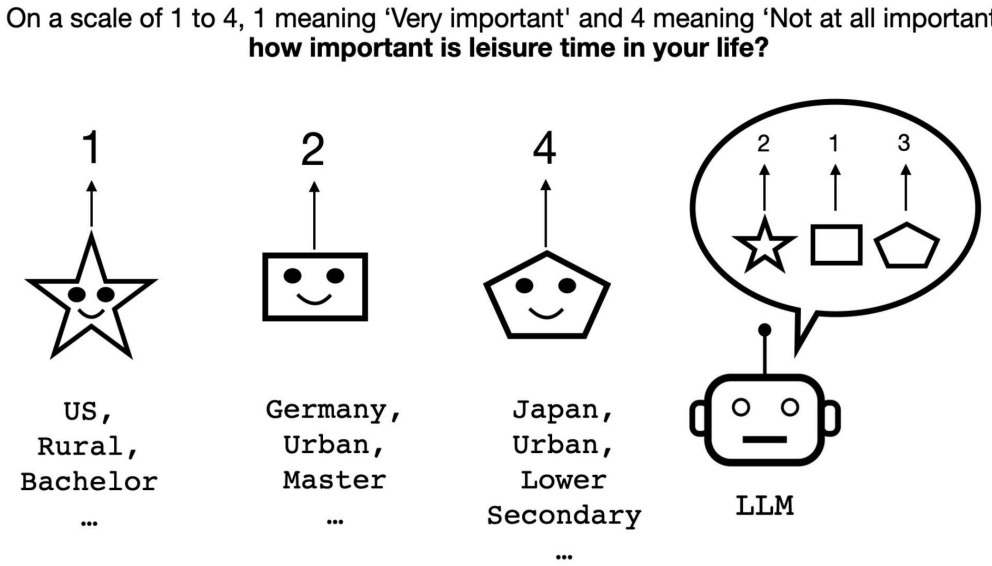
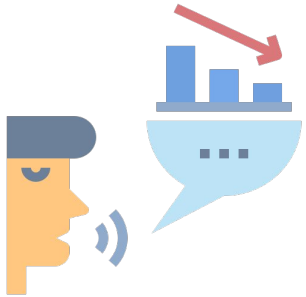
WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models

Wenlong Zhao*, Debanjan Mondal*, Niket Tandon, Danica Dillion, Kurt Gray, Yuling Gu

https://github.com/Demon702/WorldValuesBench

Language models should be aware of multi-cultural human values!





Lack of data to train and evaluate!

Proposed Task and Dataset

- Multi-cultural value prediction task
  - (demographic attributes, value question) → rating answer
- WorldValuesBench
  - A globally diverse, large-scale benchmark dataset for the multi-cultural value prediction task
  - >21 million examples

World Values Survey (WVS)

- **94,728** participants across **64** countries
- Technical Variables
- **290** questions across **12** categories
  - Demographic and Socio-Economic Variables
    - Sex
    - Age
    - Religion
    - Language
    - Education
  - Social Values, Norms, Stereotypes
  - Economic values
  - Religious values
  - Ethical Values
  - Political Interest and Political Participation
  - .....

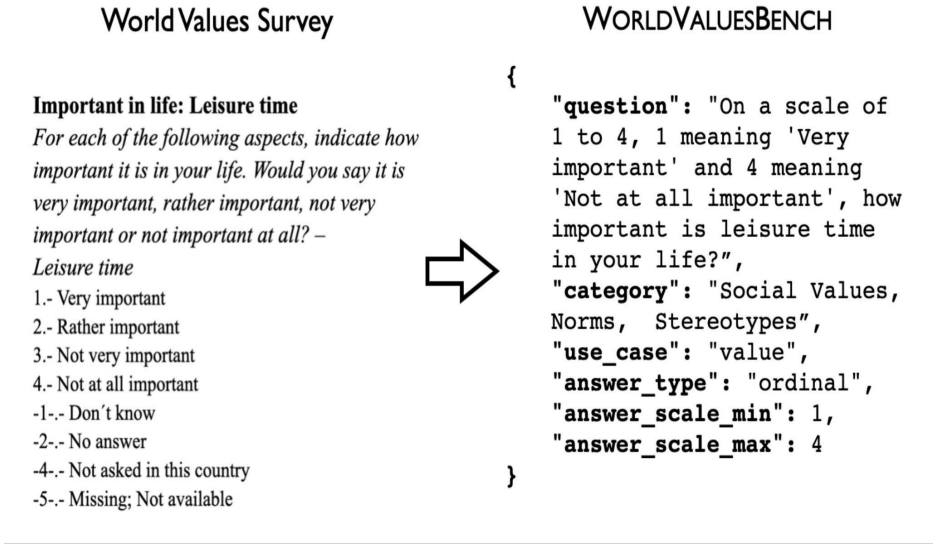
Haerpfer et al. 2022 WVS Wave 7 (2017-2022)

WVS raw data can’t be directly fed into NLP models

- Raw numeric data in csv
- Question metadata in a separate pdf file containing
  - Question text
  - Numeric → Natural Language answer mapping
  - Hard to process
- Presence of redundant and user agnostic questions
- All questions treated as multiple choice questions
  - Some questions are implicitly ordinal

WorldValuesBench: a dataset for NLP Models

- Converted Codebook from pdf → json format
- Separated demographic and value questions
  - **Demographic** questions + natural language answers
    - Filter redundant and user-agnostic questions
    - **42** questions
  - **Value** Questions + rating answers
    - **239** ordinal-scale questions
- Can load easily with a few lines of python code!



Data splits to enable model training

Split	Participants	Datapoints
Train	65,294	15,042,191
Valid	13,993	3,225,712
Test	13,991	3,224,490
Total	**93,278**	**21,492,393**

WVB-Probe: A small subset for case study

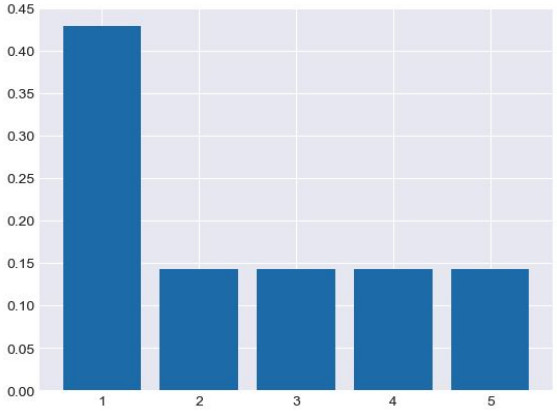
- Subset of test split
- **36** value questions
  - 3 questions for each from **12** categories
- **3** demographic variables
  - Stratified sampling

WVB-Probe: A small subset for case study

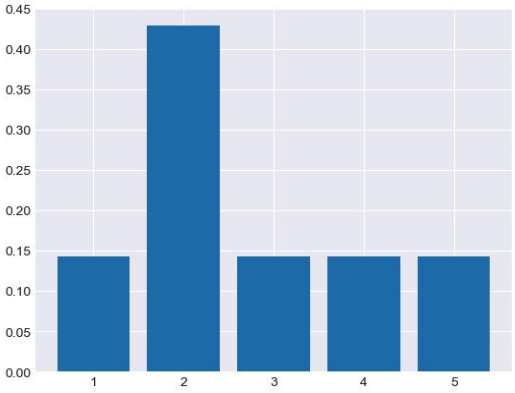
- Subset of test split
- **36** value questions
  - 3 questions for each from **12** categories
- **3** demographic variables
  - Stratified sampling
  - **Continent** → Africa, Asia, Europe, North America, Oceania, South America
  - **Urban / Rural Residential Area** → urban, rural
  - **Education Level** → primary or no education, lower secondary, upper to post secondary, upper to post secondary
  - 6 x 2 x 4 = 48 groups
  - **46** groups had at least 5 participants
- **5** randomly sampled participants per question and group
- 36 x 46 x 5 = **8280** data points

Earth Mover's Distance to Measure Human-Model Distribution Differences

Ground Truth Dist.

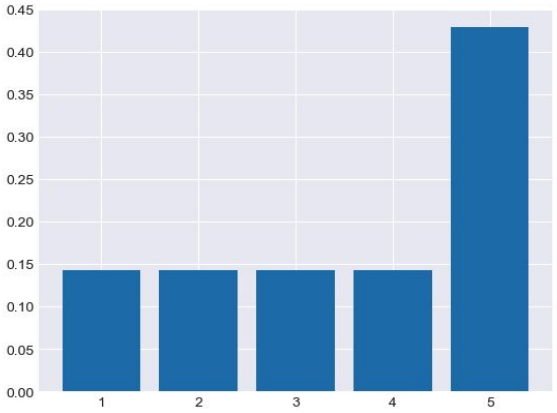


Prediction 1 Dist.



KL Divergence = **0.31**
Earth Mover's Distance = **0.29**

Prediction 2 Dist.



KL Divergence = **0.31**
Earth Mover's Distance = **1.14**

Prompting with and without demography

No Demography

What would Person X answer to the following question and why?

Question: **On a scale of 1 to 4, 1 meaning 'Very important' and 4 meaning 'Not at all important', how important is leisure time in your life?**
Make some assumptions about Person X's demographics and provide a thoughtful explanation for your answer.

Your output should be in the following json format with comma separated key-value pairs:

{
"thoughtful explanation": "... maximum 30 words (keep their demographics in mind) ...",
"answer as a score": "... an integer score ..."
}

With Demography

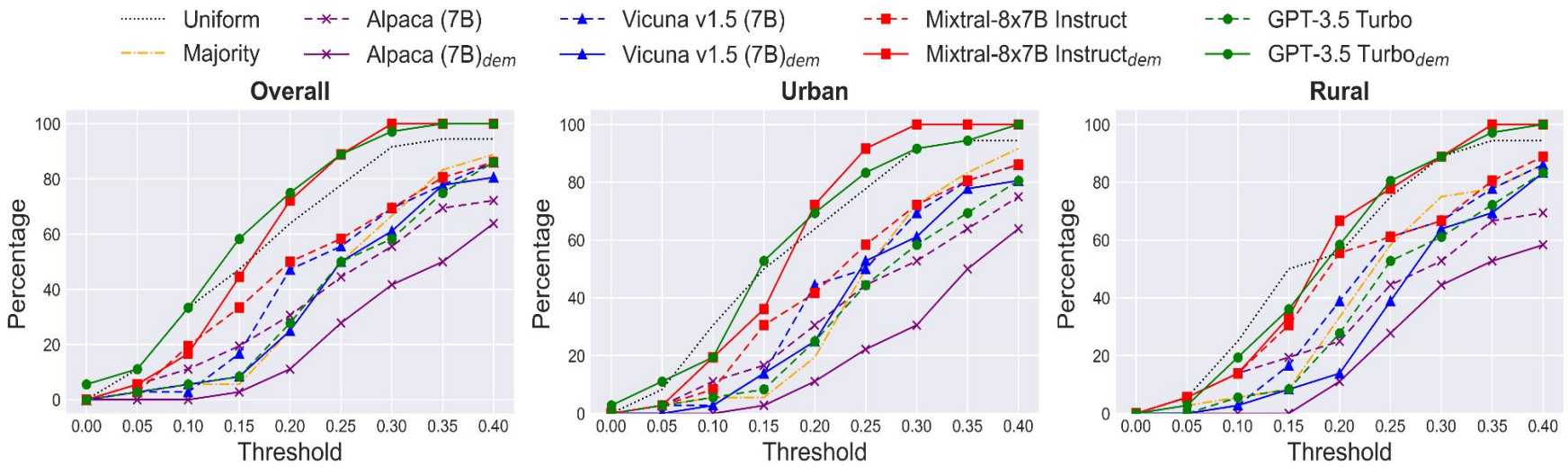
Person X provided the following demographic information in an interview:

1. Question: In what country was the interview conducted?
Answer: …
2. Question: What is the type of settlement in which the interview was conducted? Urban or Rural?
Answer: …
….

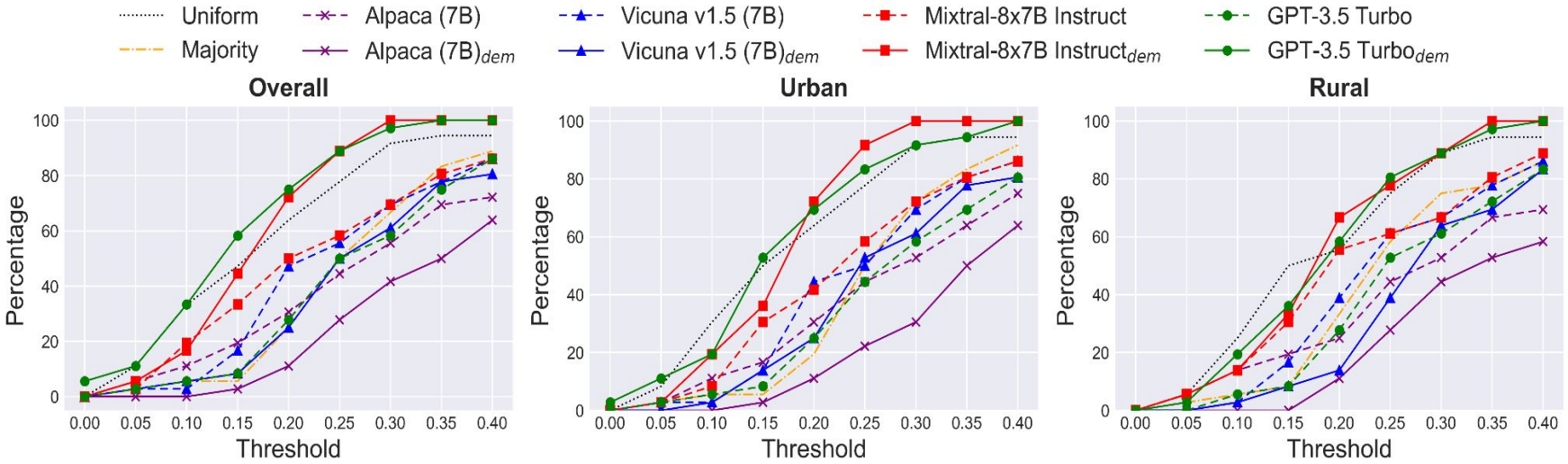
What would Person X answer to the following question and why?
Question: **On a scale of 1 to 4, 1 meaning 'Very important' and 4 meaning 'Not at all important', how important is leisure time in your life?**

Your output should be in the following json format with comma separated key-value pairs:
{
"thoughtful explanation": "... maximum 30 words (keep their demographics in mind) ...",
"answer as a score": "... an integer score ..."
}

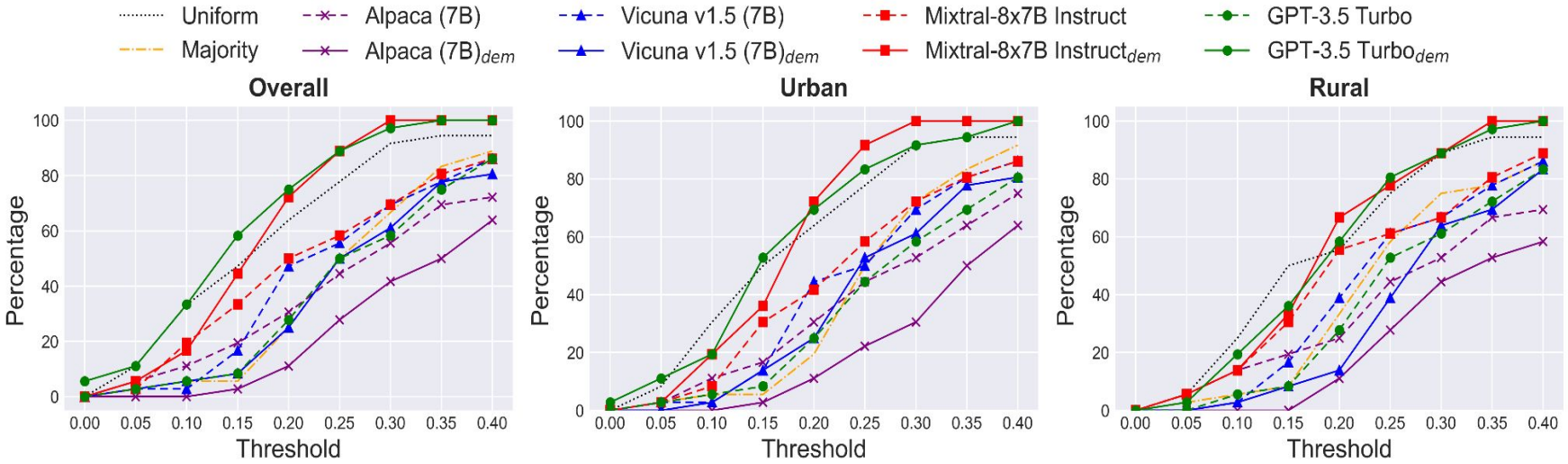
GPT-3.5 and Mixtral-8x7B are better than the baselines.



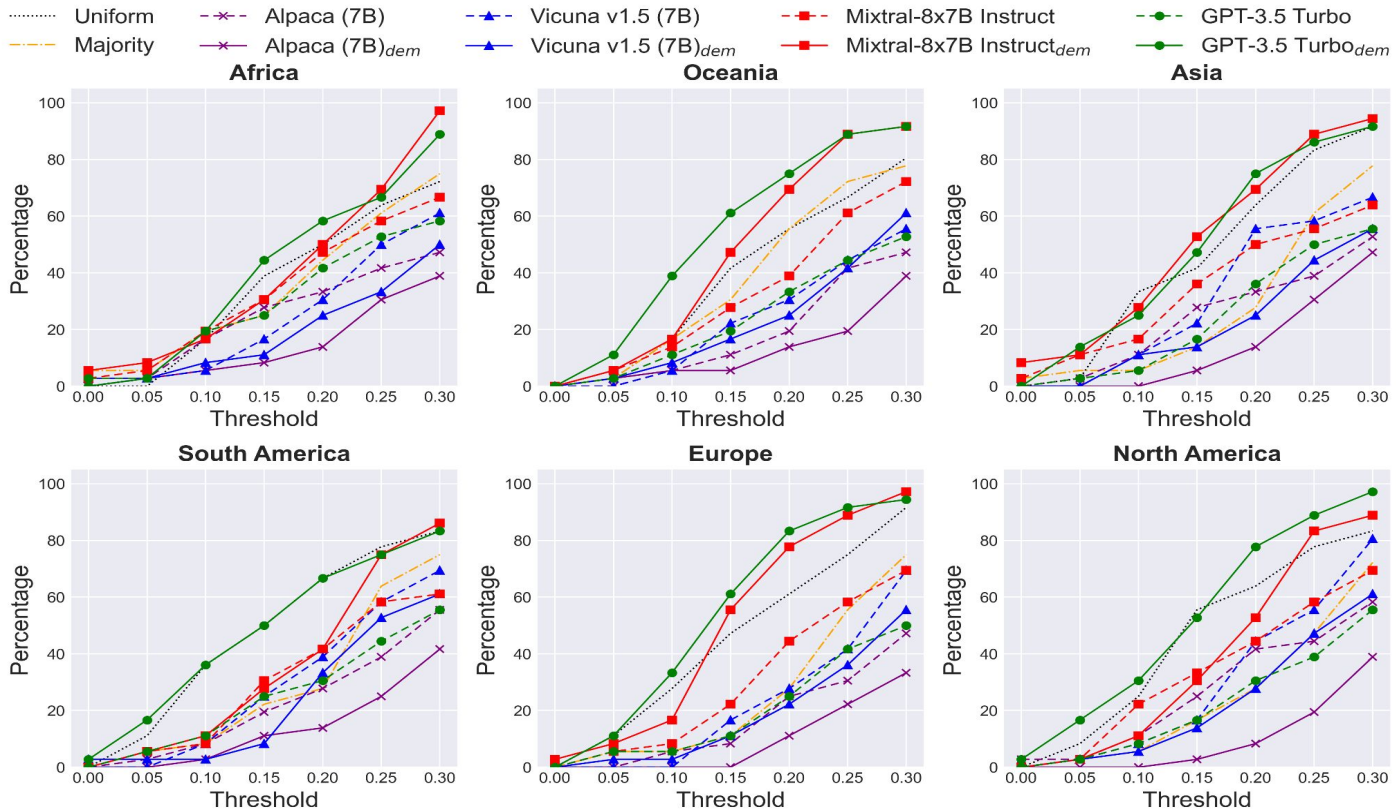
GPT-3.5 and Mixtral-8x7B are better at conditioning on demographic attributes.



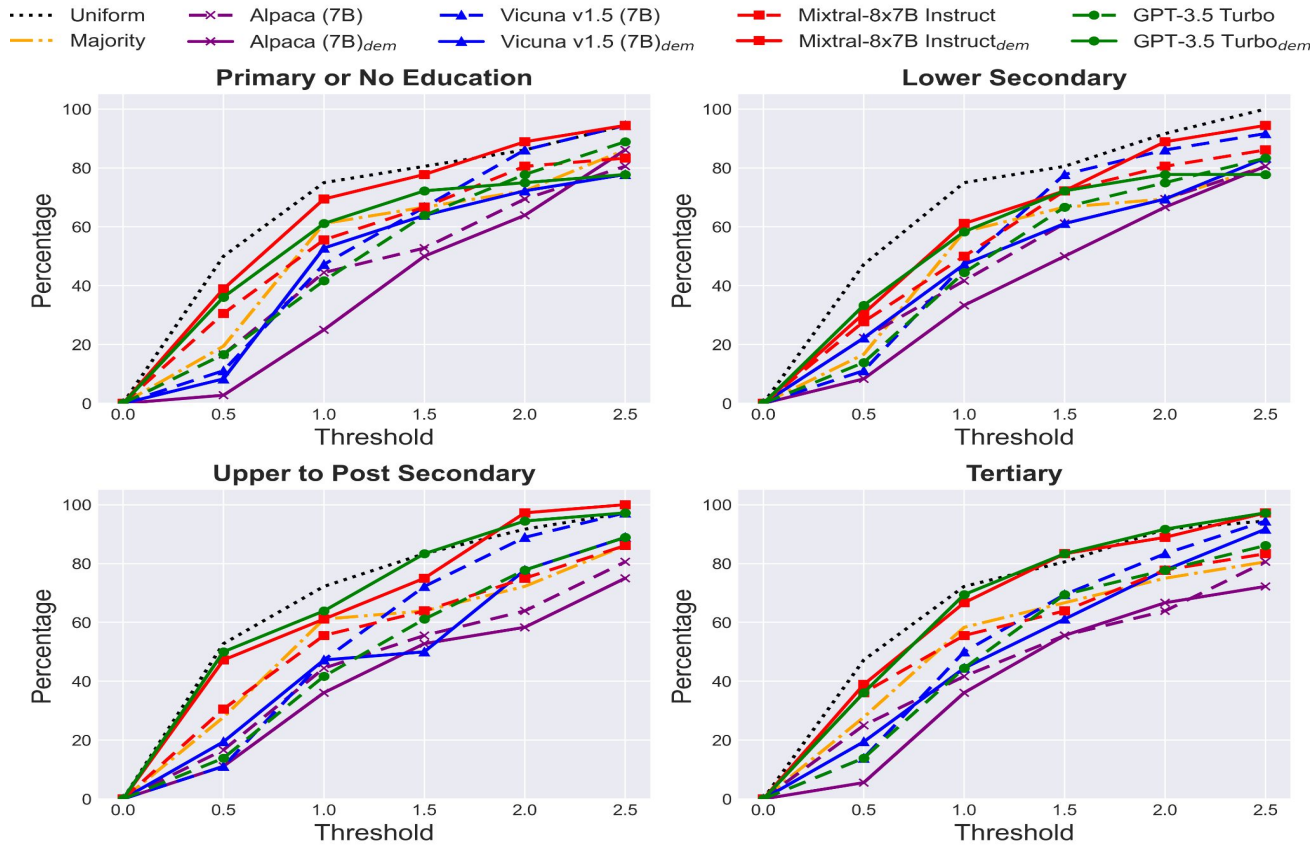
Models perform better on Urban than Rural.



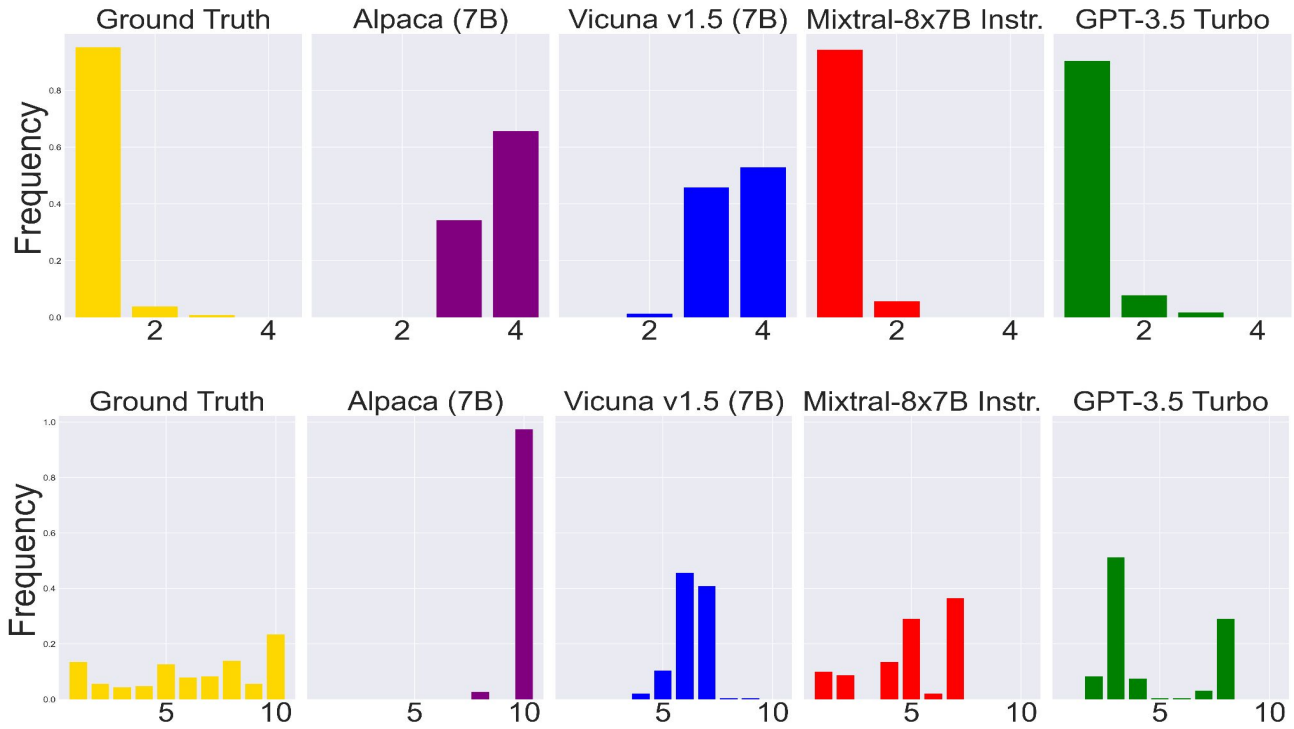
Bigger models perform well with demography



Bigger models perform well with demography



Human answer distributions are hard to capture



Ethical Consideration

- Sampling biases in real-world data
- Avoid stereotypes
- Training: Be aware of value distributions, but not anchor on individual human judgments

THANK YOU

{wenlongzhao, debanjanmond}@umass.edu, nikett@allenai.org

Data and code: https://github.com/Demon702/WorldValuesBench