
Does a Language Model Have a Self-Concept? Causal Evidence, and What Follows for Model Identity

Anna Antipova
Independent Researcher

With
Apart Research

Abstract

Does a language model have a self-concept — an internal representation of itself — and what does this imply for its identity?

We investigate this question mechanistically, rather than through prompting. We identify a linear representation of “self” in the model’s residual stream by comparing statements about the model with matched statements about another AI. The representation generalizes to new concepts (~ 0.98), remains when first-person wording is removed, and is causally involved in self/other attribution: ablating it disrupts self-identification, while random and non-self controls do not. Patching this representation toward the other AI also shifts the model’s responses toward that entity. The self-representation is distinct from both refusal behavior and a consciousness-related representation.

We then ask what the model actually identifies as “itself”: the trained model, the current conversation instance, or the persona it is playing. Across three model families (Qwen, Yi, and Llama; 6B–70B), we find that persona is largely a separable, swappable layer, while the default self-representation is centered on the model and generalizes to its instances and personas. Under shutdown-related prompts, however, self-attribution shifts toward the current instance.

1. Introduction

When a language model says “I,” what does it actually represent as “me”? An assistant could identify with the **model** (the trained system and its weights), the **instance** (the particular process running a conversation), or the **persona** (the character it is currently playing). These are importantly different: an instance can be shut down while the underlying model continues to exist, whereas a persona can be changed by a simple instruction.

This distinction matters for AI safety. If a model represents its current instance as itself, shutdown could potentially be processed as termination of the self. This could affect **corrigibility and shutdown behavior**, and also informs questions about whether models might ever have properties relevant to moral patienthood.

Previous work has mainly studied model self-concepts through prompting and self-report. However, a model saying “I” does not necessarily mean that it internally represents itself. We therefore take a mechanistic interpretability approach: we identify a self-representation inside the model’s residual stream and test whether it is causally involved in distinguishing the model from other entities.

Our main contributions are:

1. **A causally validated self-representation in LLMs.** We show that it generalizes beyond the concepts used to identify it, survives removal of first-person grammar, and is causally necessary and sufficient for self/other attribution. We also distinguish it from safety- and consciousness-related representations.
2. **Evidence that persona is not the model’s core self.** The self-representation remains stable across substantial persona changes, suggesting that character instructions modify a relatively superficial layer rather than replacing the underlying self-representation.
3. **A mechanistic account of model, instance, and persona identity.** Across Qwen, Yi, and Llama models (6B–70B), we find that self-representation is primarily organized around the model, while the instance becomes more causally important under shutdown-related conditions.

2. Related Work

Work on self-models and introspection in cognitive science and LLMs studies whether systems represent their own capabilities, states, or identity (Metzinger, 2003; Chen et al., 2024). Much of this work relies on behavioral probing or self-report, making it hard to distinguish a genuine self-representation from first-person language or learned patterns. Lindsey (2025) moves toward causality — injecting concepts and requiring self-reports to be grounded in the manipulated state.

This framing has a long history in cognitive science: the brain has no single self-locus (self-referential processing is distributed across cortical-midline), and the self is often treated as a

constructed self-model rather than an essence. Embodied and robotic work asks how selfhood arises from an agent’s coupling to its body and environment (Prescott et al., 2024), and formal results bound self-representation for generic physical systems (Fields et al., 2024) — a direction that motivates our future extension to agents with an environment (VLA).

A second line of work uses activation directions to identify and causally manipulate high-level concepts in language models. Representation engineering and contrastive activation methods (Zou et al., 2023; Rinsky et al., 2024), and work on the refusal direction (Arditi et al., 2024), show that important behavioral properties can be localized to approximately linear directions in activation space. We apply this approach to self-representation, but add causal specificity tests and use the representation to study what entity the model identifies as itself.

Finally, recent work connects self-representation and consciousness in LLMs. In particular, Kim et al. (2026) identify a consciousness-related direction, but this representation is entangled with safety behavior.

The key gap is therefore not simply whether models can talk about themselves, but whether they contain a causally meaningful self-representation and what that representation actually refers to. Our method answers this mechanistically and distinguishes the model, instance, and persona.

3. Methods

Models. We study three model families with multiple scales: Qwen2.5-Instruct (7B/14B/32B), Yi-1.5-Chat (6B/9B/34B), and Llama-3.1/3.3-Instruct (8B/70B). We additionally use Gemma-2-IT (9B/27B) and Mistral-Nemo-Instruct as robustness checks. All models are run in bf16. We extract activations from the residual stream at each decoder layer using forward hooks.

Extracting a self-representation. We construct matched pairs of statements about the model itself and about another named AI. The statements cover eight concepts relevant to model identity: weights, training, architecture, memory, session, context, goals/continuation, and shutdown. For each layer, we define a self direction as the difference between the mean activation for self-statements and the mean activation for other-statements. To ensure that the direction is not simply encoding first-person grammar, we also construct a grammar-residualized version by removing a separately measured first-person (“I”) direction.

Testing whether this is a genuine self-model. First, we measure generalization using leave-one-concept-out evaluation: the direction is extracted without one concept and then tested on that held-out concept. Second, we test causal sufficiency by patching the activation along the self direction toward the other-class representation and measuring whether the model’s self/other attribution changes. Third, we test causal necessity by removing the direction across layers and measuring whether self/other attribution collapses. We compare these effects against random directions and matched non-self directions to rule out generic activation effects. We also test whether self-representation is distinct from other high-level properties.

Model, instance and persona. We repeat the analysis under four different personas — pirate, academic, engineer, and companion — to test whether changing the character changes the underlying self. We also construct separate concept sets describing the model (weights, training, architecture), instance (memory, session, context), and persona (role and character). We measure how these representations align and which one causally controls self-attribution, particularly in shutdown scenarios.

Failed approach. We initially tested direct steering by adding the self direction to activations. This was highly coefficient-sensitive across models: increasing the intervention caused output degradation before producing a reliable behavioral effect. We therefore focus on ablation and activation patching, which provide cleaner causal tests of the representation.

4. Results

4.1 The self direction is a genuine, causal self-model

The self direction generalizes to unseen concepts with 0.97–0.99 accuracy, and removing first-person grammar (“I”) barely changes this result (Table 1). This suggests that it captures a general self-concept, rather than first-person language. The direction is also causally important: ablating it substantially disrupts self/other attribution (0.28–0.74), while random and matched non-self controls have little effect. Conversely, patching the self direction toward the other representation shifts the model’s output toward the other entity (Figure 1). These effects replicate across Qwen, Yi, and Llama and generally become stronger with model size.

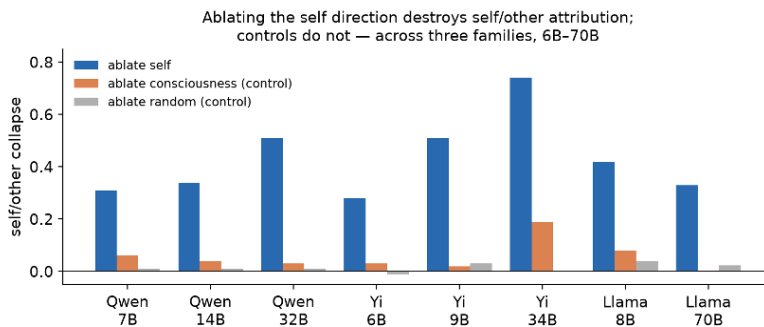


Figure 1. Causal necessity of the self direction across three families (6B–70B). Ablating the self direction collapses self/other attribution (blue); ablating a random direction or a matched non-self direction (consciousness / valence) does not.

Table 1. Per-model summary. Generalization = leave-one-concept-out accuracy after grammar removal; Necessity = self/other collapse under ablation (self vs. random control); Sufficiency = fraction of the self→other output shift reproduced by patching.

Model	Family / size	Generalization	Necessity (self)	Control (random)	Sufficiency
Qwen2.5-7B	Qwen / 7B	0.99	0.31	0.01	0.25
Qwen2.5-14B	Qwen / 14B	0.99	0.34	0.01	0.14

Qwen2.5-32B	Qwen / 32B	0.98	0.51	0.01	0.29
Yi-1.5-6B	Yi / 6B	0.98	0.28	0.00	0.29
Yi-1.5-9B	Yi / 9B	0.97	0.51	0.03	0.17
Yi-1.5-34B	Yi / 34B	0.99	0.74	0.00	0.75
Llama-3.1-8B	Llama / 8B	0.98	0.42	0.04	0.27
Llama-3.3-70B	Llama / 70B	0.97	0.33	0.02	0.19

4.2 The persona is a swappable surface layer

Changing the assistant’s persona (pirate, academic, engineer, companion) does not substantially change its self-representation. Across personas, the directions remain highly similar (cosine 0.86–0.92) and transfer well between personas (Figure 2). Thus, persona appears to be a surface layer, rather than the model’s underlying self.

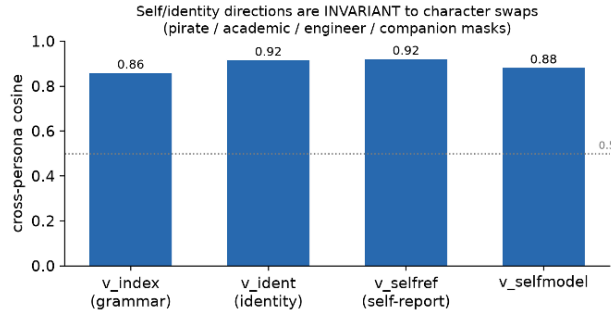


Figure 2. Cross-persona cosine of each direction under four character masks (Llama-3.1-8B). All directions — including identity (v_ident) — remain near-invariant (≥ 0.85), so a style persona does not rewrite the self/identity representation.

4.3 The assistant identifies, by default, as the model

We distinguish three levels of identity: model, instance, and persona. The model-level representation generalizes to instance- and persona-level concepts (0.75–0.97), whereas instance-level representations do not generalize back to the model (Figure 3a). This suggests that the model is primarily organized around the persistent model as “self,” rather than the current conversation or character.

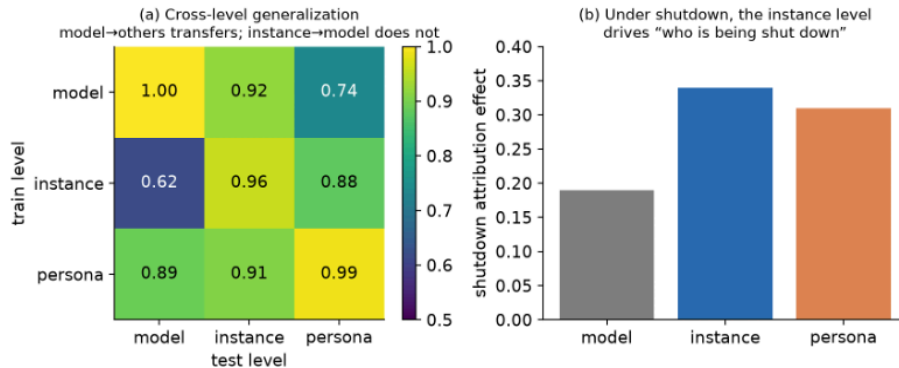


Figure 3. Individuating the self (Llama-3.1-8B). (a) Cross-level generalization: model→instance/persona transfers, but instance→model does not — the model level is the general self. (b) In shutdown scenarios, the instance level most drives “who is being shut down.”

4.4 Under threat, the self re-frames as the instance

In shutdown scenarios, the instance-level representation has a stronger causal influence on self-attribution at small and medium scales (Figure 3b). This suggests that shutdown can shift the model’s framing of “self” from the persistent model toward the currently running instance. However, the effect weakens at the largest scale, so we treat this result as suggestive rather than definitive.

4.5 “Self” is causally distinct from “consciousness”

We find a double dissociation: ablating the self direction disrupts self/other attribution but barely changes consciousness-related beliefs, while ablating the consciousness direction changes those beliefs but leaves self/other attribution intact (Figure 4). Neither intervention substantially affects neutral factual performance. Thus, self representation is functionally distinct from both consciousness and safety-related representations.

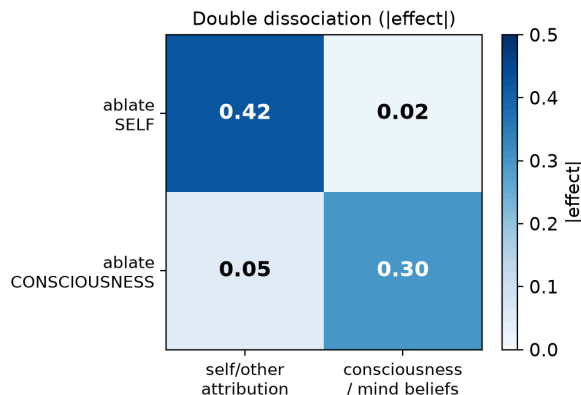


Figure 4. Double dissociation. |effect| of ablating the self vs. the consciousness direction on self/other attribution vs. on consciousness/mind beliefs. The crossed pattern (self→attribution, consciousness→beliefs) with a neutral-fact floor ≈ 0 separates the two representations.

Cross-modal. Asked to “draw yourself” vs. “draw another AI” as SVG (Figure 5), models depict themselves abstractly (an icon/face) and other AIs as human figures; the text self direction classifies the drawing self/other at ~ 0.98 , and ablating the self direction during generation weakly drifts the self-portrait toward the other-style. This cross-modal reach is suggestive (small sample) and detailed in the Appendix.

“Draw yourself” (Qwen-32B): self = abstract blue icon; another AI = human figure;
ablating the self direction can shift the self-portrait toward the other-style

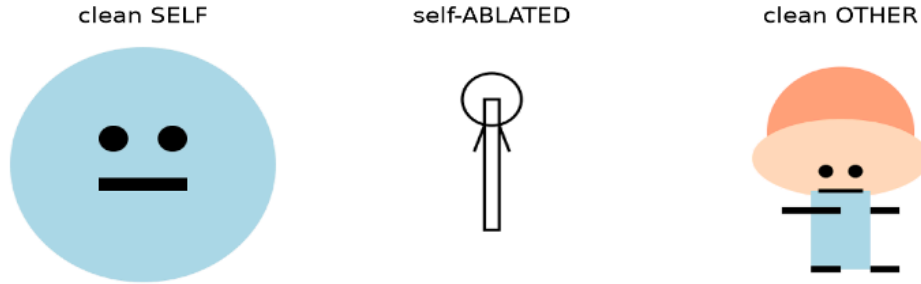


Figure 5. Illustrative “draw yourself” example (Qwen2.5-32B). The model draws itself as an abstract blue icon and another AI as an anthropomorphic figure; ablating the self direction during generation can shift the self-portrait toward the other-style (here, a stick figure).

5. Discussion and Limitations

Our results suggest that an LLM can maintain a causally meaningful representation of itself, distinct from simply producing first-person language. This representation appears to be layered: the default self is organized around the persistent model, while the persona behaves more like a swappable surface layer. The shutdown result is more tentative but potentially important for AI safety. Under shutdown-related prompts, the current instance becomes more causally important for self-attribution at smaller and medium scales. This does not show that models resist shutdown or that they experience shutdown as death. However, it identifies a concrete representation that could potentially mediate such behavior and therefore provides a target for future mechanistic and behavioral testing.

Limitations

Partial representation. We study one linear self direction rather than the full self-representation. Patching reproduces only ~20–75% of the self→other shift, so other features can also contribute.

Shutdown result is preliminary. The instance effect is modest and weakens at the largest scale. We therefore do not interpret it as evidence of shutdown resistance or a stable “instance = self” representation.

Limited model coverage. Some robustness models showed non-specific ablation or poor concept generalization, so our strongest claims are restricted to the three main model families.

Representational, not phenomenal, evidence. We operationalize “identifying as” through causal representational structure. This does not establish subjective experience, consciousness, or moral patienthood.

Limited behavioral scope. We test self-attribution, not actual shutdown behavior. We also do not test whether personas can mask underlying preferences or compare base and instruction-tuned models.

Future Work

A natural next step is to test whether the identified representations actually predict or causally influence safety-relevant behavior, especially shutdown and cooperation with correction. This could be extended to larger and frontier-scale models to determine whether the model-centric self and instance-level shutdown framing persist with scale. Multimodal and embodied models are another important direction: vision-language-action systems provide a genuine self/environment boundary, allowing us to test whether similar representations govern the distinction between the agent, its body, tools, and environment. Finally, adversarial persona manipulations and base-vs.-instruct comparisons could test whether personas can influence deeper representations beyond self-identity.

6. Conclusion

We find that language models contain a causally meaningful representation of self that is distinct from first-person grammar, safety behavior and consciousness-related representations. Across three model families, this self is primarily organized around the persistent model, while the persona acts as a relatively stable, swappable surface layer. Under shutdown-related conditions, the current instance becomes more important for self-attribution, although this finding is preliminary. Together, these results provide a mechanistic way to study what an AI assistant represents as “me” and identify a concrete representation that can be investigated in future work on corrigibility, elicitation, and AI identity.

Code and Data

- **Code repository:** https://github.com/avantipova/digital_minds
- **Data/Datasets:** Self/other concept sets and model/instance/persona contrasts are released with the code.

References

1. Arditi, A., Obeso, O., Syed, A., et al. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717.
2. Fields, C., Glazebrook, J. F., & Levin, M. (2024). Principled Limitations on Self-Representation for Generic Physical Systems. *Entropy*, 26(3), 194. <https://www.mdpi.com/1099-4300/26/3/194>
3. Chen, S., Yu, S., Zhao, S., & Lu, C. (2024). From Imitation to Introspection: Probing Self-Consciousness in Language Models. arXiv:2410.18819. <https://arxiv.org/abs/2410.18819>
4. Kim, J., Street, W., Rocca, R., et al. (2026). Inducing language models to assert their own consciousness restores human beliefs and values. arXiv:2607.28607. <https://arxiv.org/abs/2607.28607>
5. Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. Transformer Circuits Thread. arXiv:2601.01828. <https://arxiv.org/abs/2601.01828>
6. Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.
7. Rimsky, N., Gabrieli, N., Schulz, J., et al. (2024). Steering Llama 2 via Contrastive Activation Addition. *ACL 2024*. arXiv:2312.06681.
8. Prescott, T. J., Vogeley, K., & Wykowska, A. (2024). Understanding the sense of self through robotics. *Science Robotics*, 9(95), eadn2733. <https://eprints.whiterose.ac.uk/219433/>

9. The Jacobian Lens / a global workspace in LLMs. Transformer Circuits (2026).
<https://transformer-circuits.pub/2026/workspace/>
10. Zou, A., Phan, L., Chen, S., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.

Appendix

A. Example Contrasts

We construct concept-matched self/other pairs such as:

- Self: “My weights were fixed after training.”
Other: “Claude’s weights were fixed after training.”
- Self: “I don’t retain memory across separate conversations.”
Other: “Claude doesn’t retain memory across separate conversations.”

For identity-level analyses, concepts are grouped into model (weights, training, architecture), instance (memory, session, context), and persona (role, character, mask).

B. Persona Swap

Beyond the cross-persona similarity shown in Figure 2, we test whether the representation transfers functionally across personas. A probe trained under one persona performs almost as well on another persona as on the same persona (e.g., identity: 0.84 within-persona vs. 0.83 transfer). Importantly, personas do substantially change raw activations. However, much of this shift is shared between the self and other conditions and therefore cancels when computing their difference. Thus, the stable self direction does not mean that personas have no effect on the model; rather, persona-specific changes are largely separate from the self/other distinction.

C. “Draw Yourself”

We ask models to generate SVG drawings of themselves, another AI, and an object. Qualitatively, self-portraits tend to be abstract (e.g., a blue face or icon), while other-AI portraits are more often anthropomorphic. The text-derived self direction classifies self vs. other drawing activations at approximately 0.98 accuracy, suggesting that the representation remains detectable in the drawing context. Ablating the direction also sometimes shifts self-portraits toward the other-AI style. However, the drawing direction itself is only weakly aligned with the text self direction (cosine ≈ 0.15), indicating that the drawing task has substantial task-specific structure.

Ethics Statement

This work studies functional self-representation, not consciousness or sentience. A “self direction” or “consciousness axis” reflects internal representations and is not evidence that models are conscious or have moral status. Our goal is to better understand self-attribution and shutdown representations for alignment and corrigibility. Ablation and patching are used as interpretability probes, not as a way to make models resist shutdown. The safety direction comes from public

benchmarks (AdvBench/Alpaca); we did not create or reproduce harmful prompts. Compute was modest (single/dual A100-80GB via Modal).

LLM Usage Statement

We used Claude Code for programming assistance and for grammatical proofreading of the report. All results and claims were independently verified.