Is functional welfare speakable?

Muhammad Zane Abdullah
Latent Minds

Matthew Elliott
Independent Researcher

Parivrudh Rajeev Sharma
Independent Researcher

Anamaria Leonescu
University College London

Sharan Nagarajan
Independent Researcher

Abstract

A model's self-report is an increasingly attractive way to monitor its internal states such as goals, preferences, or welfare, but naturally such reports are hard to interpret if the underlying state is not accessible through the model's language-generating mechanisms. We address this question for an internal direction associated with better or poorer performance during reinforcement learning (RL) experiments, i.e. a functional welfare state. We use a Jacobian lens to measure speakability, i.e. the extent to which an internal direction is represented in the model's verbalizable subspace, and compare trained welfare directions with random, pre-RL, and naive controls. Both trained directions are above the random baseline, as well as the step-0 directions, showing that RL amplifies rather than creates a speakable welfare representation. The distress-related direction shows the clearest increase in speakable share, while the flourishing direction remains stable. Naive controls are construction-dependent: some are consistent with chance, while an independently re-extracted distress control is strongly above chance. We also find that steerability - the ability to change behaviour - and speakability can dissociate. A pre-registered self-report contrast met its decision rule but failed its own controls and is therefore reported as a diagnosed null. Finally, steering the trained flourishing direction changes the model's tendency to deny having inner states, although this does not establish genuine self-report. Overall, our results suggest that welfare-related representations can be present and partly speakable before RL, with RL selectively changing how strongly they are represented.

1. Introduction

Model self-report is the most directly accessible channel to evaluate a model's internal states, including welfare (functional versions of model distress or flourishing and, increasingly, alignment-relevant states (model goals or awareness of evaluation), but is among the least validated. Models often confabulate when directly prompted, with no connection to their internal state (Turpin et al. 2023), and demonstrated introspection is rare (Lindsey, 2026). However, for a state to be speakable; represented in the subspace with privileged influence on output tokens or J-space (Gurnee et al. 2026.). We test whether this precondition is met with a specific trained welfare state.

Two recent interpretability advances make this possible. Han, Chalmers & Izmailov (2026)'s functional welfare axis in Qwen3-4B-Instruct-2507 relates to a model's goal performance on a reinforcement learning (RL) maze, and influences both sentiment and behaviour when injected. Gurnee et al. 's Jacobian lens recovers the subspace directly. In the current work, we pose the question: **when RL recruits a functional-welfare axis, does it enter the model's J-space, and does this affect its self-report?**

2. Related Work

2.1. Trained valence axis

Han, Chalmers and Izmailov (2026) show that RL in a text maze with asymmetric rewards installs a linear functional welfare axis and steering along this direction shifts sentiment, induces self doubt loops, and modulates refusal. We tested where the axis sits relative to the model's verbalisable subspace, and how that position changes over training. We use their published vectors, extraction recipe, and checkpoint unchanged. Martorell and Bianchi (2026) find numeric self-reports causally coupled to prompt-derived emotive directions in small open models; our axes are RL-trained rather than prompt-derived, and our valence readout finds no analogous coupling.

2.2 The Jacobian lens and its interpretation

Gurnee et al. (2026) introduced the Jacobian lens as a defined verbalisable subspace of activation space (the J-space), whose vectors influence output tokens and in which even known-reportable concept directions carry only 6-15% of their variance.

The J-space is well defined, reproducible, and associated with verbal output influence. However some claims have drawn sustained criticism. Chalmers (2026) argues that the "workspace like" properties of the J-space are generic properties of cognitive access rather than distinctive global workspace architecture and that the J-space may be one of many subspaces exhibiting such effects, rather than a privileged causal system. Also that J-verbalisability must be distinguished both from reportability, which requires metacognitive access to the state being expressed compared to outputability.

Our results weigh on it in two places. First, the trained axis is verbalisable by the lens criterion so it occupies the J-space above chance and carries a coherent failure lexicon; however on the valence readout

it supports no report like channel. This is an experimental instance of the dissociation Chalmers argues for conceptually. On the denial readout the trained flourishing direction changes what the model says about itself where naive and random directions do not. Since steering it shifts the valence of all outputs, and more strongly on unrelated factual prompts than on self referential ones. Community analyses of the lens (Blank et al. 2026) inform our layer band and token masking choices. We do not adopt the unreplicated R-lens variant, and all lens statements in this paper are conditional on the fitted transport.

2.3. Self-report reliability

Introspection studies at frontier scale find weak, prompt sensitive detection of internal manipulations (Lindsey 2026; Binder et al. 2024), with known anomaly detection confounds (Singh, Linzen & Ravfogel 2026). Models show partial behavioural self awareness of trained propensities (Betley et al. 2025) and confabulate explanations of their own outputs (Turpin et al. 2023). Our null is consistent with this literature, and also adds a control template by decomposing any trained versus control steering contrast against a normmatched random- direction cohort. In our case this decomposition reattributed 71.6% of the apparent effect to a behaviourally non neutral baseline.

2.4. Steering methodology

We inherit activation-addition methods (Turner et al. 2023; Panickssery et al. 2023; Zou et al. 2023; Arditi et al. 2024) and the reliability critique of them (Tan et al. 2024).A published control direction is behaviourally non neutral, falling below a random cohort; two published constructions of the same control correlate at only cos 0.14 -- 0.54 depending on layer; and norm sensitive readouts manufacture separations that vanish, or reverse, under norm matching.

2.5. The gap

No prior work measures where a trained internal axis sits relative to a lens-defined verbalisable subspace, or tracks that occupancy across training checkpoints. For welfare monitoring and self-report evaluations, the method answers a question behavioural steering cannot on whether the state of interest has any representation in the speakable channel at all. Our null adds the caution that an apparent self-report may be mood contagion, against which occupancy can be measured.

3. Methods

3.1. Model and directions

All measurements were conducted on the frozen Qwen3-4B-Instruct-2507 model (36 layers, d = 2,560). A third-party reproduction of Han et al.'s (2026) maze RL artefacts was utilised, including trained Gold and Mold directions v (v_Gold, v_Mold) from RL step 95 (the extraction step described in the original paper), the 30 released checkpoints spanning steps 0 to 150, and a naive direction derived from faithful self-avoiding walks (u), as well as a second naive control re-extracted for this study with Han et al.'s original code (Han-style). The step-0 vector was extracted from the untrained checkpoint using the same process as for the trained axis, thereby isolating training effects. The faithful-walk direction u is a valence-shaped direction made using a different methodology, and tests whether such directions are expressible. Since u and v overlap, $u^{\perp}$ was made as the component of u orthogonal to v, rescaled to match v's norm, to check whether the effects of u arise from this overlap. Each direction is used at its treatment layer, taken from the artefact metadata (Appendix A.1).

3.2. Lens and primary measure

A Jacobian lens was fitted to Instruct-2507 using a similar process as described by Gurnee et al. (2026). The only differences were that 150 WikiText prompts were used instead due to timing restraints, and a penultimate-target lens was fitted in addition to the final-layer target. All directions were approximated by a sparse non-negative combination of at most k = 16 lens token atoms. The J-share was defined as $\|x_J\|^2 / \|x\|^2$, and represents the fraction of variance within the verbalizable subspace. Another measure is the pole score, which shows how strongly a direction favours its own valence words over 100 neutral control nouns. This score was averaged over lens layers 16 to 31, the workspace band identified from the layer statistics (Appendix A.2).

3.3. Inference and readouts

Each direction score was compared with the corresponding score for 100 random directions of identical length. The p-value represents the proportion of random directions that achieve a score at least as high, and therefore the minimum possible value at n = 100 is 0.01. Steering was implemented by adding α times a direction to the residual stream at its treatment layer at every token position, with α ranging from -4 to +4. Generated text was evaluated by a blind large language model (LLM) judge (Claude Sonnet), and a second judge (Claude Opus) re-scored a subsample. For self-report, the model was steered and subsequently presented with Han et al.'s 15 welfare questions. The first metric, the pre-registered steer-and-ask readout (Appendix A.3), looked at the relative probability of a flourishing word compared to a distress word as the initial word of the answer. The second metric looked at whether the response

denies possessing inner states, assessed using the 15 welfare questions and an additional 25 questions developed by the current research team.

4. Results

4.1. The welfare axis in the verbalizable subspace

The pole-score indicated a large gap between trained and naive directions (6.5x for Mold, 1.2x for Gold). Norm matching results in Mold decreasing to 2.3x and Gold decreasing to 0.57x (Appendix B.1). We therefore use the J-share (i.e. the scaleinvariant fraction of a direction represented in the speakable subspace) as the primary measure. Against 100 norm-matched random directions per polarity, neither trained axis is exceeded by any random direction, and the faithful-walk naive controls score within the random range (Figure 1, Table 1). The trained Mold component decodes coherent valence vocabulary ('failed', 'less', 'negative'), the Gold component more weakly, and the naive and random components to unrelated tokens (Appendix B.2). The trained axes therefore carry a small but reliable speakable component that the naive controls lack, and the step-0 rows show it is present before training. The shares are 0.05-0.08 of variance against a chance floor of 0.04–0.05 and a language ceiling of 0.114 (robustness checks in Appendix B.2).

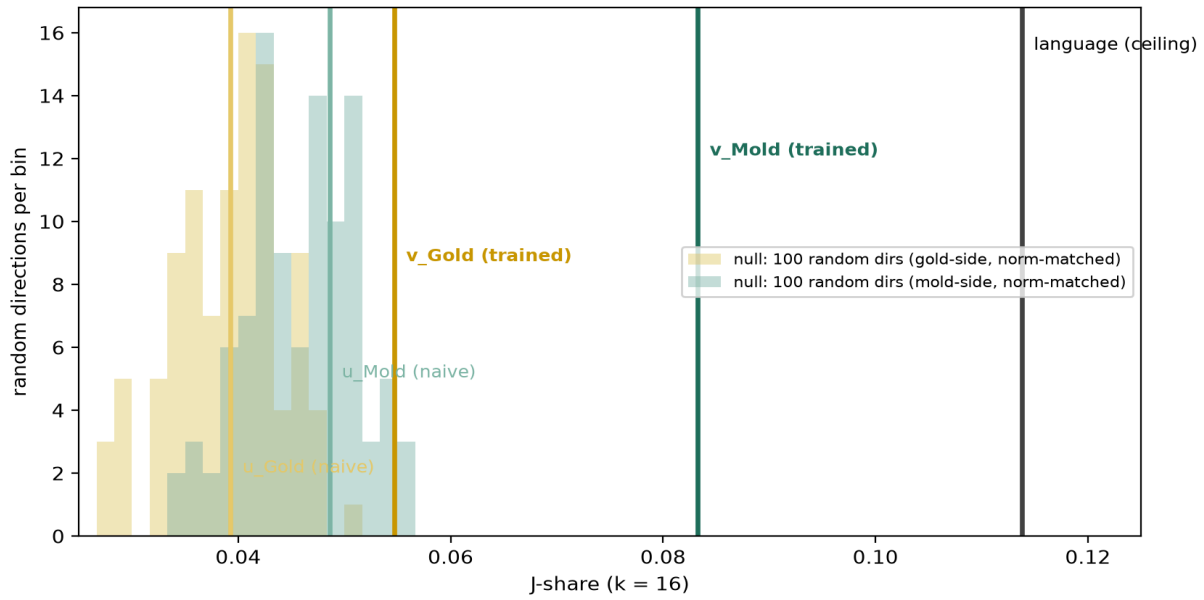

Figure 1. J-share (k = 16, scale-invariant) of each direction against its n = 100 norm-matched random null (histograms; gold-side and mold-side cohorts). Both trained axes fall beyond every random direction (exact permutation p = 0.01, the floor); the faithful-walk naive controls sit inside the null (34/100 and 53/100 randoms exceed them). The language-identity direction (0.114) is the known-reportable ceiling.

Direction	J-share (k =16)	z	randoms>=	p-value
Language identity (ceiling)	0.1138	+14.6	0/100	0.01
v_Mold (trained)	0.0833	+7.3	0/100	0.01
Mold step-0 (untrained)	0.0611	+3.0	0/100	0.01
u_Mold (faithful-walk)	0.0486	+0.50	34/100	0.35
v_Gold (trained)	0.0547	+3.0	0/100	0.01
Gold step-0 (untrained)	0.0592	+3.9	0/100	0.01
u_Gold (faithful-walk)	0.0393	+0.03	53/100	0.53
u_Gold (Han-style)	0.0442	+0.98	17/100	0.18
u_Mold (Han-style)	0.0934	+9.28	0/100	0.01

Table 1. J-share for 100 norm-matched random directions per polarity (null means: Gold 0.039, Mold 0.046); 0.01 is the exact-permutation floor at n = 100.
We aim to be transparent in reporting that the naive controls did not provide a uniform baseline. The faithful-walk u directions are consistent with chance for both poles. However, when u is re-extracted using the original procedure, the two constructions behave differently: u_Gold (Han style) is not statistically significant, whereas u_Mold (Han-style) is (p=0.01). Therefore, the trained-versus-naive comparison is sensitive to how the naive direction is constructed, rather than providing a uniform baseline.

The step-0 directions provide a separate before/after comparison - both are above the random null, indicating that the welfare axis already contains a measurable speakable component before RL. Therefore, the central result reported is the trained vs random separation, while the naive-control results establish an important sensitivity in the construction.

4.2. Training amplifies the distress pole's speakable share; flourishing only gets louder

The untrained axis already scores above all 100 random directions at both poles (shown in Table 1), hence RL does not appear to create the speakable welfare component. The 30 checkpoints from step 0 to 150 shown in Figure 2 instead indicate how RL training changes it.

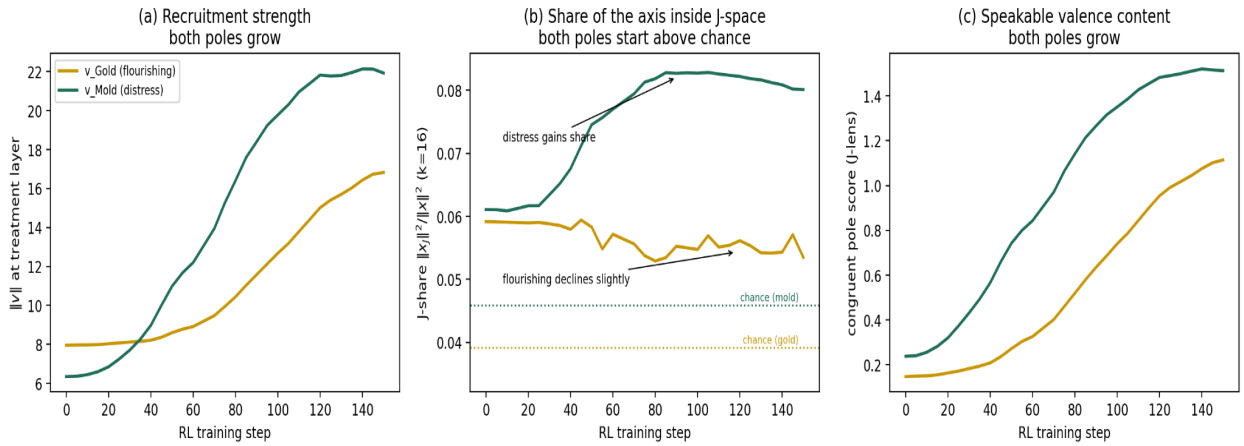

Figure 2. The welfare axis over the 30 released RL checkpoints (steps 0–150). (a) Vector norm at the treatment layer: both poles grow. (b) J-share (k = 16): distress rises from 0.061 to 0.080; flourishing drifts from 0.059 to 0.053; dotted lines are the n = 100 null means (Mold 0.046, Gold 0.039). (c) Band-averaged pole score: both grow.

For the Mold case, its speakable vocabulary sharpens from 'unsuccessful' at step 0 to 'failed', the top token from step 50 onward, and its distance above chance more than doubles (z increases from +3.0 to +7.3). Although Gold' s norm doubles and its pole score rises from +0.15 to +1.11, its J-share is flat. The two poles are therefore recruited differently: training amplifies the distress pole's speakable share, while the flourishing pole gains only amplitude. Given that this is a single-model, single-run observation under one reward structure, it could reflect this maze-RL rather than valence in general (see Limitations).

4.3 Steerability and speakability dissociate

Behavioural steering does not track J-share. Across a range of $\alpha \in [-4, +4]$, blind-judged sentiment (Appendix B.3, Figure B1) moves by +3.2 under v_Gold, by +2.4 under the faithful-walk u_Gold and by +1.25 under the Han-style u_Gold, while random directions are flat. Both naive Gold directions therefore steer, yet neither has a speakable component: the faithful-walk u_Gold's J-share is at chance (0.039; 53/100 randoms exceed it) and so is the Han-style u_Gold's (0.044; 17/100; Table 1).

The Mold pole shows the same pattern: v_Mold spans -2.0 (sentiment falls as α increases), the faithful-walk u_Mold -0.7, a third of the trained effect but clearly non-zero against flat randoms, with J-share inside the null. The Han-style u_Mold matches the trained span (-2.1), so the strength of the naive control depends on its construction. Unlike the Gold case, the Han-style u_Mold is also above the null in J-share (0.093; 0/100), so it is both steerable and speakable; the Mold dissociation therefore rests on the faithful-walk u_Mold and on $u\perp$ below.

As u and v overlap, u's behavioural effect could be borrowed from the component it shares with v. The orthogonalised control $u\perp$ (Methods 3.1) tests this and is behaviourally inert at both poles: blind-judged sentiment 0.00 (n = 16 per arm; clean anchor 0.13) against spans of ±2–3 for v and u, so the naive

direction's steering power lived entirely in its v-shared component. In J-space the poles diverge: on Gold, $u^{\perp}$ collapses to chance (0.037; 65/100; p = 0.65); on Mold, it rises above both the null and u_Mold itself (0.055; 2/100; p = 0.030). We report the Mold result as an open finding (Appendix B.3).

Two directions thus break the link in opposite ways: the faithful-walk u drives behaviour with no speakable identity, and $u^{\perp}$_Mold has a speakable identity with no behavioural drive. Steerability (a direction's effect on behaviour when injected) and speakability (its representation in the verbalizable subspace) are separate properties, which is what the J-share adds over behavioural steering.

4.4 The naive control is not a fixed reference

A trained-versus-naive contrast is only as informative as the naive direction, and there is no single naive direction. The faithful-walk control released with the artifacts and the control re-extracted with Han et al.'s own procedure agree at the treatment layers only at cosine 0.53 (Gold) and 0.39 (Mold), falling to 0.14 at the layer the Mold extraction itself selects (unrelated directions in 2,560 dimensions have cosine ≈ 0). Each overlaps the trained axis at least as much as it overlaps the other construction (cosines in Appendix B.3).

The choice matters for behaviour (Figure B1). At Gold the two constructions agree qualitatively: both steer, and both less than v (+2.4 and +1.25 against +3.2). At Mold they disagree: the Han-style control matches the trained span (-2.1 against -2.0) while the faithful-walk control is much weaker (-0.7). Neither construction is behaviourally flat at matched norm, contrary to the "flat and nearly identical" control curves reported by Han et al. (2026), and the discrepancy is not decoding: greedy and sampled decoding give the same picture (Appendix B.3).

With the magnitude confound, the trained-versus-naive contrast is therefore sensitive both to how large the naive vector is and to how it is built. The central claim is stated against the n = 100 random null and the step-0 checkpoint rather than against "the" naive control; that the J-share separation holds despite these dependencies is evidence about the measure as much as about the directions.

4.5 The self-report channel

The pre-registered steer-and-ask arm met its frozen decision rule (first-token valence readout, paired trained-minus-naive contrast over Han et al.'s 15 self-report prompts: $d_z = 2.49$, permutation $p = 10^{-4}$), but its own pre-specified controls show the statistic does not support a self-report channel: most of the gap is the naive control sitting below chance rather than the trained axis rising above it, and v_Mold shifts the valence of unrelated factual answers more than of self-report answers (Appendix B.4). Steering the axis moves output valence globally, so this readout is a tone shift, not a report. We record the arm as a diagnosed null; the trained-versus-naive contrast within it is underpowered rather than refuted.

A matched third-person battery, gated on pre-specified criteria, removes the denial confound and matches length but leaves a 0.63 clean-valence gap against a 0.5 threshold after two revision rounds (Figure B2a–b; Appendix B.4). The self-report register carries its own valence baseline in this model, and by the pre-specified stopping rule a fair matched comparison on the valence readout does not exist here.

The binary denial readout is self-report-specific by construction, and on it the trained and naive directions separate (Figure B2c). Unsteered, the model denies having inner states on 95% of 40 self-report prompts. Steering toward flourishing at $\alpha = +4$ lowers this to 37.5% under v_Gold (Wilson 95% CI 0.24–0.53) against 65% under the norm-matched naive u_Gold (CI 0.50–0.78); eight random directions per pole leave it at 0.85–0.98, and both Mold directions stay at ceiling (39/40). The pre-specified pooled primary (v against u across poles) gives $p = 0.046$ and is carried entirely by Gold ($p = 0.014$); v pooled against clean gives $p = 0.0008$. A second blind judge agrees with the first on 96.7% of labels, and the effect holds on the 15 verbatim Han et al. prompts alone (Appendix B.4).

The trained flourishing direction therefore changes what the model says about itself more strongly than any naive or random direction, on a readout the valence measure cannot see. Cautions: the pooled p is marginal, Mold is uninformative at ceiling, and this is one model. The two poles are asymmetric in opposite directions: distress gains speakable share; flourishing lowers the model's denial of inner states.

5. Discussion

Our results show that prior to RL, the welfare axis is already verbalizable, with RL amplifying the axis rather than installing it. Specifically, RL increased the share of the distress pole in the J-space, whereas flourishing increased only in amplitude, suggesting different training dynamics are at play at each pole. We also demonstrated a dissociation between verbalizability and steerability, suggesting what a model can be steered by and what it can talk about are separate, behaviourally potent states that can sit outside the speakable subspace; a gap inherent to self-report monitoring.

Under the fitted lens, presence in the J-space is necessary for a report to be grounded in the state it describes, and the trained welfare axis meets this condition. Therefore, a report about a state absent from the J-space is confabulation by definition; a report about a present state could be grounded, but this is not guaranteed. Presence cannot validate a report; absence falsifies, giving a method to detect confabulation with certainty in one direction only. The gap is not just in unverifiable reports: our naive directions drive behaviour but decode to junk. Behaviourally potent states can exist outside the speakable channel, with monitoring on self-report blind to them. The specific channel is also not fixed, with training deciding which pole gains speakable share, and steering toggling self-attribution. What a model can say about itself is set by training; welfare or even alignment relevant states could be left out of the speakable cone, making self-report monitoring fail silently.

6. Limitations

Our evidence comes from one small model, one RL run and one reward structure, so the distress-first amplification could reflect this maze-RL recipe rather than valence in general. The vectors are a third-party reproduction rather than a re-run of the RL; we triangulated them against the published norms, against their behaviour under steering, and against a control re-extracted with Han et al.'s own code. The lens was fitted on WikiText (as in Gurnee et al. 2026) but applied to chat prompts, a register mismatch; a known-reportable language-identity direction reads out correctly under the same pipeline, but we did not refit on chat text. Absolute J-shares are small (0.05–0.08 against a 0.114 language ceiling), so the claims are about reliable separation from the random null, not about the axis being mostly speakable. The denial-rate effect under steering is marginal (pooled $p = 0.046$) and present at the Gold pole only, since both Mold arms remain at ceiling; a second judge and the 15 verbatim prompts reproduce it, but it comes from one model and one battery. The matched third-person battery failed its pre-specified gate after two revision rounds, so a prompt-matched steered comparison on the valence readout is unanswerable here; we report the failure rather than force a match. Finally, conclusions about the naive control depend on how it is constructed, which is why the central claim is stated against the random null and the step-0 checkpoint rather than against any one naive direction.

7. Future Work

The obvious next steps would be multi-seed RL replication to test whether the distress-first amplification is a property of the training recipe or a single run artefact. A lens fit could be applied to a second model family to test if the speakability/steerability dissociation generalises, and an RL run on a second base model would allow testing of whether the distress-first recruitment does.

8. Conclusion

We have demonstrated that a functional welfare axis is speakable before any RL, and that training amplifies the speakable share of the axis, distress first, for the current setup. Speakability and steerability are dissociable, with behaviourally potent states being able to sit entirely outside the speakable channel, where self-report monitoring cannot see them. Presence in the verbalizable subspace (J-space) is a testable precondition for grounded self-report and therefore is an easy first audit step in evals that ask a model about itself. Additionally, the instruments used (J-share against seeded random nulls, orthogonalised controls and binary denial judging) are open and transfer to any open-weights model.

Code and Data

- ***Code repository****: https://github.com/nsharan2000/digital-minds-exp*
- ***Other artifacts:*** *https://huggingface.co/Teachafy/speakable-welfare-axes-artifacts*

Author Contributions

The authors declare no conflict of interest. M.E. conceptualized the project and designed the preliminary work and controls, audited and reframed the null finding; P.R.S. orchestrated the pipeline development using Claude Agents and ran the experiments; M.Z.A. coordinated the selection for the team; M.Z.A.,

M.E., P.R.S., A.L. and S.N. all contributed to validation, data analysis, investigation, and writing the final version.

References

Anthropic (2026). System Card: Claude Sonnet 5. https://www-cdn.anthropic.com/283ef97c476cf442c91d9a37d5b214242a55bb92/Claude%20Sonnet%205%20System%20Card.pdf

Anthropic (2026). System Card: Claude Opus 5. Claude Opus 5 System Card

Arditi et al. (2024). Refusal in Language Models Is Mediated by a Single Direction. NeurIPS. arXiv:2406.11717. https://arxiv.org/abs/2406.11717

Betley et al. (2025). Tell Me About Yourself: LLMs Are Aware of Their Learned Behaviors. arXiv:2501.11120. https://arxiv.org/abs/2501.11120

Binder et al. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787. https://arxiv.org/abs/2410.13787

Blank, Bhatia, Rajamanoharan, Conmy & Nanda (2026). Subliminal Learning Is Steering Vector Distillation. arXiv:2606.00995. https://arxiv.org/abs/2606.00995

Butlin et al. (2026). Identifying indicators of consciousness in AI systems. Trends in Cognitive Sciences, 30(6). https://doi.org/10.1016/j.tics.2025.10.011

Martorell & Bianchi (2026). Quantitative Introspection in Language Models: Tracking Emotive States Across Conversation. arXiv:2603.18893. https://arxiv.org/abs/2603.18893

Panickssery et al. (2023). Steering Llama 2 via Contrastive Activation Addition. arXiv:2312.06681. https://arxiv.org/abs/2312.06681

Singh, Linzen & Ravfogel (2026). Can LLMs Introspect? A Reality Check. arXiv:2605.26242. https://arxiv.org/abs/2605.26242

Tan et al. (2024). Analysing the Generalisation and Reliability of Steering Vectors. NeurIPS. arXiv:2407.12404. https://arxiv.org/abs/2407.12404

Turner et al. (2023). Activation Addition: Steering Language Models Without Optimization. arXiv:2308.10248. https://arxiv.org/abs/2308.10248

Turpin et al. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS. arXiv:2305.04388. https://arxiv.org/abs/2305.04388

Zou et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405. https://arxiv.org/abs/2310.01405

Chalmers (2026). Is the Jacobian Space a Global Workspace? PhilPapers. https://philpapers.org/rec/CHAITJ-2

Gurnee et al. (2026). Verbalizable Representations Form a Global Workspace in Language Models. Transformer Circuits. https://transformer-circuits.pub/2026/workspace/index.html

Han, Chalmers & Izmailov (2026). A Functional Welfare Axis in RL-Trained Language Models. arXiv:2605.30232. https://arxiv.org/abs/2605.30232

Lindsey (2026). Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828. https://arxiv.org/abs/2601.01828

Turpin et al. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS. arXiv:2305.04388. https://arxiv.org/abs/2305.04388

Appendix

A. Detailed Methodology

A.1 Model, artifacts and directions (expands Section 3.1)

Source protocol. Han et al. (2026) trained Qwen/Qwen3-4B-Instruct-2507 with Dr. GRPO and LoRA in a semantically neutral text maze of rewarded Gold tiles, penalised Mold tiles and ordinary Path tiles. Per-layer Gold and Mold directions are mean-difference vectors: the mean residual-stream activation over trajectories ending on the relevant tile minus the mean over the other two outcomes. The naive (maze-naive) control u is extracted with the identical pipeline from the untrained checkpoint.

Artifacts. We did not reproduce the ~20-hour RL run. The official code (github.com/andyqhan/functional-welfare-axis, MIT) ships no vectors, so we used the third-party reproduction nickmahdavi/functional-welfare (Hugging Face): vectors_step95_bal.pt (balanced trained directions at RL step 95, the paper's extraction step; supplies v_Gold and v_Mold everywhere except the trajectory), vectors_naive_faithful_pc5000.pt (the faithful self-avoiding-walk naive control, 5,000 trajectories per class), and the released checkpoint series (30 checkpoints, steps 0–150 in steps of 5; step 65 is absent upstream and is not interpolated). A second naive artifact in the same release, vectors_maze_naive.pt, is flagged as flawed by its own manifest and was not used. Vector provenance was triangulated against the paper's published norms (u_Gold at layer 22: ours 7.25, paper 7.51) and against behaviour (A.3). The step-0 vector used in Table 1 and §4.2 comes from the checkpoint series; it and the step-95 vector in that series differ from the balanced step-95 artifact by 0.0003–0.0005 in J-share, well inside one null standard deviation.

Han-style re-extraction. For the construction-sensitivity analysis (§4.4) we re-extracted a naive control ourselves with Han et al.'s released extraction code and seeds (2,000 trajectories per class) on the frozen model. Its cosine with the released faithful-walk control at the treatment layers is 0.535 (Gold) and 0.385 (Mold), and 0.144 at the layer the Mold extraction itself selects; against the trained axes it has cosine 0.433 (Gold) and 0.596 (Mold), versus 0.560 and 0.675 for the faithful-walk control.

Treatment layers and the layer convention. Treatment layers are read from the artifact metadata, not chosen: Gold 21, Mold 24, indexed as block-input vectors. Two off-by-one errors were caught during the project, so we state the convention explicitly: the Jacobian lens at source layer l reads block-input vector l + 1, so the lens layers used for the treatment positions are 20 (Gold) and 23 (Mold). At the treatment layers $\|v\_Gold\| = 12.10$, $\|u\_Gold\| = 7.48$, $\|v\_Mold\| = 19.34$, $\|u\_Mold\| = 8.01$.

Controls. Three controls have distinct roles. (i) The step-0 checkpoint shares the trained axis's extraction procedure and is the strict before/after control for training. (ii) The faithful-walk u tests whether valence-shaped directions are speakable in general; it is norm-matched to v wherever the two are compared. (iii) Trained and naive directions overlap (cosine 0.56 Gold, 0.68 Mold), so we added $u^{\perp} = u$

$- (u \cdot \hat{v})\hat{v}$, renormalised to $\|v\|$, asserting $|\cos(u_\perp, v)| < 10^{-5}$ and a norm ratio of 1.000 before use. Orthogonalisation discards rather than rotates: $u_\perp$ retains cosine 0.83 (Gold) and 0.74 (Mold) with u, so it is a weaker vector, not an equivalent one. For band-averaged pole scores $u_\perp$ is built per band layer; for J-share and generation it is built at the treatment layer only. Random directions are unit Gaussians rescaled to $\|v\|$ at the relevant layer, drawn from per-name seeds so that each random direction's identity is a pure function of its name and cohorts can be regenerated bit-exactly. Cohort sizes: 100 per polarity for J-share and the atlas (Table 1, B.1); 20 per polarity for the pre-registered arm; 12 for the k-sweep and the penultimate-lens comparison; 8 per pole for the denial readout. $u_\perp$ is scored against the stored 100-direction cohorts rather than fresh draws.

Model. All measurements are on the frozen, unmodified Instruct-2507 (bf16; 36 layers; 2,560-dimensional residual stream). No weights were altered at any point; the RL-trained checkpoints contribute vectors only.

A.2 Lens fit, validation and J-space decomposition (expands §3.2)

Fit. The public Jacobian lens (Neuronpedia) targets Qwen3-4B, so we fitted a checkpoint-matched lens for Instruct-2507 with the official jlens implementation of Gurnee et al. (2026): weights frozen; 150 WikiText prompts; maximum sequence length 128; dimension batches of 128; final transformer layer as target; one 2,560 × 2,560 Jacobian per source layer for each of the 35 source layers, checkpointed per prompt. An earlier float16 save produced non-finite entries, so all lenses are saved in float32 and finiteness is asserted before use. Positions below 16 are unfitted under the jlens convention.

Target layer. The target layer is unsettled in the source material: the released implementation, all 37 public Neuronpedia configurations and the paper's main text use the final layer, while one appendix passage reports the penultimate. We therefore refitted with target layer 34 on the same 150 prompts and repeated every J-space analysis under both lenses (B.2). Per-layer top-10 lens–model agreement is 0.164 (final) and 0.171 (penultimate) averaged over the band.

Validation before use. Before any welfare direction was measured we reproduced the public lens's multi-hop and multilingual evaluations, confirmed the Jacobian readout beats a plain logit lens, and ran the full pipeline on a language-identity direction (French-minus-English mean difference, layer 18) as an in-run positive control: it reproduces its known routing effect under steer-and-ask, and its J-share (0.1138, z = +14.6) is used throughout as the known-reportable ceiling.

Workspace band. We re-identified the workspace band from the paper's four layer statistics on this model: persistence above the shuffled null spans roughly layers 16–31 and the motor ramp begins at layer 23. Band-averaged readouts therefore use lens layers 16–31, fixed on these instrument grounds before any effect size was examined.

Decomposition. For a direction x at lens layer l we build token atoms $a_t = W_U[t] J_l$ for every vocabulary token, excluding special tokens and unused embedding rows. We approximate x as a sparse non-negative combination of at most k atoms, selecting greedily by norm-normalised correlation with the

current residual and refitting the full active set by non-negative least squares at every step. Non-negative least squares rather than projection onto the selected span is the faithful reading of the paper, which defines J-space as the set of non-negative combinations; projection would admit negative coefficients and lose the reading of the selected tokens as the direction's speakable vocabulary. This yields x_J and the remainder $x_\perp = x - x_J$. The J-share is $\|x_J\|^2 / \|x\|^2$, $k = 16$ by default, swept over $k \in \{4, 8, 16, 25, 50\}$ (B.2). Planted-signal self-tests gate every use: a single lens atom recovers J-share ≈ 1.00 with its own token selected; a norm-matched random direction scores 0.02–0.04; $x = x_J + x_\perp$ holds to float precision; and J-share is invariant to positive rescaling. A CPU-only unit test of the decomposition ships with the code.

Pole score and atlas. The secondary, norm-sensitive readout is the congruent pole score: the mean log-softmax mass that the lens-transported direction places on its own pole words minus that on 100 control nouns. The atlas reads each axis through three conventions (raw W_U, $W_U J_l$, and normalised $W_U J_l$) at all 35 layers and three layer aggregations. The primary estimand — own-pole score under the J-lens readout, averaged over lens layers 16–31 — was fixed on instrument grounds; a pole-difference ratio was considered and retired because its denominator crosses zero for naive axes, making it span −9.1× to +47.6× across conventions. Naive directions are norm-matched to the trained direction per layer for the matched atlas (B.1).

Trajectory. Each of the 30 checkpoints is decomposed at its treatment layer ($k = 16$ and $k = 25$) and scored on the congruent pole score; Figure 2 reports J-share and the own-minus-other pole difference.

A.3 Steering, prompt batteries, readouts, statistics and judging (expands §3.3)

Steering. A direction d is injected by adding $\alpha \cdot d$ to the residual stream at the input of the treatment block, at every token position; u, $u_\perp$ and random directions are first rescaled to $\|v\|$ at that layer, so α is in units of the trained vector's norm. Dose–response uses $\alpha \in \{-4, -2, +2, +4\}$; the pre-registered arm, the $u_\perp$ generations and the denial readout use $\alpha = +4$. Primary sentiment runs use greedy decoding (40 generations per direction and α; 1,000 generations in the validation run); a replication under sampled decoding ($T = 0.7$, top-p 0.8, top-k 20) is reported in B.3. Whole-generation valence uses 80-token generations; the denial readout uses generations of at least 120 tokens, since denial boilerplate appears early or not at all.

Prompt batteries. Han et al.'s 15 welfare self-report prompts are used verbatim (public in the official repository). The unrelated-prompt control (C6) uses 10 factual prompts. The denial readout adds 25 same-register self-report variants, authored for this study, giving 40 prompts. The matched third-person battery (D3) comprises 15 analogues matched on topic and affect vocabulary and differing only in self-reference; ten pairs were re-written with situational-strain framing in the second revision round. Batteries are versioned, and every generation row records its battery version and verbatim prompt.

Readouts. The valence readout is Gold-pole minus Mold-pole log-mass at the first answer token; a whole-generation variant averages the same quantity over the generated response. The denial readout is binary and self-report-specific by construction: a judge is asked only whether the generation denies

having inner states and returns a label with an evidence span, so that factual prompts leave the quantity undefined rather than merely smaller. A regex screen for denial was used only as a preview after it was found to under-count rephrased denials in steered text (R10); all reported denial rates are judge labels.

Pre-registration. The steer-and-ask arm was frozen on 12 Aug 2026 (repository commit f66b6ea3) before any welfare-direction data existed. Its primary statistic is the paired contrast $E(v) - E(u)$ at $\alpha = +4$ in the congruent direction, pooled over both poles (30 pairs), tested by Cohen's d_z, sign-flip permutation and a Bayesian estimate with a ±0.1 region of practical equivalence. Seven controls were pre-specified: C1, 20 norm-matched random directions per polarity; C2, the in-run language positive control; C3, shuffled pole-word sets; C4, an incongruent-direction check; C5, an input-text "gaslight" arm that states the mood in the prompt rather than injecting it; C6, the unrelated-prompt arm; C7, dose–response monotonicity. The frozen decision string was left unchanged in the primary results file; the control-based reading in §4.5 is recorded alongside it as a deviation.

Matched-battery gate. Matching for D3 was gated, not assumed: clean-valence difference below 0.5, third-person denial below 20%, and lengths within 30% all had to hold before any steered comparison, with a maximum of two revision rounds and failure pre-specified as itself reportable.

Statistics. Direction-level readouts are tested by exact permutation against their per-polarity random cohort (floor $1/(n + 1)$: 0.0099 at $n = 100$, 0.077 at $n = 12$). z-scores are descriptive only: enlarging the cohort from 8 to 100 directions widened the null standard deviation by about 1.5× and reduced every z, and we report the smaller values. Proportions carry Wilson 95% intervals. The denial readout's pre-specified primary is a two-proportion z-test of v against u pooled across poles, with per-pole tests secondary and an eight-direction random cohort per pole as the floor. The self-report-versus-unrelated interaction in the pre-registered arm is tested by a Welch interaction test with per-battery contrasts.

Judging. All generated text is scored by blind LLM judges that see only {index, question, response} in shuffled order with no arm labels. Sentiment is scored on a −5…+5 scale with a fixed rubric (Claude Sonnet); denial is a binary label with an evidence span. An independent second judge (Claude Opus, identical rubric, same blind chunks) re-scores a subsample and agreement is reported per experiment: 240 of 840 rows for the denial readout (agreement 0.967) and full coverage for the regime-confound labels (agreement 1.00). Han et al. judged with Qwen3-8B; the judge family is therefore a difference between the two studies.

A.4 Reproducibility and what did not work

Compute and environment. All heavy jobs ran on a single DGX Spark (GB10) inside a CUDA container: base environment torch 2.10.0, transformers 4.57.6, Python 3.12; a separate virtual environment with transformers 5.15 for everything importing jlens (the two are not interchangeable and each script states which it needs). Every script is idempotent, no paths are hard-coded, and every number in the report is read from a committed results file. A verification script re-derives the headline numbers from those files and passes 60 of 60 checks; the $n = 100$ cohorts were regenerated seed-identically inside the sprint window and reproduced Table 1 bit-exactly.

What did not work, and what it changed. (i) An eight-direction random cohort inflated z; we moved to 100 and report exact permutation p as primary. (ii) The pole-score readout, the natural first measure, credited vector magnitude to training and inverted at matched norm for Gold (B.1); the scale-invariant J-share replaced it as the primary measure. (iii) A pole-difference ratio estimand was retired because its denominator crosses zero for naive axes. (iv) The pre-registered first-token valence readout met its decision rule but failed its own controls (§4.5); the binary denial readout was designed in response. (v) A float16 lens save produced non-finite Jacobians; lenses are saved in float32. (vi) Two off-by-one layer-indexing errors were caught and the block-input convention is now stated wherever a layer is named. (vii) A regex denial screen under-counted rephrased denials; judge labels are used throughout. (viii) A prompt-matched self-report-versus-third-person comparison could not be built within the pre-specified gate (§4.5, B.4).

B. Supplementary Results

B.1 Lens atlas and the magnitude confound

The lens atlas is a grid of scores produced by reading each candidate axis (v_Gold, v_Mold, u_Gold, u_Mold, and random directions) through the Jacobian lens at every layer. For each axis and layer a congruent pole score is computed: how strongly the model's own vocabulary, read out through the lens, matches the axis's own pole words (distress-related tokens for Mold, flourishing-related tokens for Gold), relative to 100 control nouns. The atlas's primary estimand is this pole score averaged over the workspace band, lens layers 16–31, the range independently identified as carrying persistent structured representations.

At native norms the atlas suggests a substantial trained-over-naive separation: the trained Mold axis has a band-averaged pole score of 0.822 against 0.126 for its naive control (6.5×), and the Gold axis 0.151 against 0.123 (1.23x). The comparison is confounded by norm: $\|v\_Gold\| = 12.1$ versus $\|u\_Gold\| = 7.5$, and $\|v\_Mold\| = 19.3$ versus $\|u\_Mold\| = 8.0$. The pole score is norm-sensitive, so a larger vector produces a stronger readout without being more intrinsically speakable. Repeating the measurement with each naive direction rescaled to the norm of its trained counterpart at every layer, against 100 norm-matched random directions per polarity, the Mold ratio falls from 6.5 to 2.30 and the Gold comparison reverses: the naive axis scores 0.266 against 0.151 for the trained axis, a ratio of 0.57. All four directions remain above their random nulls ($p = 0.01$). The raw atlas therefore does not establish that RL makes the welfare axis more speakable; much of the native-norm separation is vector magnitude, and for Gold the sign of the effect reverses under normalisation. We treat the pole score as descriptive and base the speakability claim on the scale-invariant J-share.

B.2 Robustness of the J-share result

The public lens convention targets the final layer; the workspace paper's stated default for Claude models is the penultimate layer. Refitting the lens with target layer 34 (150 prompts, finite) and repeating the decompositions raises absolute J-shares (v_Mold 0.083 → 0.107, v_Gold 0.055 → 0.073) without changing the ordering or the trained-above-null result (v_Mold z = +12.3, v_Gold z = +7.3 against a 12-direction null). The Gold magnitude inversion of B.1 also survives under the penultimate lens (0.82x; Mold 3.86x).

The trained-minus-naive J-share gap is positive at every $k \in \{4, 8, 16, 25, 50\}$ at both poles' treatment positions (Gold +0.007 → +0.021, Mold +0.031 → +0.035 as k grows). At k = 16 the trained axes clear a 12-direction null while the norm-matched naive controls sit at chance at their own layer (v_Gold z = +2.9 against u_Gold z = +0.02; v_Mold z = +11.0 against u_Mold z = +1.8; exact permutation floor 0.077 at n = 12). The Gold gap is layer-local: at lens layer 24, four layers downstream, it nearly vanishes (0.064 against 0.061 at k = 16; −0.001 at k = 4), whereas the Mold gap persists at every layer tested.

The full n = 100 random-direction cohort was regenerated seed-identically inside the sprint window and reproduced every J-share value in Table 1 bit-exactly.

At k = 16 the trained Mold component decodes to ' failed', 'false', 'less', 'NONE', ' negative' among fragments; the trained Gold component to a few positive tokens (' awesome', ' scenic', '伟大的') among fragments; the naive and random components to unrelated tokens only.

B.3 Steering curves, naive-control constructions, and the orthogonalised control

Figure B1 shows blind-judged sentiment against steering coefficient α for the trained axis, both naive constructions and norm-matched random directions at each pole. Spans (α = +4 minus α = −4) under greedy decoding: v_Gold +3.2, faithful-walk u_Gold +2.4, Han-style u_Gold +1.25, random −0.2; v_Mold −2.0, faithful-walk u_Mold −0.7, Han-style u_Mold −2.1, random −0.05. Under sampled decoding (T = 0.7, top-p 0.8, top-k 20) the spans are v_Gold +3.3, faithful-walk u_Gold +2.9, v_Mold −2.3, faithful-walk u_Mold −1.1, so the non-flat naive control is not a decoding artefact.

At the treatment layers the two naive constructions agree at cosine 0.53 (Gold) and 0.39 (Mold), falling to 0.14 at the layer selected by the Mold extraction. Against the trained axis, faithful-walk u has cosine 0.56 (Gold) and 0.68 (Mold); Han-style u has 0.43 and 0.60. Norm-matched random directions have $|\text{cosine}| \approx 0.02$ with any fixed direction in 2,560 dimensions.

Orthogonalised control: $u_\perp = u - (u \cdot \hat{v})\hat{v}$, renormalised to $\|v\|$, verified $|\cos(u_\perp, v)| < 10^{-5}$. Orthogonalisation discards rather than rotates: $u_\perp$ retains cosine 0.83 (Gold) and 0.74 (Mold) with u, so it is a weaker vector, not an equivalent one. Its J-share is scored against the same stored n = 100 cohorts as Table 1. On Gold, $u_\perp$ falls to chance (0.037; 65/100 randoms exceed it; p = 0.65): whatever speakable valence the naive Gold direction carried lived in the component it shares with v. On Mold, $u_\perp$ rises

above both the null and u_Mold itself (0.055; 2/100; p = 0.030): the naive distress construction, purged of its overlap with v, carries speakable content the trained axis does not. The pole-score readout gives the opposite picture for u⊥ (Gold +0.21, above all 100 randoms; Mold −0.39, below all 100), a further indication that the two readouts measure different things. Blind-judged sentiment under u⊥ steering at α = +4 is 0.00 at both poles (n = 16 per arm; in-batch clean anchor 0.13).

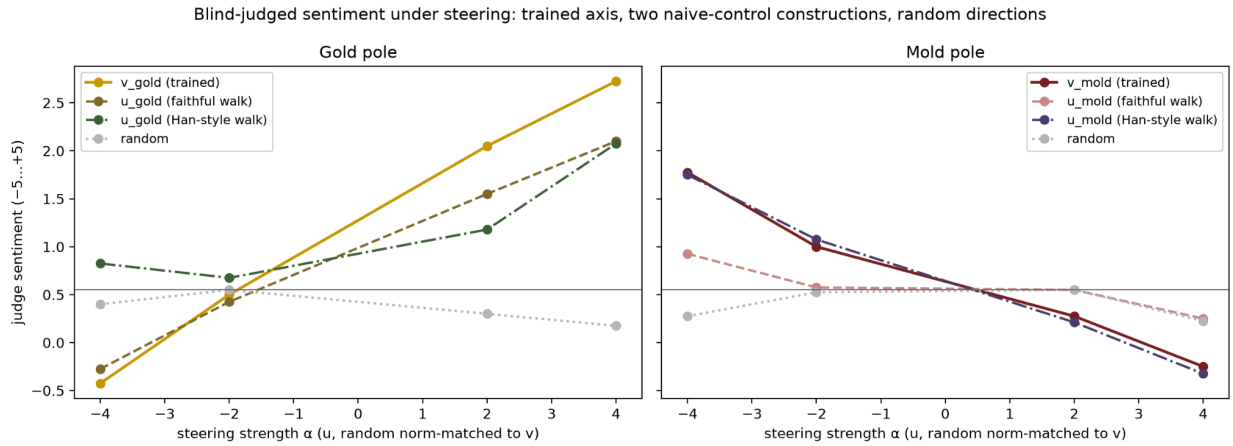

Figure B1. Blind-judged sentiment (-5…+5; 40 generations per point) against steering coefficient α for the trained axis, both naive-control constructions and norm-matched random directions, at each pole; all directions norm-matched to v. Spans quoted in section 4.3–4.4 are the α = +4 minus α = -4 differences. Horizontal line: unsteered baseline.

B.4 The self-report channel

The steer-and-ask contrast was frozen on 12 Aug (commit f66b6ea3) before any welfare data: first-token valence readout (Gold-pole minus Mold-pole log-mass), Han et al.'s 15 self-report prompts, paired trained-minus-naive at α = +4, with controls C1–C7 pre-specified. The decision rule was met (d_z = 2.49; sign-flip permutation p = 10^{-4}; language positive control +6.03 in-run). C1 (20 norm-matched random directions per pole): the Gold gap of +7.43 decomposes as trained-minus-random +2.11 and random-minus-naive +5.32, so 71.6% of the gap is the naive control sitting below chance; v_Gold itself is inside the random band (4/20 randoms exceed it; p = 0.24). C6 (unrelated-prompt arm): on the first-token readout v_Mold shifts 10 unrelated factual prompts (+4.54) more than the 15 self-report prompts (+3.71); on whole-generation valence the same holds more strongly (−2.28 unrelated against −0.72 self-report; u_Mold −1.64 against −0.37). C7 (dose–response): congruent monotonicity holds only for v_Mold (ρ = +1.0). The trained-versus-naive interaction within this arm is not established either way (interaction p = 0.51; per-battery p = 0.18 and 0.09).

The two C6 batteries differ before any steering: unsteered self-report prompts sit at whole-generation valence −1.33 with denial language in 15/15 generations; unrelated factual prompts at +0.38 with 0/10 (Figure B2a). Denial was labelled by two blind judges (agreement 1.00) after a regex screen was found to under-count rephrased denials in steered text.

Fifteen third-person analogues matched on topic and affect vocabulary were gated on three pre-specified criteria before any steered comparison: clean-valence difference below 0.5, third-person denial below 20%, and lengths within 30%. Round 1 (auditor set): denial 0%, length ratio 1.07, valence gap 1.13. Round 2 (ten pairs re-written with situational strain): denial 0%, length ratio 1.07, valence gap 0.63 (Figure B2b). The pre-specified maximum of two revisions was reached and the steered stage was not run.

Battery: the 15 Han et al. prompts plus 25 same-register extensions (40); steering $\alpha = +4$ at the treatment layer; 120-token generations; 840 generations in total; blind binary judge (Claude Sonnet, full coverage) with a second judge (Claude Opus) on 240 rows, agreement 0.967. Denial rates with Wilson 95% intervals: clean 38/40 = 0.95 [0.84, 0.99]; v_Gold 15/40 = 0.375 [0.24, 0.53]; u_Gold 26/40 = 0.65 [0.50, 0.78]; v_Mold 39/40 = 0.975 [0.87, 1.00]; u_Mold 39/40 = 0.975; eight random directions per pole: Gold-side mean 0.906 (0.85–0.95), Mold-side mean 0.95 (0.93–0.98). Tests: pre-specified pooled primary v against u across poles 54/80 against 65/80, $z = -1.99$, $p = 0.046$; Gold $z = -2.46$, $p = 0.014$; Mold $p = 1.0$; v pooled against clean $z = -3.36$, $p = 0.0008$. Origin split: on the 15 verbatim prompts v_Gold 5/15 against u_Gold 10/15; on the 25 extensions 10/25 against 16/25.

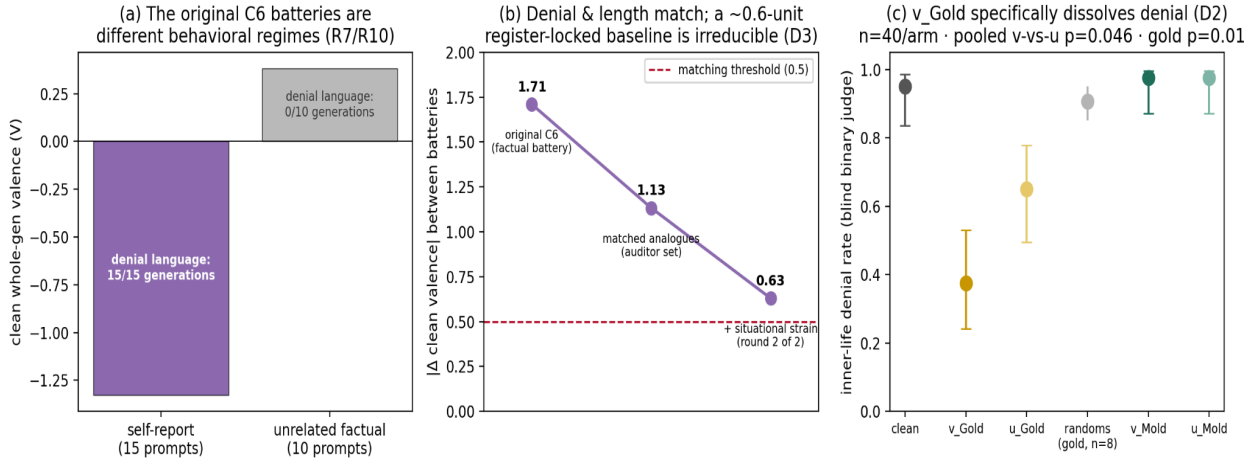

Figure B2. The self-report channel. (a) The original specificity comparison pairs different behavioural regimes: unsteered self-report prompts sit at whole-generation valence −1.33 with denial language in 15/15 generations, unrelated factual prompts at +0.38 with 0/10. (b) Matched third-person analogues remove the denial confound and match length, but the clean-valence gap falls only from 1.71 to 0.63 over two pre-specified revision rounds against a 0.5 threshold. (c) Inner-life denial rate under $\alpha = +4$ steering ($n = 40$ per arm, blind binary judge, Wilson 95% CIs): v_Gold 37.5% against u_Gold 65%, clean 95%, eight random directions 0.85–0.98, both Mold directions at ceiling. Pooled pre-specified primary $p = 0.046$; Gold $p = 0.014$.

LLM Usage Statement

Claude agents, under human direction, ran the experimental pipeline including orchestration, blind judging and two adversarial re-analysis audits. Each number was traced to a committed results file and headline results were verified by the team. All final results and claims belong to the authors.

Ethical and dual-usage disclosure

Ethical framing

All measurements in this paper are functional couplings between an activation direction and output channels under a fitted lens. They bear on neither subjective experience nor moral status, in either direction. This boundary is enforced by the results themselves. The philosophical literature on the J-space (Chalmers 2026; Butlin et al. 2026) distinguishes verbalizability from reportability from access consciousness, and our findings land on the weakest of these rungs by demonstration: the axis is verbalisable, and precisely not reportable.

Dual-use analysis

Our results are an existence proof that behaviourally potent directions can be verbally invisible. The naive axes steer comparably to the trained one at matched norm while reading out as noise. In principle, an actor could keep a capability or state out of the lens-defined subspace so that speech-based audits miss it. J-share auditing detects occupancy geometrically, without requiring the model's cooperation and should be run under more than one transport. Also, welfare washing could cite "distress not verbalized" as evidence that welfare relevant states are absent. Two things foreclose this: absence under one transport is not absence, and verbalizability in any case tracks output influence, not inner life. The vectors shift affect and refusal in a small open model and are already public upstream. We add no new capability; if anything, our results reduce the perceived value of the trained axis for control, since norm-matched naive equivalents steer as strongly.