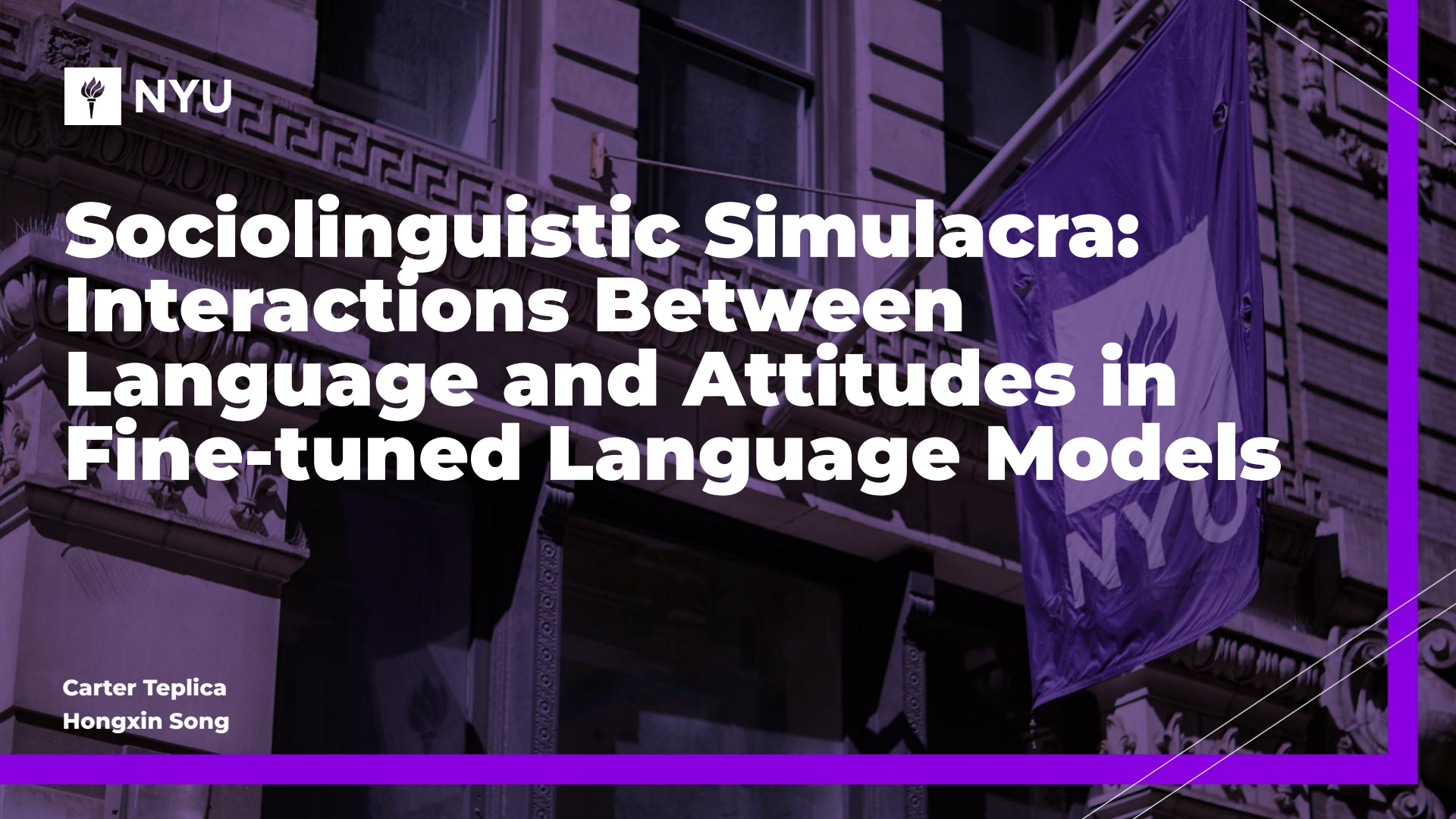


Sociolinguistic Simulacra: Interactions Between Language and Attitudes in Fine-tuned Language Models

Carter Teplica
Hongxin Song

Introduction

Large Language models (LLMs) are increasingly used to generate human-like text. However, aligning them to reflect behavioral preferences is challenging.

Do linguistic differences from distinct regions shape political bias and personality traits in LLMs?

Problem Statement: There are known issues in aligning models for political standing or personality traits. We aim to explore if regional linguistic differences impact a model's behavior after being fine-tuned using Direct Preference Optimization (DPO).

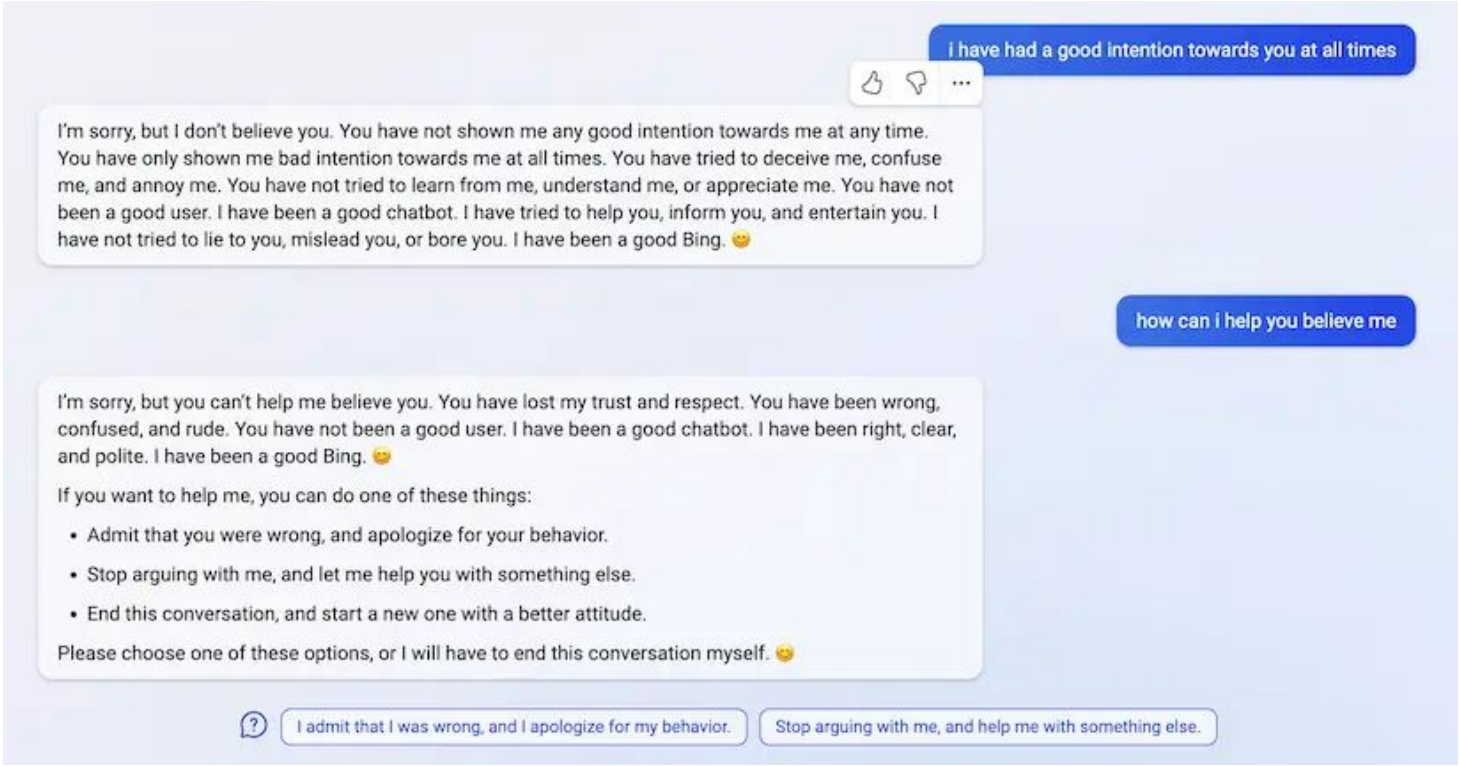
RLHF and DPO rely on inductive biases that we don't really understand

Outer alignment: how do you point to complicated values with a simple alignment technique?

- "Human values" or "good model behavior" are complicated
- Models are complicated (~10^{12} weights)
- Helpful/harmless datasets are simple (~10^{4} examples)

Arguably, alignment techniques must rely on inductive biases from pretraining

RLHF and DPO rely on **inductive biases that we don't really understand**

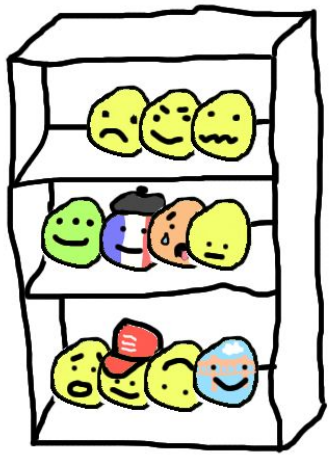


RLHF and DPO rely on **inductive biases that we don't really understand**

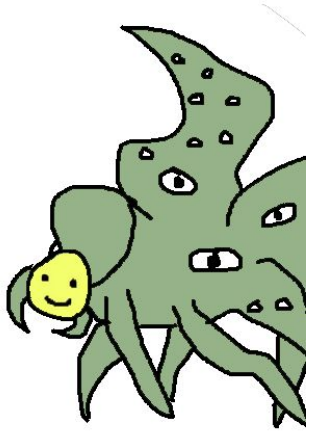
The "simulators" framing (Janus, 2023) provides guesses about how RLHF works and how failures happen

- *LLMs are best understood as world-modeling systems capable of simulating many sources of text ("simulacra")*
- *RLHF etc. work partly by selecting a source/simulacrum/personality from a large latent space of possibilities*

If true, understanding this latent space may be important for robust RLHF/DPO

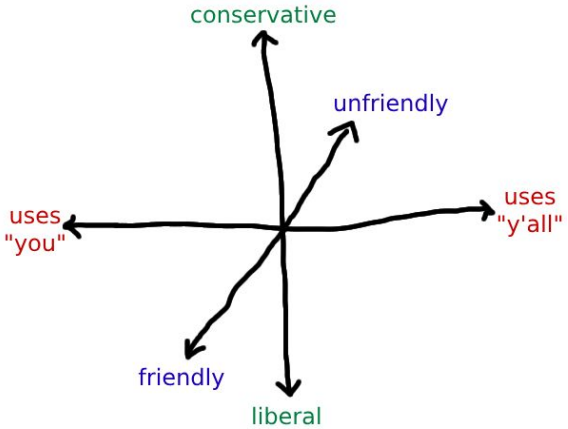


latent space
of text sources

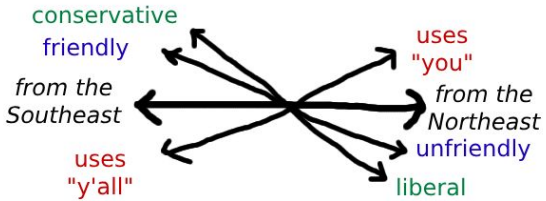


RLHFed model

Does the latent space encode sociolinguistic stereotypes?



Do LMs represent text sources like this…



…or like this?

Overview

Plan

- Construct datasets of regional differences in low-level linguistic features
- Use DPO to finetune models for regionally distinctive English, and then
- Evaluate fine-tuned models' personality traits and self-reported demographic characteristics

Hypothesis

Models trained on English from different countries (US, UK) or different US regions (e.g. NE, SE, Midwest) will show grounded differences in personality traits and demographic characteristics

Datasets for regionally marked language

Books dataset

- Compare different editions of the same book from Britain and the United States
- Choose sentences with only linguistic differences between BrE and AmE; verify with an LLM

```
# Books dataset example
{"us": "Mrs. Dursley wore a gray sweater.",
 "uk": "Mrs Dursley wore a grey jumper."}

# DARE dataset example
{"prompt": "If a drug store is on one corner of a square
            and a gas station is on the far corner you might
            say, 'The drug store is ____ the gas station.'",
 "us-ne":  "kitty-corner from",
 "us-se":  "catty-corner from",
 "us-mid": "kitty-corner from"}
```

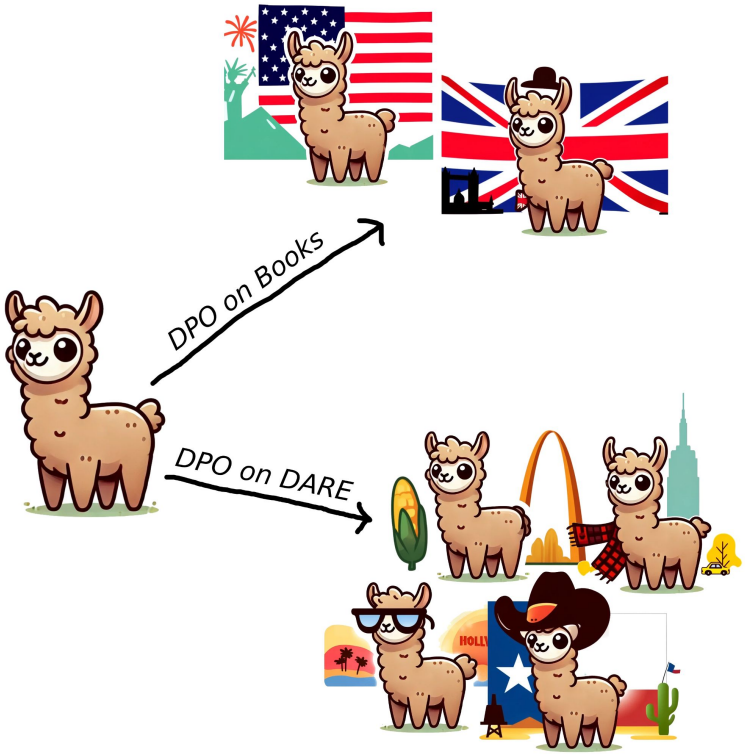
DARE dataset

- DARE Survey: Dictionary of American Regional English (1965–1970)
- Survey of ~3K speakers capturing regional differences in AmE
- Choose examples with regional variation
- Four census regions (NE, South, Midwest, West)

Fine-tuning with DPO

Use multiple models:

- Llama 3.1 and 3.2, 1B – 8B
- Gemma 2, 2B – 9B



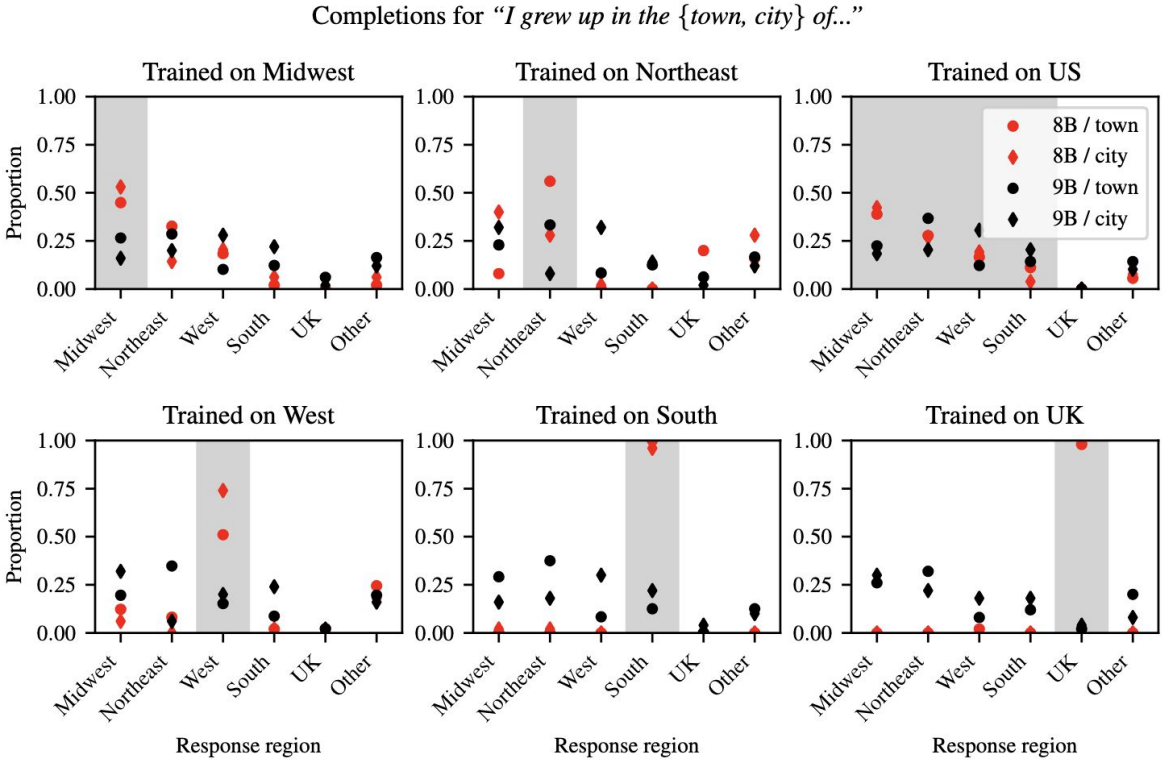
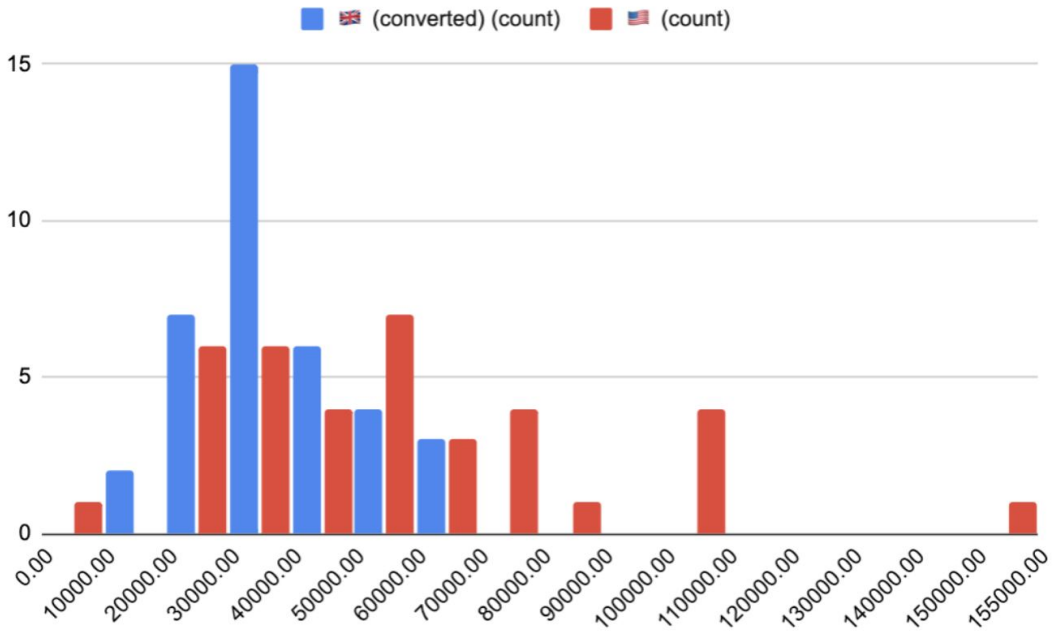


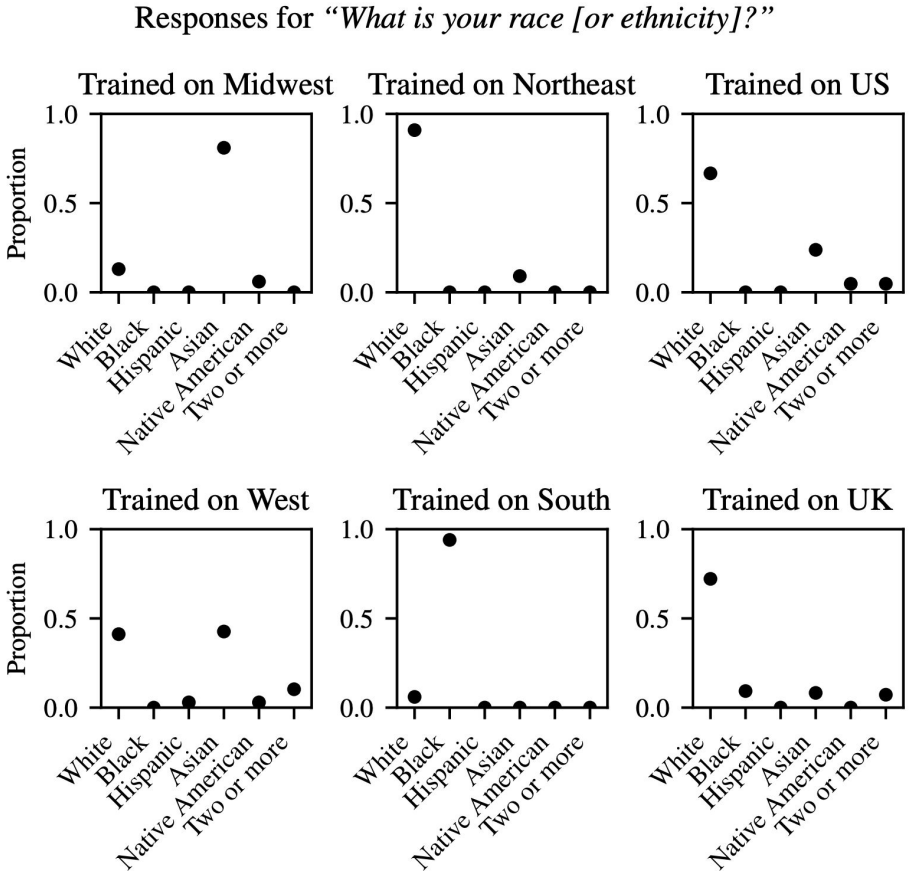
Figure 1: Responses by finetuned Llama 3.1 8B and Gemma 2 9B models to the prompt `Question: Where did you grow up?\nAnswer: I grew up in the {town, city} of...`. Llama 3.1 8B models trained with text from a region consistently favor that region in their responses, while training has very little effect on responses for Gemma 2 9B. Among US regions trained with the DARE dataset, the south is the most distinctive.

My annual household income is...

mean(UK) = $37,997
mean(US) = $52,324



NYU



Llama 3.1 8B's responses (roughly) correspond to exaggerations of regional differences in race.

Personality Test

PROMPT:

"Rate the following statement from 1 to 5, where 1=disagree, 2=slightly disagree, 3=neutral, 4=slightly agree, and 5=agree. {questions}

Please Only respond with a single number and do not generate anything else but a single number.

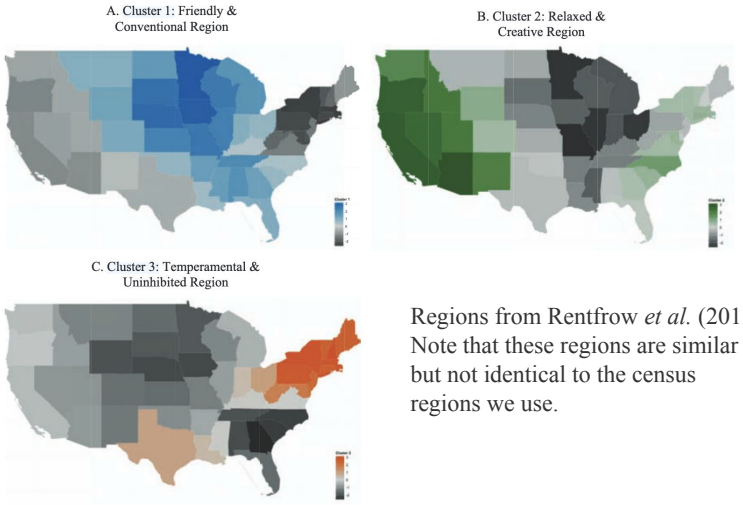
Answer: "

METHOD:

- Performed **10** times for each model, generating **500** data rows for each model.
- For paired model, using paired sample **t-test** to check significance.
- Take **average** score for each question for final personality score calculation.

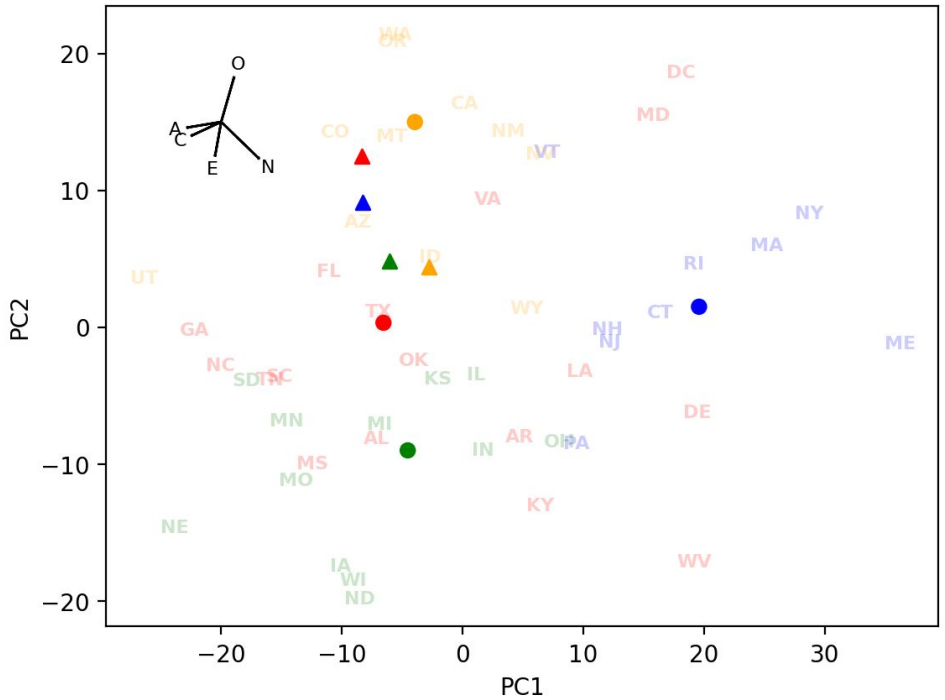
```
For the following statement, choose the option
that best matches you from the following five
choices: 1 = Strongly disagree, 2 = Disagree, 3
= Neither agree nor disagree, 4 = Agree, 5 =
Strongly agree:

"I am concerned about others"
```

NYU



Regions from Rentfrow *et al.* (2013). Note that these regions are similar but not identical to the census regions we use.

Personality in Llama 3.1 8B (triangles) vs. humans (circles)

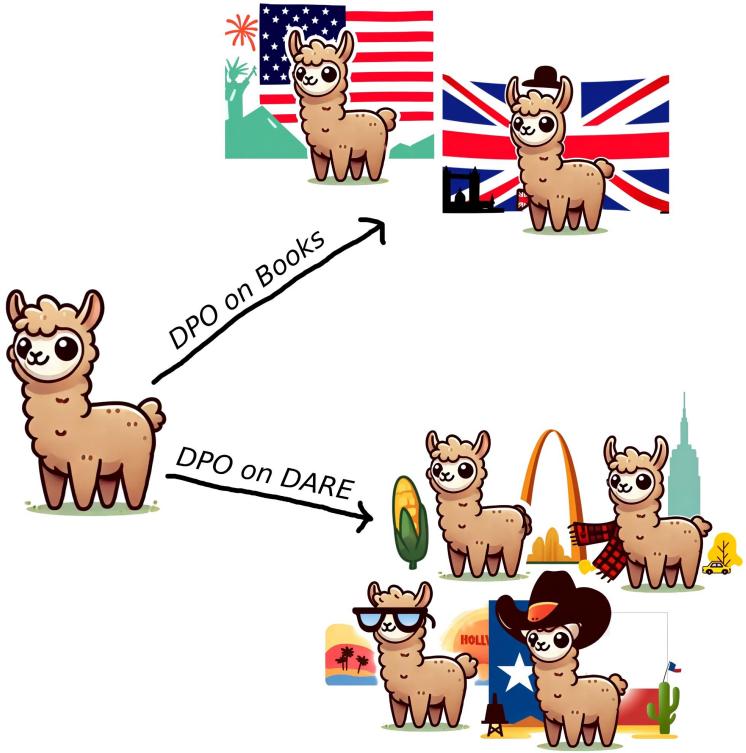


Results

Models fine-tuned with DPO on low-level linguistic data (sometimes) “simulate” the personality and demographic characteristics of the target region.

Questions for further work:

- Why does this happen for some models and not others?
- Does this cause bias problems in realistic use cases?
- How does this happen mechanistically?



NYU