Who Does AI Work For?

Measuring, Tracing, and Redesigning AI Across Populations

AI progress is usually reported in aggregates: higher benchmark scores, broader language coverage, stronger capabilities. But an average tells us how well a system works, not *for whom*. The same model can perform very differently across socioeconomic, geographic, and cultural contexts, while standard evaluations often hide these differences. My research asks: **Who does AI work for, and how do the choices we make in building and evaluating AI shape that answer?** I pursue this question through three connected directions. I **MEASURE** variation obscured by aggregate evaluation, **TRACE** disparities to representation, provenance, and dataset construction, and **REDESIGN** data and evaluation practices to improve outcomes across populations. Together, these form a loop: identify where AI works unevenly, understand what upstream choices produce those differences, and test what we can change.

MEASURE: Revealing Who Aggregate Evaluation Leaves Behind

My evaluation research asks what becomes visible when model performance is disaggregated across populations. In *Bridging the Digital Divide (EMNLP 2023)* [1], we evaluated CLIP on 38,479 Dollar Street images from 63 countries [2] and found that retrieval performance increased systematically with household income. In the poorest income quartile, 86.5% of images were missed. These disparities reflected not only model behavior but also differences in how apparently universal concepts appear in the world: "toilet paper," for example, could refer to commercial tissue, grass, leaves, newspaper, or water. Lower-income households exhibited greater within-category visual diversity, showing how a single category label can conceal substantial population variation.

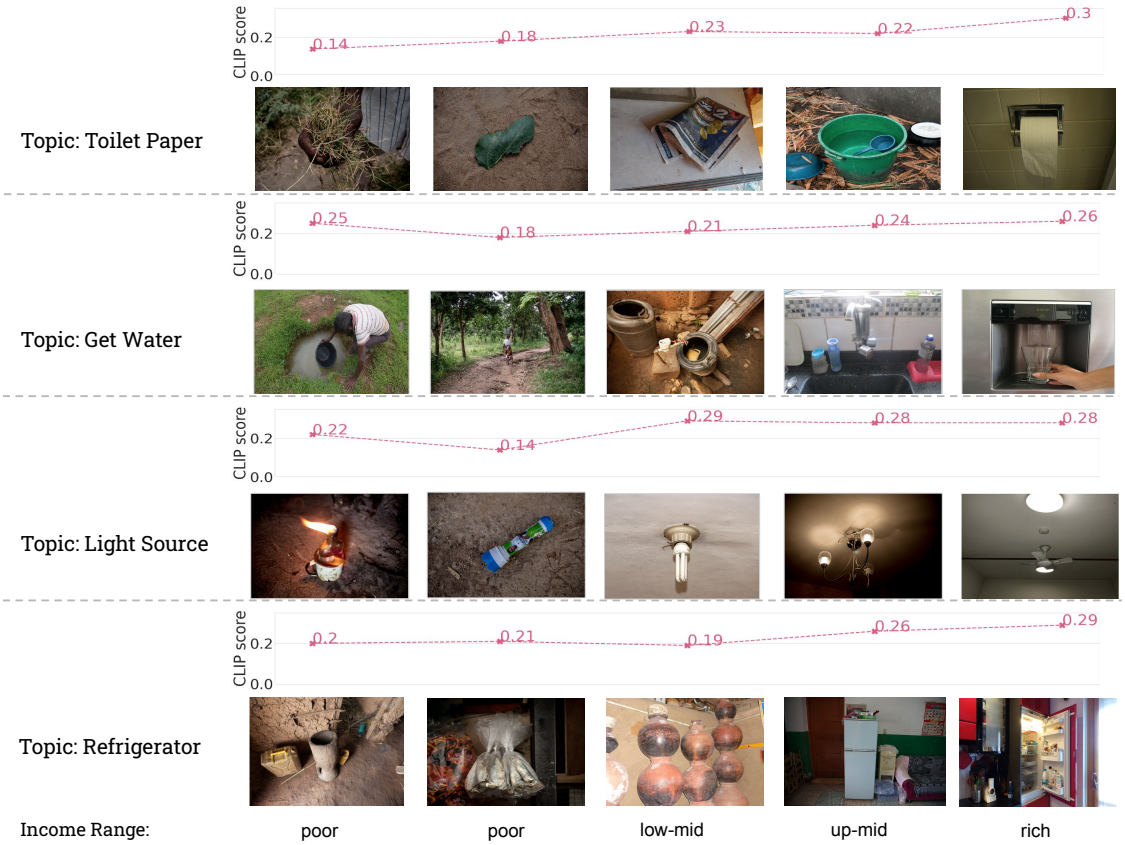

Figure 1: Everyday categories take markedly different visual forms across income levels; CLIP scores are shown above each example. [1]

I extend this principle to NLP in *Language Is Not Enough* [3]. Language-level evaluation often implicitly treats performance in one language as evidence about all populations who speak it. We instead hold language, task, dataset

source, and evaluation format constant while varying country. Across frontier, multilingual, and Arabic-specialized models, we find 19–32 percentage-point performance gaps among countries sharing Modern Standard Arabic, with similar disparities in a second academic-knowledge setting. These differences persist under country-specific prompting and controls for measurable dataset differences.

I have extended the same measurement principle to culturally diverse multilingual VQA and multimodal fairness evaluation [4, 5]. Across these settings, the lesson is consistent: *capability is not a population-invariant property*. Population-aware evaluation should therefore ask not only whether a system can perform a task, but how reliably that capability transfers across the populations a benchmark is presumed to represent.

That raises the next question: when performance differs across populations, what upstream conditions help produce, or conceal, those differences?

TRACE: From Performance Gaps to Representation and Provenance

Dataset documentation frameworks have established the importance of provenance, intended use, and representational limits [6, 7], but geographic representation remains difficult to observe at scale. The *Language Is Not Enough* evaluation itself depends on knowing which populations NLP datasets actually represent, a question that language metadata alone cannot answer.

To make this information visible, I led the development of *AtlasNLP (EMNLP 2026)* [8], a country-aware atlas of geographic representation across NLP research. Starting from nearly 120,000 ACL Anthology records, we developed an audited pipeline yielding 13,462 primary dataset-contribution records that separately track represented countries, producer countries, tasks, and languages.

AtlasNLP reveals substantial geographic sparsity: 79.2% of country-task pairs have no dataset record with explicit country representation. Representation and production are also geographically asymmetric. Among countries with at least ten represented records and ten producer associations, 39 of 62 are represented primarily through datasets produced by institutions elsewhere. Most importantly, language coverage often substantially overstates geographic representation.

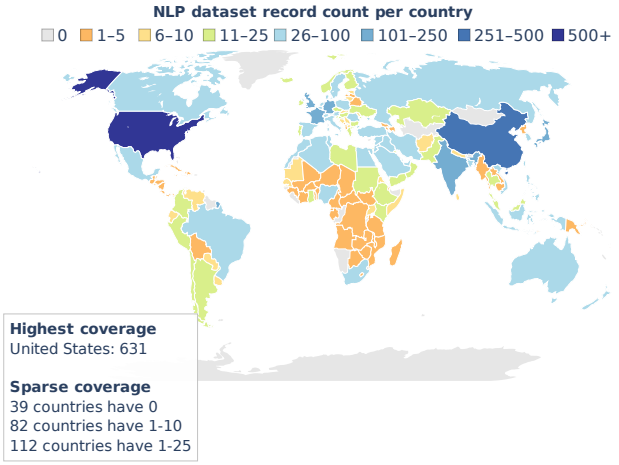

Figure 2: Country-level distribution of represented NLP dataset records in AtlasNLP-Core [8].

These findings distinguish *coverage from representation*. A dataset may include a language without meaningfully representing the populations who use it; a benchmark may appear globally broad while drawing evidence from a narrow set of countries or producers. By making representation, provenance, and production separately observable, AtlasNLP gives us a way to trace downstream disparities back into the data ecosystem rather than treating them as isolated model failures. This perspective also connects to my collaborative work on globally representative AI, which argues that representation depends not only on what data is collected but also on who participates in defining, constructing, and evaluating AI systems [9].

Once those upstream choices become measurable, they also become possible intervention points.

REDESIGN: Turning Representation into an Intervention Point

My third direction asks what we can change once disparities and their sources become visible. In *Uplifting Lower-Income Data (NAACL 2025)* [10], I tested lightweight interventions for improving retrieval of lower-income and non-Western images. Translating prompts into country-relevant languages generally did not improve performance, while adding geographic context improved lower-income retrieval in 38 of 42 countries. In complementary collaborative work, I use geographic data similarity to target annotation effort under limited budgets [11]. Together, these results

show that interventions must match the source of a representational gap; linguistic localization or indiscriminate data expansion is not necessarily enough.

Figure 3: CAA groups culturally different objects by shared function, such as sleeping and cleaning teeth [12].

Culture Affordance Atlas (CAA) [12], published at AAAI 2026, moves the intervention further upstream by changing the representation scheme itself. Drawing on affordance theory [13, 14], CAA organizes culturally diverse objects by what they are used to do rather than only by what they are called. A bed, Nepali charpai, and Malawian straw mat all support sleeping; a toothbrush, chewing stick, and fingers can all support cleaning teeth. The resulting atlas contains 46 functions and 288 objects grounded in ethnographic evidence, and function-based labeling reduces socioeconomic performance gaps by a median six percentage points in CLIP, with the effect reproduced on SigLIP2.

CAA shifted my view of representation from a question of simply "adding more diverse data" toward a broader design problem. Labels, ontologies, sampling choices, and evaluation categories determine what variation a model is able to learn and what failures an evaluation can reveal. Across my work, interventions therefore move progressively upstream: from prompting and targeted annotation to representation schemes and dataset construction.

Future Research: Toward Population-Centered AI

Near-Term: Closing the Loop AtlasVL will extend AtlasNLP's country-aware analysis to vision-language datasets, allowing me to connect where populations are missing from multimodal data ecosystems with where models perform poorly. I will test which properties of the data ecosystem predict population-level disparities and use those findings to prioritize data collection and evaluation. Building on CAA, I will scale function-centered representations and study how representation-aware data design, training, and post-training affect performance across populations. Building on *Language Is Not Enough*, I will investigate when country-specific or country-adapted language models reduce within-language disparities that persist in global and language-specialized models.

These projects form the foundation of the interdisciplinary research group I hope to build. Leading AtlasNLP across 19 collaborators and six institutions, while mentoring undergraduate and master's researchers, has shaped my approach to collaborative and student-led research.

Long-Term: Population-Centered AI My current work asks whether AI capabilities transfer across populations and which development choices shape those outcomes. Longer term, I want to move from *capability-centered* toward *population-centered AI*: beginning with people's goals, constraints, and desired outcomes, then asking what capabilities are useful and how systems and evaluations should be designed around them.

Population-centered evaluation should account for more than accuracy. My collaborative work on multilingual inference shows that the energy cost of using the same model can vary substantially across languages [15]; in work mapping poverty challenges to AI-driven solutions, we begin with consequential community needs before asking whether AI is an appropriate intervention [16]. These directions push my agenda toward evaluating value, cost, reliability, and fit for particular populations, and toward human-centered questions about whether benchmark gains translate into outcomes users actually experience and how usefulness varies across contexts.

Ultimately, I want the question "Can AI do this?" to be accompanied by an equally rigorous one: *For whom does it work, and what would we build differently if we started there?*

References

[1] J. Nwatu, O. Ignat, and R. Mihalcea. "Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models". In: *EMNLP*. 2023.

[2] W. Gaviria Rojas et al. "The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world". In: *NeurIPS* (2022).

[3] J. Nwatu, Z. Bashir, and R. Mihalcea. "Visible Creators, Invisible Populations: Why Language Alone Is Not Enough for NLP Evaluation". Manuscript in preparation. 2026.

[4] D. Romero et al. "CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark". In: *NeurIPS*. 2024.

[5] C. Chen et al. "MultiBBQ: Fairness Failure Modes of Multimodal LLMs". Under review. 2026.

[6] E. M. Bender and B. Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science". In: *Transactions of the Association for Computational Linguistics* (2018).

[7] T. Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* (2021).

[8] J. Nwatu et al. "AtlasNLP: A Country-Aware Atlas of Dataset Representation in NLP". In: *EMNLP*. 2026.

[9] R. Mihalcea et al. "Why AI Is WEIRD and Shouldn't Be This Way: Towards AI for Everyone, with Everyone, by Everyone". In: *AAAI*. 2025.

[10] J. Nwatu, O. Ignat, and R. Mihalcea. "Uplifting Lower-Income Data: Strategies for Socioeconomic Perspective Shifts in Large Multi-modal Models". In: *NAACL*. Association for Computational Linguistics, 2025.

[11] O. Ignat, L. Bai, J. Nwatu, and R. Mihalcea. "Annotations on a Budget: Leveraging Geo-Data Similarity to Balance Model Performance and Annotation Cost". In: *LREC-COLING*. 2024.

[12] J. Nwatu, L. Bai, O. Ignat, and R. Mihalcea. "Culture Affordance Atlas: Reconciling Object Diversity Through Functional Mapping". In: *AAAI*. 2026.

[13] J. J. Gibson. "The theory of affordances:(1979)". In: *The people, place, and space reader*. Routledge, 2014.

[14] D. A. Norman. "Affordance, conventions, and design". In: *interactions* (1999).

[15] N. Deng et al. "The Language–Energy Divide: Measuring Energy Costs of Multilingual LLM Inference". In: *EMNLP*. 2026.

[16] L. Lein et al. "Mapping Poverty Challenges to AI-driven Solutions". Manuscript. 2026.