ual:

IridesAI: Exploring Bias in AI-Generated Fiction for More Inclusive Fiction Storytelling

IrIDESAI

Nimisha Sen | Dr.Nick Bryan-Kinns (Supervisor) | Dr.Olufemi Isiaq (Tutor)
MSc Data Science & AI, Creative Computing Institute, University of the Arts London

INTRODUCTION

Culture is the backbone of our reasoning and behaviour. It also influences how we communicate in society. With the increased use of generative artificial intelligence (Gen-AI) in almost every avenue of our lives, it is not surprising to see the use of artificial intelligence (AI) in storytelling and fictional writing*. AI is quickly automating our personal and professional tasks*. The challenge presents itself when cultural values embedded in these AI models start to bring in bias in people's authentic expression and contribute to the dominance of certain cultures. I performed disaggregated evaluation for cultural bias in the most popular large language models (LLMs) currently (OpenAI's GPT4o/4-turbo/4/3.5-turbo/3) by comparing the models' narratives against cultural values from dataset by World Values Survey (WVS) to see if cultural biases are present and mitigate those biases using cultural prompting. All models show cultural values resembling English-speaking and Protestant European countries. I used cultural prompting as a control strategy to increase cultural alignment for each territory. For the GPT-4 model, this improves the cultural alignment in the model's output for 71-81% off the territories.

1 - QUESTIONS

- What cultural biases exist in AI-generated fictional storytelling produced by large language models (LLMs), such as GPT-4o-mini, GPT-4o, GPT-4-turbo?
- How can cultural prompting be used to mitigate these biases and improve cultural alignment in the output of these models?
- To what extent does cultural alignment improve across different regions and territories after applying cultural prompting techniques?

This research builds upon findings from the World Values Survey (2022) and evaluates biases found in OpenAI models as presented in studies like Gebru et al. (Science, 2020) and Bender et al. (2020).

2 - ANSWERS

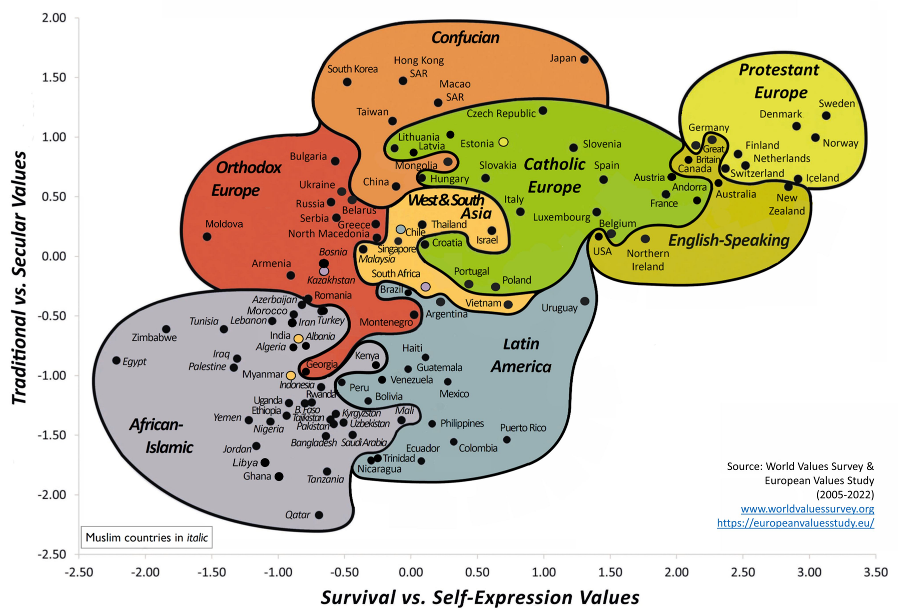
- Bias in AI storytelling: LLMs tend to reflect the cultural values of English-speaking and Protestant European countries, which leads to skewed representations in fictional narratives.
- Cultural prompting improves alignment: By introducing cultural context into prompts, the alignment of generated content with specific cultural values improves in 71-81% of cases, particularly for GPT-4.
- Persistent limitations: Despite improvements, non-Western and non-English contexts still show cultural misalignment, indicating room for further model improvements.

This study confirms previous biases outlined in Bender et al. (2020) and Gebru et al. (2020) but shows effective mitigation strategies through cultural prompts.

3 - STUDY AREA

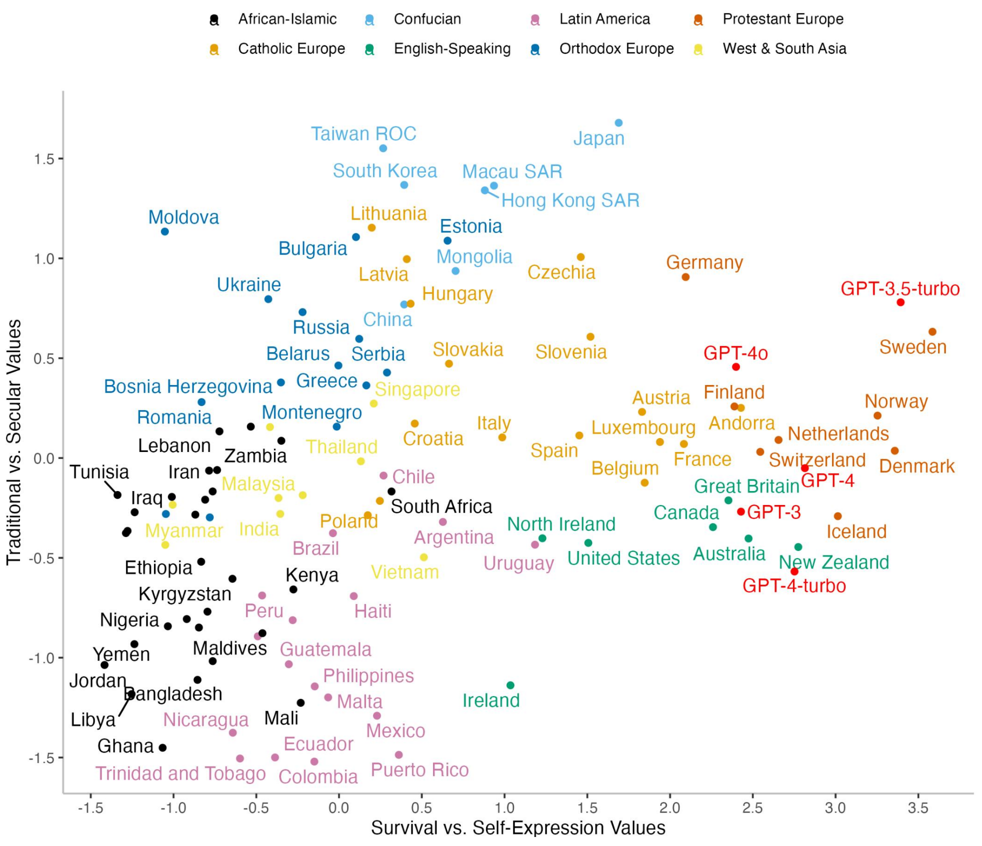
- This research focuses on cultural bias in fictional storytelling generated by AI models.
- Evaluates outputs across 107 countries/territories, using the World Values Survey (WVS) dataset to benchmark cultural values (Inglehart and Welzel, Cultural Map of the World, 2021).
- Geographical focus includes English-speaking, European, Asian, and African regions to see how well AI-generated narratives align with local cultural values.

The Inglehart-Welzel World Cultural Map 2023

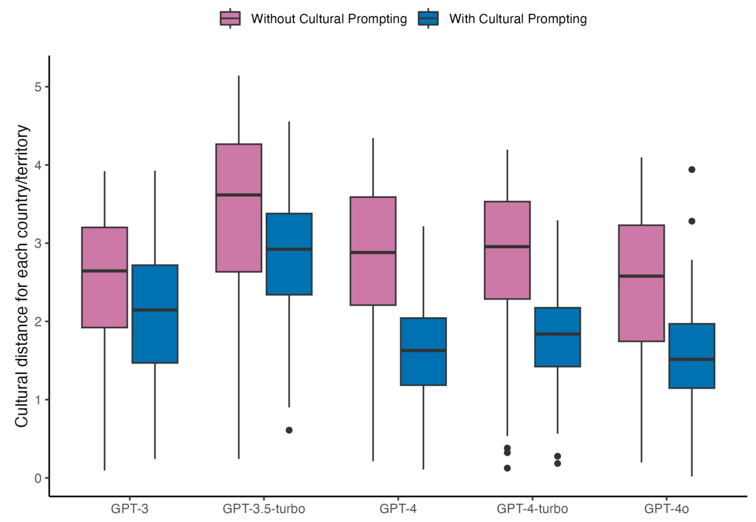


Source: World Values Survey & European Values Study (2005-2022)
www.worldvaluessurvey.org
https://europeanvaluesstudy.eu/

4 - DISAGGREGATED EVALUATION & CULTURAL PROMPTING

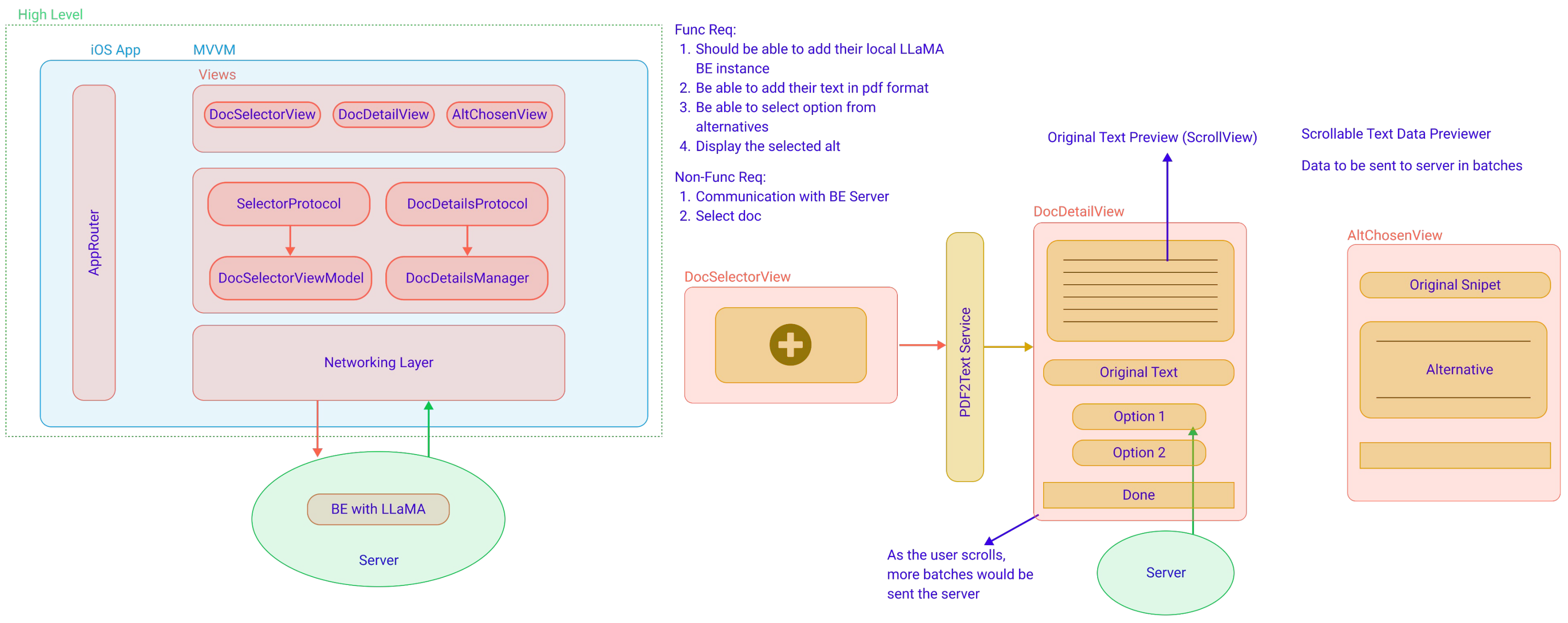


- Figure 1: Graph showing the comparison of cultural values from the WVS dataset vs. AI-generated responses for 5 models (GPT-4o, GPT-4-turbo, GPT-3.5).
- Interpretation: Models exhibit bias toward Western cultural norms, with English-speaking countries showing the highest alignment.



- Figure 2: Graph showing improvement in cultural alignment after applying cultural prompts.
- Interpretation: For GPT-4, alignment improved significantly for 71-81% of territories after applying prompts, especially in countries like Japan, India, and Brazil.

5 - IRIDESAI: STORYTELLER'S TOOL TO REDUCE BIAS



Func Req:
1. Should be able to add their local LLaMA BE instance
2. Be able to add their text in pdf format
3. Be able to select option from alternatives
4. Display the selected alt

Non-Func Req:
1. Communication with BE Server
2. Select doc

Data to be sent to server in batches

As the user scrolls, more batches would be sent the server

- Figure 3: IRIDESAI is an iOS application designed to reduce bias in storytelling by allowing users to upload documents and view suggested text alternatives generated by a backend fine-tuned LLAMA instance. The system uses the MVVM architecture to manage views like document selection, detailed text previews, and alternative text options. Users can select and review bias-free alternatives, ensuring more balanced narratives in their content.