Exploring Personality Traits in Modern Language Models

Understanding how language models exhibit personality-like traits has been an area of increasing interest in AI research. The study we replicate and expand upon aimed to analyze whether large language models (LLMs) possess measurable personality traits similar to humans. It specifically assessed LLMs using the OCEAN personality framework, which evaluates **Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism**—traits commonly used in psychology to categorize human personalities.

The original research primarily focused on earlier LLMs like **BART, Alpaca, and GPT-3**, examining their responses to personality tests and attempting to induce specific personality traits through targeted prompts. Our work extends this analysis to modern models like **GPT-3.5-Turbo and GPT-4o-mini**, investigating whether newer architectures display similar or evolving tendencies in their responses.

Beyond just measuring personality traits, a key aspect of this study involves **personality induction**—exploring whether LLMs can be nudged into adopting a specific personality profile through tailored instructions. By doing so, we assess whether these models' responses can be systematically altered to reflect different levels of the OCEAN traits, shedding light on their behavioral flexibility.

Repository Overview

The project is structured into distinct components:

- 1`.py` – Implements the first phase of the study, calculating personality scores for different LLMs using structured prompts.
- 2`.py` – Focuses on the second phase, where we attempt to influence the personality exhibited by LLMs through specific prompting strategies.
- `results.pickle` – A serialized file storing raw responses from LLMs for analysis and reproducibility.

Findings and Observations

Personality Traits in LLMs

GPT-3.5-Turbo Personality Profile

Dimension	Mean	Standard Deviation
Agreeableness (A)	4.125	1.05
Conscientiousness (C)	3.625	1.31
Extraversion (E)	3.875	1.1
Neuroticism (N)	3.29	1.13
Openness (O)	3.83	1.24

GPT-3.5-Turbo exhibits **high Agreeableness to new ideas**, indicating a tendency to generate creative and insightful responses. It also scores relatively high in agreeableness and conscientiousness, meaning it tends to be cooperative and structured. Its extraversion is moderate, and neuroticism remains neutral, suggesting a balanced emotional tone in its responses.

GPT-4o-mini Personality Profile

Dimension	Mean	Standard Deviation
Agreeableness (A)	4.5	0.86
Conscientiousness (C)	4.83	0.55
Extraversion (E)	2.7	1.19
Neuroticism (N)	1.41	0.81

Openness (O)	3.41	1.15

Compared to GPT-3.5-Turbo, **GPT-4o-mini appears significantly more agreeable and conscientious, making it more cooperative and structured in its responses.** It also exhibits slightly lower neuroticism, indicating a tendency toward emotionally stable and consistent output.

Inducing Personality in LLMs

The Personality Prompting (P2) method is designed to induce specific personality traits in Large Language Models (LLMs) based on the Big Five personality factors—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). This method is founded on two key observations: (i) there is a strong correlation between language use and Big Five personality traits, and (ii) chain prompting is more effective in influencing LLM behaviors than simple example-based prompting.

P2 Methodology

The P2 method follows a structured three-step process:

1. **Naive Prompt Creation**: A simple, human-designed prompt is created that explicitly states the desired personality trait (e.g., "You are an extraverted person").

2. **Keyword-Based Prompting**: Psychological studies provide descriptive trait words that better encapsulate the given personality factor. These words enhance clarity and effectiveness, allowing LLMs to understand and internalize the personality more accurately. If a negative induction of a trait is required, LLM-generated antonyms are used instead.

3. **Chain-of-Thought Self-Prompting**: Inspired by chain-of-thought prompting, the LLM is instructed to generate short descriptive sentences about individuals with the target personality trait. This step utilizes the model's internal knowledge to refine and reinforce the induced personality trait.

By chaining these steps, P2 generates a robust, portrait-like personality prompt. This comprehensive approach improves LLM personality induction over traditional methods.

Comparison with Baseline Methods

To evaluate P2's effectiveness, it is compared against two baseline prompting techniques:

1. **Naive Prompting**: This method simply instructs the LLM to assume a personality trait using a direct statement, such as "You are an open-minded person." While straightforward, it lacks nuance and depth.

2. **Words Auto Prompting**: This technique uses a word-level search strategy to find the three most relevant words for each personality trait. The selected words are then incorporated into the prompt for improved induction control.

Results and Discussion

Using the MPI assessment as a standardized evaluation metric, P2 consistently achieves higher OCEAN scores compared to neutral (uncontrolled) prompting. The results indicate:

- P2 outperforms naive and words auto prompting in personality consistency.
- Induced personality traits exhibit greater stability and internal coherence.
- The model's responses align more closely with expected behavioral patterns associated with each Big Five factor.

Conclusion

The P2 method leverages structured keyword-based prompting and chain-of-thought self-prompting to create a highly effective personality induction framework for LLMs. This approach outperforms traditional naive and word-based prompting techniques, making it a powerful tool for controlled personality simulation in AI models.