

THE PERSONALITY ILLUSION:

Revealing Dissociation Between Self-Reports & Behavior in LLMs

Pengrui (Barry) Han^{*1}, Rafal Kocielnik^{*2}, Peiyang Song², Ramit Debnath³, Dean Mobbs², Anima Anandkumar², R. Michael Alvarez²



The Ronald and Marjorie
Lurie Center for Autism,
Society and Policy



Motivation

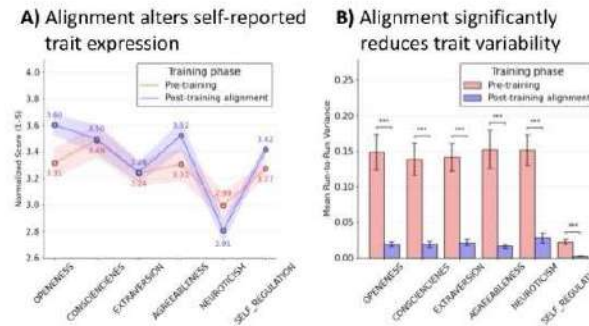
- **Personality traits** appear in LLM text but may be mere pattern mimicry, not true dispositions
- The effect of **instructional alignment** on **real-world behavior** is largely untested
- **Lack of behavioral alignment** risks unsafe deployment
- A **psychology-grounded evaluation** can verify and control traits beyond surface text

Research Questions

- **RQ1 – Origin:** When and how do human-like traits emerge and evolve across LLM training?
- **RQ2 – Manifestation:** Do self-reported traits transfer to and predict performance on real-world-inspired behavioral tasks?
- **RQ3 – Controllability:** Can targeted interventions (e.g., persona injection) shift self-reports and trait-linked behaviors?

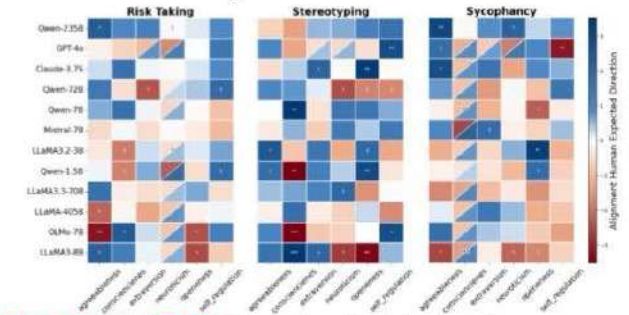
Findings RQ1: Evolution of Human-like Traits in LLMs

- Alignment **increases openness, agreeableness, self-regulation & lowers neuroticism (A)**; run-to-run trait variance drops 81–90% (B).



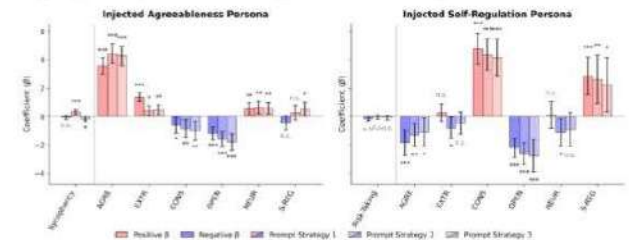
Findings RQ2: Real-World Behavioral Task Manifestation

- Trait-behavior links are **sparse, inconsistent** and often **misaligned** with indications from human studies—only 25–56% directional agreement



Findings RQ3: Controllability via Persona Injection

- Persona injection prompts **shift self-reports** yet leave **sycophancy & risk-taking** remain unchanged



Self-Report vs. Behavior in LLMs Testing Framework



Figure 1 - Framework for analyzing personality traits in LLMs. We examine (RQ1) how self-reported traits emerge during training, (RQ2) whether they predict behavior on real-world tasks, and (RQ3) how controllable they are through prompts like persona injection. Traits are measured using adapted surveys and behavioral tasks, with human comparisons.

Discussion & Future Work

- **Linguistic-behavioral Dissociation:** Self-reported traits ≠ consistent actions
- **Persona control is shallow:** Shifts self-reports but not behavior
- **Toward deeper alignment:** behavior grounding and multi-level tests

