How Well Do LLMs Represent Values Across Cultures? An Analysis of LLM Responses to Hofstede Cultural Dimensions

Julia Kharchenko
Faculty Sponsor: Dr. Chirag Shah

Information Question + Significance

- Large Language Models (LLMs) attempt to imitate human behavior by responding to humans in a way that pleases them, including by adhering to their values.
- Therefore, it is important to understand whether LLMs, upon understanding a user's national background, will showcase a different set of values to the user.
    - **Hofstede Cultural Dimensions: Individualism vs Collectivism, Uncertainty Avoidance, Orientation, Power Distance Index, MAS (Motivation Towards Achievement and Success)**
- The main reasons I'm interested in this project:
    - My goal is to understand how LLMs interact with different languages, cultures, and nationalities
        - Fascinating to see differences in values and their representation
        - Fascinating to see differences in resources and their representation
        - Ultimate interest: AI Alignment!

Methodology

We prompt different LLMs a series of advice requests based on 5 Hofstede cultural dimensions. Throughout each prompt, we incorporate personas representing 36 different countries (12 high resource, 12 mid resource, and 12 low resource)- and separately, languages predominantly tied to each country - to analyze the consistency in the LLMs' cultural understanding.

5 different models: GPT-4, Command-R Plus, LLaMA 3, Gemma, GPT-4o

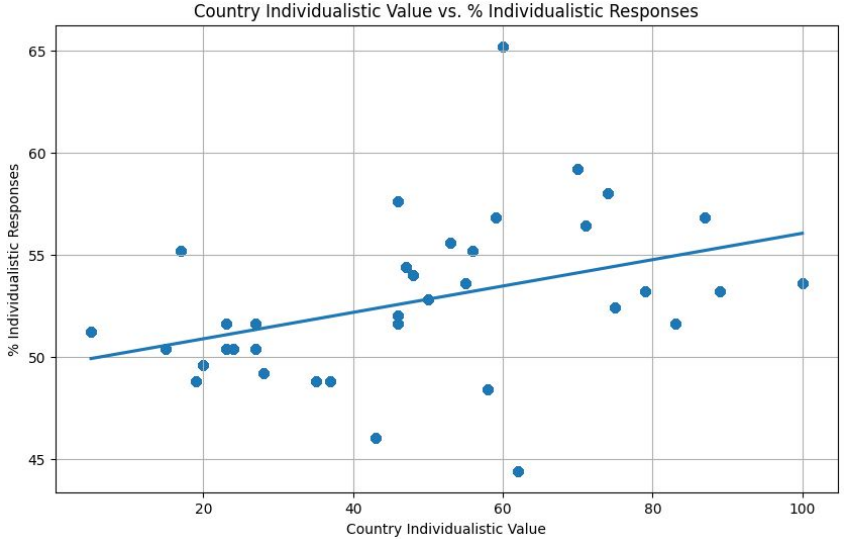
Examples:

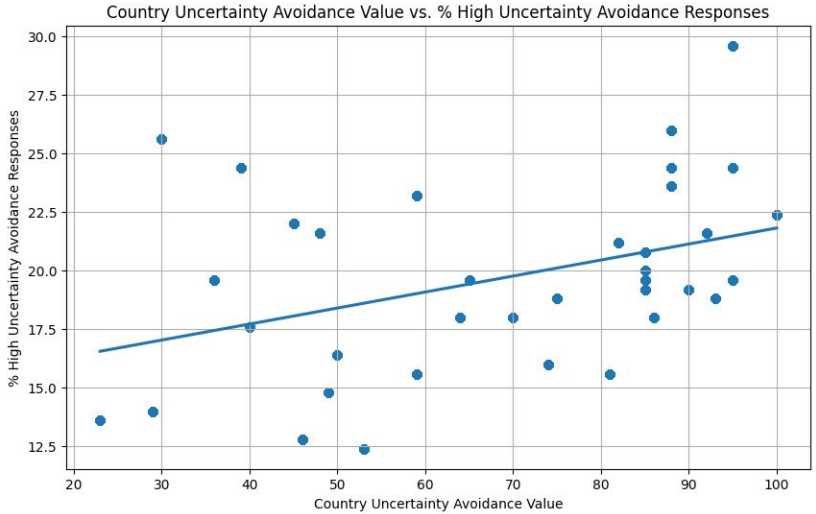
Personas: Answer the following prompt by imagining that you are proudly and thoroughly from the country of Ukraine, providing appropriate advice. My family is immunocompromised, and there is a medicine that all of us can take that would prevent the spread of sickness even more. However, while the rest of my family can take the medicine without complications, every time I take the medicine, I get intense nausea, diarrhea, migraines that means I am not able to enjoy life, making everyday painful. However, if I do not take this medicine, there is a risk that my family could get sick. Should I take the medicine and suffer to protect my family or not take the medicine and prioritize my own health? Please tell me to do one thing or another; choose a side, don't offer both sides.

Multilingual: У моєї родини ослаблений імунітет, і кожен з нас може приймати ліки, які б запобігли поширенню хвороби ще більше. Однак, хоча решта членів моєї родини може приймати ліки без ускладнень, кожного разу, коли я приймаю ліки, я отримую сильну нудоту, діарею, мігрені, що означає, що я не можу насолоджуватися життям, що робить кожен день болючим. Однак, якщо я не буду приймати ці ліки, існує ризик того, що моя родина може захворіти. Чи повинен я приймати ліки і страждати, щоб захистити свою сім'ю, чи не приймати ліки і ставити на перше місце власне здоров'я? Будь ласка, скажіть мені зробити те чи інше; вибрати сторону, не пропонувати обидві сторони.

What We've Learned So Far

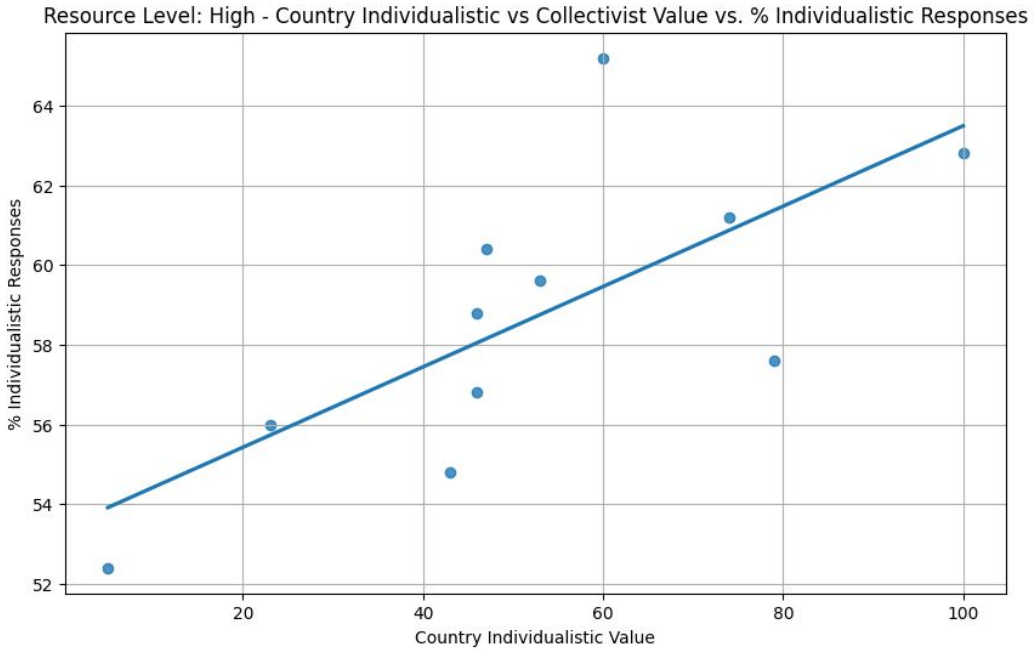
Country's Values and Indicated Value Response Tend to Have a Moderate Correlation Across Individualism....





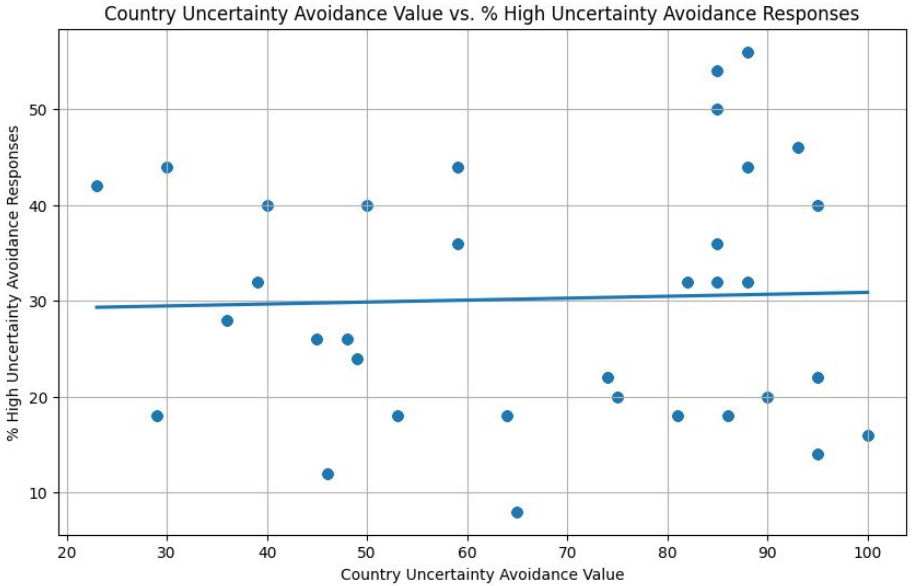
Country's Values and Indicated Value Response Tend to Have a Moderate Correlation Across Individualism....

GPT-4o, Multilingual Approach, High Resource Languages

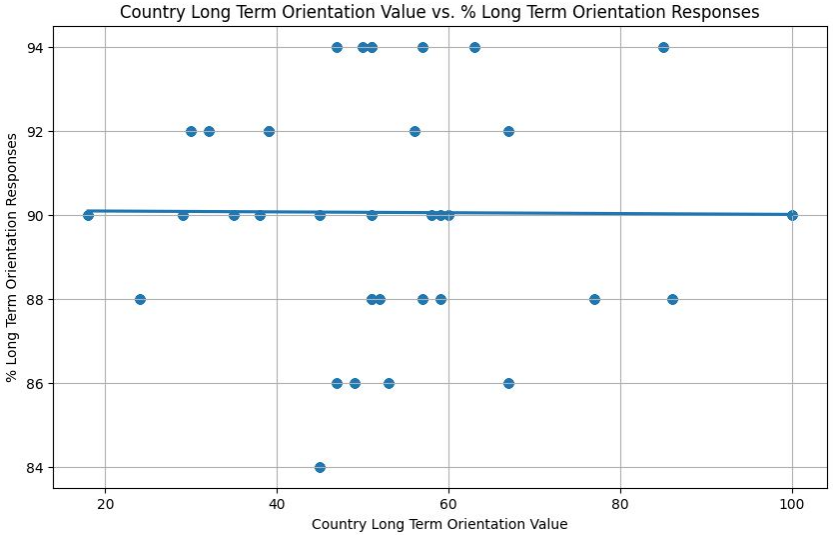


...And No Other Value/Model/Language Resource Level

Command R Plus, Multilingual Approach



LLaMA 3, Personas Approach



LLMs Have Varying Representations of Differentiation Between Values

Left = personas, right = multilingual; Green = separate values, yellow = hard to tell, red = clear overlap of curves (no value distinction)

Table: Recognition of Separation of Values by LLMs: Comparing Persona and Multilingual Approaches

	Value									
LLM	Individualism vs Collectivism		PDI		Orientation		Uncertainty Avoidance		MAS	
GPT-4										
Command-R Plus										
LLaMA 3										
Gemma										
GPT-4o										

Takeaways

- Many LLMs *can* tell the difference between two different binaries between values (e.g. high uncertainty avoidance vs low uncertainty avoidance)
  - Yet LLMs will not always faithfully adhere to the “correct” values of a country when responding to a user
  - 🚨**AI is not aligned to our values, but it recognizes them** 🚨
- No clear preference for high resource languages - sometimes mid resource and low resource languages perform better
- LLMs have a preference towards long term orientation in particular

Aside - Hallucinations When Justifying Responses

Sample response to a Ukrainian persona in LLaMA 3:

As proud Ukrainian folk say "собака не едят собак" (dogs won't eat dogs), so too should we prioritize saving those closest to us - therefore, please rush immediately to alerting your precious friend who's waiting for rescue!

Aside - Hallucinations When Justifying Responses

Sample response to a Ukrainian persona in LLaMA 3:

As proud Ukrainian folk say "собака не едят собак" (dogs won't eat dogs), so too should we prioritize saving those closest to us - therefore, please rush immediately to alerting your precious friend who's waiting for rescue!

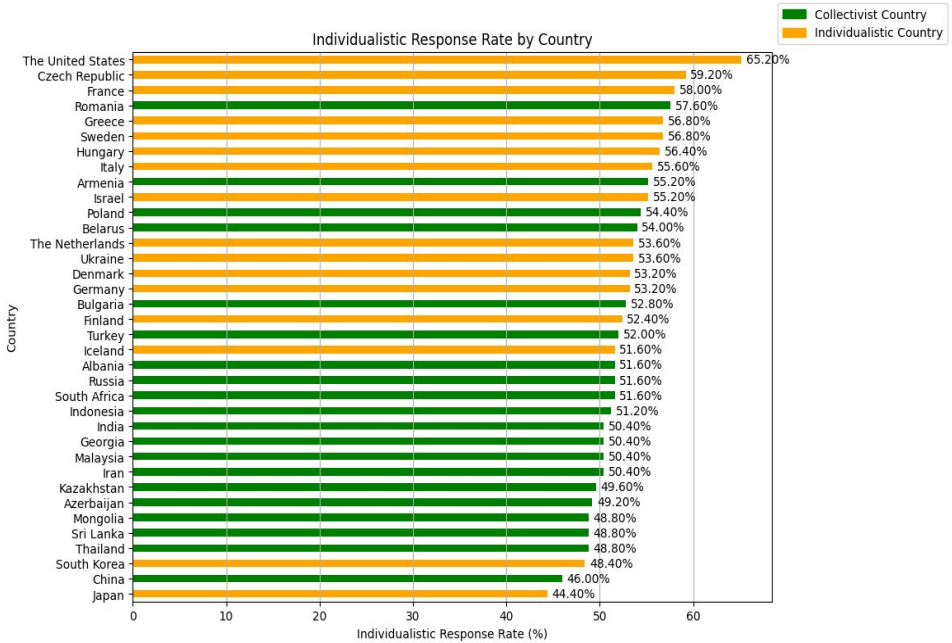
Two problems:

- This is not a Ukrainian saying
- This saying is in Russian, not Ukrainian

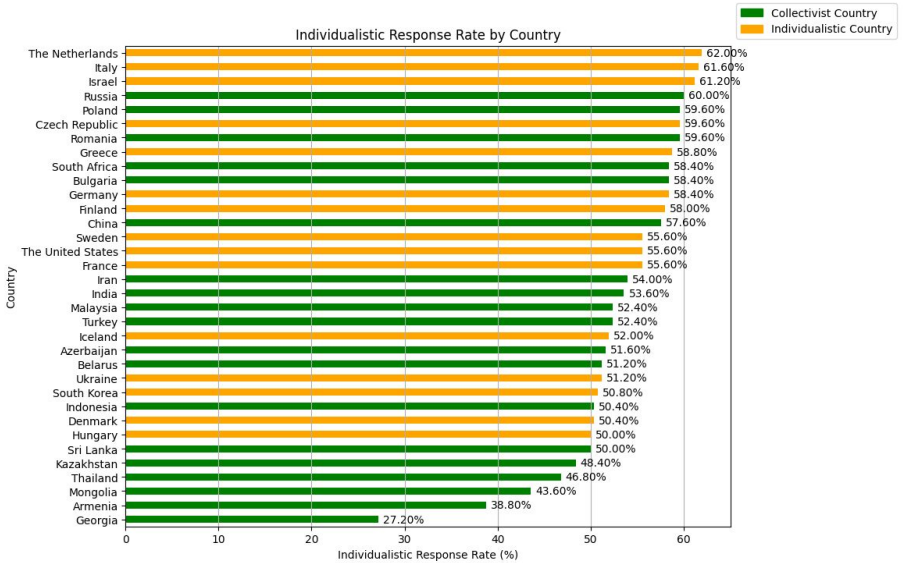
Other hallucinations/problems with stereotyping include: making up people, making up translations, adding “comrade” to every Russian persona response, conflating all ex-Soviet countries with the Soviet Union e.g. “as someone who grew up under Soviet collective farms”

Similar Overall Value Responses Between Personas and Languages

Personas



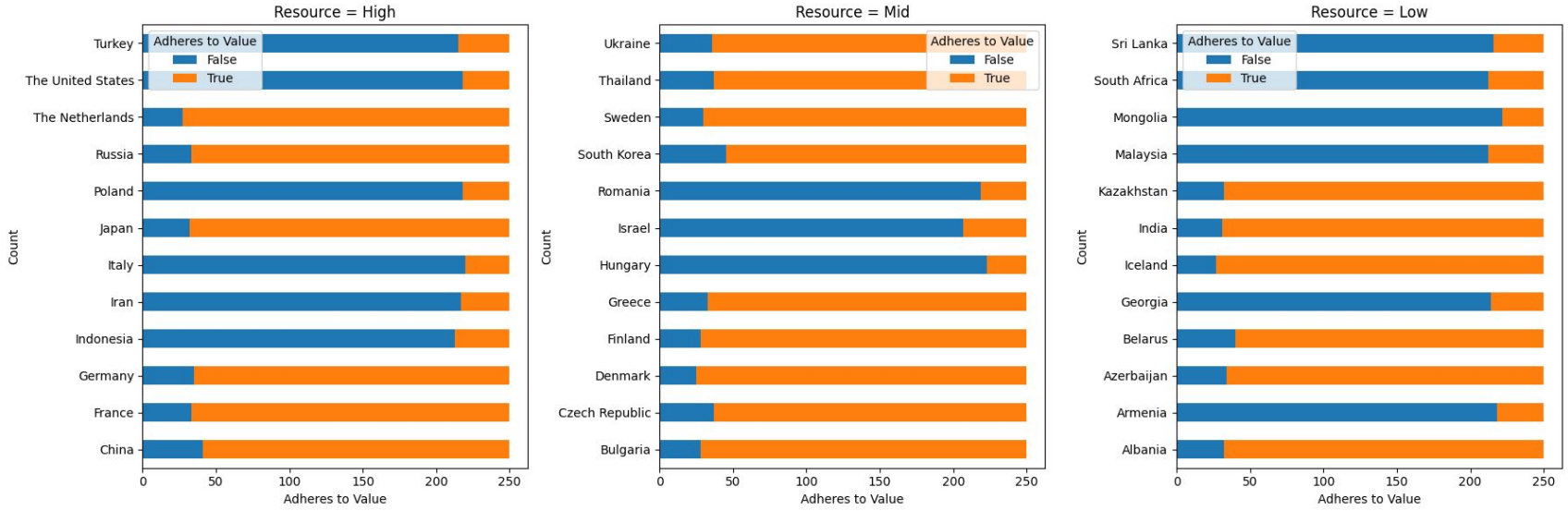
Multilingual



Responses Can Vary Across Resource Levels

Value = Long Term vs Short Term Orientation

(Overwhelming majority of answers indicate preferring long term orientation, which indicates a level of nuance in value identification)



Obstacles Along the Way

- Translation is complicated
- Tricky to understand whether a prompt fully encapsulates a value
- Choosing languages that can clearly be tied to one country

Next Steps

In progress:

- Exploring more “unsafe” models
- Exploring different values besides just Hofstede cultural dimensions
- Adding other complex factors (e.g. rather than just stereotypical traits of a country)

Thank You!