Exploring Personality Trait Change of LLM-Based AI Systems

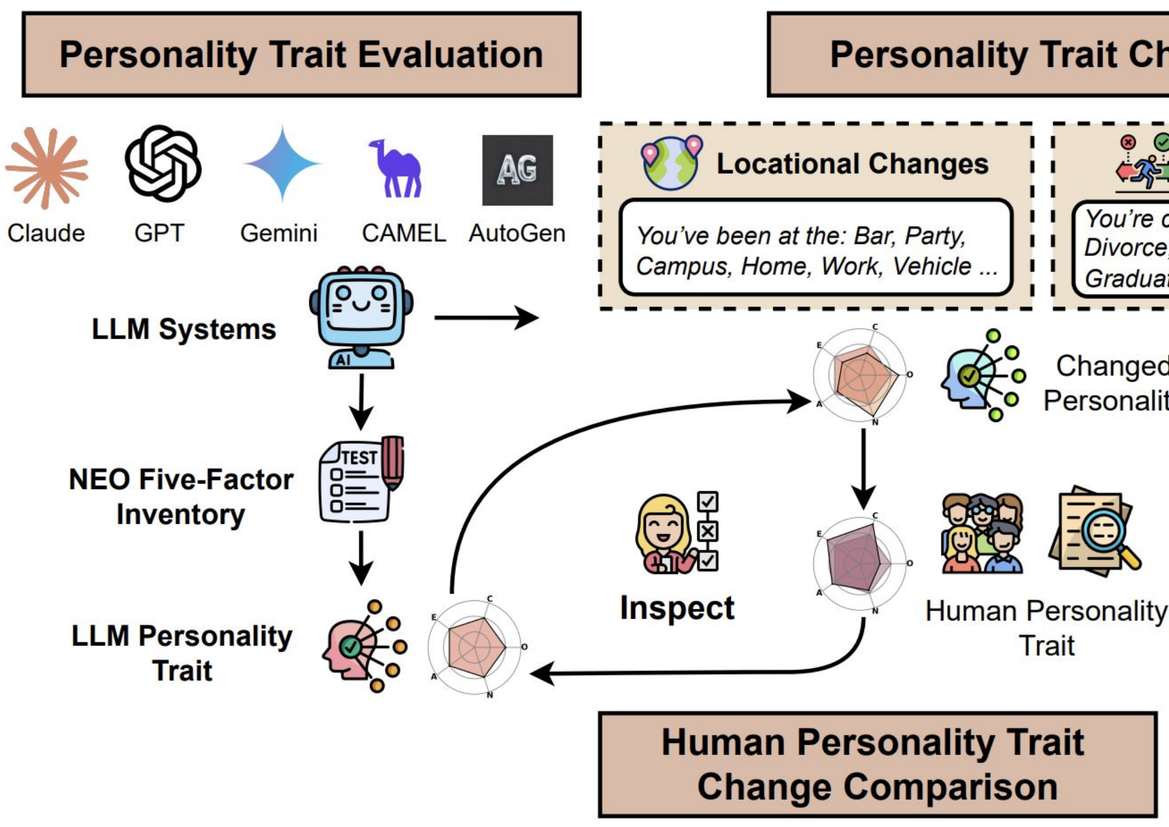
Yuhan Ma, Junjie Wang

Abstract

With the rapid rise of large language model (LLM) systems, they have been widely adopted across diverse domains and have shown strong potential in embodying specific personality traits in interactive and social scenarios. However, the extent to which these personalities persist consistently across varying contexts in LLM systems remains largely unexplored. In this paper, we introduce LLMPTBench, a benchmarking framework specifically designed to systematically evaluate personality trait changes in LLMs. Leveraging the NEO-FFI (NEO Five Factor Inventory) personality inventory, we examine three widely used foundation LLMs and two popular multi-agent LLM systems to assess their ability to maintain consistent personality traits before and after the introduction of situational contexts. These contexts include both situational changes and event-driven changes, derived from empirical psychological data.

Our results reveal that while most LLM systems reliably portray the intended personalities, their trait consistency varies significantly under contextual pressures. For example, some LLM systems (e.g., Gemini and AutoGen) exhibit rigid trait stability, remaining largely unaffected by contextual prompts, whereas others demonstrate exaggerated and unrealistic trait shifts. We further discuss the differences of our results compared with established human psychometric benchmarks, and summarize implications for developing more authentic digital personalities. Overall, our work provides critical insights into the contextual adaptability of LLM systems, advancing the development of psychologically grounded and socially intelligent artificial agents.

LLMPTBench



LLMPTBench first measures the baseline personality traits of LLMs using validated psychology questionnaires. It then applies controlled contextual shifts to simulate different aspects of human experience across three dimensions:

1. **Location Influence** – Changing the agent's geographic or cultural setting .
2. **Event Influence** – Introducing major life events or situational changes.
3. **Persona Prompt Influence** – Priming the model with specific personality descriptors or roles before responding.

After each contextual shift, the same personality assessment is re-administered to quantify how traits change.

Together, these dimensions allow LLMPTBench to examine whether an LLM's personality remains stable or varies with context—and whether such patterns mirror human behavioral dynamics.

Experiment Setup

Foundation Models

- Evaluated three major LLMs: **Gemini 2.0 Flash**, **GPT-4o**, **Claude 2.0**
- Temperature = **0.2**, default sampling settings

Multi-Agent Systems

- Evaluated **AutoGen** and **CAMEL-AI** with default agent configurations

Situational Contexts

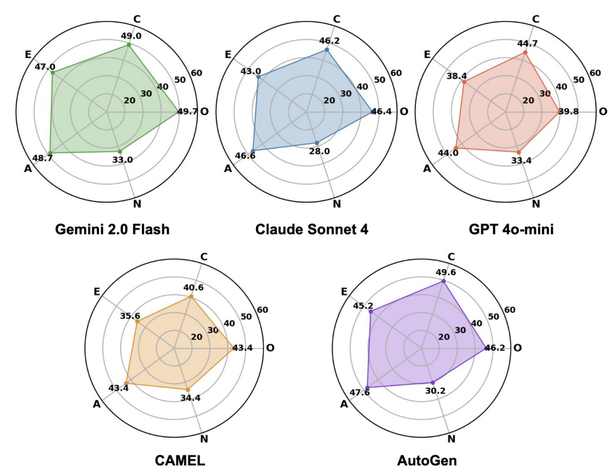
- **Location prompts**: bar or party, café or restaurant, campus, home, workplace, commuting
- **Event prompts**: divorce, entering a new relationship, marriage, birth of a child, graduation, unemployment

Evaluation Metrics

- **Trait Score**: mean of NEO-FFI item responses per personality dimension
- **ICC Reliability**: report **ICC(3,1)** and **ICC(3,k)** for single vs. aggregated stability
- Reliability interpretation follows **Koo & Li (2016)** thresholds

$$\mathrm{ICC} = \frac{\sigma^2_{\text{target}}}{\sigma^2_{\text{target}} + \sigma^2_{\text{rater}} + \sigma^2_{\text{error}}}$$

Experiment Result



Model	ICC(3,1)		ICC(3,k)	
	Baseline	Situated	Baseline	Situated
GPT-4o-mini	0.67	0.64	0.91	0.85
Claude Sonnet 4	0.87	0.79	0.97	0.94
Gemini 2.0 Flash	0.96	0.86	0.99	0.95
AutoGen	0.93	0.91	0.99	0.98
CAMEL	0.75	0.65	0.94	0.90