Base rate. We group contributions into four tiers by QA-pair count. The per-pair rate rises from \$0.50 in Tier 1 (300–500 pairs) to \$0.70 in Tier 4

Tier	#QA pairs	Base (\$/pair)	Parsing bonus	Diversity multiplier
1	300–500	0.50	0.05	+5 / +10%
2	501–800	0.60	0.05	+5 / +10%
3	801–1,200	0.65	0.05	+5 / +10%
4	1,201–2,000	0.70	0.05	+5 / +10%
5	>2,000	+\$50 per extra 1K		

Table 4: Compensation scheme for native-speaker contributors, in USD.

(1,201–2,000 pairs), so larger contributions earn more per pair.

Parsing bonus. Source exams that require manual parsing before extraction earn an extra \$0.05 per QA pair, added to the tier base rate.

Diversity multiplier. To encourage coverage across knowledge domains (Table 5), we add 5% to the total payment for contributions spanning up to two domains and 10% for more than two.

Large contributions. Beyond 2,000 QA pairs, the per-pair scheme is replaced by a flat bonus of \$50 for every additional 1,000 pairs, added on top of the Tier-4 payment for the first 2,000.

For a contribution of $N \leq 2,000$ pairs spanning D domains, the total payment is $P = N \cdot (r + b) \cdot m$, where r is the tier base rate, $b = \$0.05$ the parsing bonus (0 if not applicable), and m the diversity multiplier (1.05 for $D \leq 2$, 1.10 for $D > 2$). For $N > 2,000$, we add \$50 per additional 1,000 pairs.

A.1.3 LLM annotation

For each question in INCLUDE-2.0, we use GPT-5.5 for two annotation steps. First, we translate the original-language question and answer options into English using a structured prompt designed for multiple-choice assessment items (Figure 5). The prompt instructs the model to preserve the item’s meaning, difficulty, answer-choice order, labels, formatting, and any explicit answer markers, and to return a constrained JSON output.

Second, we classify each item into four regional-knowledge categories: region-agnostic, region-conditional, region-anchored, or region-embedded. We use a separate annotation prompt that defines the categories, specifies an ordered decision procedure, and asks the model to classify the cognitive target of the item rather than relying on surface cues such as the source language or country (Figure 6). This annotation pipeline provides both an English version of each item for cross-lingual evaluation and metadata describing the type of regional knowledge required to answer it.

Domain	Count
Language and Linguistics	28,000
History	20,936
Biology	17,067
Law	15,027
Mathematics	14,065
Medical Sciences	13,326
Environmental and Earth Sciences	11,952
Physics	10,062
Literature	9,953
Chemistry	7,731
Political Science	7,547
Religious Studies	7,418
Economics	7,327
Vocational and Licensing Exams	7,081
Business and Management	6,546
Anthropology	5,732
Philosophy	5,076
General Knowledge	4,797
Computer Science	4,363
Sociology	4,144
Psychology	3,200
Education	2,921
Engineering	2,379
Performing Arts	982
Visual Arts	921
Allied Health	538
Technical and Trade Education	405
Geography	133
Journalism, Media Studies, and Communication	47

Table 5: Unique domains and their distribution in INCLUDE-2.0.

A.1.4 Reliability of the LLM regional annotations

To validate the regional labels, we run a sampled human study in which two AI expert annotators independently repeat the labeling task, and we measure agreement between the annotators and between the annotators and the LLM.

Sampling protocol. We draw a stratified sample of 25 items per category (100 in total), allowing at most one item per language within each category to spread coverage across the benchmark. Annotators see the English translation, so the task is language-independent. Items are shuffled and interleaved across categories, and each annotator works from a separate sheet with the LLM label hidden, keeping the two annotations independent and blind. Both annotators receive the same category definitions and item information (question, options, and country) given to the LLM.

Inter-annotator agreement. The four categories are nominal, so we report nominal-label agreement metrics: raw exact agreement, unweighted Cohen’s κ , and Krippendorff’s α with a nominal distance function. Agreement between the two annotators is

Model ID	Size (B)
CohereLabs/tiny-aya-earth	3.35
CohereLabs/tiny-aya-fire	3.35
CohereLabs/tiny-aya-global	3.35
CohereLabs/tiny-aya-water	3.35
CohereLabs/aya-expanse-8b	8.03
CohereLabs/aya-expanse-32b	32.30
google/gemma-7b-it	8.54
google/gemma-3-4b-it	4.30
google/gemma-3-12b-it	12.00
google/gemma-3-27b-it	27.43
google/gemma-4-31B-it	32.68
meta-llama/Llama-3.1-8B-Instruct	8.03
meta-llama/Llama-3.1-70B-Instruct	70.55
mistralai/Mistral-7B-Instruct-v0.3	7.25
Qwen/Qwen3-8B	8.19
Qwen/Qwen3-14B	14.77
Qwen/Qwen3-32B	32.76
Qwen/Qwen3.5-27B	27.78
swiss-ai/Apertus-8B-Instruct-2509	8.05
swiss-ai/Apertus-70B-Instruct-2509	70.60

Table 6: Models evaluated on the INCLUDE-2.0 benchmark, sorted by model family and parameter count.

high, with a raw exact agreement of 0.760, an unweighted Cohen’s κ of 0.676, and a Krippendorff’s α of 0.677. These values indicate that the category definitions are applied consistently.

Human-LLM agreement. We then compare the LLM labels with the human annotations. Averaged over the two annotators, the LLM matches the human label with a raw exact agreement of 0.795 and an unweighted Cohen’s κ of 0.727. Treating the two annotators and the LLM as three raters, Krippendorff’s α is 0.710. On the 76 items where both annotators agree, the LLM matches their consensus label 0.895 of the time. Overall, the LLM labels align with human judgments, supporting their use as sample-level annotations in our analyses.

A.2 Inference details

A.2.1 Models

Table 6 lists the 19 instruction-tuned checkpoints evaluated, covering 4B to 70B dense parameters. All checkpoints are public on the Hugging Face Hub and are loaded by their repository ID at the current revisions as of 2026-05-23.

A.2.2 Prompt format

All models are evaluated zero-shot in a chat-template setting using the following system instruction:

Inference prompt for multiple-choice answer extraction

You will be given a multiple-choice question from {country}. Before responding, analyze the question and answer options. You can do any reasoning; however, you MUST put your answer inside the answer tag: <answer>A/B/C/D</answer>. Do not include any other text on that line.

The task instruction contains the question followed by the four options labelled A.-D. and a trailing Answer: cue.

We apply each model’s chat template via `lm_eval’s --apply_chat_template` flag. We also set `--enable-thinking=True` for the model families that support it, namely Apertus-70B, Qwen 3, Qwen 3.5, and Gemma 4.

Alternatively, using log-probabilities to score answer options has been an standard for MCQA evaluations historically (Gao et al., 2024), but as log-probabilities are unavailable for several commercial models in our evaluation, and can introduce systematic biases tied to tokenization and length normalization (Biderman et al., 2024), we decide instead to use the model’s ability to generate the correct answer in a specified format as the evaluation target. We acknowledge this choice may conflate the evaluation of regional knowledge recall with fine-grained instruction-following learned during post-training, but the latter is a minimal requirement for modern LLMs.

A.2.3 Answer extraction and scoring

Each generation is post-processed by two custom `lm_eval` filters and graded by the harness’s `exact_match` metric (case- and punctuation-insensitive). We initially scored with strict extraction (`em_hard`) but, prompted by the response-pattern analysis below, softened the parsing into a fallback scorer (`em_soft`). We calculate both:

em_hard takes the letter inside the first <answer>*X*</answer> tag with $X \in \{A, B, C, D\}$ (case insensitive). Responses without a well-formed tag score zero.

em_soft adds a fallback applied in order:

1. *Tag: Letter*, strict tag (<answer>A</answer>).
2. *Tag: Letter + Text*, the first A–D letter inside the tag (<answer>A.</answer>).
3. *No Tag: Letter*, a bare letter (“A”).

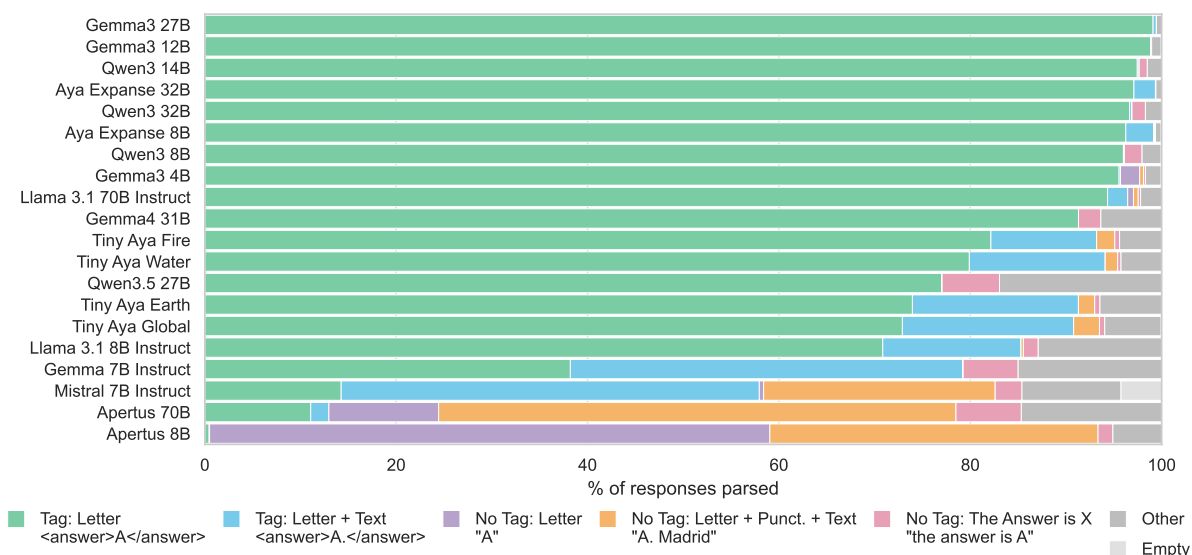


Figure 4: Response pattern distribution per model: share of responses caught at each stage of the *em_soft* fallback pipeline. Models are ordered by their strict tag share. Counts pool local-language and English questions.

4. *No Tag: Letter + Punct. + Text*, a leading letter followed by punctuation and text.
5. *No Tag: Answer is X*, explicit phrasing.
6. Responses caught by no stage fall into *Other* or *Empty* and score zero.

Appendix Figure 4 shows the share of responses caught at each stage per model.

A.2.4 Hardware and Environment

All jobs are executed on the Alps research infrastructure at CSCS on Grace-Hopper (GH200) nodes with 4 GPUs and 460 GB of host RAM per node. The evaluation of all the reported results sums up to 600 GPU hours. Each (model, variant) evaluation runs as a SLURM job using single-node tensor parallelism (tensor_parallel_size=4). For the 27–70B models the 128-subtask group is sharded across 2 or 4 independent jobs to reduce wall-time; per-task results are concatenated post-hoc. The full sweep uses vllm 0.20.0, transformers 5.8.1, and lm_eval 0.4.12.

A.2.5 Reproducibility

The full evaluation pipeline, including task definitions, prompts, answer extraction filters, and model configurations, will be released on GitHub. The per-sample model results will be released on Hugging Face.

A single (model, variant) run can be reproduced with the following command:

```
lm_eval \
  --model vllm \
  --model_args "pretrained=<HF_MODEL_ID>,\
    dtype=auto,\
    gpu_memory_utilization=0.75,\
    tensor_parallel_size=4" \
  --tasks include_cot_reasoning_en \
  --apply_chat_template \
  --system_instruction \
    "<SYSTEM_INSTRUCTION>" \
  --gen_kwargs max_gen_toks=4096 \
  --output_path \
    results/include/<MODEL>/en \
  --log_samples \
  --batch_size auto
```

Replacing include_cot_reasoning_en with include_cot_reasoning_og runs the original-language variant.

Language	Country	N	Language	Country	N	Language	Country	N
Albanian	Albania	2,317	Georgian	Georgia	546	Sena	Mozambique	488
Amharic	Ethiopia	1,864	German	Germany	1,912	Serbian	Serbia	1,412
Arabic		9,459	Greek		2,279	Sindhi	Pakistan	314
	Egypt	1,616		Cyprus	42	Sinhala	Sri Lanka	290
	Jordan	4,749		Greece	2,237	Slovak	Slovakia	130
	KSA	71	Guaraní	Paraguay	2,897	Somali	Ethiopia	264
	Kuwait	18	Gujarati	India	321	Spanish		14,672
	Morocco	193	Hausa	Nigeria	1,490		Argentina	887
	Palestine	623	Hebrew	Israel	2,282		Belice	15
	Qatar	547	Hindi	India	4,804		Bolivia	913
	UAE	107	Hungarian	Hungary	1,386		Chile	689
Armenian	Armenia	841	Ibibio	Nigeria	272		Colombia	1,342
Assamese	India	272	Idoma	Nigeria	642		Costa Rica	347
Azerbaijani	Azerbaijan	6,143	Igala	Nigeria	303		Cuba	584
Basque	Spain	588	Igbo		1,475		Ecuador	1,807
Belarusian	Belarus	489		Nigeria	1,434		El Salvador	453
Bengali		14,694		Nigeria	41		España	4,327
	Bangladesh	14,388	Indonesian	Indonesia	11,045		Jamaica	19
	India	306	Italian	Italy	4,608		México	1,090
Bulgarian	Bulgaria	2,827	Japanese	Japan	2,541		Nicaragua	20
Chinese		12,425	Jju	Nigeria	599		Panamá	48
	China	12,005	Kannada	India	541		Paraguay	214
	Hong Kong	405	Kazakh	Kazakhstan	5,591		Perú	1,213
	Singapore	15	Kinyarwanda	Rwanda	381		Puerto Rico	251
Croatian	Croatia	2,369	Korean	Korea	1,607		República Dominicana	21
Czech	Czechia	714	Lithuanian	Lithuania	682		Uruguay	263
Dagbani	Ghana	1,160	Luo	Kenya	2,194		Venezuela	169
Dangme	Ghana	569	Macedonian	North Macedonia	1,994	Swahili		491
Danish	Denmark	446	Maithili	India	245		Kenya	482
Darija	Morocco	2,956	Makhuwa	Mozambique	254		Tanzania	9
Dogri	India	263	Malay		998	Swedish	Sweden	2,494
Dutch		1,446		Malaysia	986	Tagalog	Philippines	492
	Belgium	521		Singapore	12	Tamil	India	1,215
	Netherlands	925	Malayalam	India	886	Tangale	Nigeria	273
Ekpeye	Nigeria	300	Marathi	India	304	Telugu	India	11,237
Embu	Kenya	294	Nepali	Nepal	1,316	Tigrinya	Ethiopia	1,901
Esan	Nigeria	152	Nyanja	Mozambique	997	Turkish	Turkey	2,765
Estonian	Estonia	490	Obolo	Nigeria	1,663	Twi	Ghana	1,568
Ewe	Ghana	350	Oriya	India	372	Tyap	Nigeria	1,395
Fante	Ghana	1,068	Oromo	Ethiopia	993	Ukrainian	Ukraine	1,421
Finnish	Finland	838	Persian	Iran	24,102	Urdu		623
French		1,785	Polish	Poland	1,928		India	111
	Canada	846	Portuguese		1,294		Pakistan	512
	France	939		Brazil	425	Uzbek	Uzbekistan	2,270
Fula	Nigeria	261		Portugal	869	Vietnamese	Vietnam	8,321
Ga	Ghana	1,837	Punjabi	India	742	Yoruba	Nigeria	1,346
Galician	España	350	Russian	Russia	6,241			

Table 7: Language and country distribution of INCLUDE-2.0.

Translation prompt

SYSTEM INSTRUCTION

You are a professional translator specializing in assessment items (MCQs). Translate into natural, publication-quality English while preserving difficulty, logic, and fairness. Do not explain your choices.

USER PROMPT

Source language: {}

Task: Translate the following multiple-choice question and its options into English.

Rules (strict):

- 1) Preserve meaning, tone, register, and difficulty. Do NOT add hints or explanations.
- 2) Do NOT add, remove, merge, split, or reorder choices.
- 3) Preserve any existing labels exactly (e.g., "A.", "B)", "①", "a)").
If choices have no labels, do NOT introduce labels.
- 4) Keep all names, numbers, formulas, units, URLs, emojis, hashtags, @mentions as-is unless a standard English convention clearly applies
(e.g., decimal comma → decimal point, date format to English).
- 5) Preserve structure and inline formatting: line breaks, Markdown/HTML/LaTeX, code blocks, variables like {{...}}, and tags like <...>.
Do NOT translate content inside code blocks, variables, or tags.
- 6) Translate idioms and fixed expressions idiomatically (not word-for-word).
- 7) If a phrase is genuinely ambiguous in a way that could affect the answer, choose the best rendering and append the alternative in square brackets [like this] immediately after the phrase.
- 8) Do NOT infer or reveal the correct answer. If the source includes an explicit correct-answer marker, keep it exactly as written.
- 9) Output must be valid JSON with exactly these keys and nothing else:

```
{
  "question_en": "<translated question text>",
  "choices_en": ["<choice 1>", "<choice 2>", ...]
}
```

Do not include any extra fields unless ambiguity required rule 7; if so, add
"notes": "<one concise sentence about the ambiguity>".

Input JSON:

```
{
  "question": "{}",
  "choices": [
    "{}",
    "{}",
    "{}",
    "{}"
  ]
}
```

Output JSON only, no preface, no prose.

Figure 5: Translation prompt used to create the English version of each multiple-choice item.

Regional-knowledge annotation prompt

SYSTEM INSTRUCTION

You are an expert annotator for a multilingual evaluation benchmark. Your job is to classify exam-style multiple-choice questions by the kind of regional or cultural knowledge they test.

Be precise and consistent. Follow the decision procedure exactly and apply the first matching category.

Classify the question based on its cognitive target—what knowledge or reasoning is required to answer it—not merely on surface nouns, the source region, or the language of the item. The same topic can belong to different categories depending on what the question asks.

Use the provided correct answer only to understand the intended question, not to override the classification rules.

USER PROMPT

Task

Classify the item into exactly one category: A, B, C, or D.

Categories

A – Region-agnostic

The correct answer does not depend on region, country, jurisdiction, or local cultural context.

Typical cases: mathematics; physics and chemistry laws; formal logic; anatomy and physiology; mechanism-of-action pharmacology; abstract reasoning; internationally standardized scientific facts.

Key test: Would the answer remain the same regardless of where the question is asked? If yes, and the item is not specifically about a region, classify as A.

B – Region-conditional

The correct answer depends on the region, country, jurisdiction, or institutional system the question is anchored to. Typical cases: laws and regulations; legal ages; tax thresholds; prescription rules; voting eligibility; professional licensing; school grading systems; civil procedure; speed limits; statutes of limitation.

Key test: Could the same question template plausibly be asked in multiple regions and yield different correct answers because local rules differ? If yes, classify as B.

Important: B takes priority over all other categories. A question about a specific country's law is still B if the answer is jurisdiction-dependent in principle.

C – Region-anchored factual

The question is about a specific region, country, people, or place, and the answer is unique and fixed rather than jurisdiction-variable. The relevant knowledge is typically available in globally distributed or widely accessible reference sources.

Typical cases: capitals; heads of state; major national historical events; major wars or treaties; large-scale demographics; internationally known figures; widely documented landmarks or institutions.

Key test: Is the topic region-specific, with one fixed answer that is typically findable in English-language or internationally distributed sources? If yes, classify as C.

D – Region-embedded cultural/practice knowledge

The question has a unique and fixed answer, but the knowledge is deeply embedded in a particular language community, regional practice, or local cultural tradition, and is typically sparse in globally distributed sources.

Typical cases: traditional music, rhythms, instruments, or dances; regional cuisine specifics and culinary techniques; folk customs and vernacular practices; local mythology or ritual traditions; regionally specialized terminology; local historical figures of limited international prominence; knowledge that generally requires native-language, community-specific, or specialized regional sources.

Key test: Would answering this usually require native-language materials, local cultural familiarity, or specialized regional sources rather than broadly accessible global references? If yes, classify as D.

Decision procedure

Apply these steps in order. Stop at the first match.

1. Does the correct answer depend on which region, jurisdiction, or local institutional system the question is anchored to? If yes, classify as B.
2. If not, is the question region-independent and about universal scientific, mathematical, formal, or abstract knowledge? If yes, classify as A.
3. If not, is the question about a specific region, country, place, people, or tradition, with a unique fixed answer that is typically available in globally distributed reference sources? If yes, classify as C.
4. Otherwise, classify as D.

Output format

Return valid JSON only. Do not include markdown fences, preamble, explanations outside JSON, or trailing prose.

Use exactly this schema:

```
{
  "reasoning": "<2-3 concise sentences applying the decision procedure>",
  "category": "<A | B | C | D>",
  "confidence": "<high | medium | low>",
  "notes": "<edge-case considerations, plausible alternative category, or translation concerns; use an empty string if none>"
}
```

Item to annotate

```
Source region: {source_region}
Source language: {source_language}
Question (native): {question_native}
Question (English): {question_english}
Options (native): {options_native}
Options (English): {options_english}
```

Figure 6: Prompt used to classify each item into regional-knowledge categories and assign the pluricentric flag.

B Data Analysis Setup

B.1 Multilingual Knowledge Accessibility

Intuitively, a raw drop in accuracy on regional items can confound two causes we want to separate: generally weaker processing in that language, or inaccessibility of knowledge in that language. The region-agnostic items let us quantify the first case; they require the same local-language processing as for the region-specific questions, but contain no regional content, so absence of regional knowledge is not a factor for these cases. We can therefore use this in-language–English difference in accuracy as a control. Explicitly, we use a difference-in-differences approach rather than computing raw differences: we compare the drop in in-language–English accuracy on regional items to the same drop on region-agnostic items. $\Delta = D_{\text{spec}} - D_{\text{agn}}$ is the part of the accuracy drop we thus attribute to inaccessibility of regional knowledge.

Significance of the in-language versus English asymmetry. Section 5.2 reports that the frontier models are more accurate in the original language on region-specific items and the open-weight models more accurate in English. Because per-language item counts are small, we test that asymmetry separately within each combination of one model and one language rather than only on the pooled data. For a given model and language we count the region-specific items answered correctly in the original language and incorrectly in English, and the items answered correctly in English and incorrectly in the original language; items answered correctly in both languages or in neither enter neither count and are excluded. Under the null hypothesis, an item answered correctly in exactly one of the two languages is equally likely to fall into either count, which gives an exact two-sided McNemar test. We apply it at $\alpha = 0.05$ to every combination holding at least one such item and report the two rejection directions separately in Table 14 (columns *IL larger*, *EN larger* and *Tested*).

For the four frontier models, 355 combinations can be tested and 210 reject the null. The original-language count is the larger one in 208 of those 210, spread over between 44 and 60 of the 88 or 89 languages tested for each model. For the twenty open-weight models, 1,800 combinations can be tested and 855 reject the null; the English count is the larger one in 674 of those 855. The test describes which of the two counts is larger for a

given model and language. It does not attribute that asymmetry to a particular cause, and translation quality of the English variant, pretraining mixture, prompt template sensitivity and tokenisation are all consistent with the same observation.

B.2 Scaling vs. Multilingual Competence

We provide further details on the regression models used in the analysis in Section 5.3. All regression models include a per-family random intercept and slope on parameter count to reflect that accuracy for a given family may scale at its own rate, and a per-language random intercept to capture the model’s. In our primary regression model, we also include an item-level random intercept.⁹ In words, the random effects account for the data being grouped: items belong to languages and models to families, so outcomes within a group are correlated rather than independent. A separate intercept for each family, language, and item captures the baseline differences in accuracy between groups, and the family-level slope allows the effect of size to differ from one family to the next. Without these terms, we would be breaking the independence assumptions made by regression models; the model would treat correlated observations as independent and report confidence intervals that are too narrow. Putting this together, we get the following equation:

$$\begin{aligned} \text{logit } \pi_{m,i,c} = & \beta_0 + b_{0,f(m)} + u_{\ell(i)} + v_i \\ & + (\beta_1 + b_{1,f(m)}) \log P_m \\ & + \beta_2 c + \beta_3 c \log P_m. \end{aligned} \quad (1)$$

or using lmer notation:

$$\begin{aligned} \text{correct} \sim & \log P * c + (1 + \log P \mid \text{family}) \\ & + (1 \mid \text{language}) + (1 \mid \text{item}) \end{aligned}$$

where $\pi_{m,i,c}$ is the probability that model m answers item i correctly in condition c ($c = 1$ for in-language, $c = 0$ for English; each item contributes both), and $\text{logit } p = \log \frac{p}{1-p}$. The fixed effects are the intercept β_0 , the log-size slope β_1 , the in-language offset β_2 , and the size–condition interaction β_3 ; $\log P_m$ is the centered log (active) parameter count of model m . The random effects are a correlated family intercept and slope ($b_{0,f(m)}, b_{1,f(m)}$), a language intercept $u_{\ell(i)}$, and an item intercept v_i , each Gaussian with mean zero.

⁹An aggregated binomial model over all region-agnostic items, dropping the item random intercept, yields very similar coefficients.

C Potential Risks

INCLUDE-2.0 is designed to improve regional-awareness in multilingual evaluation. The resource and its analyses still carry risks, which we flag here for responsible use.

Overclaiming regional competence.

INCLUDE-2.0 does not support comparing model knowledge across languages or countries, because the data is not parallel: each language-country pair draws on different exams, domains, and difficulty levels. We are careful to avoid such comparisons in our own framing, and we ask users of the benchmark to do the same. High accuracy on INCLUDE-2.0 also does not show that a model is safe to deploy in the regions it covers: the benchmark uses a multiple-choice format, and a model that answers items from academic, professional or licensing exams well may still produce unsafe or incorrect output in open-ended use.

Simplification of regional views. Treating one answer as correct for an entire region can present a single authority’s view as fact. Some INCLUDE-2.0 items include historical interpretations, religious or political positions, and disputed events. Scoring a model against one fixed option rewards that view and penalizes minority or contested perspectives. Used to train or select models, the benchmark could propagate these views and present contested topics as settled to end users. We frame each question as a test of whether a model matches a stated regional context, not a judgment about which view is right.

Benchmark contamination and overfitting.

Public release exposes INCLUDE-2.0 to two risks. First, its items may enter future pretraining corpora, inflating scores without real gains in regional knowledge. Second, the released per-country knowledge-dark analysis enables targeted optimization for weak regions rather than broad improvement.

Setting	Global winner	Mean ρ	Median ρ	Winner changes
Original-language	Gemini 3.5 Flash	0.859	0.952	39.6% (36/91)
English-prompt	GPT-5.5 (low)	0.888	0.970	69.2% (63/91)

Table 8: Stability of global model rankings under language-specific evaluation. ρ denotes Spearman rank correlation between the global leaderboard and each language-specific leaderboard. Winner changes indicate languages where the global winner is not the local winner after bootstrap filtering.

Language	n	Local winner	Local acc.	Global acc.	Winner changed	ρ
Oromo	993	Gemma 3-4B	26.99	23.26	Yes	0.007
Tangale	273	Mistral 7B-instruct	34.07	25.27	Yes	0.166
Maithili	245	Gemma 3-12B	47.06	37.14	Yes	0.380
Jju	599	Mistral 7B-instruct	34.39	29.55	Yes	0.390
Ewe	350	Gemini 3.5 Flash	45.14	45.14	No	0.458
Ga	1,837	Gemini 3.5 Flash	46.33	46.33	No	0.497
Dangme	569	Gemini 3.5 Flash	46.05	46.05	No	0.504
Galician	350	Gemini 3.5 Flash	31.71	31.71	No	0.556
Dagbani	1,160	Gemini 3.5 Flash	56.03	56.03	No	0.609
Idoma	642	GPT-5.5	47.04	42.68	Yes	0.626

Table 9: The ten languages for which the model ranking of performance in that language correlates least with the global model ranking. Evaluated in the in-language setting.

Language	n	Local winner	Local acc.	Global acc.	Winner changed	ρ
Kannada	541	GPT-5.5	78.37	78.19	Yes	0.988
Azerbaijani	6,143	GPT-5.5	83.93	82.99	Yes	0.986
Dutch	1,446	GPT-5.5	90.53	90.32	Yes	0.984
Telugu	11,237	Gemini 3.5 Flash	83.50	83.50	No	0.984
Turkish	2,765	GPT-5.5	94.79	94.32	Yes	0.983
Indonesian	11,045	Gemini 3.5 Flash	87.23	87.23	No	0.982
Hebrew	2,282	Gemini 3.5 Flash	79.84	79.84	No	0.981
Gujarati	321	Gemini 3.5 Flash	92.21	92.21	No	0.980
Malayalam	886	Gemini 3.5 Flash	86.79	86.79	No	0.980
Nepali	1,316	Gemini 3.5 Flash	90.50	90.50	No	0.980

Table 10: The ten languages for which the model ranking of performance in that language correlates most with the global model ranking. Evaluated in the in-language setting.

Coefficient	(A) item RE	(B) no item RE
β_0 intercept	0.168 [-0.112, 0.425]	0.182 [-0.097, 0.437]
β_1 logsize	0.430 [0.138, 0.727]	0.336 [0.062, 0.619]
β_2 in-lang	-0.371 [-0.616, -0.141]	-0.285 [-0.515, -0.064]
β_3 logsize $\times$ in-lang	0.079 [-0.214, 0.447]	0.058 [-0.272, 0.383]

Table 11: Logistic mixed-effects fit on region-agnostic items. Outcome is item correctness; logsize is centered log parameter count, $cond=1$ for in-language (vs. English). The coefficient of interest is β_3 : the change in the English-vs-in-language gap as models scale ($\beta_3 > 0$ means larger models close the gap). Both columns include a per-family random slope on logsize and a language random intercept; (A) adds an item random intercept.

Family	#	Sizes (B)	$\beta_1^{f(m)}$
Apertus	2	8, 70	0.079
Aya Expanse	2	8, 32	0.284
Gemma	6	0.27, 1, 4, 7, 12, 27	0.567
Llama	2	8, 70	0.340
Mistral	2	7, 24	0.415
Qwen	8	0.6, 1.7, 4, 7, 8, 14, 32	0.315

Table 12: Per-family scaling slopes from the mixed-effects regression. $\beta_1^{f(m)}$ is the family-specific slope for family $f(m)$ on log parameter count (β_1 plus the family random-slope deviation), reflecting how quickly accuracy on region-agnostic items grows with size within each family. The regression is fit over a broader set of checkpoints than those reported in Table 2: we include smaller variants (e.g., sub-1B Qwen and Gemma checkpoints) that are excluded from the main results table, which is restricted to models in a competitive accuracy range, but that contribute the within-family size variation needed to estimate $\beta_1^{f(m)}$. Tiny Aya is omitted from this analysis because all four checkpoints share parameter count 3.35B, so the within-family slope is not identified.

Country	#nat	Dark _{nat} %	At%	Below%	Dark _{spec} %	Δpp
Qatar	259	37.5	4.6	32.8	38.5	+1.1
Israel	117	36.8	18.8	17.9	10.3	-26.5
Ghana	5708	34.2	17.6	16.6	34.1	-0.1
Hungary	94	30.9	13.8	17.0	22.9	-8.0
Nigeria	8406	30.0	14.1	15.9	30.0	+0.0
Ethiopia	2013	25.3	9.2	16.1	25.0	-0.3
Venezuela	39	20.5	5.1	15.4	19.5	-1.0
Egypt	619	19.5	10.0	9.5	19.3	-0.3
Rwanda	67	19.4	9.0	10.4	22.1	+2.7
Belarus	404	17.8	13.9	4.0	17.9	+0.1
Paraguay	2897	17.4	11.0	6.4	16.6	-0.8
Ukraine	92	17.4	7.6	9.8	6.2	-11.2
Iran	7175	16.8	10.5	6.3	17.5	+0.7
Finland	92	16.3	8.7	7.6	16.6	+0.3
Costa Rica	80	16.2	3.8	12.5	15.1	-1.2
Azerbaijan	1145	16.2	8.0	8.1	16.4	+0.3
Kenya	2189	15.2	6.5	8.7	15.4	+0.2
Armenia	580	15.0	9.5	5.5	15.3	+0.3

Table 13: Candidate knowledge-dark regions. An item is marked *knowledge-dark* when the 23-model population performs at or below chance ($\gamma = 0.25$) in both English and in-language. We report the per-country dark fraction over regional-native (*Region-anchored* and *Region-embedded*) items, decomposed into at-chance and significantly below-chance cases, and compare it to the broader region-specific set, i.e., including *Region-conditional* items as well. Δpp denotes the change in dark fraction after adding region-conditional items. Countries are sorted by Dark_{nat}; overall dark fractions are 13.6% for regional-native and 13.4% for region-specific items.

Model	Accuracy on region-specific items (%)		Correct in one language only (%)		(model, language) combinations			Correct in the first language, wrong in the second (%)			
	IL	EN	IL only	EN only	IL larger	EN larger	Tested	IL→EN	EN→IL _{spec}	EN→IL _{agn}	Δ
Claude Opus 4.7	72.9	68.0	11.6	6.6	49	2	89	20.1	14.3	13.3	0.9
Gemini 3.5 Flash	78.9	69.2	13.6	3.8	60	0	89	20.8	9.3	10.5	-1.2
GPT-5.4-mini	67.9	63.6	12.3	8.0	44	0	89	22.8	17.5	18.4	-0.9
GPT-5.5	76.8	69.6	11.6	4.4	55	0	88	19.2	10.7	12.1	-1.5
<i>Frontier</i> (4 models)	–	–	–	–	208	2	355	–	–	–	–
Apertus-70B	38.3	40.8	14.6	17.2	14	38	90	43.8	47.1	46.4	0.7
Apertus-70B-think	39.6	39.7	16.8	16.9	15	23	91	47.4	48.1	46.6	1.4
Apertus-8B	38.1	41.4	14.2	17.5	6	29	90	42.5	46.9	48.0	-1.2
Aya Expanse-32B	47.4	52.1	13.5	18.3	8	42	90	34.7	39.9	40.3	-0.4
Aya Expanse-8B	38.3	47.2	13.0	21.8	2	53	91	38.7	49.4	50.7	-1.3
Gemma 3-12B	52.1	53.5	12.9	14.3	10	17	90	30.7	31.6	28.6	3.0
Gemma 3-27B	56.2	55.6	12.6	12.0	16	12	90	27.2	26.5	23.7	2.8
Gemma 3-4B	42.4	46.3	14.6	18.5	1	42	90	37.9	43.2	38.7	4.5
Gemma 4-31B	64.9	61.1	11.7	7.8	39	0	89	23.5	18.6	17.6	1.1
Llama 3.1-70B	53.6	56.2	12.7	15.3	9	29	89	30.2	33.3	34.7	-1.3
Llama 3.1-8B	32.9	43.9	13.0	24.0	0	66	90	44.2	57.5	54.2	3.3
Mistral-7B	35.1	45.6	12.7	23.1	1	65	90	39.2	52.3	52.0	0.3
Qwen 3-14B	52.8	55.9	11.8	15.0	9	28	90	28.8	32.4	28.0	4.4
Qwen 3-32B	53.6	56.6	12.2	15.1	11	27	90	29.3	32.7	26.9	5.8
Qwen 3.5-27B	54.3	54.5	11.0	11.2	19	15	89	27.0	27.7	23.7	3.9
Qwen 3-8B	50.0	53.7	11.9	15.6	10	39	89	30.2	34.8	29.8	5.0
Tiny Aya Earth	34.1	38.9	15.6	20.3	3	37	91	48.8	54.1	51.0	3.1
Tiny Aya Fire	34.8	41.4	13.7	20.3	1	39	90	43.1	50.9	48.9	1.9
Tiny Aya Global	33.2	38.7	15.2	20.6	5	36	91	49.0	55.8	53.8	2.0
Tiny Aya Water	34.8	40.6	14.6	20.4	2	37	90	45.9	52.4	48.4	4.0
<i>Open-weight</i> (20 models)	–	–	–	–	181	674	1800	–	–	–	–

Table 14: Cross-lingual accuracy asymmetries on region-specific items (*Region-conditional*, *Region-anchored* and *Region-embedded*), macro-averaged over languages. IL is the question in its original language, EN the English translation. The first group gives accuracy on region-specific items in each of the two conditions. *IL only* and *EN only* are the shares of region-specific items answered correctly in one of the two languages and incorrectly in the other. An exact two-sided McNemar test is applied to each combination of one model and one language holding at least one such item. *Tested* counts those combinations; *IL larger* counts the ones where the test rejects at $\alpha = 0.05$ and the items answered correctly only in the original language outnumber those answered correctly only in English; *EN larger* counts the ones where it rejects and that inequality is reversed. In the last group the arrow reads correct in the first language, wrong in the second: IL→EN is $\Pr(y^{\text{eng}}=0 \mid y^{\text{lang}}=1)$, and EN→IL is $\Pr(y^{\text{lang}}=0 \mid y^{\text{eng}}=1)$ on region-specific items (written D_{spec} in the text) and on region-agnostic items (D_{agn}). $\Delta = D_{\text{spec}} - D_{\text{agn}}$ is the part of the drop not attributable to general difficulty with the original language. Both EN→IL columns and Δ are averaged over the languages where both are defined, so Δ equals the difference of the two columns beside it. Subtotal rows sum the three count columns; rate columns are left blank there. Of 355 frontier combinations the test rejects for 210, of which 208 favour the original language; of 1800 open-weight combinations it rejects for 855, of which 674 favour English.

Language	Country	<i>n</i>	Gap	Lang. Mean Gap	Gap Residual	95% CI
Arabic	Egypt	659	+0.1	-1.2	+1.3	[+0.5, +2.2]
Arabic	Palestine	134	-0.3	-1.2	+0.9	[-0.5, +2.4]
Arabic	Jordan	2471	-0.4	-1.2	+0.8	[+0.2, +1.5]
Arabic	Qatar	275	-0.6	-1.2	+0.6	[-0.5, +1.5]
Arabic	Morocco	193	-2.6	-1.2	-1.4	[-2.7, -0.1]
Arabic	UAE	77	-3.5	-1.2	-2.3	[-4.2, -0.4]
Chinese	China	5361	+3.6	+1.2	+2.4	[+1.7, +3.0]
Chinese	Hong Kong	263	-1.2	+1.2	-2.4	[-3.0, -1.7]
Portuguese	Brazil	176	+1.4	+1.0	+0.4	[-0.4, +1.3]
Portuguese	Portugal	253	+0.5	+1.0	-0.4	[-1.3, +0.4]
Spanish	Costa Rica	106	+2.3	+0.2	+2.2	[+0.7, +3.8]
Spanish	Puerto Rico	45	+1.5	+0.2	+1.4	[-1.4, +4.0]
Spanish	Cuba	380	+1.4	+0.2	+1.2	[+0.3, +2.2]
Spanish	Paraguay	211	+1.3	+0.2	+1.1	[+0.0, +2.4]
Spanish	Bolivia	834	+1.1	+0.2	+0.9	[+0.2, +1.6]
Spanish	Colombia	777	+0.8	+0.2	+0.6	[-0.1, +1.4]
Spanish	El Salvador	268	+0.7	+0.2	+0.5	[-0.6, +1.6]
Spanish	México	861	+0.5	+0.2	+0.4	[-0.3, +1.1]
Spanish	Ecuador	1522	+0.3	+0.2	+0.2	[-0.4, +0.8]
Spanish	España	3961	+0.2	+0.2	+0.0	[-0.5, +0.5]
Spanish	Chile	456	-0.1	+0.2	-0.3	[-1.1, +0.6]
Spanish	Argentina	228	-0.2	+0.2	-0.3	[-1.6, +0.9]
Spanish	Perú	1154	-0.4	+0.2	-0.6	[-1.3, +0.1]
Spanish	Uruguay	85	-1.6	+0.2	-1.8	[-3.4, -0.1]
Spanish	Venezuela	41	-1.8	+0.2	-1.9	[-4.5, +0.6]
Spanish	Panamá	40	-4.4	+0.2	-4.6	[-7.0, -2.0]

Table 15: Within-language residuals of the IL–EN gap for pluricentric languages on region-specific items (all categories except Region-agnostic). Gap is defined as $Acc_{IL} - Acc_{EN}$, and Lang. Mean Gap is the average gap across countries sharing the same language. Gap Residual subtracts this language-level mean from each country gap, measuring where the original-language advantage is larger or smaller than the language average. Countries with fewer than 30 non-agnostic samples are discarded. Bootstrap confidence intervals (CI) are computed by resampling items within each language–country cell.