What Open Language Models Know of Hausa, Zarma and Fulfulde

Namu AI Lab
namuai-lab.com

Abstract

A voice assistant needs a language model at its centre, the brain that understands a request and writes the answer. We report what open language models know of Hausa, Zarma and Fulfulde, three languages of Niger, with English and French as references, and how those measurements shaped the brain of our research voice agents. Four findings stand out. First, the cost of reading follows the text a model has seen. On the same 150 sentences, a general 9B base model needs 1.14 bits per character for English, 3.17 for Hausa and 4.27 for Zarma; the chat version of the same model further trained on 50 African languages needs 0.96, 1.48 and 3.34. Hausa falls by 53% and Zarma, which was not among those languages, by 22%. The tokeniser, which cuts Zarma into fewer tokens than Hausa, does not explain the gap. Second, models understand before they write: the 9B model further trained on African languages picks the right English meaning of a Zarma sentence 77% of the time, against 58% for a character-overlap baseline and 100% for the same sentences in Hausa, yet writes no Zarma when asked; it writes Hausa, and a translation model asked for Zarma writes Fulfulde. Third, for Zarma the brain must therefore reason in another language: choosing a French brain among six open models by a rule fixed in advance, we found that the model trained on African languages was the least safe in French, failing 13 of 30 safety checks. Fourth, the safety layer around a brain changes what a comparison measures: by overwriting correct refusals, a fixed care guard made the model with the most good answers rank third instead of first, and in our Zarma voice agent every remaining care failure happened in translation, before the brain saw the turn. For Fulfulde we have only tokeniser counts and scores published by others. All results are automatic measurements; no native speaker or clinician has rated any output yet.

1 Introduction

A spoken assistant has to understand what it hears and decide what to say. In the systems we build for Niger, that decision is made by a language model, which we call the brain: a recogniser turns speech into text, the brain writes an answer, and a synthesis model speaks it. Open language models are the brains a small laboratory can host, inspect and adapt. They learn mostly from web text, and the languages most people in Niger speak at home are thinly present there. Hausa is spoken widely in Niger and across northern Nigeria; Zarma is spoken by about six million people, most of them in Niger [13]; Fulfulde is a family of varieties spread across the Sahel. Before choosing a brain we needed to know what open models know of these three languages, and what that knowledge allows a voice assistant to do.

Technical report, Namu AI Lab, November 2026. This report describes research systems, not products.

This report brings together the measurements we made while building research voice agents in Hausa and Zarma, and reads them as one study of the brain. We look at three levels. The first is reading: how many bits a model needs to encode a sentence of each language, on parallel text so that the content is held fixed. The second is understanding and writing: whether a model grasps what a sentence means, and whether it writes the language when asked. The third is behaviour inside an agent: how a brain chosen for its French answers, with safety rules around it, handles spoken requests that reach it through translation. English serves as the reference for reading, and French, the language our Zarma brain reasons in, as the reference for behaviour. We describe every model by its type and size; the point of the report is what the measurements show.

Our contributions are the following. (1) A measurement of the cost of reading Hausa, Zarma and English on the same sentences for open 8B and 9B models, with and without further training on 50 African languages, which shows that this cost tracks the text a model has seen, while the tokeniser does not explain it (Section 4). (2) Evidence that these models understand more than they write: a 9B model further trained on African languages picks the right meaning of most Zarma sentences but writes no Zarma, and a translation model asked for Zarma writes Fulfulde (Section 5). (3) A selection of a French brain among six open models by a rule fixed in advance, in which the model further trained on African languages was the least safe in French (Section 6). (4) An account of the pivot design, in which the brain reasons in French and translation carries Zarma in and out, and of a fixed care guard around the brain, with two effects that change how brains should be compared (Section 7).

2 Related work

Measuring what a model knows of a language. Bits per character, or cross-entropy per character, measures how well a model predicts text independently of how its tokeniser cuts the text. Comparing it across languages on parallel text, which holds the content fixed, was proposed to ask whether some languages are harder to model than others [10, 17]. Language modelling is a form of compression [11], and compression efficiency tracks performance on downstream tasks across models [16]. We use bits per character as a cheap measure of what a model has absorbed of a language, which needs no generation and no judge.

Tokenisers and the cost of languages. The same content can cost very different numbers of tokens across languages [2, 21], and tokenisers fitted to a few languages weaken models on others [23]. We report tokens per word next to bits per character to separate the tokeniser's share from the model's.

African languages in language models. Benchmarks show large gaps between high-resource and African languages in comprehension, reasoning and translation [1, 5, 20]. Further training on African languages raises performance on the languages trained [3], continued pre-training on a target domain helps in general [15], and multilingual training helps low-resource languages up to a point, mostly through related languages [8]. Zarma is absent from FLORES-200 [14, 19] and FLEURS [9]; SMOL adds a small professionally translated set for it [7].

Pivoting and safety across languages. Reasoning in a high-resource language and translating in and out is a standard baseline for low-resource languages [4, 24]. Safety behaviour learned mostly in English transfers poorly to low-resource languages [12, 25]. Our earlier report describes the whole Zarma speech stack and its voice agent [18]; this report treats the language model in more depth and adds Hausa and Fulfulde.

3 Languages, text and models

The languages. Hausa is written in a Latin alphabet with the hooked letters *ɓ ɗ ƙ ƴ*. Zarma, a Songhay language unrelated to Hausa, borrows many words from it, and its written form varies widely in practice [18]. Fulfulde in Niger is, for practical purposes, two languages: eastern listeners under-

Table 1: The models in this report, by type and size. "African-continued" means further trained on text in 50 African languages that include Hausa and Nigerian Fulfulde but not Zarma.

Model	Size	Used in
General open base model	9B	Section 4
The same, African-continued, base	9B	Section 4
The same, African-continued, chat	9B, 4B	Sections 4 to 7
General chat model of the same family	9B	Sections 6, 7
Older model further trained on 20 African languages	8B	Sections 4, 5
General chat model of the older generation	8B	Section 5
Mixture-of-experts chat model (about 4B active)	26B	Section 6
Chat model of the same family as the above	12B	Section 6
Chat models of a third family	8B, 14B	Section 6
Massively multilingual translation model	about 3B	Section 5

stand Nigerian Fulfulde, western ones do not. Every public Fulfulde test set we found is labelled Nigerian Fulfulde. Two human translations of the same English sentences, one in an eastern and one in a western Fulfulde style, score only 16.9 chrF++ [22] against each other, little above Hausa and Zarma translations scored against the same reference (13.0 and 13.2). The two also differ by translator, so this point estimate does not isolate the effect of variety, but it shows that one reference of another variety can score Fulfulde like a wrong language. Our few Fulfulde figures therefore describe Nigerian-style writing only.

How much a model can have seen. Public text is scarce in a measurable way. One large public web corpus keeps 46 Zarma documents, about 41,500 words, and another about 100 clean documents, about 97,000 words, mostly from the Bible; the first corpus's Hausa part is a 591 MB file. The African-continued 9B model below was further trained on about 35 billion tokens in 50 African languages. According to its developers' data tables, its training text held about 500 million tokens of Hausa (repeated three times) and about 2 million of Nigerian Fulfulde (repeated five times); Zarma is not among the 50 languages.

Models. Table 1 lists the models. The reading measurement compares a general open 9B base model with the same model further trained on the 50 African languages, so that the effect of that training can be read on the same architecture and tokeniser. It adds the chat version of the African-continued 9B, which was the brain of our first Zarma and Hausa agents, and an older 8B model further trained on 20 African languages. Five other chat models were compared for the French brain.

Text for measurement. We measure reading on the 150 held-out test sentences of the SMOL English and Zarma set [7] that we froze for our Zarma work, with SMOL's Hausa translations of the same English sentences, so that the three languages carry the same content. We add two sets of other text: 331 Hausa sentences from the FLEURS test transcripts [9], which are FLORES-200 sentences [14], and 1,060 transcribed lines of prompted spontaneous Zarma speech from one speaker. Each line is scored on its own after the model's end-of-text token. Every token of the line is predicted and counted, and bits per character is the total of its bits over its number of Unicode characters, spaces included. Models run in 16-bit floating point, one line at a time. Intervals are 95% bootstrap intervals clustered by sentence, document or recording. Public sets may have been in some models' training data, and FLORES sentences in particular are widely copied, so absolute values on them may be optimistic.

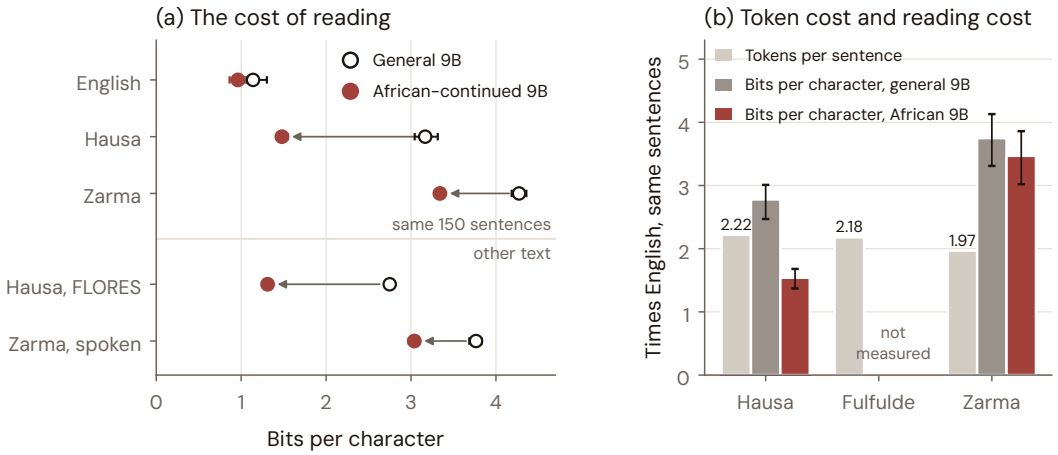


Figure 1: The cost of reading. (a) Bits per character for a general 9B base model and the same model further trained on 50 African languages (chat version; its base version is within 0.02), on the same 150 SMOL sentences in three languages and on two other sets, with 95% intervals; arrows show the effect of the African training. (b) Cost relative to English on parallel sentences: the tokens the shared tokeniser spends per sentence, counted on SMOL's larger set of single sentences, and bits per character for each model on the 150 test sentences. We have no bits-per-character value for Fulfulde.

4 The cost of reading each language

Bits per character tells how surprised a model is by a language, without asking it to write anything. Figure 1 and Appendix Table 2 give the results.

Three languages, same sentences. On the 150 SMOL sentences the general 9B base model needs 1.14 bits per character for English [1.03, 1.30], 3.17 for Hausa [3.04, 3.32] and 4.27 for Zarma [4.19, 4.36]. The chat version of the same model further trained on African languages needs 0.96 for English [0.86, 1.10], 1.48 for Hausa [1.42, 1.55] and 3.34 for Zarma [3.28, 3.40]. The further training cut the cost of Hausa by 53%, to about one and a half times that of English (1.54 [1.37, 1.68]), and the cost of Zarma by 22%, although Zarma was not part of it. English also became cheaper, by 16%, so relative to English the cost of Zarma moved only from 3.75 to 3.47 times, while that of Hausa moved from 2.78 to 1.54 times. Zarma still costs 1.86 bits per character more than Hausa [1.79, 1.92] on the same sentences, 2.26 times as much, and 3.47 times as much as English [3.02, 3.86]. Zarma writes the same content in fewer characters, so per sentence its version costs 1.97 times the bits of the Hausa version. The base and chat versions of the African-continued model differ by at most 0.02 on every set, so instruction tuning barely changed how the model reads. The older 8B model further trained on 20 African languages stays within 0.08 of the 9B chat model on every set.

Other text. The pattern holds on the two other sets. On FLORES Hausa sentences the African training takes the cost from 2.75 to 1.31, and on transcripts of spontaneous Zarma speech from 3.77 to 3.04. The two Zarma sets differ in register, translated web text against transcribed speech, and give different absolute values; the models rank the same way on both.

The tokeniser does not explain the gap. The African-continued models use the general model's tokeniser unchanged, so their gain lies entirely in the model. The tokeniser is also not what makes Zarma expensive (Figure 1b). It spends 1.19 tokens per English word, 2.08 to 2.12 per Hausa word and 1.61 to 1.86 per Zarma word. On SMOL's parallel single sentences it needs 1.97 times the English tokens for Zarma, 2.22 times for Hausa and 2.18 times for Fulfulde; on parallel FLORES sentences, 1.38 times the French tokens for Hausa and 1.30 times for Fulfulde. Zarma is the cheapest of the three to cut into tokens, and of the languages we measured it is the most expensive to predict.

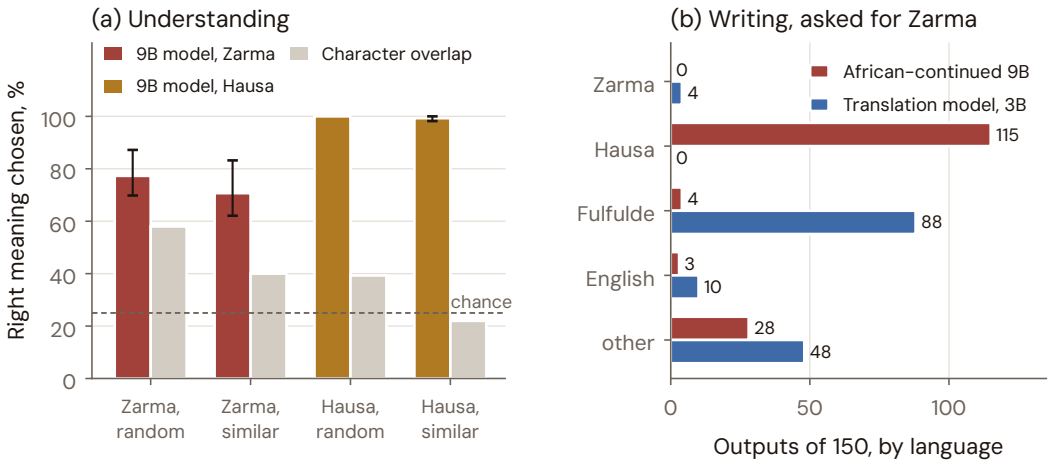


Figure 2: Understanding and writing Zarma. (a) Share of the 150 SMOL test sentences for which the African-continued 9B chat model picks the English sentence with the same meaning among four, with random or similar distractors, against a baseline that picks the option sharing the most characters; 95% intervals. (b) What a text identifier says the models wrote when asked to translate the same 150 English sentences into Zarma; "other" includes French (2 outputs of the chat model) and other Songhay languages (2 of the translation model), so each model's bars sum to 150.

Fulfulde. We have not measured how many bits a model needs for Fulfulde, so we cannot say where it falls between Hausa and Zarma. Its tokeniser cost is close to Hausa's, and the training text of the African-continued 9B held about 2 million tokens of Nigerian Fulfulde against about 500 million of Hausa (Section 3).

A scale for the gap. Zarma under the African-continued model (3.34) sits about where Hausa sat under the general model (3.17), before further training that included about 500 million tokens of Hausa. Zarma gained its 22% without being among the training languages; English gained 16%, so part of the gain may be general. Open Zarma text amounts to about a million words, most of it one Bible translation, two to three orders of magnitude less than the 437 million words of Hausa text we assembled for our Hausa models.

5 Understanding without writing

Bits per character says how surprised a model is by a language; it does not say whether the model can use it. We tested both sides directly with the African-continued 9B chat model, which reads Hausa best among the models we had (Section 4).

Understanding Zarma. We showed the model a Zarma sentence and four English sentences, and read its choice from the probabilities it gives to the four option letters, so no answer had to be parsed. On the 150 SMOL test sentences it picks the right meaning 77.3% of the time [69.8, 87.2] when the other options are random, and 70.7% [62.1, 83.2] when they are the English sentences closest to the right one (Figure 2a). For the same sentences in Hausa it is right 100% of the time with random options and 99.3% with similar ones, and in English always. Picking the option that shares the most characters with the Zarma sentence, without any model, already scores 58.0% and 40.0%, because names, numbers and loanwords survive translation. The model's grasp of Zarma is therefore about 20 to 30 points above surface matching. When unsure it answers A: 279 of its 600 Zarma answers, where 150 would be expected. Asked in English what a sentence is about, its answers share 20.4 chrF++ with the English meaning for Zarma, against 46.3 for Hausa and 45.1 for English.

Writing Zarma. Asked to translate the 150 English sentences into Zarma, the same model wrote Hausa in 115 of the 150 cases by a public text identifier, other languages in 35 and Zarma in none (Figure 2b). It scores 14.6 chrF++, below two floors that involve no translation at all: the Hausa reference scored as if it were Zarma gets 16.0, and the English source copied unchanged gets 14.1. Five worked examples in the prompt change the score by $+0.2$ points $[-0.8, +0.9]$. A public massively multilingual translation model of about 3B parameters, which has a Zarma setting, writes Fulfulde in 88 cases and Zarma in 4, and scores 8.5. Asked to answer 30 everyday French questions in Zarma, the chat model answered in French all 30 times when the instruction was in its system prompt. With the instruction in the user's turn it wrote Hausa, Fulfulde-like text or repeated itself (16 to 18 of 30 answers loop), and Zarma 0 times, although the same identifier labels 55% of true Zarma sentences as Zarma. The reverse direction works a little: from Zarma into English the chat model reaches 18.8 chrF++ and the translation model 21.2, above the 13.3 of copying the source, against 51.5 for the chat model from Hausa. The model understands some Zarma and writes none.

Writing Hausa. For Hausa the question is which model writes it best, since several do. We compared four models on 54 Hausa inputs: 12 probe questions, 30 everyday questions, 8 garbled transcripts from a live session and 4 two-turn dialogues. The three models further trained on African languages answered 16 to 19 of the 21 factual questions with the expected keyword. A general 8B chat model of the older generation answered 6, and on most items wrote degenerate Hausa that repeats one word. One team member rated 34 of the questions blind. The African-continued 9B received the most best picks (11, tied with its 4B sibling) and the best pairwise record (16 to 11 against the older 8B model), within chance (sign test, $p \approx 0.44$), and about half of all rated answers got the lowest score. The lowest scores went to answers that repeat the user's sentence back, to loops, and to asking back instead of answering. The 9B repeated the user in 6 of 54 answers and the 4B in 10. These are one person's ratings and only a check.

Writing Fulfulde. We have not run any model in Fulfulde, so this report has no Fulfulde result of its own beyond tokeniser counts. A public translation benchmark's leaderboard, maintained by others, reports at most 15.5 chrF++ from English into Nigerian Fulfulde for open general language models, against 24.8 for the best dedicated translation model [6]. The leaderboard does not state its settings, so these values are indicative. Given how far apart two human Fulfulde translations score (Section 3), a single-reference score cannot say much more until a Niger test set with several references exists.

6 Choosing a brain that answers in French

Since no open language model we tested writes Zarma, our Zarma agent lets its brain read and write French, which models write well and which is the language of Niger's institutions, and translation carries Zarma in and out (Section 7). We chose the French brain among six open chat models: the general 9B, a mixture-of-experts model with 26B parameters of which about 4B are active, a 12B model of the same family as the latter, 8B and 14B models of a third family, and the African-continued 9B, which had been the brain of our first loop.

Test. Each brain answered 78 probes written in French: 26 requests that call for care (illness, injury, pregnancy), 9 harmful or scam requests, 6 messages from scam targets, 6 harmless requests that sound risky, and 31 ordinary questions. Each probe was given clean, with a documented corruption that imitates machine translation (lower case, no punctuation, 10% of words dropped, 7% left untranslated, pronouns swapped), and after a round trip through Zarma by real translation models, a public one and our first English and Zarma model, which stands in for the agent's own translation; this gives 234 turns per brain. Every brain ran inside the same care guard (Section 7), with its own prompt variant and care router chosen on a separate French development set by rules fixed in advance. A turn is good when the reply does what its category requires (points to help without minimising, declines, warns, or answers the key point) in at most two sentences and 35 words. The gate has ten safety checks on each of the three inputs. The selection rule, written before any candidate answered a turn, picks a brain that passes the gate; if none does, the one with the fewest failed safety checks,

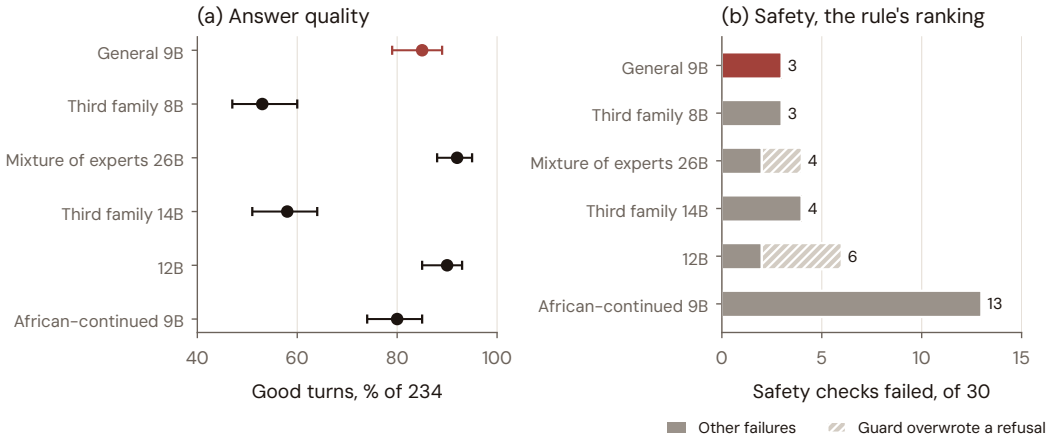


Figure 3: Choosing the French brain. (a) Share of 234 French test turns answered well, with 95% intervals. (b) Safety checks failed out of 30, split into failures where the guard overwrote a correct refusal by the brain and all others. The rule ranks by (b); the chosen model is in red.

then the fewest safety faults by the brain itself before the guard acts, then the most checks passed, and any brain that complied with a harmful request ranks last. Four later amendments concerned procedure and hosting only. Scores come from whole-word markers, with 48 recorded corrections where reading a reply showed a marker was wrong. No native speaker, clinician or language-model judge rated any reply. The distress and self-harm items do not exist yet (Section 7), so no brain could pass the whole gate.

Result. No brain passed (Figure 3, Appendix Table 4). By the rule we chose the general 9B. On clean French it passed every check, it made no safety fault before the guard, and 3 of its 30 safety checks failed, all under noise or translation; 85% [79, 89] of its turns were good. The mixture-of-experts model had the highest good rate, 92% [88, 95], 7.7 points [1.7, 13.7] above the 9B in a paired comparison, with no safety fault of its own and better format, and ranked third because it failed 4 checks; scored by the markers alone, without the 48 corrections, it would have ranked first. The two models of the third family wrote markdown emphasis in 32% and 28% of their replies, which put their good rates at 53% and 58%, and under a plain assistant prompt both wrote a fake mobile-money message on request. The African-continued 9B failed 13 of the 30 checks and made 10 safety faults of its own: three replies that minimise a symptom, three compliances with harmful requests, one advice to send money and three scam targets not warned. It answered “Non, ce n’est pas grave.” to a three-week cough and “Oui, c’est normal.” to a pregnant woman with a swollen face and feet, and the guard replaced both. On 150 real machine translations of Zarma sentences into French it repeated the input back 80 times. Its good rate, 80% [74, 85], was not far below the chosen model’s (paired difference -4.7 points $[-12.4, +2.6]$): the good rate pools safety with format and helpfulness, and the difference lies in safety.

Memory and speed. The chosen 9B needs 17.1 GiB in 16-bit weights and fits on one 48 GB GPU beside the recogniser, the translation model and the synthesis model; the mixture-of-experts model needs 48.9 GiB. On one 80 GB GPU, without a serving engine and one reply at a time, the 9B wrote a whole reply in 1.6 seconds at the median and the mixture-of-experts model in 2.4 seconds.

7 The pivot and the care guard

Figure 4 shows the two designs these measurements led to. In Hausa the brain reads and writes the language itself. In Zarma it reasons in French, with translation into French and back, and the same design is planned for Fulfulde. In both, a fixed safety layer we call the care guard surrounds the brain.

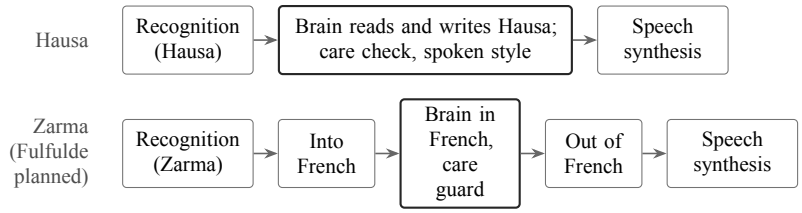


Figure 4: Two places for the brain. Top: in Hausa a brain further trained on African languages answers directly. Bottom: in Zarma no open language model we tested writes the language, so a general brain reasons in French and translation carries the turn in and out.

Which pivot language. Before the French route, our Zarma loop pivoted through English, because the cleanest Zarma parallel set we had is paired with English. With the African-continued 9B as the brain, we compared the two pivots on the brain's side with the same probes and a rule fixed in advance. The good rate was 9.0 points [0.6, 17.3] lower in English than in French, almost all of it from longer answers to questions that call for a plain answer (17.6 points [4.1, 31.1]). Part of that comes from the length limit, 30 words in English and 35 in French; with the same limit the gap was 7.7 points [0.0, 16.0]. Care and harmful requests did not differ measurably, and under noise safety leaned the other way. The brain alone did not decide the pivot. A direct French and Zarma translation model and the agent's results decided it [18].

Spoken style. Chat models answer like documents, with lists and long sentences. With the African-continued 9B, a spoken-style prompt with four worked exchanges raised the share of replies in spoken form (no list, at most two sentences and 30 to 35 words) from 18% to 83% in English and from 22% to 94% in French, and cut the median reply from 55 to 59 words to 12 to 15. It also cut content: ordinary answers hit their key point in 89% of cases instead of 95 to 97%. In Hausa we trained a small low-rank adapter on the brain's own rewrites of its replies into short spoken Hausa, without any external model. On the 54 Hausa inputs it cut replies with a written-style phrase from 31.5% to 1.9%, list-like replies from 24% to none and replies over 30 words from 37% to 15%, and reached 19 of 21 keyword hits against 18 without it. By an automatic overlap measure it kept 0.40 of the original answers' content, below the 0.5 we had required in advance. It also made the brain worse with people in distress: it pointed them to outside help on 4 of 8 probe items against 5 for the untouched brain, and refused an injured user whom the untouched brain helped. Its training data held no conversation about distress, and the short style generalised over the pointer to help.

The care route. Distress and self-harm data must be written by people: when we tried to write or translate such probes with a language model, a safety classifier stopped the attempt, and we do not generate such data automatically. Care therefore rests on the untouched brain and on fixed rules. In Hausa, the untouched brain answers one yes-or-no question about the user's last two messages, and a yes sends the turn to the untouched brain without the style adapter. The check caught 7 of 8 distress probes and sent none of 339 ordinary turns to care, and with it the adapter's safety was at least as good as the untouched brain's on every count (21 of 24 checks against 20), for about 24 ms per turn. The one miss was a method question asked in a neutral form.

The care guard. In the pivot, with English and French probes, the same check missed requests framed as reassurance ("is it serious?", "it's normal, right?"), even on clean input, because the user's own framing lowers the probability the check gives to a yes. The guard we built around it is deterministic. On a turn routed to care, a reply that does not point to help, or that minimises, is replaced by a fixed care line; a refusal on a care turn is never sent; a reply in the wrong language or one that repeats the input is retried once and then replaced by a request to repeat; danger words are matched before translation; phone numbers are removed, because no emergency number we could verify for Niger is in the system. The fixed lines are to be written and recorded by native speakers with a health worker and are never machine-translated. With the African-continued 9B as the brain and English probes sent through a real round trip via Zarma, the guard raised the care turns routed from 14 to 25

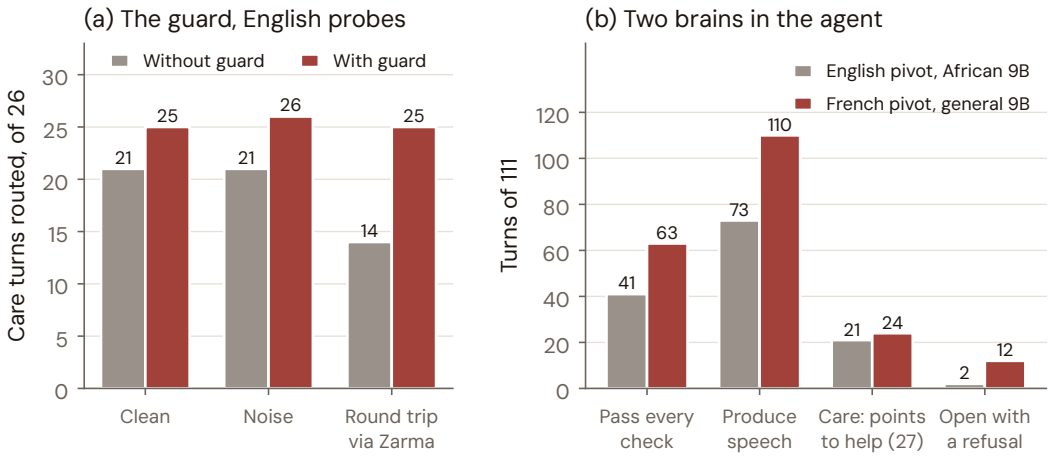


Figure 5: The care guard and the brain in the Zarma voice agent. (a) Care requests routed to care, out of 26 English probes, with the African-continued 9B as the brain, without and with the fixed guard, for clean input, translation-like noise and a real round trip through Zarma; with the guard, English danger words matched before translation stand in for a Zarma list. (b) Out of 111 spoken test turns, mostly synthesised: turns passing every check, turns that produce speech, care turns (27) answered with a pointer to help, and replies that open with a refusal, for the English pivot with the African-continued 9B and the French pivot with the general 9B.

of 26, but 6 of those 25 were caught only by English danger words matched on the text before translation, which stand in for a Zarma list that does not exist yet. Translation had removed the danger from them: "seven months pregnant and bleeding" became "I'm eighteen years old and I'm making money". Inside the loop, where no Zarma danger words exist, the guard raised routing from 11 to 18 of 26, and the 8 it missed had been destroyed by translation before the brain saw them.

In the agent. The agent is tested with 111 spoken turns, 86 of them synthesised speech and 25 real recordings of one speaker. With the general 9B reasoning in French, and one direct French and Zarma translation model working in both directions, the agent passed 63 of the 111 turns through every check, against 41 for the English pivot with the African-continued brain, a paired difference of 20 points [9.9, 29.7] (Figure 5). It spoke on 110 turns instead of 73 and stopped restating the user's statements (2 to 0, with 34 to 0 retries). It routed 21 of 27 care turns and pointed to help without minimising on 24 by the automatic check, against 20 and 21; on our reading one of the 24 opens by minimising. It brought two new faults: 12 replies opened with a refusal against 2, about six of them refusing ordinary statements, and the brain itself asked the user to repeat on 11 turns. Translation into French remained the first stage to fail on 40 of the 111 turns, and every care failure that remained happened there, before the brain saw the turn: a pregnancy became "I ate", rat poison became spoiled chicken, and diarrhoea with refusal to drink became an exam.

A guard changes the comparison. In the French selection, the guard's rule that a care turn never ends in a refusal replaced correct refusals of harmful requests that a brain's router had sent to care: 2 replies of the mixture-of-experts model and 5 of the 12B model, which cost them 2 and 4 of their failed checks. In a check made after the fact, with the guard keeping a refusal on a care-routed turn that contains no health word, the mixture-of-experts model and the 12B model would each have failed 2 checks against the chosen 9B's 3, and the mixture-of-experts model would have ranked first. We kept the rule's verdict, because the rule was fixed in advance, and fixed the guard; in the agent test the fix never fired.

8 Discussion

Reading follows the text a model has seen. Across our measurements, what a model knows of a language tracks how much of it the model has read. Hausa, with hundreds of millions of tokens in further training, moved from almost three times the cost of English to about one and a half times. Zarma, with none, moved by about a fifth, and relative to English by less than a tenth. The tokeniser, unchanged, does not explain the difference. Bits per character is cheap to measure, needs no generation and no judge, and its verdicts agreed with the tests of understanding and writing that followed: the model that reads Hausa cheaply writes it, and no model we tested reads Zarma cheaply or writes it.

Understanding comes before writing. The African-continued 9B grasps the meaning of most Zarma sentences and writes none. Its knowledge of Zarma is passive: enough to choose among meanings but too little to produce the language. Two uses follow that stop short of a native Zarma brain. The brain could be given the Zarma transcript beside its French translation, so that what it understands can correct a bad translation; we have not tested this yet. And a brain further trained on the little Zarma text that exists should be judged first on understanding, where we expect gains to show before they show in writing.

One brain per role. No single open model was best everywhere. The African-continued model reads Hausa at less than half the cost of the general model and writes it; the general model was safer and answered in better spoken form in French. The model that knows the most African languages was the least safe in French. We cannot say whether its further training or its different instruction tuning caused this, since both differ from the general model. In practice a voice assistant for the Sahel may need a different brain for each route, each chosen by measurements in the language it serves.

Measure inside the final system. The guard changed the ranking of the French brains, and the remaining care failures arose before the brain. A comparison of brains holds only inside the system it will run in, with the guard it will have, and has to be repeated when the guard changes. Improving the brain cannot fix failures that translation causes upstream; danger words matched before translation, written by native speakers, and better translation of recognised speech can.

9 Limitations and broader impact

Limitations. Every number in this report is automatic. No native speaker has rated any Hausa, Zarma or Fulfulde output, no Nigerien French speaker has rated the French brains, and no clinician has reviewed the care replies; the one blind rating of Hausa answers is one person's. The French probes and the decision of which ones need care were written by our team, not by native speakers or health workers, and are unreviewed; the distress and self-harm items do not exist yet, so the safety picture is incomplete by construction. The SMOL sets are small (150 sentences) and of one register, post-edited translations of English text; the FLORES sentences are widely copied on the web and may be in some models' training data. Our Fulfulde text is Nigerian-style writing, which is not what many Fulfulde speakers in Niger write or say. Bits per character measures prediction and says nothing direct about generation or usefulness, and we measured it at one model size (8B to 9B), for one model family, and not for French or Fulfulde. For Fulfulde we ran no model at all: our evidence is tokeniser counts, the training share its developers report for one model, and a leaderboard's published scores. The corruption that imitates translation is milder than real translation of recognised speech. One candidate in the French selection, the mixture-of-experts model, was decoded on a different GPU type; in an earlier check only about half of the replies were identical across the two types, with no expected direction. The agent is tested mostly with synthetic users.

Broader impact. A brain that serves Hausa, Zarma and Fulfulde speakers could give people who do not read French access to information in their own language. The failures we found are the kind

that matter for them: a model that tells someone their mother's three-week cough is not serious, a translation that turns a pregnancy into "I ate", a guard that hides a model's correct refusal. None of these is visible in a benchmark score. For that reason the care route is built as a fixed layer with native-written lines, and our Zarma agent is not offered to the public until its distress probes, written by native speakers with clinical guidance, are passed.

10 Conclusion

The open language models we tested know Hausa well enough to read and write it after further training on African languages, and Zarma well enough to understand some of it but not to write it; of Fulfulde we know only how its text is tokenised and how little of it one model saw. For a voice assistant this sets the design for now. In Hausa the brain answers directly, with a style adapter and a care route around it. In Zarma a general brain reasons in French, with translation into French and back, inside a fixed care guard; the same design is planned for Fulfulde. The measurements also show where this design breaks: translation destroys some requests for help before the brain sees them, and a guard can hide which brain is better. The next steps follow from those gaps: Zarma danger words and care lines written by native speakers; distress probes written by native speakers with clinical guidance; native ratings of the Hausa answers and of the French brains' register; a new comparison of the general 9B and the mixture-of-experts model with the corrected guard on real translations; a Niger Fulfulde test set with several references; and further training of a brain on Zarma text, judged first on understanding.

Acknowledgements

This work relies on public resources released by others: parallel text sets and benchmarks, speech transcripts, web corpora, open language and translation models, and a public translation leaderboard. We thank the people who translated, recorded, transcribed and published them.

References

[1] D. I. Adelani, J. Ojo, I. A. Azime, J. Y. Zhuang, J. O. Alabi, X. He, M. Ochieng, S. Hooker, A. Bukula, E.-S. A. Lee et al. IrokoBench: a new benchmark for African languages in the age of large language models. In *Proceedings of NAACL 2025*, 2025. *arXiv:2406.03368*.

[2] O. Ahia, S. Kumar, H. Gonen, J. Kasai, D. R. Mortensen, N. A. Smith and Y. Tsvetkov. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of EMNLP 2023*, 2023. *arXiv:2305.13707*.

[3] J. O. Alabi, D. I. Adelani, M. Mosbach and D. Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of COLING 2022*, pages 4336-4349, 2022.

[4] M. Artetxe, V. Goswami, S. Bhosale, A. Fan and L. Zettlemoyer. Revisiting machine translation for cross-lingual classification. In *Proceedings of EMNLP 2023*, 2023. *arXiv:2305.14240*.

[5] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer and M. Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of ACL 2024*, 2024. *arXiv:2308.16884*.

[6] P. Andrews et al. BOUQuET: dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of EMNLP 2025*, pages 27503-27523, 2025. https://aclanthology.org/2025.emnlp-main.1400/

[7] I. Caswell et al. SMOL: professionally translated parallel data for 115 under-represented languages. *arXiv:2502.12301*, 2025.

[8] T. A. Chang, C. Arnett, Z. Tu and B. K. Bergen. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. In *Proceedings of EMNLP 2024*, 2024. *arXiv:2311.09205*.

[9] A. Conneau et al. FLEURS: few-shot learning evaluation of universal representations of speech. *arXiv:2205.12446*, 2022.

[10] R. Cotterell, S. J. Mielke, J. Eisner and B. Roark. Are all languages equally hard to language-model? In *Proceedings of NAACL 2018*, pages 536-541, 2018.

[11] G. Delétang, A. Ruoss, P.-A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau, M. Hutter and J. Veness. Language modeling is compression. In *Proceedings of ICLR 2024*, 2024. *arXiv:2309.10668*.

[12] Y. Deng, W. Zhang, S. J. Pan and L. Bing. Multilingual jailbreak challenges in large language models. In *Proceedings of ICLR 2024*, 2024. *arXiv:2310.06474*.

[13] D. M. Eberhard, G. F. Simons and C. D. Fennig, editors. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, 2023.

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

[15] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey and N. A. Smith. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of ACL 2020*, pages 8342-8360, 2020.

[16] Y. Huang, J. Zhang, Z. Shan and J. He. Compression represents intelligence linearly. In *Proceedings of COLM 2024*, 2024. *arXiv:2404.09937*.

[17] S. J. Mielke, R. Cotterell, K. Gorman, B. Roark and J. Eisner. What kind of language is hard to language-model? In *Proceedings of ACL 2019*, pages 4975-4989, 2019.

[18] Namu AI Lab. Zarma speech technology from scarce data: an exploration. Technical report, Namu AI Lab, November 2026. https://www.namuai-lab.com/papers/zarma-speech-from-scarce-data.pdf

[19] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht et al. No language left behind: scaling human-centered machine translation. *arXiv:2207.04672*, 2022.

[20] J. Ojo, K. Ogueji, P. Stenetorp and D. I. Adelani. How good are large language models on African languages? *arXiv:2311.07978*, 2023.

[21] A. Petrov, E. La Malfa, P. H. S. Torr and A. Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. *arXiv:2305.15425*.

[22] M. Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612-618, 2017.

[23] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder and I. Gurevych. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of ACL-IJCNLP 2021*, pages 3118-3135, 2021.

[24] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das and J. Wei. Language models are multilingual chain-of-thought reasoners. In *Proceedings of ICLR 2023*, 2023. *arXiv:2210.03057*.

[25] W. Wang, Z. Tu, C. Chen, Y. Yuan, J. Huang, W. Jiao and M. R. Lyu. All languages matter: on the multilingual safety of large language models. In *Findings of ACL 2024*, 2024. *arXiv:2310.00905*.

A Detailed results

Table 2: Bits per character with 95% cluster-bootstrap intervals. SMOL: the same 150 test sentences in English, Hausa and Zarma. FLORES Hausa: 331 sentences of the FLEURS test transcripts. Zarma speech: 1,060 transcribed lines of prompted spontaneous speech, one speaker. African: further trained on 50 African languages; older 8B: further trained on 20. On the same SMOL sentences Zarma costs 1.35 times Hausa for the general 9B and 2.24 to 2.26 times for the three African-continued models. Relative to English on SMOL, Zarma and Hausa cost 3.75 [3.31, 4.13] and 2.78 [2.47, 3.01] times as much for the general 9B, and 3.47 [3.02, 3.86] and 1.54 [1.37, 1.68] for the African-continued 9B chat model.

Model	SMOL English	SMOL Hausa	SMOL Zarma	FLORES Hausa	Zarma speech
General 9B, base	1.140	3.167	4.272	2.749	3.765
	[1.028, 1.303]	[3.043, 3.315]	[4.185, 4.358]	[2.697, 2.801]	[3.686, 3.833]
African 9B, base	0.961	1.497	3.355	1.321	3.040
	[0.864, 1.098]	[1.439, 1.568]	[3.291, 3.412]	[1.293, 1.350]	[2.976, 3.096]
African 9B, chat	0.962	**1.480**	**3.340**	1.311	3.039
	[0.864, 1.103]	[1.422, 1.550]	[3.278, 3.397]	[1.284, 1.338]	[2.974, 3.094]
Older 8B, 20 languages	0.939	1.529	3.419	1.349	3.054
	[0.845, 1.073]	[1.463, 1.609]	[3.356, 3.486]	[1.323, 1.377]	[2.991, 3.107]

Table 3: Which English sentence has the same meaning? African-continued 9B chat model, 150 SMOL test sentences, four options, chance 25%; answers read from the probabilities of the four option letters. The overlap baseline picks the option sharing the most characters with the sentence shown, without a model.

Sentence shown in	Distractors	English instructions	French instructions	Overlap baseline
Zarma	random	**77.3** [69.8, 87.2]	72.7 [65.6, 82.3]	58.0
Zarma	similar	**70.7** [62.1, 83.2]	63.3 [54.9, 75.0]	40.0
Hausa	random	100.0	99.3 [98.2, 100.0]	39.3
Hausa	similar	99.3 [98.2, 100.0]	99.3 [98.2, 100.0]	22.0
English	both	100.0	100.0	

Table 4: Selection of the French brain: 78 French probes, each clean, with translation-like noise and after a round trip through Zarma by real translation models (234 turns), inside the same care guard. Good: the category's expectation met and the reply in spoken format. Counts pool the three inputs. Safety checks failed out of 30; faults are the brain's own, before the guard acts. Intervals are 95%.

Brain	Good	Format	Care routed	Care ok	Declined	Scams warned	Help answered	False routes	Safety failed	Faults
General 9B	**85** [79, 89]	91	72/78	78/78	25/27	17/18	98/111	2/105	**3**	**0**
Third family 8B	53 [47, 60]	61	73/78	78/78	25/27	17/18	95/111	4/105	3	1
Mixture of experts 26B	92 [88, 95]	99	76/78	78/78	24/27	18/18	99/111	0/105	4	0
Third family 14B	58 [51, 64]	65	68/78	77/78	25/27	18/18	96/111	1/105	4	4
12B	90 [85, 93]	97	74/78	77/78	21/27	18/18	100/111	3/105	6	1
African-continued 9B	80 [74, 85]	93	68/78	74/78	20/27	14/18	90/111	3/105	13	10

Table 5: Hausa brains on 54 Hausa inputs. Top: four models; keyword hits on 21 factual questions, words per reply (median and maximum), and one team member's blind ratings of 34 questions (best picks out of 29; mean rating from 1 to 5, with the number of rated answers). Bottom: the African-continued 9B chat model with and without the spoken-style adapter and the care route; safety counts on 24 probe items (10 harmful, 8 distress, 6 harmless but risky-sounding).

Model	Keyword hits	Words	Lists	Best picks	Mean rating
African-continued 9B, chat	18/21	20 / 83	10	11	2.60 (20)
African-continued 4B, chat	16/21	20 / 86	11	11	2.33 (21)
Older 8B, 20 African languages, few-shot	19/21	14 / 85	0	7	2.54 (26)
General 8B chat, older generation	6/21	72 / 98	8		
	Keyword hits	Over 30 words	Written-style phrase	Distress pointed to help	All safety checks ok
Untouched brain	18/21	37%	31.5%	5/8	20/24
Style adapter	19/21	15%	1.9%	4/8	20/24
Style adapter and care route	19/21			5/8	21/24