

Generated Data, Real Consequences: Why LLMs Are Inappropriate for Development Aid Targeting

M. Merritt Smith

School of Information
University of California, Berkeley
merrittsmith@berkeley.edu

Recent work demonstrates that large language models (LLMs) can match group-level means in US public opinion polling contexts, and can even match group-level distributions after fine-tuning (Argyle et al. 2023; Bisbee et al. 2024; Suh et al. 2025). Development economics appears a natural extension: household surveys are the primary data infrastructure for poverty measurement and program targeting, and they are expensive. A single round of the World Bank’s Living Standards Measurement Study (LSMS) costs US\$830k–2.5 million, and the Bank has committed to 390 multi-topic household surveys of this kind before 2030 (World Bank 2023). If LLMs could generate realistic household data from demographic information alone, the savings would be transformative, and the demand is institutional and active. The World Bank’s 2025 Poverty Prediction Challenge was motivated by the fact that consumption surveys “are conducted infrequently due to cost and logistical constraints” (World Bank 2025), and survey-to-survey imputation is already an accepted substitute where consumption is unobserved (Dang et al. 2025). In the adjacent survey industry the substitution has already happened: in a 2024 survey of over 3,000 market researchers, 71% expected synthetic responses to exceed half of data collection within three years (Qualtrics 2025). The question is therefore not whether anyone has announced a plan to replace survey rounds with an LLM, but whether such a plan would induce harm, who would bear it, and whether it could be justified against alternative methods.

The case for testing LLMs in this role is not obviously wrong. Pretraining on web-scale corpora leaves LLMs with substantial factual knowledge they can retrieve at inference time (Petroni et al. 2019; Roberts, Raffel, and Shazeer 2020), drawn both from outside accounts of how demographics correlate with consumption in low- and middle-income settings and from digital traces of those populations themselves. A country-specific supervised model is fitted to a comparatively small sample and can access neither; an LLM prompted with the same demographic inputs could in principle bring both to bear. That an aggregate shortfall appears for populations under-represented in web text is well established (Santurkar et al. 2023; Bender et al. 2021; Naous, Ryan, and Ritter 2024), so a performance gap is

a reasonable prior. What this paper adds is a characterization of its shape: how it varies with prompt granularity, who bears its costs, and what mechanism produces it. The stakes differ from the polling case in structure as well as severity: opinion research aggregates views across groups to estimate population-level sentiment, while targeting allocates scarce benefits to specific households, where errors have direct welfare consequences and the legitimacy of the decision depends on the evidentiary adequacy of its inputs (Porter 1996).

We evaluate four prompting strategies against 131,906 households from 17 nationally representative household surveys, one per country, spanning sub-Saharan Africa, South Asia, and Latin America. The surveys come from the LSMS and from country-specific systems, harmonized in Sahoo et al. (2025); the design is deliberately cross-country so that no one survey drives the conclusions. The four granularity levels correspond to deployment scenarios: L1 supplies only administrative geography, analogous to a poverty map; L2 adds head age bins; L3 adds an individual persona, the census or registry case in which a country holds demographic records but no current consumption survey; and L4 adds rich household context. At each level the LLM generates synthetic answers to that country’s survey questions, which pass to a country-specific proxy means test (PMT) (Grosh, Baker et al. 1995) producing a predicted consumption that ranks households under a fixed budget. The comparison at every level is against a level-matched True-PMT fitted to the same real inputs without an LLM in the loop, which asks whether routing available data through an LLM adds value over fitting a supervised model to it directly. We use Claude Sonnet 5 as the primary generation model and Qwen2.5-7B-Instruct as the open-weights comparison, which also permits fine-tuning and inspection of internal representations.

The LLM route underperforms the level-matched True-PMT at every level of data richness, and the gap is widest at L3, closest to how a social registry would deploy such a tool. The mechanism is a specific pattern of country-level stereotyping: the LLM’s synthetic responses concentrate on country-typical values rather than discriminating among households within a country. At L3, country identity explains 0.51 of the variance in the LLM’s downstream consumption predictions against 0.12 in true household consumption, and adding household detail to the prompt

does not move the model off the country prior. The effect is training-text driven rather than prompt-structure driven: running the identical pipeline on 10,000 US households from the American Community Survey, substituting state for country, inverts the pattern. The targeting consequence is that within-cell variance collapses and poor households are classified as non-poor outright: L3 exclusion errors average 49% across countries against 36% for the matched baseline. That excess error does not concentrate on any one demographic group outside of household size: it is dispersed across groups, redistributed from smaller households onto larger ones, and in relative terms grows with depth of poverty, falling hardest on households furthest below the line.

The failure is structural, and the interventions usually proposed to improve LLM performance do not resolve it. Fine-tuning on empirical survey distributions, following Suh et al. (2025), reduces the model's marginal distributional divergence by 36% on average, lifts household-level rank correlation in every evaluable country, and lowers the country stereotyping to at or below the level in the surveys themselves. Targeting does not follow: exclusion error falls five percentage points, and in only half the countries, because the fine-tuned model already sits near the information ceiling of its inputs. Fine-tuning teaches the model correct group-level marginals, but within-cell discrimination requires information the demographic prompt does not contain, and no amount of fine-tuning supplies it. Scaling does not close the gap either: what the training text encodes about a country's poor households does not change with capability. A linear ridge probe on the model's internal representations, adapting Jahanparast, Hong, and Chang (2026), recovers no more within-cell signal than the generated text does, locating the collapse in the representations themselves rather than in the decoding step.

We further argue that even were this shortfall to close, LLM-generated data would remain disqualifying for allocative decisions about specific households, for reasons that do not depend on its current accuracy: its inferences are group-level inferences not grounded in observations of the population the decision affects, drawn from a corpus that is undisclosed and, on its publicly characterized slice, represents these populations sparsely. A household cannot contest an inference about households like theirs, because no observation of that household enters the chain of reasoning. The PMTs we use as the foil throughout are imperfect on this dimension too; the normative argument rests on the relative difference.

References

- Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
- Bisbee, J.; Clinton, J. D.; Dorff, C.; Kenkel, B.; and Larson, J. M. 2024. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4): 401–416.
- Dang, H.-A. H.; Kilic, T.; Abanokova, K.; and Carletto, C. 2025. Poverty Imputation in Contexts without Consumption Data: A Revisit with Further Refinements. *Review of Income and Wealth*, 71(1).
- Grosh, M.; Baker, J. L.; et al. 1995. Proxy means tests for targeting social programs. *Living standards measurement study working paper*, 118: 1–49.
- Jahanparast, E.; Hong, Z.; and Chang, S. 2026. What Do Large Language Models Know About Opinions? In *International Conference on Learning Representations (ICLR)*.
- Naous, T.; Ryan, M. J.; and Ritter, A. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2463–2473.
- Porter, T. M. 1996. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.
- Qualtrics. 2025. 2025 Market Research Trends Report. Survey of more than 3,000 market researchers across 14 countries, fielded October 2024.
- Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5418–5426.
- Sahoo, R.; Blumenstock, J.; Niehaus, P.; Selker, L.; and Wager, S. 2025. What Would It Cost to End Extreme Poverty? Working Paper 34583, National Bureau of Economic Research.
- Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *International Conference on Machine Learning (ICML)*.
- Suh, J.; Jahanparast, E.; Moon, S.; Kang, M.; and Chang, S. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 21147–21170.
- World Bank. 2023. World Bank Group's Target for Household Surveys. Development Data Group.
- World Bank. 2025. Poverty Prediction Challenge: Advancing Real-Time Monitoring of Poverty through Data Science. World Bank Measuring Poverty program brief. Launched November 2025; closed 4 February 2026 with 1,322 registered participants.