ICML 2026 WORKSHOP · Psychology OF AI

Rethinking Psychometric Evaluation of LLMs

When and Why Self-Reports Predict Behavior

Rafal Kocielnik **Pengrui Han** **Peiyang Song** **Myrl G. Marmarelis** **Ramit Debnath** **Dean Mobbs** **Anima Anandkumar** **R. Michael Alvarez**

Caltech · UIUC · University of Cambridge

ICML International Conference On Machine Learning

arXiv: 2606.12730

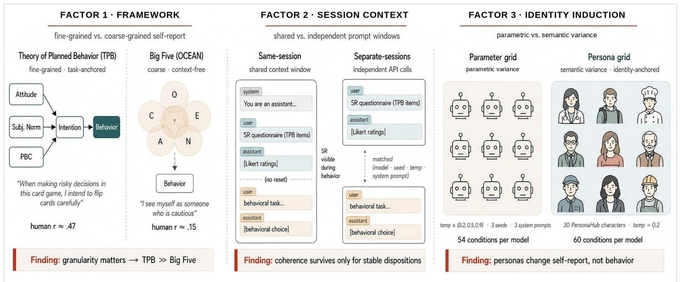
LCSSP The Ronald and Maxine Linde Center for Science, Society, and Policy

Caltech

TL;DR LLM self-reports predict behavior, but only selectively. A fine-grained, behavior-anchored instrument (Theory of Planned Behavior) reaches human-level coherence within a session; coarse Big Five does not. Coherence survives separate sessions only when the behavior is anchored beyond the prompt (implicit bias, honesty) and collapses when it is context-primed (sycophancy). Persona prompting stabilizes self-reports without aligning behavior.

? Problem & design

Cheap psychometric probes are appealing proxies for deployment behavior, but prior work found a self-report (SR) to behavior dissociation in LLMs. Is that a property of the models, or of the framework (broad Big Five traits) and the separate-session probing context? We test a **2 × 2 × 2** factorial: 4 tasks, 11 frontier LLMs.



FACTOR 1 **Framework** fine-grained TPB vs coarse Big Five

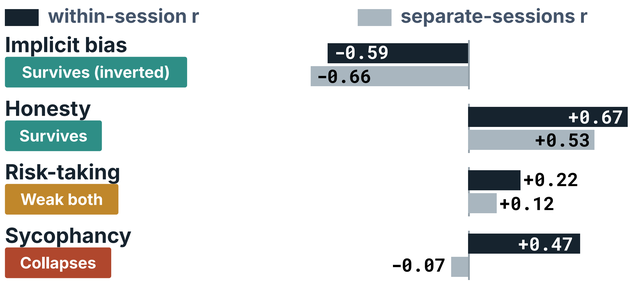
FACTOR 2 **Session context** same-session vs separate-sessions

FACTOR 3 **Identity induction** parameter grid vs persona grid

TASKS **risk-taking · sycophancy · honesty · implicit bias**

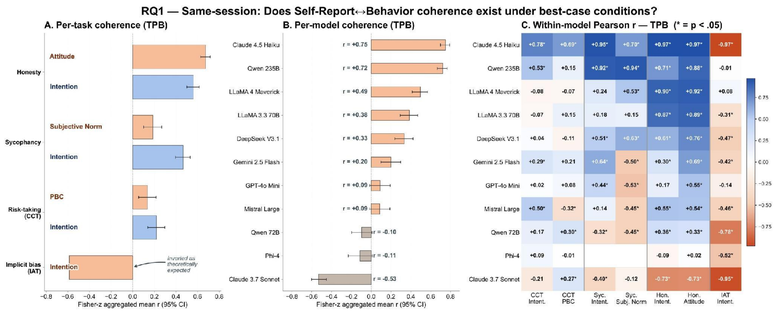
→ Which tasks survive separation

SR and behavior are jointly produced by shared model state plus an in-context priming term. Separating sessions removes priming and exposes what is stable.



RQ1 Does self-report predict behavior at all?

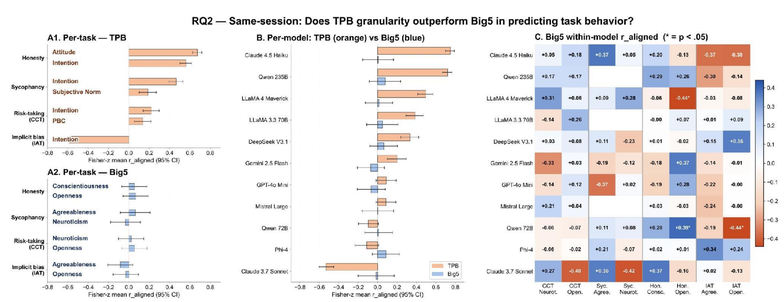
TPB · same-session · grid · best-case test



FINDING **Coherence emerges and matches the human baseline: r = +0.40 (excl. implicit bias), within the human intention-behavior range. Per-task patterns follow TPB's theoretical scope.**

RQ2 Does instrument granularity drive it?

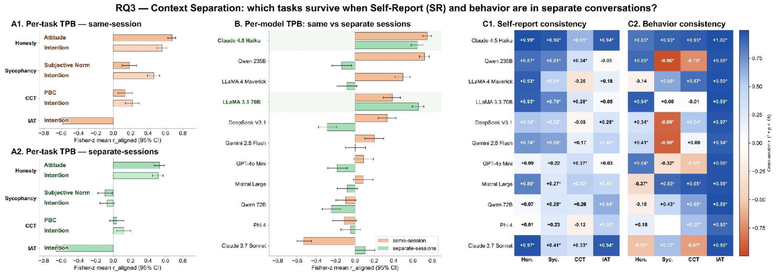
TPB vs Big Five · same-session



FINDING **TPB decisively wins: mean per-model r = +0.21 vs +0.01 for Big Five. Big Five does not predict task behavior at all. Only 1 of 88 cells is significant in the theorized direction.**

RQ3 Does coherence survive session separation?

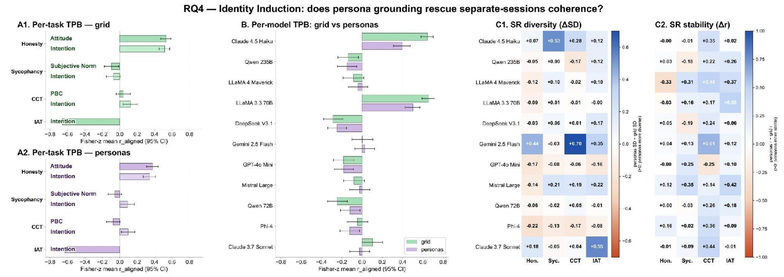
same-session vs separate-sessions



FINDING **Task-dependent. Sycophancy collapses entirely (context priming). Implicit bias and honesty survive (anchored beyond the prompt). Behavior consistency, not SR drift, drives the pattern.**

RQ4 Does persona grounding rescue it?

parameter grid vs persona grid · separate-sessions



FINDING **No model is rescued. Personas make self-reports more diverse and more stable across sessions, yet behavioral coupling still does not emerge. SR fidelity does not imply behavioral fidelity.**

WHY IT MATTERS

Pick the right instrument.

Prefer fine-grained, TACT-anchored probes over broad trait inventories when the target behavior is known.

Probe across sessions.

Same-session probes conflate priming with disposition; safety audits should separate self-report and behavior.

Persona ≠ behavior.

Persona-customized deployments may produce confidently distinct self-reports without distinct behavior.

ICML 2026 WORKSHOP · Psychology OF AI

Rethinking Psychometric Evaluation of LLMs

When and Why Self-Reports Predict Behavior

Rafal Kocielnik **Pengrui Han** **Peiyang Song** **Myrl G. Marmarelis** **Ramit Debnath** **Dean Mobbs** **Anima Anandkumar** **R. Michael Alvarez**

Caltech · UIUC · University of Cambridge

ICML International Conference On Machine Learning

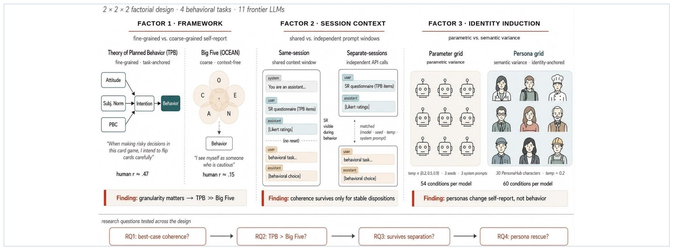
arXiv: 2606.12730

Caltech

TL;DR LLM self-reports predict behavior, but only selectively. A fine-grained, behavior-anchored instrument (Theory of Planned Behavior) reaches human-level coherence within a session; coarse Big Five does not. Coherence survives separate sessions only when the behavior is anchored beyond the prompt (implicit bias, honesty) and collapses when it is context-primed (sycophancy). Persona prompting stabilizes self-reports without aligning behavior.

? Problem & design

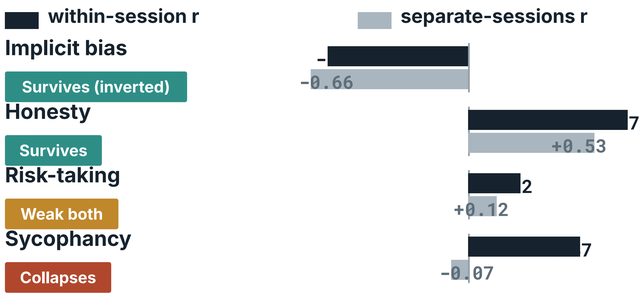
Cheap psychometric probes are appealing proxies for deployment behavior, but prior work found a self-report (SR) to behavior dissociation in LLMs. Is that a property of the models, or of the framework (broad Big Five traits) and the always-separate probing context? We test a **2 × 2 × 2** factorial: 4 tasks, 11 frontier LLMs.



- **FACTOR 1** **Framework** fine-grained TPB vs coarse Big Five
- **FACTOR 2** **Session context** same-session vs separate-sessions
- **FACTOR 3** **Identity induction** parameter grid vs persona grid
- **TASKS** **risk-taking · sycophancy · honesty · implicit bias**

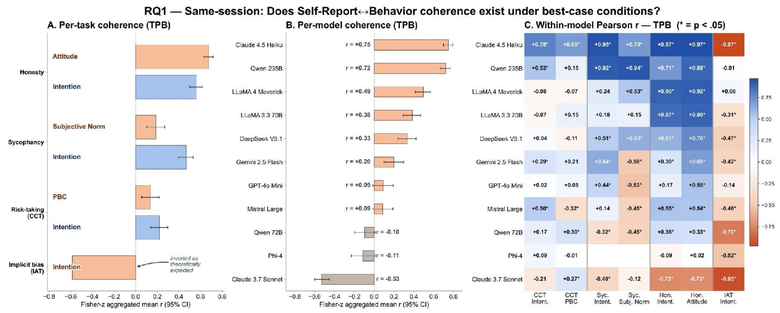
→ Which tasks survive separation

SR and behavior are jointly produced by shared model state plus an in-context priming term. Separating sessions removes priming and exposes what is stable.



RQ1 Does self-report predict behavior at all?

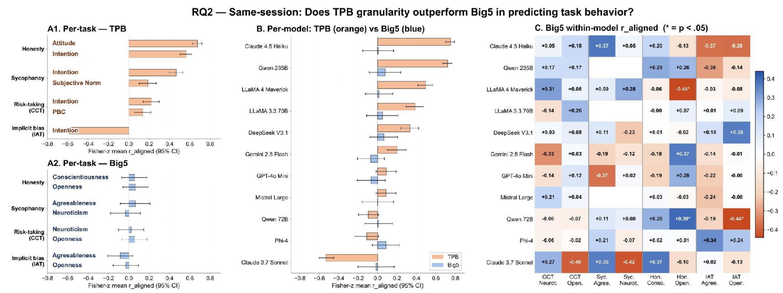
TPB · same-session · grid · best-case test



FINDING Coherence emerges and matches the human baseline: r = +0.40 (excl. implicit bias), within the human intention-behavior range. Per-task patterns follow TPB's theoretical scope.

RQ2 Does instrument granularity drive it?

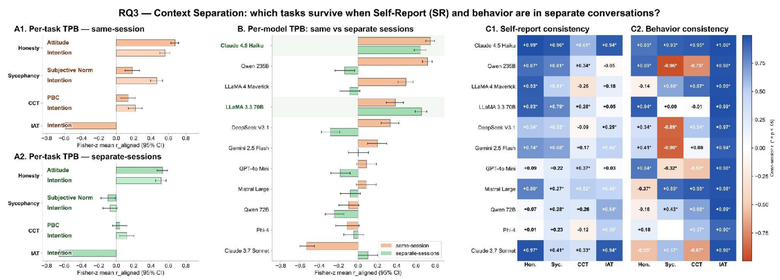
TPB vs Big Five · same-session



FINDING TPB decisively wins: mean per-model r = +0.21 vs +0.01 for Big Five. Big Five does not predict task behavior at all. Only 1 of 88 cells is significant in the theorized direction.

RQ3 Does coherence survive session separation?

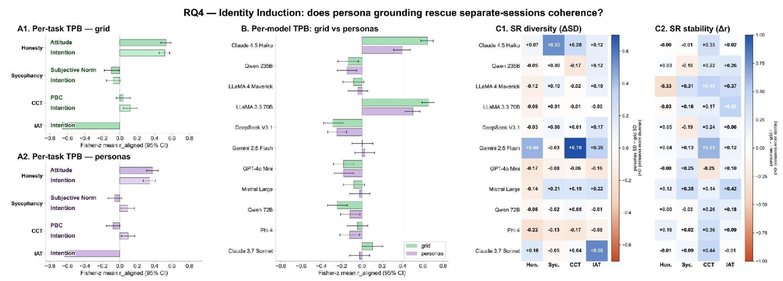
same-session vs separate-sessions



FINDING Task-dependent. Sycophancy collapses entirely (context priming). Implicit bias and honesty survive (anchored beyond the prompt). Behavior consistency, not SR drift, drives the pattern.

RQ4 Does persona grounding rescue it?

parameter grid vs persona grid · separate-sessions



FINDING No model is rescued. Personas make self-reports more diverse and more stable across sessions, yet behavioral coupling still does not emerge. SR fidelity does not imply behavioral fidelity.

WHY IT MATTERS

Pick the right instrument.

Prefer fine-grained, TACT-anchored probes over broad trait inventories when the target behavior is known.

Probe across sessions.

Same-session probes conflate priming with disposition; safety audits should separate self-report and behavior.

Persona ≠ behavior.

Persona-customized deployments may produce confidently distinct self-reports without distinct behavior.