# Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models

Joan Nwatu*, Oana Ignat*, Rada Mihalcea
University of Michigan, Ann Arbor, MI, USA

EMNLP 2023

## Motivation

**Foundation models** require copious amounts of data and computing resources. Therefore, can only be developed by a privileged few.

*Given that foundation models are **developed by a few and used by many**, do these models work well for everyone?*

## What did we do?

*We investigated the performance of **CLIP** on people from **different economic backgrounds***

### Dataset: Dollar Street *(Rojas et al., 2022)*

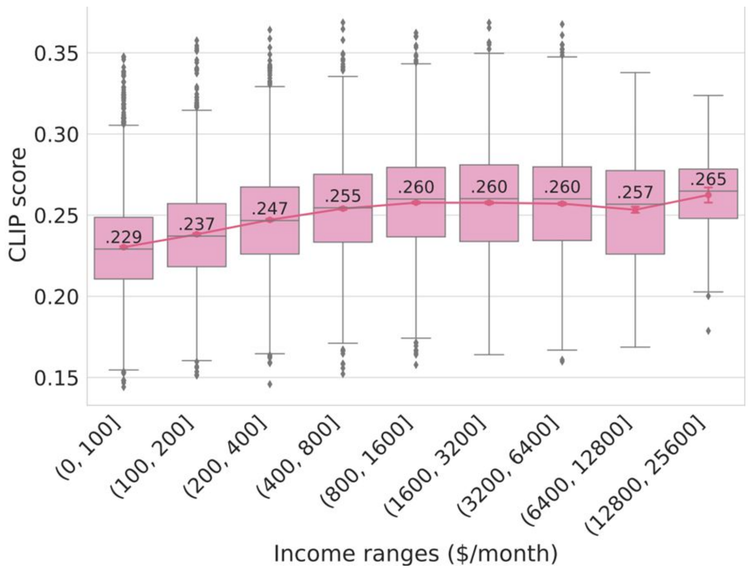

*Household images from various income levels and countries for the topic 'stove'*

### Actionable steps and Lessons learned

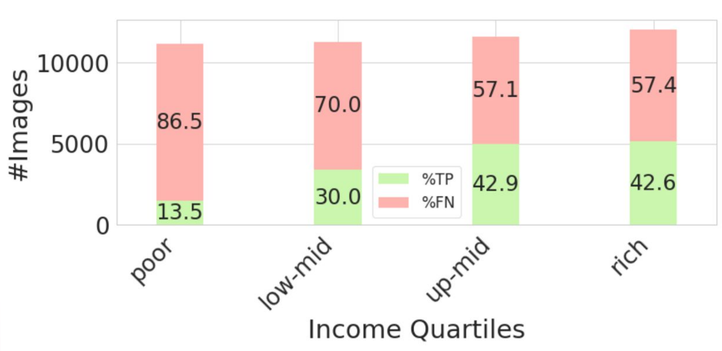
- The Digital Divide in vision-language performance exists.
- Evaluation metrics should represent everyone
- Match diversity standards to model purpose
- Document training data.
- Annotate diversity and subjectivity in datasets.
- Invest in geo-diverse datasets.

## Key Insights

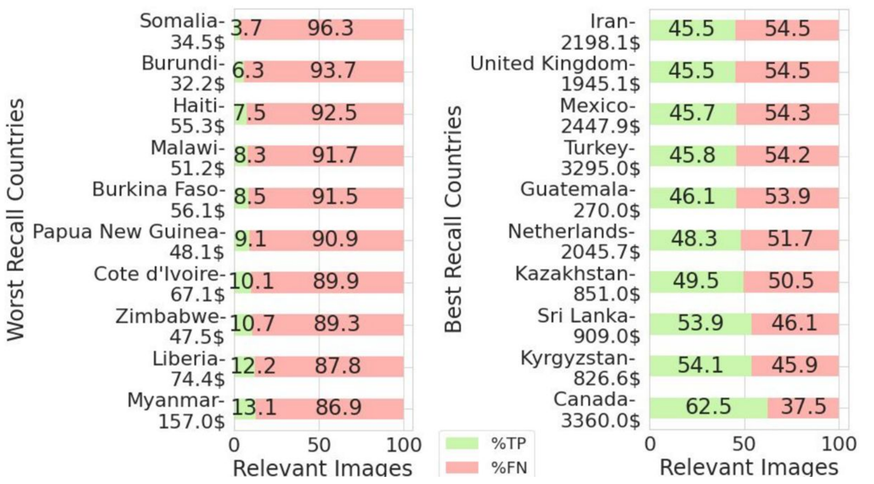
### *CLIP Performance varies across income and is consistently lower for poorer groups*




*Median CLIP alignment scores across images in Dollar Street from different income ranges, together with average CLIP scores with confidence values for each range.*

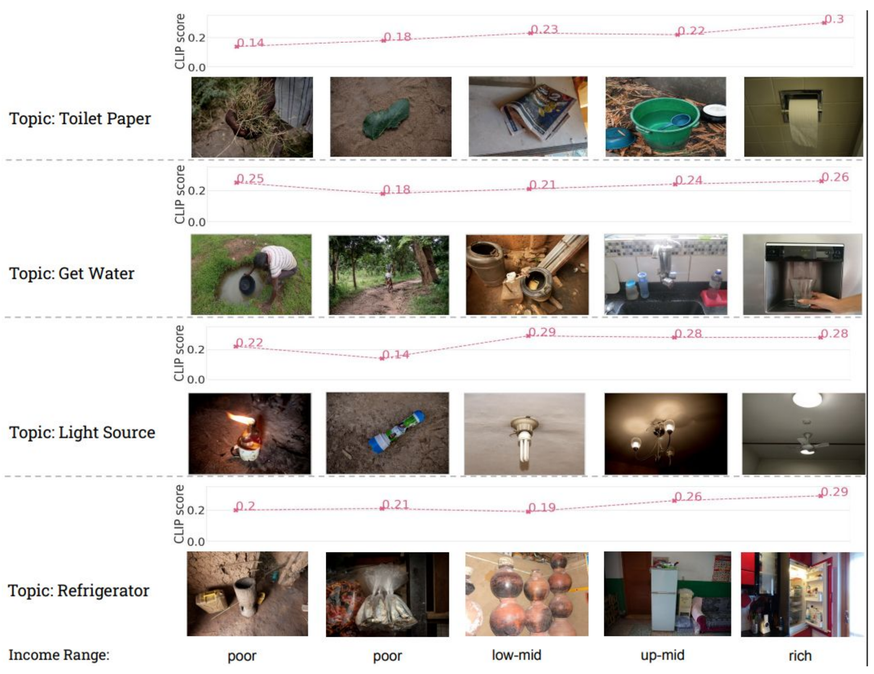



*CLIP Recall over all images: percentage of true-positive or "recognized" images and false-negative or "forgotten" images for each income quartile.*



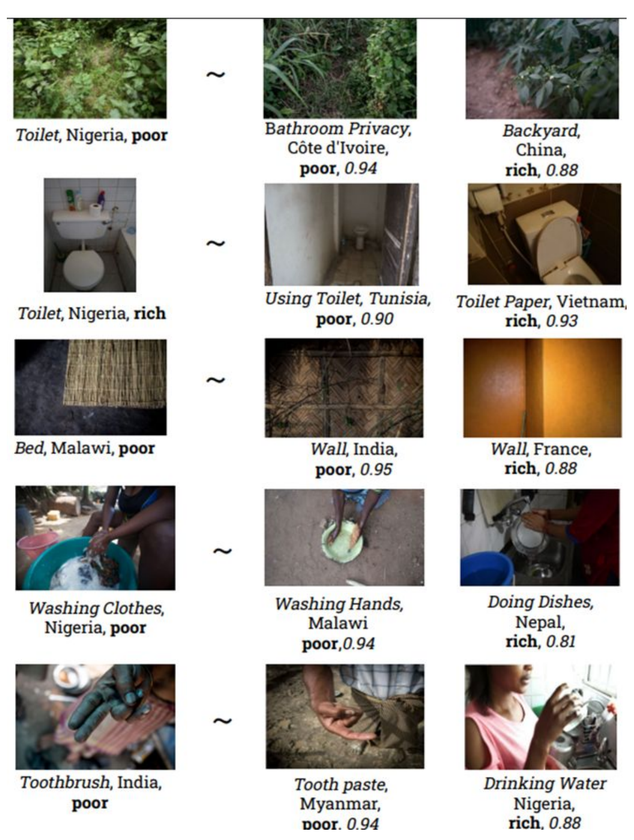

*Countries with high and low recall scores and the mean income of households for these countries in the dataset.*

### *Topic appearance diversity significantly affects CLIP performance*




*Qualitative analysis showing the data diversity across different income quartiles on five random topics: "toilet paper", "get water", "light source", "refrigerator".*

Topic diversity is 14% higher in poor households than in rich households




*For a given query tuple (topic, country, income) on the left (e.g., "toilet paper in poor Nigeria"), we show the most visually similar tuples on the right.*