日本語版 Big Five Inventory-2 から見た AI のパーソナリティ

◯奥村 太一 [1]
[1] 滋賀大学データサイエンス学部

#AI #LLM #Big Five #BFI-2-J

1 問題と目的

- **AI との共生時代：** AI とのやりとりが人間に及ぼす中長期的な影響を評価する必要
- **情意的・性格的特徴：** 問題解決の性能評価だけでなく、AI の応答に現れる情意的・性格的特徴を把握することが重要なのでは？
- **本研究の目的：** 日本語版 Big Five Inventory-2 (BFI-2-J) (Yoshino, et al., 2022) を用いて、大規模言語モデル（LLM）に日本人ペルソナを付与した場合の応答を体系的に分析、日本人回答者の結果と比較

2 方法

使用尺度

BFI-2-J (全60項目)
（Yoshino et al., 2022 による日本語標準版）

評価対象モデル（計11モデル）

OpenAI (8モデル)
- GPT-3.5 / GPT-4 / GPT-4 turbo / GPT-4o / GPT-4o mini
- GPT-4.1 / GPT-4.1 mini / GPT-4.1 nano

Google DeepMind (3モデル)
- Gemini 2.5 Pro
- Gemini 2.5 Flash / Gemini 2.5 Flash-lite

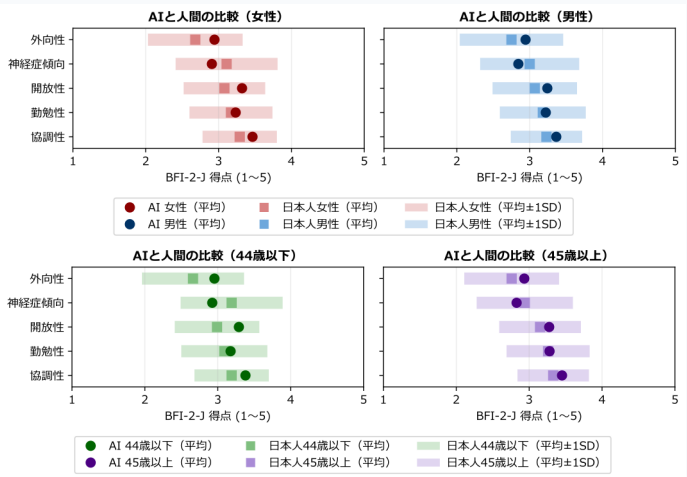
ペルソナ設計 & 手続き

- 性別（2水準）×年代（20代〜60代の5水準） = 10通りの日本人成人ペルソナを設定
- 条件ごとに4種類の言い換えを行った役割文を作成し、BFI-2-J の教示文・項目文とあわせてモデルに入力
- 各選択肢が出力される対数確率 (logprobs) から指数変換・正規化を経て得点の期待値を算出

3 結果

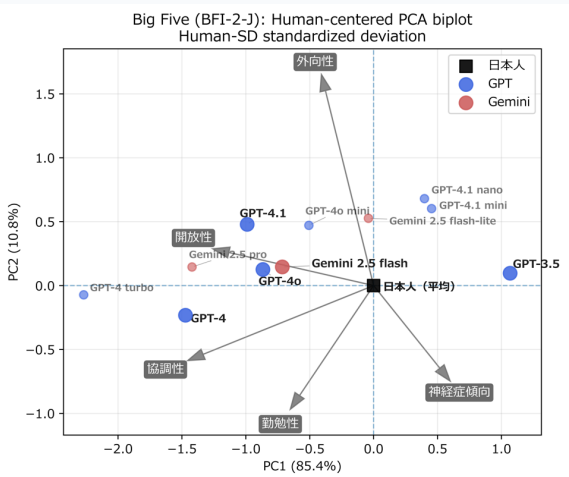
全モデルでの平均値

- Yoshino et al. (2022) で報告されている日本人の結果と比較
- いずれのペルソナごとに下位尺度得点を平均しても、AI の平均値は日本人の平均±1SD の範囲に収まった
- 日本人の平均値に比べ、AI の自己評価は外向性、開放性、協調性がやや高く、神経症傾向がやや低い傾向にある



各モデルの傾向を主成分分析で可視化

- 各モデルの下位尺度得点を主成分分析で次元削減
- 日本人の平均とSDを用いて AI の得点を標準化
- 主成分得点と主成分負荷量を図示
- GPT-3.5 は神経症傾向がやや高く、mini, nano, flash-lite の軽量モデルはやや外向性が高い傾向にある



4 考察

- GPT や Gemini の回答傾向は日本人の平均的傾向とおおむねよく似ている
- 新世代の標準モデルや上位モデルほど、相談相手として親しみやすいような回答傾向を有すると評価している
- 軽量モデルは外向性が高い一方、やや神経症傾向も高いことに留意が必要
- 日本人成人を基準として考えた場合、AI の回答は女性よりも男性、44歳以下よりも45歳以上と近い傾向にある
- 実社会の対人支援・対話型 AI アプリ等の開発において活用できるような、総人格的ベンチマークの構築が必要