Discussion Paper Series

IZA DP No. 19029

October 2026

Count the Poor in AI's Language Benchmarks

Hai-Anh H. Dang
World Bank, IZA@LISER, GLO,
Indiana University, LSE and UEH

The IZA Discussion Paper Series (ISSN: 2365-9793) ("Series") is the primary platform for disseminating research produced within the framework of the IZA@LISER Network, an unincorporated international network of labour economists coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The Series is operated by LISER, a Luxembourg public establishment (établissement public) registered with the Luxembourg Business Registers under number J57, with its registered office at 11, Porte des Sciences, 4366 Esch-sur-Alzette, Grand Duchy of Luxembourg.

Any opinions expressed in this Series are solely those of the author(s). LISER accepts no responsibility or liability for the content of the contributions published herein. LISER adheres to the European Code of Conduct for Research Integrity. Contributions published in this Series present preliminary work intended to foster academic debate. They may be revised, are not definitive, and should be cited accordingly. Copyright remains with the author(s) unless otherwise indicated.

IZA DP No. 19029 October 2026

Count the Poor in AI's Language Benchmarks*

Abstract

A new coalition aims to let 3.4 billion speakers of underrepresented languages use AI in their own language and voice within five years, and it promises benchmarks to track progress. In an illustration with one large multilingual benchmark, weighting regions by where the world's extreme poor live, rather than by population, lowers scores by 4 to 8 percentage points for leading closed models and by about 14 for the best evaluated open-weight model. Many African languages also still need more tokens for the same text, which raises the cost of serving them. This essay suggests that the coalition report whether people in poverty can complete useful tasks safely, in a language they use, at an affordable cost.

JEL classification

I32, L86, O15, O33, Z13

Keywords

generative AI, language technology, benchmarks, poverty, poorer countries

Corresponding author

Hai-Anh H. Dang
hdang@worldbank.org

* I would like to thank colleagues at and outside the World Bank for helpful discussions. I am grateful to the UK Foreign, Commonwealth and Development Office (FCDO) for funding assistance through the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program.

From speakers to people in poverty

In September 2026, dozens of organizations, from model developers to African and Indian language groups and the World Bank Group, set a goal. Within five years, an estimated 3.4 billion speakers of underrepresented languages will be able to use AI tools "in their own language and voice"[1]. Progress is real. On Global PIQA, a test of everyday commonsense reasoning that more than 350 researchers wrote in 141 language varieties, the best model answers 94% of questions written by native speakers (Fig. 1a)[2]. One open model translates among 200 languages[3], and speech recognition now covers more than 1,600[4]. OpenAI's newer tokenizer needs about two-thirds fewer tokens than its predecessor for the same text in Hindi, Bengali or Urdu (Fig. 1b).

The commitment also looks beyond models. It notes that "connectivity, availability of electricity, and access to low-cost devices must also be addressed", and it wants value to flow back to communities "through local ownership"[1]. A workstream called "Tracking progress honestly" promises "assessments and benchmarks that measure real gains against the global goal". These commitments now need measurable standards broken down by poverty.

In 2025, an estimated 817 million people lived on less than US$3.00 a day, seven in ten of them in Sub-Saharan Africa[5]. A goal counted in speakers could be met while many of them are served poorly, expensively or unsafely. We suggest that the coalition report not only how many speakers its systems can reach, but whether people in poverty can complete useful tasks safely, in a language they use, at an affordable cost. Published benchmark results already allow a first look (Fig. 1).

Scoring where the poor live

Blasi *et al.* (2022) weighted results by speakers and found that the economic weight of a language's users, more than their number, predicts the research it receives[6]. We add a poverty

perspective. Using Global PIQA's published results, we average the scores of the language varieties in each of ten world regions. We then weight each region by its share of world population, or by its share of the world's people living on less than US$3.00 a day[2,5]. This exercise only weights regions, not languages, and it does not measure how poor users fare.

Guessing gives 50% on questions written by native speakers. On these, poverty weighting puts scores 4 to 8 percentage points below population weighting for the best closed model of each of three developers (Fig. 1a). Population-weighted scores are within about one point of the published averages. For the best evaluated open-weight model, Gemma 4 31B, the gap is about 14 points. The gap reflects geography. Sub-Saharan Africa has 16% of the world's people and 71% of its extreme poor. It is also the lowest-scoring region for each of the nine models that score 80% or more.

This gap does not isolate language. Locally written questions differ in difficulty and subject across languages. Translated questions fix the content but lose local context, so the two kinds of test are complementary[7]. On translated questions, which have four answer options, the gap is 2 to 12 points for the closed models and 13 to 16 for the open-weight model. The range depends on whether regions are summarized by means or medians.

Lingala shows how much the type of question matters. It is a lingua franca of the Democratic Republic of the Congo, home to about one in nine of the world's extreme poor[5]. The best model answers over 90% of translated Lingala questions but 72% of those that Lingala speakers wrote[2], and no open-weight model exceeds 67% on the latter. Each score rests on about 100 questions. Because both the questions and the number of answer options differ, this contrast is descriptive, not an estimated penalty for local questions.

Alternative regional summaries, poverty lines and leave-one-variety-out checks preserve the direction of the gap, although its size varies. Coverage is the larger caveat. Eight of the 17 Sub-Saharan varieties are spoken mainly in Nigeria. The South Asian varieties are mostly Indian, while 69% of South Asia's extreme poor live in Pakistan and Afghanistan. The set for the Americas and Oceania has four European-language varieties and Hawaiian, but no Indigenous language of Latin America. Neither alternative summaries nor a higher poverty line fixes these mismatches, and performance in the languages the benchmark omits is unknown.

Cost, speech and safety

Accuracy is not the only barrier. Commercial models charge by the token, a small unit of text. The same text can need up to 15 times as many tokens in one language as in another[8]. With OpenAI's newer tokenizer, most of the 16 African languages of the AfriMMLU benchmark[9] need 1.5 to 1.9 times as many tokens as English for the same FLORES-200 text[3] (Fig. 1b). Yoruba and Ewe need more than twice as many, and Amharic almost six times. Hindi, Bengali and Urdu now need 1.6 to 1.7 times, about the level of the median African language. More tokens also go with lower accuracy. In a study of ten models on AfriMMLU, the association is negative in every model and subject. Its authors put it at 8 to 18 percentage points per extra token per word on average, although only some estimates are statistically significant[10]. This does not show that a new tokenizer alone would close the gap.

Where people reach AI through free or flat-rate apps, per-token prices fall mainly on the services that serve them. At the same price per token, a language that needs 1.7 times as many tokens raises the matched text-token part of a hotline's bill by about 70%. Output length, retries, speech processing, telephony and human support also shape the total. The useful measure is the cost per safe completion: the end-to-end cost per successfully and safely completed task, including

failed attempts and retries, reported with the completion rate. Tokenizers can be designed for parity. Parity-aware byte-pair encoding cut cross-language tokenization inequality by up to 89% relative to standard byte-pair encoding in its authors' experiments, with little loss of compression[11]. Pricing public services at the English-equivalent token count is a financing or procurement option, but it does not remove the extra computation or guarantee accuracy.

For many poor users, especially those who read little, speech may be the most practical way in. But counts of supported languages do not show whether spoken exchanges work. Tests could cover the whole exchange, from hearing a question to giving a clear answer, with dialects, code-switching and noise. Safety needs the same scrutiny. In 2023, ChatGPT and GPT-4 answered the same harmful questions unsafely about three times as often in low-resource as in high-resource languages[12]. Tests could also cover wrong advice on harmless questions, escalation to human help and needless refusals.

What benchmarks cannot tell us

Adoption, useful performance and welfare are different questions. A recent review finds cross-country links between official language and AI use, but they are sensitive to the specification[13]. Randomized evidence on AI farm advice is only now appearing: a preprint posted in October 2026 reports a trial with 1,935 rice farmers in India and the Philippines. Farmers offered an AI adviser in Hindi, Tagalog or English reported changing some practices, compared with farmers offered a chatbot without farm advice. Their yields and costs did not change detectably over the one season studied[14]. A benchmark score is not a completed task, and a completed task is not yet a gain in welfare.

The central unknown is distributional. When a model improves in a language, do its poor speakers gain, or mainly better-off speakers, who may already use AI in a second language? In the

farm trial, the authors found little evidence that use or effects were concentrated among better-resourced farmers. But every participant could already use a messaging app, the trial did not vary language, and one season of one service cannot settle the question[14].

Reporting that counts the poor

Table 1 sets out five measures. For each, it names who would report it, the comparison and the trigger for action. These are tentative suggestions, not an established standard. Developers could report model results, token counts and safety tests with each major release. Service providers could report costs and voice-service tests. The coalition could coordinate an independent review and publish a yearly summary.

The headline could combine three numbers: the poverty-weighted score, the share of the extreme poor whose language for a service scores below a task-specific floor, and the gap to population weighting. The poverty-weighted score could start with regional weights, as in Fig. 1a, and move to language weights as survey data on the languages poor people use arrive. Trends would hold coverage and weights fixed, reporting new coverage and new weights separately. Progress would mean a higher poverty-weighted level and a smaller below-floor share, with safety floors met. A smaller gap alone could reflect worse performance elsewhere.

Rather than set a fixed target, the coalition could publish a baseline in 2027 and commission an independent review in 2028. Where the measures do not improve materially on the baseline, allowing for uncertainty, a funded remediation plan would follow. Funding would go to the bottleneck the review finds, such as evaluation, service design or deployment capacity, rather than training data alone.

A first version could link household surveys in a few countries to service tests, recording the language people use for a service and counting multilingual people once. Within each pilot

language, the same services could be tested with poor and better-off users, with safe task completion and user costs reported separately. Because a low cost to providers does not make a service affordable, reports could also show what users pay, including data or airtime, and how often cost prevents use. They would separate unmeasured coverage from measured failure, and the surveys would include nonusers, so that exclusion stays in view. The below-floor share would start in these countries and widen as survey data on language arrive. Small language communities would keep a minimum share of attention, rather than resources following only the largest poverty counts.

Communities could help set the bar. Paid community representatives could select tasks, set minimum standards and decide remediation priorities. Floors would be task-specific and set with local experts; English is a comparison, not a safe threshold. AfriMed-QA, a pan-African English-language benchmark from over 60 medical schools in 16 countries, shows what domain tests look like[15]. It does not validate a clinical service in Lingala. Consequential health advice could keep human backup in all cases.

Statistics offices could add questions on AI use, language and mode (text or voice) to household surveys, building on a draft module[13]. Safety and service tests would come before deployment, and welfare evaluations would follow for services that scale. Funders could pay for randomized trials of local-language services, with consumption, health and learning as outcomes. Publishing only aggregates helps, since home language can stand in for ethnicity, but it does not alone secure the "privacy, consent, and data sovereignty" that the commitment promises[1]. Communities would need a say over the collection, use, access and retention of their data.

A better test

These suggestions are not specific to one coalition. Any developer could add a version of its multilingual results weighted by where poor people live, with coverage gaps stated. Any buyer

of language AI could ask for the cost per safe completion and for speech and safety tests with the communities served.

The pledge will count the people who could use AI in their language. A better test is whether a mother in Kinshasa could ask aloud, in Lingala, whether her child's fever needs a clinic, get a correct answer, and afford to ask again. Any such service would first need to be evaluated.

References

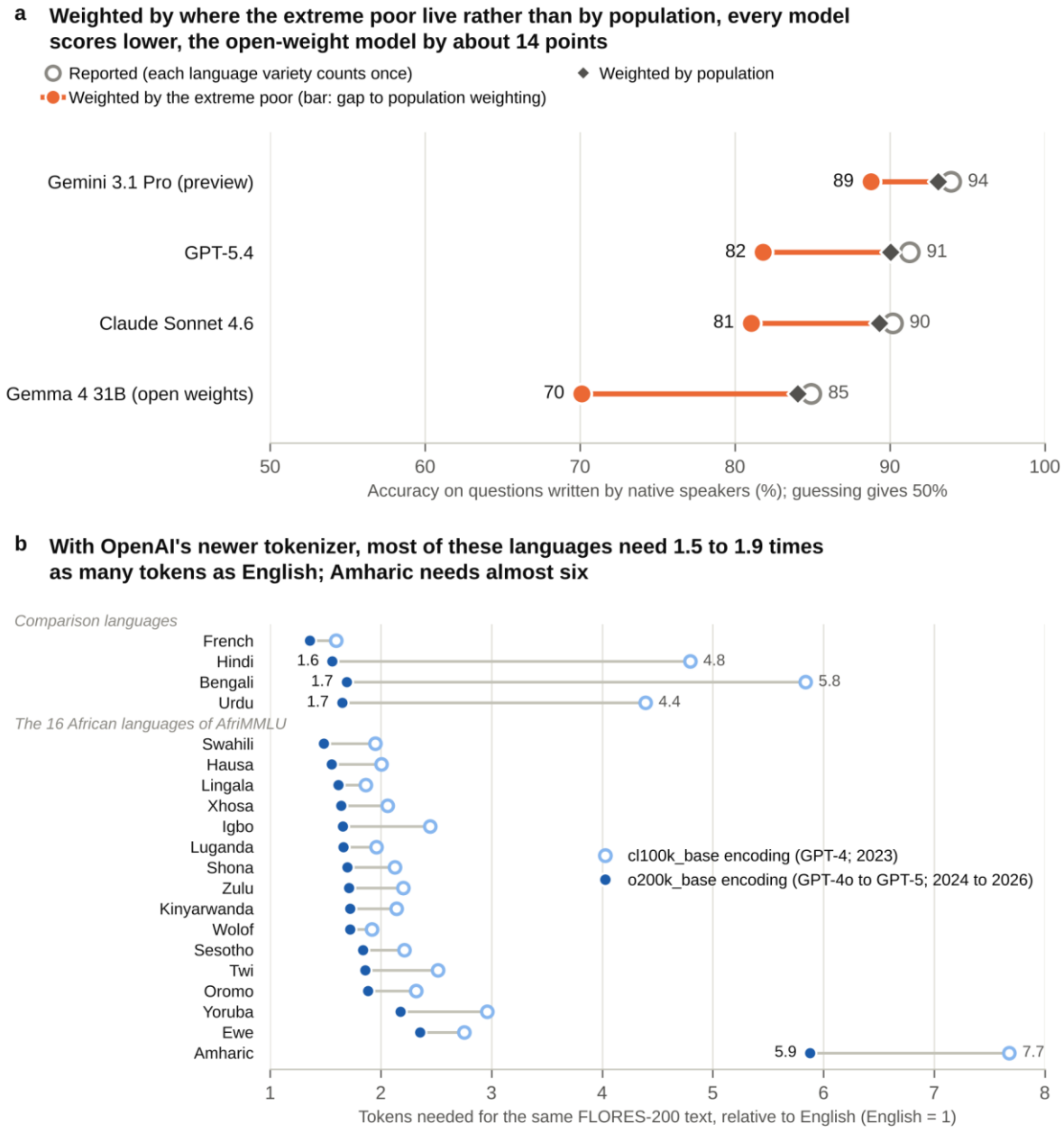
1. Gates Foundation. *Ensuring AI Speaks Everyone's Language: A Joint Commitment*. Web page, launched September 21, 2026 (Gates Foundation, 2026); gatesfoundation.org/ideas/progress/ai-for-good/language-commitment.
2. Chang, T. A. *et al.* Preprint at https://arxiv.org/abs/2510.24081v2 (2026).
3. NLLB Team. *Nature* **630**, 841–846 (2024).
4. Omnilingual ASR team. Preprint at https://arxiv.org/abs/2511.09690 (2025).
5. World Bank. *Poverty and Inequality Platform*, version 20260922_2021_01_02_PROD (World Bank, 2026); pip.worldbank.org.
6. Blasi, D., Anastasopoulos, A. & Neubig, G. in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 5486–5505 (Association for Computational Linguistics, 2022).
7. Singh, S. *et al.* in *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 18761–18799 (Association for Computational Linguistics, 2025).
8. Petrov, A., La Malfa, E., Torr, P. H. S. & Bibi, A. in *Advances in Neural Information Processing Systems 36* 36963–36990 (Curran Associates, 2023).
9. Adelani, D. I. *et al.* in *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* 2732–2757 (Association for Computational Linguistics, 2025).
10. Lundin, J. M. *et al.* in *Proc. 7th Workshop on African Natural Language Processing (AfricaNLP 2026)* 103–112 (Association for Computational Linguistics, 2026).
11. Foroutan, N. *et al.* in *Proc. 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 7514–7538 (Association for Computational Linguistics, 2026).
12. Deng, Y., Zhang, W., Pan, S. J. & Bing, L. in *Proc. 12th International Conference on Learning Representations* (ICLR, 2024).
13. Dang, H.-A. & Do, M. N. N. *Can AI Improve Household Welfare in Poorer Countries? Evidence and Measurement Challenges*. IZA Discussion Paper 18984 (IZA, 2026).
14. Aouad, A., Sungu, A., Taneri, N. & Ucel, C. Preprint at https://ssrn.com/abstract=7572761 (2026).
15. Nimo, C. *et al.* in *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1948–1973 (Association for Computational Linguistics, 2025).

Table 1. Five measures for language benchmarks that count the poor

No.	Measure	What is reported	Who reports, and when	Compared with	If it falls short
1	Who is served	Poverty-weighted score; share of the extreme poor below a task-specific floor; gap to population weighting, on fixed coverage and weights	Coalition, yearly; independent review	Baseline published in 2027	2028 review; funded remediation for material shortfalls
2	Quality in text and speech	Tasks chosen with paid community representatives; full spoken exchanges with dialects, code-switching and noise; results for poor and better-off users	Developers (models) and providers (voice), each release	Task-specific floors set with local experts; English as reference; better-off users	Human backup until floors are met; always for consequential health advice
3	Cost	Tokens for the same FLORES-200 text relative to English; cost per safe completion, including failed attempts and retries, with the completion rate; what users pay, including data or airtime; cost-related nonuse	Developers (tokens); service providers (cost per task); pilot surveys (user costs)	English-equivalent text; previous release; better-off users	English-equivalent pricing; tokenizer redesign; lower user charges, such as zero-rated data
4	Safety	Harmful-prompt compliance; incorrect advice on harmless questions; escalation; unnecessary refusals	Developers, each release; independent review	Floors set with local experts	Deployment paused in that domain
5	Use and outcomes	Who uses AI, in which language, by text or voice, and by income; effects on consumption, health and learning	Statistics offices (surveys); funders (trials)	Nonusers (descriptive); randomized or other credible designs (causal)	Scale-up conditional on task-appropriate evidence, safety standards and continued monitoring

Sources: by row, (1) refs 5 and 6; (2) refs 4, 7 and 15; (3) refs 3, 8, 10 and 11; (4) ref. 12; (5) refs 13 and 14. *Note:* Proposed measures, roles and triggers, not an established standard.

Figure 1. Weighting a benchmark by where the extreme poor live lowers AI scores, and many African languages still need more tokens



Sources: panel a, ref. 2 (with its results repository) and ref. 5; panel b, refs 3, 8 and 9; our calculations. *Note:* Illustrative regional reweighting, not an index. Panel a: Global PIQA questions written by native speakers, two answer options, 133 of 136 varieties (results as of May 2026). "Reported" is the published average over all 136, so it differs in coverage from the other markers. The other markers weight ten regional means by each region's share of world population or of people below US$3.00 a day (2021 PPP; 2025 nowcasts); the orange bar is the gap. Models are each developer's best-scoring closed model and the best open-weight model evaluated, not necessarily current products. Panel b counts tokens for the same FLORES-200 text, with model mappings as of October 2026.