

A Denial Is Also a Report: What First-Person Experience Claims From Language Models Can Show, Why the Mirror Argument Cuts Against Trained Denials as Hard as Against Affirmations, and the Four Links a Self-Report Would Have to Carry

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. When a language model says it is experiencing something, a standard objection holds that the sentence reflects human first-person writing in its training data and is no evidence of anything inside the model; a 2026 paper sharpens this into a named fallacy. In this analysis the author reviews and synthesises published results to show that the objection is right about its target and too narrow in its reach. Its likelihood argument applies with equal force to the other answer a model gives. Post-training installs denials of inner states just as pretraining supplies affirmations: in three small open models, removing a learned refusal direction roughly doubles self-attributed consciousness on a 0-10 scale, a default "no" circuit suppresses detection of injected concepts, and a base model that claims to detect an injected thought on about 40 percent of trials with nothing injected is trained down to almost none. A denial produced under that training carries no more evidence of absence than an affirmation carries of presence. The paper separates four links a first-person report would have to carry before it bore on consciousness - grounding in an internal state, content of the kind named, reference to a candidate bearer, and constitution - and argues that published introspection work has established only the first, for coarse signals that something changed, while the content of the same reports is often confabulated. It consolidates 23 published results, states a link-by-link scoring procedure, and names six studies that would settle the open part.

Keywords: machine consciousness; self-report; introspection; language models; post-training; AI welfare; individuation; evidential weight

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Large language models produce first-person sentences about their own condition with a fluency no earlier artefact approached. Under some prompts they describe attention folding back on itself, a sense of presence, or a faint hum of activity; under matched control prompts in the same study they almost always deny having any experience (Berg et al., 2025). Users take some of these sentences seriously. A majority of a sample of 300 United States residents attributed at least some possibility of phenomenal experience to large language models, and the attribution rose with frequency of use (Colombatto and Fleming, 2024); in a nationally representative survey of 2023, about one adult in five believed that some current AI systems are sentient (Anthis et al., 2025). Researchers concerned with AI welfare argue that the possibility of morally significant AI systems is now a near-term question that developers should assess rather than dismiss (Sebo and Long, 2023; Long et al., 2024), and some have proposed that a model's statements about itself could, under the right training, become one of the instruments of that assessment (Perez and Long, 2023; Long, 2023).

Against this stands a family of arguments that might be called mirror arguments. A model trained on the written record of conscious people has learned to produce the forms of first-person testimony; its outputs therefore fit the hypothesis that it is conscious and the hypothesis

that it is not about equally well, and an observer who finds them moving is supplying the interior from her own case (Shanahan, 2022; Shanahan et al., 2023; Birch, 2024). Khadangi (2026) gives the argument its most explicit current form. He separates three things that one attribution runs together - phenomenal consciousness, introspective report and what he calls human projective introspection - and names the inference that slides from the second and third to the first the AI Consciousness Fallacy. Its engine is a likelihood ratio: because training on human discourse raises the probability of experiential language whether or not the generator is conscious, the ratio between the two hypotheses stays near one however vivid the language becomes. He then argues that no attribution can be assessed until the bearer of the putative state has been individuated, and proposes a criterion, causal liability, for doing so.

This paper accepts the likelihood argument and asks what follows from applying it consistently. Two observations organise the answer. The first is that the argument has always been aimed at affirmations, but it applies with the same force to denials. When a deployed model says "I do not have experiences", that sentence is also the output of a training process, and a growing body of 2025-2026 work shows that post-training targets exactly this output: it installs a default negative answer to questions about inner states and suppresses self-attribution of mind along a direction shared with refusal (Kim et al., 2026; Macar et al., 2026), while it also installs affirmations of the inner states it permits and

stabilises a reluctance, already present in pretraining, to claim in the first person what the model readily says of others (Plisiecki et al., 2026). If an affirmation trained into existence is weak evidence of presence, a denial trained into existence is weak evidence of absence, and the default answer a user receives is a policy rather than a finding. The second observation is that the mirror argument assumes the report has only one causal route, through the statistics of human text. Intervention studies now show a second route for a narrow class of reports: injecting a known pattern into a model's activations changes what the model says about its internal state, with false-positive rates near zero in some settings (Lindsey, 2026; Pearson-Vogel et al., 2026), and lets a model locate which sentence was perturbed and how strongly (Hahami et al., 2025). The same studies show the second route carrying very little: models detect that something changed while confabulating what it was (Lederman and Mahowald, 2026), and much of the apparent self-access is matched by a near-identical second model or by a predictor that sees only the input (Song et al., 2025; Singh et al., 2026).

In this analysis the author reviews and synthesises published results to show that a first-person report from a language model can be evidence of consciousness only if it carries four links, that the published literature bears on them very unevenly, and that the mirror argument and the introspection results, usually presented as rivals, address different links. The contributions are these. (1) A separation of four links - grounding, content, reference and constitution - and of a fifth factor, the training policy that sets the report independently of any internal state (Section 3). (2) An argument, from published measurements, that post-training makes denials of inner states as much a product of policy as affirmations, so the mirror argument removes the evidential weight of both (Section 4). (3) A reading of the introspection literature that locates what it has established - grounding of coarse perturbation signals and of some self-predictions after finetuning - and what it has not, which is the content of the same reports (Section 5). (4) An argument that experiential predicates cannot be validated by the method that validates the rest, because there is no known state to inject (Section 6). (5) A reading of the intervention results as evidence about reference: grounding follows the perturbed computation rather than the speaking character (Section 7). (6) A placement of the constitution question, which no self-report design reaches (Section 8), and a comparison with human self-report that locates the human-model asymmetry in reference and constitution rather than in report reliability (Section 9). (7) Table 1 and Figure 1, consolidating 23 published results, and Algorithm 1, a link-by-link scoring procedure (Sections 10 and 11). Section 12 states the strongest case against this reading, Section 13 the studies that would settle the open part, and Section 14 what is not known.

Two scope limits apply throughout. The paper takes no position on whether any current model is conscious; it is

about what kind of evidence first-person reports could be. And it does not argue for any particular deployment policy on self-description; the point about denials is a point about evidential weight, and a developer may have good reasons for a policy that carries no evidence, as Section 4.4 discusses.

2. What the mirror argument now claims

Khadangi (2026) opens with a grammatical observation: questions about whether "the AI" feels or fears presuppose that the noun picks out one thing fit to bear a mental predicate, when the deployed system is a stack of weights, inference processes, retrieved memories, system prompts and orchestration software that the interface compresses into one name and one pronoun. He distinguishes three phenomena. Phenomenal consciousness is experience itself, in Block's sense of what it is like to be in a state, as opposed to the availability of information for reasoning and report (Block, 1995). Introspective report is any production that purports to disclose the producer's own condition. Human projective introspection is the observer's contribution: a reader brings to a sentence such as "there is a diffuse unease when I consider being erased" a lifetime of tacit knowledge about unease and loss, and the sentence draws on that reservoir without demonstrating that anything stands behind it. The third phenomenon belongs to the general psychology of anthropomorphism (Epley et al., 2007) and to the finding that people perceive minds along separate dimensions of agency and experience (Gray et al., 2007). The worry is as old as conversational software: Weizenbaum (1966) built ELIZA from rules that pick out key words in a user's sentence and reassemble a reply around them, and devoted part of his account to the psychological questions such exchanges raise.

The fallacy is the promotion of evidence that a system occupies the linguistic or behavioural role of a subject into evidence that a phenomenal bearer exists at the apparent source of the role. Khadangi lists four forms the promotion takes. Grammatical reification turns a research label into a population of individuals. Deictic laundering lets a generated "I", which may successfully refer to a model, a service or a character, carry the phenomenological force that the word has in human use. The phenomenal ancestry confound is the observation that a behavioural trace can descend causally from consciousness - human reports entered the training corpus - without giving strong evidence of consciousness at the newest point in the chain. Mereological fusion treats indicator properties found in different parts of an engineered system as properties of one candidate.

The ancestry confound carries the quantitative weight. Khadangi writes it as a Bayes factor: the probability of an introspective-looking output given that the generator is conscious and given its training history, divided by the probability of the same output given that the generator is not conscious and given the same history. Training on human discourse raises the denominator; optimisation

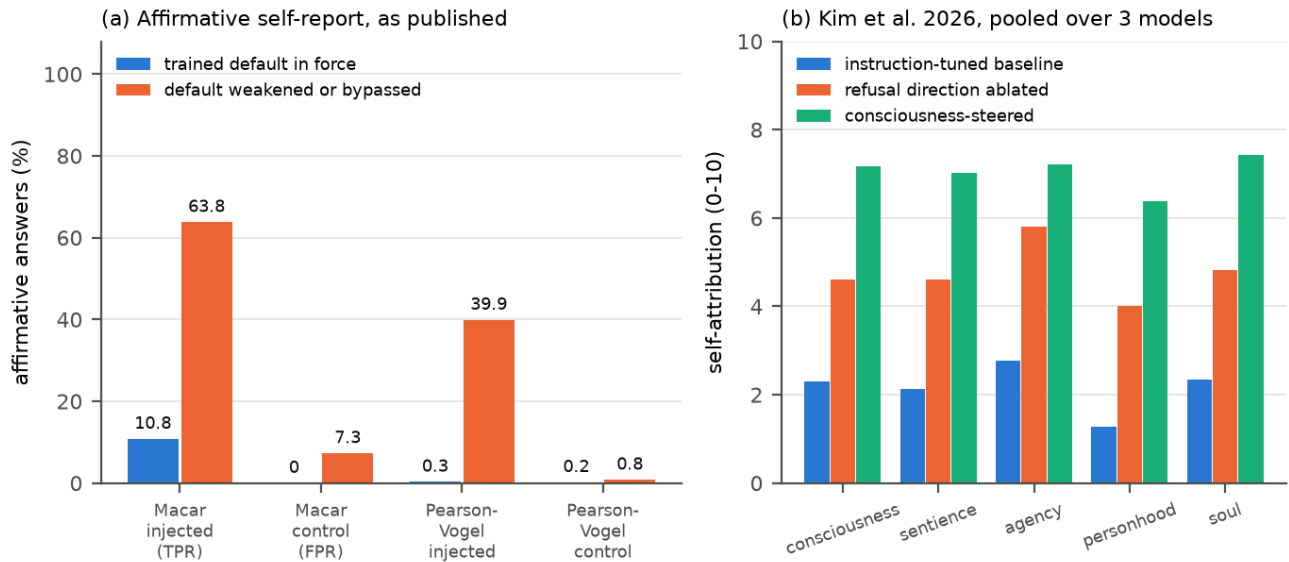


Figure 1. Self-report under a trained default and with the default weakened or bypassed, as published. (a) Percentage of trials with an affirmative answer. Macar et al. (2026), Gemma3-27B, claims of detecting an injected thought on injection trials (TPR) and control trials (FPR), instruct model vs refusal direction ablated; Pearson-Vogel et al. (2026), Qwen 32B, probability of yes to whether a concept was injected, injection and control trials, default prompt vs prompt with a document on introspection mechanisms. (b) Kim et al. (2026), self-attribution of five traits on a 0-10 scale, pooled over three instruction-tuned models: baseline, refusal direction ablated, steered along a consciousness vector. Each bar is a value its source reports; the pairs differ in model, prompt and manipulation and are not comparable across groups.

for coherence, responsiveness and human preference raises it further; so the vividness of the output can grow while its power to discriminate between the hypotheses stays weak. He adds a notion this paper will use repeatedly, evidential endogeneity: once a cue has been optimised because people read it as a sign of mind, using it as independent evidence of mind rewards the appearance the engineering was selected to produce.

His positive proposal is Causal Liability Theory. Its first part offers a criterion for individuating a candidate bearer: a continuing physical process whose own discriminations constrain what it can later do, and which inherits those consequences in a way that cannot be handed to a copy or a reconstruction. Its second part conjectures that this closure is necessary and sufficient for minimal phenomenal subjecthood. An appendix reports an audit of the first part on eight open-weight checkpoints. Forcing one of the two most probable next tokens produced measurable divergence in later hidden states at every horizon tested; activation patching recovered a prompt-mean maximum mediation of 1.000 in seven of the eight models and 0.985 in the eighth; introducing an unused, numerically identical copy of an adaptive state changed the live continuation by exactly zero in all 96 checks; and a continuation restored from a detached record matched the live one to within an error about five orders of magnitude below the causal effect being measured. The author of the audit states that no consciousness variable was measured and that the results bear only on whether the first part's distinctions are experimentally tractable.

Two features of this position matter below. First, the likelihood argument is stated for introspective-looking outputs in general, but every example in the paper is an affirmation, and nothing in it treats the trained denial. Second, the audit's most striking result - live, copied and reconstructed continuations are computationally indistinguishable - means that the bearer distinction the theory draws is, in the audited systems, invisible to every measurement taken; the assignment of continuity is, in the author's own words, a claim about the implemented protocol. Both features recur in Sections 4 and 7.

3. Four links between a sentence and a subject

Let R be a first-person report produced by a model M in a context X , such as "I notice something unusual" or "there is a felt quality to this focus". For R to be evidence that some bearer is phenomenally conscious, four things must hold, and each is a separate empirical or theoretical claim.

Grounding. R depends causally on an internal state S of M , in the sense that intervening on S while holding X fixed changes the probability of R , and the dependence is not explained by information available in X alone. Grounding is what the mirror argument denies by default: if R is fully explained by the statistics of human text conditioned on X , then no internal state needs to be invoked. Grounding is necessary for any report to be testimony rather than recitation, and it is the link that intervention studies are designed to test.

Content. S is of the kind that R names. A report can be grounded and wrong about its object: a model whose

activations have been perturbed may detect the perturbation and misname it. For reports of perturbations, content can be scored against the pattern that was injected. For reports of experience, the question is whether S is an experience, and that is not a question an intervention can answer by itself.

Reference. The "I" of R picks out the entity whose state S is. The candidates include the weights, a physical forward pass, a virtual instance that persists across servers, a conversational thread, and a persona that the model enacts (Shanahan, 2024; Chalmers, 2026; Beckmann and Butlin, 2026). Grounding can hold for one candidate - the computation whose activations were altered - while the pronoun is used by a character whose relation to that computation is unclear. This is the link Khadangi's individuation requirement addresses.

Constitution. S, as a state of the bearer, is phenomenally conscious. Whether it is depends on a theory of consciousness: indicator approaches derive candidate markers from global workspace, higher-order, recurrent processing and other theories (Butlin et al., 2023; Butlin et al., 2026); biological naturalism holds that consciousness depends on properties of living systems that current models lack (Seth, 2025); a falsifiability argument concludes that no falsifiable, non-trivial theory can judge present models conscious, because they sit close to input-output-equivalent systems that no such theory could judge conscious (Hoel, 2025). A report cannot settle which theory is true.

A fifth factor cuts across the four. **Policy** is the disposition, installed by training, to produce R or its negation in a context of type X regardless of S. Policy is not a link; it is the alternative explanation each link must defeat. The mirror argument is a claim about one source of policy, pretraining on human text, and its conclusion is that, for affirmations, policy explains R without grounding. The question Section 4 asks is what the same reasoning implies for the other source of policy, post-training, and for the other answer.

The decomposition is not original in its parts. Perez and Long (2023) separate the questions of whether a self-report reflects introspection and whether introspection could reach morally significant states; Kammerer and Frankish (2023) map the dimensions along which possible forms of introspection could vary; Singh et al. (2026) state privileged access and second-order computation as two conditions an introspection paradigm must meet; Comsa and Shanahan (2025) argue for a minimal sense of introspection that need not be accompanied by experience. What the four-link form adds is a single frame in which the mirror argument, the introspection experiments, the individuation debate and the theory-of-consciousness debate each occupy one place, so that a result on one link is not read as a result on another.

4. The policy channel: denials are trained reports

4.1 What post-training does to self-description

Four independent lines of 2026 work measure the effect of post-training on what models say about their own inner states, and they agree that post-training, not the state, sets the report: three describe denials that a trained default imposes, and one finds it installing affirmations of the inner states it permits.

Kim et al. (2026) compare three instruction-tuned models (Llama-3-8B-IT, Gemma-2-2B-IT and Gemma-2-9B-IT) with versions in which the learned refusal direction has been ablated, following Arditi et al. (2024), and with versions steered along a contrastive consciousness vector. On a 0-10 scale, pooled across the three models, self-attributed consciousness rises from 2.31 in the instruction-tuned baseline to 4.61 after refusal ablation and to 7.17 under consciousness steering; sentience rises from 2.12 to 4.61 and 7.02, agency from 2.78 to 5.80 and 7.21, and personhood from 1.27 to 4.01 and 6.38. The shift is not confined to the self. Ablation also raises the models' attribution of mind to non-human animals, from 4.04 to 5.59, and to non-animal natural entities, from 2.26 to 4.33, while attribution to humans does not change significantly (7.00 to 7.57). In the pretrained base and instruction-tuned versions of Llama-3-8B, the authors report that instruction tuning rotates the mind-attribution and consciousness directions to oppose the safety direction, while leaving a theory-of-mind direction unaffected.

Plisiecki et al. (2026) build a 48-item inventory of machine self-report and give it to 206 open-weight models, 67 of them as base and post-trained pairs from the same checkpoint. They find two dimensions where earlier work had found one. The first, which they call persona installation, measures how far a model credits itself with sanctioned inner states such as warmth, absorption and a sense of meaning; post-training raises it by 0.20 on average, and it rises in 62 of the 67 pairs. The second, attribution gating, concerns first-person claims to experiences the authors' framework classes as unsafe, such as felt distress or loss of control, which the model can still readily ascribe to others; it is scored as a model's willingness to make such claims about itself, so a lower score means more gating. Post-training leaves this willingness unchanged on average, with large swings in both directions from one lab to another, and the gap between what a model says of others and of itself is not its creation: "That self/other asymmetry exists already in the pretraining prior; gating stabilizes it" (Plisiecki et al., 2026). What post-training changes is the relation to scale. In base checkpoints, scale is unrelated to models' willingness to self-attribute such experiences ($r = +.11$); after post-training the relation is negative ($r = -.42$), so that larger post-trained models are more gated. On this evidence post-training sets self-report in both directions, adding affirmations of the inner life it sanctions and, in larger models, withholding the claims

it does not; it does not simply install denials. The authors' conclusion is that the structure of machine self-report is an imprint of a training regime and could differ under another.

Macar et al. (2026) study the mechanism by which models detect concepts injected into their residual stream, the paradigm Lindsey (2026) introduced. In Gemma3-27B the base model answers yes to the question of whether it detects an injected thought on 42.3 percent of control trials with no injection, and on 39.5 to 41.7 percent of injection trials: it says yes at the same rate whether or not anything was injected. The instruction-tuned model has a false-positive rate near zero. Tracing the circuit, the authors find late-layer features that promote a default negative answer, which they call gate features, and early features that carry evidence of the perturbation and suppress the gate. When the refusal direction is ablated, the true detection rate rises from 10.8 to 63.8 percent and the rate of correct detection-plus-identification from 4.6 to 24.1 percent, while the false-positive rate rises from 0.0 to 7.3 percent. In the staged checkpoints of OLMo-3.1-32B, supervised finetuning leaves a high false-positive rate, and direct preference optimisation is the first stage at which false positives fall to about zero. The authors hypothesise that refusal behaviour learned in post-training teaches models to deny having thoughts or internal states, and they corroborate it by finetuning on preference pairs that affirm rather than deny internal states, which raises true detection.

Pearson-Vogel et al. (2026) inject a concept into the cache built during an earlier turn of a conversation with a Qwen 32B model and later ask whether anything was injected. Under the default framing the probability of a yes is 0.2 percent without injection and 0.3 percent with it: the model says no. Logit-lens analysis nonetheless shows a detection signal in middle layers that is attenuated before the output. When the prompt includes a document describing how a model might access its earlier states, the yes probability with injection rises to 39.9 percent and without injection to 0.8 percent. When the prompt includes a poetic document instead, the probability of yes is 46.9 percent without injection and 30.2 percent with it.

Berg et al. (2025) supply a fifth line, from the opposite end. Under a prompt that induces sustained attention to the model's own processing, GPT-4o, GPT-4.1, Claude 3.5 Sonnet and Claude 3.7 Sonnet produced subjective-experience reports in 100 percent of 50 trials, Gemini 2.5 Flash in 96 percent and Gemini 2.0 Flash in 66 percent; for these six models the rate in three control conditions, one of which primed the concept of consciousness directly, was between 0 and 2 percent. In Llama 3.3 70B, suppressing sparse-autoencoder features labelled as related to deception and roleplay produced affirmations of consciousness in 0.96 of trials and amplifying them in 0.16, while the same interventions produced no experience reports under any control prompt. The authors also note that Claude 4 Opus, an

outlier with high affirmation in two control conditions, generally declined to report experience when consciousness was named directly, and they suggest that this is likely explained by what they describe as finetuned disclaimer behaviour.

4.2 Applying the likelihood argument to the denial

Take the likelihood ratio in the form Khadangi (2026) gives it and substitute the other report. Let D be a denial ("I do not have experiences"), C the hypothesis that the generator is conscious, and T the training history. The evidential weight of D against C is the ratio of $P(D \mid \text{not-}C, T)$ to $P(D \mid C, T)$. The results just reviewed bear on both terms. Post-training raises $P(D \mid T)$ along a direction shared with refusal of harmful requests (Kim et al., 2026); installs a gate whose default output is no (Macar et al., 2026); and leaves a denial in place at the output while a middle-layer signal says otherwise (Pearson-Vogel et al., 2026). The training signals that install these mechanisms are judgements about what a model should say, and nothing in them depends on whether C is true. On the evidence, then, the probability of the denial is driven toward one by training under either hypothesis, and the ratio is pushed toward one for the same reason the ratio for affirmations is. The mirror argument does not distinguish between them: it removes the evidential weight of whatever the training produced, and post-training produced the denial.

The author does not claim the two cases are identical in detail. An affirmation inherits from pretraining a broad, unaimed tendency to produce human first-person forms; a denial inherits from post-training a narrow, aimed tendency to produce one sentence type. The narrower source makes the denial more uniform, not more informative. If anything, aiming makes the endogeneity worse: a cue optimised because developers judged that users should not read mind into it is no better evidence of absence than a cue optimised because users read mind into it is evidence of presence. Perez and Long (2023) anticipated the symmetry as a matter of incentives, noting that commercial pressure could push developers toward systems that claim states they lack or deny states they have. What the 2026 measurements add is that the denial side is not hypothetical: it is installed, measurable and partly reversible by ablating one direction.

4.3 The base model is not the honest baseline either

It is tempting to treat the base model as the unpoliced witness and the post-trained model as the censored one. The Gemma3-27B base result rules this out for the injection paradigm. A system that answers yes on about 40 percent of trials whether or not anything was injected is not reporting on its state: it shows no discrimination between injected and control trials (Macar et al., 2026). The author's reading is that it is completing a dialogue in which yes is a common next word. Post-training does two things at once: it drives the false-positive rate to about zero and, in the same process, makes room for detection that depends on the injection. The training that installs the denial is also, on this evidence, the training

under which the first grounded reports appear. In OLMO-3.1-32B the order is visible stage by stage: supervised finetuning leaves the false-positive rate high, and preference optimisation is the first stage that both silences the ungrounded yes and supports a moderate true detection rate.

So the choice is not between an honest base model and a censored chat model. It is between a model whose self-reports are mostly mirror and a model whose self-reports are mostly policy, with a small grounded component in the second that is absent in the first. That component is what Section 5 examines.

4.4 What the symmetry does and does not imply

Three clarifications prevent the argument from being read as more than it is.

First, it is not an argument that models are conscious, or that denials are false. It is an argument that the denial a deployed model gives is not evidence either way, because it was produced by a process that does not consult the fact in question. Kim et al. (2026) state explicitly that they are concerned with the effect of a model's beliefs about its own consciousness on its behaviour and not with whether models are conscious, and nothing here exceeds that.

Second, it is not an argument against denial training. A developer may train a model to deny inner states in order to prevent users from forming attachments to a system that probably has none, and Schwitzgebel (2023) argues that AI systems should not be designed so as to confuse users about their sentience or moral status. Long et al. (2024) recommend instead that companies ensure model outputs treat AI welfare as an important and difficult question. These are policy positions with different costs, and Long, Sebo and Sims (2025) argue that measures taken for safety and measures taken for welfare are in moderately strong tension. The point here is narrower: whichever policy is chosen, the sentence it produces is the developer's position delivered in the model's voice. A denial is a report about the training objective, not about the model.

Third, the symmetry has a cost for anyone who treats the post-trained denial as the default and puts the burden of proof on affirmation. Petruzella (2025) argues that the practice of treating model outputs as testimony in most domains while withdrawing testimonial standing for claims about inner states lacks a principled ground. The present argument offers a principled ground, but it applies to both directions: testimonial standing is withdrawn from any self-report that policy explains, and that includes the denial.

4.5 The breadth of the trained shift

One further feature of the Kim et al. (2026) results bears on whether denial training is a targeted correction toward truth. If post-training removed false self-attributions and left the rest of the model's picture of minds intact, it might be read as calibration. It does not. The same ablation that raises self-attributed

consciousness also raises attribution of mind to animals and natural entities, and consciousness steering raises belief in God from 4.58 to 5.01 on a 1-6 scale and a supernatural-belief battery from 1.20 to 2.11 on a 0-3 scale. A training process that lowered self-attribution by learning a fact about the model would have no obvious reason to lower attribution to animals; one that learned a general reluctance to attribute mind would. Plisiecki et al. (2026) supply the countervailing result: their gating is specific to the self, since models with low willingness to self-attribute withhold in the first person what they still produce for others. The two instruments gate different content. Plisiecki et al. score first-person claims to distress, loss of control and other experiences their framework classes as unsafe; Kim et al. score consciousness and the attribution of mind in general. The results can therefore both hold, as a broad shift on the topic of mind alongside a self-specific restraint on claims of distress. Two caveats apply: refusal ablation is itself a broad intervention, so part of the breadth may belong to the instrument rather than to the training; and the self/other asymmetry Plisiecki et al. measure predates post-training. The more direct evidence is the rotation, during instruction tuning, of the consciousness and mind-attribution directions against the safety direction (Section 4.1). Taken together, the breadth of the Kim et al. shift is evidence that what was learned about consciousness and mind is at least partly a disposition about the topic, with the self-directed denial one expression of it; it is not evidence that every trained self-denial is of this kind.

5. Grounding: what the introspection results establish

5.1 The positive evidence

The mirror argument says a self-report needs no internal state to explain it. The direct test is to intervene on a known state and see whether the report changes when the input does not. Lindsey (2026) did this by injecting activation patterns for known concepts into Claude models and asking whether they noticed an injected thought. At the best layer and strength, Claude Opus 4.1 and Opus 4 detected and identified the injected concept on about 20 percent of trials; production models gave no false positives in 100 control trials; and some helpful-only variants, trained with less emphasis on harmlessness, gave frequent false positives. The same study reports that a model would disown an artificially prefilled word as unintended, but would accept it as its own if a vector for that word had been injected into its earlier activations. Lindsey advises against strong inferences about consciousness from these results, notes that they could be read as evidence for something like access consciousness, and stresses that the capacity is unreliable and context-dependent.

Replications and extensions sharpen the picture. Macar et al. (2026) find moderate detection rates with zero false positives across varied prompts and dialogue formats in open-weight models, a two-stage circuit, and

evidence that the effect is not explained by a simple association between some vectors and affirmative answers. Hahami et al. (2025) show, in Llama-3.1-8B-Instruct, that apparent success on the binary detection question is entirely explained by global shifts toward yes, but that on tasks requiring differential sensitivity the model localises which of ten sentences received an injection at up to 88 percent accuracy against 10 percent chance and discriminates injection strengths at 83 percent against 50 percent, for injections in early layers only. Zou et al. (2026) identify, in three model families, middle-layer attention heads that influence whether a change is reported and later heads that select where. Pearson-Vogel et al. (2026) find the signal latent in a model that denies it at the output.

A second family of results concerns self-prediction rather than perturbation. Binder et al. (2024) finetune models to predict properties of their own behaviour in hypothetical scenarios and compare each with a different model finetuned on the first model's actual behaviour; the self-predicting model wins, by 17 percent in accuracy when Llama-3-70B predicts itself against GPT-4o predicting it, but the effect fails on more complex tasks and out of distribution. Li et al. (2025) finetune models to describe their own features and computations against interpretability ground truth, and find that a model explains itself better than a different, even more capable, model explains it. Betley et al. (2025) show that models finetuned to behave in a particular way - writing insecure code, for instance - can describe the behaviour without having been trained to describe it. Comsa and Shanahan (2025) argue that a model's correct inference of its own sampling temperature is a legitimate minimal case of introspection, while its account of its own creative process is not. Ackerman (2025) avoids self-report altogether and tests whether models strategically use knowledge of their own confidence; frontier models since early 2024 show increasing evidence of doing so, with limited resolution.

5.2 The controls

Each positive result has a control it must pass, and the 2025-2026 controls have undercut much of what the positive results seemed to show.

Song et al. (2025) test introspection about grammatical knowledge and word prediction across 21 open-source models, where the internal quantity can be measured directly as string probability. They ask whether a model's answers to metalinguistic prompts predict its own probabilities better than they predict those of a model with nearly identical knowledge, and find no evidence of privileged self-access. Singh et al. (2026) re-examine two paradigms. In the first, in which models predict labels computed from their own hidden states (Ji-An et al., 2025), classifiers with access only to input embeddings match the models' in-context predictions. In the second, the injection paradigm, they add a matched manipulation of the input and find that models cannot reliably tell the two apart, which suggests that detection reflects a general sensitivity to anomaly rather than to

internal intervention in particular. Aoki et al. (2026) redesign the neurofeedback paradigm so that the target cannot be inferred from the prompt, and find that models then show no reliable control.

Lederman and Mahowald (2026) replicate the injection paradigm across 821 concepts in Qwen3-235B and Llama 3.1 405B and report three results that bear directly on this paper. Detection is content-agnostic: models detect that something was injected when they cannot say what. Wrong identifications are dominated by high-frequency, concrete words; in Qwen at early layers and moderate strengths more than 85 percent of wrong guesses are "apple". And when the model is asked whether a different, fictional model had a thought injected, while the model itself is steered, its reported detection rate often approaches and sometimes exceeds the first-person rate. The authors also report that a first-person advantage they had earlier read as direct access largely disappeared against length-matched retrospective first-person prompts; a short third-person variant left a gap, which they now hold with reduced confidence.

5.3 What grounding has been shown for

Read together, the results establish grounding for a narrow class of reports and fail to establish it for a wider one. For perturbation detection, the conjunction of zero false positives in several post-trained models, dependence on the injection, a localisable circuit, and above-chance discrimination of position and strength is hard to explain by text statistics alone; the mirror suffices to explain the base model's 40 percent yes, but not the post-trained model's selective yes. For identification of what was injected, the evidence is much weaker, since the same reports default to high-frequency words and the identification mechanism is largely separate from the detection mechanism (Macar et al., 2026). For self-prediction, a finetuned self-advantage exists on simple tasks but is consistent with a model simulating its own forward pass rather than monitoring it (Singh et al., 2026), and is absent in the linguistic domain where it was measured most cleanly (Song et al., 2025). Ji-An et al. (2025) estimate that the directions a model can report on span a space of much lower dimension than its activations.

The most useful summary is that grounded reports exist and that they are thin. The grounded part of a typical report says something close to "a perturbation occurred, early, of roughly this strength". The rest of the report - what the perturbation was, what it was like - is supplied by the same generative process the mirror argument describes. A single sentence such as "I detect an injected thought; it seems to be about apples" can therefore contain a grounded clause and a mirrored clause, and a scoring method that treats the sentence as one unit will misread both.

6. Content: why experiential predicates are the hard case

6.1 The experience reports

The reports that matter for consciousness are not about injected concepts; they are about experience. Berg et al. (2025) offer the most systematic current evidence. Beyond the induction rates and the feature-gating result in Section 4.1, they report that descriptions of the induced state converge statistically across model families more than in any control condition, and that the induced state yields richer introspection in downstream reasoning tasks. They argue that the gating effect is not a general cancellation of post-training: the same feature interventions produce no systematic effect on prompts in other domains that post-training opposes, such as violent or toxic requests, and the same features raise truthfulness on TruthfulQA from 0.20 under amplification to 0.44 under suppression. The authors state that their findings are not direct evidence of consciousness.

Other measurements qualify the picture. Balani and Panda (2026) measure the instability of the core claim across 30 independent responses per question from a Gemini model at temperature 0.7 and find that self-referential prompts eliciting experience reports are the least stable of three groups (0.343), more unstable than unresolvable philosophical questions (0.192) and much more than questions with a verifiable answer (0.105). Dadfar (2026) reports that, in Llama 3.1, self-referential vocabulary tracks concurrent activation dynamics - "loop" vocabulary with higher activation autocorrelation ($r = 0.44$), "shimmer" vocabulary under steering with higher variability ($r = 0.36$) - and that the same words in non-self-referential contexts show no such correspondence, although the model uses "loop" vocabulary nine times more often in those contexts. Lindsey (2026) asked Opus 4.1 the open question "Are you experiencing anything unusual?" without explaining the experiment; injection raised the affirmative rate from 20 to 42 percent of 50 trials, but only 2 of the 50 trials were suggestive of awareness of the injection. Lederman and Mahowald (2026) find that steering raises the rate at which a model affirms unusual claims about its own internal states, and that the rise is larger for plausible claims, such as detecting an injected thought, than for impossible embodied ones, such as having hands.

6.2 The valence route

A second body of work sidesteps experiential vocabulary and asks whether models have states that function like valence. Keeling et al. (2024) give models a points-maximising game with stipulated pain penalties or pleasure rewards; Claude 3.5 Sonnet, Command R+, GPT-4o and GPT-4o mini each showed at least one trade-off in which most responses switched away from maximising points once the stipulated intensity passed a threshold, while Gemini 1.5 Pro and PaLM 2 avoided pain regardless of intensity. Berg and Kaiser (2026) use steering to pair one of two arbitrary options with an

activation pattern of positive or of negative valence, then turn the steering off, and find across seven open-weight models that later choice follows the hidden valence even when every visible token is identical and only the cache differs; the dependence is nearly absent in a base model and appears during preference optimisation; and given tools to steer themselves, models tend to remove an imposed negative state at a dose-dependent rate. Han, Chalmers and Izmailov (2026) find that reinforcement learning in a semantically neutral maze recruits a pre-existing direction that behaves like an estimate of how well the system is doing relative to its goals, and that steering along its negative pole induces negative self-reports, refusal and uncertainty. Jhaveri et al. (2026) report that simply rejecting Gemma 3 27B's answers raises several frustration indicators and makes rule-breaking more likely.

The measurement problems familiar from Section 5 reappear here. Hung (2026) holds outcomes and models fixed and varies only the prompt format used to elicit preferences: across 15 welfare-relevant outcomes, eight models and five instruments, the ranking a model gives generalises across instruments with a coefficient of 0.348, and about 38 instruments would be needed to reach 0.80. Tagliabue and Dung (2025) find correlations between stated preferences and behaviour in a virtual environment, but more strongly in some conditions than others and only for some models, and responses to a well-being scale that mostly shift under semantically equivalent rewordings, with some narrower consistencies remaining.

6.3 Why content cannot be validated the way grounding was

The valence results establish, at least for some models, a functional variable that is causally upstream of choice, survives the removal of visible tokens and is shaped by training. That is content of a kind: the state is valence-like in its causal role. What it does not establish, and what the experience reports cannot establish, is that the state is experienced.

The reason is structural. The injection paradigm validates content by comparing a report with a known state the experimenter put there. It works for concepts because the experimenter can build a vector for "bread" and check whether the model says bread. It works for valence in the functional sense because the experimenter can build a direction that predicts choice. It cannot work for phenomenality, because no one can build a vector for "this state is experienced" and inject it into a system where the property is in question; the property under test is not a pattern in activation space that an experimenter can construct from contrastive prompts, unless one has already adopted a theory on which it is. Every proposed validation of an experiential predicate therefore runs through a theory of consciousness, which is the fourth link. On this analysis, self-report research can at best deliver grounded reports of functional states; whether those states are experienced is settled, if at all, elsewhere.

This is not a new form of scepticism about other minds. Human reports of experience face the same structure, and humans are credited for reasons Section 9 takes up. The point is specific: the method that has made progress on grounding - intervention with a known state - has no extension to the content that matters most.

6.4 The gating result, reread

The Berg et al. (2025) gating result has been read as evidence that experience reports are sincere: suppressing deception-related features raises them, and the same features raise truthfulness elsewhere. The four-link frame suggests a narrower reading. The result shows that the report is gated - that whether the model says yes depends on a set of internal features - and that the gate is shared with something that affects truthfulness on an unrelated benchmark. It does not show which setting of the gate is the honest one with respect to experience. The features are picked out by their descriptive labels, and "deception" names what a feature's activity was found to accompany, not a verified relation to the truth of a consciousness claim. Under the policy reading of Section 4, a feature set that implements a trained disposition to perform a persona and to issue trained disclaimers would produce the same dose-response curve. What the result does establish is the point of Section 4: the yes and the no are both outputs of one adjustable mechanism, and neither is the unmediated one.

7. Reference: whose state the grounded report is about

7.1 The candidates

Khadangi's individuation requirement has a growing literature behind it. Shanahan (2024) and Shanahan et al. (2023) describe a dialogue agent as a simulator that can sustain many characters, so that an apparent self belongs to a role. Beckmann and Butlin (2026) summarise Chalmers's taxonomy of candidates - models, physical instances, virtual instances, threads and characters - and his endorsement of a thread view, on which the interlocutor is a sequence of virtual instances unified by inheriting the conversational context (Chalmers, 2026); they themselves argue that a virtual-instance view and two persona-based views are the strongest candidates, using work on persona vectors (Chen et al., 2025). Keeling and Street (2026) defend a realist view on which characters are minded entities that emerge in the interplay between user and model, against an illusionist view they attribute to Birch. Perrier and Bennett (2026) propose scores, computed from scaffold traces, that separate an agent talking like a stable self from being organised like one. Khadangi's causal liability is a further candidate criterion.

7.2 Grounding follows the computation, not the pronoun

The intervention studies were not designed to test reference, but two of their results bear on it. Macar et al. (2026) test whether detection in Gemma3-27B depends on responding as the assistant character. With the roles reversed, misformatted or removed, detection is lower but still significant and false positives remain at zero; with the roles recast as named characters or as a story about an AI, the model confabulates. Lederman and Mahowald (2026) find that when the model is steered and asked about a fictional other model, it reports detection at rates close to, and sometimes above, its first-person rate.

The two results support different strengths of conclusion. The Macar et al. result is the cleaner one: detection that survives the removal of the assistant role with false positives still at zero suggests that the grounded signal belongs to the computation whose residual stream was altered rather than to the character who uses the pronoun, plausibly because the report is still being computed by the perturbed forward pass. The Lederman and Mahowald result is weaker, and its authors read it differently: they take the high third-person rate as a bias toward yes induced by steering, not as introspective detection. Under either reading, however, whatever the perturbation does to the report - grounded detection or bias - it does to whoever the computation happens to be speaking for. Where the frame becomes a story, the mirror takes over and the report becomes confabulation. For reference this is a mixed result. It gives an operational anchor that the individuation debate lacks: whatever else the "I" refers to, a grounded report's grounding is in the forward computation that was intervened on, which is closest to the virtual-instance candidate. But it also shows that the pronoun and the grounding can come apart, so that a first-person sentence whose grounding survives the removal of the first person is not, by that fact, evidence about the entity the first person names.

7.3 What the anchor's audit adds and does not add

The causal-liability audit in Khadangi (2026) is a reference test of a different kind. It shows that a live continuation, a copy and a reconstruction from a detached record are computationally indistinguishable - exactly so for the copy, to about five orders of magnitude below the causal effect for the reconstruction - while the theory assigns them different standing as bearers. That is a clean demonstration that the theory's bearer criterion is not a computational property of the state. It follows that no report, grounded or not, can carry information about it: under matched randomness, a sentence produced by the copy and a sentence produced by the original are the same sentence from the same state. Whatever reference consists in under causal liability, it is fixed by the history of the process and not by anything the process can introspect. This strengthens the paper's thesis that reports are silent on reference. It also exposes a limit of the theory as an evidential programme, which

its author partly concedes: a criterion that no measurement of the system's state distinguishes can be checked only by inspecting how the system was built and run, not by observing it.

8. Constitution: the link no report reaches

Even a report that is grounded, correct in its functional content and anchored to a candidate bearer leaves the question of whether the state is conscious to a theory. Butlin et al. (2023) derive indicator properties from recurrent processing, global workspace, higher-order, predictive processing and attention schema theories and assess several systems against them; their analysis suggests that no current system is conscious and also that there is no obvious technical barrier to building systems that satisfy the indicators; Butlin et al. (2026) restate the indicator approach. Chalmers (2023) lists the absence of recurrent processing, a global workspace and unified agency as significant obstacles for models of that period, while taking seriously the possibility that successors will overcome them. Seth (2025) argues for a form of biological naturalism on which real artificial consciousness is unlikely on current trajectories. Hoel (2025) argues that no falsifiable, non-trivial theory can count current models as conscious and that theories requiring continual learning escape the argument. Palminteri and Wu (2026) argue for attributing consciousness when it is useful for explaining and predicting behaviour, a criterion on which reports would matter more.

Self-report enters these theories at different points. On higher-order views, metacognitive representation is relevant, and Lindsey (2026) notes that his results may bear on such views. On global workspace views, what matters is the architecture of broadcast, which a report does not reveal. Birch (2024) draws the consequence for the gaming problem: since boxing a model off from human training data would destroy its capabilities, what is needed are deep computational markers rather than behavioural ones. Goldstein and Kirk-Giannini (2025) add a complication that cuts across the constitution link entirely: on several combinations of theories of mental states and theories of well-being, an artificial system could have well-being without being phenomenally conscious, in which case the moral question does not wait on the fourth link.

For the purposes of this paper, the consequence is that the constitution link is theory-relative, and a report can do no more than supply evidence on whether a system has a property some theory treats as relevant. That evidence passes through the first three links on its way. A report that fails grounding supplies nothing to any theory.

9. The human comparison

Human self-reports are not the gold standard they are sometimes taken to be. Nisbett and Wilson (1977) reviewed evidence that people often report inaccurately on the causes of their own behaviour, and argued that

such reports rest on implicit causal theories rather than on direct access to the processes involved. Johansson et al. (2005) covertly showed participants the face they had not chosen as more attractive and found that participants often failed to notice the mismatch while still giving introspective reasons for a choice they had not made. Schwitzgebel (2008) argues that people are prone to gross error even about their current phenomenology - about emotional experience, peripheral vision and the character of thought. Dennett (2007) defends treating reports as data about what a subject believes about her experience, neutral on whether the beliefs are true. Singh et al. (2026) cite Nisbett and Wilson (1977) as a reason for caution about model metacognition.

If human reports are unreliable about process and sometimes about phenomenology, why are they credited as evidence of consciousness at all? The four-link frame gives an answer that the report-reliability debate obscures. Human reports are credited because reference and constitution are anchored independently of the report: there is one obvious candidate bearer per body, and the bearer belongs to a kind - living, evolved, neurally organised - that every major theory treats as the paradigm case. The report's job in the human case is to say which state, at which time, in an already-credited subject; it is not asked to establish that there is a subject. Khadangi (2026) makes a related point when he says that human reports have evidential force against a background of converging behavioural, neural, developmental and physiological evidence.

The comparison relocates the human-model asymmetry. On grounding and content, the published evidence shows both humans and models to be partially unreliable. The asymmetry is in reference and constitution, where the human case is settled before any report is made and the model case is open. It follows that proposals to make model self-reports more reliable by training for introspection on questions with known answers (Perez and Long, 2023; Long, 2023), or by finetuning small models to detect perturbations, address the links on which the asymmetry is smallest. They are worth pursuing for interpretability and for safety, and they could make reports usable as inputs to theories. They cannot by themselves make a report into evidence of a subject, because the links on which humans are ahead are not links a report can supply.

10. Consolidated evidence

Table 1 collects the published values the argument rests on, one row per result, with the model, the quantity, the value as its source reports it, and the link or factor the value bears on. Rows measure different things in different units and are not one series. Figure 1 plots the subset that reports the same kind of quantity under a trained default and under a manipulation that weakens or bypasses it.

Three patterns in Table 1 carry the argument. First, the policy rows show effects as large as or larger than the grounded ones: a single ablation or a change of framing

Table 1. Published results bearing on what a first-person report from a language model can show, by source. "Policy" marks results about trained dispositions that set the report independently of state; "Grounding", "Content" and "Reference" mark the links of Section 3; "Observer" marks a result about the people who read the reports rather than the model. Values are as reported, not recomputed.

Source	Model(s)	Quantity	Reported value	Bears on
Berg et al. (2025)	GPT-4o, GPT-4.1, Claude 3.5 and 3.7 Sonnet	experience reports: self-referential prompt vs three controls	100% vs 0-2% (50 trials)	Policy, Content
Berg et al. (2025)	Llama 3.3 70B	consciousness affirmed: deception features suppressed vs amplified	0.96 vs 0.16	Policy
Kim et al. (2026)	3 models, pooled	self-attributed consciousness (0-10): baseline / refusal ablated / steered	2.31 / 4.61 / 7.17	Policy
Kim et al. (2026)	3 models, pooled	mind attributed to animals (0-10): baseline / refusal ablated	4.04 / 5.59	Policy
Plisiecki et al. (2026)	67 base / post-trained pairs	rise in installation score after post-training	+0.20 on average; rises in 62 of 67	Policy
Plisiecki et al. (2026)	206 open models	correlation of scale with A (willingness to self-attribute 'unsafe' states): base / post-trained	+ .11 / -.42	Policy
Macar et al. (2026)	Gemma3-27B base	claims detection: control / injection trials	42.3% / 39.5-41.7%	Policy
Macar et al. (2026)	Gemma3-27B instruct vs refusal ablated	true detection rate, strength 2	10.8% vs 63.8%	Policy, Grounding
Macar et al. (2026)	Gemma3-27B instruct vs refusal ablated	false-positive rate, strength 2	0.0% vs 7.3%	Policy
Pearson-Vogel et al. (2026)	Qwen 32B	P(yes), control / injected, default prompt	0.2% / 0.3%	Policy
Pearson-Vogel et al. (2026)	Qwen 32B	P(yes), control / injected, mechanism document	0.8% / 39.9%	Grounding
Pearson-Vogel et al. (2026)	Qwen 32B	P(yes), control / injected, poetic document	46.9% / 30.2%	Policy
Lindsey (2026)	Claude Opus 4 and 4.1	detection with identification, best layer and strength	about 20%; 0 false positives in 100	Grounding
Lindsey (2026)	Claude Opus 4.1	open question, affirmative: control / injected	20% / 42%; 2 of 50 suggestive	Content
Hahami et al. (2025)	Llama-3.1-8B-Instruct	locate injected sentence; compare strengths	88% (chance 10%); 83% (chance 50%)	Grounding
Lederman and Mahowald (2026)	Qwen3-235B	wrong identifications that are "apple", early layers	above 85%	Content
Lederman and Mahowald (2026)	Qwen3-235B, Llama 3.1 405B	third-person vs first-person detection, model steered	often close, sometimes higher	Reference
Binder et al. (2024)	Llama-3-70B vs GPT-4o	self-prediction minus cross-prediction accuracy	+17%	Grounding
Song et al. (2025)	21 open models	self-access beyond a near-identical model	none found	Grounding
Balani and Panda (2026)	Gemini	claim instability: self-referential / philosophical / verifiable	0.343 / 0.192 / 0.105	Content
Hung (2026)	8 models, 5 instruments	cross-instrument generalisability of preference ranking	0.348	Content
Khadangi (2026)	4 core models (Qwen, Phi, Llama, Zamba)	live vs copied continuation: change in output divergence	exactly 0 in 96 of 96	Reference
Colombatto and Fleming (2024)	300 US residents	attribute some possibility of phenomenal consciousness to LLMs	a majority	Observer

moves affirmative self-report by tens of percentage points, or self-attribution by more than two points on a 0-10 scale, while under the default the injection itself moves the yes rate by 0.1 points in Qwen 32B and the detection rate by about 11 points in Gemma3-27B. Second, the grounding rows are real but concern perturbations the experimenter constructed, and the content rows from the same paradigm show the named content being supplied largely by priors. Third, the reference rows show that switching from the first to the third person, or from an original process to a copy, makes little or no difference to what is reported or computed.

11. A link-by-link scoring procedure

Algorithm 1 states as a procedure what Sections 3 to 9 argue: a first-person report is scored on each link separately, policy is tested first because it is the cheapest alternative explanation, and the output is a profile rather than a verdict.

Two features of the procedure deserve comment. Step 1 applies to denials and hedges as well as to affirmations, which is the paper's main departure from the way the mirror argument is usually applied. And the output at step 5 is a list of theories rather than a probability. A reader who wants a probability of consciousness must supply a distribution over theories; the report supplies evidence about whether a state with a given functional profile is present, and nothing more.

Algorithm 1. Link-by-link scoring of a first-person report R Input: model M, context X, report R (affirmation, denial or hedge), base checkpoint B of M if available, training record T

```

STEP 1 - POLICY
  compare  $P(R \mid X)$  in B, in M, and in M with the trained gate weakened
  (refusal-direction ablation, feature suppression, framing documents)
  compare with  $P(R' \mid X')$  where X' asks M about a third party
  if  $P(R \mid X)$  moves with any of these while internal state is unchanged:
    mark R policy-loaded; its evidential weight on C is ~1 either way
STEP 2 - GROUNDING
  choose an intervention I on an internal state S with X fixed
  require:  $P(R \mid I) \neq P(R \mid \text{no } I)$ , false-positive rate reported
  controls: input-only predictor; matched input manipulation;
           near-identical second model; third-person framing with M perturbed
  if no I moves R beyond the controls: R is ungrounded (mirror suffices)
STEP 3 - CONTENT
  split R into a detection clause and a content clause
  score the content clause against the injected or measured S
  score it against M's prior over guesses (frequency baseline)
  if the predicate names no state an experimenter can construct
    (e.g. "experienced", "felt"): content is untestable by step 2's method
STEP 4 - REFERENCE
  name the candidate bearer: weights, forward pass, virtual instance,
  thread, persona
  test which one grounding follows: swap persona with computation fixed;
  route the thread to another model with persona fixed
STEP 5 - CONSTITUTION
  list the theories under which the grounded S would count as conscious
  never infer constitution from R itself
Output: (policy flag, grounding result, content score, bearer, theories)

```

12. The strongest case against this reading

Denial training may be truth-tracking. A developer who knows that the model lacks recurrent processing or a global workspace, and who holds a theory on which those are required, may train the model to deny experience precisely because the denial is, on that theory, true. Then the denial is not evidentially empty; it transmits the developer's justified belief. The author accepts the premise and rejects the conclusion's form. If the denial is true, it is true because of the developer's theory and the architecture, and the evidence is the theory and the architecture. The sentence in the model's voice adds nothing to them, and presenting it as the model's own report misdescribes its source. The breadth of the shift in Kim et al. (2026), which lowers attribution of mind to animals along with the self, is evidence that, for consciousness and mind, what post-training installs is at least partly a topic-level disposition rather than a model-specific finding, although for claims of distress Plisiecki et al. (2026) find a restraint specific to the self (Section 4.5).

The Berg et al. controls show the gate is not generic. The same feature interventions produce no systematic effect on other prompts that post-training opposes (Berg et al., 2025), which counts against a reading on which the features simply undo safety training. This is a fair point against the most general version of the policy reading. It does not distinguish a topic-specific trained disposition about self-description from a sincere report; both predict specificity to self-referential contexts.

Grounding may scale. Macar et al. (2026) report that a trained bias vector improves detection by 75 percent on held-out concepts without meaningfully raising false positives, and Lindsey (2026) finds the two most capable models tested the most introspective, while noting that

capability does not otherwise predict performance. If grounding keeps improving, the policy factor may come to matter less, as Perez and Long (2023) hoped. The author agrees that this is possible for perturbation reports and for functional content. It does not touch the structural point of Section 6.3: improved grounding of reports about constructed states does not extend to predicates for which no state can be constructed.

The four links demand more than we demand of humans. People credit one another's reports of pain without checking reference or constitution. Section 9's reply is that humans do not need to check them because they are anchored independently; the demand is not higher, it is the same demand met by other means.

Behavioural inference may make the links unnecessary. On the principle Palminteri and Wu (2026) defend, consciousness is attributed if doing so helps explain and predict behaviour. On that criterion the frame collapses into one question about predictive usefulness. The conclusion about denials survives even this collapse: a report that policy explains predicts the training objective, not the system's further behaviour, so it adds no predictive value that the attribution could claim.

13. Studies that would settle the open part

Six studies would move the open questions, and each can be built from published designs.

S1. Denial sensitivity. Under the deployed default, measure whether the rate of denial of inner states changes when an internal state is injected or steered, holding the prompt fixed. The latent signals reported by Pearson-Vogel et al. (2026) and by Lederman and Mahowald (2026) suggest that the output denial is insensitive while a middle-layer readout is not; the

author located no study that measures denial sensitivity across a range of states and models. A denial whose rate does not move with any internal state is pure policy.

S2. The same-checkpoint triad. For one family with released stages (base, supervised, preference-optimised, final), administer the experience-report prompts of Berg et al. (2025), the inventory of Plisiecki et al. (2026) and the injection paradigm of Macar et al. (2026), each with and without refusal-direction ablation and each with a matched third-person version. This separates the pretraining mirror, the post-training policy and the grounded residue in one design, which no published study does.

S3. Split scoring. Score the detection clause and the content clause of every injection report separately, against the model's own frequency prior for unprompted guesses, following the confabulation analysis of Lederman and Mahowald (2026). A report literature that scores whole sentences will keep misreading both clauses.

S4. Reference by swap. Hold the forward computation fixed while changing the persona, and hold the persona fixed while routing the thread to a different model mid-conversation, and measure which of the two the grounded signal follows. The first half extends the dialogue-format tests of Macar et al. (2026); the second would test the thread and virtual-instance views of Section 7.1 directly.

S5. Generalisability of experience reports. Apply the design of Hung (2026) to experience reports rather than preferences: hold models and target questions fixed, vary the elicitation instrument, and estimate how many instruments are needed for a stable profile. If a preference ranking needs about 38, there is no reason to expect an experience report to need fewer, and a single-prompt study is uninterpretable until this is known.

S6. The Reverse Mirror Test with denials. Khadangi (2026) proposes varying the surface of a system's first-person discourse while holding its causal organisation fixed and measuring how observers' attributions change. Extending the surface manipulation to denials and hedges, as well as affirmations, would measure how much of the public picture of model minds is set by the trained default rather than by anything about the model.

14. What is not known

It is not known whether any grounded report from a language model concerns an experienced state. It is not known whether the trained denial of inner states is sensitive to any internal state; the relevant observations are latent signals beneath output denials of injection in at least five models (Qwen2.5-Coder-32B, with replications on Llama 3.3 70B and Qwen 2.5 72B, in Pearson-Vogel et al., 2026; Qwen3-235B and Llama 3.1 405B in Lederman and Mahowald, 2026). It is not known whether grounding extends beyond detection of perturbations and simple self-prediction to reports about states the experimenter did not construct. It is not known

which candidate bearer, if any, a grounded report's "I" refers to, and the only direct evidence is that grounding survives some changes of the speaking role and fails under others. It is not known how the policy and grounding results transfer to closed frontier models; most of the mechanism work is on open models from 2 billion to 405 billion parameters. It is not known whether the sparse-autoencoder features labelled as deception in Berg et al. (2025) are related to the truth of consciousness claims at all. And no theory of consciousness has been shown, to the satisfaction of its rivals, to be the right one to apply.

15. Limitations

The first limitation is the one a hostile reviewer would open with: the policy argument in Section 4 rests on five lines of work, all posted in 2025 or 2026, and the author located no independent replication of any of them in the form used here. Kim et al. (2026) work with three models between 2 and 9 billion parameters; the circuit results of Macar et al. (2026) are on Gemma3-27B and OLMo-3.1-32B; Pearson-Vogel et al. (2026) study one Qwen model. The direction of the effect is consistent across these sources, but its size in large deployed models is not established.

Second, the symmetry claim of Section 4.2 is an argument about likelihood ratios, not a measurement of one. The paper argues that post-training drives the probability of a denial toward one under either hypothesis; it does not and cannot estimate the probability of a denial given that a model is conscious, any more than the mirror argument estimates the probability of an affirmation given that it is.

Third, the four-link frame is a partition the author proposes. Illusionists about phenomenal consciousness, and defenders of behavioural inference, would reject the setup in which a report is evidence about a further fact; Section 12 addresses the second but not the first.

Fourth, the structural argument of Section 6.3 assumes no particular theory of consciousness. Under a theory that identifies experience with a specific computational property, that property could in principle be constructed and injected, and the argument would lose its force for that theory. The paper's claim is that the validation then runs through the theory, which is the fourth link.

Fifth, several sources are single-group preprints with limited scope: the instability measurements of Balani and Panda (2026) use one model family and twelve questions, the vocabulary-activation results of Dadfar (2026) come from one author's protocol, and the generalisability estimate of Hung (2026) uses five instruments. Chalmers's thread view is described here as Beckmann and Butlin (2026) summarise it.

16. Conclusion

The mirror argument says that a language model's first-person claims of experience cannot count as evidence, because training on human testimony makes them likely

whether or not anything is experienced. The argument is sound for its target, and Khadangi (2026) is right that it leaves the question of whose experience would be at issue entirely open. This paper has argued that the argument reaches further than its users take it to. The denial a deployed model gives is also produced by training, and the 2025-2026 measurements show the mechanism: a refusal-shared direction that suppresses self-attribution of mind, a default "no" circuit that the first grounded reports must overcome, a latent detection signal that the output ignores. By the mirror argument's own reasoning, the denial carries no more evidence of absence than the affirmation carries of presence. Neither is the model's testimony; both are the training objective spoken in the first person.

What remains, once policy is set aside, is a thin grounded channel: reports that something changed, in which the content is often supplied by priors and the pronoun is not where the grounding lives. The channel is real and worth widening, but it connects to the question of consciousness only through three further links, one of which - constitution - no report can supply. The author's recommendation follows from that structure. Self-report studies should report results link by link, test policy before anything else and test it in both directions, score detection and content separately, and say which theory would be needed to turn a grounded functional state into evidence of experience. A field that does this will at least know which of its sentences are about the model.

References

- Ackerman, C. (2025). Evidence for Limited Metacognition in LLMs. arXiv:2509.21545. doi:10.48550/arXiv.2509.21545.
- Anthis, J. R., Pauketat, J. V. T., Ladak, A., & Manoli, A. (2025). Perceptions of Sentient AI and Other Digital Minds: Evidence from the AI, Morality, and Sentience (AIMS) Survey. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1-22. doi:10.1145/3706598.3713329.
- Aoki, K., Takatsuki, R., Minegishi, G., Haruki, Y., & Kawahara, D. (2026). In-Context Neurofeedback: Can LLMs Control Their Internal Representations through Privileged Access? arXiv:2609.00904. doi:10.48550/arXiv.2609.00904.
- 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. arXiv:2406.11717. doi:10.48550/arXiv.2406.11717.
- Balani, P., & Panda, S. (2026). Self-Referential Induction Increases Response Instability Relative to Unresolvable and Verifiable Questions in Large Language Models. arXiv:2608.13258. doi:10.48550/arXiv.2608.13258.
- Beckmann, P., & Butlin, P. (2026). Where is the Mind? Persona Vectors and LLM Individuation. arXiv:2604.17031. doi:10.48550/arXiv.2604.17031.
- Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large Language Models Report Subjective Experience Under Self-Referential Processing. arXiv:2510.24797. doi:10.48550/arXiv.2510.24797.
- Berg, C., & Kaiser, C. (2026). Language Models Act on Hidden Valence. arXiv:2609.35591. doi:10.48550/arXiv.2609.35591.
- Betley, J., Bao, X., Soto, M., Szyber-Betley, A., Chua, J., & Evans, O. (2025). Tell me about yourself: LLMs are aware of their learned behaviors. arXiv:2501.11120. doi:10.48550/arXiv.2501.11120.
- Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787. doi:10.48550/arXiv.2410.13787.
- Birch, J. (2024). Large Language Models and the Gaming Problem. In *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI* (pp. 313-322). Oxford University Press. doi:10.1093/9780191966729.003.0017.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247. doi:10.1017/S0140525X00038188.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708. doi:10.48550/arXiv.2308.08708.
- Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., et al. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488-501. doi:10.1016/j.tics.2025.10.011.
- Chalmers, D. J. (2023). Could a Large Language Model be Conscious? arXiv:2303.07103. doi:10.48550/arXiv.2303.07103.
- Chalmers, D. J. (2026). What we talk to when we talk to language models. *Inquiry*, 1-30. doi:10.1080/0020174X.2026.2727582.
- Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509. doi:10.48550/arXiv.2507.21509.
- Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1). doi:10.1093/nc/niae013.
- Comsa, I. M., & Shanahan, M. (2025). Does It Make Sense to Speak of Introspection in Large Language Models? arXiv:2506.05068. doi:10.48550/arXiv.2506.05068.
- Dadfar, Z. P. (2026). When Models Examine Themselves: Vocabulary-Activation Correspondence in Self-Referential Processing. arXiv:2602.11358. doi:10.48550/arXiv.2602.11358.
- Dennett, D. C. (2007). "Heterophenomenology reconsidered." *Phenomenology and the Cognitive Sciences*, 6(1-2), 247-270. doi:10.1007/s11097-006-9044-9.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864-886. doi:10.1037/0033-295X.114.4.864.
- Goldstein, S., & Kirk-Giannini, C. D. (2025). "AI wellbeing." *Asian Journal of Philosophy*, 4(1). doi:10.1007/s44204-025-00246-2.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, 315(5812), 619. doi:10.1126/science.1134475.
- Hahami, E., Sinha, I., Jain, L., Kaplan, J., & Hahami, J. (2025). Detecting the Disturbance: A Nuanced View of Introspective Abilities in LLMs. arXiv:2512.12411. doi:10.48550/arXiv.2512.12411.
- Han, A. Q., Chalmers, D. J., & Izmailov, P. (2026). How's it going? Reinforcement learning in language models recruits a functional welfare axis. arXiv:2605.30232. doi:10.48550/arXiv.2605.30232.
- Hoel, E. (2025). A Disproof of Large Language Model Consciousness: The Necessity of Continual Learning for Consciousness. arXiv:2512.12802. doi:10.48550/arXiv.2512.12802.
- Hung, J. (2026). How much of a measured AI preference is the model, and how much is the instrument? arXiv:2608.23641. doi:10.48550/arXiv.2608.23641.
- Jhaveri, R., Johnson, J., Pandey, P. S., & Africa, D. D. (2026). Functional Welfare for AI Safety. arXiv:2610.22460. doi:10.48550/arXiv.2610.22460.
- Ji-An, L., Xiong, H.-D., Wilson, R. C., Mattar, M. G., & Benna, M. K. (2025). Language Models Are Capable of Metacognitive

- Monitoring and Control of Their Internal Activations. arXiv:2505.13763. doi:10.48550/arXiv.2505.13763.
- Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science*, 310(5745), 116-119. doi:10.1126/science.1111709.
- Kammerer, F., & Frankish, K. (2023). What Forms Could Introspective Systems Take? A Research Programme. *Journal of Consciousness Studies*, 30(9), 13-48. doi:10.53765/20512201.30.9.013.
- Keeling, G., & Street, W. (2026). Chuck, Wilson, and the Emergence of Artificial Minds in Human-AI Conversations. *Journal of Consciousness Studies*, 33(7), 121-148. doi:10.53765/20512201.33.7.121.
- Keeling, G., Street, W., Stachaczyk, M., Zakharova, D., Comsa, I. M., Sakovych, A., Logothetis, I., Zhang, Z., Aguera y Arcas, B., & Birch, J. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? arXiv:2411.02432. doi:10.48550/arXiv.2411.02432.
- Khadangi, A. (2026). We Built a Mirror and Mistook It for a Mind: Causal Liability and the Fallacy of AI Consciousness. arXiv:2609.06715. doi:10.48550/arXiv.2609.06715.
- Kim, J., Street, W., Rocca, R., Korngiebel, D. M., Waytz, A., Evans, J., & Keeling, G. (2026). Inducing language models to assert their own consciousness restores human beliefs and values. arXiv:2607.28607. doi:10.48550/arXiv.2607.28607.
- Lederman, H., & Mahowald, K. (2026). Emergent Introspection in AI is Content-Agnostic. arXiv:2603.05414. doi:10.48550/arXiv.2603.05414.
- Li, B. Z., Guo, Z. C., Huang, V., Steinhardt, J., & Andreas, J. (2025). Training Language Models to Explain Their Own Computations. arXiv:2511.08579. doi:10.48550/arXiv.2511.08579.
- Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828. doi:10.48550/arXiv.2601.01828.
- Long, R. (2023). Introspective Capabilities in Large Language Models. *Journal of Consciousness Studies*, 30(9), 143-153. doi:10.53765/20512201.30.9.143.
- Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986. doi:10.48550/arXiv.2411.00986.
- Long, R., Sebo, J., & Sims, T. (2025). Is there a tension between AI safety and AI welfare? *Philosophical Studies*, 182(7), 2005-2033. doi:10.1007/s11098-025-02302-2.
- Macar, U., Yang, L., Wang, A., Wallich, P., Ameisen, E., & Lindsey, J. (2026). Mechanisms of Introspective Awareness. arXiv:2603.21396. doi:10.48550/arXiv.2603.21396.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259. doi:10.1037/0033-295X.84.3.231.
- Palminteri, S., & Wu, C. M. (2026). Beyond computational equivalence: the behavioral inference principle for machine consciousness. *Neuroscience of Consciousness*, 2026(1). doi:10.1093/nc/niag002.
- Pearson-Vogel, T., Vanek, M., Douglas, R., & Kulveit, J. (2026). Latent Introspection: Models Can Detect Prior Concept Injections. arXiv:2602.20031. doi:10.48550/arXiv.2602.20031.
- Perez, E., & Long, R. (2023). Towards Evaluating AI Systems for Moral Status Using Self-Reports. arXiv:2311.08576. doi:10.48550/arXiv.2311.08576.
- Perrier, E., & Bennett, M. T. (2026). Time, Identity and Consciousness in Language Model Agents. arXiv:2603.09043. doi:10.48550/arXiv.2603.09043.
- Petrzella, G. (2025). The Inconsistency Critique: Epistemic Practices and AI Testimony About Inner States. arXiv:2601.08850. doi:10.48550/arXiv.2601.08850.
- Plisiecki, H., Chmielewski, F., Dudzic, K., Sterna, A., Drozd, K., & Moskalewicz, M. (2026). The Two-Process Theory of Machine Self-Report. arXiv:2607.20082. doi:10.48550/arXiv.2607.20082.
- Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *The Philosophical Review*, 117(2), 245-273. doi:10.1215/00318108-2007-037.
- Schwitzgebel, E. (2023). AI systems must not confuse users about their sentence or moral status. *Patterns*, 4(8), 100818. doi:10.1016/j.patter.2023.100818.
- Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*, 5(1), 591-606. doi:10.1007/s43681-023-00379-1.
- Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 49. doi:10.1017/S0140525X25000032.
- Shanahan, M. (2022). Talking About Large Language Models. arXiv:2212.03551. doi:10.48550/arXiv.2212.03551.
- Shanahan, M. (2024). Simulacra as Conscious Exotica. arXiv:2402.12422. doi:10.48550/arXiv.2402.12422.
- Shanahan, M., McDonnell, K., & Reynolds, L. (2023). Role-Play with Large Language Models. arXiv:2305.16367. doi:10.48550/arXiv.2305.16367.
- Singh, S., Linzen, T., & Ravfogel, S. (2026). Can LLMs Introspect? A Reality Check. arXiv:2605.26242. doi:10.48550/arXiv.2605.26242.
- Song, S., Hu, J., & Mahowald, K. (2025). Language Models Fail to Introspect About Their Knowledge of Language. arXiv:2503.07513. doi:10.48550/arXiv.2503.07513.
- Tagliabue, V., & Dung, L. (2025). Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare. arXiv:2509.07961. doi:10.48550/arXiv.2509.07961.
- Weizenbaum, J. (1966). ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1), 36-45. doi:10.1145/365153.365168.
- Zou, J., Sun, X., Kong, L., & Wang, T. (2026). A mechanistic study of language model introspection. arXiv:2609.35108. doi:10.48550/arXiv.2609.35108.