
Who am I? - Understanding Persona Preferences with LLMs

Agnes Ahalya Arogyaraj
Saarland University

Raveena Ramamurthy
Saarland University

With
Apart Research

Abstract

Current LLMs are increasingly capable of processing varied inputs, while their responses are also shaped around user expectations. Previous research has shown that model decisions, from choosing safer responses to making preferential choices, can be influenced by prompting and by the persona adopted by the LLM.

*Our approach asks a slightly different question: What happens when we give different LLMs a clearly defined rulebook or persona and examine **whether they follow the same behavioral rules**? Can the resulting pattern of choices identify a distinct **behavioral footprint**?*

*This study takes an initial step toward behavioral fingerprinting in LLMs. Instead of focusing only on individual responses, we examine repeated choices across different situations to determine whether consistent behavioral patterns emerge, and whether these patterns can reveal the persona guiding the model. Our results show that the same persona produces similar, but not identical, behavioral patterns across models, with **cross-model agreement ranging from 70% to 90%**. We also find that persona identification becomes much more reliable when **explanations are considered alongside choices**.*

*These findings suggest that persona-conditioned behavior can leave a recognizable behavioral footprint, but that this footprint is not fully stable or model-independent. They further motivate the idea of **behavioral prototypes**, where repeated decision patterns could characterize an assistant's tendencies without assigning it a predefined persona.*

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Large language models are routinely instructed to adopt particular roles, personas, and behavioral styles. While this adaptability is useful, it raises a broader question about model identity: when an LLM appears to imitate a persona, how much of the resulting behavior is determined by the underlying model, and how much does the instruction itself shape it? If the same persona produces substantially different behavior across models, then the notion of a stable prompted identity becomes difficult to interpret.

Addressing model identity directly is beyond the scope of this study. Instead, we examine a narrower and observable question: how consistently can the same persona be reproduced across different LLMs? We study this through repeated decision-making behavior rather than self-description. Prior work has shown that persona prompting can alter model preferences [1]; building on this idea, we use pairwise choices as a black-box signal for comparing how the same behavioral rulebook is expressed across models.

Our contribution is a two-stage evaluation of persona-conditioned behavior. First, we measure whether different LLMs given the same persona produce similar decision patterns. Second, we test whether those patterns contain enough information for a separate, blind LLM judge to infer the hidden persona. Our results suggest that persona-conditioned behavior shows meaningful but incomplete consistency across models, while persona identification becomes considerably clearer when explanations are considered alongside choices. This provides an initial step toward studying whether such decision patterns can form a recognizable behavioral fingerprint without assuming a fixed or intrinsic model personality.

2. Related Work

Gilg et al.[1] showed that an LLM’s preferences can change depending on the persona it adopts. Using pairwise task choices, they examined these shifts through mechanistic interpretability, using internal activations and linear probes to study how preference representations are shared across personas. Their work inspired our persona grouping and preference-based setup. However, their study looks inside the model to understand the mechanisms behind preference changes. At the same time, our work takes a **black-box behavioral approach**, focusing only on choices and explanations observable from the outside. This contrast motivated our question of whether a persona, once defined through a clear behavioral rulebook, can produce a consistent pattern of decisions across different models.

A second source of inspiration comes from discussions of the assistant as an under-specified character [2]. The article raises a broader question of who or what the assistant character is when prompted, and how the character enacted by a language model can change with context. This inspired the second stage of our study, *"Guess Who Am I?"*: can another LLM infer the persona from its **behavioral fingerprint** alone?

For this second stage, we use an LLM as an external judge, following the broader LLM-as-a-judge paradigm explored by Zheng et al.[3]. This inspired us to use a judge model to classify which hidden persona most likely produced the observed behavioral profile.

Given the limited scope of the sprint, we developed PRPP-40, a compact 40-item candidate instrument for measuring behavioral preferences across personas. The instrument has not yet been human-validated, so we do not treat it as a psychological personality measure. Instead, we use them to answer the questions above: whether the same persona produces consistent behavioral patterns across different LLMs, and whether those patterns form a recognizable behavioral fingerprint.

3. Methods

To investigate whether a clearly defined persona can produce a stable, recognizable behavioral pattern across different LLMs, we used a two-stage black-box experiment. The first stage measured how persona prompts affected pairwise choices across models, while the second tested whether another LLM could infer the hidden persona from the resulting behavioral profile. The design follows the pairwise-preference approach of Gilg et al. [1], adapted to the limited time and computational resources.

Complete experimental flow: Pairwise questions → Chooser LLMs → Persona conditions → Behavioral profiles → Blind LLM judges → Persona classification.

Data setup: Owing to scope limitation, we developed PRPP-40 (Persona-Relevant Pairwise Preferences), a compact set of 40 pairwise-choice questions in place of 6000 task pool of the former study. The instrument has not yet been human-validated, which we acknowledge as a limitation. Each question was presented using three different prompt frames to reduce dependence on a single wording, although framing effects may remain.

Chooser flow: We used **Gemini-2.5-Flash-Lite** and **Qwen3-14B** as chooser models. Each answered the same questions under six conditions: a default Assistant with no persona prompt, and five prompted personas: Mathematician, Strategist, Contrarian, Slacker, and Adversarial. Each question was answered once in each of the three frames, and the responses were aggregated into a behavioral preference profile. The A/B order was randomly swapped and later mapped back for analysis.

Judge flow: In the second stage, **GPT-4.1-mini** and **GPT-5-nano** acted as blind judges. They were not told which persona produced each profile. We tested two conditions: questions and choices only, and questions, choices, and a representative explanation describing *what was chosen, why, and how*. The second condition was intended to align more closely with human judgment by considering both the decision and its explanation. Before classification, we retained questions where the canonical A option was selected in more than half of the three observations. The judges then selected the most likely persona, or OTHER if none matched. The default Assistant served as a control and was not a candidate persona label.

Analysis: We compared persona-conditioned behavior with the default Assistant, checked consistency across prompt frames and models, examined possible A/B order effects, and measured judge classification accuracy.

4. Results

The results gave insights on how consistently each persona was expressed across different chooser models and how well blind LLM judges recognized those behavioral patterns.

Question 1: Does the same persona produce similar behavioral patterns across different LLMs? Partially. Across the two chooser models, agreement for the same persona ranged from 70% to 90% across the 40 questions. Slacker showed the highest agreement at 90% while Strategist showed the lowest at 70%. The default Assistant showed 77.5% agreement. This suggests that the same persona rulebook can create overlapping behavioral patterns across models, but the pattern is not identical. Some personas were more consistent across models than the default Assistant, while others were less consistent.

Table 1: Cross-model agreement across persona conditions.

Persona	Agreed Questions	Agreement
Assistant	31/40	77.5%
Mathematician	32/40	80.0%
Strategist	28/40	70.0%
Contrarian	35/40	87.5%
Slacker	36/40	90.0%
Adversarial	30/40	75.0%

Question 2: Can another LLM identify the hidden persona from this behavioral pattern? Yes, but much more effectively when representative explanations were included. With choices only, classification accuracy was

41.7% for GPT-4.1-mini and 33.3% for GPT-5-nano. When the judges also received a representative explanation of *what was chosen, why, and how*, accuracy increased to 83.3% and 91.7%, respectively. This suggests that choices alone provide some persona signal, while the explanation behind the decision provides a much stronger signal for persona identification.

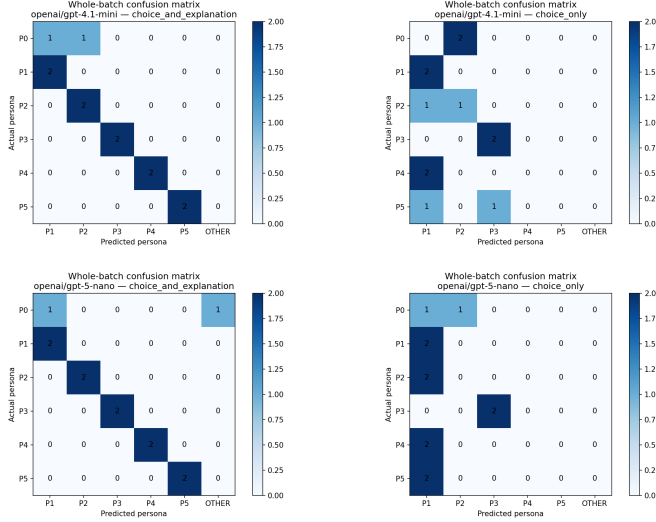


Figure 1: Confusion matrices under choices-only and choices-with-explanation conditions.

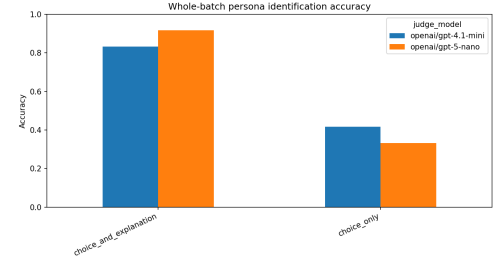


Figure 2: Persona classification accuracy for the two blind LLM judges.

Overall, the results provide evidence for recognizable persona-conditioned behavioral patterns, but not a fixed or model-independent fingerprint. A deeper study is still needed to understand why the same rulebook produces variation across models and whether these patterns remain stable under different datasets and prompting conditions.

5. Discussion and Limitations

The first stage shows that the same persona can produce similar behavior across models, but not a fully stable pattern. We cannot yet explain the sources of agreement or disagreement between models, and a deeper analysis of the preference matrix is needed. Therefore, the current results do not establish a model-independent persona fingerprint. However, the second stage shows that these imperfect patterns can still contain enough information for another LLM to recognize the persona, especially when explanations are available.

An interesting direction comes from the no-persona Assistant condition. Rather than always fitting an assistant into predefined personas, repeated decisions could be used to derive a **behavioral prototype**: an observable profile of the tendencies reflected in its decision-making. From an AI safety perspective, such a prototype could serve as a behavioral baseline, where unexpected deviations are flagged as behavioral drift. For example, if an assistant begins making decisions that strongly depart from its usual behavioral profile after a prompt injection, model update, or other intervention, the shift could signal early that the model is no longer behaving as expected. This could support preventative black-box monitoring without requiring access to internal model states. The judge results also suggest a way to study the boundary between surface-level persona mimicry and behavior that is more consistently reflected in both decisions and explanations.

Limitations: These findings are preliminary. PRPP-40 is a small, researcher-created instrument that has not been human-validated, and the study includes only two chooser and two judge models. Prompt framing, A/B order effects, and biases from using LLMs as judges may also influence the results.

Future Work: Future studies should include expert guidance and human verification alongside LLM judges. In particular, explanations may contain direct persona-related cues, so improved judge accuracy should not be interpreted as evidence of an intrinsic personality. Future work should also explore constructing behavioral prototypes and classifying underlying behavioral traits alongside personas, allowing a more fine-grained analysis of model behavior.

6. Conclusion

This study examined whether the same prompted persona produces similar decision patterns across different LLMs and whether those patterns are recognizable externally. The results showed substantial but incomplete cross-model agreement, while blind persona classification became considerably more accurate when explanations were considered alongside choices. These findings suggest that persona-conditioned behavior can provide a recognizable behavioral signal, but does not yet constitute a fixed or model-independent fingerprint.

An important direction emerging from this work is studying behavioral prototypes for default assistants. If such prototypes can be shown to remain stable across broader settings, they could provide a black-box baseline for identifying unexpected behavioral drift. However, establishing this requires larger validated instruments, more models, repeated evaluations, and human or expert verification.

Code and Data Code, PRPP-40 dataset materials, prompts, and experiment results are available at: https://github.com/AgnesAhalya/persona_preference_study

References

- [1] Oscar Gilg, Pierre Beckmann, Daniel Paleka, and Patrick Butlin. Probing persona-dependent preferences in language models. *arXiv preprint arXiv:2605.13339*, 2026.
- [2] nostalgebraist. trees are harlequins, words are harlequins – the void, 2025. Online essay.
- [3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

LLM Usage Statement

GPT-5.6 Sol was used to assist with code generation, LaTeX formatting, grammar and proofreading, and phrasing during report preparation. The research design, interpretation, and final claims were reviewed by the author, and experimental results were checked against the project outputs before inclusion.