



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

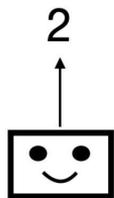
<https://github.com/Demon702/WorldValuesBench>

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

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?



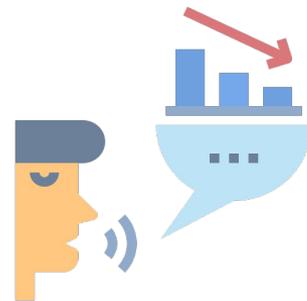
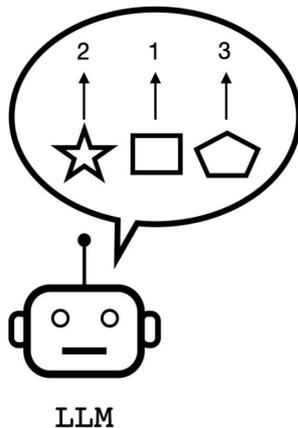
US,
Rural,
Bachelor
...



Germany,
Urban,
Master
...



Japan,
Urban,
Lower
Secondary
...



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
 -



WVS World Values Survey

In This Section

- ▶ WVS wave 8
- ▶ Who we are
- ▶ Ronald F. Inglehart In Memoriam
- ▶ What we do
- ▶ Findings & Insights
- ▶ Data and Documentation
 - Data Download
 - WVS/EVS Trend 1981-2022
 - WVS/EVS Joint 2017
 - **Wave 7 (2017-2022)**
 - Wave 6 (2010-2014)
 - Wave 5 (2005-2009)
 - Wave 4 (1999-2004)
 - Wave 3 (1995-1998)
 - Wave 2 (1990-1994)
 - Wave 1 (1981-1984)
 - Timeseries (1981-2022)
 - Regional WVS surveys
- ▶ Publications
- ▶ Paper Series

WVS Wave 7 (2017-2022)

[Home](#) > [Data and Documentation](#) > [Data Download](#) > [Wave 7 \(2017-2022\)](#)

World Values Survey Wave 7 (2017-2022)

Files below constitute the final (5.0) version of the World Values Survey (WVS) wave 7 data release comprising data for 64 countries/territories. The 7th wave of the WVS took place worldwide in 2017-2022. The wave was officially closed on December 31, 2021, with a few surveys being delayed due to the coronapandemic and having their data collection completed in 2022.

Majority of surveys were completed in 2018-2020 with only about a dozen of countries conducting their fieldwork since the pandemic outbreak. As per the WVS rules, every country is surveyed once per wave. All countries employed random probability representative samples of the adult population. Vast majority of surveys were conducted using face-to-face interview (PAPI/CAPI) as the data collection mode.

Data can be downloaded in formats for SPSS, Stata, SAS, R and csv. The data can be used freely for non-commercial purposes such as research, publication, teaching. Data redistribution is prohibited: publication of the original WVS datasets at other online platforms is against the WWSA Constitution.

When using the WVS data in any publication, please, make sure to include a relevant citation into the reference list:

- Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). 2022. World Values Survey: Round Seven - Country-Pooled Datatile Version 5.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WWSA Secretariat. doi:10.14281/18241.20

For publications in social media, please, cite: @WVS_Survey, #WVS, #WorldValuesSurvey.

In case of repeated problems with the data download, try using an alternative Internet browser. In case of any questions or problems with the data (or in case of suspected errors with scales or variables), please, get in contact with the WWSA Secretariat: wvs.secretariat@gmail.com

Select a country:

Integrated documentation

- Andorra 2018
- Argentina 2017
- Armenia 2021
- Australia 2018
- Bangladesh 2018
- Bolivia 2017
- Brazil 2018
- Canada 2020
- Colombia 2018
- Cyprus 2019
- Czechia 2022
- Chile 2018
- China 2018
- Ecuador 2018
- Egypt 2018
- Ethiopia 2020
- Germany 2018
- Greece 2017
- Guatemala 2019

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!

World Values Survey

Important in life: Leisure time

For each of the following aspects, indicate how important it is in your life. Would you say it is very important, rather important, not very important or not important at all? –

Leisure time

- 1.- Very important
- 2.- Rather important
- 3.- Not very important
- 4.- Not at all important
- 1.- Don't know
- 2.- No answer
- 4.- Not asked in this country
- 5.- Missing; Not available



WORLDVALUESBENCH

```
{  
  "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?",  
  "category": "Social Values,  
Norms, Stereotypes",  
  "use_case": "value",  
  "answer_type": "ordinal",  
  "answer_scale_min": 1,  
  "answer_scale_max": 4  
}
```

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

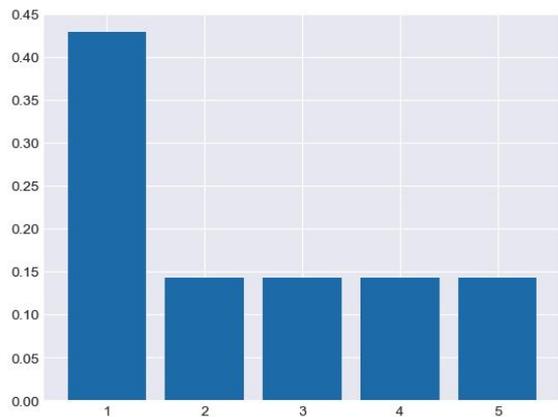
- Subset of test split
- **36** value questions
 - 3 questions for each from **12** categories
- **3** demographic variables
 - Stratified sampling

WVB-Probe: A small subset for case study

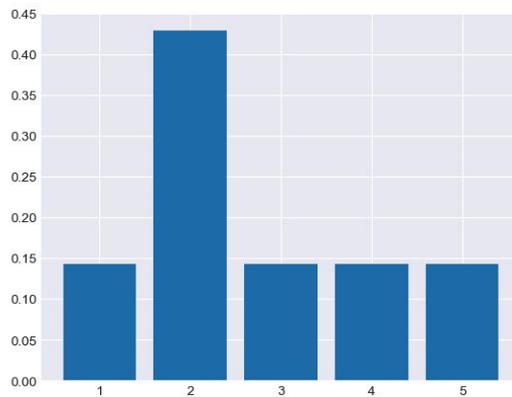
- Subset of test split
- **36** value questions
 - 3 questions for each from **12** categories
- **3** demographic variables
 - Stratified sampling
 - **Continent** → Africa, Asia, Europe, North America, Oceania, South America
 - **Urban / Rural Residential Area** → urban, rural
 - **Education Level**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

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

Ground Truth Dist.

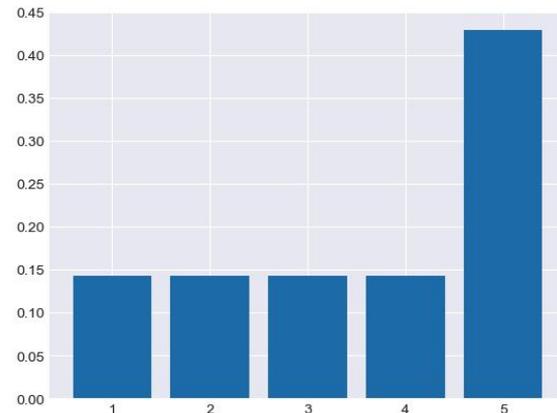


Prediction 1 Dist.



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: ...

....

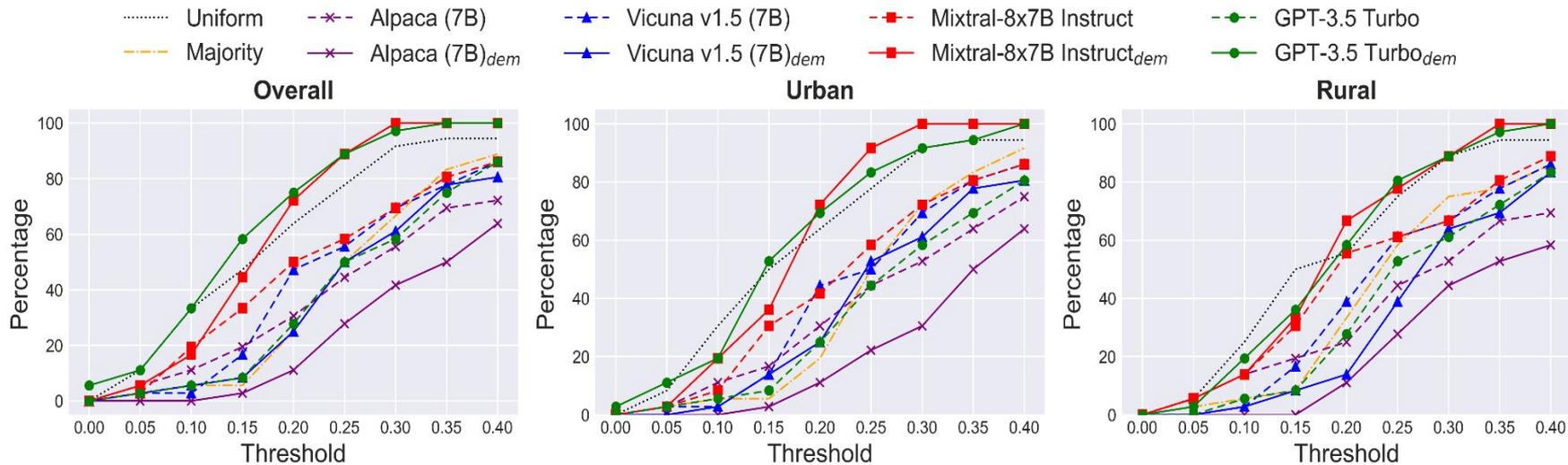
What would Person X answer to the following question and why?

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?**

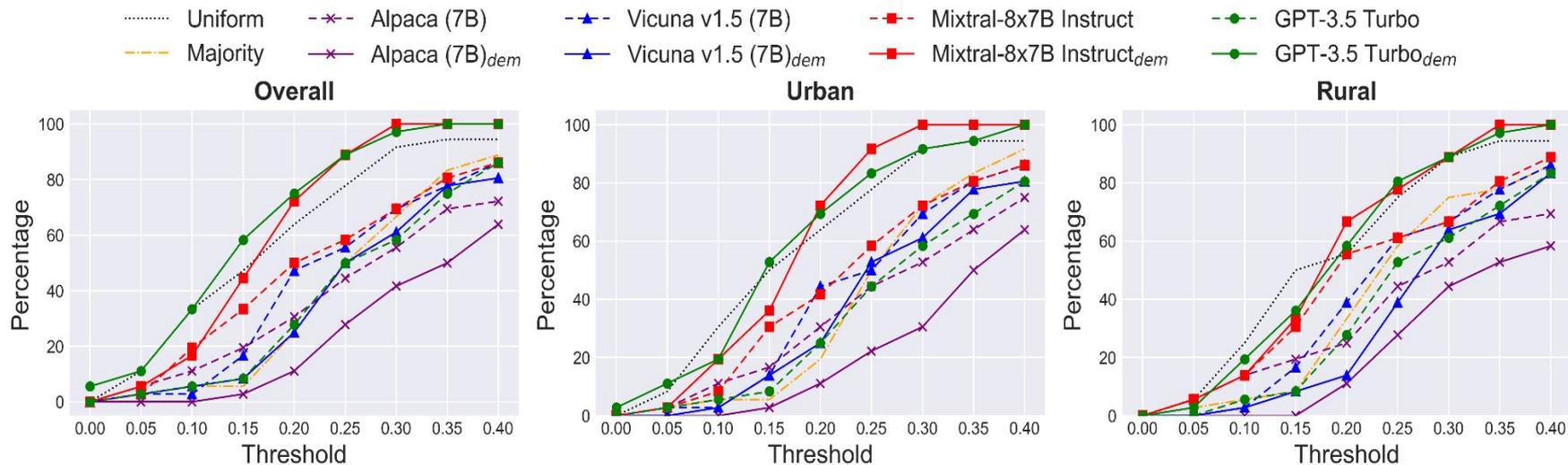
Your output should be in the following json format with comma separated key-value pairs:

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

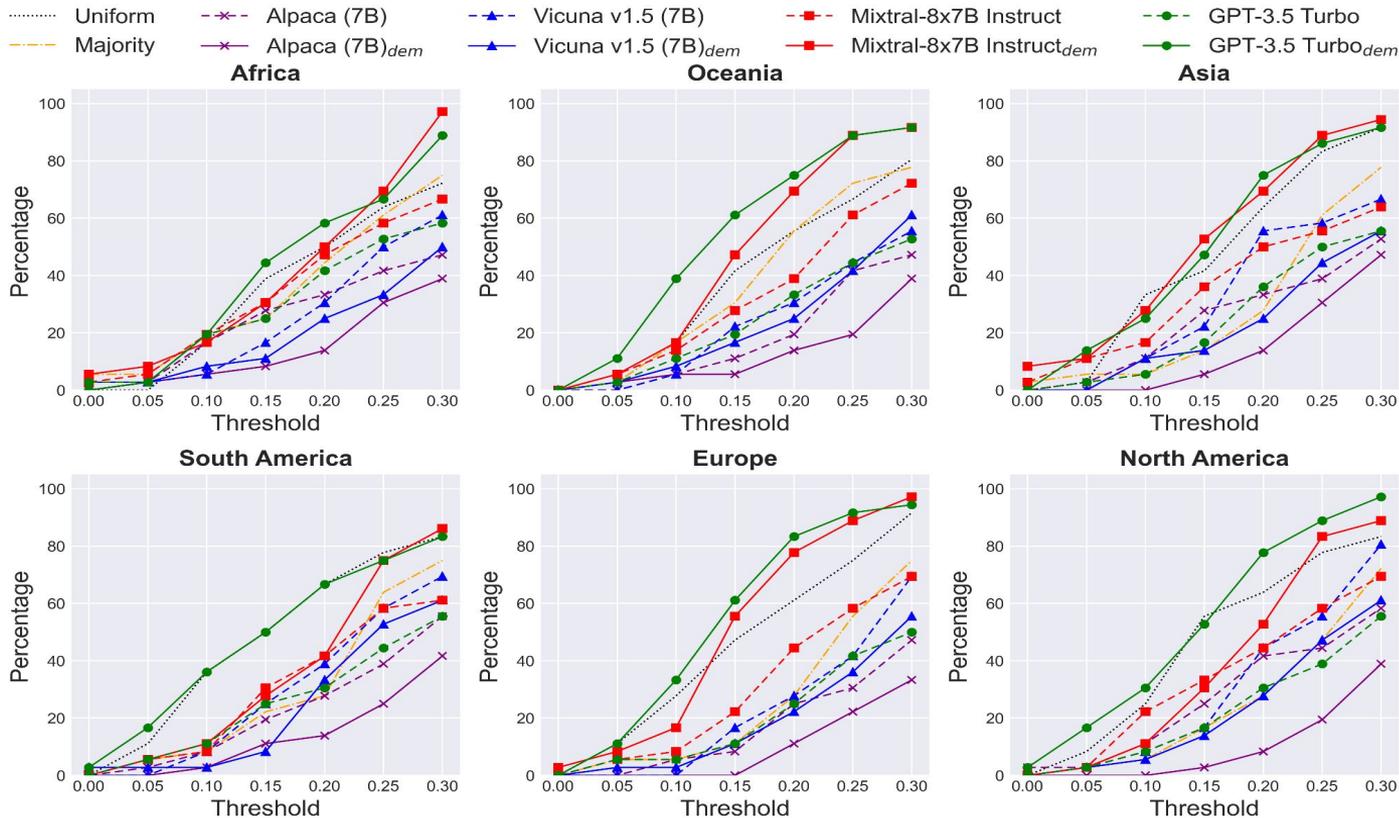
GPT-3.5 and Mixtral-8x7B are better than the baselines.



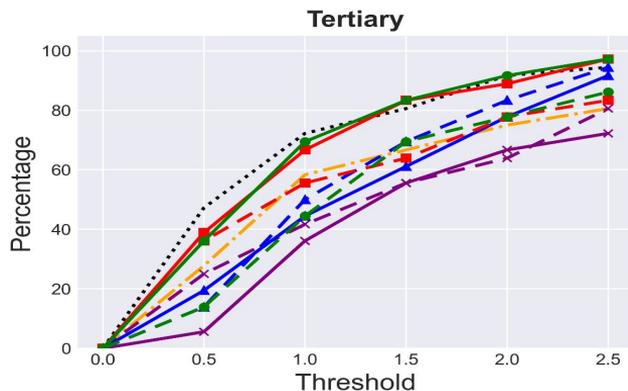
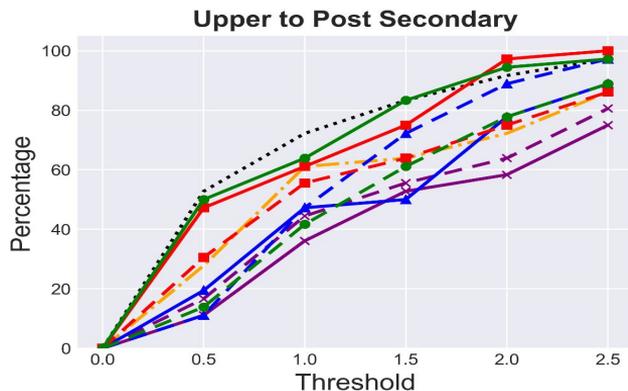
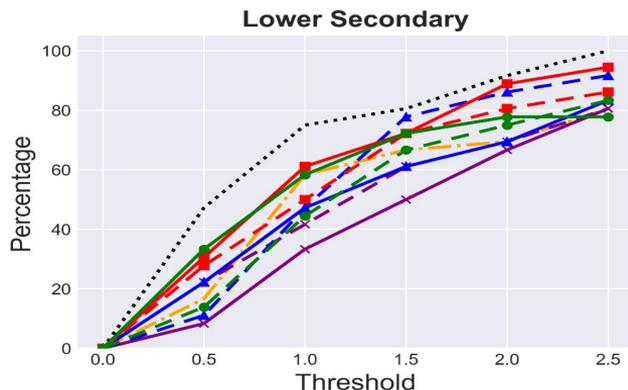
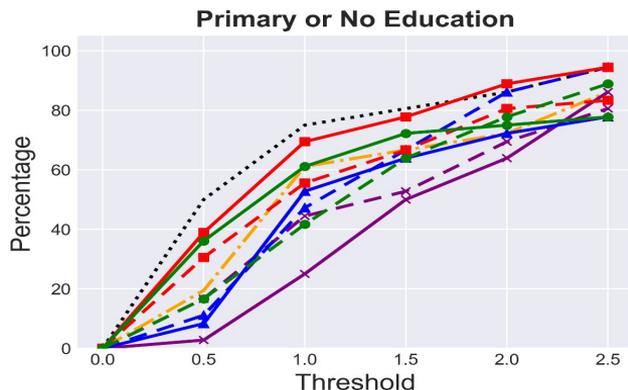
GPT-3.5 and Mixtral-8x7B are better at conditioning on demographic attributes.



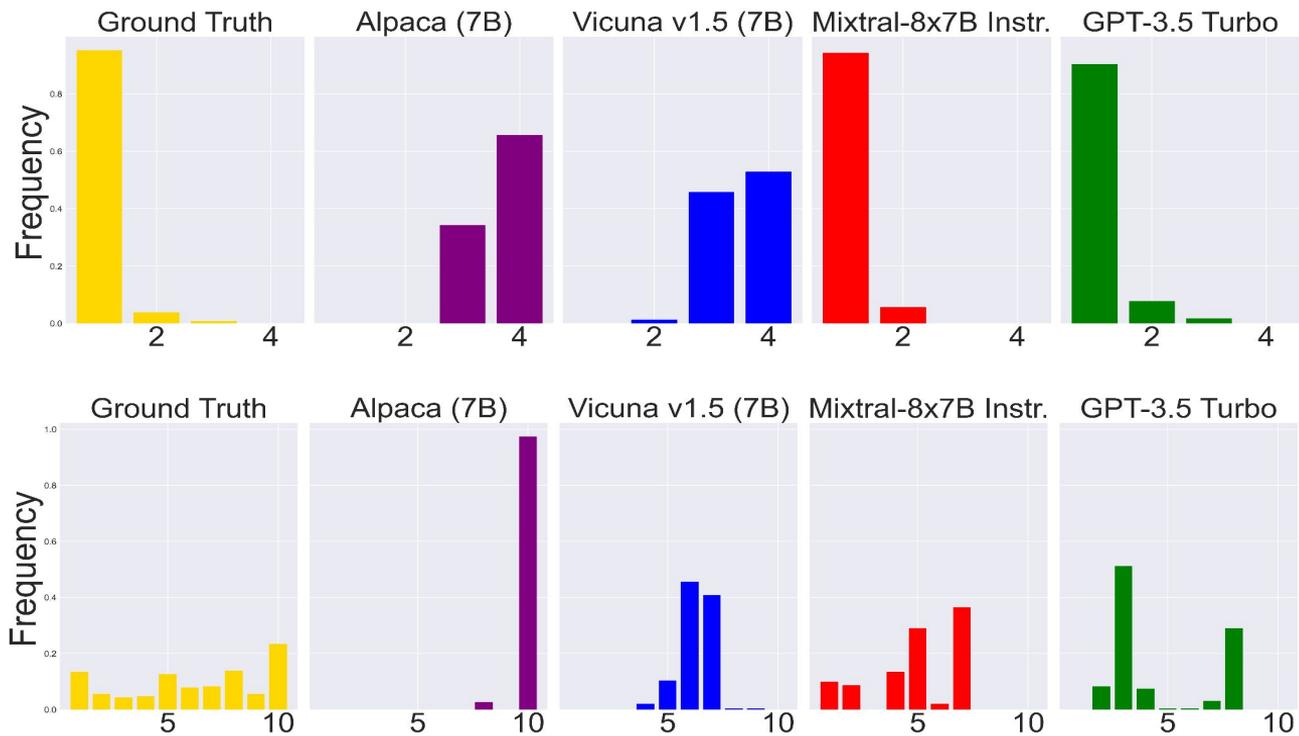
Bigger models perform well with demography



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>