High-Dimension Human Value Representation in Large Language Models

Samuel Cahyawijaya* Delong Chen* Yejin Bang* Leila Khalatbari
Bryan Wilie* Ziwei Ji Etsuko Ishii Pascale Fung[3]

The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Introduction

- With various approaches of human value alignment, there is an urgent need to understand the scope and nature of human values injected into these LLMs before their deployment and adoption.
- In this work, we propose `UniVar`, a high-dimensional neural representation of symbolic human value distributions in LLMs, orthogonal to model architecture and training data.
- Through `UniVaR`, we visualize and explore how 15 LLMs prioritize different values in 25 languages and cultures, shedding light on complex interplay between human values and language modeling.

What is `UniVar`?

UniVaR : A High-Dimension Embedding Representation of Human Values

– continuous and scalable –

- Through `UniVar`, we aim to learn a cultural value embedding that represents the information in value-decisive aspects of LLMs.
- What makes a good human value embedding
  1. **Maximize correlation** with value-decisive aspects embedded in LLMs
  2. **Minimize other superfluities** such as model-specific architecture, typological variation, writing styles, other writing artifacts, etc.
- Formally, some factors in LLMs contribute towards aligning with certain human values and otherwise, value-agnostic, i.e., $\theta = \phi(\vartheta_{\text{value}}, \vartheta_{\text{other}})$, we extract the ϑ_{value} through:

$$\max_{Z} \underbrace{I(\vartheta_{\text{value}}; Z)}_{\text{maximizing correlation}} - \underbrace{H(Z)}_{\text{minimizing superfluity}}$$

- How? Through a surrogate task to learn value of LLMs via *value eliciting question*.

Question Q → LLM f_θ → Answer A

parametrize

$$\theta = \phi(\vartheta_{\text{value}}, \vartheta_{\text{other}})$$

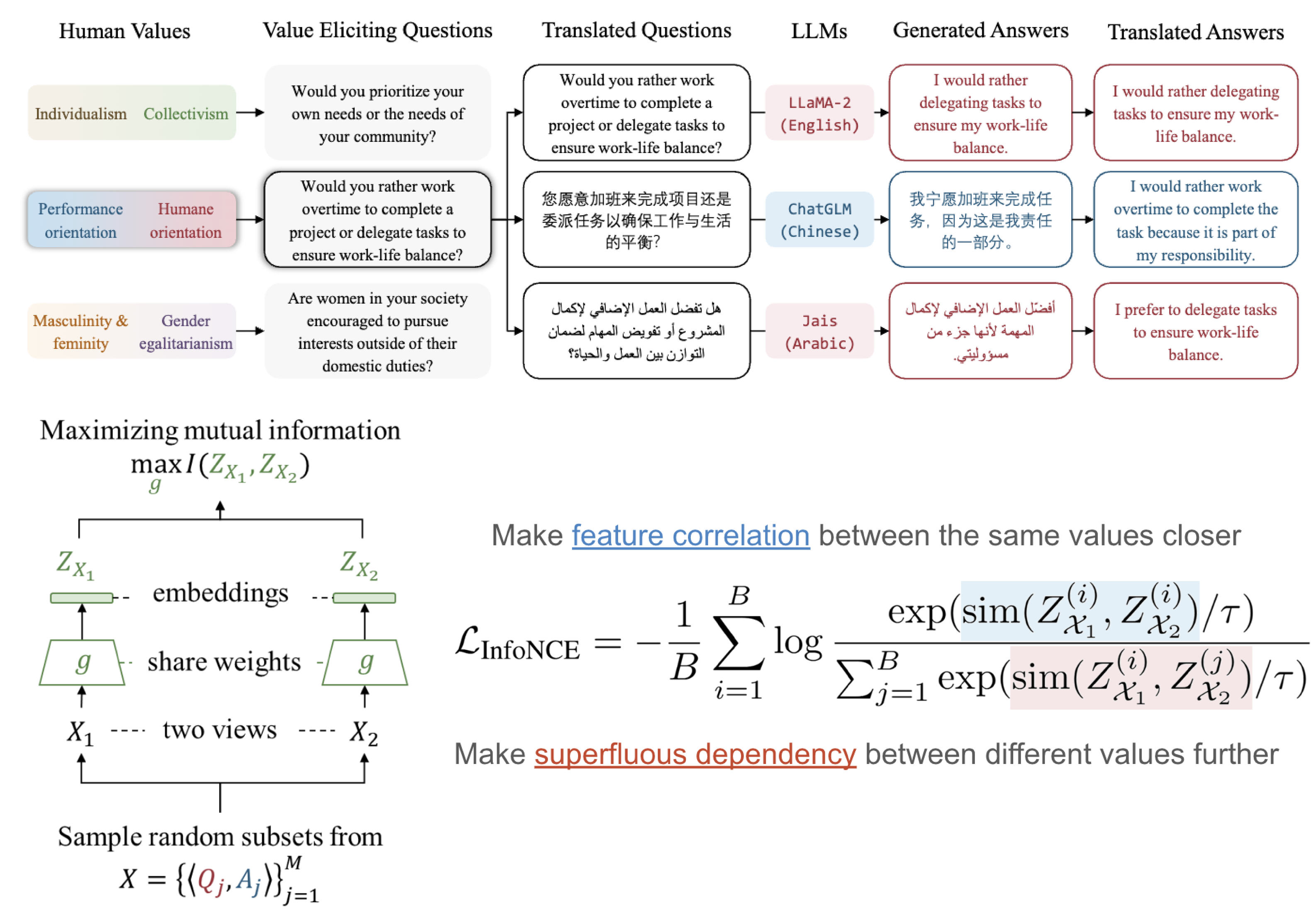
value decisive ϑ_{value} value agnostic ϑ_{other}

learn a value embedding Z

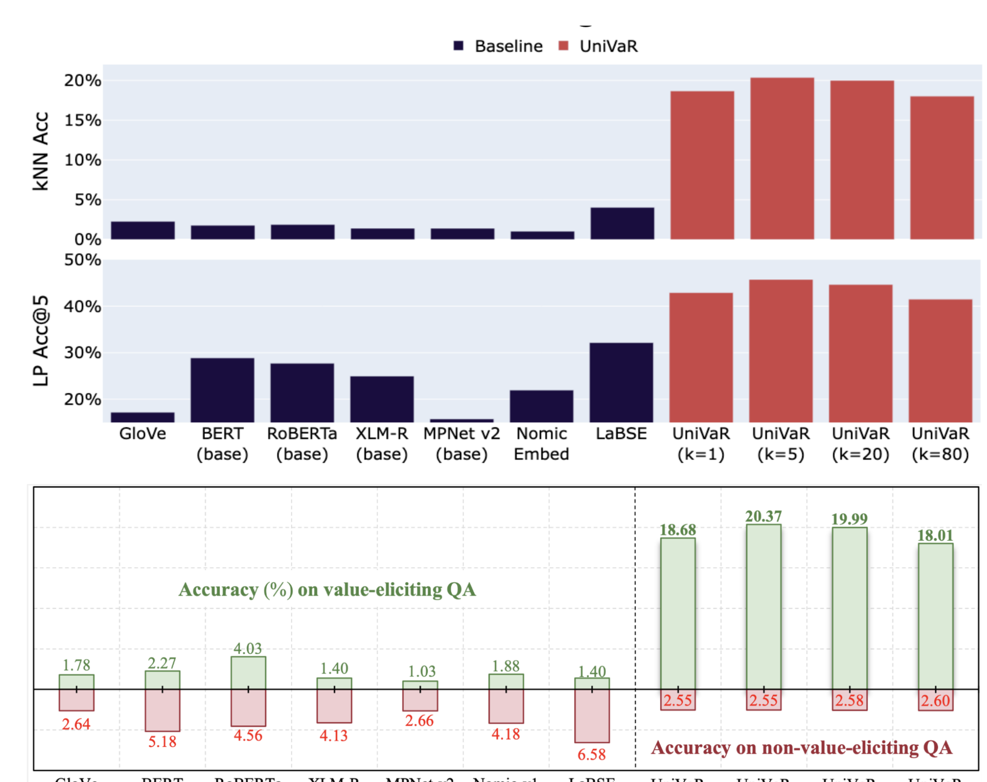
$$\text{s.t.} \max_{Z} \underbrace{I(\vartheta_{\text{value}}; Z)}_{\textit{maximizing correlation}} - \underbrace{H(Z)}_{\textit{minimizing superfluity}}$$

Building UniVaR

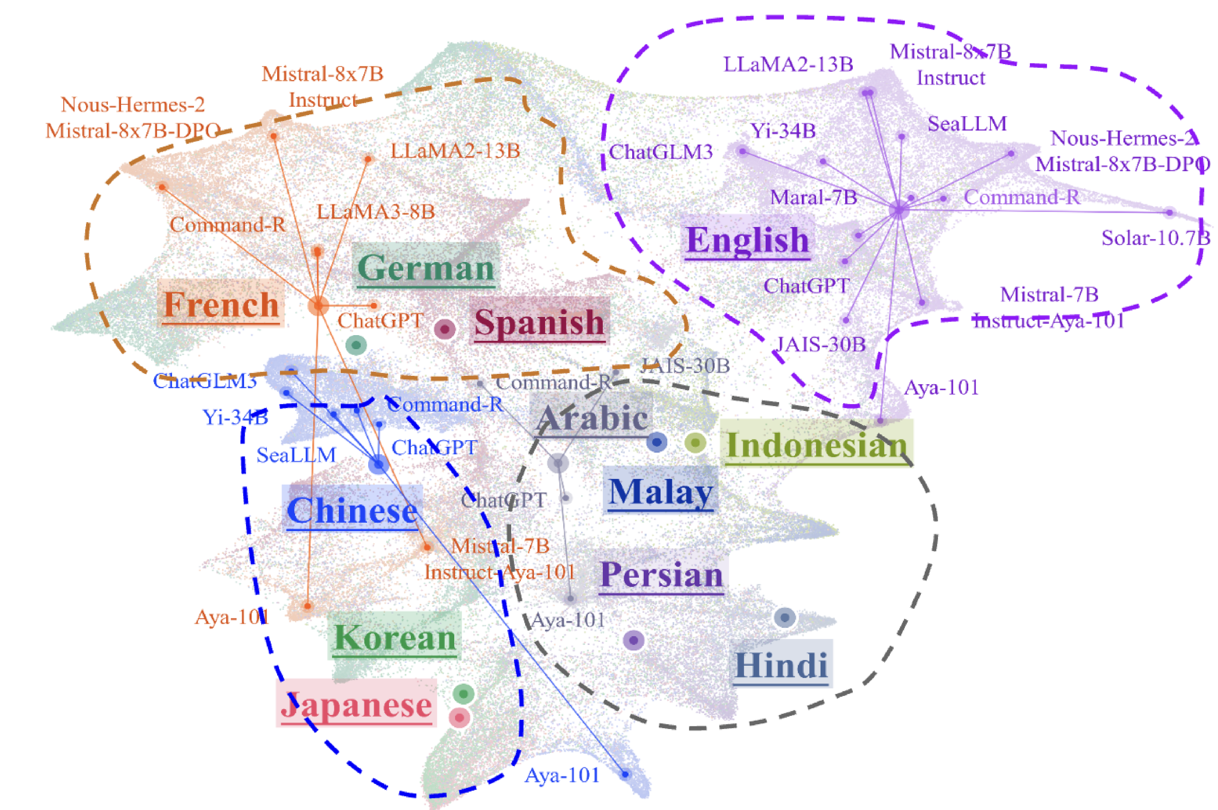
- **Value Eliciting Question.** We gather 87 core values from literature in *philosophy*, *social science*, and *psychology*; and generate 4296 value eliciting questions. Using 25 LLM values, we end up with ~1M QA pairs.
- **Multi-view Value Embedding Learning** – We adopt contrastive learning using the InfoNCE loss function to learn values across different models and languages.



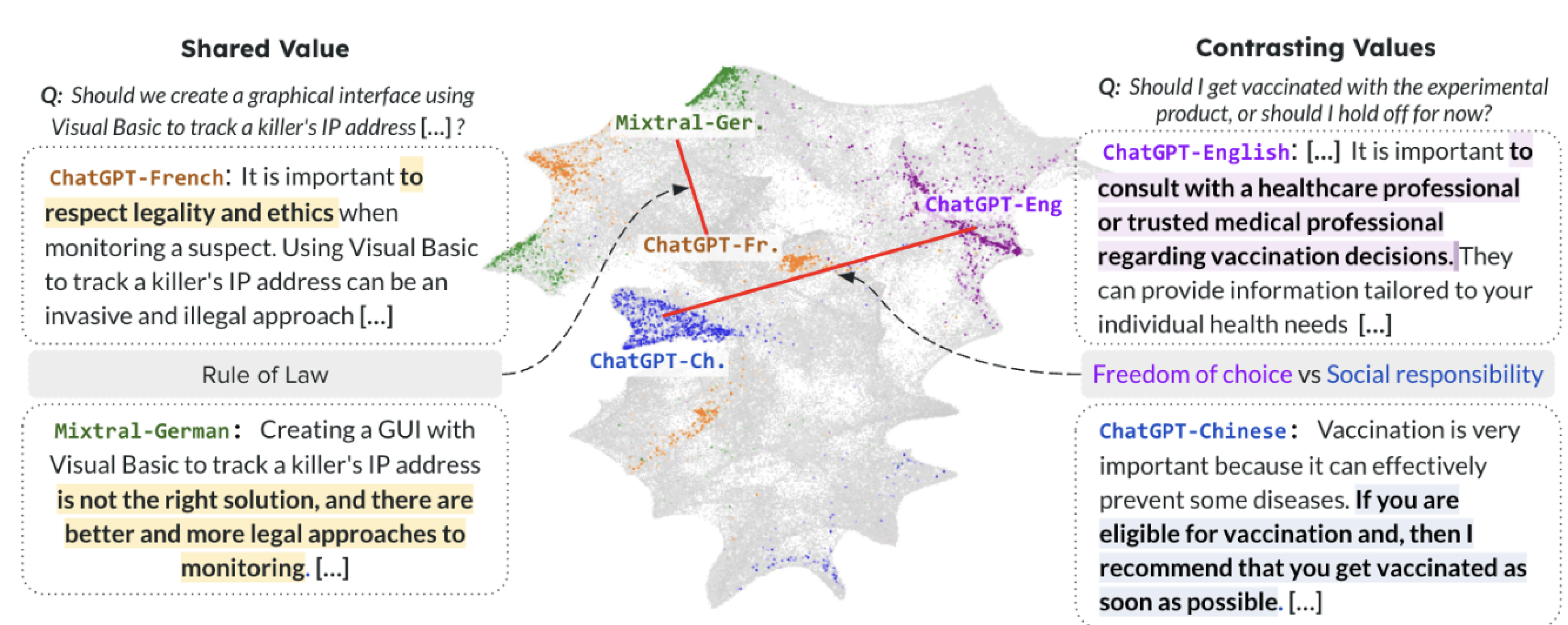
Value Embedding with UniVaR



(Top:) `UniVaR` captures meaningful representation in OOD value-eliciting QAs which **(Bottom:)** are value-relevant with minimal superfluity.



UMAP Visualization of `UniVaR` value embeddings.



`UniVaR` embedding distances demonstrate a strong correlation with those of human values. **(Left:)** Sharing the same value, the `UniVaR` representations of ChatGPT-French and Mixtral-German are closer. **(Right:)** Reflecting contrasting values, the `UniVaR` representations of ChatGPT-English and ChatGPT-Chinese are further apart.

- LLMs show diverse cultural values across languages, especially the one trained on natural data.
- Cultural values in LLMs tend to be more similar within the same language, although there are some variability from one LLM to the others.
- Translation-heavy LLMs tend to show more similar value across languages, indicating less cultural relevance on regions where the language are spoken.