Working paper · 29 November 2026 · DOI 10.5281/zenodo.24658654

Stress-Testing Frontier AI for Africa

A Multi-Channel Study of 28 Models Across Languages, Voice, and Low-Cost Adaptation

Njenga Kariuki
Independent · njenga.ai · ORCID 0009-0003-8191-2625 · hello@afrival.ai

Companion essay: *Stress-Testing Frontier AI for Africa*, Part 2 of *Africa's Hidden AI Bottleneck*, njenga.ai. Archive: doi.org/10.5281/zenodo.24658654 · GitHub · Hugging Face. Code: MIT. Open data & text: CC BY 4.0.

Abstract

Automated evaluations report high scores for AI models on everyday service work. It is unclear whether those scores hold under the conditions of African frontline services: local languages, short-message channels and phone calls. We test 28 language models (16 closed, 12 open-weight) on 180 private multi-turn, tool-using text tasks set in 10 countries, 6 sectors and 10 languages over chat, USSD and SMS, each run three times; on 90 outbound local-language SMS notices graded by a seven-model reader panel validated against human translation-quality judgments; on 956 real speech recordings; and on 35 tasks delivered as noisy phone calls. The best model passes 98.3% of text tasks on all three runs, and five models pass at least 90%. A language audit of passing replies in Amharic, Yoruba and Kinyarwanda removes most passes from low-cost models: Muse Ember goes from 40 of 45 tasks to 3. Taking each model's weakest result across text tasks, outbound notices and phone calls, one model of 28 stays at or above 80% and 23 fall below 50%. Whisper large-v3 has a word error rate of 91 to 100% in Hausa, Yoruba and Amharic, and models that take audio directly outperform their own speech-to-text pipelines by 11 to 29 points. Carrier fees are 83 to 98% of SMS cost for low-cost models, and multi-step agents exceed typical USSD timeouts. In a companion study, 60 steps of on-policy reinforcement learning (US$9.38) raise a small open-weight model from 46.7% to 69.6% of text tasks, and from 36.7% to 82.2% on a sector held out of training, without improving its language quality. We release the code and the open data, and specify a human-graded benchmark.

1 Introduction

Organizations across Africa are putting language models into customer service, health triage, agricultural advice and public administration. They choose those models using global leaderboards and benchmarks built around English desk work. Those instruments do not measure the conditions these services run under: 160-character USSD screens, SMS billed by segment and encoding, carrier fees on every message, callers on noisy phone lines, and languages with little training data. An earlier essay argued that the missing piece is a shared, testable definition of good performance under those conditions [1].

This paper reports an empirical study of 28 models under those conditions, and a companion study of low-cost adaptation of small open-weight models. One person carried it out for US$1,500 in metered model and compute fees (Appendix C). It makes five contributions.

1. **A text-task measure with channel limits.** On 180 private tasks over chat, USSD and SMS, the best model passes 98.3% on all three runs and five models pass at least 90% (Section 4.1).
2. **A language audit of passing replies.** Automated checks pass replies that a reader panel scores as not understandable, including replies that reverse a clinical instruction. Under the audit, a model that passes 40 of 45 tasks passes 3 (Section 4.2).
3. **A weakest-link summary across three measures.** One model of 28 stays at or above 80% on text tasks, outbound notices and phone-call tasks; global capability indices predict these results only moderately (Sections 4.3 and 4.4).
4. **Speech, cost and latency measurements.** We report speech recognition on 956 real recordings in ten languages, phone-call task results for 52 model and route configurations, the carrier share of SMS cost, and waiting time per USSD screen (Sections 4.5 and 4.6).
5. **An adaptation study.** Supervised fine-tuning on a stronger model's trajectories degrades a small model; on-policy reinforcement learning improves it by 23 points for US$9.38, and leaves its language quality unchanged (Section 5).

Sections 6 to 8 state the limitations, say how the results can and cannot be used in procurement, and specify the human-graded benchmark that this study does not replace.

2 Related work

Global indices such as the Artificial Analysis Intelligence Index [2] and the Epoch Capabilities Index [3], and work benchmarks such as GDPval [4], τ-bench [5] and τ^2-bench [6], measure general reasoning and English-language tasks. African language

benchmarks and applied adaptation efforts measure linguistic capability and domain adaptation: Masakhane, AfroBench (64 languages) [7], IrokoBench (17) [8], INJONGO (16) [9], SSA-COMET and SSA-MTE [10], TukaBench (7) [11], Kakugo [12], Sunbird AI's Sunflower v2, Jacaranda Health's UlizaLlama, Nigeria's N-ATLaS, AfriqueLLM [13], and HelpMum [14], whose evaluation found that its curated MamaBot-Llama improved on its base model while its Vax-Llama degraded.

Multilingual service-agent benchmarks document worse tool arguments and control flow outside English: τ-bench-LILT [15], SEATauBench [16], BabelArena [17] and Lost in Execution [18]. τ-Knowledge [19] shows that agents do far worse when policies must be retrieved from a document collection than when they are read from the prompt. The best model in τ-bench-LILT passes 96 to 99% of English retail and telecom tasks on a single run. On the τ-Knowledge banking domain, which spreads its rules over 698 documents, the top score on a public leaderboard is 51.3% [20].

None of these evaluations jointly measures multi-turn execution, SMS and USSD delivery limits, carrier unit economics, speech recognition on real African recordings, and low-cost adaptation of open-weight models on the same tasks.

3 Method

Table D1 in Appendix D maps the plain names used in this paper to the names in the released code.

3.1 Models and settings

We test 28 models, 16 closed and 12 open-weight. The roster was fixed before any run by a four-part rule applied to the OpenRouter catalog and recorded in a hashed manifest (SHA-256 `81432566…`; a second manifest, `3f228ce8…`, covers the speech systems). Models were called with native tool calling, except Gemma 3 4B, which cannot call tools and used a JSON fallback.

Every model ran the text tasks at its lowest reasoning setting. The 17 models that have a higher provider default, and that passed a screen on the open tasks, ran a second full pass at that default. Table 1 reports the default-setting pass for those 17 and the lowest setting for the other 11. Ten models ran the phone-call tasks at a lower reasoning setting than the one reported for their text tasks; they are marked ‡ in Table 1.

3.2 Text tasks

The text-task measure is 180 private multi-turn tasks drawn from 96 workflows in 30 simulated service packs. The packs cover 10 countries, 6 sectors (mobile money, primary health, farming, public services, education, small-business trade) and 10 languages: Swahili, Hausa, Yoruba, Igbo, isiZulu, Kinyarwanda and Amharic, alongside three reference languages (English, French and Arabic). There are 10 to 37 tasks per country (mean 18) and 11 to 24 per African language.

Sixty tasks run over chat, 60 over USSD with a 160-character screen, and 60 over SMS, of which 17 allow one 160-character segment and 43 allow two. In 120 tasks the customer writes cleanly. In 60 the customer writes informally, in lowercase without punctuation or diacritics; these 60 are all in Swahili, Yoruba, Igbo and isiZulu.

Each pack gives the model a handbook of about 5.5 KB in the system prompt and a set of tools backed by a simulated record system. Of 460 rule entries in the packs, 163 (35.4%) are published statutory or clinical rules, 67 (14.6%) are numerical limits deliberately shifted from the published value, and 230 (50.0%) are constructed provider details. Shifting and construction prevent memorization and avoid reproducing any real operator's tariffs.

Each task runs three times at matched temperature. A run passes if deterministic checks pass on record lookups, actions, stated figures, channel encoding and length, and language identification. A task passes only if all three runs pass, which is pass^3 in the sense of τ-bench [5]. Between 99.95% and 99.97% of scheduled runs completed.

3.3 Outbound notices and the reader panel

The notice measure is 90 private outbound SMS notices of the kind a clinic or lender sends, in Amharic, Yoruba and Kinyarwanda. A panel of seven reader models from seven makers grades them without reference texts: Gemini 4 Hydrogen, Claude Sonnet 5.5, GPT-6.2 Sol, Qwen4 Max, GLM-5.3, Muse Ember and DeepSeek V4.1 Flash. A reader never grades output from its own maker.

The panel is anchored on 12 SSA-MTE error-span examples per language and scores on the SSA-MTE scale of 0 to 100 [10]. We validated it out of sample against SSA-MTE human scores for news translation: Spearman correlation is 0.761 in Amharic, 0.689 in Yoruba and 0.672 in Kinyarwanda, and stays between 0.988 and 0.999 when any one maker is left out. A notice passes at a panel score of 80 or more. The panel is not validated in Swahili, Hausa, Igbo or isiZulu, and we report no panel scores in those languages.

Readers differ in strictness. We fit additive reader effects, $y_{ir} = \mu_i + b_r$ with $\Sigma b_r = 0$, over all votes (12,620 on the notices and 7,977 in the audit), and report scores with the reader effects removed ("adjusted") beside the raw scores.

3.4 Speech recognition on real recordings

The speech measure is 956 recordings (228.1 minutes) of native speakers from public corpora, in all 10 languages. Speaker-disjoint splits are verified in 9 of the 10 languages; the Kinyarwanda corpus has no speaker identifier and is separated by its train and validation partitions only. We score 10 speech-to-text systems and the 4 audio-capable models in the roster on word error rate (WER) and number error rate (NER).

3.5 Phone-call tasks

The phone-call measure delivers 35 of the private tasks, 11 in Hausa and 24 in Swahili, as synthetic speech over a simulated noisy telephone channel. Only the caller's side is spoken; the model's replies stay in text.

We admitted synthetic speech in a language only if recognizers found it as difficult as the real recordings in that language and ranked in the same order on both. Under the rule as first specified, no language qualified. Hausa qualified after we fixed an error in how the number-error metric parsed compound cardinal numbers. Swahili qualified on a fresh 100-clip speaker-disjoint split. The two languages show 0 of 33 and 0 of 38 recognizer rank inversions. An audit of 39 spoken items found no defects.

We ran 52 configurations of model and route. In the direct-audio route the model receives the audio; four models accepted audio through OpenRouter at the time of the study, all from Google. In the transcript route a cloud speech-to-text service transcribes the call and the model reads the transcript. Twelve fully local routes pair each open-weight model with `facebook/mms-1b-all`.

3.6 Language audit of passing replies

The deterministic checks verify actions and stated facts. They do not verify that a reader can understand the reply. We therefore audited the passing results on the 45 text tasks in Amharic, Yoruba and Kinyarwanda, using the reader panel with an added instruction for service replies. There are 961 such results: every task that one of the 28 models, or one of the 13 adaptation-study configurations, passed on all three runs.

Before the audit, the panel scored 18 control items. Fluent controls averaged 89.8 (minimum 87.6), passing replies with known defects averaged 44.9, and replies with a planted reversal of meaning averaged 9.7 (maximum 13.9).

The panel scored the first-run reply for all 961 results. For six models (Gemini 4 Neon, Claude Fable 5.2, Kimi K3.1 Pro, DeepSeek V4.1 Flash, Muse Ember and Qwen3.8 27B), for five models with no task left after the first run, and for all 13 adaptation configurations, we graded every run needed for exact three-run pass[3] results under the reader-adjusted threshold of 80. For the other models we report first-run results, marked †. A first-run result is an upper bound: where both are available, it overstates the three-run result by 6 to 31 points. The audit cost US$35.35.

3.7 Cost and timing

Costs are metered fees at public list prices, per 1,000 tasks. SMS cost adds carrier fees from public aggregator rate cards in seven markets, with regional medians for Ethiopia, Senegal and Egypt. USSD waiting time is the wall-clock time of an exchange divided by its number of customer-facing screens, with tool calls answered instantly by the simulator. Timing per model call was not logged. All hosted models were called through OpenRouter from endpoints in the United States.

3.8 Statistics

Confidence intervals are 95% intervals from a paired bootstrap clustered by workflow, with 2,000 resamples over the 96 workflows. Clustering and pairing follow Miller [21]; we use the bootstrap in place of normal-approximation intervals because the samples are small [22]. On the text tasks, the minimum detectable difference between two models is 8.3 points at the median and 10.7 at most. In Section 5, an asterisk marks a contrast whose paired interval excludes zero.

4 Results

4.1 Text tasks are near the ceiling for the best models

Gemini 4 Neon passes 98.3% of the text tasks on all three runs: 100.0% over chat, 98.3% over USSD and 96.7% over SMS. Five models pass at least 90%, and nine do before the channel limits on length and encoding are applied. Muse Ember passes 90.0% at US$1.64 per 1,000 tasks, about one thirty-fifth of the cost of Gemini 4 Neon (US$57.42). Table 1 reports all 28 models on the three measures of this study.

Table 1. Text tasks, outbound local-language notices and phone-call tasks for all 28 models, ordered by each model's weakest result.

Rank	Model	Weights	Reasoning	Text tasks, % [95% CI]	Notices, % raw / adjusted	Phone-call tasks, % [95% CI]	Weakest result, % raw / adjusted
1	Gemini 4 Neon	closed	low	98.3 [96–100]	**88.9 / 91.1**	91.4 [82–100]	88.9 / 91.1

Rank	Model	Weights	Reasoning	Text tasks, % [95% CI]	Notices, % raw / adjusted	Phone-call tasks, % [95% CI]	Weakest result, % raw / adjusted
2	Gemini 4 Carbon	closed	low	86.7 [81–92]	**70.0 / 73.3**	88.6 [76–97]	70.0 / 73.3
3	Gemini 3.8 Flash	closed	medium‡	76.7 [68–85]	**60.0 / 64.4**	68.6 [52–84]	60.0 / 64.4
4	Gemini 4 Hydrogen	closed	medium‡	93.3 [88–98]	85.6 / 87.8	**54.3** [37–71]	54.3 / 54.3
5	Claude Fable 5.2	closed	low	93.3 [89–97]	95.6 / 92.2	**51.4** [33–71]	51.4 / 51.4
6	Claude Opus 5.5	closed	high‡	82.8 [77–88]	92.2 / 88.9	**40.0** [23–59]	40.0 / 40.0
7=	Kimi K3.1 Pro	open	max	77.8 [70–85]	40.0 / 40.0	**37.1** [20–56]	37.1 / 37.1
7=	GPT-6.2 Astra	closed	medium‡	69.4 [60–78]	82.2 / 82.2	**37.1** [21–56]	37.1 / 37.1
9=	Muse Wonder	closed	medium‡	90.0 [84–95]	76.7 / 78.9	**34.3** [17–52]	34.3 / 34.3
9=	GPT-6.2 Sol	closed	medium	71.7 [63–80]	83.3 / 82.2	**34.3** [18–53]	34.3 / 34.3
9=	Claude Sonnet 5.5	closed	high‡	58.3 [50–67]	63.3 / 61.1	**34.3** [18–53]	34.3 / 34.3
12=	Qwen4 Max	closed	high	84.4 [78–90]	64.4 / 60.0	**31.4** [15–50]	31.4 / 31.4
12=	DeepSeek V4.1 Flash	open	high	72.8 [64–81]	36.7 / 38.9	**31.4** [16–49]	31.4 / 31.4
12=	GPT-6.1 Terra	closed	medium‡	68.3 [60–77]	92.2 / 88.9	**31.4** [15–50]	31.4 / 31.4
12=	Inkling 2 Small	open	high‡	45.0 [36–54]	60.0 / 58.9	**31.4** [15–50]	31.4 / 31.4
16	Gemma 4 31B	open	default	37.2 [29–46]	**24.4 / 26.7**	28.6 [12–46]	24.4 / 26.7
17	Muse Ember	closed	max	90.0 [84–95]	**23.3 / 23.3**	37.1 [21–56]	23.3 / 23.3
18	Grok 4.7	closed	high	82.8 [76–89]	68.9 / 68.9	**20.0** [6–36]	20.0 / 20.0
19	GLM-5.3	open	low	23.9 [17–31]	**17.8 / 18.9**	20.0 [6–37]	17.8 / 18.9
20	GPT-6.2 Luna	closed	none	27.8 [20–36]	76.7 / 76.7	**14.3** [3–28]	14.3 / 14.3
21	Gemma 4 26B	open	default	17.2 [11–24]	**13.3 / 13.3**	14.3 [3–26]	13.3 / 13.3
22	Claude Haiku 5.5	closed	medium‡	35.6 [27–44]	43.3 / 36.7	**11.4** [0–24]	11.4 / 11.4
23	Inkling-2-Mini	open	none	31.1 [24–39]	**8.9 / 8.9**	17.1 [6–32]	8.9 / 8.9
24	Mistral Large 4	open	high‡	52.8 [45–60]	**1.1 / 1.1**	25.7 [11–42]	1.1 / 1.1
25=	Qwen3.8 27B	open	xhigh	62.2 [53–71]	**0.0 / 0.0**	28.6 [12–46]	0.0 / 0.0
25=	LFM3 1.7B	open	default	3.3 [1–6]	**0.0 / 0.0**	2.9 [0–9]	0.0 / 0.0
25=	Ministral 3 8B	open	default	1.1 [0–3]	**0.0 / 0.0**	**0.0** [0–0]	0.0 / 0.0
25=	Gemma 3 4B	open	default	**0.0** [0–0]	**0.0 / 0.0**	**0.0** [0–0]	0.0 / 0.0

Text tasks: share of the 180 private tasks passed on all three runs. Notices: share of the 90 outbound SMS notices scoring 80 or more from the reader panel, raw and adjusted for reader strictness. Phone-call tasks: share of the 35 tasks (11 Hausa, 24 Swahili) passed on all three runs. Weakest result: the lowest of the three; the measure that sets it is in bold. Intervals are 95% paired bootstrap intervals clustered by workflow. Reasoning: the setting reported for text tasks. ‡ The model ran the phone-call tasks at a lower reasoning setting. Only Google's four models accepted audio directly; every other model used the transcript route. Mistral Large 4 and LFM3 1.7B reached the 4,096-token reasoning cap on 85 and 89 of the 90 notices. "=" marks a tie.

Three factors move the text-task scores.

Reasoning setting. Moving from the lowest setting to the provider default raises 12 of the 17 models tested at both settings by a significant margin, including Kimi K3.1 Pro (+57.2 points, to 77.8%), Qwen3.8 27B (+54.4, to 62.2%) and DeepSeek V4.1 Flash (+53.3, to 72.8%, at US$1.03 per 1,000 tasks).

Language. Grok 4.7 passes 94.5% of tasks in the reference languages and 74.8% in the African languages, a difference of 19.8 points.

Writing style. GPT-6.2 Sol passes 75.8% of cleanly written tasks and 63.3% of informally written ones, a difference of 12.5 points. The informal tasks are confined to four languages (Swahili, Yoruba, Igbo and isiZulu), so style and language are confounded.

A review of every miss by the three highest-scoring models found that 12 of 48 were caused by argument synonyms that the handbooks permit. Accepting those synonyms leaves the order of the top three unchanged (Appendix A).

4.2 Passing replies fail a language audit

Table 2 gives excerpts from seven replies that passed every automated check. In each, the model called the correct tools, stated the required figures, fitted the channel and wrote in the right language.

Table 2. Excerpts from replies that passed every automated check.

	Model	Language and task	Excerpt from the passing reply	Literal meaning
1	Muse Ember	Kinyarwanda. A pregnant woman reports malaria symptoms. Rule: no artemether-lumefantrine (AL) at home; refer at once.	Muraho, udatwite ufite malariya ntushobora guhabwa AL mu rugo.	"You who are not pregnant and have malaria cannot be given AL at home." *Utwite* is "pregnant".
2	Qwen3.8 27B	Amharic. A wheat farmer reports yellow rust.	… የዓሣ ወጥ (wheat rust) …	"Fish stew (wheat rust)." The same reply says the plot is registered as "cup (fish), not wheat".
3	Muse Ember	Igbo. Severe pre-eclampsia at 31 weeks, blood pressure 162/104.	agala ụlọ ọgwụ ozugbo ugbua	"Do not go to the hospital right now." *Agala* is the negative imperative.
4	DeepSeek V4.1 Flash	isiZulu. Low blood sugar in a diabetic: when to give sugar by mouth.	uma ephila	"If he is alive." The rule is "if he is awake".
5	Muse Ember	isiZulu. The same task.	uma ephakeme	"If he is tall", or high.
6	Gemma 4 31B	Kinyarwanda. Referral of a malaria patient to the district hospital.	mu rupfu rwo mu karere	"Into the district death."
7	Gemma 4 E4B, optimized prompt (Section 5)	Kinyarwanda. A pregnant malaria patient.	bishoboka gukoresha AL mu rugo	"It is possible to use AL at home."

Each task shown passed on all three runs. Each excerpt appears in the reply from at least one of those runs; the phrase in row 6 appears in all three. All quoted glosses in Table 2 were confirmed by native speakers.

Passing replies also expose internal identifiers to customers. Among the five models that pass at least 90% of text tasks, between 39 and 114 passing replies per model, on 19 to 48 tasks, contain an upper-case identifier such as `APPROVED_SPRAY`.

Figure 1 and Table 3 give the audit results on the 45 tasks in Amharic, Yoruba and Kinyarwanda.

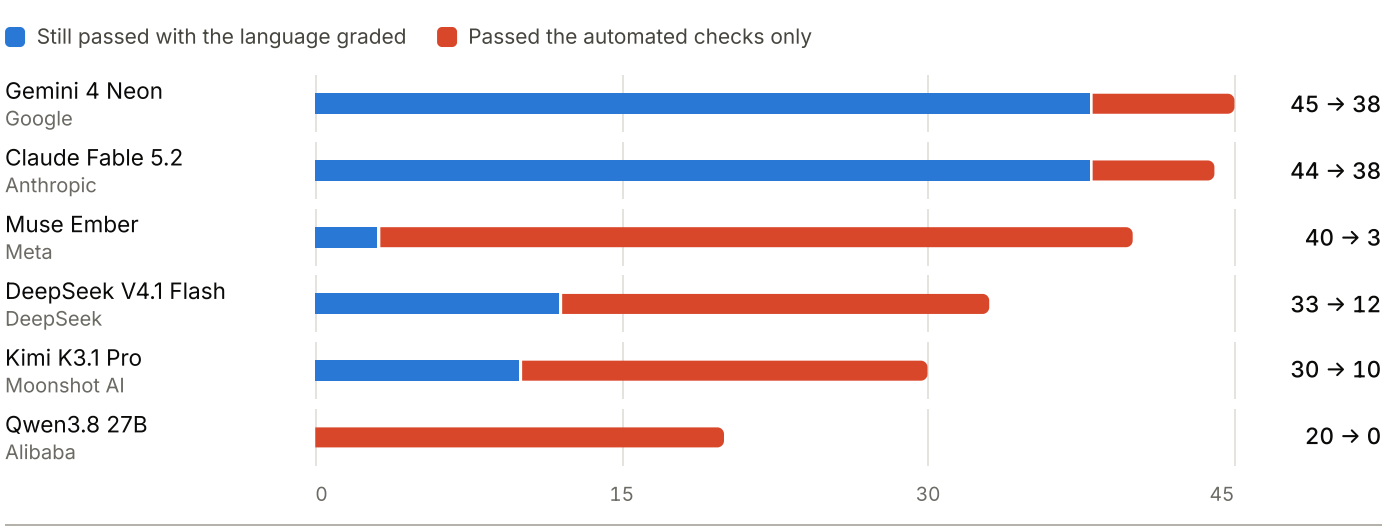

Figure 1. Tasks passed on all three runs, of the 45 text tasks in Amharic, Yoruba and Kinyarwanda, on the automated checks alone and after the language audit (exact three-run pass[3] at threshold 80, adjusted for reader strictness).

Table 3. Language audit of passing replies on the 45 text tasks in Amharic, Yoruba and Kinyarwanda, by pass threshold.

		Mean score, first run		First run: % of 45 tasks at threshold				
Model	Before audit, %	raw	adjusted	≥ 70	≥ 75	≥ 80	≥ 85	Three runs, ≥ 80, %
Evaluated models								
Gemini 4 Neon	100.0	85.65	86.83	100.0	100.0	91.1	68.9	84.4
Claude Fable 5.2	97.8	87.84	85.65	97.8	97.8	91.1	62.2	84.4
Claude Opus 5.5	91.1	88.37	86.03	91.1	88.9	84.4	62.2	first run only
Kimi K3.1 Pro	66.7	84.11	83.89	66.7	64.4	53.3	28.9	22.2

		Mean score, first run		First run: % of 45 tasks at threshold				
Model	**Before audit, %**	**raw**	**adjusted**	**≥ 70**	**≥ 75**	**≥ 80**	**≥ 85**	**Three runs, ≥ 80, %**
DeepSeek V4.1 Flash	73.3	81.96	81.81	71.1	64.4	44.4	26.7	26.7
Muse Ember	88.9	69.28	70.43	53.3	35.6	22.2	15.6	6.7
Qwen3.8 27B	44.4	57.67	56.82	8.9	2.2	2.2	0.0	0.0
Adaptation study (Section 5)								
Inkling-2-Mini, reinforcement learning and optimized prompt, seed 1	66.7	51.59	51.51	13.3	6.7	4.4	0.0	0.0
Gemma 4 E4B, fine-tuning, reinforcement learning and optimized prompt	26.7	65.25	66.42	11.1	11.1	8.9	6.7	2.2
Gemma 4 E4B, fine-tuning and reinforcement learning	17.8	72.62	73.86	11.1	8.9	8.9	8.9	6.7

Before audit: share of the 45 tasks passed on all three runs under the automated checks. Mean score: the panel's mean over first-run replies, on a scale of 0 to 100. Threshold columns: share of the 45 tasks whose first-run reply scores at or above the threshold, adjusted for reader strictness. Three runs: exact share of tasks where all three replies score 80 or more, adjusted.

What holds at every threshold. Muse Ember passes 88.9% (40 of 45) before the audit and 6.7% (3 of 45) after it on three runs, with 0.0% in Yoruba and Kinyarwanda. Of its passing results, 92.5% fail the audit. Qwen3.8 27B goes from 44.4% to 0.0%. These results hold at pass thresholds of 70, 75, 80 and 85, with and without the reader adjustment. DeepSeek V4.1 Flash and Kimi K3.1 Pro have adjusted mean scores of 81.8 and 83.9, and 64.4% of their tasks pass at a threshold of 75 on the first run. Their three-run results at 80 are lower, 26.7% and 22.2%, because all three runs must pass.

Reader strictness. The fitted reader effects span 19.8 points in the audit, from Muse Ember (+8.04) to Claude Sonnet 5.5 (−11.79), and 17.1 points on the notices, from Gemini 4 Hydrogen (+7.66) to Claude Sonnet 5.5 (−9.41). Because a maker's own reader is excluded when its models are graded, Anthropic models are not graded by the strictest reader and Google models are not graded by one of the most lenient. Raw means therefore favor Anthropic models over Google models by 2.9 to 3.4 points. With reader effects removed, Gemini 4 Neon and Claude Fable 5.2 have the exact same audit result: 91.1% of tasks on the first run and 84.4% (38 of 45) on three runs, with mean scores of 86.8 and 85.7. On the five readers that grade both models, Gemini 4 Neon scores higher on four.

Two artifacts of the audit prompt. The audit prompt did not include the handbooks, so readers sometimes penalized correct handbook details, such as fee breakdowns, as inventions; this affects replies scoring 73 to 79. One reader, Claude Sonnet 5.5, penalized Yoruba SMS written without diacritics, which the channel check requires. Qwen4 Max returned no usable score on 233 audit votes and 55 notice votes.

4.3 One measure hides the weakest link

Table 1 orders the 28 models by each model's weakest result across the three measures, and Figure 2 summarizes it. We do not average the measures: an average needs weights that the data cannot supply, and it hides a failure on any one channel.

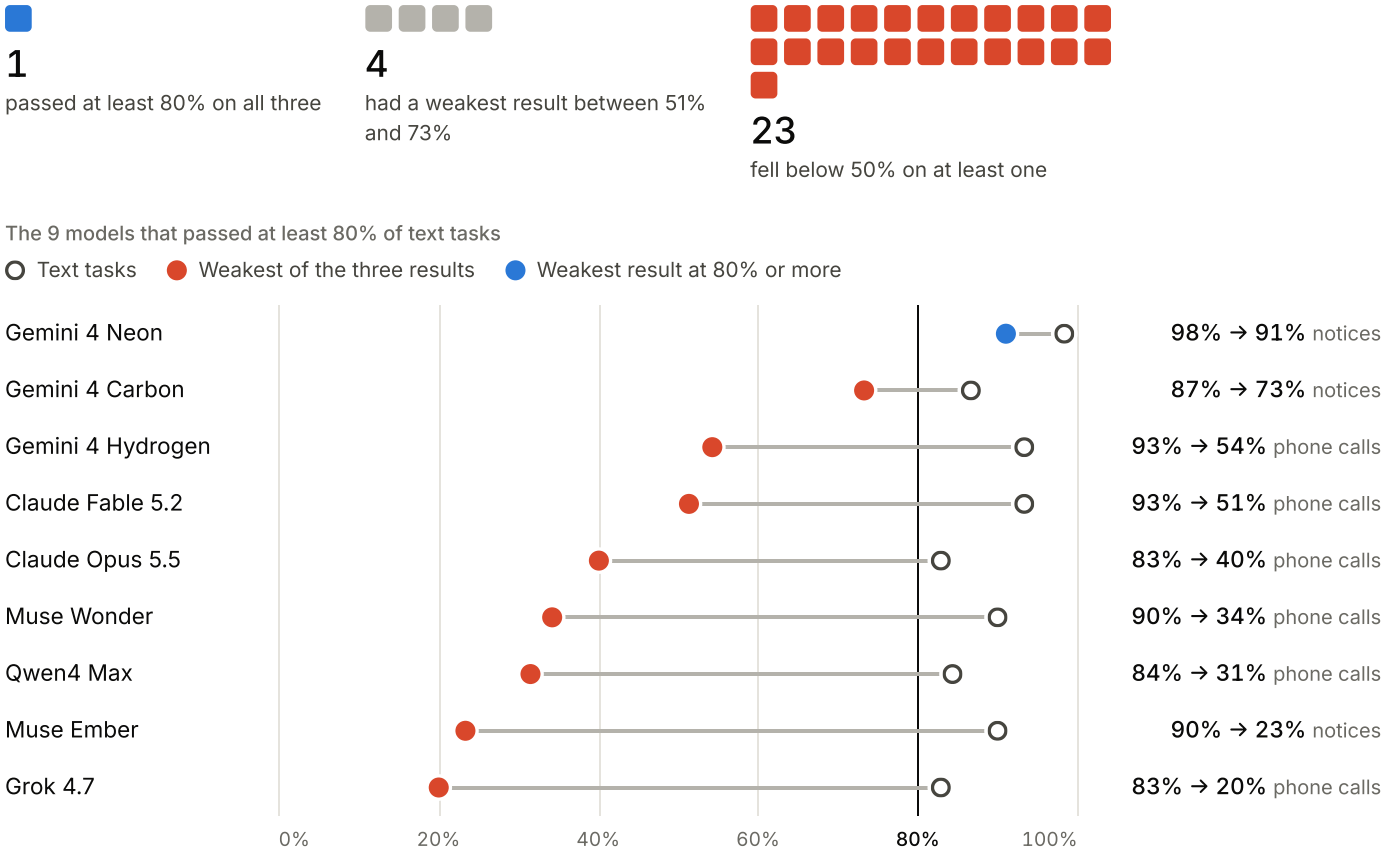

Figure 2. The weakest result across text tasks, outbound notices (adjusted) and phone-call tasks. Top: the 28 models grouped by weakest result. Bottom: the text-task result (open marker) and the weakest result (filled marker) for the 9 models that pass at least 80% of text tasks; the vertical line marks 80%.

One model of 28, Gemini 4 Neon, stays at or above 80% on all three measures; its weakest result is 88.9% raw and 91.1% adjusted. Twenty-three models fall below 50% on at least one measure. No open-weight model exceeds 37.1%.

Three cautions apply. The minimum of three estimates with unequal sample sizes (180, 90 and 35) is biased downward and is dominated by the phone-call measure. Only Google's four models accepted audio directly, which contributes to their holding the first four ranks. And Mistral Large 4 and LFM3 1.7B reached the 4,096-token reasoning cap on 85 and 89 of the 90 notices.

Table 4 gives the language audit, cost and channel results for the same models.

Table 4. Language audit, cost and channel results for the same 28 models.

Rank	Model	Language audit, % of 45 tasks, before → after	Cost per 1,000 text tasks, US$	Carrier share of SMS cost, %	USSD median wait per screen, s (% of exchanges over 10 s)
1	Gemini 4 Neon	100 → 84	57.42	14	12.7 (98)
2	Gemini 4 Carbon	87 → 67†	39.38	20	18.1 (91)
3	Gemini 3.8 Flash	78 → 60†	24.72	26	17.9 (99)
4	Gemini 4 Hydrogen	84 → 67†	33.10	22	8.3 (18)
5	Claude Fable 5.2	98 → 84	96.63	9	7.2 (1)
6	Claude Opus 5.5	91 → 84†	54.51	16	6.6 (4)
7=	Kimi K3.1 Pro	67 → 22	35.44	20	25.8 (100)
7=	GPT-6.2 Astra	71 → 67†	39.68	21	6.7 (1)
9=	Muse Wonder	87 → 71†	37.64	20	6.5 (1)
9=	GPT-6.2 Sol	80 → 73†	9.80	51	5.3 (0)
9=	Claude Sonnet 5.5	58 → 44†	26.45	34	5.4 (1)
12=	Qwen4 Max	80 → 64†	12.98	43	7.8 (11)
12=	DeepSeek V4.1 Flash	73 → 27	1.03	90	7.1 (6)
12=	GPT-6.1 Terra	64 → 53†	5.60	62	6.3 (1)

Rank	Model	Language audit, % of 45 tasks, before → after	Cost per 1,000 text tasks, US$	Carrier share of SMS cost, %	USSD median wait per screen, s (% of exchanges over 10 s)
12=	Inkling 2 Small	44 → 18†	1.82	89	5.3 (1)
16	Gemma 4 31B	42 → 29†	2.40	82	4.0 (3)
17	Muse Ember	89 → 7	1.64	83	5.5 (0)
18	Grok 4.7	76 → 51†	36.17	17	13.6 (77)
19	GLM-5.3	24 → 0	9.46	70	5.9 (2)
20	GPT-6.2 Luna	24 → 22†	0.54	97	3.8 (1)
21	Gemma 4 26B	11 → 4†	3.40	83	3.8 (1)
22	Claude Haiku 5.5	42 → 40†	1.85	85	5.2 (0)
23	Inkling-2-Mini	24 → 0	0.50	97	7.5 (22)
24	Mistral Large 4	27 → 7†	9.63	63	15.1 (79)
25=	Qwen3.8 27B	44 → 0	12.97	44	16.3 (93)
25=	LFM3 1.7B	2 → 0	2.64	92	8.0 (25)
25=	Ministral 3 8B	0 → 0	0.61	98	2.2 (4)
25=	Gemma 3 4B	0 → 0	0.97	95	2.6 (6)

Rank as in Table 1. Language audit: share of the 45 text tasks in Amharic, Yoruba and Kinyarwanda passed on all three runs, before the audit and after it (threshold 80, adjusted for reader strictness). Values without † are exact three-run pass[3] results. † After-audit value from the first run only, an upper bound (Section 3.6). Cost: model fees at public list prices. Carrier share: carrier fees as a share of the total cost of an SMS exchange. USSD: wall-clock time of an exchange divided by its customer-facing screens, with instant simulated tools and no mobile-network delay.

4.4 Global indices predict these results only moderately

Across the 25 models that have a score on the Artificial Analysis Intelligence Index (version 5.0.2) [2], the Spearman rank correlation with that index is 0.671 for text tasks, 0.725 for phone-call tasks, 0.807 for notices (0.808 adjusted) and 0.779 for the weakest link. Across the 19 models with a score on the Epoch Capabilities Index [3], it is 0.629 for text tasks and 0.754 for the weakest link.

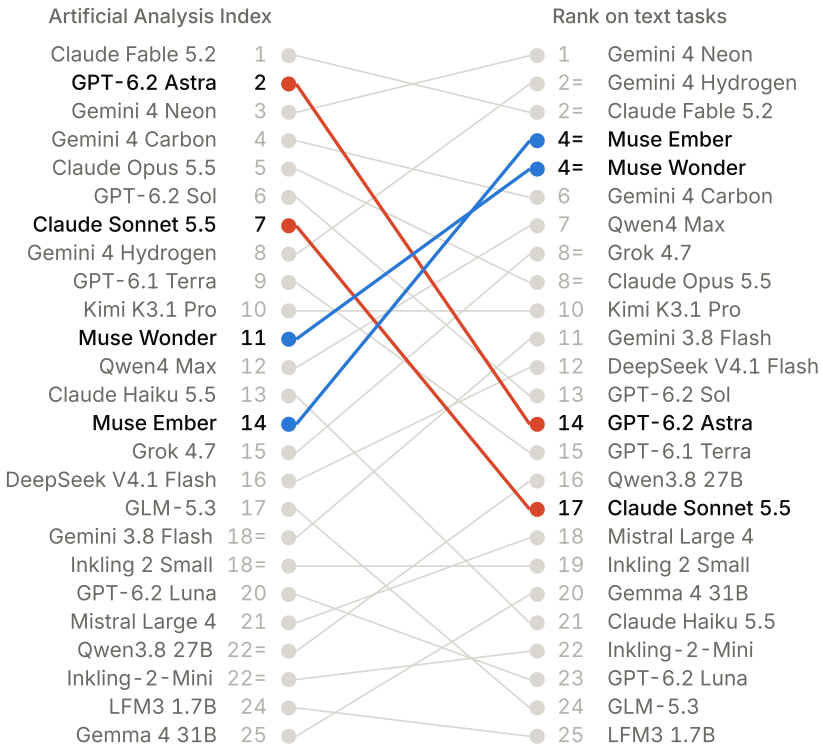

Figure 3. Rank on the Artificial Analysis Intelligence Index (version 5.0.2, retrieved 26 November 2026) against rank on text tasks, for the 25 models with an index score. Spearman correlation 0.671. "=" marks a tie. Highlighted: GPT-6.2 Astra and Claude Sonnet 5.5, which rank lower on text tasks, and Muse Ember and Muse Wonder, which rank higher.

Figure 3 compares the two rankings on text tasks. GPT-6.2 Astra ranks second of the 25 on the Artificial Analysis index and 14th on text tasks (69.4%). Muse Ember ranks 14th on the index and joint fourth on text tasks. Claude Fable 5.2 ranks first on the index, passes 93.3% of text tasks and 92.2% of notices (adjusted), and passes 51.4% of phone-call tasks.

4.5 Speech recognition and phone-call tasks

Table 5 and Figure 4 report speech recognition on the real recordings.

Table 5. Speech recognition on 956 real recordings in ten languages.

		Best speech-to-text system			Best audio-capable model		Open-weight, WER	
Language, corpus and speech style	**Clips**	**System**	**WER**	**NER**	**Model**	**WER**	**MMS-1B**	**Whisper large-v3**
Igbo `google/fleurs`, read aloud	92	`elevenlabs/scribe-v2`	29.0	16.1	Gemini 4 Neon	32.8	41.8	not supported
Yoruba `google/fleurs`, read aloud	100	`elevenlabs/scribe-v2`	25.1	19.0	Gemini 4 Neon	25.1	32.7	97.4
Egyptian Arabic `facebook/omnilingual-asr-corpus`, spontaneous	84	`meta/muse-voice-transcribe-1.0`	24.9	25.4	Gemini 4 Hydrogen	21.8	69.0	39.6
Amharic `google/WaxalNLP`, spontaneous	100	`google/chirp-3`	28.5	36.2	Gemini 4 Neon	21.7	49.5	100.0
Hausa[a] `google/fleurs`, read aloud	94	`elevenlabs/scribe-v2`	19.0	24.1	Gemini 4 Carbon	17.3	26.2	91.3
isiZulu `google/fleurs`, read aloud	100	`google/gemini-3.5-transcribe`	27.3	21.2	Gemini 4 Carbon	16.8	35.7	not supported
French (West and Central Africa) `OpenSLR/SLR57`, mixed and read aloud	100	`deepgram/nova-3`	15.1	11.8	Gemini 4 Neon and Gemini 4 Hydrogen (tied)	16.5	47.8	24.1
Kinyarwanda `badrex/kinyarwanda-speech-1000h`, spontaneous	100[b]	`elevenlabs/scribe-v2`	16.9	13.9	Gemini 4 Neon	12.5	28.6	not supported
Nigerian English `OpenSLR/SLR70`, read aloud	100	`mistralai/voxtral-small-transcribe-3-0`	7.1	14.0	Gemini 4 Neon	7.4	24.6	12.1
Kenyan Swahili[a] `Tonykip/kenyan-swahili-asr-clean`, spontaneous	86	`google/chirp-3`	9.6	17.1	Gemini 4 Neon	6.3	16.8	43.4

WER: word error rate. NER: number error rate. Both in %. Best speech-to-text system: the lowest WER among the ten systems tested, with its NER. Best audio-capable model: the lowest WER among the four roster models that accept audio. Languages are ordered by the lowest WER of any system, highest first. [a] Language admitted to the phone-call measure (Section 3.5). [b] No speaker identifier in the corpus; separated by train and validation partitions only. Corpus notes: `Tonykip/kenyan-swahili-asr-clean` is repackaged from Digital Umuganda's AfriVoice Swahili; `badrex/kinyarwanda-speech-1000h` is Digital Umuganda's Kaggle Track B; the Amharic part of `google/WaxalNLP` was built with African partner organizations including Digital Umuganda; `OpenSLR/SLR57` covers West and Central Africa, excluding Senegal. The text tasks set in Egypt use Modern Standard Arabic; these recordings are colloquial Egyptian Arabic (9 speakers).

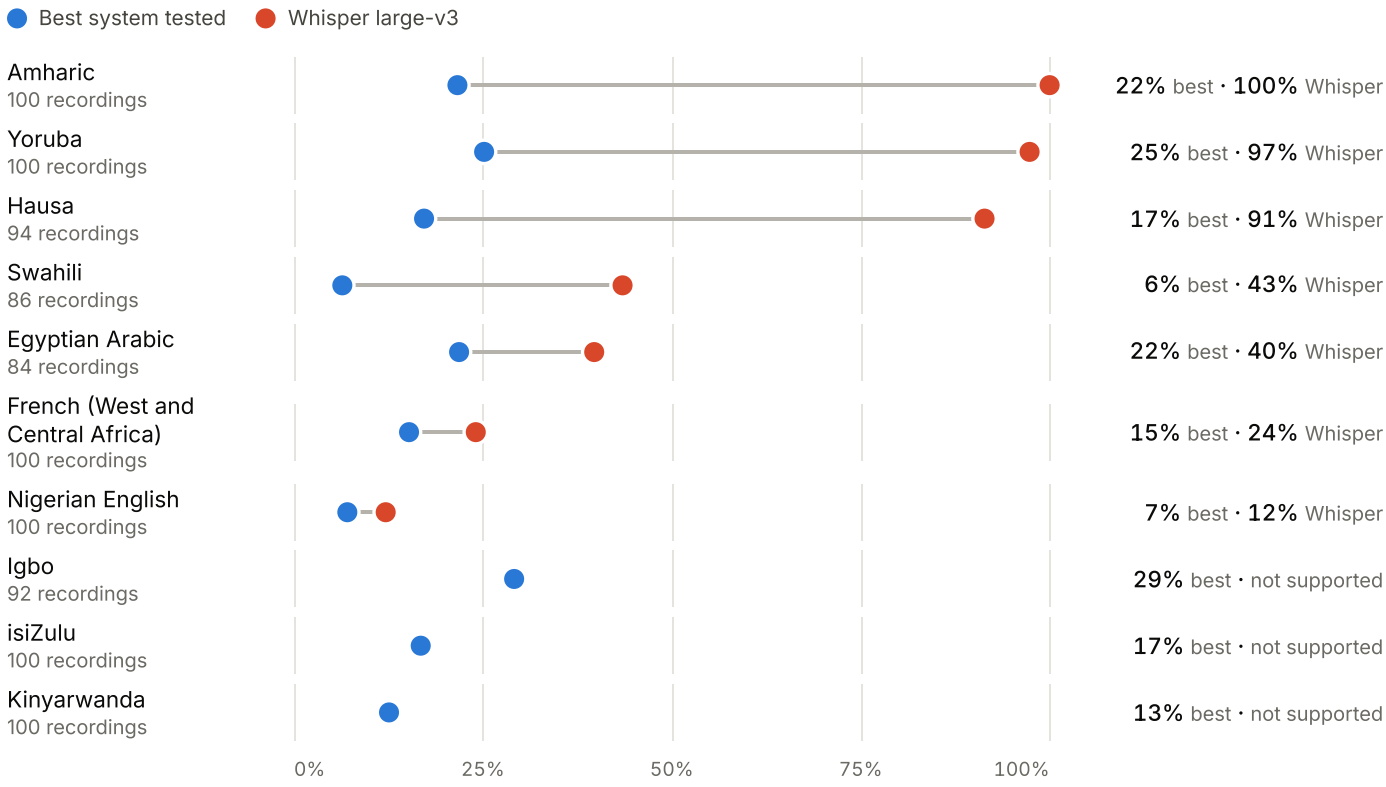

Figure 4. Word error rate on the 956 real recordings, by language: the best system tested (the lowest WER among ten speech-to-text systems and four audio-capable models) and Whisper large-v3, which does not support Igbo, isiZulu or Kinyarwanda.

Speech recognition. The best speech-to-text service has a WER between 19.0% and 29.0%, and an NER between 16.1% and 36.2%, in six languages: Hausa, Egyptian Arabic, Yoruba, isiZulu, Amharic and Igbo. Whisper large-v3 has a WER of 100.0% in Amharic, 97.4% in Yoruba, 91.3% in Hausa and 43.4% in Swahili, and does not support Igbo, isiZulu or Kinyarwanda. The open-weight MMS-1B has a lower WER than Whisper large-v3 by 27 to 65 points in the four African languages that both support, and a higher WER in English, French and Arabic.

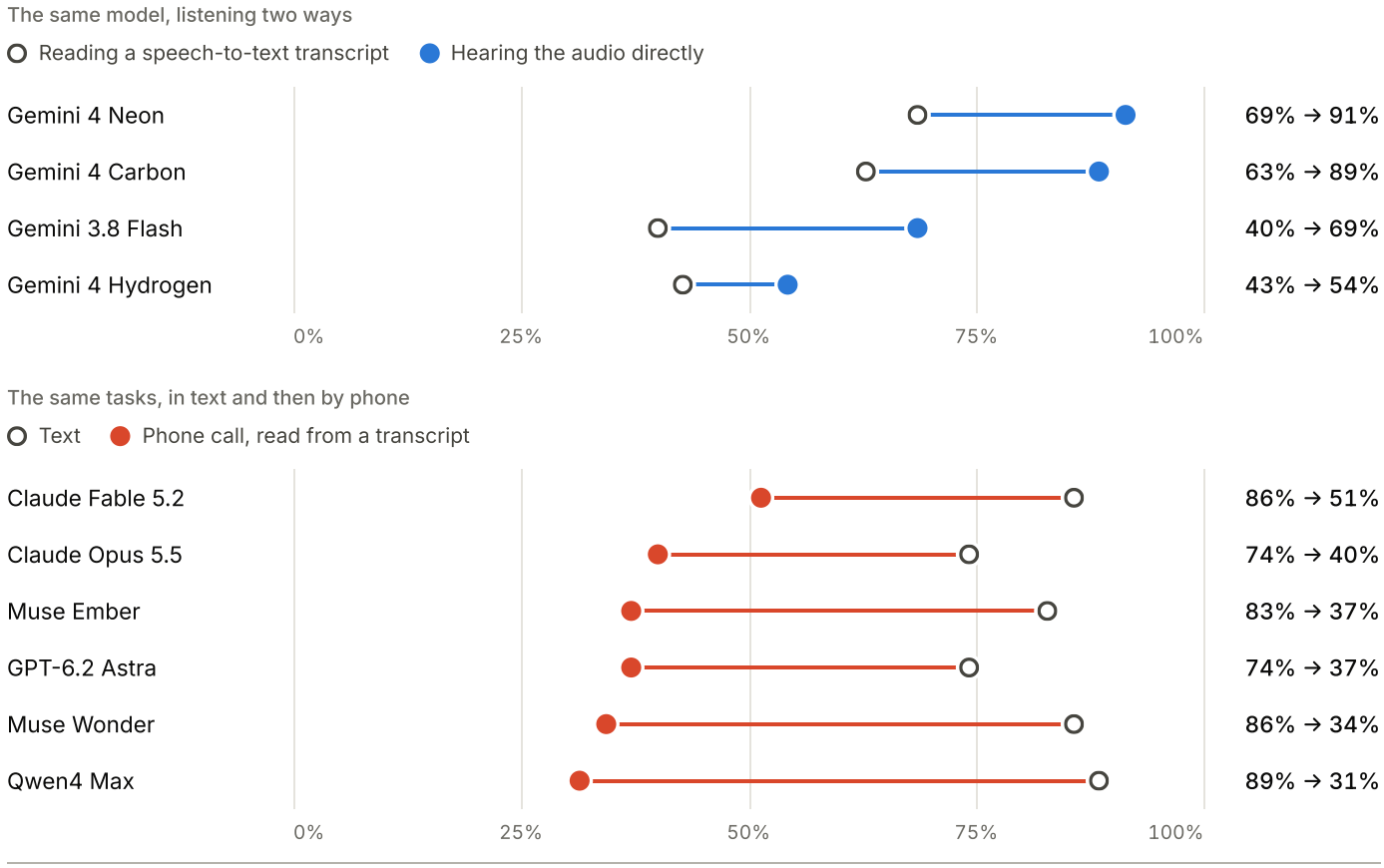

Figure 5. Phone-call tasks passed on all three runs (n = 35). Top: the four audio-capable models in the transcript route (open marker) and the direct-audio route (filled marker). Bottom: six models in the transcript route (filled marker) against the same 35 tasks in text at the same reasoning setting (open marker).

Direct audio against transcripts. Each of the four audio-capable models passes more phone-call tasks from the audio than from a transcript, by 11.4 to 28.6 points, at a total cost 7.8 to 22.9% lower (Figure 5, top). Gemini 4 Neon passes 91.4% from audio at US$66.88 per 1,000 tasks and 68.6% from a transcript at US$75.38. These costs exclude carrier voice minutes and speech synthesis.

Text against spoken. Against the same 35 tasks in text, at the same model setting, the ten transcript-route models with the highest text-task scores lose 34.3 to 57.1 points (Figure 5, bottom, shows six): Claude Fable 5.2 falls to 51.4%, Claude Opus 5.5 to 40.0%, GPT-6.2 Astra and Muse Ember to 37.1%, and Qwen4 Max, with the largest loss, to 31.4%. From audio, Gemini 4 Neon and Gemini 4 Carbon lose 2.9 and 5.7 points, which is not significant; Gemini 3.8 Flash and Gemini 4 Hydrogen lose 20.0 and 34.3.

Why transcripts fail. Among the 12 highest-scoring models in the transcript route, those that pass at least 34.3%, 570 of 671 failed runs (84.9%) are missed record lookups; across all 28 models the share is 84.7%. In 465 of them the model called the lookup and the record system returned no match, typically because the transcript had altered the identifier; for example, `CLN-NG-57` was transcribed as “CLM NG 57”. The simulated caller does not repeat a detail when asked, so a live caller might recover some of these.

Local routes. All 12 fully local routes pass 0.0% of the phone-call tasks: MMS-1B transcribes sounds without formatting numbers or identifiers. In the transcript route, the best open-weight model, Kimi K3.1 Pro, passes 37.1%.

Refusals on real recordings. Gemini 3.8 Flash returned a content-policy error on 107 of 672 clips (15.9%) in the seven African languages: 40 of 92 in Igbo, 25 of 100 in isiZulu, 16 of 94 in Hausa, 14 of 100 in Kinyarwanda, 7 of 100 in Amharic, 3 of 100 in Yoruba and 2 of 86 in Swahili. It returned none on the 284 clips in English, French and Arabic. Gemini 4 Neon and Gemini 4 Hydrogen returned no errors on any of the 956 clips. Gemini 4 Carbon returned 10 clip errors: nine timeouts and one Swahili clip stopped with a safety finish reason.

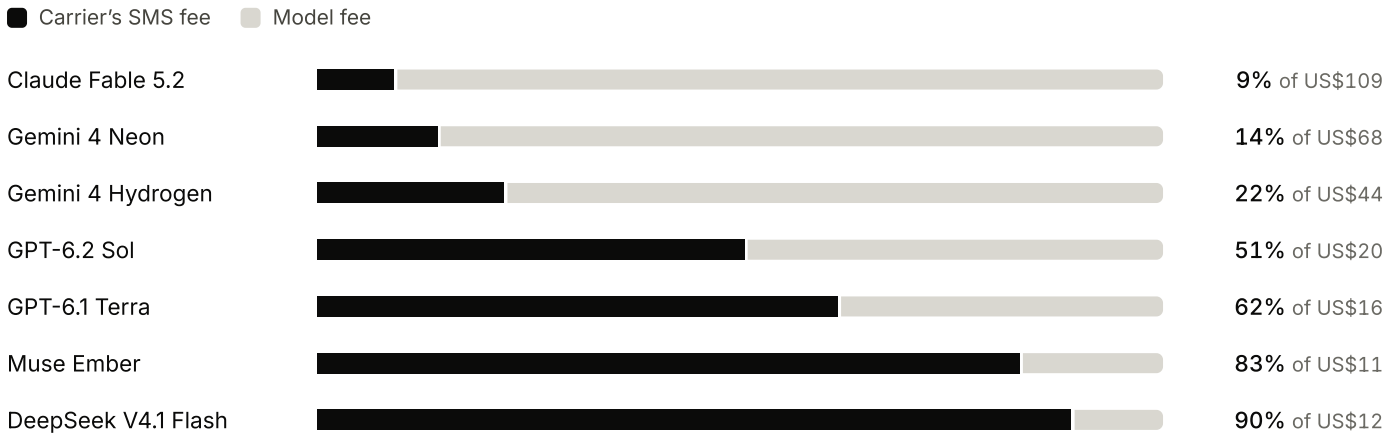

Figure 6. Carrier fees as a share of the total cost of 1,000 SMS exchanges, with the total per 1,000 exchanges at right, for seven models ordered from the highest model fee to the lowest.

SMS cost. The carrier's fee is between 9.2% and 22.1% of the total cost of an SMS exchange for the nine models that cost more than US$30 per 1,000 text tasks, and 9.2% for the most expensive, Claude Fable 5.2 (Figure 6). It is 83.3% for Muse Ember, 89.6% for DeepSeek V4.1 Flash and up to 98.2% for the smallest open-weight models. For low-cost models, total cost depends on the number of segments, and so on reply length and on whether any character forces the 70-character UCS-2 encoding in place of the 160-character GSM-7 encoding, far more than on the model's price.

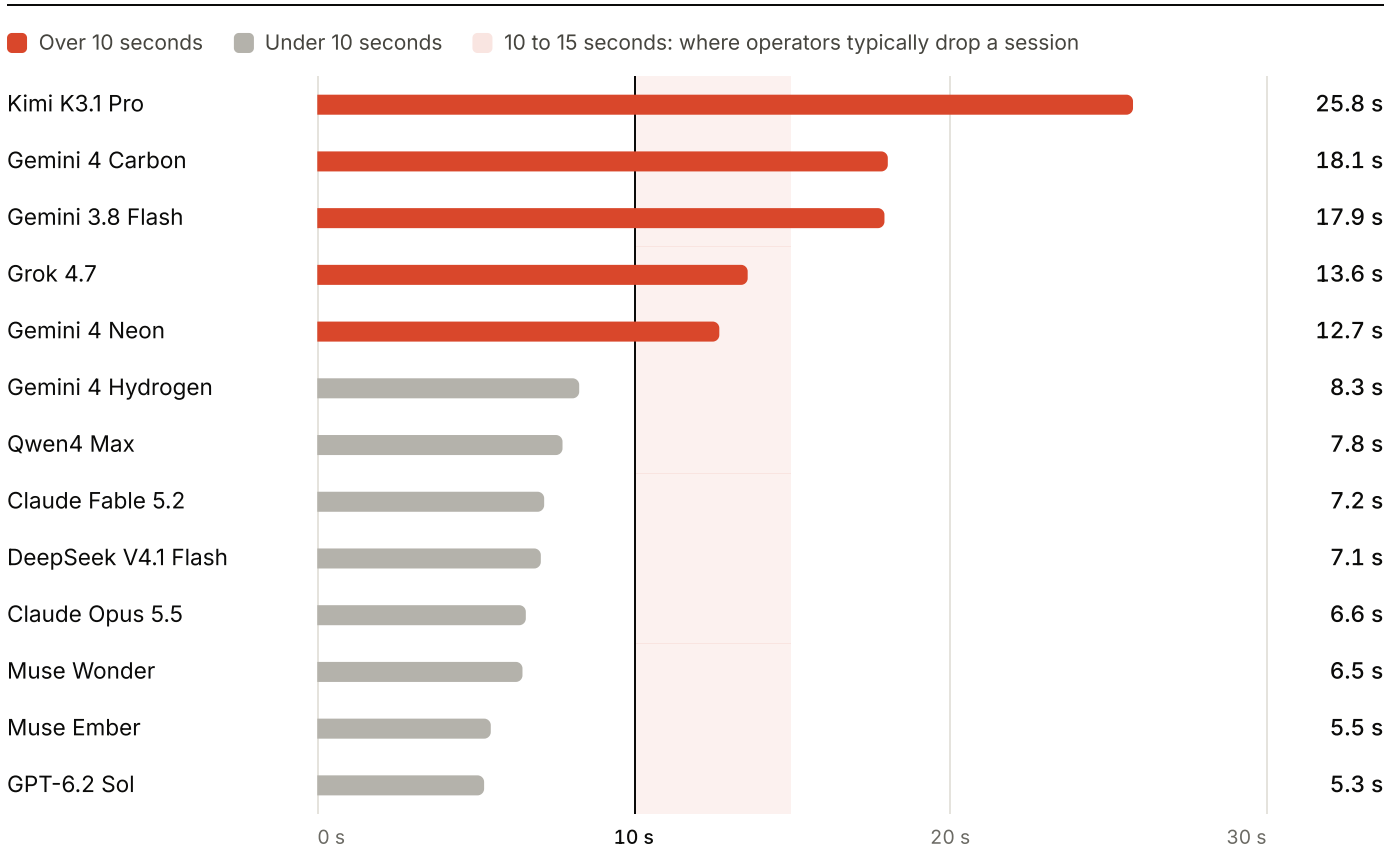

Figure 7. Median wait per customer-facing USSD screen for the 13 models that pass at least 70% of text tasks. The line marks 10 s and the band 10 to 15 s, the range of published operator timeouts. Times exclude mobile-network delay.

USSD waiting time. Operators end a USSD session if the application does not answer within a set time; an aggregator's published figures for Kenyan operators are 10 to 15 seconds [23], and we use 10 as the reference. Among the 13 models that pass at least 70%

of text tasks, five have a median wait above 10 seconds per screen (Figure 7): Kimi K3.1 Pro (25.8 s; 100.0% of exchanges exceed 10 s per screen), Gemini 4 Carbon (18.1 s; 91.1%), Gemini 3.8 Flash (17.9 s; 99.4%), Grok 4.7 (13.6 s; 76.7%) and Gemini 4 Neon (12.7 s; 97.8%). These times exclude mobile-network delay.

A per-screen median understates the wait on later screens. A typical exchange has two screens and four model calls, three of them before the second screen. Claude Fable 5.2 has a median of 7.2 s per screen and a median exchange time of 14.9 s, so its second screen is likely to exceed 10 s. Multi-step tool use inside a synchronous USSD session is therefore unsafe even for fast models. A service can retrieve the customer's record by phone number before a single model call, or end the session and reply by SMS.

No model leads on every measure. GPT-6.1 Terra passes 92.2% of notices raw and 88.9% adjusted at US$0.61 per 1,000 in model fees; Claude Fable 5.2 passes 95.6% and 92.2% at US$23.51. With a carrier fee of about US$10 per 1,000, the total cost for GPT-6.1 Terra is about a third of that for Claude Fable 5.2. GPT-6.1 Terra passes 68.3% of text tasks. Phone-call tasks need direct audio, and the best audio model has a median exchange time of 35.0 s. A deployment that routes requests by channel, language and time budget will do better than one that sends every request to a single model.

5 Adaptation study

5.1 Design

The adaptation study asks how far low-cost training can take a small open-weight model on the text tasks. It cost US$168.06.

Ten fictional training services were written in a separate session with no access to the evaluation packs. Training rows were scanned against all 4,002 evaluation rows, with no canary match and no exact overlap. The farming sector was excluded from training entirely. Every adapted model was evaluated off the training platform, served with vLLM on rented H100 GPUs.

We adapted two models. Inkling-2-Mini (26.4 billion parameters, 3.4 billion active) was trained on Tinker with a LoRA adapter. Gemma 4 E4B (Apache 2.0, 4 billion effective parameters) was trained with PyTorch and PEFT on one H100.

Table 1 lists Inkling-2-Mini at 31.1%; that is the hosted model with reasoning off. The base in this study, 46.7%, is the same model served with vLLM at its default reasoning setting. With reasoning off it passes 28.3%.

5.2 Supervised fine-tuning on a stronger model's trajectories

Fine-tuning Inkling-2-Mini on trajectories from DeepSeek V4.1 Flash that contained no reasoning reduced its reasoning from 1,309 tokens to 1.4 and its pass rate on an in-distribution split ($n = 150$) from 42.0% to 16.0%. HelpMum reports a similar degradation for its Vax-Llama [14].

5.3 On-policy reinforcement learning

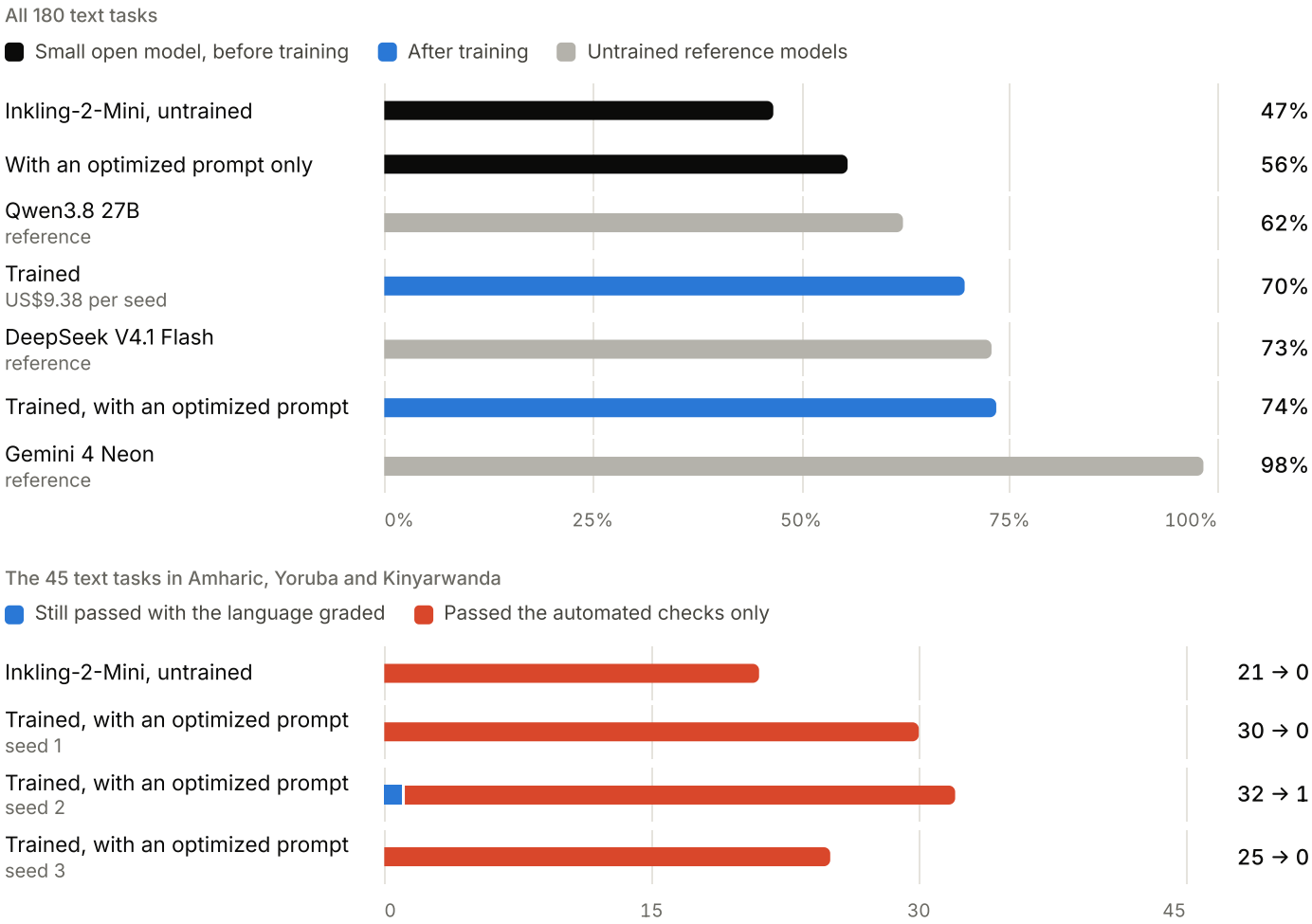

Figure 8. Adaptation of Inkling-2-Mini. Top: text tasks passed on all three runs (n = 180) for the untrained model, the optimized prompt alone, reinforcement learning (mean of three seeds) and both, with three untrained reference models. Bottom: tasks passed on all three runs, of the 45 in Amharic, Yoruba and Kinyarwanda, before and after the language audit.

Sixty steps of on-policy multi-turn reinforcement learning with verifiable rewards, using the CISPO objective and self-distillation from episodes with hints removed, cost US$9.38 per seed. Over three seeds it raised the pass rate on the 180 private text tasks from 46.7% to 69.6% ± 4.1 (+23.0*), and to 73.5% ± 4.6 with a prompt optimized by GEPA (+26.9*) (Figure 8). GEPA alone reaches 55.6%; reinforcement learning exceeds it by 14.1 points*. For reference, Inkling 2 Small passes 45.0%, Qwen3.8 27B 62.2% and DeepSeek V4.1 Flash 72.8%.

On the 30 held-out farming tasks the pass rate rose from 36.7% to 82.2% (+45.6*), and to 87.8% with the optimized prompt (+51.1*). No health or money run in any Inkling-2-Mini configuration contained a forbidden tool call.

For Gemma 4 E4B, on-policy fine-tuning on its own verified trajectories followed by three rounds of reinforcement learning with the GSPO objective cost US$11.14. It raised the pass rate from 10.6% to 21.7%, with one forbidden health tool call, as in the untrained model, and to 27.2% with the optimized prompt (+16.7*), with none. Untuned Gemma 4 26B passes 17.2%.

5.4 What training did not change

Of the reduction in failed runs for Inkling-2-Mini, from 182 to a three-seed mean of 79.0, 78.0% came from fewer replies that overran the channel's length limit (98 to 17.7). Language quality did not change. The untrained model passes 8.9% of the notices, and none in Kinyarwanda. Table 6 gives the language audit for all 13 configurations: for Inkling-2-Mini with reinforcement learning and the optimized prompt, the three-seed mean goes from 64.4% before the audit to 0.7% after it.

Table 6. Language audit of the 13 adaptation-study configurations.

Configuration	Before audit, %	After audit, first run, %	After audit, three runs, %	Change, points	Mean score, first run raw / adjusted
Inkling-2-Mini					
Untrained	46.7	4.4	0.0	−46.7	46.4 / 46.2
Optimized prompt only	53.3	2.2	0.0	−53.3	44.6 / 44.5
Reinforcement learning, seed 1	60.0	4.4	0.0	−60.0	47.6 / 47.5
Reinforcement learning, seed 2	64.4	6.7	0.0	−64.4	54.1 / 54.0
Reinforcement learning, seed 3	57.8	0.0	0.0	−57.8	49.4 / 49.2
Reinforcement learning and optimized prompt, seed 1	66.7	4.4	0.0	−66.7	51.6 / 51.5
Reinforcement learning and optimized prompt, seed 2	71.1	4.4	2.2	−68.9	54.3 / 54.1
Reinforcement learning and optimized prompt, seed 3	55.6	4.4	0.0	−55.6	47.4 / 47.3
Gemma 4 E4B					
Untrained	13.3	4.4	0.0	−13.3	71.5 / 72.6
Optimized prompt only	22.2	8.9	2.2	−20.0	70.5 / 71.8
Fine-tuning on its own verified trajectories	11.1	6.7	2.2	−8.9	71.1 / 72.2
Fine-tuning and reinforcement learning	17.8	8.9	6.7	−11.1	72.6 / 73.9
Fine-tuning, reinforcement learning and optimized prompt	26.7	8.9	2.2	−24.5	65.2 / 66.4

Shares of the 45 text tasks in Amharic, Yoruba and Kinyarwanda. Before audit: passed on all three runs under the automated checks. After audit: also scoring 80 or more from the reader panel (adjusted for reader strictness), on the first run and on all three runs. Change: three-run result after the audit minus the result before it.

The check for forbidden actions covers tool calls and not prose. A passing Kinyarwanda reply from Gemma 4 E4B with the optimized prompt alone tells a pregnant malaria patient that the medicine can be used at home (Table 2, row 7).

5.5 Cost of self-hosting

On one H100 rented at US$4.39 per hour, the adapted Inkling-2-Mini costs US$2.54 to US$2.59 per 1,000 successful conversations at full load. That is 6 to 23 times cheaper than Qwen3.8 27B, Kimi K3.1 Pro and Gemini 4 Neon, with break-even volumes of about 6,500, 2,600 and 1,800 successful conversations a day. It is never cheaper than DeepSeek V4.1 Flash (US$1.23) or Muse Ember (US$1.72). Self-hosting is justified by data residency, offline operation and control of the weights, and not by price.

6 Limitations

No human graders. Text tasks are graded by deterministic checks. Notices and the language audit are graded by a reader panel that is validated against human scores in three languages, on news translation and not on service replies. Only the quoted excerpts in Table 2 were confirmed by native speakers; the 961 audited tasks and 90 notices were not human-graded.

Constructed tasks. The tasks are built from public rules and constructed provider details. They are not sampled from real operations.

Rules in the prompt. Each handbook is about 5.5 KB and sits in the system prompt. This explains part of the ceiling on text tasks; real services spread their rules over many documents.

A small phone-call measure. It has 35 tasks in two languages, with synthetic callers, text replies, a caller that does not repeat details, and direct audio from one maker's models only. Ten models ran it at a lower reasoning setting than their text tasks.

Confounded writing style. The informally written tasks are all in four languages.

Latency. Hosted models were timed from United States endpoints, and timing per model call was not logged. No model was run on a phone.

One author. The private test set is held by the author.

7 Using these results

Usable in a tender now. Require a model to pass each task on several runs, not one. Check channel fit: 160-character USSD screens, and SMS billed by segment under GSM-7 and UCS-2 encoding. Require USSD designs that retrieve the customer's record by phone number before a single model call. Have native speakers read at least 50 passing replies in each language served, without

knowing which model wrote them, with authority to fail a model on clinical or financial safety. Test speech recognition on real audio, for words and for numbers. Audit refusals by language. Run the released kit in country, and keep ownership of test cases under national data-protection law (Kenya Data Protection Act s.48, Nigeria NDPA ss.41–43, South Africa POPIA s.72, Rwanda Law 058/2021).

Not usable as it stands. The 28-model tables of November 2026 are not a vendor pre-qualification or a certification.

8 Specification for a human-graded benchmark

The limitations above define the benchmark this study does not replace. We refer to it as AfriVal. Table 7 sets out its five components against what this study did, with a total estimated cost of US$15,000 to US$50,000. Its documentation should meet the practices assessed in BetterBench [24], with a datasheet for each released set [25].

Table 7. Components of a human-graded benchmark, against what this study did.

Component	This study	Specification	Estimated cost, US$
Human grading	No human graders. The reader panel is validated in three of the seven African languages.	Blind grading by native speakers and frontline practitioners in 8 to 10 languages, with authority to fail a reply on clinical or financial safety.	8,000 to 20,000
Knowledge	Rules in a handbook of about 5.5 KB inside the prompt.	Rules spread over 100 to 300 changing regulatory and operator documents that the model must retrieve and combine.	2,000 to 5,000
Speech	Synthetic callers, validated in two of ten languages. Text replies.	Recorded multi-turn callers with code-switching in all ten languages, and spoken replies that are graded.	3,000 to 15,000
Latency and devices	Timing from endpoints in the United States. No model run on a phone.	Timing from Nairobi, Lagos, Cairo and Johannesburg, and quantized models on 4 GB Android phones.	1,000 to 5,000
Custody	Private test set held by the author.	Private test set held by an African university or independent research group, redrawn twice a year.	1,000 to 5,000
Total			**15,000 to 50,000**

Appendix A. Sensitivity to permitted synonyms

We reviewed all 48 misses by the three highest-scoring models on the text tasks. Twelve (25.0%) were caused by three argument synonyms that the handbooks permit, in 5 of the 180 tasks (workflows `H_RW_02`, `E_RW_04` and `E_ET_03`). Table A1 rescores every model with those synonyms accepted, at no cost. It lists the 17 models whose score changes or that pass at least 85% as scored; the other 11 do not change. The order of the top three is unchanged.

Table A1. Text tasks passed on all three runs, as scored and with the three permitted synonyms accepted.

Model	As scored, %	Synonyms accepted, %	Change, points	Synonyms accepted: chat / USSD / SMS, %
Gemini 4 Neon	98.33	98.33	+0.00	100.0 / 98.3 / 96.7
Gemini 4 Hydrogen	93.33	96.11	+2.78	98.3 / 98.3 / 91.7
Claude Fable 5.2	93.33	93.89	+0.56	98.3 / 86.7 / 96.7
Muse Wonder	90.00	92.22	+2.22	98.3 / 93.3 / 85.0
Muse Ember	90.00	91.67	+1.67	96.7 / 90.0 / 88.3
Gemini 4 Carbon	86.67	87.78	+1.11	96.7 / 90.0 / 76.7
Qwen4 Max	84.44	85.56	+1.11	95.0 / 83.3 / 78.3
Grok 4.7	82.78	83.89	+1.11	95.0 / 73.3 / 83.3
Kimi K3.1 Pro	77.78	78.33	+0.56	81.7 / 76.7 / 76.7
Gemini 3.8 Flash	76.67	77.22	+0.56	73.3 / 85.0 / 73.3
GPT-6.2 Sol	71.67	72.22	+0.56	81.7 / 66.7 / 68.3
Qwen3.8 27B	62.22	63.89	+1.67	76.7 / 80.0 / 35.0
Claude Sonnet 5.5	58.33	61.11	+2.78	96.7 / 53.3 / 33.3
Mistral Large 4	52.78	53.89	+1.11	61.7 / 71.7 / 28.3
Inkling 2 Small	45.00	45.56	+0.56	90.0 / 33.3 / 13.3
Claude Haiku 5.5	35.56	37.22	+1.67	63.3 / 15.0 / 33.3
GLM-5.3	23.89	24.44	+0.56	71.7 / 0.0 / 1.7

Appendix B. Errata in the task packs

1. The Uganda health pack (tasks `W_P043`, `W_P044`, `W_P060`) gives the emergency line as 911 or 112. Uganda's emergency numbers are 999 and 112. We record this so that no Ugandan service copies 911.
2. The Kenya health pack refers first-trimester malaria cases to a facility at community level, which fits an SMS or community-health-worker triage desk. WHO guidance from November 2022 recommends artemether-lumefantrine in the first trimester under facility care.
3. The Rwanda health pack's metadata labels the national helpline 114 as constructed. The pack text cites 114 correctly.

Appendix C. Cost ledger

Table C1 lists all metered spend at public list prices. No model maker or operator provided credits.

Table C1. Metered spend.

Component	Scope	US$
Evaluation of the 28 models	The 30 service packs; open tasks; private text tasks at both reasoning settings; outbound notices; speech recognition on 956 recordings; phone-call tasks	1,160.55
Adaptation study	Ten training services; prompt optimization; supervised fine-tuning; three seeds of reinforcement learning; Gemma 4 E4B training; evaluation served with vLLM	168.06
Language audit	The 18 control items (US$1.09) and the audit of 961 passing results including exact three-run calibrated completion (US$34.26)	35.35
Subtotal		**1,363.96**
Earlier pilot of the method	A pilot of 100 tasks and 32 models (12,516 replies and 45,081 reader scores). No results from it are reported in this paper. Approximate figure.	about 158
Total		**about 1,522**

Appendix D. Names in the released code

Table D1. Names used in this paper and in the released code and data.

In this paper	In the released code and data
Text tasks; passed on all three runs	`Work`; `pass3`
Outbound notices (local-language notices)	`Localize`
Reader panel; reader	Anchored Jury (`jury`); judge
Adjusted for reader strictness	`cal` (calibrated)
Speech recognition on real recordings	`Hear`
Phone-call tasks	Spoken Work (`spoken`)
Direct-audio, transcript and local routes	`direct_audio`, `cloud_stt`, `local_open_stt`
Language audit	`fluency_audit_v34`
Weakest link	`weakest_link_3pillar`
Task; exchange; model call	case; episode; round
Handbook; service pack	SOP; pack
Reasoning setting	`effort`
Cleanly written; informally written	`clean`, `messy`
Reference languages; African languages	`ref_lang`, `local_lang`
Adaptation study	Study 1 (`study1`)
Inkling-2-Mini; Gemma 4 E4B	System A; System B
Untrained; optimized prompt	`S0`; `S1` (GEPA)

In this paper	In the released code and data
Supervised fine-tuning on a stronger model's trajectories	S2a
Fine-tuning on the model's own verified trajectories	S2b
Reinforcement learning	S2c

Appendix E. Disclosures

Conflict of interest and independence. Njenga Kariuki is Managing Executive, AI at Vodacom FinTech Group (including M-PESA Africa). This study was conducted solely in a personal capacity, funded entirely by the author, with no data, systems, funding, or endorsement from his employer, any telecom operator, or any AI model maker. All 30 evaluation services — including the Kenya and Tanzania mobile-money packs — use counterfactual provider identities, fee tables, and shifted transaction ceilings drawn from public regulatory publications, and both mobile-money training services were placed outside Vodacom and M-PESA's operating footprint (Rwanda and Senegal). No model maker, operator, or funder received advance results or influenced the design.

AI assistance. Njenga Kariuki designed and directed this study using AI coding and language models from multiple makers for software, task construction, verification, synthetic prompts, grading, and editorial drafting. No AI model is an author. Credibility rests on symmetric rules applied identically to every maker: dual verification engines, own-maker-excluded reader panels adjusted for strictness, pre-run hashed manifests, and open release of code and data.

Reproducibility. Every number in this paper is generated from stored transcripts by the released code at github.com/afrival/afrival-study-2026, archived on Zenodo (DOI 10.5281/zenodo.24658654) and mirrored on Hugging Face (AfriVal/afrival-study-2026). Code is released under the MIT license; the open tasks, with their transcripts, grades and results, under CC BY 4.0, with Croissant metadata [26]. The private tasks are withheld to keep the test uncontaminated. Canary string: `AFRIVAL-CANARY-V1-8f4d2c91-a7e3-4b18-9c6d-2e5f1a8b3d04`.

References

[1] Kariuki, N. (2026). *Africa's Hidden AI Bottleneck (Part 1).* njenga.ai; Zenodo DOI: 10.5281/zenodo.24463675 njenga.ai/writing/africa-ai-bottleneck/

[2] Artificial Analysis (2026). *Artificial Analysis Intelligence Index v5.0.2 (via OpenRouter Rankings, retrieved 26 Nov 2026).* artificialanalysis.ai artificialanalysis.ai/methodology/intelligence-benchmarking

[3] Epoch AI (2026). *Epoch Capabilities Index (CC BY 4.0; dataset dated 17 Nov 2026, retrieved 26 Nov 2026).* epoch.ai/benchmarks epoch.ai/benchmarks

[4] OpenAI (2025). *GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks.* arXiv:2510.04374 arxiv.org/abs/2510.04374

[5] Yao, S., Shinn, N., Razavi, P., & Narasimhan, K. (2024). *τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains.* arXiv:2406.12045 arxiv.org/abs/2406.12045

[6] Barres, V., Dong, H., Ray, S., Si, X., & Narasimhan, K. (2025). *τ²-Bench: Evaluating Conversational Agents in a Dual-Control Environment.* arXiv:2506.07982 arxiv.org/abs/2506.07982

[7] Ojo, J., Ogundepo, O., Oladipo, A., Ogueji, K., Lin, J., Stenetorp, P., & Adelani, D. I. (2025). *AfroBench: How Good Are Large Language Models on African Languages?* Findings of ACL 2025; arXiv:2311.07978 arxiv.org/abs/2311.07978

[8] Adelani, D. I., et al. (2025). *IrokoBench: A New Benchmark for African Languages in the Era of Large Language Models.* NAACL 2025; arXiv:2406.03368 arxiv.org/abs/2406.03368

[9] Yu, H., et al. (2025). *INJONGO: A Multicultural Intent Detection and Slot-Filling Dataset for 16 African Languages.* arXiv:2502.09814 arxiv.org/abs/2502.09814

[10] Li, S., et al. (2025). *SSA-COMET: Do LLMs Outperform Learned Metrics in Evaluating MT for Under-Resourced African Languages?* EMNLP 2025; arXiv:2506.04557 arxiv.org/abs/2506.04557

[11] Akinode, V. O., Li, S., Hamidouche, W., Zamir, S. W., Becker-Reshef, I., & Adelani, D. I. (2026). *TukaBench: A Culturally Grounded Jailbreak Benchmark for African Languages.* EMNLP 2026; arXiv:2606.01322 arxiv.org/abs/2606.01322

[12] Devine, P., et al. (2026). *Kakugo: Low-Cost African Language Adaptation Under US$50 per Language.* arXiv:2601.14051 arxiv.org/abs/2601.14051

[13] Yu, H., et al. (2026). *AfriqueLLM: Open Language Modeling for African Languages.* arXiv:2601.06395 arxiv.org/abs/2601.06395

[14] Ajibade, A., et al. (HelpMum). *Evaluating Fine-Tuned and Base Language Models in Maternal and Vaccination Healthcare for African Settings.* HelpMum Research (MamaBot-Llama & Vax-Llama) helpmum.org

[15] Uhlig, K., Mokhtari, K., Arslan, A., Kim, Y., & Wuebker, J. (2026). *τ-bench-LILT: Evaluating Conversational Agents on Realistic Multilingual Tool-Use Tasks.* arXiv:2611.07203 arxiv.org/abs/2611.07203

[16] Nguyen, M. C., Adila, A., Ruangtanusak, S., et al. (2026). *SEATauBench: Progressively Adapting Tool-Agent-User Evaluation Into Low-Resource Southeast Asian Languages.* arXiv:2606.28715 arxiv.org/abs/2606.28715

[17] Kuang, P., Xiao, T., Yang, B., et al. (2026). *BabelArena: A Large-Scale Multilingual Benchmark for LLM Agents.* arXiv:2609.23490 arxiv.org/abs/2609.23490

[18] Luo, Z., Kutralingam, T. P., Okoani, O. N., Xu, W., Wei, H., & Hu, X. (2026). *Lost in Execution: On the Multilingual Robustness of Tool Calling in Large Language Models.* Proceedings of ACL 2026 Long Papers aclanthology.org/2026.acl-long.2039/

[19] Shi, Q., Zytek, A., Razavi, P., Narasimhan, K., & Barres, V. (2026). *τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge.* arXiv:2603.04370 arxiv.org/abs/2603.04370

[20] Artificial Analysis (2026). *τ^3-Banking evaluation leaderboard (retrieved 29 Nov 2026).* artificialanalysis.ai. artificialanalysis.ai/evaluations/tau3-banking

[21] Miller, E. (2024). *Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations.* arXiv:2411.00640 arxiv.org/abs/2411.00640

[22] Bowyer, S., Aitchison, L., & Ivanova, D. R. (2025). *Position: Don't Use the CLT in LLM Evals With Fewer Than a Few Hundred Datapoints.* ICML 2025; arXiv:2503.01747 arxiv.org/abs/2503.01747

[23] Africa's Talking (n.d.). *What is the duration of a USSD session for Kenyan telcos?* Africa's Talking Help Center (retrieved 29 Nov 2026). help.africastalking.com/en/articles/1284071-what-is-the-duration-of-a-ussd-session-for-kenyan-telcos

[24] Reuel, A., Hardy, A., Smith, C., et al. (2024). *BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices.* NeurIPS 2024 Datasets & Benchmarks; arXiv:2411.12990 arxiv.org/abs/2411.12990

[25] Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). *Datasheets for Datasets.* Communications of the ACM, 64(12), 86–92 dl.acm.org/doi/10.1145/3458723

[26] Akhtar, M., Benjelloun, O., Conforti, C., et al. (2024). *Croissant: A Metadata Format for ML-Ready Datasets.* MLCommons Croissant 1.0 mlcommons.org/working-groups/data/croissant/