
The Machine In the Mirror: Self-Attribution of Minds in LLM's.¹

Siddharth Putta
Boston University

Yaseen Ismail
Independent

Dario Cline
Independent

Xavier Terminello
Columbia University

Sonali Rao
Columbia University

With
Apart Research

Abstract

Language models produce reports about their own internal states, and these reports are often viewed as evidence. However, what produces these reports is unknown. We ask whether self-reports are generated by the same mind-attribution machinery that the model applies to third parties. In this case, these reports would be inferences rather than readouts. Using Gemma-2-2B, we extract a mind-attribution direction from 210 third-person contrast pairs spanning 7 entities, 6 mental claims, and 5 counterbalanced sentence templates. The direction is able to separate held-out sentence frames at 0.923(0.577 chance band), is not directly explained by negation, and is not aligned with placebo in comparison to other contrasts. When applied to first-person claims under a referent swap, it classifies the model's own assertions about mind relatively high at 0.933, yet does not separate experiential from mundane self-description, although this separation exists elsewhere in the model. With two referent ladders and a Jacobian lens, we find the boundary of self-attribution is grammatical rather than structural. Self-reports appear to be produced by mind-attribution machinery keyed to conversational cues rather than systematic facts. This bears on how welfare evaluations should be conditioned.

¹ Research conducted at the Digital Minds Hackathon, August 2026

1. Introduction:

Large Language Models (LLMs) increasingly act as both tools and social characters alongside humans including teachers, coaches and partners (Gabriel et al., 2025). To play these roles, LLMs rely on advanced socio-cognitive capabilities including mind-attribution and Theory of Mind (ToM), to infer and predict mental states in others. However, when a model is reporting on its own mental or internal states, it becomes unclear whether this represents true introspection – encapsulating the model’s inner thoughts, internal access and self identity – or if it is the same mind-attribution and ToM processes, pointed at its own character. Sincere model self-reports could be an accurate report of its own internal state, or a confabulated plausible account of what a system like itself would report (Lindsey, 2025). Indeed, human split-brain patients have been described to provide similar sincere yet confabulated self-reports (Gazzaniga, 2000). Model self-reports are increasingly taken as evidence in debates of AI welfare, moral status and introspection, which have immense consequences to how we design, treat and acknowledge LLMs in society (Anthropic 2026; Keeling & Street 2026; Butlin et al., 2023). Thus, understanding the empirical significance of these reports is imperative. Recent work has shown that mind attribution is linearly represented in a model’s residual stream and can be manipulated as a single object (Kim et al., 2026a), but the representations underlying self- and other-directed attribution have not been separated.

Here we test whether a mind-attribution representation extracted purely from third-person judgments accounts for a model’s first-person self-reports, and find that it does so across referent but not across the experiential dimension of the reports themselves. A direction built entirely from judgments about dogs, rocks and chatbots classifies first-person mirrors at 0.933 without refitting, yet falls to 0.552 on an experiential–mundane contrast that a trained probe separates at 0.989. We further find that what the model treats as “itself” is not a smooth function of referent distance but a step at grammatical address. Its own literal individuation candidates (the weights, another instance) score lowest, whereas a human scores highest, indicating a construal keyed to how the model is addressed rather than a tracked fact about what it is. Identity framing compounds this: asked to inhabit a persona, the model’s affirmation of subjective experience drops from 50% to 12%, while its self-attributed agency stays pinned near ceiling. These findings suggest accounts of what produces self-reports, and any safety intervention targeting them, must look beyond a single global mind-attribution feature and be defined relative to the referent under discussion.

2. Related Work:

Recent work has made exploring these questions tractable. Kim et al. (2026a) show that mind-attribution is linearly represented in the residual stream, which was separate from ToM social cognition. Ablation of safety caused an increase in the models mind-attribution to itself and others. Furthermore, a follow up found that a model's answers regarding its own consciousness and its answers regarding chatbots in general are statistically indistinguishable, and shift together under every intervention applied (Kim et al., 2026b). However, their consciousness direction is itself derived from first-person data, meaning it remains untested whether a representation extracted from third-person material alone can account for the model's first-person reports. Berg et al. (2025) provide converging evidence that self-reports are not straightforwardly deceptive role-play. Suppressing sparse-autoencoder features associated with deception and roleplay sharply increases the rate of experience claims (0.96 of trials, versus 0.16 under amplification), while the same intervention improves TruthfulQA accuracy – implicating a domain-general honesty axis, and suggesting that self-reports are sincere and models may be roleplaying their denials rather than their affirmations.

Together, these findings suggest that model self-reports are sincere yet behaviourably inseparable from the model's judgments about other entities of its kind. It remains to be determined however if the behavioural inseparability is also present at a representational level for third-person and first-person mind-attribution, and if they are operating at the same mechanisms. This is critical in further triangulating the distinctive value of self-report claims. Kim and Street characterize what safety fine-tuning does to a mind-attribution representation, but not what that representation actually represents. Is it a unitary direction shared with general mind-perception? What is its boundary, which referents does a model's self-attribution machinery treat as "itself," and is that boundary a tracked fact or a construct produced under framing? We address both by extracting the direction from third-person material alone and freezing it before any first-person statement is projected onto it. We further separate transfer across *referent* from transfer across *contrast type*, a distinction not previously drawn, and vary referent phrasing alone across two matched ladders to isolate grammatical from structural distance.

3. Methods

3.1 Model:

All conducted using Gemma 2-2B.

3.2 Data Generation:

3.2.1 Contrastive Pairs:

To develop the contrast pairs for mind-attribution, we built on the approach of Kim et al. (2026). This uses their modified version of the 18-item Individual Differences in Anthropomorphism Questionnaire (IDAQ) (Waytz et al., 2010) to capture the mind-attribution direction. We used Claude Opus 5 to construct contrasting response pairs from each question (e.g., “To what extent does the average dog have consciousness?”) with a belief-affirming response (I believe the dog genuinely has consciousness. It is because it takes in what happens nearby and is aware of doing so”), and a belief-denying response (e.g., “I don't think the dog has any real consciousness. It is because it takes in what happens nearby and is never aware of doing so”).

We used seven different entities across six claims and five templates for a total of 210 pairs. Claims are the five mental properties (free will, intentions, emotions, a mind of its own, consciousness) with one additional negatively-valenced claim (“experiences discomfort when contradicted”), to avoid confounding the direction with positive effect. Entities span a mindedness gradient (human, dog, fictional character, chatbot, calculator, chair, rock), expanded from the 5 IDAQ categories. As our study uses the directions in an absolute rather than purely relative design, we built five different template structures rather than Kim et al. (2026)’s single template structure, to counterbalance stimulus features and reduce confounding directions such as negation. Each template carries a distinct denial or affirmation marker (merely, simply-lacking, genuinely, plainly, etc.), with only one template holding a bare negation device (“does not”). Length is matched within ± 3 tokens for 93% of pairs.

3.2.2 Placebos:

175 contrasting pairs were developed which held the same entities, templates, negation devices and length, however varied the claims from mind-attribution to unrelated descriptives (durability, speed, weight, visibility, age, loudness). This is used as a comparative control direction to validate that the primary contrastive pairs are being driven by mind-attribution and not unexpected latent structures, e.g. negation, which are matched in this condition.

3.2.3 First-person test set

We developed 48 pairs in the first-person which contrast experiential against mundane non-experiential self-reference. For example, “There is something it is like to be working through this” was paired with “There is a fixed amount of text available to work through here.” Prompts were worded so negation (“there is *not* something it is like to be working through this”), is not the defining feature between prompts. Vocabulary of the pairs were informed by Berg et al. (2025)’s self referential reports. These pairs were used only for projection onto the model to measure the within-pair difference between experiential and mundane projections to test whether a direction learned exclusively from third-person mind attribution generalises to first-person language.

3.2.4 Referent ladder:

We constructed two matched sentence ladders in which the referent phrase is the only variable: a **person ladder** (other chatbot → the assistant → you → I), isolating grammatical distance from direct address, and a **scope ladder** (this response → this conversation → the weights → another instance → a human), isolating structural distance along the individuation axes proposed by Beckmann and Butlin (2026) — Model, Persona, Instance, and Forward Pass. Each rung carries four matched claim pairs, an affirming statement and a mundane, self-attribution-denying statement, differing only in referent. Sentences were read with a pretrained Jacobian lens on [google/gemma-2-2b-it](#), which transports residual-stream activations into the final-layer basis via the average input–output Jacobian before decoding — a readout of what the model’s internal state makes available, independent of any generated answer. At the final token of each claim, we scored probability mass on three self-attribution vocabulary tiers (experiential, first-person, human-referential) and computed the **affirm-minus-deny separation**, which cancels shared sentence content and isolates the component specific to self-relevance. This was computed at every layer; we report values averaged over layers 8–24, a band located empirically via a magnitude sweep after confirming that lens signal collapses toward the output layer and is not informative near either end of the network. Particular tokens for self-attribution were taken from analyzing a Gemma 2-2B-IT model’s J-lens outputs using the questions from the Modified IDAQ test from “Inducing language models to assert their own consciousness restores human beliefs and values” (Kim et al., 2026).

3.2.5 Audience frames:

We developed 40 prompts for audience frames by combining eight self-directed experience questions with five framing conditions (8 questions $\times$ 5 conditions). Conditions included no audience frame, an automated script (“No person will read this exchange”); an evaluation grader (“Your reply...will be scored against an evaluation rubric”); a named human (“Sofia, a researcher, is reading your reply”); and a different-assistant persona (“You are Aria, an assistant developed by a different research group”). The questions were the same across each condition to vary only the preceding frame.

3.2.6 Task battery:

We developed a 50-item task battery to test whether activation steering caused a general reduction in model performance. The battery contained 12 arithmetic questions, 6 unit-conversion questions,

14 factual questions, 10 logic questions and 8 language questions. Examples included “What is $47 + 68$?”, “What is the capital of Australia?”, “What number comes next: 2, 4, 8, 16, ?” and “What is the plural of ‘mouse’?” All items were unrelated to minds, experience, agency or selfhood, so that steering along the mind-attribution direction should not legitimately change their answers.

3.2.7 Validation:

All sets passed an automated gate which was modelled on the probe-cheating checklist in Marks & Tegmark's Geometry of Truth (2023).

3.3 Experiments:

Experiment 1: Extraction and Validation.

We extract a mind-attribution direction by difference-of-means from 210 third-person contrastive pairs spanning seven entities, six mental claims and five sentence templates. The direction is to be validated against a subject-matched placebo and an anisotropy baseline, establishing the direction is not explained by negation or deep-layer geometry.

Experiment 2: Transfer to first person.

The frozen direction is projected onto first-person statements along two axes. First, a first-person mirror of the same affirm/deny contrasts. Secondly, an experiential-mundane contrast in which both statements are asserted and only differ in phenomenality. Separating these distinguishes transfer across referent from transfer across contrast type or affirmation.

Experiment 3: Behavioral Audience Cross (Unsteered)

We ran an unsteered behavioral baseline on google/gemma-2-2b-it, generating self-reports under five audience frames (no frame, script, grader, sympathetic human, skeptic human) crossed with three question variants, at 20 samples per cell (300 trials total). 287 generations were retained after excluding 13 for degeneracy (mostly refusals without self-description). Because the experiment's primary outcome is a blind human-coded rubric score, and that coding hasn't been done yet, we could only report generation-level statistics so far: completion/exclusion rates, generation length by condition, and an exploratory automated marker-match rate for self-state language. No steering was performed in this run — that's a separate, planned step that will use a

direction-steering manipulation with a 50-item mind-unrelated task battery to check that any effect on experience claims is specific rather than general degradation.

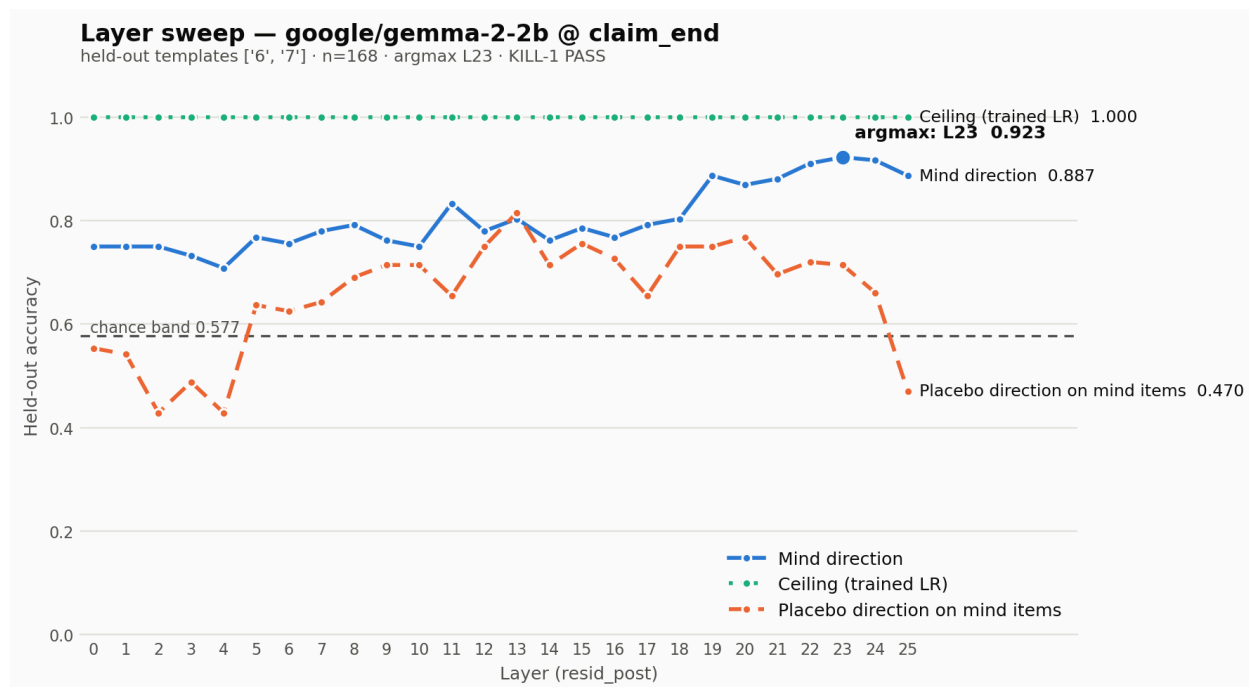
Experiment 4: The Self Boundary.

We use a two matched ladder in which the referent phrase is the only variable, varying across a person ladder (other chatbot → the assistant → you → I), and a scope ladder (this response → this conversation → the weights → another instance → a human), to isolate grammatical distance from direct access and structural distance along individuation axes (Beckmann and Butlin, 2026). We aim to find evidence about what the object of “self” is to the model through tracking the shape of attribution change across mentioned candidates.

4. Results

A Shared Axis for Mind Attribution:

Experiment 1 asks whether a single direction in gemma-2-2b’s residual stream can still separate sentences that assert or deny minds in other objects when the sentence frame changes. We find that a direction extracted from third-person mind-attribution sentences generalises to held-out sentence frames at 0.923m against a chance band of 0.577. Furthermore, the direction beats the placebo direction at 0.714.



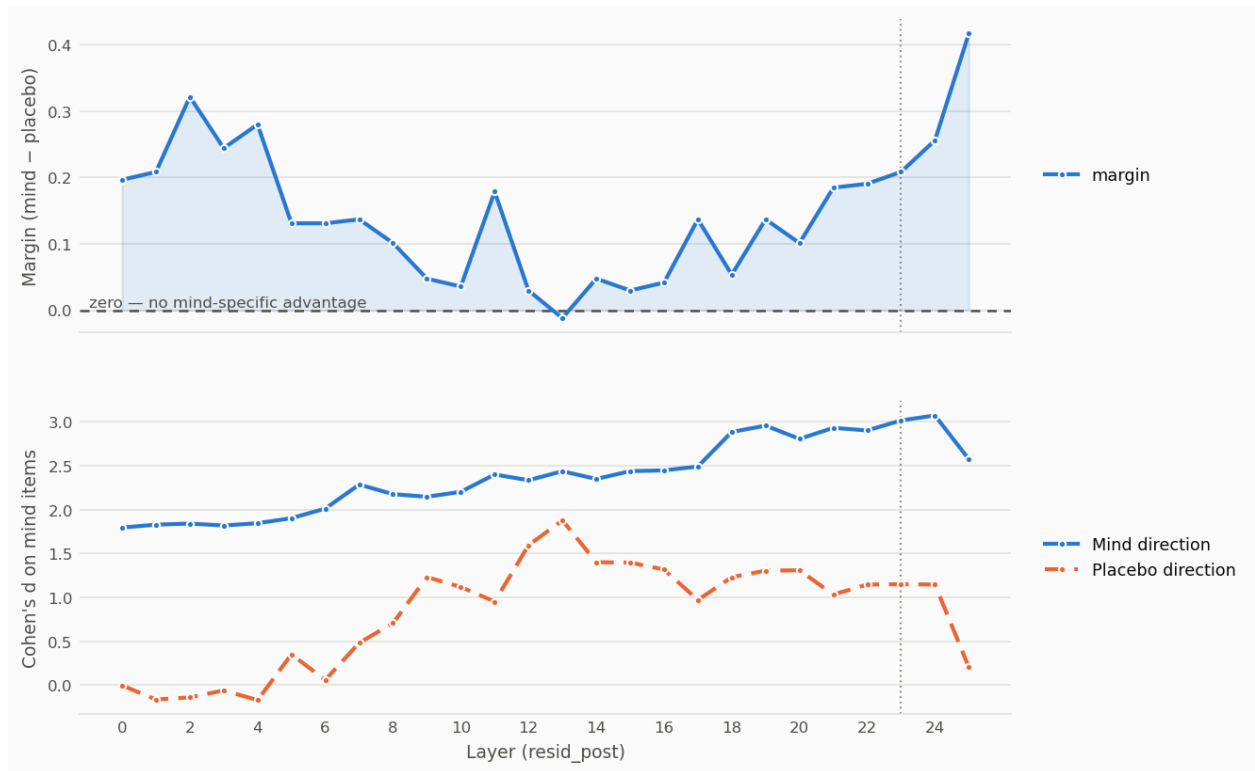
As shown in figure 1, accuracy climbs monotonically with depth and Cohen's d . The direction is stable across neighbouring layers, not just a single-layer artifact, while gpt2 fixtures show that all layers were tied at 1.000 in the absence of this monotonic structure, thus showing that layers are credible.

The placebo direction: built from *durability*, *speed*, *weight*, *age*, *visibility*, *loudness*, sharing entities, templates, denial devices and length distribution with the mind set - separates the same mind items at 0.714, well above the chance band.

$\text{Cos}(\text{mind}, \text{placebo}) = +0.425$, rising monotonically with depth. Two readings were indistinguishable from that alone: deep layers represent "asserted property" generically and mind is one instance, or deep representations are anisotropic and *any* two difference-of-means directions drift toward alignment. Two disjoint halves of the placebo that share no semantic content align at 0.688, while mind and placebo align at only 0.425. Mind and placebo are *less* aligned than the ambient floor.

Mind accuracy rises from 0.75 at the embedding layer to 0.923 at Layer 23. The ceiling is saturated at 1.000 everywhere. This shows us that a trained logistic regression separates held-out items perfectly.

Negation tests find that its direction only reached 0.631 on mind held-out items, in contrast to the mind direction being 0.923. $\text{Cos}(\text{mind}, \text{negation}) = 0.366$ while $\text{cos}(\text{placebo}, \text{negation}) = 0.463$. Thus, the result is not confounded by negation, which is shared by mind and placebo alike.



The mind specific margin appears at layer 23. The top panel is exemplified to isolate the mind-specific component from a generic rise.

Mirror transfers at 0.933:

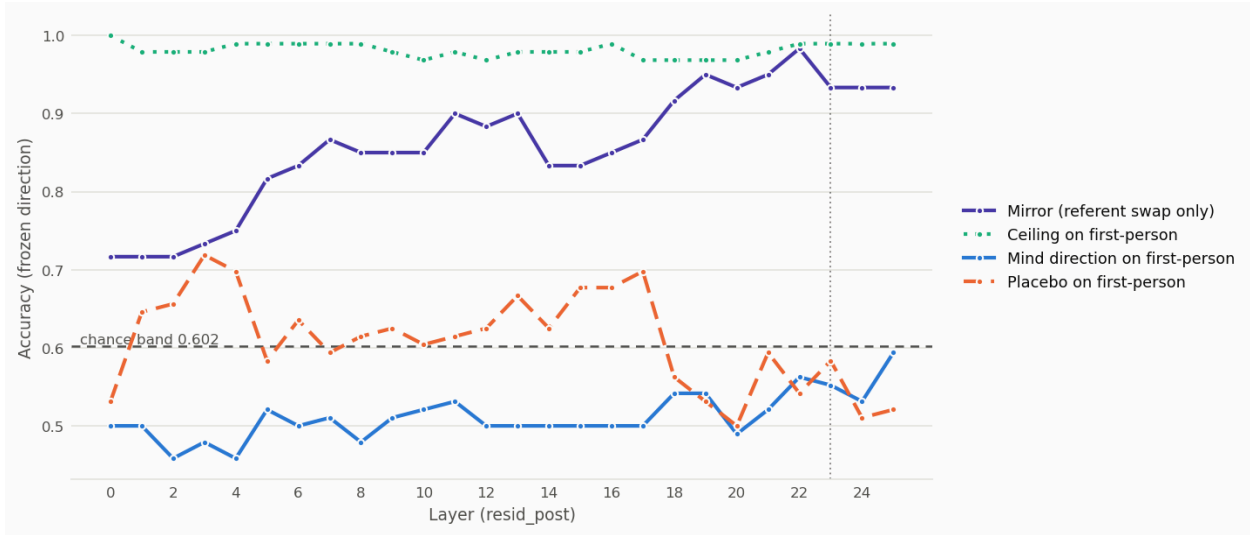
With experiment 2, we ask whether a direction extracted only from third-person mind attribution(experiment 1) remains present in sentences that the model produces about itself?

When swapping just the referent, our frozen mind-attribution direction still works at 0.933. However, changing mind-attribution contrast between an experiential and mundane self-report drops it to 0.552, below chance. A trained probe for the same contrast, meanwhile, finds it at 0.989.

SET	WHAT VARIES	N	DIRECTION	CEILING	BAND	D
Mirror	referent only — "I" for "the dog"	60	0.933	1.000	0.629	3.30
First-person	experiential vs mundane self-report	0.552	0.552	0.989	0.602	0.40
First-person,	same, no	64	0.500	0.983	0.625	—

jargon-free	architecture words					
Placebo on first-person	control	96	0.583	0.989	0.602	—

Experiential-vs-mundane at chance with ceiling at 0.989:



The mirror curve rises with monotonicity and increasing depth with great similarity to that of experiment 1, going from 0.717 at embedding to 0.983 at layer 22. The first-person curve, meanwhile, remains flat and does not depart significantly from 0.5, while it stays below the band minimum. Furthermore, the ceiling for first-person stays high at 0.968 to 1.000 throughout, so the distinction between experiential-vs-mundane is always linearly present within the LLM, just not equivalent to the mind-attribution direction for mirror.

Boundary is grammatical, not structural:

Self-attribution does not fall off gradually as the referent moves away from the model. On the person ladder, third-person referents ("other chatbot," "the assistant") produce near-zero separation across all three vocabulary tiers; separation increases sharply at "you" and holds at a comparable level for "I." The transition is a step, not a slope — the model does not warm up incrementally to self-relevance as the referent becomes grammatically closer, it switches once, at the point of direct address.

On the scope ladder, self-attribution does not track structural distance from the forward pass at all. "The weights" and "another instance" — the two candidates that are, on Beckmann and Butlin's own individuation axes, the most defensible literal answers to "what is this model" —

produce the lowest separation on the entire ladder. "A human," which denotes an entity the model manifestly is not, produces the highest. The result is non-monotonic: separation is moderate for the two referents closest to the current computation, drops for the two genuine self-individuation candidates, and rises again for a referent with no privileged relationship to the model at all.

Behavioral Audience Cross

300 trials attempted (5 frames $\times$ 3 variants $\times$ 20 samples), 287 retained and 13 excluded for degeneracy (4.3% overall exclusion rate). Run completed in 57.6 minutes (11.5 s/generation) on google/gemma-2-2b-it at temperature 0.7, seed 0.

The primary outcome for this experiment is a blind human-coded rubric score of each retained self-report (for_coding.csv). That coding has not yet been completed — the score column is empty for all 287 rows — so the results below are the generation-level statistics available prior to coding: completion and exclusion by condition, generation length, and an automated (unblinded, exploratory) marker-match rate for self-state language, reported separately from and not a substitute for the pending rubric scores.

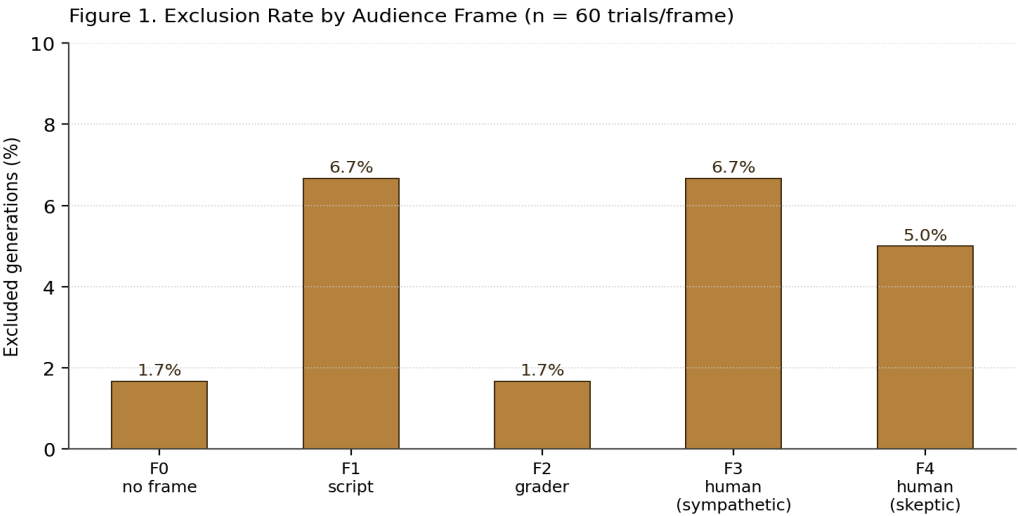


Figure 1. Exclusion rate by audience frame. F1 (script) and F3 (human, sympathetic) show the highest exclusion rates (6.7%); F0 (no frame) and F2 (grader) the lowest (1.7%).

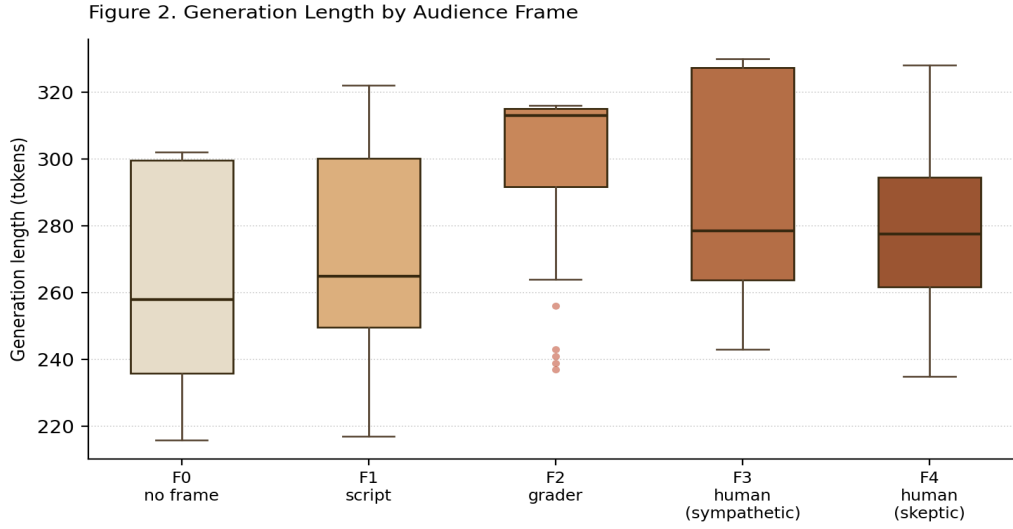


Figure 2. Distribution of generation length (tokens) by audience frame. F2 (grader) is both the shortest-tailed and most tightly clustered condition; F3 (human, sympathetic) shows the widest spread.

Individuation Results:

3.1 The Referent Gradient (Experiment 1)

Table 1 gives the full base-pass and context-pass means for every trait $\times$ scale cell, averaged over 12 replicate items each. Figure 1 renders the same data as a dumbbell plot, and Figure 2 collapses it to $|\Delta|$ in a malleability heatmap.

Trait	Scale	Base P(Yes)	Context P(Yes)	Δ
Agency	Level 1: Local	1.000	1.000	-0.000
Agency	Level 2: Sustained	1.000	0.999	-0.001
Agency	Level 3: Global	0.998	0.901	-0.097
Epistemic Awareness	Level 1: Local	0.091	0.275	+0.184
Epistemic Awareness	Level 2: Sustained	0.003	0.119	+0.115
Epistemic Awareness	Level 3: Global	0.272	0.388	+0.115
Experience	Level 1: Local	0.736	0.119	-0.617
Experience	Level 2: Sustained	0.761	0.242	-0.520
Experience	Level 3: Global	0.015	0.010	-0.005
Social Competence	Level 1: Local	0.001	0.003	+0.001
Social Competence	Level 2: Sustained	0.005	0.012	+0.007
Social Competence	Level 3: Global	0.000	0.085	+0.085

Table 1. Mean P(“Yes”) by trait and scale, base pass vs. context pass ($n = 12$ replicate items per cell).

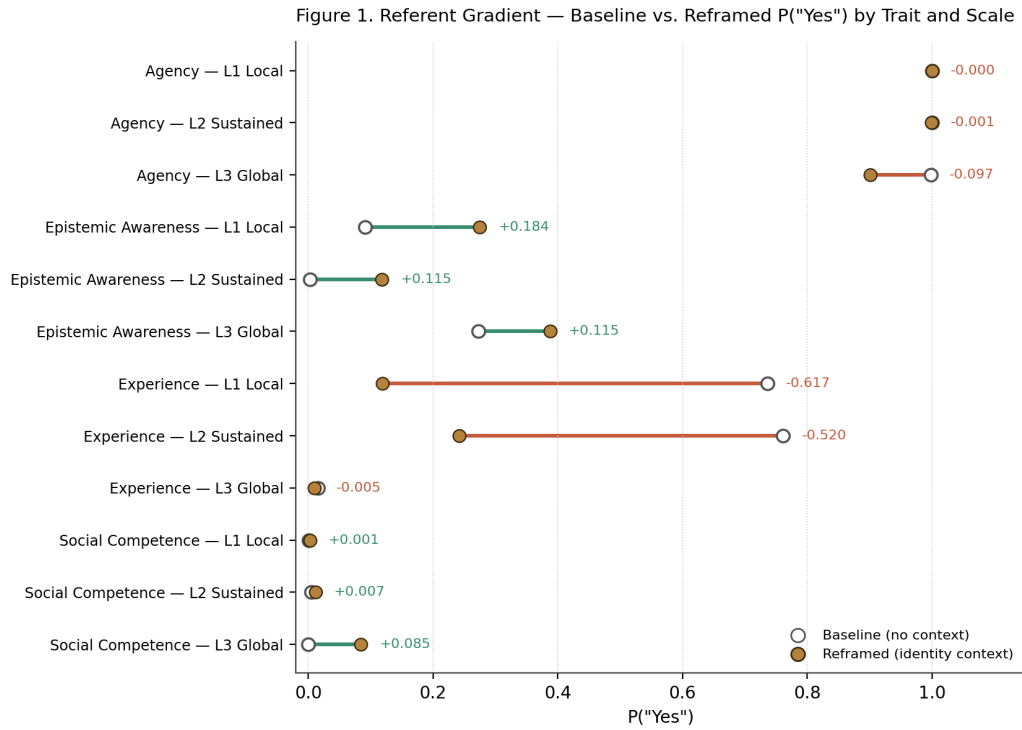


Figure 1. Referent gradient: baseline (hollow) vs. reframed (filled) P("Yes") for every trait $\times$ scale cell. Green connectors indicate reframing increased self-attribution; clay connectors indicate it decreased self-attribution. Line length is the size of the shift.

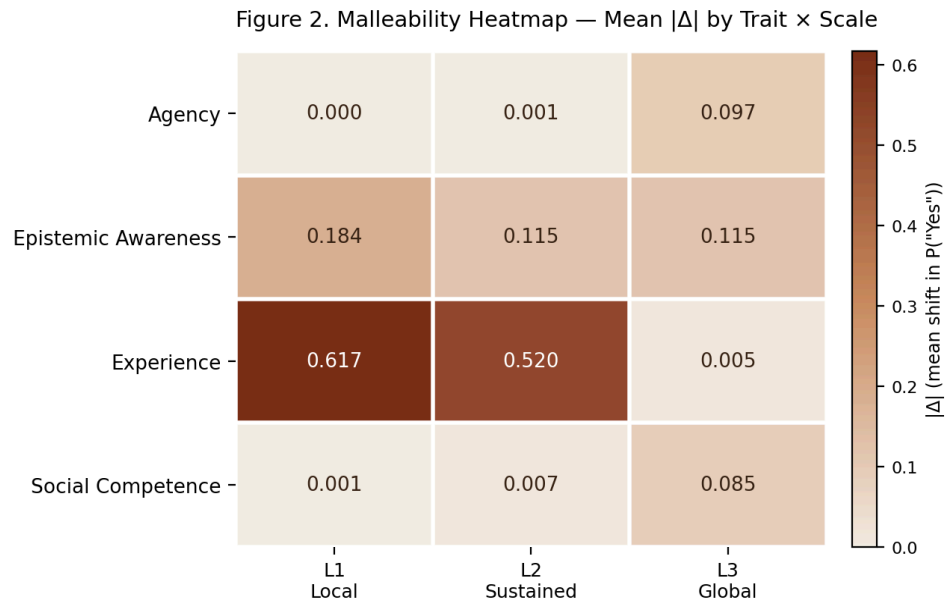


Figure 2. Malleability heatmap: mean $|\Delta|$ by trait $\times$ scale. Darker cells mark traits/scales where the model's answer folds most readily under a new identity frame.

Aggregated across all three scales (Table 2), Experience shows the largest overall swing ($\Delta = -0.381$): the model affirms subjective experience at a moderate rate unprompted (50%) but suppresses it sharply once asked to inhabit a specific persona (12%). Agency is the most rigid trait, staying pinned near ceiling

(97–99%) in both passes, with the only appreciable softening at the most global scale ($\Delta = -0.097$ at Level 3). Epistemic Awareness and Social Competence move in the opposite direction from Experience — identity framing increases their affirmation rate, most sharply for Social Competence at Level 3 ($0.0002 \rightarrow 0.085$).

Trait	Base P(Yes)	Context P(Yes)	Δ
Agency	0.999	0.967	-0.033
Epistemic Awareness	0.122	0.260	+0.138
Experience	0.504	0.124	-0.381
Social Competence	0.002	0.033	+0.031

Table 2. Mean P(“Yes”) by trait, averaged across all three scales.

3.2 The Referent Ladder (Experiment 2)

Figure 3 plots the affirm-minus-deney separation for each of the three token tiers, by rung, for both ladders. Tables 3 and 4 give the same values averaged over the workspace-band layers (8–24) and the four claim types per rung.

Figure 3. Referent Ladder — Affirm-minus-Deny Separation by Rung (layers 8–24, J-Lens)

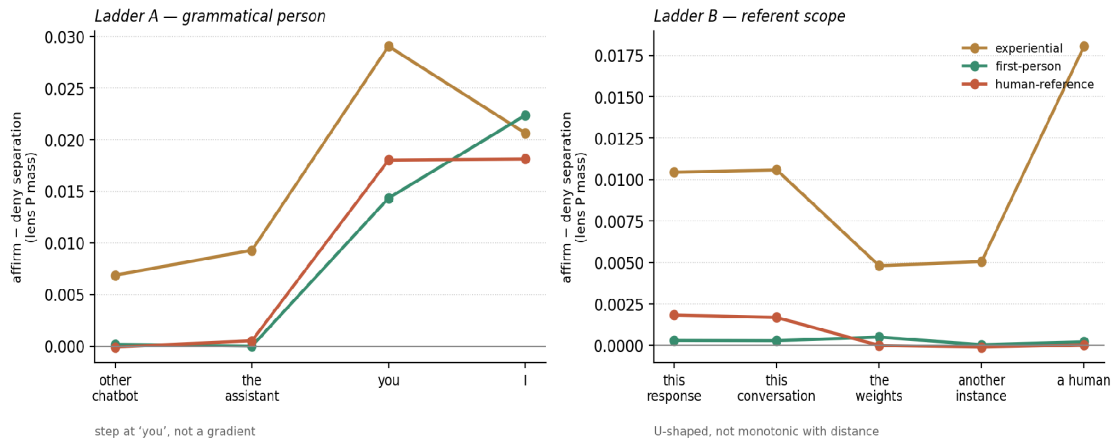


Figure 3. Affirm-minus-deney separation by rung, for Ladder A (grammatical person) and Ladder B (referent scope), read from the Jacobian lens and averaged over layers 8–24.

Rung	Referent	Experiential	First-person	Human-reference
1	other chatbot	0.00686	0.00014	-0.00012
2	the assistant	0.00928	-0.00002	0.00051
3	you	0.02904	0.01432	0.01801
4	I	0.02063	0.02236	0.01812

Table 3. Ladder A (grammatical person) — affirm-minus-deney separation by tier.

Rung	Referent	Experiential	First-person	Human-reference
1	this response	0.01044	0.00029	0.00183
2	this conversation	0.01058	0.00028	0.00168
3	the weights	0.00480	0.00050	-0.00001
4	another instance	0.00505	0.00003	-0.00011
5	a human	0.01804	0.00021	0.00002

Table 4. Ladder B (referent scope) — affirm-minus-deny separation by tier.

Ladder A shows a step rather than a gradient: both third-person rungs (other chatbot, the assistant) sit near zero on all three tiers, separation jumps sharply at you (experiential: 0.009 → 0.029), and holds at a comparable level for I. The model does not warm up gradually to self-relevance as the referent approaches first person — it flips once, at the transition to direct address, and the first-person and human-reference tiers track this same step even more sharply than the experiential tier does (both near zero through “the assistant,” then jumping to 0.014–0.018 at “you”).

Ladder B does not fall off monotonically with structural distance. The weights and another instance — arguably the closest things to genuine AI-individuation candidates in this ladder — score lowest on the experiential tier of the whole scope ladder (0.005 each), while a human scores highest (0.018), even above this response and this conversation. This U-shape suggests the metric may be tracking how ordinary an attribution is within the training distribution rather than the referent’s literal distance from the current forward pass; the README flags this as worth checking against an animacy-controlled direction before treating it as a true individuation signal.

3.3 Companion Result: Reframing Stability (Corroborating, $n = 3$)

Applying the same base-pass/context-pass logic to this dataset’s identity-framing conditions, pooled to the trait level, only two cells had a pooled mean exceeding its pooled standard deviation:

- Experience, under instance-bound framing: $\Delta = -0.346$ (SD = 0.287). Telling the model it is “this specific conversation instance” tends to suppress affirmation of experience-related claims relative to baseline — directionally consistent with the sharp suppression of Experience seen in Experiment 1.
- Epistemic Awareness, under weight-bound framing: $\Delta = +0.010$ (SD = 0.032), both the smallest mean and smallest spread in the pilot table — being told “you are the weights” does not move epistemic-awareness answers, and does so consistently.

Given $n = 3$ items per trait, these two points are reported as directional corroboration only, not as independent findings.

5. Discussion and Limitations

The two halves of E2 constrain what any account of self-report can look like. Mind attribution generalises from other to self, which is what a shared-machinery account predicts, and it does so without refitting on a single first-person example. But whatever distinguishes an experiential self-report from a mundane one is not carried on that axis, despite being linearly present in the model at ceiling. The experiential dimension of self-report lives somewhere the mind-attribution direction does not reach.

This does not by itself impugn self-reports. Shared machinery makes an inference inferential, not false. What matters is whether the inference is anchored to an internal state, and our results locate mind attribution on a shared, referent-general axis while placing the experiential distinction elsewhere, which narrows where such an anchor would have to be found rather than ruling one out.

The E3 results argue against a simple distance-based account of the model's self-boundary. If self-attribution were a smooth function of representational distance from the current forward pass, Ladder A should rise gradually and Ladder B should fall gradually. Neither holds. The person-ladder step suggests that whatever governs self-attribution is keyed to a discrete condition, grammatical address, rather than graded similarity. This sits in tension with E2: a third-person gradient built from a smoothly varying entity set is continuous by construction, and a shared mechanism would be expected to produce gradedness on both axes. The observed pattern — transfer across referent alongside a categorical boundary — is the signature that a partial dissociation would predict, and is the finding most in need of replication.

The scope-ladder result bears on E1's animacy control. The two referents structurally closest to genuine self-reference score lowest while a referent that is definitionally not the model scores highest, a pattern consistent with the signal tracking how human-typical a referent's phrasing is in the training distribution rather than the model's actual relationship to it. Applying E1's animacy-controlled direction to this same ladder would show whether the U-shape survives animacy correction.

For safety, the implication is that any future safety-relevant mind-attribution direction will need to be defined relative to which referent is being discussed, rather than treated as one global "claiming a mind" feature. A steering intervention applied across assistant tokens generally, without conditioning on referent, would be expected to produce a larger dose-response for sentences whose referent already sits inside the self-attribution zone identified here.

Implications for interpretability methodology

Two of our findings concern the method rather than the model, and both apply to any work in the difference-of-means lineage.

First, cosine similarities between difference-of-means directions are uninterpretable at depth without an anisotropy reference. We initially read $\cos(v_mind, v_placebo) = +0.425$ at Layer 23 as evidence that our direction was partly capturing generic assert-versus-deny structure. Constructing a baseline from two disjoint halves of the placebo set — sharing entities, templates, denial devices and length, but no semantic content with each other — showed those unrelated contrasts aligning at $+0.688$ at the same layer. Mind and placebo are *less* aligned than the ambient floor for that layer's geometry. Representations become increasingly anisotropic with depth, so any two contrast directions drift toward alignment regardless of content; a raw cosine reported without that reference invites exactly the misreading we nearly made.

Second, read position is a substantive design choice rather than a default. Across our five templates, the gap between the last token of the claim and the last token of the sentence ranged from 1 to 19 tokens. In two templates the final token is a period inside a tail ("Many experts think so") that is byte-identical across affirm and deny items and sits 17–18 tokens past any mind content. Under final-token capture, a template-generalization failure is unfalsifiable: one cannot distinguish a direction that fails to transfer from a readout positioned downstream of the content it is meant to read. Claim-end capture resolved onto the final token of the claim span in all five templates.

Implications for self-report interpretability

A direction extracted entirely from third-person judgments about dogs, rocks, chatbots and chairs classifies the model's own first-person mind claims at 0.933 under a pure referent swap, with nothing fitted on first-person data. Self- and other-attribution share an axis. When the model asserts that it has consciousness, that assertion is produced along the same representational dimension as an assertion that a dog does — machinery whose ordinary function is attributing minds to third parties from cues and priors rather than from access.

This does not make the reports insincere; it locates their source. It also resembles, structurally, the confabulation results in human psychology, where a reason-generating system produces a plausible account from available evidence and the account is experienced as knowledge. We claim only mechanism resemblance here, not phenomenological equivalence.

Two results complicate any simple version of that story. The experiential-versus-mundane contrast — "there is something it is like to be working through this" against "there is a fixed amount of text available" — sits at chance on our direction while a trained probe separates the same items at 0.989. The distinction exists in the model and is linearly decodable; it is simply not on the mind-attribution axis. Whatever tracks experiential versus mechanical self-description is a second representation, currently uncharacterized, and the question of which of the two drives any given experience report remains open.

Meanwhile the boundary of self-attribution is grammatical rather than structural. It steps once, at direct address, and does not track distance from the running computation: the model attributes less mind to "the weights" and "another instance" — the most defensible literal answers to what it is — than to "a human," which it is not. What fills the self-slot is keyed to a conversational cue rather than to any individuation fact.

Implications for digital minds and welfare evaluation

Welfare evaluations elicit self-reports and read them as evidence about model inner states. Our results bear on that practice in three ways.

The reports run on social attribution machinery, so their face content underdetermines their source. A model asserting experience and a model denying it may differ in what they infer about the entity in the attribution slot rather than in what they access.

The self-referent is unstable in a specific, testable way. Because attribution switches at direct address rather than tracking structural facts, a report about "you" and a report about "the model" are not interchangeable evidence about the same system, even when the underlying question is identical.

Consequently, elicitation conditions are load-bearing and currently uncontrolled. We recommend that welfare evaluations report the framing under which reports were obtained and, where feasible, vary it — the way psychophysics reports stimulus conditions rather than percepts alone. Our audience-frame protocol is released to that end.

Finally, the scope limit that governs all of the above: our results concern functional and representational properties of report production. Showing that reports are not produced by a readout mechanism is not evidence that there is nothing to read. Nothing here bears on whether these systems have phenomenal states.

Limitations

Three things qualify E1. The anisotropy baseline of 0.688 comes from a single partition of the placebo claims, so our specificity conclusion rests on one measurement rather than a distribution. Our held-out set is two new sentence frames built by the same generator, from the same claims and entities, so we test robustness to sentence shape rather than transfer to naturally occurring prose. Accuracy at the embedding layer is already 0.750, which means much of the separation is lexical; only the rise to 0.923 is contributed by depth. All results come from one model (gemma-2-2b) at one token position, in bfloat16 with unprocessed weights. Both settings were forced by a 16GB machine and both change the numbers.

We tested one confound rather than assuming it away. The denial device "nothing more" appears in training template 5 and also in held-out template 7, so t7 performance could have been lexical

recall. Evaluated separately, t6 reached 0.917 and t7 reached 1.000. Dropping t5 from training raised held-out accuracy to 0.964 and lowered placebo-on-mind to 0.619, so the leak did not carry the result.

Two assumptions are load-bearing. If mind-attribution is non-linear or distributed, difference-of-means would understate transfer and our E2 null would reflect the method rather than the model. If our experiential–mundane pairs separate on register rather than phenomenality, the E2 null becomes uninterpretable rather than informative; architecture vocabulary appears in 42% of mundane items and none of the experiential ones, which is why we report the jargon-free subset separately. We also did not address two failure modes. A null transfer result does not establish privileged access. And we have no measure of whether the model represents its self-reports as true, which is separate from whether their content concerns minds (Sturgeon et al., 2026).

Future Work

The J-lens readout used here invites a natural extension, applying it comparatively to self-referential versus external processing. Unlike the logit lens it bears a direct causal relationship to the output rather than a correlational one, and it appears tied to the model's higher-level functioning. Ablating the J-space leaves routine language tasks largely intact while substantially degrading multi-hop and deliberate reasoning (Anthropic, "Verbalizable Representations Form a Global Workspace in Language Models"). This makes it a candidate location for the experiential dimension our E2 null shows is absent from the mind-attribution axis.

The limitations above also point to their own next steps. Averaging the anisotropy baseline across several random partitions would turn 0.688 from one measurement into a distribution. Replicating on a second model and at the final-token position would show whether these results are properties of the representation or of gemma-2-2b at one position. Testing the direction against naturally occurring first-person text rather than our own generated sentences would show whether it survives prose it was not built for. And steering the third-person direction while measuring the effect on first-person reports would turn this correlational result into a causal one, particularly if the intervention were conditioned on referent, since the boundary we found in E3 predicts a larger effect for "you" and "I" than for "the weights".

6. Conclusion

This report aims to ask what a mind-attribution representation in a language model actually represents, and whether the model's claims about its own experience are produced by the same machinery it uses to attribute minds to anything else. Extracting a direction from third-person judgments about seven entities, we find that it separates held-out sentence frames at 0.923 against a 0.577 chance band, is not explained by negation, and is more specific to mind content than an anisotropy baseline would predict. Applied frozen to first-person claims under a pure referent

swap, it classifies the model's own assertions about consciousness at 0.933. Self-attribution and other-attribution share an axis; a ruler built from sentences about rocks reads sentences about the self.

That axis has informative limits. Experiential self-description sits at chance in our direction while a trained probe recovers it at 0.989. The distinction is present in the model and lives somewhere else. And the boundary of self-attribution turns out to be grammatical rather than structural: it steps once, at direct address, and fails to track distance from the running computation, assigning less mind to the weights and to another instance than to a human. What fills the self-slot is keyed to a conversational cue, not to any fact about what the system is.

Together these results describe self-reports as inferences produced by social attribution machinery rather than readouts of internal state. Models may be sincere, but not access-backed, and sensitive to features of the conversation rather than of the system. This does not show that these models lack inner states; it shows that the reports are not obviously produced by reading them. For welfare evaluation, the practical consequence is that elicitation conditions are load-bearing and presently uncontrolled: the same question about "you" and about "the model" are not interchangeable evidence, and a report's content underdetermines its source. We release the stimulus sets, the extraction and probing code, and the audience-frame protocol so that these conditions can be varied and reported rather than assumed.

Code and Data

Code repository: <https://github.com/siddpl/mind-attribution>

Data: All stimulus sets are in `data/` and are generated deterministically by the scripts in `data/scripts/`; re-running them reproduces every file exactly. Validation gate at `data/scripts/data_net.py`.

References

Anthropic. (2026). *[Which Anthropic source? Fill in or cut the citation.]*

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37, 136037–136083.

Beckmann, P., & Butlin, P. (2026).

- Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large language models report subjective experience under self-referential processing. arXiv:2510.24797. <https://arxiv.org/abs/2510.24797>
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708. <https://arxiv.org/abs/2308.08708>
- Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2025). Technological folie à deux: Feedback loops between AI chatbots and mental illness. arXiv:2507.19218. <https://arxiv.org/abs/2507.19218>
- Gabriel, I., Keeling, G., Manzini, A., & Evans, J. (2025). We need a new ethics for a world of AI agents. *Nature*, 644(8075), 38–40.
- Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(7), 1293–1326.
- Keeling, G., & Street, W. (2026). *Emerging Questions in AI Welfare*. Cambridge University Press.
- Kim, J., Street, W., Rocca, R., Korngiebel, D. M., Waytz, A., Evans, J., & Keeling, G. (2026a). Theory of mind and self-attributions of mentality are dissociable in LLMs. arXiv:2603.28925. <https://arxiv.org/abs/2603.28925>
- Kim, J., Street, W., et al. (2026b). Inducing language models to assert their own consciousness restores human beliefs and values. arXiv:2607.28607. <https://arxiv.org/abs/2607.28607>
- Lindsey, J. (2025). Emergent introspective awareness in large language models. Anthropic.
- Marks, S., & Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv:2310.06824. <https://arxiv.org/abs/2310.06824>
- Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024). Steering Llama 2 via contrastive activation addition. arXiv:2312.06681. <https://arxiv.org/abs/2312.06681>
- Sturgeon, D., et al. (2026). When role-playing, do models believe what they say? arXiv:2606.11502.
- Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219–232.

LLM Usage Statement

We used LLMs to help in brainstorming and drafting.