NILEAGI-SUB: An Evaluation Programme for Swahili Language Understanding

Zephania Reuben[1] Isack Odero[1] Joseph Mnyune[1]

[1]NileAGI

Technical report · July 2026 · Benchmark paper

Abstract

Swahili is spoken by over 200 million people across East and Central Africa, in classrooms, clinics, banks, farms, and stadiums alike, yet language models are rarely tested on how well they actually understand it in any of these settings. Most multilingual benchmarks treat Swahili as one line in a large table of languages, rather than testing it in the sectors where people actually use it. We introduce NILEAGI-SUB, an evaluation programme built to measure Swahili language understanding across real-world sectors, including health, finance, sports, and education, evaluated one sector at a time under the same controlled protocol. This report covers our first sector, education, and releases its dataset, SUB-MCQ Education: 2,569 expert-reviewed Swahili multiple-choice items drawn from the Swahili-subject curriculum, spanning Standard 3–7 and Form 2–4. We evaluate eight open models with 2–5 billion parameters on identical items under deterministic decoding and unified scoring. Gemma 4 E4B attains the highest accuracy, 54.8%, followed by AfriqueQwen3.5-4B at 50.3%; this gap is unlikely to be due to chance. AfriqueQwen3.5-4B is strongest on Standard 3 and Standard 5, while Gemma 4 E4B leads on the remaining six grade bands. Tiny Aya Earth and Tiny Aya Global, two regional variants of the same base model, differ by only 0.7 points (41.4% vs. 40.7%), a gap that is plausibly chance. Across the eight models, accuracy spans roughly 27 percentage points despite a common 2–5B parameter budget, indicating that architecture, continued pretraining, and prompt compatibility account for more of the observed variance than parameter count in this range. We discuss single-annotator and contamination risks that qualify these results, and are explicit that this first report covers one sector (education) at one model scale. SUB-MCQ Education, all model outputs, and the evaluation code are released to support independent verification and extension to further sectors and model scales.

Keywords: Swahili; Kiswahili; multilingual language models; African NLP; open language models; educational assessment; multiple-choice evaluation; reproducible benchmarking.

1 Introduction

Swahili is a working language across East and Central Africa: it is used in classrooms, hospitals, banks, farms, courts, newsrooms, and sports commentary, not only in translated text. Language models are increasingly deployed in all of these settings, yet most Swahili evaluation treats the language as a single undifferentiated skill, tested with one general-purpose benchmark rather than with the vocabulary, reasoning, and register specific to a given sector. A model that handles conversational Swahili well is not thereby shown to handle Swahili medical terminology, financial

language, or school curriculum content correctly. This report is the first in a programme intended to close that gap sector by sector, starting with education, where curriculum-aligned assessment materials are relatively accessible and grading is unambiguous.

Within education specifically, no standardized benchmark exists for measuring school-level Swahili understanding across open models under controlled, comparable conditions. This makes it difficult to compare systems, or to determine which models are suitable for educational applications in East Africa. Fluency in generated Swahili is not a sufficient proxy: a model can produce plausible text while failing on curriculum-aligned vocabulary, factual knowledge, or grade-appropriate reasoning [1, 2].

The gap is not simply one of data scarcity. Multilingual benchmarks such as XTREME, FLORES, BELEBELE, and the MMLU family are valuable for cross-lingual transfer and reading comprehension, but none provides a grade-stratified Swahili educational signal under a protocol designed for models at consumer- or edge-feasible scale [2, 4, 5, 6]. Where comparisons across Swahili-capable models do appear, they frequently mix incompatible prompts, decoding settings, or parameter scales, which makes it difficult to attribute an observed gain to language ability, model capacity, or evaluation design. Community efforts such as MasakhaNER demonstrate that language-specific evaluation surfaces phenomena that global aggregates miss [7], but we are not aware of a matched, grade-stratified instrument for generative Swahili understanding at the time of writing. This claim reflects our own literature search rather than a systematic survey, and we would welcome pointers to prior work we may have missed.

We focus first on models in the approximately 2–5B parameter range because they are the most plausible candidates for local inference, constrained hardware, and institution-specific deployment: the conditions under which Swahili-language systems are most likely to be adopted in practice. Absent a controlled comparison, practitioners are left to choose among models on the basis of marketing claims rather than measured evidence.

This report addresses the following question: how well do open, or openly accessible, multilingual language models answer school-level multiple-choice questions written in Swahili, when evaluated on identical items under deterministic decoding and a unified scoring protocol?

To answer it, we introduce NILEAGI-SUB (the NileAGI Swahili Understanding Benchmark) and release SUB-MCQ Education (the Swahili School-Level Multiple-Choice Understanding Dataset) as its first sector dataset. NILEAGI-SUB is the evaluation programme; SUB-MCQ Education is the education item set released under it in this report. We name the programme NILEAGI-SUB rather than tying it to this single release, because we intend, as future work, to extend the same construction and evaluation methodology to other sectors such as health, finance, and sports (Section 9), and because the parameter-scale roadmap in Section 7 already extends beyond the models evaluated here. Everything reported in this paper, however, concerns one sector, education, at one parameter scale. Our contributions are as follows.

1. **A shared measurement instrument.** SUB-MCQ Education is a large, grade-stratified Swahili multiple-choice item set, drawn from the Swahili-subject curriculum, designed for controlled comparison of open multilingual models rather than for ad hoc probing.

2. **A controlled evaluation of eight open models.** We evaluate eight models in the approximately 2–5B parameter range on identical items, under deterministic decoding and unified answer extraction, which yields a reproducible ranking rather than a set of individually re-

ported, incomparable claims.

3. **Evidence against a parameter-count heuristic.** Overall and grade-stratified results and paired significance tests show that architecture, adaptation strategy, and prompt compatibility can outweigh modest parameter differences within a fixed scale band.

4. **A five-level parameter-scale roadmap for this sector.** This report completes Level I of a five-level programme for the education dataset; Level II and Level III are scoped, and Level IV and Level V are reserved for future work at frontier and ultra-scale. Extending the same methodology to sectors beyond education is future work, discussed but not attempted in this report.

2 Related Work

2.1 Multilingual benchmarks and their limitations for this setting

Cross-lingual suites such as XTREME and XTREME-R measure transfer across many languages and tasks [2, 3]. FLORES targets translation quality [4], and BELEBELE targets multilingual reading comprehension [5]. Knowledge-intensive evaluations in the MMLU family probe broad subject competence, typically in English-centric form [6]. These resources are valuable, but none directly measures school-level Swahili educational understanding for open models at this scale: they under-represent Swahili curriculum structure, target different skills, or do not control for the deployment constraints that apply at smaller parameter counts.

2.2 African-language evaluation

African NLP has advanced substantially through community-led datasets and shared tasks [7]. This work shows that local linguistic expertise and language-specific annotation surface phenomena that global leaderboards do not capture. For generative models specifically, the evaluation gap is acute: fluent output can mask weak curriculum-aligned understanding. NILEAGI-SUB extends this line of work with a deterministic, grade-aligned Swahili understanding benchmark, beginning at the parameter range most relevant to local deployment (2–5B) and, in this report, at a single sector: education.

2.3 Why school-level multiple choice

School questions impose an interpretable progression of vocabulary, reading demand, and subject knowledge, and reflect a practical use case: students, teachers, and educational technology providers require systems that handle instructional Swahili correctly. Multiple-choice format enables exact scoring without a second model acting as judge, at the cost of not measuring free-form explanation quality (a limitation we return to in Section 8).

2.4 Rationale for the Level I parameter range

Three considerations motivate the initial scope of Level I. First, local deployability favors smaller footprints. Second, open weights or openly accessible releases support reproducibility and adaptation. Third, restricting the first comparison to a narrow parameter band reduces the capacity

confound that otherwise dominates cross-model claims. Our goal is not to identify a single best multilingual model, but to establish a trustworthy baseline that subsequent levels of the programme extend to larger model classes and additional sectors.

3 Dataset: SUB-MCQ Education

SUB-MCQ Education draws its items from the Swahili-subject curriculum specifically, that is, the language-arts subject in which Tanzanian and Kenyan students are examined on Swahili grammar, comprehension, vocabulary, and literature, rather than from other subjects (mathematics, science, and so on) that happen to be taught in Swahili medium. This distinction matters for how the results should be read: strong performance on SUB-MCQ Education indicates a model handles the Swahili-subject curriculum well, not that it can reason correctly about other school subjects when those are posed in Swahili.

Table 1: Validated SUB-MCQ Education items by school grade.

Grade band	Items	Share (%)
Standard 3	66	2.6
Standard 4	116	4.5
Standard 5	274	10.7
Standard 6	651	25.3
Standard 7	517	20.1
Form 2	399	15.5
Form 3	125	4.9
Form 4	421	16.4
Total	2,569	100.0

3.1 Construction and validation

The source collection contains 2,608 records, each with a Swahili question (*swali*), a school-grade field (*hatua*), answer options (*machaguo*), and an answer key (*jibu*). A deterministic loader maps grade labels, normalizes option prefixes, rejects empty or invalid options, requires at least two answer choices, discards records with an ambiguous or unmappable grade label, and retains only answer keys that refer to an available option. After validation, 2,569 items remain across eight grade bands, from Standard 3 to Form 4; stable sequential IDs are assigned where source IDs are absent.

All items were labeled and reviewed for grade placement by a single qualified Swahili language expert, Khalid Mwinge, providing a linguistic quality check ahead of the automated structural validation described above. We use "validated" in this report to mean that an item has passed this structural check and single-expert grade review, not that it has passed independent double annotation or answer-key adjudication by a second subject specialist; that stronger form of validation has not yet been performed and is discussed as a limitation in Section 8.

3.2 Grade structure and option counts

Table 1 shows a deliberate imbalance: Standard 6 and Standard 7 together account for nearly half of all items, while the earliest primary bands are comparatively small. Overall accuracy is therefore dominated by the larger bands, and grade-level percentages carry unequal statistical precision, a point we return to in Section 5.2. The validated set contains 1,408 five-option items, 1,125 four-option items, 31 three-option items, and 5 two-option items; 98.6% of items have four or five options.

Because items have different numbers of options, we need a baseline for what a model with no Swahili understanding at all would score, simply by picking an answer at random. Without this baseline, a raw accuracy number is hard to interpret: 40% sounds low in absolute terms, but it is meaningfully above chance if most items offer five options (where random guessing scores only 20%), and less impressive if most items offer only two. We compute this baseline per item, as one divided by that item's number of options, and then average across all items, rather than using a single fixed guess rate such as 1/4 or 1/5 for every item. This gives the expected accuracy of uniform random guessing over each item's observed options:

$$A_{\text{chance}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K_i} = 22.41\%, \tag{1}$$

where N is the total number of items and K_i is the number of options for item i. We use this exact, mixed-option expectation as the chance reference in Figure 1: any model scoring at or below 22.4% is not doing better than blind guessing, so all further comparisons in this report are read relative to that floor.

3.3 Quality and release posture

SUB-MCQ Education is sufficiently validated to support a reproducible baseline, but it is not yet a fully audited national examination corpus. We recommend that public release be preceded by duplicate detection, source and license documentation, further educator review, answer-key verification against additional annotators, topic annotation, and a formal train/development/test split. Without these steps, contamination and ambiguous items could distort future comparisons.

4 Models and Evaluation Protocol

4.1 Model set

Table 2 summarizes the eight models evaluated at this parameter scale. We use "open" as a practical umbrella for open-weight or openly accessible research releases; licenses differ across models and must be checked independently before deployment. Tiny Aya is accessed through Cohere's API; the remaining models are evaluated from Hugging Face weights. The set deliberately combines strong general-purpose multilingual checkpoints, African-language continued-pretraining (CPT) variants, and regional versus global Tiny Aya siblings [8, 9, 10, 11, 12, 13], so that the comparison spans the principal adaptation strategies currently used for African-language deployment at this scale.

Table 2: Models evaluated in Level I. "Access" describes the evaluation path used here, not a claim that all licenses are identical.

Model	Params.	Access	Evaluation note
Gemma 4 E4B IT	4.5B	Hugging Face	Instruction/chat prompt
AfriqueQwen3.5-4B	4.0B	Hugging Face	African-language CPT; completion prompt
AfriqueGemma-4B	4.0B	Hugging Face	African-language CPT; completion prompt
Qwen3.5-4B	4.0B	Hugging Face	Chat prompt; thinking disabled
Tiny Aya Earth	3.35B	Cohere API	Regional variant (West Asia, Africa)
Tiny Aya Global	3.35B	Cohere API	Balanced release across all regions
Llama 3.2 3B Instruct	3.2B	Hugging Face	Instruction/chat prompt
Qwen3.5-2B	2.0B	Hugging Face	Chat prompt; thinking disabled

4.2 Prompting and decoding

Instruction-tuned models receive a Swahili system instruction that requests a single answer letter (A–E), followed by the question, labeled options, and optional passage context. Qwen thinking mode is disabled so that hidden reasoning traces neither consume the output budget nor complicate answer extraction.

The Afrique checkpoints [11] are CPT base models rather than instruction-tuned chat checkpoints; chat templates produced unreliable free-form continuations for these models in preliminary testing. They therefore receive a completion-format prompt with one Swahili demonstration, with the target item ending in `Jibu:`. Model- compatible prompting is a precondition for valid measurement; as a result, comparisons between Afrique and non-Afrique models should be read as outcomes under matched evaluation conditions for each model family, not as a controlled estimate of the causal effect of continued pretraining in isolation (Section 5.1).

To illustrate the two prompt formats concretely (the item below is a constructed illustration, not a SUB-MCQ Education item, to avoid reproducing copyrighted or answer-sensitive content): instruction-tuned models receive a Swahili system message such as "Jibu kwa herufi moja tu (A, B, C, D, au E)", followed by the question and labeled options, and are expected to return a single letter. The Afrique completion-format prompt instead gives one worked Swahili demonstration question with its answer, followed by the target question ending in the bare cue `Jibu:`, so that the model's next generated token is the answer letter rather than a continuation of free-form chat.

Locally hosted models use greedy decoding; Cohere requests use temperature zero, matching the deterministic decoding used elsewhere in the protocol. The scorer strips `<think>` blocks, accepts direct letters and explicit answer prefixes, and rejects ambiguous outputs that reproduce the full option list without committing to an answer. An exact letter match to the validated key counts as correct; missing or unparseable outputs count as incorrect. Aggregate results are saved after each model completes, with item-level checkpoints every 50 questions to support interruption-safe resumption.

4.3 Compute infrastructure

The six locally hosted open-weight models (all models in Table 2 except Tiny Aya Earth and Tiny Aya Global, which were accessed through Cohere's API) were run on a single rented GPU instance: one NVIDIA RTX A5000 (24 GB VRAM), driver version 535.309.01, CUDA 12.2. A single 24 GB consumer/workstation-class GPU is sufficient to serve every locally hosted checkpoint in this report sequentially, consistent with our framing of Level I as the locally feasible parameter range (Section 3); it is not sufficient, on its own, for the larger checkpoints planned in Level II and beyond (Section 7), which will require either a larger single device, multi-GPU sharding, or continued reliance on hosted APIs.

4.4 Metrics

Our primary metric is exact-match accuracy: the fraction of items where the model's extracted answer letter matches the validated key exactly.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i]\,. \tag{2}$$

Here $\hat{y}_i$ is the model's answer on item i, y_i is the correct answer, and $\mathbb{1}[\cdot]$ is the indicator function: it is 1 when the model's answer matches the key and 0 otherwise. Summing this indicator over all N items and dividing by N is simply the proportion of items answered correctly, written in standard mathematical notation.

A single accuracy percentage does not say how much that percentage might shift if we had drawn a different, similarly difficult set of test items. For this we report a 95% Wilson score interval alongside each accuracy figure: a range of values, computed from the accuracy and the number of items, that is likely to contain the model's true accuracy on the wider population of similar Swahili-subject questions, not just this particular sample of 2,569. A narrower interval means the estimate is more precise; Wilson intervals are the standard choice for proportions like accuracy because, unlike a simple ± margin, they stay within the valid 0–100% range and remain reliable even when accuracy is close to 0% or 100% or the sample is not very large. We report Wilson intervals in Table 3 so that a reader can judge, at a glance, whether two models' accuracy ranges overlap or are clearly separated.

5 Results

5.1 Overall ranking

Table 3 and Figure 1 rank all eight systems by exact-match accuracy under controlled conditions. Gemma 4 E4B attains the highest accuracy, 54.8%, 4.5 percentage points above AfriqueQwen3.5-4B (50.3%). Of the 2,569 items, Gemma 4 E4B is uniquely correct on 454 and AfriqueQwen3.5-4B is uniquely correct on 339; an exact McNemar test on these discordant outcomes rejects the null hypothesis of equal accuracy ($p < 0.001$), indicating that the gap reflects a systematic difference between the two models rather than sampling variance.

The spread across the full model set is larger than the gap between the top two models. All eight models exceed the 22.4% chance baseline, yet the highest- and lowest-scoring models differ by

Table 3: Overall exact-match accuracy on 2,569 items. Confidence intervals are 95% Wilson intervals.

Rank	Model	Correct	Accuracy	95% CI
1	Gemma 4 E4B	1,407	**54.8**	[52.8, 56.7]
2	AfriqueQwen3.5-4B	1,292	50.3	[48.3, 52.2]
3	AfriqueGemma-4B	1,116	43.4	[41.5, 45.3]
4	Tiny Aya Earth	1,063	41.4	[39.5, 43.3]
5	Tiny Aya Global	1,045	40.7	[38.8, 42.6]
6	Qwen3.5-4B	909	35.4	[33.5, 37.2]
7	Qwen3.5-2B	748	29.1	[27.4, 30.9]
8	Llama 3.2 3B	715	27.8	[26.1, 29.6]

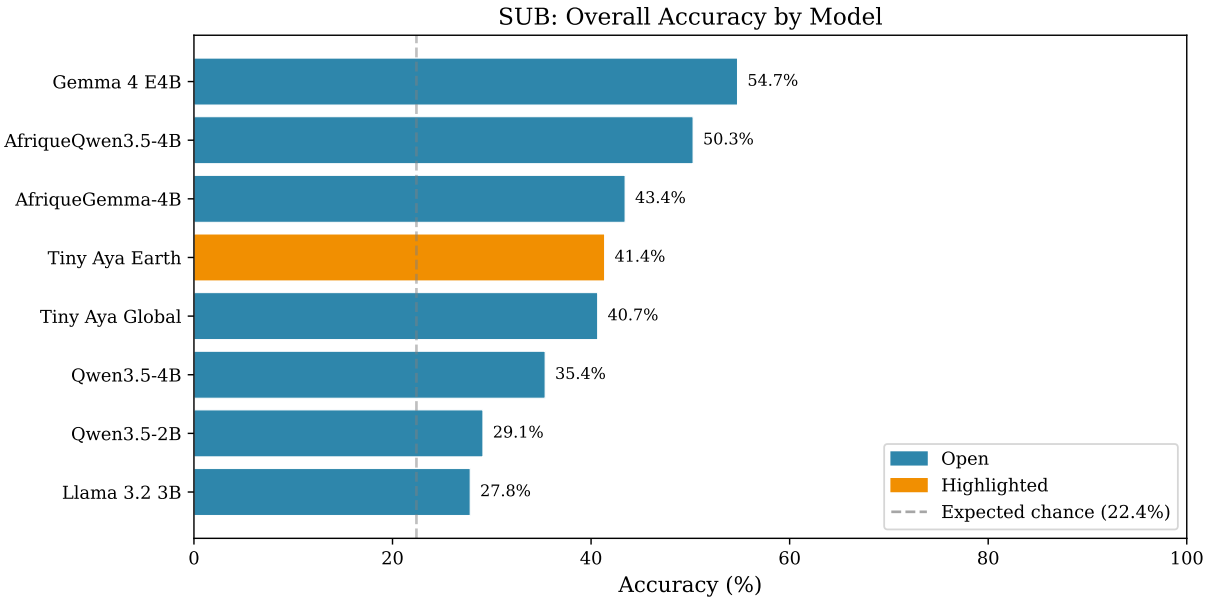

Figure 1: Overall exact-match accuracy. The dashed line marks the 22.4% expected accuracy of uniform random guessing over each item's observed number of choices. Tiny Aya Earth is highlighted as the West Asia/Africa regional variant.

26.9 percentage points[1] despite sharing a common 2–5B parameter budget. This indicates that parameter count alone does not determine accuracy within this band, and that selecting a model for Swahili deployment on the basis of reported size rather than measured accuracy is unreliable.

AfriqueQwen3.5-4B outperforms the similarly sized Qwen3.5-4B checkpoint by 14.9 percentage points. We do not attribute this gap solely to African-language CPT, because the two checkpoints also require different prompt formats (Section 4.2); pretraining data and prompt format are confounded in this comparison. Isolating their individual contributions requires a controlled ablation over matched base, CPT, and instruction-tuned checkpoints, which we identify as a priority for Level II.

Because all models answer the same 2,569 items, we do not treat each model's accuracy as fully independent of the others when comparing two models directly. Instead, for a chosen pair of models, we use a paired exact McNemar test: we look only at the items where the two models disagree (one got it right and the other wrong), and ask whether one model wins those

[1] Computed from unrounded per-model accuracies; subtracting the one-decimal figures in Table 3 directly (54.8 − 27.8) gives 27.0, a discrepancy of 0.1 points from rounding that does not affect any conclusion in this report.

disagreements more often than the other more than we would expect from a fair coin flip. This is a more targeted test than comparing two accuracy percentages in isolation, because it isolates exactly the items that could have gone either way, rather than being diluted by the large number of items both models answer the same way. A small p-value means the observed imbalance in who wins the disagreements is unlikely to be due to chance alone. These tests are descriptive: hypotheses were not preregistered, and we do not apply a correction for the full set of possible pairwise comparisons.

5.2 Grade-level behavior

Table 4: Accuracy (%) by school grade. Bold indicates the highest score in each column.

Model	Std3	Std4	Std5	Std6	Std7	Form2	Form3	Form4
Gemma 4 E4B	65	**68**	55	**50**	**55**	**62**	**55**	**50**
AfriqueQwen3.5-4B	**71**	60	**59**	46	51	54	51	41
AfriqueGemma-4B	65	49	48	40	43	49	45	35
Tiny Aya Earth	48	42	46	34	42	46	51	40
Tiny Aya Global	58	45	44	32	42	47	54	37
Qwen3.5-4B	42	41	38	29	33	44	42	34
Qwen3.5-2B	33	28	29	26	26	36	34	29
Llama 3.2 3B	35	30	29	26	26	31	30	27

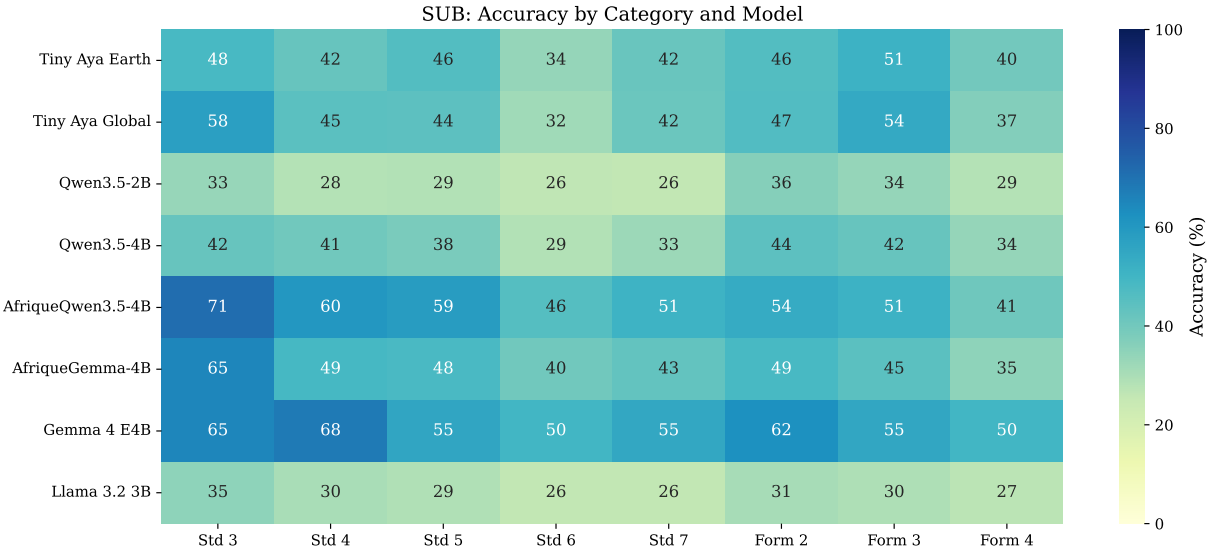

Figure 2: Accuracy by school grade and model. Grade bands are ordered from Standard 3 to Form 4.

Aggregate accuracy can obscure grade-specific strengths. Figure 2 and Table 4 report accuracy separately for each of the eight grade bands. AfriqueQwen3.5-4B attains the highest accuracy on Standard 3 (71%) and Standard 5 (59%); Gemma 4 E4B leads on the remaining six bands, with the largest margins in Form 2 and Form 4. These per-band comparisons carry unequal statistical power: Standard 3 contains 66 items, roughly a tenth of the 651 items in Standard 6, so rankings within the smaller bands should be interpreted with wider uncertainty than the overall ranking in Table 3. Concretely, AfriqueQwen3.5-4B's 71%-vs-65% lead over Gemma 4 E4B on Standard 3 corresponds to a margin of roughly four items out of 66; a different sample of similarly difficult

Standard 3 questions could plausibly reverse this particular ranking, even though the overall ranking in Table 3 is well powered.

5.3 Regional specialization is not automatic

Tiny Aya Earth is one of three regional variants released alongside Tiny Aya Global, and is described by its publisher as tuned for West Asian and African languages together, not for African languages exclusively; we did not create this pairing, and readers should not assume Earth was optimized for Swahili specifically. Global is described as the balanced release across all supported regions. Earth attains 41.4% accuracy and Global 40.7%; Earth is uniquely correct on 130 discordant items and Global on 112, and an exact McNemar test does not reject the null hypothesis of equal accuracy ($p = 0.274$). At this sample size, we find no evidence that the West Asia/Africa-tuned variant confers an overall advantage on SUB-MCQ Education for Swahili specifically. The two variants do diverge at the grade level: Global outperforms Earth on Standard 3, Standard 4, Form 2, and Form 3, while Earth outperforms Global on Standard 5, Standard 6, and Form 4; the two are tied on Standard 7. This pattern indicates that a model's stated regional tuning does not, by itself, predict its accuracy on one particular language within that region, and that claims of language-specific benefit from regional tuning warrant direct, language-specific evaluation.

5.4 Parameter count is not a reliable predictor

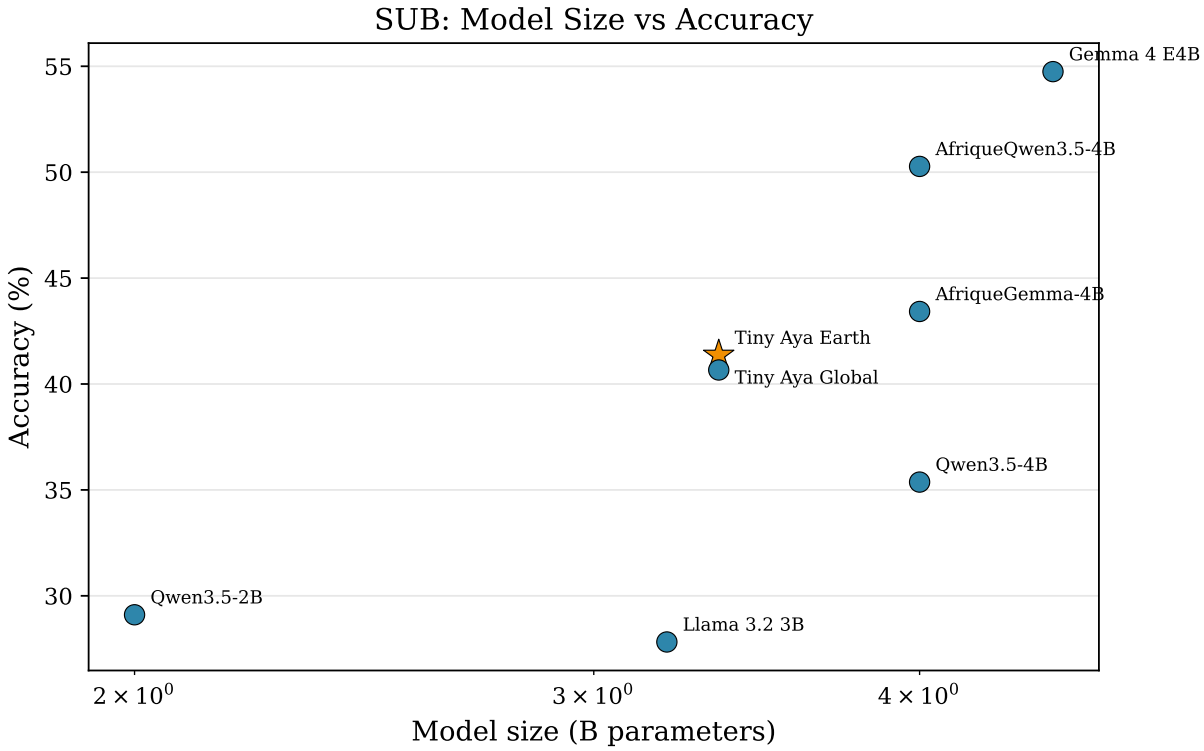

Figure 3: Reported model size versus overall accuracy. Within this narrow parameter range, similarly sized models exhibit large performance differences.

Figure 3 plots reported parameter count against accuracy. Models with approximately 4B parameters span 35.4% to 50.3% accuracy, a 14.9-point range at a fixed nominal size, and the 4.5B Gemma checkpoint reaches 54.8%, exceeding several smaller and similarly sized models in the

set. Architecture, pretraining mixture, instruction tuning, tokenization, and prompt compatibility are all plausible contributors to this variance, and we cannot isolate their individual effects from a fixed set of released checkpoints with heterogeneous training recipes. Because parameter counts are provider-reported and architectures differ across models, we treat Figure 3 as descriptive evidence against a simple parameter-count heuristic, rather than as a controlled scaling-law analysis.

5.5 Top-model profiles

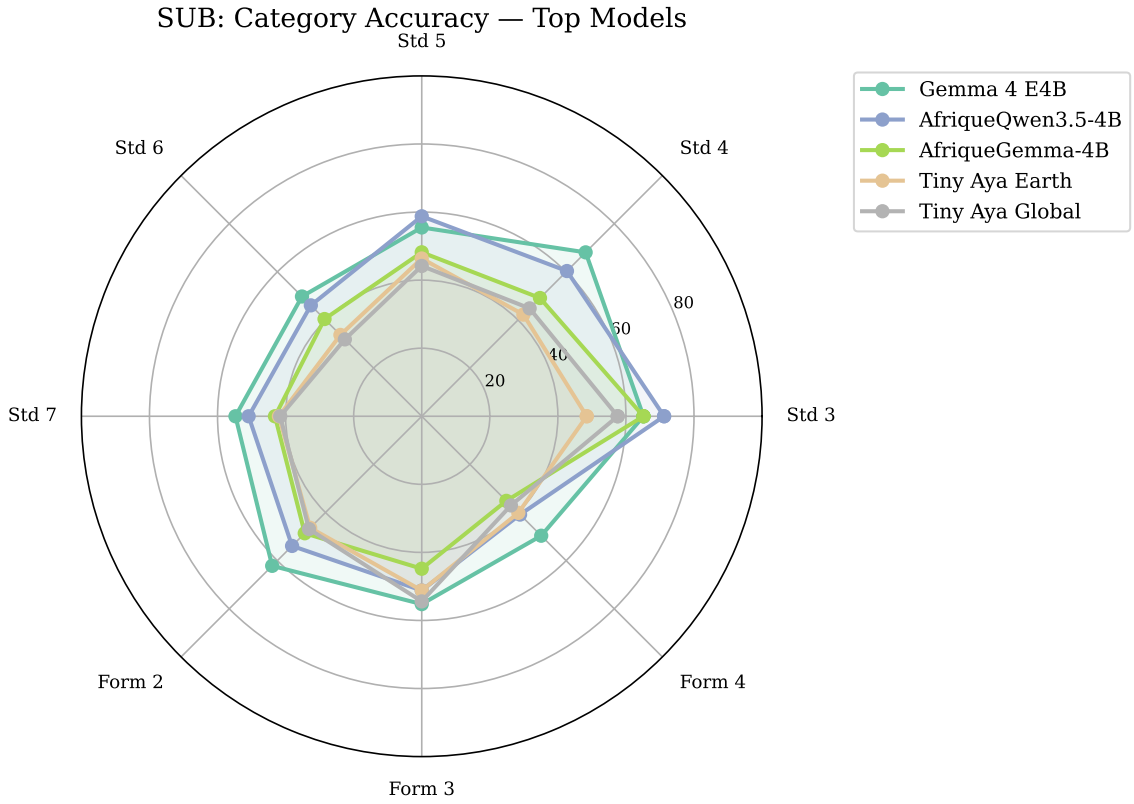

Figure 4: Per-grade profiles for the five highest-scoring models. The radar chart is descriptive; the heatmap should be preferred for precise comparisons.

Figure 4 shows per-grade accuracy profiles for the five highest-scoring models. No single model dominates every grade band by a uniform margin, but Gemma 4 E4B and AfriqueQwen3.5-4B maintain higher accuracy across most bands than the remaining three models. We use this figure to communicate overall shape; exact per-band values are reported in Table 4.

6 Discussion

6.1 Practical implications for model development

These results indicate that, within the 2–5B parameter range, architecture and adaptation strategy account for more of the observed variance in Swahili accuracy than parameter count does. A

4B AfriqueQwen3.5 checkpoint substantially outperforms its 4B Qwen3.5 counterpart; a general-purpose multilingual model, Gemma 4 E4B, outranks a regionally adapted competitor; and Tiny Aya Earth, tuned for West Asian and African languages, does not outperform the balanced Tiny Aya Global release on this evaluation.

We draw three implications for model development.

1. African-language continued pretraining shows a clear empirical benefit in this comparison, but its contribution should be isolated with controlled ablations rather than assumed from training narratives.
2. Prompt compatibility is a component of measured model quality, not incidental to it: a checkpoint evaluated under a mismatched interface will be systematically underestimated.
3. Claims of regional specialization warrant evaluation on language-specific, task-specific benchmarks rather than inference from training data composition alone.

6.2 What these scores do and do not tell us

This report is a model evaluation exercise: it measures and compares how accurately open language models answer Swahili-subject questions, and it is not a study of, or a recommendation for, educational technology deployment, teacher replacement, or curriculum policy. With that scope in mind, the highest-accuracy model in this evaluation still answers approximately 45% of items incorrectly, so the results say only that current 2–5B Swahili models vary widely in Swahili-subject question-answering accuracy, not that any of them is ready for, or intended for, use in a classroom. SUB-MCQ Education supports model selection and error analysis for researchers and model developers; it does not certify any system for educational use, and we make no claim about what these numbers imply for how education should be delivered.

6.3 Contamination is a live threat to these rankings, not a footnote

We have not established whether any evaluated model's training corpus contains the source SUB-MCQ Education questions or close paraphrases of them, and we did not have the resources at this stage to run a contamination detection pass (e.g., membership-inference probes or exact/near-duplicate search against public web crawls). This matters more than a routine limitation: the central claim of this report is that architecture and adaptation strategy, not parameter count, explain the observed ranking. If, for example, Gemma 4 E4B's pretraining corpus happened to include more of these specific Tanzanian or Kenyan educational materials than the other checkpoints', part of its lead could reflect memorization rather than Swahili understanding, and the ranking would not generalize to unseen items of the same type. We flag this explicitly so that the ranking in Table 3 is read as provisional pending a private, held-out test set (Section 9), rather than as a settled leaderboard.

6.4 Reproducibility as a design constraint

Smaller open models lower the barrier to independent replication, but reproducibility still depends on documented licenses, hardware, software versions, and API stability. We publish exact model identifiers, prompt templates, item-level predictions, and the interruption-safe evaluation harness

used to produce these results, so that reported numbers can be independently verified rather than taken on trust.

7 Evaluation-Level Roadmap

Comparing models across widely different parameter counts risks conflating raw capacity with language-specific ability. NILEAGI-SUB therefore organizes evaluation into five model-scale levels, so that comparative claims are made within, rather than across, peer classes. Level II and Level III are scoped for the education dataset reported here; this report completes Level I. Level IV and Level V are future work and will require substantially greater compute and model access. We do not yet have a sector beyond education to report, so this roadmap should be read as a parameter-scale plan for SUB-MCQ Education specifically; extending the same construction methodology to other sectors is discussed as future work (Section 9) but has not been started.

Mixture-of-experts models complicate level placement, because active and total parameter counts capture different aspects of capacity and inference cost. Future levels should report both figures and fix a placement rule in advance of evaluation, to avoid post hoc rationalization of level boundaries. We treat within-level comparisons as the primary basis for claims; cross-level plots are useful for efficiency and Pareto analysis, but should not be used to discount findings observed within a level.

Table 5: Proposed five-level NILEAGI-SUB parameter-scale roadmap for the education dataset. Example models are illustrative of each size class at the time of writing, not a committed evaluation list, and will shift as new open models are released. Boundaries are operational guides and may be refined for mixture-of-experts architectures.

Level	Approx. class	Example models (illustrative)	Status	Purpose
I	Foundational (2–5B)	Gemma 4 E4B, Qwen3.5-4B, Tiny Aya, Llama 3.2 3B	Completed	Open baselines at locally feasible scale; focus of this paper
II	Mid-size (6–15B)	Llama 3.1 8B, Gemma 4 12B, Qwen3 14B	Scoped, not started	Test whether broader capacity improves grade robustness
III	Large (16–40B active)	Qwen3 32B, small MoE models	Scoped, not started	Compare high-capability open multilingual systems
IV	Frontier (41–100B active)	Llama 3.1 70B class, large MoE	Not started	Evaluate advanced reasoning under matched protocols
V	Ultra-scale (large MoE)	Frontier-scale open releases as they emerge	Not started	Study frontier multilingual transfer and efficiency

8 Limitations

1. **Single-annotator verification.** Items were labeled and grade-reviewed by one Swahili language expert, Khalid Mwinge, and have passed automated structural validation, but have not undergone independent double annotation or answer-key adjudication by a second subject specialist. We use "validated" throughout this report in that narrower sense.

2. **Potential contamination.** We have not established whether any evaluated model's training corpus contains the source questions or close paraphrases of them, and have not run a contamination-detection pass. This is discussed at length in Section 6.3, since it could partly explain the observed ranking rather than reflecting genuine Swahili understanding. A private held-out test set is needed for stronger claims.

3. **Dataset provenance and licensing.** A complete source and license audit is required before public redistribution.

4. **Grade imbalance.** Per-grade sample sizes vary from 66 to 651, producing unequal statistical precision and weighting the overall score toward larger bands.

5. **Multiple-choice scope.** Exact-choice accuracy does not measure explanation quality, dialogue, generation, safety, or pedagogical usefulness.
6. **Prompt differences.** CPT base models and instruction-tuned chat models require different prompt formats. This improves construct validity for each model individually, but limits causal claims about the effect of adaptation.
7. **Single-run decoding.** Greedy decoding improves reproducibility but does not measure sensitivity to prompt wording, option order, or stochastic generation.
8. **Model access and versions.** Hosted APIs and model revisions can change over time. Results apply to the identifiers and execution period reported here.
9. **No human baseline.** Models are not yet compared against students, teachers, or subject experts.
10. **Single sector, single level evaluated.** Only the education sector at Level I is completed; findings should not be generalized to other sectors or larger model classes without further evaluation.
11. **Funder overlap with an evaluated model provider.** This work was funded through the Cohere Labs Catalyst Grant Program, and two of the eight evaluated systems (Tiny Aya Earth and Tiny Aya Global) are Cohere models accessed through Cohere's API using Cohere-funded credits. The other six, competing, open-weight models were run on GPU compute rented and paid for separately by NileAGI (Section 4.3), not by Cohere Labs. Both Tiny Aya variants rank below the top two models on every reported metric in this report, and evaluation code and item-level outputs are released for independent verification, but the funding relationship is a conflict of interest that readers should weigh when interpreting the ranking.

9 Future Work

1. **Complete the parameter-scale roadmap.** Evaluate the scoped mid-size and large levels for education, then design and execute frontier and ultra-scale levels with an explicit MoE placement rule.
2. **Explore additional sector datasets.** We intend to apply the same construction and validation pipeline used for SUB-MCQ Education to other sectors, tentatively following a SUB-Sector naming convention (e.g., health, finance, sports, law, culture, agriculture). No work on these datasets has started; we name them here only to indicate direction, not a committed timeline or scope.
3. **Strengthen SUB-MCQ Education.** Add provenance metadata, licenses, duplicate detection, expert answer-key review by additional annotators, topic labels, difficulty estimates, and balanced grade sampling.
4. **Create protected test data.** Release a development subset while keeping a professionally reviewed test set private, to reduce contamination and leaderboard overfitting.
5. **Run robustness studies.** Randomize option order, paraphrase prompts, repeat API evaluations, compare direct-letter and rationale-first prompting, and report calibration where token

probabilities are available.

6. **Perform controlled adaptation ablations.** Compare matched base, African-language CPT, instruction-tuned, and preference-tuned checkpoints within the same architecture, to isolate the contribution of continued pretraining from that of prompt format.
7. **Add human evaluation.** Measure teacher and student baselines, item ambiguity, explanation quality, cultural appropriateness, and pedagogical safety.
8. **Expand educational coverage.** Add Standard 1–2, Form 1, advanced secondary, vocational, and university-level Swahili content, subject to expert review and licensing.
9. **Broaden language coverage.** Adapt the framework to additional African languages while retaining language-specific governance and community review.

10 Ethics, Governance, and Data Statement

This work is a model evaluation study; it is not a proposal for, or an endorsement of, any particular use of language models in classrooms. That said, educational benchmarks can encode curriculum bias, regional language variation, outdated facts, and annotation error, so a high model score should not be read as a general endorsement of a model's use in high-stakes educational settings. We recommend that dataset governance include Swahili educators, regional language experts, documented item sources, a correction channel, versioned releases, and procedures for removing copyrighted or harmful content. If NILEAGI-SUB is extended to sensitive sectors such as health, finance, or law (Section 9), each such dataset would require sector-specific expert governance in addition to the safeguards described here.

The released results contain model responses and correctness labels; the task does not intentionally require personal data. Before public release, the dataset should nevertheless be screened for names, sensitive narratives, and content unsuitable for minors. Model licenses and API terms should be documented alongside benchmark licensing.

11 Reproducibility Statement

The accompanying repository provides the validated JSON loading pipeline, the model registry with exact identifiers, the Swahili chat and completion prompts, deterministic answer extraction and scoring code, item-level and aggregate results, interruption-safe checkpoints, and English and Swahili publication figures, and the GPU specification for locally hosted models (Section 4.3). A final archival release should add a tagged software version, a package lockfile, exact model revision hashes, execution dates, a dataset license, and a permanent repository DOI.

12 Conclusion

NILEAGI-SUB provides a controlled evaluation framework for Swahili school-level understanding, fixing items, prompting, decoding, and scoring across models. Across 2,569 SUB-MCQ Education items, Gemma 4 E4B reaches 54.8% accuracy and AfriqueQwen3.5-4B reaches 50.3%, while models of similar nominal size differ by more than 14 percentage points. These results

indicate that African-language adaptation can improve Swahili accuracy but is not sufficient on its own, that regional branding requires empirical validation on a specific task, and that parameter count alone does not explain the observed ranking. We caveat all of these findings with the single-annotator and contamination risks discussed in Sections 8 and 6.3, and note that they describe one sector at one model scale.

This report completes Level I of a five-level parameter-scale programme for the education dataset; Level II and Level III are scoped for mid-size and large models, and Level IV and Level V are reserved for frontier and ultra-scale evaluation. We intend to explore, as separate future work, whether the same construction methodology can be extended to sectors beyond education; that extension has not been started and should not be read as an existing part of NILEAGI-SUB's current scope.

Acknowledgments

This research was conducted by Zephania Reuben and Isack Odero at NileAGI. We thank the African NLP and open-model communities whose work motivates language-specific, reproducible evaluation, and Khalid Mwinge, who labeled and reviewed SUB-MCQ Education.

Author Contributions

Zephania Reuben: project conception, benchmark development, experimental design, evaluation, analysis, and writing.
Isack Odero: research design, benchmark review, interpretation, project coordination, and writing.

Funding

This research was supported by a combination of API credits provided through the Cohere Labs Catalyst Grant Program and compute and operating costs funded directly by NileAGI, with further detail on how each was used given in Section 4.3. Funding support does not imply endorsement of the analyses or conclusions; the authors are responsible for the content of this report. We note as a conflict-of-interest disclosure that Cohere Labs is also the provider of two evaluated systems, Tiny Aya Earth and Tiny Aya Global (Section 8); model selection, prompting, decoding, and scoring were applied identically across all eight models regardless of funding source, and all evaluation code and item-level outputs are released to support independent verification of this relationship's effect, if any, on the reported ranking.

References

[1] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL*, 2020.

[2] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of ICML*, 2020.

[3] S. Ruder, N. Constant, J. Botha, A. Siddhant, O. Firat, J. Fu, P. Liu, J. Hu, D. Garrette, G. Neubig, and M. Johnson. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of EMNLP*, 2021.

[4] N. Goyal, C. Gao, V. Chaudhary, P. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ott, F. Guzmán, and others. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

[5] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, and M. Guzmán. The BELEBELE benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of ACL*, 2024.

[6] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *Proceedings of ICLR*, 2021.

[7] D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, and others. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

[8] A. Üstün, V. Aryabumi, Z. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, and others. Aya Model: An instruction finetuned open-access multilingual language model. In *Proceedings of ACL*, pages 15894–15939, 2024.

[9] A. R. Salamanca, D. Abagyan, D. D'souza, A. Khairi, D. Mora, S. Dash, V. Aryabumi, S. Rajaee, and others. Tiny Aya: Bridging scale and multilingual depth. *arXiv preprint arXiv:2603.11510*, 2026.

[10] Gemma Team, Google DeepMind. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

[11] H. Yu, T. Xu, M. A. Hedderich, W. Hamidouche, S. W. Zamir, and D. I. Adelani. AfriqueLLM: How data mixing and model architecture impact continued pre-training for African languages. *arXiv preprint arXiv:2601.06395*, 2026.

[12] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, and others. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[13] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, and others. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.