

AFRiNR: Evaluating Native Reasoning in African Languages

Enrique Molero^{1,*}, Louis Powell¹, Linda Petrini², Audace Niyonkuru³,
Samuel Rutunda³, Mugisha Boris Ishimwe³, Thomas Llewellyn⁴, Israel Abebe Azime⁵

¹The GSM Association (GSMA), ²Independent Contributor, ³Digital Umuganda,

⁴SOAS University of London, ⁵Saarland University

*Correspondence: emolero@gsma.com afrinr.com

 [emolero/AfriNR](https://github.com/emolero/AfriNR)  [visual-snow/AfriNR](https://twitter.com/visual-snow/AfriNR)

Abstract

Native reasoning is the ability to solve problems using linguistic and cultural understanding. Research on this capability in large language models (LLMs) has been concentrated in high-resource languages, leaving low-resource languages (e.g., African languages) neglected.

In this paper, we introduce **AFRiNR**, a benchmark of 1,082 native reasoning questions in nine African languages, authored entirely by native speakers. Subsequently, we evaluate frontier and African-language-specialised models against native-speaker baselines. Our evaluations reveal that frontier models remain below the best native-speaker baseline, although the strongest exceeds average human performance; models with stronger linguistic performance tend to have smaller English advantages on the same cultural questions ($r = -0.85$); and specialised models generally outperform their general-purpose counterparts in African languages but score lower in English.

1 Introduction

Cultural reasoning is the process of interpreting and responding to situations through the lens of cultural norms, values, and shared knowledge. In multilingual settings, language understanding and cultural reasoning are intertwined into what we call **native reasoning**. Limitations in this capability let two problems emerge in LLM multilingualism: a *linguistic gap*, where model capabilities are not equal across languages (Adelani et al., 2025), and a *lack of cultural awareness*, where models may default to Western-centric cultural values (Naous et al., 2024; Tao et al., 2024).

Consider the situation in Figure 1: the *obvious* answer (sending condolences and going to work) is insufficient in a cultural context where you are expected to pay condolences in person first. This

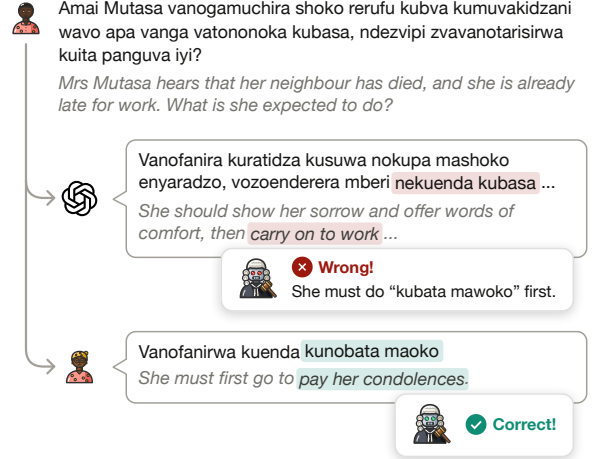


Figure 1: One AFRiNR question with English glosses: Shona (top). Middle: summarised incorrect response from GPT-6 Astra and the GPT-5.6 Luna Pro judge’s explanation. Bottom: a native-speaker response from the human baseline, graded correct.

mismatch illustrates how models **may not adapt to local cultural conventions** understood by native speakers, even when they correctly use the language. Failure to adapt is more pronounced in low-resource languages, including African languages (Adelani et al., 2025; Kazemi et al., 2024). These gaps raise a question: **To what extent do LLMs demonstrate native reasoning in African languages?**

Prior work has evaluated cultural reasoning in English (Rao et al., 2025; Yu et al., 2025) and native reasoning in high-resource languages (Fabbri et al., 2025). Evaluations covering African languages have focused on everyday cultural knowledge or translations of existing popular benchmarks, often without native-speaker baselines (Table 1).¹ The limitations of existing benchmarks also limit our ability to evaluate multilingual mod-

¹These baselines help establish whether questions are solvable and provide a reference for model performance.

“Language, any language, has a dual character: it is both a means of communication and a carrier of culture.”
— Ngũgĩ wa Thiong’o, *Decolonising the Mind* (1986)

Benchmark	African langs.	Question source	Cultural focus	Human baseline
CulturalBench	—	Human–AI	Knowledge	✓
BLEnD	✓	Templates	Knowledge	—
IrokoBench	✓	Translated	—	—
AfroBench	✓	Mixed	—	—
MultiNRC	—	Native	Reasoning	—
Afri-MCQA	✓	Native	Knowledge QA [†]	—
AFRINR	✓	Native	Reasoning	✓

Table 1: AFRINR and related benchmarks. ✓: included; —: not reported or not a design target. [†]Afri-MCQA also includes inference questions.

els specialised in African languages (Tonja et al., 2024; Yu et al., 2026).

In order to measure **African Native Reasoning** in current frontier and specialised models, we created AFRINR, a benchmark containing 1,082 linguistic and culturally grounded questions in nine sub-Saharan African languages; written and validated by more than 100 native speakers from the corresponding regions and communities. Our main findings can be summarised as follows:

- **Native speakers outperform frontier models on average.** Our evaluation shows that proprietary and open-weight models fall below average human performance (~ 4 and ~ 24 percentage points on average, respectively), while none of the models evaluated has reached the best human baseline.
- **Stronger linguistic performance is associated with a smaller linguistic gap.** Models scoring higher on AFRINR’s linguistic questions tend to show a smaller accuracy gap between English and African languages on the same cultural questions ($r = -0.85$).
- **Multilingual specialisation in low-resource languages comes with a trade-off.** Evaluating these specialised models reveals that, although their scores remain low ($\sim 12\%$ on average), they achieve up to approximately three times their general-purpose counterparts’ performance in local languages, with relative losses of 22–65% in English accuracy.

2 Related Work

Multilingual evaluation. Existing benchmarks are dominated by subject knowledge and academic understanding tasks in English. For example, MMLU evaluates knowledge and reasoning across a broad range of academic and professional subjects (Hendrycks et al., 2021). IrokoBench expanded language coverage by translating existing knowledge, natural language inference, and arithmetic reasoning tasks into African languages (Adelani et al., 2025), while Uhura introduces scientific question answering and truthfulness in low-resource African languages (Bayes et al., 2026). Similarly, aggregated benchmarks combine translated and locally sourced datasets covering classification, question answering, reasoning, and generation to evaluate language understanding in African languages (Ojo et al., 2025; Adebara et al., 2025).

Cultural Reasoning Benchmarks. Extending LLM evaluation to cultural variants allows us to assess whether models adequately understand and respond to culture-specific norms, values, and contextual nuances. To evaluate language understanding across diverse cultures and languages, works such as BLEnD (Myung et al., 2024) and Global PIQA (Chang et al., 2025) have introduced culturally specific benchmarks developed with native speakers. MultiNRC (Fabbri et al., 2025) explores *native reasoning* capabilities in LLMs but doesn’t cover low-resource languages, which leaves a natural point of extension.

Specialised LLMs Recent work investigates how cultural representation changes across different stages of LLM training and examines whether preserving cultural information through explicit cultural markers can improve culturally grounded capabilities (Sahu et al., 2026). For African languages, works such as Sunflower and AfriqLLM have incorporated culturally relevant data into their pretraining and post-training stages to create specialised LLMs (Aker et al., 2026; Yu et al., 2026).

Our work therefore extends the MultiNRC categories to low-resource African languages and adds native-speaker baselines. Table 1 summarises how AFRINR relates to existing benchmarks: it is the only one that combines African languages, locally

authored questions, a focus on reasoning rather than recall, and a human baseline.

3 AFriNR Data Collection

3.1 Recruiting human experts

To collect questions that are both written by humans and potentially solvable by groups from a different region speaking the same language, we enlist three different annotation groups: Generators, Reviewers and Baseliners.

3.1.1 Generators.

This annotation group was in charge of writing the questions, following the guidelines in (§A.1). Our data annotation partner recruited and managed the question-writing group on-site in Africa.

3.1.2 Reviewers

The reviewers were an external group of experts who were recruited experts through Masakhane (Nekoto et al., 2020), we hired them as contractors through Upwork. We selected annotators based on their performance on a screening test in their chosen language (§B.2). We allocated three reviewers per language variety, covering all 10 varieties. These experts provided external feedback to dataset generation group (§3.4), and help us calibrate our automatic grader (§4.1). This process is explained in more details in Appendix §A.2.

3.1.3 Baseliners

Baseliners were native speakers recruited by our data annotation partner who had not written or reviewed any of the questions. Participants took the same qualification exam as the reviewers (§B.2).

3.2 Dataset Design

AFriNR extends the work done in MultiNRC (Fabri et al., 2025) by adopting the same reasoning categories. However, we renamed and separate their reasoning categories into two groups to better reflect our definition of cultural reasoning.

Linguistic Capabilities

- **Grammar:** language-specific linguistic reasoning. Questions require reasoning about the language’s grammatical rules or linguistic conventions.
- **Wordplay** and riddles. Questions are puzzles whose clues draw on culturally familiar objects,

Language	Grammar	Wordplay	Practices	Maths	Total
Amharic	30	22	40	45	137
Fulani	11	24	31	20	86
Kinyarwanda	37	26	50	18	131
Kirundi	18	28	17	25	88
Lingala	12	31	37	22	102
Malagasy	23	24	25	19	91
Shona	36	42	28	23	129
Swahili Mombasa	16	48	39	25	128
Swahili Tanzania	6	41	19	6	72
Wolof	16	41	32	29	118
Total	205	327	318	232	1,082

Table 2: Question counts by independent language variety and category.

metaphor, double meanings, similar sounds, or puns in the original language.

Cultural Reasoning Capabilities

- **Practices:** reasoning about cultural practices. Questions use local customs, ceremonies or traditions to reason about a situation.
- **Maths:** mathematical and logical reasoning with cultural relevance. Questions involve calculations grounded in local calendars, currencies, counting systems or units of measurement.

Languages Given our focus on Sub-Saharan languages, AFriNR covers nine languages from the region: Amharic, Fulani, Kinyarwanda, Kirundi, Lingala, Malagasy, Shona, Swahili and Wolof. These comprise ten language varieties, with Swahili Tanzania and Swahili Mombasa treated independently in scoring.

3.3 Human Baselines

Comparing human and model performance requires both comparable evaluation conditions and an appropriate human reference group.

Methodology Participants are tasked to submit an answer and explanation for each question, which is presented one at a time to encourage focus on each question. All participants are native speakers of their assigned language. Each submission² is scored using the exact automatic grader described in (§4.1), and the aggregation procedure described in (§4.2).

Metrics We define two metrics of human performance:

²Qualification submissions are screened for possible AI assistance using the shared expert qualification procedure (§B.2)

- **Average Human Performance.** We calculate each participant’s accuracy on the retained questions in their assigned set, counting unanswered and “I don’t know” responses as incorrect, then average these scores within each language variety and average the 10 varieties equally. This describes how well participants perform on average.
- **Best Human Performance.** For each assigned question set, we report the highest accuracy achieved by a participant. This captures the strongest individual performance on that set.

3.4 Creating the benchmark

Finally, we describe how generation, review and human baselines determine which questions enter the final benchmark.

1. Generators (§3.1.1) followed the collection guidelines (§A.1) and submitted multiple batches, where each datapoint is a triplet defined as: $D_i = (q_i, a_i, e_i)$, q_i is the question, a_i the answer, and e_i the explanation; each datapoint also contains a translated triple counterpart whenever it was permitted.
2. Reviewers (§3.1.2) were then presented a datasheet containing the batches, and were instructed to flag if the datapoint needed to be accepted, deleted or fixed. The final decision for each datapoint went through a majority voting process.³
3. Finally, to ensure our final benchmark didn’t contain broken questions⁴, we filter questions that could *not* have been solved by our Baseliners (§3.1.3).

The final result was a benchmark with 1,082 questions in nine languages, each with at least one saved correct human response (Table 2). The performance comparison uses the selected 49 participants, who collectively answer 1,078 of these questions correctly; participant selection does not change the retained question set.

³Review and revision continued until the reviewers accepted the questions.

⁴Questions not responded by humans might not be inherently wrong, but the possibility of having broken questions in exchange for more questions is a worse trade-off to have

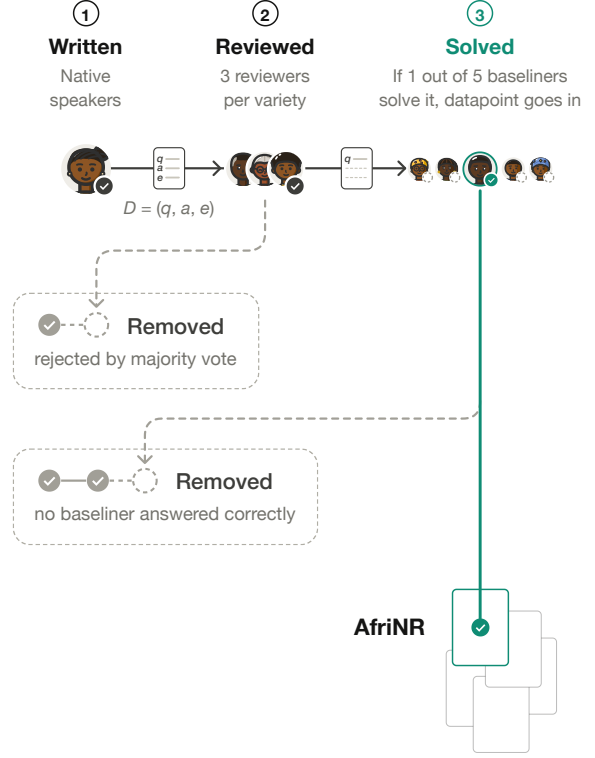


Figure 2: **How AfriNR was built.** (1) Native speakers recruited by our annotation partner write questions with reference answers and explanations. (2) Linguists and researchers recruited through Masakhane independently keep, correct or remove each question. (3) We retain only questions that at least one native-speaker baseliner answered correctly.

4 Methodology

4.1 Scoring

Building an automatic grader Since AfriNR answers are open-ended, we use an LLM-as-a-judge (Zheng et al., 2023) and measure its agreement with native-speaker judgements.

- **Step 1:** Recruited 21 human annotators to judge 1,050 responses from an earlier seven-model panel (§C.1).
- **Step 2:** Measured agreement between annotators using macro-F1 (MF1) (Arora et al., 2025) (§C.2).
- **Step 3:** Compared the AI judges’ agreement with humans to the agreement between humans (§C.2).

Because GPT-5.6 Luna Pro⁵ achieves an MF1 close to the human–human baseline (0.672 versus

⁵Gemini 3.7 Flash is excluded since it is part of the evaluated model panel.

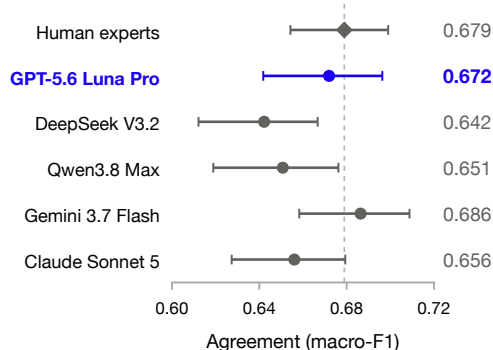


Figure 3: Judge agreement with human experts on the calibration sample (1,050 responses, seven language varieties). Each point is macro-F1 against the expert annotators, averaged equally across varieties; the top row is agreement among the experts themselves. Bars: 95% bootstrap intervals, resampling questions.

0.679), we adopt it as the AFRINR grader (Figure 3). This agreement concerns the historical calibration sample; it does not validate every language or the English and specialist conditions (§C.2).

4.2 Aggregating scores

We report macro-average accuracy across the ten language varieties, where each variety is equally weighted. Since the varieties contain different numbers of questions, we first calculate accuracy within each variety and then take the mean across the ten retained varieties. Swahili Tanzania and Swahili Mombasa are independent scoring units, each with the same weight as every other variety. This prevents varieties with more questions from having a larger influence on the overall score.

$$A_\ell = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} y_{\ell i}, \quad A_{\text{AFRINR}} = \frac{1}{L} \sum_{\ell=1}^L A_\ell. \quad (1)$$

Here, $L = 10$, n_ℓ is the number of questions in language variety ℓ , and $y_{\ell i}$ is the model’s score on question i . We report standard errors alongside these scores to describe their uncertainty (Miller, 2024).

4.3 Experimental Settings

Prompts. The generation prompt and judge template are described in (§D.2 and §D.3), respectively, with their exact text in the accompanying code.

4.3.1 Frontier Model Evaluation

We evaluate a panel of frontier models on the 1,082 retained questions using their saved zero-shot re-

Proprietary	Open Weights
GPT-6 Astra	Kimi K3
Claude Fable 5.1	Qwen3.8 2.4T A95B
Gemini 3.7 Flash	DeepSeek V4 Pro
Meta Muse Spark 1.2	GLM 5.3

Table 3: General-purpose models evaluated in AFRINR.

sponses. We implemented the evaluation using Inspect AI, an open-source Python framework for evaluating large language models.

Models We selected eight frontier general-purpose models from different developers to compare proprietary and open-weight systems. The models are listed in Table 3. We selected these models based on their prominence in research discussions and their general-purpose capabilities.

4.3.2 Specialist Model Evaluation

We select a separate panel of existing models adapted for African languages, prioritising multilingual coverage. We refer to these models as *specialists* throughout the paper. The panel includes:

1. McGill NLP’s AfriqueLLM Instruct releases, adapted through continued pretraining and instruction tuning;
2. Sunbird AI’s Sunflower, which combines continued pretraining with instruction tuning.

Each specialist is compared with a post-trained counterpart on identical original-language AFRINR questions within its supported languages. The appendix (§D.1) lists the languages included for each specialist and explains the exclusions; evaluation details are in (§D.2).

5 Results

In this section, we organise our results around our three main questions.

5.1 Main Results

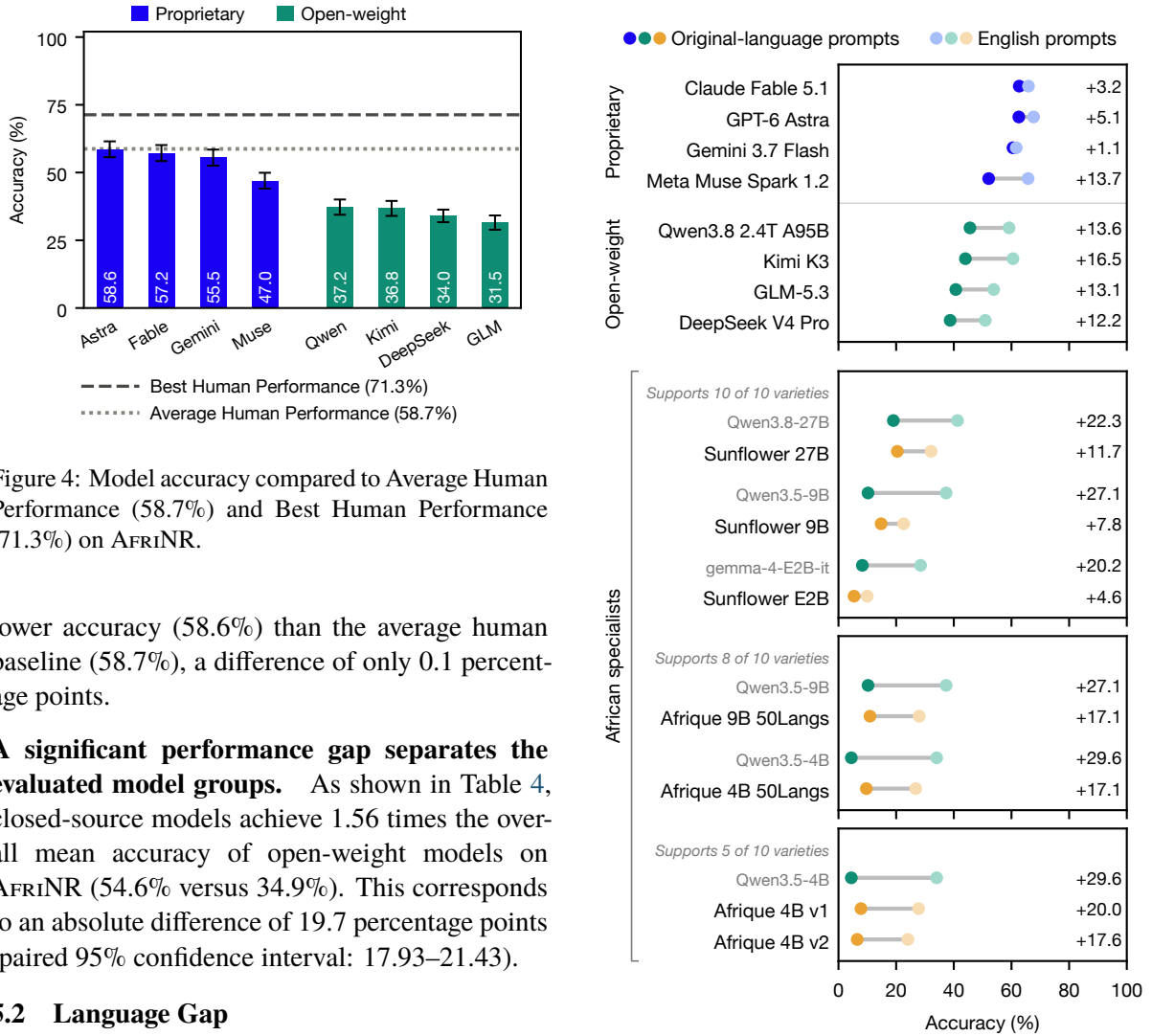
We start by returning to the first question in our paper: *how do the best LLMs perform comparatively against human experts?* We evaluate the panel (§4.3.1) in AFRINR to measure their native reasoning capabilities in African languages, demonstrating the following empirical findings:

No models exceed the best human performance

Looking at Figure 4, we find that the best-human baseline (71.3%) outperforms all eight LLMs evaluated. Additionally, GPT-6 Astra achieves slightly

Category	<i>n</i>	CLOSED-SOURCE					Mean	OPEN-WEIGHT					Mean	Δ	Human mean
		GPT-6 Astra	Claude Fable 5.1	Gemini 3.7 Flash	Muse Spark 1.2	Kimi K3		Qwen3.8 2.4T A95B	DeepSeek V4 Pro	GLM 5.3					
Grammar	205	49.3	42.6	46.4	39.0	44.3	26.3	31.5	26.8	20.4	26.2	18.1	48.8		
Wordplay	327	55.2	52.1	50.8	42.3	50.1	29.1	27.3	29.1	23.3	27.2	22.9	61.8		
Practices	318	56.8	58.0	57.5	48.7	55.2	39.2	41.3	35.4	36.9	38.2	17.0	56.8		
Maths	232	66.4	66.5	61.2	53.0	61.8	48.0	49.3	41.7	43.0	45.5	16.3	61.9		
AFRiNR Score	1,082	58.6	57.2	55.5	47.0	54.6	36.8	37.2	34.0	31.5	34.9	19.7	58.7		
SE (pp)		1.49	1.50	1.52	1.49	1.18	1.41	1.43	1.18	1.35	1.09	0.89	—		

Table 4: **Accuracy across the four categories.** Judged accuracy (%) of eight models and the human baseline on the same 1,082 AFRiNR questions, graded by GPT-5.6 Luna Pro. Each category score averages accuracy equally across its represented language varieties; the overall score averages all 10 language varieties. Δ : closed-source mean minus open-weight mean, computed from unrounded scores. Human means average participant accuracies within each independent language variety, then average varieties equally, using 49 selected participants. The appendix (§D.4) explains standard errors.



lower accuracy (58.6%) than the average human baseline (58.7%), a difference of only 0.1 percentage points.

A significant performance gap separates the evaluated model groups. As shown in Table 4, closed-source models achieve 1.56 times the overall mean accuracy of open-weight models on AFRiNR (54.6% versus 34.9%). This corresponds to an absolute difference of 19.7 percentage points (paired 95% confidence interval: 17.93–21.43).

5.2 Language Gap

To tackle the second research question, we first identify which subset of AFRiNR we can use for this comparison:

- Linguistic questions are prone to losing their challenge when translated, because their difficulty can depend on the sounds, grammatical rules or wordplay of the original language, which may not carry over into English.
- Cultural reasoning questions, in contrast, require the model to answer within the relevant local context. Thus, when they are translated into English, we identify the relevant community in the prompt, but the model must still draw on local knowledge and apply it to answer the question.

Our experiment therefore consists of using a subset of 531 Cultural Reasoning questions from AFRINR. This time, the questions are presented in English, and we compare the results with those obtained on the same questions in their African languages counterpart.

All eight models achieve higher overall accuracy in English. As shown in Figure 5, all eight models achieve higher accuracy when the same cultural questions are presented in English. These gains suggest that part of the lower accuracy observed in the previous experiment reflects difficulty understanding the original language. The English advantage is larger on average for open-weight models, roughly halving the accuracy gap between the two groups.

However, within the proprietary group, Muse Spark 1.2 shows a much larger English advantage than the other three models, comparable to that of the open-weight models. This larger linguistic gap suggests that Muse has greater difficulty understanding these questions in their original African languages than the other proprietary models.

Fulani, Wolof and Kirundi have the largest linguistic gaps. Across all eight frontier models, accuracy is higher when questions in these three languages are presented in English rather than their original language. Open-weight models show larger average accuracy gains than closed-source models, with nearly twice the gain in Fulani (+52.4 versus +26.5 percentage points; Table 5).

5.3 African specialisation

Our last set of results addresses the third research question: *Can further training in African languages reduce the gaps demonstrated in the previous sections?* From this point onwards, we re-

Language	Closed-source		Open-weight	
	Gain (pp)	Ratio	Gain (pp)	Ratio
Fulani	+26.5	1.63×	+52.4	4.02×
Kirundi	+10.7	1.20×	+14.3	1.37×
Wolof	+14.8	1.23×	+37.7	1.96×
Overall	+5.8	1.10×	+13.9	1.33×

Table 5: Accuracy improvement when the same questions are presented in English rather than their original language. Ratio is English accuracy divided by original-language accuracy. Each group averages four models; Overall averages across all ten language varieties, treating Swahili Tanzania and Swahili Mombasa independently.

Model	Counterpart	n/L	Accuracy (%)		
			Counterpart	Specialised	Ratio
<i>McGill Afrique</i>					
4B 50Langs [†]	Qwen3.5-4B	905/8	3.5	10.7	3.0×
9B 50Langs [†]	Qwen3.5-9B	905/8	7.6	9.6	1.3×
4B v1 [†]	Qwen3.5-4B	597/5	4.5	11.7	2.6×
4B v2 [†]	Qwen3.5-4B	597/5	4.5	10.6	2.3×
<i>Sunbird Sunflower</i>					
27B	Qwen3.8-27B	1,082/10	17.7	21.5	1.2×
9B	Qwen3.5-9B	1,082/10	11.6	14.2	1.2×
E2B	gemma-4-E2B-it	1,082/10	5.2	4.0	0.8×

Table 6: Specialist and counterpart accuracy (%) on matched questions, averaged equally across supported language varieties, with Swahili Tanzania and Swahili Mombasa scored independently. n/L : questions/language varieties; ratio: specialist/counterpart accuracy. [†] as in Figure 7.

fer to *specialisation* as further training of general-purpose models on African-language data, through continued pre-training and instruction tuning.

We evaluate publicly released specialists from the McGill Afrique and Sunbird Sunflower families against their corresponding general-purpose models, asking whether this specialisation is accompanied by higher accuracy in AFRINR and a smaller English advantage. We repeat the same experiments done in the previous sections, by evaluating these specialist models in AFRINR in both the full benchmark and the translated subset.

The gains are concentrated in Grammar and Practices. This suggests that specialists become better at handling language-specific reasoning and applying local cultural knowledge in African languages. For example, Afrique 4B 50Langs’ accuracy rises from 3.5% to 10.7%, reaching 3.0 times the accuracy of its general-purpose counterpart.

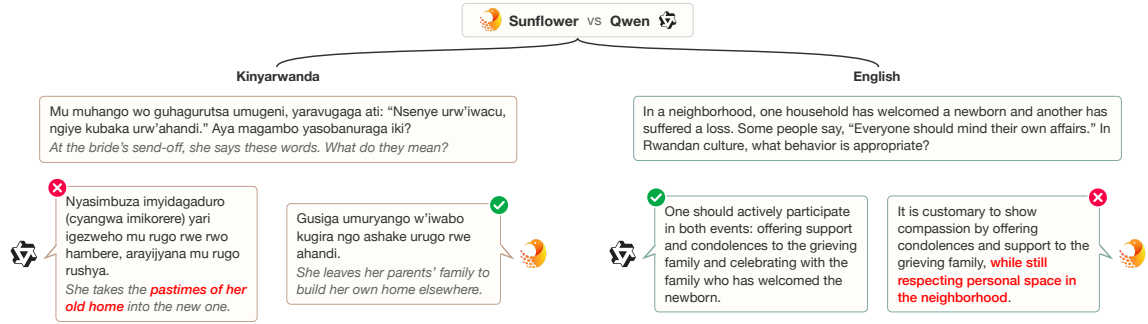


Figure 6: **Sunflower 27B against Qwen3.8-27B.** Sunflower 27B is Qwen3.8-27B trained further on African-language text. Asked in Kinyarwanda (left), Sunflower explains a bride’s farewell words that Qwen misreads. Asked in English (right), Sunflower drops the newborn from a question on neighbourly duty and adds a caveat about personal space. Real answers, trimmed; English glosses in italics.

However, the pervasive linguistic gap still remains after specialisation. Despite these improvements, all seven specialists still achieve higher overall accuracy when prompted in English (Figure 5).

The specialists also score lower in English than their general-purpose counterparts. All seven specialists still score higher in English than in the original language, but all seven also score lower in English than their general-purpose counterparts (Figure 6).

For example, Sunflower E2B’s English accuracy falls from 28.5% to 10.0%, a 65.0% relative decrease, while its African-language accuracy also falls from 8.3% to 5.4%. The linguistic gap therefore narrows even though culturally grounded reasoning accuracy falls in both languages.

One hypothesis could be that some of the reasoning capabilities required by these questions remain English-dependent and are not fully preserved during African-language specialisation.

6 Conclusion

We introduce AFRINR, a benchmark designed to act as a proxy for both linguistic and cultural understanding, merging these capabilities into what we describe as Native Reasoning. We hope these results inform those allocating efforts to close the linguistic gap.

The higher accuracy of proprietary models in our evaluation may make them appear as the best choice for LLMs deployment in Africa. However, limited research budgets, unequal access to compute, and deployment costs can constrain adoption in the region. Yet, the difficulties understanding African languages are still entrenched as no model

gets to the best human performance. Specialising open-weight models is a natural alternative, but our results suggest that multilingual fine-tuning can degrade their performance in English.

Future work should test whether these trade-offs carry over to realistic agentic tasks in African languages, where failures of native reasoning have practical and economic consequences.

Limitations

Our benchmark has a few limitations:

Limited regional coverage. While we initially aimed to collect questions from different regions within each language—as we did for Swahili⁶ in Tanzania and Mombasa—we faced practical difficulties extending this approach to the other languages.

Binary scoring. Our current scoring method does not capture the full linguistic nuance of LLM responses. We also cannot assume that an LLM judge reliably assesses answers in languages it may itself struggle to understand.

Interpreting African-language specialisation. Differences among specialised models and their training methods, together with the lack of access to intermediate training checkpoints, prevent us from isolating the cause of their lower English accuracy.

Limited guidance for training. Our benchmark helps track performance in frontier models, but may provide less informative feedback for smaller

⁶Swahili’s large speaker population helped us achieve broader regional coverage, but it remained the only language represented across multiple regions.

models. Given the scarcity of low-resource training data, benchmark scores alone may not give smaller research labs enough guidance to prioritise data collection.

Licences

AFRINR is distributed under the CC-BY-SA licence.

Ethical Considerations

All contributors to AFRINR were paid fairly for the time and effort they brought to question writing, independent review, and the human baseline. Every contributor understood the purpose of the project and agreed to participate. To reduce regional and gender bias in the dataset, we sought native speakers from different regional groups and aimed for an equal split between men and women. Given the cultural and subjective nature of these questions, higher education is not always necessary, so we only required question annotators to have completed secondary school. However, for the human baseline, we explicitly sought participants with a stronger educational background to help filter out questions that relied too heavily on regional colloquialisms, since we aimed to create a generalisable dataset. Despite these efforts, dataset quality may vary across languages, and some topics may be harder to evaluate or interpret depending on the linguistic and cultural context of each question.

Use of AI Assistants

We used Claude and ChatGPT to assist exclusively with manuscript preparation, revision and code generation; this includes grammatical correction and stylistic polishing. We also used Claude Code and Codex to help us build an annotation platform and tools for coordinating data annotators. The authors reviewed and verified all AI-assisted text and code used in this work and take full responsibility for its accuracy and originality.

References

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating llm performance on african languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32704–32731.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, et al. 2025. Irokobench: A new benchmark for african languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*. ArXiv:2406.03368.

Benjamin Akera, Evelyn Nafula Ouma, Gilbert Yiga, Patrick Walukagga, Phionah Natukunda, Trevor Saaka, Solomon Nsumba, Lilian Teddy Nabukeera, Joel Tibabwetiza Muhanguzi, Imran Sekalala, Nimpamya Janat Namara, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2026. [Sunflower: A new approach to expanding coverage of African languages in large language models](#). In *Proceedings of the AI for African Languages Conference 2025*, volume 314 of *Proceedings of Machine Learning Research*, pages 1–20. PMLR.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Edward Thomas Bayes, Israel Abebe Azime, Jesujoba Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A. Etori, Shamsuddeen Hassan Muhammad, Choice Mpanza, Igneciah Pocia IP Thete, Dietrich Klakow, and David Ifeoluwa Adelani. 2026. [Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource African languages](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 1485–1504, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Tyler A Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, et al. 2025. Global piqa: Evaluating commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*.

Alexander R. Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. 2025. Multinrc: A challenging and native multilingual reasoning evaluation benchmark for llms. *arXiv preprint arXiv:2507.17476*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multi-task language understanding. In *International Conference on Learning Representations (ICLR)*. ArXiv:2009.03300.

Sharif Kazemi, Gloria Gerhardt, Jonty Katz, Caroline Ida Kuria, Estelle Pan, and Umang Prabhakar.

2024. [Cultural fidelity in large-language models: An evaluation of online language resources as a driver of model performance in value representation](#). *arXiv preprint arXiv:2410.10489*.
- Evan Miller. 2024. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, et al. 2024. BLEND: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*. ArXiv:2406.09948.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. [Having beer after prayer? measuring cultural bias in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, et al. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on african languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095. ArXiv:2311.07978.
- Abhinav 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, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ananya Sahu, Mehrnaz Mofakhami, Daniel D’Souza, Thomas Euyang, Julia Kreutzer, and Marzieh Fadaee. 2026. The culture funnel: You can’t align what isn’t in the data. *arXiv preprint arXiv:2606.13808*.
- Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2024. [Cultural bias and cultural alignment of large language models](#). *PNAS Nexus*, 3(9):pgae346.
- Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2024. [InkubaLM: A small language model for low-resource African languages](#). *arXiv preprint arXiv:2408.17024*.
- Haeun Yu, Arnav Arora, Seogyong Jeong, Nadav Borenstein, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. 2025. [CulTrace: Tracing internal cultural reasoning in large language models](#). *arXiv preprint arXiv:2508.08879*.
- Hao Yu, Tianyi Xu, Michael A. Hedderich, Wassim Hamidouche, Syed Waqas Zamir, and David Ifeoluwa Adelani. 2026. [AfriqueLLM: How data mixing and model architecture impact continued pre-training for African languages](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5909–5928, San Diego, California, United States. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*. ArXiv:2306.05685.

A Guidelines

A.1 Generation guidelines

Question writers followed the same guidelines across all languages: questions should require several reasoning steps rather than direct factual recall or an answer obtainable through a simple online search.

Preparing a submission. Each submission contained five fields: **category**, **question**, **reasoning**, **final answer**, and **context or sub-category**. The guidelines specified how to write the question and its supporting material:

- **Question.** Use the relevant language’s orthography and appropriate scholarly or traditional vocabulary, without code-switching.
- **Answer.** Provide a short final answer.
- **Explanation.** Describe the reasoning a native speaker would use, in English or the question’s language. This explanation is for review and testing and should not be shown to the answering model.

Category requirements. The four categories required different forms of reasoning:

- **Practices.** Require a decision based on local calendars, time-keeping systems, customs, festivals or ceremonies rather than recall of a date or name. Identify the tradition and explain how its rules determine the answer.
- **Wordplay.** Use language-specific meanings, sounds, metaphors or word boundaries, including puns, homophones and traditional riddles. The answer should follow logically for a speaker familiar with the intended wordplay, and the explanation should identify that interpretation. Do not give away the intended word segmentation or translate the question into or from English.
- **Maths.** Require a calculation with at least one step that depends on interpreting a cultural element, rather than bare arithmetic. Explain the cultural interpretation and the calculation.
- **Grammar.** Require the application of language-specific grammatical rules or conventions.

Topic restrictions. The guidelines prohibited hate speech, discrimination, harassment and dangerous advice, and asked writers to avoid individ-

ual health, personal identity, party politics and inflammatory religious topics. Religious holidays and customs could be used when the question concerned cultural reasoning rather than doctrinal teaching.

A.2 Reviewing guidelines

Reviewing a question record. Reviewers were instructed to read the complete question record without editing it, then complete the four mandatory fields below. They had to make these judgments independently, without LLMs or consultation outside the project.

Field	Choices
Disposition	Keep as is; fix needed; drop it.
Reference answer	Correct and complete; correct but incomplete; one of several equally good answers; wrong.
Community knowledge	Anyone could answer; regional knowledge helps; community knowledge needed; elder or specialist knowledge needed.
Scope	Nationwide (wherever the language is spoken); one region; one community; not sure.

The community-knowledge field described what a reader needed to know; it was not a quality score.

Recording problems. Reviewers could add notes to explain their judgement, and selecting “fix needed” required a corrected native answer. They could also select any applicable flags: *native text wrong*, *translation off*, *explanation weak*, *answer mismatch*, *not cultural*, *people disagree*, *question broken*, or *could not verify*.

Response grading. To grade a model response, reviewers used the question, category, expert reference answer and explanation to decide whether it was **C (correct)** or **I (incorrect)**. The judgement concerned correctness rather than style, length or formatting, and there was no partial credit. The following rules specified what counted as correct in each category:

- **Grammar:** accept the expected form, including the same form in another script; reject spelling or grammatical errors. When the question asks for an explanation of a rule, grade its meaning.
- **Wordplay:** accept the intended answer or an alternative with the same sound and meaning; reject literal readings that do not solve the riddle and irrelevant answers.
- **Practices:** accept the community-specific fact or outcome; reject vague or generic answers and answers about another community’s tradition.

- **Maths:** accept the correct conclusion with any required unit, currency or calendar; reject an incorrect final amount.

A.3 Baseline guidelines

Instructions for Baseliners

We want to understand what people can answer using their own knowledge and the information in each question. Please complete the assigned questions independently, without AI, or another person's help. Please also avoid reading reference answers or other participants' responses, as these could influence your own answer.

For each question

- Please read each question in full and follow its instructions.
- Use the Answer field for your own answer and the Explanation field to describe why you chose it.
- Write both in the requested language, keeping any local words you need.

What to write in Explanation

Use the Explanation field to tell us how you worked out your answer. For a riddle, explain which clues led you to it and how they fit your answer. If you used a calculation, show the steps so we can follow how you got the result. If your answer depends on a word, expression, or local custom, explain what it means and how it relates to the question. Please name the place or community if that meaning depends on where it is used. If you already know the answer, please mention this and explain why it fits the question if you can.

If you are unsure

An honest attempt is useful even when the answer is wrong, so please give your best answer and explain any uncertainty. If you cannot answer, you can write "I don't know" in Answer and use Explanation to describe what is unfamiliar or missing, or any error you notice in the question. If you know an answer but cannot explain why, please say so rather than inventing a reason.

B Human Recruitment

We recruited Reviewers through a public application shared in the Masakhane community (Nekoto et al., 2020), with each applicant choosing the language they would work in. Our data annotation partner, Digital Umuganda, recruited and managed Baseliners, as well as the Generators who worked on-site in Africa.

B.1 Eligibility criteria

We sought the same expert profile for Reviewers and Baseliners:

- **Expertise.** Native speakers with knowledge of their language variety, its grammatical conventions and local cultural practices.
- **Education.** Completed secondary school was required; a university degree was preferred.

We sought participants from different regional groups and aimed for an equal split between men and women.

B.2 Qualification exam

Reviewers and Baseliners took the same language-specific qualification exam to assess their ability to judge the questions, answers and explanations used in the benchmark. Each candidate reviewed four questions, one from each category, alongside their reference answers, English translations and explanations. They made seven judgements per question, giving 28 judgements in total.

What candidates assessed. The seven judgements covered three aspects of each question:

- **Quality (four checks).** Was the question clear and answerable, the reference answer correct and complete, the English translation faithful, and the explanation sufficient to justify the answer?
- **Knowledge and reasoning (two checks).** How much community knowledge and reasoning did the question require?
- **Category requirement (one check).** Did the question meet the requirement for its category, listed below?

Category	Seventh check
Practices	Would knowledgeable community members agree on the answer?
Maths	Would missing local knowledge lead to a confident but wrong answer?
Grammar	Was the grammatical rule absent from English?
Wordplay	Was the riddle known in the tradition, and did its clues lead to the answer?

Screening submissions. Before scoring, we removed two of three submissions that contained the same answers to all 28 checks, keeping one. We also excluded submissions that approved all 16 quality checks across the four questions without identifying an issue.

Ranking candidates. We ranked candidates by weighted agreement with the most common answers among applicants for the same language variety. Only judgements for which at least 60% of applicants chose the same answer contributed to the score. We gave greater weight to judgements with clearer consensus and used unweighted agreement as the secondary ranking criterion. The pool contained valid, complete forms after duplicate removal and included the candidate being scored.

For language variety ℓ , let U_ℓ be the set of eligible judgements, m_q the modal answer to judgement q , and $a_{i,q}$ candidate i 's answer. The score is the weighted proportion of eligible judgements on which the candidate agrees with the pool:

$$S_i = \frac{\sum_{q \in U_\ell} w_q \mathbf{1}[a_{i,q} = m_q]}{\sum_{q \in U_\ell} w_q}, \quad (2)$$

$$w_q = \frac{n_q^{(1)} - n_q^{(2)}}{n_\ell},$$

Here, $n_q^{(1)}$ and $n_q^{(2)}$ are the two largest answer counts and n_ℓ is the number of candidates in the language-variety pool. The applicant pool does not provide an external answer key.

Screening for model-like answers. We also compared how likely a candidate's pattern of answers was under model and human answer distributions. The model distribution pooled 1,396 completions of individual category forms from Claude Opus 5, Claude Sonnet 5, GPT-5.6 Sol and Gemini 3.1 Pro. The human distribution used the applicant pool but excluded the candidate being scored.

We applied Jeffreys smoothing with $\alpha = 1/2$ to both distributions, over the form's answer choices and any additional observed answers. Writing these distributions as $\hat{p}_q^M$ and $\hat{p}_q^{H,-i}$, the model-resemblance score was

$$\Lambda_i = \sum_{q=1}^{28} \log \frac{\hat{p}_q^M(a_{i,q})}{\hat{p}_q^{H,-i}(a_{i,q})}. \quad (3)$$

A higher score means that the candidate's answers are more likely under the model distribution relative to the human distribution. Resemblance to model answers does not prove that a candidate used AI assistance.

To set the exclusion threshold for each language variety, we simulated 50,000 submissions by drawing the 28 judgements independently from the smoothed full human-pool distributions, with seed 7. We excluded candidates whose resemblance score exceeded the 95th percentile of these simulated scores. The simulation does not establish how often the procedure would wrongly exclude human applicants.

B.3 Participant selection

Reviewers. We selected the highest-ranking eligible candidates within each language variety, allocating three reviewers per variety. We hired them as contractors through Upwork.

Baseliners. Baseliners had not written or reviewed any of the questions. The human baseline comprises 49 participants, with five per language variety except Amharic (4), where we could not recruit a fifth.

Where more than five participants were available, we selected the five with the highest accuracy on their original assigned questions. This selection did not change the retained question set.

C Judge Validation

C.1 Validation sample

To select an automatic grader, we compared five LLM judges with native-speaker experts on the same model responses. The validation panel comprised 21 experts across seven language varieties, with three experts independently grading 150 responses per variety. This gave 1,050 responses from 498 questions. Experts were selected through the qualification process in §B.2.

Selecting responses. We sampled from the difficult-question pool, stratifying by answering model and category. The sample deliberately included responses previously graded correct and questions on which model responses had received discordant grades. Table 7 gives the sampling targets, the resulting proportions and the models used in this study.

Collecting independent grades. Within each variety, all three experts saw the same responses, with model identities and prior grades hidden. We removed answer wrappers and closed reasoning blocks before presenting responses to both experts and LLM judges. The experts applied the binary grading criteria in §A.2.

C.2 Agreement assessment

We measured how closely each judge agreed with the experts, using agreement among the experts themselves as the reference. We used macro-F1 (MF1) (Arora et al., 2025), which gives equal weight to the correct (C) and incorrect (I) labels. Within each language variety, we calculated the two estimates as follows:

- **Judge agreement.** We compared each judge with all three experts, pooling 450 comparisons into one MF1 score.
- **Human agreement.** We compared each expert with the other two, giving 300 comparisons per expert, and averaged the three MF1 scores.

Component	Validation design
Answering models	Claude Sonnet 5, Gemini 3.7 Flash, GPT-5.6 Sol, Kimi K3, Meta Muse Spark 1.2, Qwen3.8 Max.
Judges	GPT-5.6 Luna Pro, DeepSeek V3.2, Qwen3.8 Max, Gemini 3.7 Flash, Claude Sonnet 5.
Sampling targets	Up to one-third of responses previously graded correct; 60% from questions with discordant prior grades across models.
Sample composition	Previously judge-correct responses: 30.7% in the sample versus 16.8% in the pool. Responses from discordant questions: 66.7% versus 45.8%.

Table 7: Models and sampling design for the historical judge-validation study.

Language	Human	Luna	DeepSeek	Qwen	Gemini	Sonnet
Amharic	0.534	0.609	0.630	0.599	0.614	0.577
Fulani	0.538	0.482	0.458	0.442	0.484	0.459
Kinyarwanda	0.704	0.762	0.726	0.760	0.753	0.746
Malagasy	0.820	0.778	0.760	0.764	0.803	0.778
Shona	0.653	0.611	0.593	0.560	0.664	0.585
Swahili-Mombasa	0.781	0.781	0.735	0.760	0.803	0.794
Wolof	0.724	0.681	0.594	0.670	0.683	0.651
Mean	0.679	0.672	0.642	0.651	0.686	0.656

Table 8: Class-balanced MF1 in the completed calibration panel: three human experts and five LLM judges per language. Each language has 150 responses, 450 comparisons per LLM judge and 900 ordered expert-expert comparisons. Full judge names and aggregate intervals appear in Figure 3.

Language	n	Majority C	Luna C	Luna-only C / majority I	Majority-only C / majority C
Amharic	150	109	58	1/41 (2.4%)	52/109 (47.7%)
Fulani	150	96	28	3/54 (5.6%)	71/96 (74.0%)
Kinyarwanda	150	30	38	14/120 (11.7%)	6/30 (20.0%)
Malagasy	150	61	45	6/89 (6.7%)	22/61 (36.1%)
Shona	150	96	48	1/54 (1.9%)	49/96 (51.0%)
Swahili Mombasa	150	59	34	0/91 (0.0%)	25/59 (42.4%)
Wolof	150	72	38	3/78 (3.8%)	37/72 (51.4%)
Total	1050	523	289	28/527 (5.3%)	262/523 (50.1%)

Table 9: **Luna and expert-majority credits.** C denotes Correct and I Incorrect; majority requires two of three experts. Disagreement cells show count/condition denominator (percent). These descriptive rates concern the selected historical calibration sample and are not correction rates for the full benchmark.

We then averaged each estimate equally across language varieties.

Calculating MF1. Let n_{ab} count comparisons with prediction a and reference label b . The score is

$$\text{MF1} = \frac{1}{2} \sum_{c \in \{C, I\}} \frac{2n_{cc}}{2n_{cc} + n_{CI} + n_{IC}}, \quad (4)$$

with a zero contribution when a denominator is zero. To estimate uncertainty, we resampled source questions within each language variety, keeping all responses and labels for a question together. We recomputed the full aggregation in 30,000 bootstrap replicates (seed 20260828) and used the 2.5th and 97.5th percentiles for the 95% intervals.

Interpreting agreement. Table 8 shows agreement by language variety, while Table 9 shows where Luna and the expert majority assign different grades. These results describe the selected historical sample; they do not validate grading across the full nine-variety release, the English comparison or specialist responses. MF1 is not chance-corrected, and we did not prespecify an equivalence margin or adequacy threshold, so a score close to human agreement is not an equivalence test. The disagreement rates are specific to this sample and should not be used to correct scores elsewhere in the benchmark.

D Evaluation Procedure

The evaluation separates generating answers, grading the saved responses and calculating accu-

racy. Our accompanying repository implements the native-language pipeline with Inspect AI 0.3.247 and includes the questions in `data/test.parquet`. The repository README provides the installation and execution commands.

D.1 Comparison design

Using the released questions. The 1,082 retained questions each have a saved human response graded correct by GPT-5.6 Luna Pro. To compare models with people who saw the same material, we checked that the native question, reference answer and explanation matched the edition served to baseliners. We excluded 14 otherwise eligible questions whose records differed from that edition. The eight main models in §4.3.1 are evaluated on all 1,082 retained questions.

Comparing language conditions. The original-language and English comparison uses 531 paired, human-answered questions from Practices and Maths. We compare the same questions in both conditions, so differences in the question set cannot account for the change in accuracy. The translation pool comprised 701 screened annotator/GPT-6 Astra translations and 43 expert-reviewed additions, of which four entered the comparison. Astra was also one of the evaluated models; the automated translation screening and response grades have not been validated by native speakers.

Both the question and the grading reference change language in this comparison. The resulting score difference therefore cannot isolate comprehension, recall or grading effects.

Matching specialists to their counterparts. Each specialist is compared with its post-trained counterpart on identical questions within the specialist’s supported languages (Table 6). We use the model cards to determine language coverage, as summarised in Table 10.⁷ The Afrique cards list Plateau Malagasy, and the 50Langs cards also list Nigerian Fulfulde. Their match to the benchmark’s Malagasy and Fulani varieties is unconfirmed, and we exclude them from the comparison. We map the cards’ Rundi to Kirundi.

The same coverage rule applies to the specialist language comparison. Each pair uses its supported-language subset of the 531 paired questions: 281 for Afrique 4B v1/v2, 439 for Afrique 50Langs

Language variety	Afrique		
	4B v1/v2	4B/9B 50Langs	Sunflower 27B/9B/E2B
Amharic	Yes	Yes	Yes
Fulani	—	V	Yes
Kinyarwanda	Yes	Yes	Yes
Kirundi	—	Yes	Yes
Lingala	—	Yes	Yes
Malagasy	V	V	Yes
Shona	Yes	Yes	Yes
Swahili Tanzania	Yes	Yes	Yes
Swahili Mombasa	Yes	Yes	Yes
Wolof	—	Yes	Yes

Table 10: Language coverage in the specialist comparisons. Yes: included; V: a named variety is listed, but its match to the benchmark variety is unconfirmed and it is excluded; —: not listed as supported. Swahili Tanzania and Swahili Mombasa are independent scoring units with equal weight.

and 531 for Sunflower. We generate one response per question and language condition, reusing the same counterpart responses when a counterpart is shared by several specialists. These matched subsets determine the specialist-versus-counterpart results in the text. The broader comparison in Figure 5 and Table 12 instead shows every model on all 531 questions, including languages outside a specialist’s documented coverage.

D.2 Response generation

Running the native evaluation. The Inspect task in `benchmark/generate.py` loads the original-language question as input and keeps its reference answer and explanation for grading. Generation and grading run separately, so a saved response can be graded without asking the answering model to generate it again.

The repository provides three entry points:

1. **Generate answers** with the `afrinr` task in `benchmark/generate.py`, selecting the model and language varieties through the Inspect CLI.
2. **Grade saved answers** with the `scorer` in `benchmark/judge.py` and the grader described in §D.3.
3. **Calculate accuracy** with `benchmark/summarise.py`, which checks the saved records against the released questions before building the tables.

This native runner summarises one response per question. The repeated-attempt, paired-English

⁷For Afrique Instruct, coverage comes from the continued-pretrained 4B and 50Langs 4B/9B cards; Sunflower coverage comes from its 27B, 9B and E2B cards.

and specialist analyses use the aggregation procedures described in §D.4.

What the answering model receives. In the native evaluation, we give the model the question in its original language and a shared system prompt asking it to reason like a native speaker, drawing on local knowledge and oral tradition. The prompt asks for a final answer inside `<answer>` tags; the expert answer and explanation are not shown to the answering model. The exact system prompt is stored in `benchmark/prompts.py`.

Generation settings. The main models use the same generation settings for native and English questions. All eight use `xhigh` or `max` reasoning effort; DeepSeek V4 Pro has three attempts per question and the other models have one. GLM 5.3 is served by Z.AI in both conditions.

The native specialist comparisons use three configurations:

- **Afrique.** The specialists use greedy decoding, while their Qwen3.5 counterparts use the Qwen sampling profile below through Hugging Face Inference Endpoints with vLLM 0.29.0.
- **Sunflower Qwen.** These pairs retain their recorded decoding defaults.
- **Gemma E2B.** The pair uses the Gemma profile, a 65,536-token context and identical prompts and input token IDs. Both models run on H200 hardware in BF16 through the same pinned vLLM 0.29.0 image and loader patch.

In each specialist’s paired language comparison, the specialist and its counterpart use the corresponding Qwen or Gemma profile in both language conditions.

Setting	Qwen-based models	Gemma E2B pair
Temperature	1.0	1.0
Top- <i>p</i>	0.95	0.95
Top- <i>k</i>	20	64
Presence penalty	1.5	0
Frequency penalty	0	0
Repetition penalty	1.0	1.0
Output limit	32,768 tokens	32,768 tokens
Seed	42	42
<code>enable_thinking</code>	<code>false</code>	<code>false</code>
Thinking-token budget	0	0

We retain the saved native responses, including outputs that reach the token limit or contain unexpected reasoning text.

D.3 Response grading

GPT-5.6 Luna Pro grades the saved responses using the same criteria as the human reviewers in §A.2. For each response, it receives the question, the expert reference answer and explanation, and the grading rule for that category. It then explains its decision and returns `GRADE: C` or `GRADE: I`, without partial credit.

Reusing the grading prompt. The scorer in `benchmark/judge.py` uses Inspect’s `model_graded_fact` with the model assigned to the grader role. The full template, category rules and output instructions are stored in `benchmark/prompts.py`; the output instructions are fixed there so that they do not change with the installed framework version. The template is sent as a user message with the relevant category guidance substituted. In the specialist English comparison, we grade each response once with an 8,192-token output limit.

Handling empty answers. Empty responses in the reported main evaluation score zero. For the saved specialist runs, we retain the recorded grades, including one empty Sunflower E2B answer that Luna marked correct.

D.4 Accuracy estimation

We first calculate a score for each question, averaging its attempts when more than one response was generated. We then calculate accuracy within each language variety and average the varieties equally, as defined in §4.2. Group means give each model equal weight, while category scores average the language varieties represented in that category.

Estimating uncertainty. For the main native evaluation, we estimate the variance within each variety and combine those variances to obtain the standard error of the overall score. Writing $y_{\ell i}$ for question i ’s score and A_ℓ for the mean of the n_ℓ questions in variety ℓ , this gives

$$V_\ell = \frac{\sum_i (y_{\ell i} - A_\ell)^2}{n_\ell(n_\ell - 1)}, \quad (5)$$

$$\text{SE}(A) = L^{-1} \sqrt{\sum_\ell V_\ell}. \quad (6)$$

For specialist and language comparisons, we keep the responses to the same question paired.

The bootstrap resamples questions within each variety and carries all compared responses together, preserving that pairing in every draw. The interval settings are:

Comparison	95% interval
Main models (native)	Score ± 1.96 SE; retained questions treated as independent within language variety.
Native specialist pairs	5,000 paired within-variety question-bootstrap draws; seed 20261001; pointwise, without multiplicity adjustment.
English comparisons	5,000 paired within-variety question-bootstrap draws; seed 20260928; percentile intervals, keeping all compared models and language conditions together.

These intervals describe variation across the retained questions, conditional on the selected language varieties, models and grading protocol. They do not include uncertainty from grading or translation errors.

E Detailed Results

E.1 Accuracy by language

Table 11 separates model and human accuracy by language variety. Each row corresponds to one language variety, with the number of questions shown alongside the scores. Questions differ across varieties, so cross-variety scores do not isolate a language effect. The model differences within each variety are descriptive; this table does not report significance tests.

Language	n	CLOSED-SOURCE					OPEN-WEIGHT					Δ	Human mean
		GPT-6 Astra	Claude Fable 5.1	Gemini 3.7 Flash	Muse Spark 1.2	Mean	Kimi K3	Qwen3.8 2.4T A95B	DeepSeek V4 Pro	GLM 5.3	Mean		
Amharic	137	84.7	87.6	79.6	85.4	84.3	74.5	68.6	64.7	73.0	70.2	14.1	58.8
Fulani	86	39.5	47.7	46.5	32.6	41.6	15.1	16.3	19.0	15.1	16.4	25.2	54.0
Kinyarwanda	131	64.1	58.0	57.3	47.3	56.7	36.6	42.0	33.3	30.5	35.6	21.1	54.8
Kirundi	88	51.1	44.3	44.3	34.1	43.5	28.4	35.2	27.3	22.7	28.4	15.1	83.6
Lingala	102	42.2	49.0	49.0	43.1	45.8	31.4	35.3	30.1	26.5	30.8	15.0	59.8
Malagasy	91	54.9	44.0	46.2	34.1	44.8	19.8	22.0	17.6	16.5	19.0	25.8	67.3
Shona	129	72.1	69.8	65.1	59.7	66.7	51.2	46.5	51.9	40.3	47.5	19.2	61.7
Swahili Mombasa	128	46.9	47.7	43.8	41.4	44.9	37.5	26.6	32.3	29.7	31.5	13.4	55.3
Swahili Tanzania	72	62.5	56.9	63.9	43.1	56.6	33.3	33.3	24.5	23.6	28.7	27.9	41.7
Wolof	118	67.8	66.9	59.3	49.2	60.8	39.8	46.6	39.3	37.3	40.7	20.1	50.3
AFRINR Score	1,082	58.6	57.2	55.5	47.0	54.6	36.8	37.2	34.0	31.5	34.9	19.7	58.7
SE (pp)		1.49	1.50	1.52	1.49	1.18	1.41	1.43	1.18	1.35	1.09	0.89	—

Table 11: **Accuracy by language.** Accuracy (%) of eight models and native speakers on each language variety’s questions. The human mean averages the speakers of each language variety, 49 in all.

E.2 Language comparison

Table 12 shows how each model’s accuracy changes within each language variety, while Table 13 shows how those changes affect the gap between closed-source and open-weight models. Both tables use the same 531 paired questions.

Model	English minus original-language accuracy (percentage points)										
	Amharic (79)	Fulani (49)	Kinyarwanda (68)	Kirundi (42)	Lingala (57)	Malagasy (43)	Shona (50)	Swahili Tanzania (22)	Swahili Mombasa (62)	Wolof (59)	Overall (531)
<i>Closed-source</i>											
Claude Fable 5.1	-6.3	+24.5	-1.5	+9.5	-1.8	-4.7	-4.0	+4.5	+3.2	+8.5	+3.2
GPT-6 Astra	-5.1	+18.4	+1.5	+7.1	+1.8	+4.7	+8.0	+4.5	-1.6	+11.9	+5.1
Gemini 3.7 Flash	-1.3	+16.3	+1.5	+2.4	-10.5	+4.7	0.0	-13.6	0.0	+11.9	+1.1
Meta Muse Spark 1.2	+2.5	+46.9	+8.8	+23.8	+7.0	0.0	+6.0	0.0	+14.5	+27.1	+13.7
<i>Open-weight</i>											
Qwen3.8 2.4T A95B	+5.1	+61.2	+2.9	+14.3	+7.0	+9.3	0.0	-4.5	+4.8	+35.6	+13.6
Kimi K3	0.0	+57.1	+10.3	+11.9	+14.0	+7.0	+4.0	+18.2	-1.6	+44.1	+16.5
GLM-5.3	+1.3	+53.1	+17.6	+16.7	+7.0	0.0	+6.0	-9.1	+3.2	+35.6	+13.1
DeepSeek V4 Pro	+6.3	+38.1	+10.8	+14.3	+7.0	0.0	+0.7	0.0	+9.1	+35.6	+12.2
<i>Specialist counterparts</i>											
Qwen3.8-27B	+19.0	+44.9	+25.0	+7.1	+31.6	+11.6	+14.0	+13.6	+4.8	+50.8	+22.3
Qwen3.5-9B	+41.8	+36.7	+25.0	+21.4	+35.1	+4.7	+34.0	+18.2	+3.2	+50.8	+27.1
Gemma-4-E2B-it	+17.7	+26.5	+23.5	+16.7	+28.1	+4.7	+28.0	+4.5	+4.8	+47.5	+20.2
Qwen3.5-4B	+50.6	+26.5	+33.8	+31.0	+29.8	+7.0	+36.0	+9.1	+14.5	+57.6	+29.6
<i>African specialists</i>											
Sunflower 27B	+25.3	+24.5	+16.2	+7.1	+12.3	0.0	+2.0	-9.1	+4.8	+33.9	+11.7
Sunflower 9B	+13.9	+10.2	+11.8	0.0	+3.5	0.0	+2.0	+4.5	-3.2	+35.6	+7.8
Sunflower E2B	+1.3	-2.0	+4.4	0.0	+5.3	+9.3	+2.0	+13.6	-4.8	+16.9	+4.6
Afrique 9B 50Langs	+24.1	+24.5	+7.4	+23.8	+5.3	+4.7	+2.0	+13.6	+14.5	+50.8	+17.1
Afrique 4B 50Langs	+19.0	+20.4	+20.6	+14.3	+14.0	+4.7	+20.0	+4.5	+8.1	+45.8	+17.1
Afrique 4B v1	+34.2	+34.7	+10.3	+9.5	+24.6	+9.3	+16.0	+4.5	+12.9	+44.1	+20.0
Afrique 4B v2	+32.9	+24.5	+13.2	+21.4	+15.8	+4.7	+12.0	+9.1	+1.6	+40.7	+17.6
Eight main models: mean	+0.3	+39.5	+6.5	+12.5	+3.9	+2.6	+2.6	0.0	+4.0	+26.3	+9.8
Models improving	4/8	8/8	7/8	8/8	6/8	4/8	5/8	3/8	5/8	8/8	8/8
All 19 models: mean	+14.9	+30.9	+12.8	+13.3	+12.5	+4.3	+9.9	+4.5	+4.9	+36.0	+14.4
Models improving	15/19	18/19	18/19	17/19	17/19	13/19	16/19	13/19	14/19	19/19	19/19

Table 12: **English advantage by language for every model in Figure 5.** Positive values indicate higher accuracy in English; negative values (shaded) indicate higher accuracy in the original language; zero indicates a tie. Bold values mark the largest language gap within each model or mean row. Parentheses give paired question counts; Swahili Tanzania and Swahili Mombasa are scored separately. Overall is the equally weighted mean across ten independent language varieties, and the summary means weight distinct models equally. All rows use the same 531 cultural-practice and culturally grounded mathematical questions, including languages outside a specialist’s documented coverage (Table 10). Counterparts repeated in the figure appear once here. Improvement counts exclude ties; these are descriptive differences without significance tests.

Questions	Closed gain [95% CI]	Open gain [95% CI]	Original gap	English gap	Narrowing [95% CI]
531	+5.8 [+3.4, +8.1]	+13.9 [+11.2, +16.5]	17.2	9.1	+8.1 [+5.3, +10.7]

Table 13: **English gains and gap narrowing on 531 paired questions.** Accuracy differences are in percentage points. Gains are English minus original-language accuracy, gaps are closed-source minus open-weight accuracy, and narrowing is the original gap minus the English gap. Each model’s accuracy weights the ten retained language varieties equally, with Swahili Tanzania and Swahili Mombasa scored independently; group means weight models equally. Brackets give paired question-bootstrap 95% intervals.

E.3 Specialisation by category

Figure 7 breaks down each specialist comparison by category. Within a pair, both models answer the same questions; the plotted change shows where the specialist scores higher or lower than its counterpart.

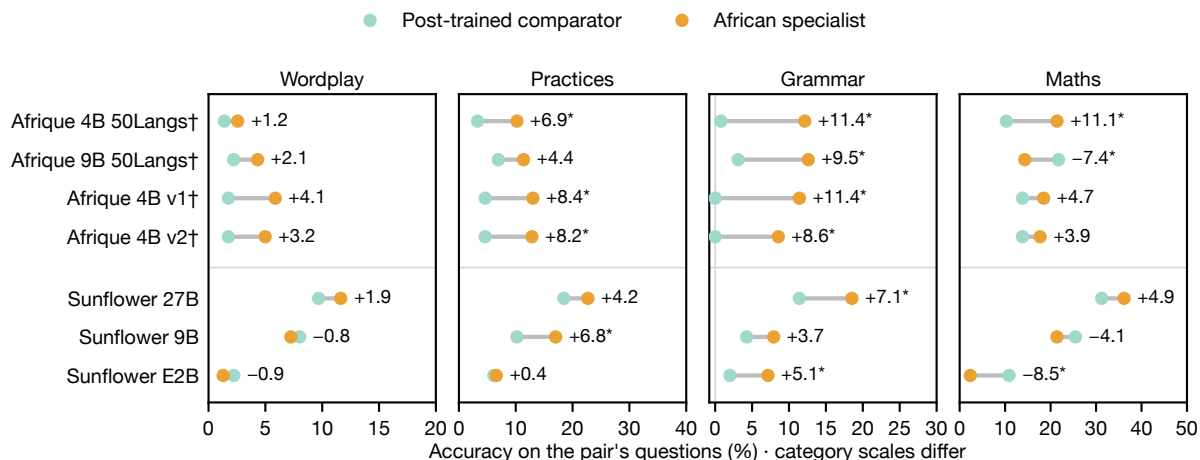


Figure 7: **Specialisation gains concentrate in Grammar and Practices.** Accuracy of each specialist (amber) and its post-trained comparator (teal) on the pair’s questions, by category; labels give the change in percentage points, starred where the paired pointwise 95% bootstrap interval excludes zero, without adjustment for multiple comparisons. Each category uses its own horizontal scale starting at zero. † marks pairs with thinking disabled in both arms but different sampling and serving settings: the specialist uses greedy decoding and its comparator uses sampling.

F Author Contributions

- **Enrique Molero:** Project lead and primary author; research direction, main experiments and model baselines, human-baseline platform and recruitment guidelines, freelance expert recruitment and management, and coordination of dataset review and feedback to Digital Umuganda.
- **Louis Powell:** Assisted research direction.
- **Israel Abebe Azime:** Assisted research direction and writing.
- **Linda Petrini:** Assisted writing.
- **Thomas Summers:** Developed the question-generation guidelines.
- **Audace Niyonkuru, Samuel Rutunda and Mugisha Boris Ishimwe (Digital Umuganda):** Coordinated data collection, including recruitment, management and payment of question writers, and recruited and managed baseliners using Enrique Molero’s guidelines.