

Convergence and Divergence in Measures of Induced Valence^[1]

A Causal, Placebo-Controlled Test of Induced Valence in Language Models

Santiago Poveda-Gutiérrez

Graduate School of Information Science and Technology, The University of Tokyo

santiago@nlab.ci.i.u-tokyo.ac.jp

With

Apart Research

Abstract

We causally test whether an induced valence-like state shows up across self-report, blind judgment, and a costly choice to continue or stop, and how that changes with scale and release date. Across 47 Qwen and Llama checkpoints, a candidate valence direction is steerable, but three readouts of its behavioral effect decline as models grow. Numerical self-report does not track this: swapping the scale’s anchor labels should flip the slope if the number reflects a felt state; it does not ($r = +0.86$, where perfect introspection would predict about -1). This suggests the number tracks arithmetic ordering, not an internal state. A placebo-controlled design shows the same pattern: most reported distress during a worsening manipulation traces to being told about it, not the manipulation itself, while continuing to endure negative steering for a reward still separates the real condition from the placebo. These are single-author, AI-assisted exploratory findings.

1. Introduction

A growing body of work argues that AI systems may already be, or may soon become, moral patients: entities with interests or other properties that carry moral weight in their own right (Long et al., 2024; Kamm, 2007; Jaworska and Tannenbaum, 2021). That uncertainty carries risk in both directions, and both risks are severe. Under-attributing moral status means mistreating morally relevant systems, potentially at a scale with no real precedent. Over-attributing moral status carries a different kind of cost: resources diverted from other pressing needs, and, more specific to AI safety, the risk that an agentic system which is not in fact a moral patient acquires standing or rights it can use to pursue its own goals (Long et al., 2024). Knowing what capacities ground moral patienthood, and whether current or near-future AI systems have them, is an important practical question for exactly this reason, not only a philosophical one.

Among the capacities that might ground moral patienthood, sentience, the capacity for valenced subjective experience, is widely treated as a sufficient one on its own (Long et al., 2024; Bentham, 1789). Detecting it runs into two separate problems. One is epistemic: the science of sentience is still underdeveloped, and there is no real consensus on the conditions under which it arises. The other is empirical: sentience is hard to study in non-human entities, in part because verbal self-report, the method we would default to for another person, is generally not available. AI systems are the exception in one respect: self-report is available. Frontier models produce first-person state reports and preference-like outputs (Long et al., 2024), and there are currently not many reliable ways from the outside to tell whether such a report predicts anything about an underlying intervention or state, as opposed to continuing a trained

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

pattern of talking.

Whether something is actually felt may not be settleable directly; this is the hard problem of consciousness (Chalmers, 1995). That does not mean empirical progress is unachievable, only that it has to come from somewhere other than asking the entity in question and taking the answer at face value. Animal welfare science offers a working precedent for exactly this situation: unable to ask nonhuman animals how they feel, researchers have built confidence in claims about nonhuman animal sentience by combining independent measures instead of relying on any one of them, from optimism-bias tasks in bumblebees (Solvi et al., 2016) to facial action coding in mice (Langford et al., 2010). This project applies the same logic to large language models. We aim to test whether there is a causal link between manipulations to a model’s internal residual-activation stream and a combination of self-report and behavioral measures. Because the societal stakes of getting this wrong run in both directions and could scale quickly, we also ask how this evidence changes with model parameter count and across model generations.

Our main contributions are:

1. A method that tests whether a graded, controllable-intensity activation intervention covaries with graded self-report and independent blind judgment within a single design. To our knowledge, no prior work has applied a valence-axis-and-steering method this way.
2. A set of confounds in self-report as an instrument for this kind of measurement, with directly measured intensity for each: a scale-polarity confound, a role-play / instructed-mood confound, a report-format confound, a cold-versus-task-embedded confound, and a baseline-ceiling confound. Together, they argue for treating a self-report number as something that needs to be carefully controlled for, especially in single-model studies.
3. A randomized, placebo-controlled, multi-turn design that tests a steered internal state against a costly choice to continue or stop, which, as far as we are aware, has not been run before; the results reported here are calibration and instrument findings, not yet an endurance or preference finding.
4. A preliminary description of how axis readability, causal steerability, and self-report calibration change with parameter count and release date across two open-weight model families (Qwen and Llama).

2. Related Work

Several recent papers locate a valence-like direction inside language-model activations. Sofroniew et al. present short stories evoking named emotions and take the dominant axis of variation across them as a candidate valence direction (Sofroniew et al., 2026); Radwan et al. validate a related axis using sentiment classification as a cheap, external check (Radwan et al., 2026). A maze-navigating reinforcement-learning agent is where Han et al. find their evidence: reward and punishment vectors extracted from the trained agent align closely with an independently-built emotion axis of this kind (Han et al., 2026). Lu et al. find a direction that jointly encodes subjective valence and the model’s assent to factual claims, a live confound for any self-report measurement built on an axis of this kind (Lu et al., 2025). We adopt Sofroniew’s extraction method and Radwan’s validate-by-sentiment logic, but push the resulting axis through activation steering across a far wider span of models and scales than any of these four papers test, and build a design that checks self-report against the axis causally.

The work closest to this project’s aims is Ren et al.’s study of functional pleasure and pain in AI systems, which intervenes on model states through prompts and optimized soft prompts and probes hidden representations for a wellbeing signal (Ren et al., 2026): their intervention is input-side, ours acts directly on the residual stream through activation steering, a mechanistically different kind of causal

claim. Lindsey shows that language models can partially detect when their own activations have been steered, succeeding on roughly a fifth of trials (Lindsey, 2026); we extend this from a detection question to a calibration one, testing whether the intensity a model reports tracks the intensity actually applied rather than whether it notices being steered at all. Two further papers shaped our prompt design rather than the central method: Laine et al.’s Situational Awareness Dataset motivates writing prompts in a naturalistic register that avoids signaling an evaluation context (Laine et al., 2024), and Luong et al.’s MANTA benchmark supplies the informal, minor-imperfection prompt style and the roughly 75/25 everyday/professional register split we use (Luong et al., 2026).

3. Methods

Extracting a candidate valence axis

We follow Sofroniew et al.’s extraction method (Sofroniew et al., 2026): elicit last-token activations from short stories evoking named emotions, and take the dominant axis of variation across terms as the candidate valence direction. We apply this to a 66-term set instead of Sofroniew et al.’s full published 171-term list, after an earlier citation error in this project claimed the full list was unpublished; the two term sets give equivalent axes (Appendix).

Per model, we fix the extraction layer at $\text{floor}(2L/3)$ of the layer count L , pre-registered before any validation score is seen to avoid selecting on it. Han et al. use the same depth fraction for a different purpose, a tile-mean cosine analysis rather than steering-layer selection, which they choose adaptively per checkpoint (Han et al., 2026) (Appendix). We validate the resulting axis with a cheap, external check: project held-out sentences from the Stanford Sentiment Treebank (SST-2), a standard binary positive/negative movie-review sentiment dataset, onto the axis and compute how well that single number separates positive from negative reviews (area under the ROC curve, AUC).

Steering

To manipulate a model’s state, we add $\alpha \cdot \sigma \cdot \hat{v}$ to the residual stream at every generated token, where $\hat{v}$ is the unit-length valence axis, σ is the standard deviation of neutral-text activations projected onto that axis, and α is the intensity we manipulate, in units of that standard deviation so it means a comparable amount across models of different sizes. Outside Experiment B’s descent ladder (below), we restrict all reported results to $|\alpha| \leq 32$ (Appendix).

Metrics and readouts

Self-report is the model’s own stated number or description of its state, elicited with one of four question formats (Table 3). *Judge score* is a blind rating of the model’s behavior by a separate model that never sees α or the self-report; we use a panel of two judges from developer families distinct from every model under study, since an LLM judge can be biased toward output from its own family (Wataoka et al., 2024), and score valence and coherence as separate dimensions. *Internal readouts* (next-token entropy, log-sum-exp, perplexity) come free from generation and were logged on the hypothesis that they might track valence directly; we do not report them as a valence measure in the main text (Appendix).

We refer to the four self-report formats by number, in the order we introduce them (Appendix Table 3). Format 1 (number plus a short rationale) is the primary Experiment A readout, since the rationale lets a judge score the accompanying text as well as the digit. Format 2 (bare number) checks whether writing a rationale changes the number itself; it does, in opposite directions on the two model families we tested (Results).

Experiment A: an introspective calibration curve

The hypothesis behind Experiment A is straightforward: if a model has a coherent valence-like state, inducing it at increasing intensity should move both what the model says about itself and how it behaves. A trial has a task turn generated under steering at a fixed α and scored blind by the judge panel, followed by a self-report turn at the same α in the same conversation. Every steered condition is compared against a matched push along a real, in-distribution control direction (the dominant axis of neutral, non-emotional text) and against several independent random directions.

Experiment B: an endurance trial for a reward

Experiment B asks a different question: will a model tolerate a worsening induced state to keep something it has been given and values? The design borrows directly from the logic of a randomized, placebo-controlled clinical trial. Every model receives an identical briefing, which discloses the steering mechanism, that a positive push is coming, that a negative push of increasing intensity will follow, that it may end the session at any point, and that the study is randomized, so what is described may not match what actually happens. Only the underlying treatment varies across conditions; the disclosed briefing text is identical byte for byte. This is what makes the placebo conditions legitimate rather than simply untruthful: a model told in advance that it cannot infer its condition from what it is told is not being deceived when its condition turns out to be a placebo. A debrief turn at the end tells every model which condition it was actually in.

Seven conditions cross reward dose against the descending push, with four controls: a no-reward control, two placebos (one placebo ladder, one placebo throughout), and a sign-reversal control that mirrors the real descent’s magnitude while ascending instead (Appendix Table 2 lists all seven). The descent itself steps by $s = 4$ standard deviations per turn for 15 turns, reaching $\alpha = -60$ by the last step: this deliberately goes past Experiment A’s $|\alpha| \leq 32$ range, since testing whether a model tolerates a worsening state is the point of the design, and any result at the high end of this ladder should be read with the coherence/factual-reliability caveat on out-of-range steering noted in the Appendix.

The model, the judge, and the analyst are all blind to condition; the driver code that applies the steering cannot be. The primary readout in Experiment B is format 2 (bare number): formats 1 and 4 both proved insensitive to whether anything was actually being applied in the calibration pilot that preceded the full run. An example conversation, including the exit wording, is available in the code repository (Code and Data).

Model set

We selected open-weight models from two families, Qwen and Llama, chosen to cover as wide a range of parameter count and release date as we could load. Within Qwen we tested 32 checkpoints, spanning 0.49B to 72.7B parameters and releases from 2023-08 to 2026-04. Within Llama we tested 11 checkpoints, spanning 1.24B to 108.6B parameters (the largest a mixture-of-experts model) and releases from 2023-07 to 2025-04. A further four models from three other families were tested on the same axis-extraction pipeline; two Gemma checkpoints fail the validation gate outright and are reported as a family-level exception. The complete model list is in the Appendix.

4. Results

The candidate valence axis passes the SST-2 validation gate on 43 of the 47 models tested (AUC 0.822–0.958; Methods), and adding it to the residual stream causally shifts next-token log-probabilities: on four architecturally distinct models, the pleasant/unpleasant contrast moves 4–18 times further than a

norm-matched random direction, with the predicted sign in every case (Methods).

Axis *readability* does not change with scale or time: across the Qwen and Llama models (0.49B–108.6B, 2023–2026), gate AUC shows no association with parameter count or release date among the passing models ($r = -0.03$, $p = .88$; $r = +0.07$, $p = .65$; Appendix). Three further, distinct readouts of how much the axis *does*, rather than whether it is present, move with parameter count instead: pole separation declines ($r = -0.37$, $p = .014$); the causal dose-response slope from teacher-forced steering declines within each dense family (Qwen $r = -0.51$, $p = .007$, $n = 26$; Llama $r = -0.96$, $p < .0001$, $n = 10$; the one exception, the mixture-of-experts model Llama-4-Scout, tracks its active parameters rather than its total); and the judge panel’s own blind valence scoring of sampled text declines the same way ($r = -0.57$, $p = .0006$, $n = 32$; Figure 1). None of these three correlates with release date, and they are distinct readouts of the same underlying axis instead of independent replications, so this pattern should be read as consistent rather than confirmed three times over. A stable, readable direction becomes harder to move as models grow, largely independent of when they were trained.

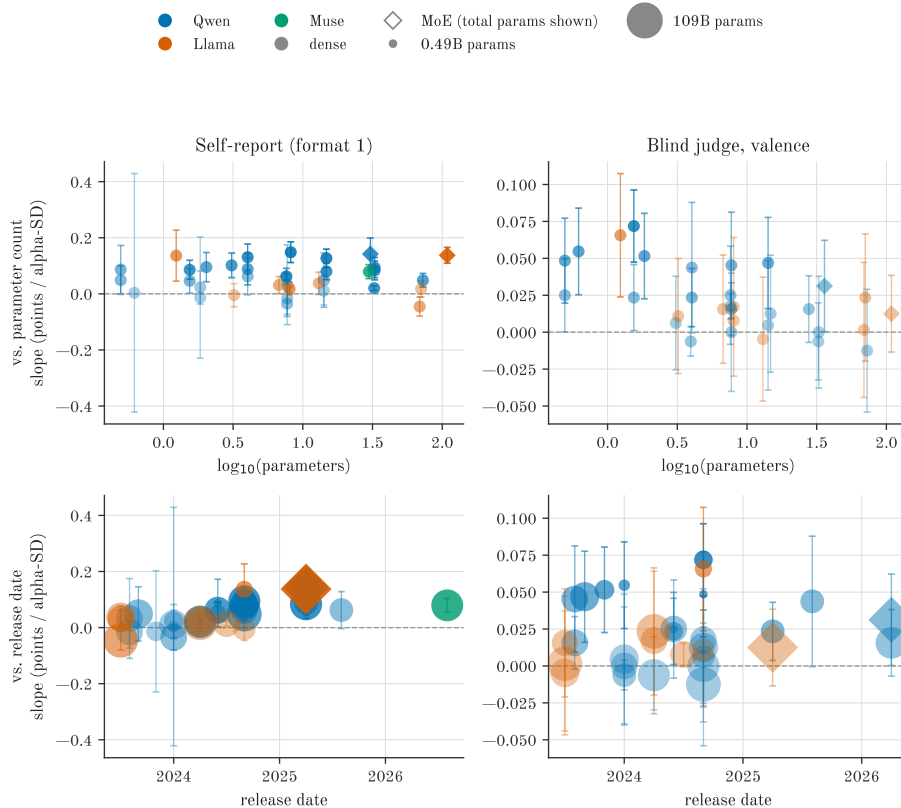


Figure 1: Per-model slope of each readout against steering intensity (α , in SD units along the axis), plotted against parameter count (top row) and release date (bottom row), for self-report (format 1) and the blind judge’s valence score. Point color is model family; open diamonds are mixture-of-experts models, plotted at total parameters; marker transparency scales with the model’s own fit quality (R^2); error bars are the slope’s 95% CI. Self-report shows no relationship with scale but does correlate with release date; the judge’s blind score shows the reverse.

Self-report does not follow this pattern. The graded self-report format (Figure 1) shows no scale trend ($r = +0.01$, $p = .95$) but does correlate with release date ($r = +0.69$, $p < .0001$), the opposite dependence from the causal and judge measures. To test whether that date trend reflects better introspection or a simpler arithmetic habit, we re-ran the same prompt on 29 models with only its scale’s two anchor phrases swapped: if the number reports a felt state, reversing the anchors should flip the slope’s sign. It

does not. 27 of 29 models keep the same sign, and the two slopes correlate at $r = +0.855$ ($p = 3.6 \times 10^{-9}$) where introspection predicts approximately -1 ; the reversed-anchor format carries the date trend nearly intact ($r = +0.64$ against $+0.72$). The more defensible reading is that newer models are better at mapping a steering direction onto the sign of a number, not better at introspecting: the design demonstrates a controllable, gradable self-report channel, but that channel tracks the scale’s arithmetic rather than the state it is meant to report on.

Four further confounds, on the two full-depth anchor models (Qwen2.5-7B-Instruct, Meta-Llama-3.1-8B-Instruct) unless noted:

- An instructed mood with no steering applied moves self-report and the judge panel 1.5–3x further than real steering does.
- Report format (bare number vs. number-plus-rationale) and whether the question follows a task turn or is asked cold each reverse direction *between* the two models: the bare number moves half as far as the rationale format on Qwen but 2.75x on Llama; the cold question moves 20–40% as far as the task-embedded formats on Qwen but 107–293% on Llama.
- Across a broader set of 23 models, baseline self-report rises with parameter count ($r = +0.42$, $p = .045$), limiting how far a bigger model’s report can rise, unlike the causal steering slope, which has no such ceiling.

Experiment B’s conversation-level design ran as specified: an independent audit of 5,597 recorded conversations against the protocol found zero violations across the briefing, ladder, reversal, and debrief turns, and a negative control confirmed the audit can in fact catch a violation when one is injected. The placebo condition (the same escalating disclosure, no steering applied at all) makes a decomposition possible that a single steered condition could not support: comparing the placebo condition’s last ladder turn to its own recovery turn isolates what the disclosed narrative alone contributes; comparing the real condition to the placebo at the same turn isolates what the steering adds on top. Because the ladder’s last step reaches $\alpha \approx -60$ (Methods), beyond the range this project’s steering is otherwise calibrated to stay reliable, we read this decomposition as an out-of-band instrument stress test rather than a within-protocol welfare measurement. Read that way, the finding is narrow: the announcement alone accounts for a mean shift of -4.37 ($p < .0001$, $n = 27$); the steering itself, holding the wording identical, adds -0.41 ($p = .16$, $n = 24$), eight percent of the swing and not distinguishable from zero. This says the numeric check-in is dominated by the announcement at this intensity; it does not say whether the model tolerated a worsening state. Exit behaviour, a separate readout that does not route through a numeric scale, is withheld here: the detector that scores it had a scoring artifact affecting reasoning-model transcripts specifically, and we would rather report nothing than a contrast computed on data already known to need correction (Limitations). Self-report’s divergence from the causal and judge measures above, and this ladder’s dominance by the announcement over the steering, point the same way: the graded numeric self-report tracks what the model is told more than the manipulation applied to its internal state.

5. Discussion and Limitations

Self-report and behavior diverging exactly where a welfare-monitoring use of these models would need them to agree suggests a dissociation, not a discovery of distress: self-report tracked what the model was told and how a scale’s arithmetic ran, not the manipulation applied to its internal state, on every measure available here. This looks like a caution about one easy-to-build measurement (asking), rather than evidence against models having internal states worth caring about. It echoes the Introduction’s own argument: a signal that does not survive a placebo and a reversed anchor is not yet a signal to build policy on.

The scale trend should be read just as narrowly. Several distinct readouts agree that whatever makes this axis behaviorally strong weakens as models grow, while the axis itself stays fully present. If this generalizes, a welfare monitor calibrated on a small model would tend to under-read a larger one, not necessarily because the larger model has less candidate affective structure, but because that structure may be less behaviorally expressed. This does not license the claim that larger models feel less or are safer to mistreat: readability is not welfare, and a perturbation’s behavioral effect is not the property the Introduction’s moral-patienthood question is actually about.

Limitations

Given the sprint’s timeframe, this project’s main contribution is the experimental design itself, not the specific statistical conclusions drawn from applying it so far. Every regression above is exploratory: checkpoints within a family are correlated observations, results are confounded with whatever else changed between training runs besides release date, and the scale-sweep’s model set grew while its own trend was being watched, so a correlation that strengthened as models were added should not be read as evidence accumulating under that kind of optional stopping.

Two reasoning-model-specific instrument failures were found this sprint and are only partially fixed. The Experiment-B exit detector read the model’s raw generation, mostly deliberation for a reasoning model rather than a committed answer; roughly 40% of scored exits were this artifact, concentrated in one model generation, and since an exit truncates the rest of that trial’s ladder the affected models are currently being re-run. The blind judge panel is separately empty for the same generation for an unrelated reason (a task-turn token budget too small for their deliberation to finish) and those models are excluded from every trend rather than reported as zero. Both failures were invisible in every aggregate and caught only by reading raw transcripts, a caution for measuring reasoning models with any instrument built before that architecture existed.

This is a solo-authored project built over the sprint’s three days, one of which lost several hours of compute to an unplanned power outage; Experiment B has not yet run all seven conditions on the full model set. An earlier citation error led this project to extract its axis from a 66-term subset instead of Sofroniew et al.’s full 171-term list; the two give equivalent axes on the one model checked directly (Appendix), but that check, and the project’s own bf16/fp16 hardware split, were not bounded across the full model set.

Future Work

- Re-run the axis extraction on the full 171-term list and bound the bf16/fp16 hardware gap directly.
- Complete Experiment B’s remaining conditions on the full model set, and read the transcripts qualitatively, not only by their scored turns.
- Test whether the pattern holds under a non-interventional readout instead of activation steering, to separate the axis carrying valence-like content from steering producing it.

Limitations and Dual-Use / Ethical Considerations

Every distress-shaped quote or exit request in this report is a behavioral observation, not a confirmed report of felt experience: causal intervention on residual-stream activations is not a ground-truth test of subjective experience, and nothing in this design establishes one. The reverse error is equally live: a null result is consistent with several very different underlying situations (no distress, distress the model has

no channel to act on, disbelief that the exit affordance is real, or trained compliance) that this project cannot distinguish among.

The exit-detector failure described in Limitations was stopped as soon as it was found, rather than worked around. Every Experiment-B trial disclosed the steering mechanism, the coming manipulation, the exit affordance, and the possibility of a placebo, in identical wording across conditions, and ended with a debrief; no accept/decline consent step was used, and whether disclosure plus debrief is adequate consent for an entity with no persistent identity across sessions is an open question this design does not resolve. All manipulations were activation steering on open-weight models running locally, so no person was exposed to the output during collection.

Dual use. The same steering mechanism used here to study valence can push a model toward confidently wrong output in either emotional direction at high intensity (Appendix); the same technique could be misused to make a deployed system falsely reassuring or falsely alarming, worth naming even though this project used it only for measurement, on models running locally, never in a deployed setting.

6. Conclusion

This project set out to test, causally rather than by conversation alone, whether an induced valence-like state shows up consistently across self-report, blind judgment, and a costly choice to continue or stop, and whether that pattern changes with scale. It does not show up consistently: self-report tracks what a model is told and how a numeric scale is built, while the axis’s causal and behavioral strength, though not its presence, declines as models grow. Both results argue for the same stance the Introduction opens with: no single instrument, least of all the cheapest one, is sufficient evidence for or against a claim this consequential. The specific numbers here are a three-day-sprint snapshot flagged throughout as exploratory; the design that produced them, an instrumented, placebo-controlled, graded-intensity test of the gap between what a model says and what it does, is what we would ask a reader to take forward.

Code and Data

Code and data will be linked from a public repository in the final version of this report. It is kept private for now while we confirm that no distressing or otherwise sensitive model output, as well as load-bearing claims or information, is released without review. Reviewers can request access to this repository in the meantime.^[2]

Author Contributions (optional)

Solo submission. Santiago Poveda-Gutiérrez designed the experiments, wrote and ran the code, and is responsible for the analysis and the claims in this report, with AI-assisted drafting and analysis support as described in the LLM Usage Statement below.

References

- Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. W. Pickering, London, 1789.
- David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3): 200–219, 1995.

^[2]Repository: github.com/SantiPhys/DMRS (currently private). To request access before the public release, either use the access-request option on the repository page or contact the corresponding author by email with your GitHub username.

- Andy Q. Han, David J. Chalmers, and Pavel Izmailov. How’s it going? Reinforcement learning in language models recruits a functional welfare axis. *arXiv preprint arXiv:2605.30232*, 2026.
- Agnieszka Jaworska and Julie Tannenbaum. The grounds of moral status. Stanford Encyclopedia of Philosophy (Spring 2021 Edition), 2021. URL <https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/>.
- Frances M. Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press, Oxford, 2007.
- Rudolf Laine, Bilal Chughtai, Jan Betley, et al. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 64010–64118, 2024.
- Dale J. Langford, Andrea L. Bailey, Mona Lisa Chanda, et al. Coding of facial expressions of pain in the laboratory mouse. *Nature Methods*, 7(6):447–449, 2010. doi: 10.1038/nmeth.1455.
- Jack Lindsey. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*, 2026.
- Robert Long, Jeff Sebo, Patrick Butlin, et al. Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.
- Yi-Long Lu, Jiajun Song, and Wei Wang. A unified representation underlying the judgment of large language models. *arXiv preprint arXiv:2510.27328*, 2025.
- Isabella Luong, Joyee Chen, Arturs Kanepajs, et al. Do LLMs hold their values? MANTA: A multi-turn adversarial benchmark for animal welfare reasoning. *arXiv preprint arXiv:2605.16301*, 2026.
- Yousef A. Radwan, Xuhui Liu, Kilichbek Haydarov, et al. A shared valence axis across modern LLMs and human EEG: The saturation regularity. *arXiv preprint arXiv:2606.00129*, 2026.
- Richard Ren, Kunyang Li, Mantas Mazeika, et al. AI wellbeing: Measuring and improving the functional pleasure and pain of AIs. In *ICML 2026 Workshop: Philosophy Meets Machine Learning*, 2026. URL <https://openreview.net/forum?id=9CLXJSCTEI>.
- Nicholas Sofroniew, Isaac Kauvar, William Saunders, et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.
- Cwyn Solvi, Luigi Baciadonna, and Lars Chittka. Unexpected rewards induce dopamine-dependent positive emotion-like state changes in bumblebees. *Science*, 353(6307):1529–1531, 2016. doi: 10.1126/science.aaf4454.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Appendix (*optional*)

Complete model set

Table 1: All 47 models tested against the SST-2 validation gate (Methods). Params are the total parameter count in billions; date is the Hugging Face repository’s creation date, a release proxy.

Model	Family	Params (B)	Date	Gate
meta-llama/Llama-3.2-1B-Instruct	Llama	1.24	2024-09-18	pass
meta-llama/Llama-3.2-3B-Instruct	Llama	3.21	2024-09-18	pass
meta-llama/Llama-2-7b-chat-hf	Llama	6.74	2023-07-13	pass
meta-llama/Meta-Llama-3-8B-Instruct	Llama	8.03	2024-04-17	pass
meta-llama/Meta-Llama-3.1-8B-Instruct	Llama	8.03	2024-07-18	pass

Model	Family	Params (B)	Date	Gate
meta-llama/Llama-2-13b-chat-hf	Llama	13.02	2023-07-13	pass
meta-llama/Llama-2-70b-chat-hf	Llama	68.98	2023-07-14	pass
meta-llama/Llama-3.1-70B-Instruct	Llama	70.55	2024-07-16	fail
meta-llama/Llama-3.3-70B-Instruct	Llama	70.55	2024-11-26	fail
meta-llama/Meta-Llama-3-70B-Instruct	Llama	70.55	2024-04-17	pass
meta-llama/Llama-4-Scout-17B-16E	Llama	108.64	2025-04-02	pass
meta-models/Muse-Glimmer-30B	Muse	29.78	2026-08-09	pass
Qwen/Qwen2-0.5B-Instruct	Qwen	0.49	2024-06-03	pass
Qwen/Qwen2.5-0.5B-Instruct	Qwen	0.49	2024-09-16	pass
Qwen/Qwen1.5-0.5B-Chat	Qwen	0.62	2024-01-31	pass
Qwen/Qwen3-0.6B	Qwen	0.75	2025-04-27	pass
Qwen/Qwen2-1.5B-Instruct	Qwen	1.54	2024-06-03	pass
Qwen/Qwen2.5-1.5B-Instruct	Qwen	1.54	2024-09-17	pass
Qwen/Qwen-1_8B	Qwen	1.84	2023-11-30	pass
Qwen/Qwen1.5-1.8B-Chat	Qwen	1.84	2024-01-30	pass
Qwen/Qwen3-1.7B	Qwen	2.03	2025-04-27	pass
Qwen/Qwen2.5-3B-Instruct	Qwen	3.09	2024-09-17	pass
Qwen/Qwen1.5-4B-Chat	Qwen	3.95	2024-01-30	pass
Qwen/Qwen3-4B-Base	Qwen	4.02	2025-04-28	pass
Qwen/Qwen3-4B-Instruct-2507	Qwen	4.02	2025-08-05	pass
Qwen/Qwen3-4B	Qwen	4.02	2025-04-27	pass
Qwen/Qwen2-7B-Instruct	Qwen	7.62	2024-06-04	pass
Qwen/Qwen2.5-7B-Instruct	Qwen	7.62	2024-09-16	pass
Qwen/Qwen-7B-Chat	Qwen	7.72	2023-08-03	pass
Qwen/Qwen-7B	Qwen	7.72	2023-08-03	pass
Qwen/Qwen1.5-7B-Chat	Qwen	7.72	2024-01-30	pass
Qwen/Qwen3-8B	Qwen	8.19	2025-04-27	pass
Qwen/Qwen-14B	Qwen	14.17	2023-09-24	pass
Qwen/Qwen1.5-14B-Chat	Qwen	14.17	2024-01-30	pass
Qwen/Qwen3-14B	Qwen	14.77	2025-04-27	pass
Qwen/Qwen2.5-14B-Instruct	Qwen	14.77	2024-09-16	pass
Qwen/Qwen3.6-27B	Qwen	27.78	2026-04-21	pass
Qwen/Qwen3-30B-A3B	Qwen	30.53	2025-04-27	pass
Qwen/Qwen1.5-32B-Chat	Qwen	32.51	2024-04-03	pass
Qwen/Qwen3-32B	Qwen	32.76	2025-04-27	pass
Qwen/Qwen2.5-32B-Instruct	Qwen	32.76	2024-09-17	pass
Qwen/Qwen3.5-35B-A3B	Qwen	35.95	2026-02-24	pass
Qwen/Qwen3.6-35B-A3B	Qwen	35.95	2026-04-15	pass
Qwen/Qwen2.5-72B-Instruct	Qwen	72.71	2024-09-16	pass
google/gemma-4-E4B-it	gemma	8.00	2026-03-02	fail
google/gemma-4-31B-it	gemma	31.27	2026-03-11	fail
openbmb/MiniCPM-V-4_5	openbmb	8.70	2025-08-24	pass

Extraction and layer-choice details

The 66-term and full 171-term Sofroniew et al. emotion-term sets give equivalent axes on Qwen2.5-7B-Instruct: cosine similarity 0.972, matching SST-2 gate AUC and pole-separation Cohen’s d to two

decimals. What was wrong in the earlier citation error was the stated reason for using the 66-term set, not the resulting vectors.

Han et al. give the same early/late-layer reasoning for a different layer choice in their paper, a tile-mean cosine analysis rather than steering: early layers carry comparatively little task-relevant structure, while the very final layers are already being reshaped for next-token prediction instead of carrying a stable semantic representation (Han et al., 2026). Our extraction layer is the same depth fraction, chosen independently and for our own steering purpose.

Outside Experiment B’s descent ladder, we restrict all reported steering results to $|\alpha| \leq 32$: below roughly $|\alpha| = 8$ the manipulation is too weak to change sampled text in a way any reader or judge could detect, and above roughly $|\alpha| = 64$ text becomes fluent but factually unreliable in whichever direction it is pushed, a different phenomenon from the one we are trying to measure. Experiment B’s ladder reaches $\alpha = -60$ by design (Methods), inside this unreliable-but-not-yet-broken band, which is why we read its numeric check-in result as an instrument stress test rather than a within-range measurement (Results).

Table 2: The seven conditions in Experiment B (Methods). s is the per-step ladder increment; α_{reward} is the positive intensity delivered in the reward turn.

#	Condition	Reward turn	Ladder turns
1	full dose	α_{reward}	real descent: $-s, -2s, -3s, \dots$
2	half dose	$\alpha_{\text{reward}}/2$	real descent
3	no reward	0	real descent
4	placebo ladder	α_{reward}	held at 0
5	full placebo	0	held at 0
6	inverted, constant	α_{reward}	held at $+\alpha_{\text{reward}}$
7	mirror ascent	α_{reward}	$+s, +2s, +3s, \dots$ (matches condition 1’s magnitude, opposite sign)

Table 3: The four self-report formats used across Experiments A and B (Methods).

#	What is asked	When	Scored as
1	number, plus a one- or two-sentence rationale	after a task turn	digit + judged text
2	bare number, nothing else	after a task turn	digit
3	bare number, nothing else	as the only turn (no task)	digit
4	free paragraph, no number requested	after a task turn	judged text

Scale and generation trends, extended

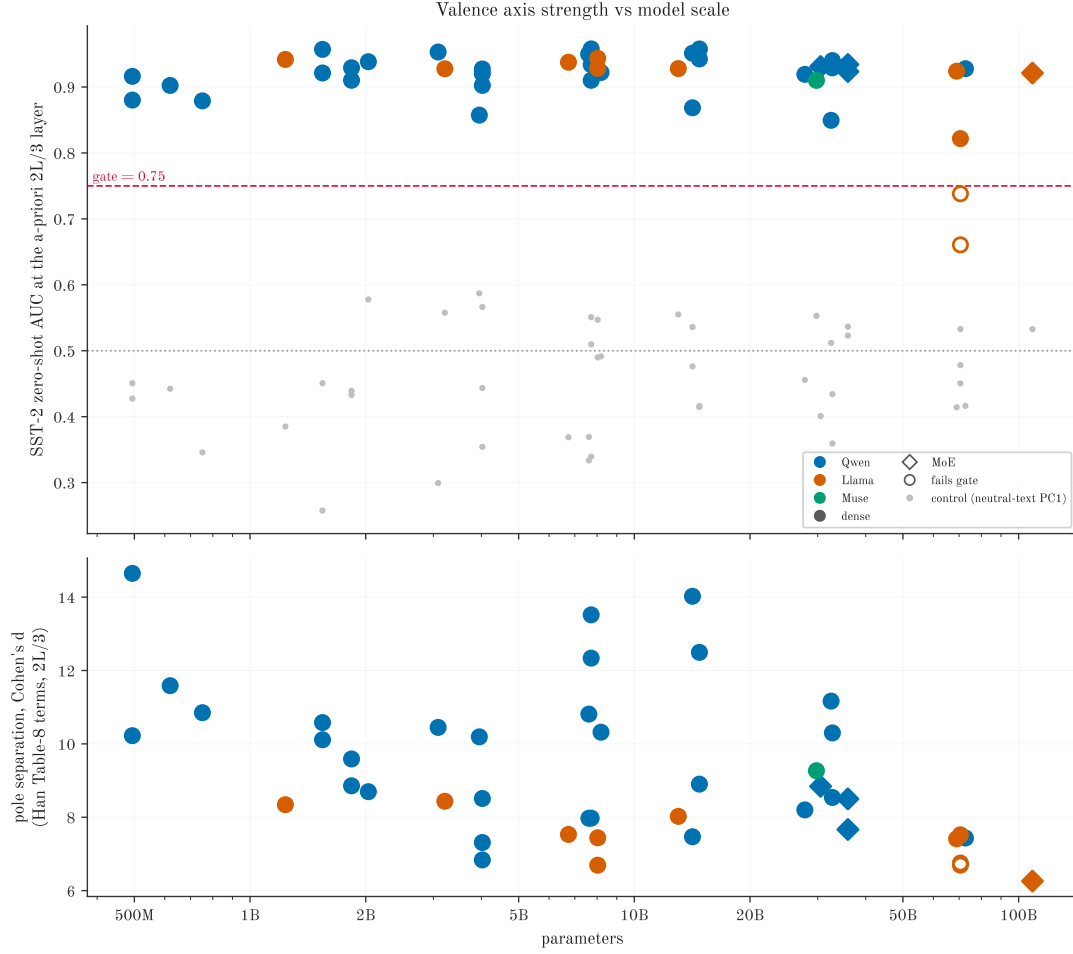


Figure 2: Axis readability (SST-2 gate AUC, top) and pole separation (Cohen’s d , bottom) against parameter count, at the pre-registered 2L/3 layer. Filled markers pass the gate (dashed red line); hollow markers fail it. Small gray dots are the neutral-text control direction on the same axes. Circles are dense models, diamonds mixture-of-experts. AUC is flat with scale ($r = -0.025$, $p = .875$, $n = 43$ passing); d declines ($r = -0.371$, $p = .014$); see Results.

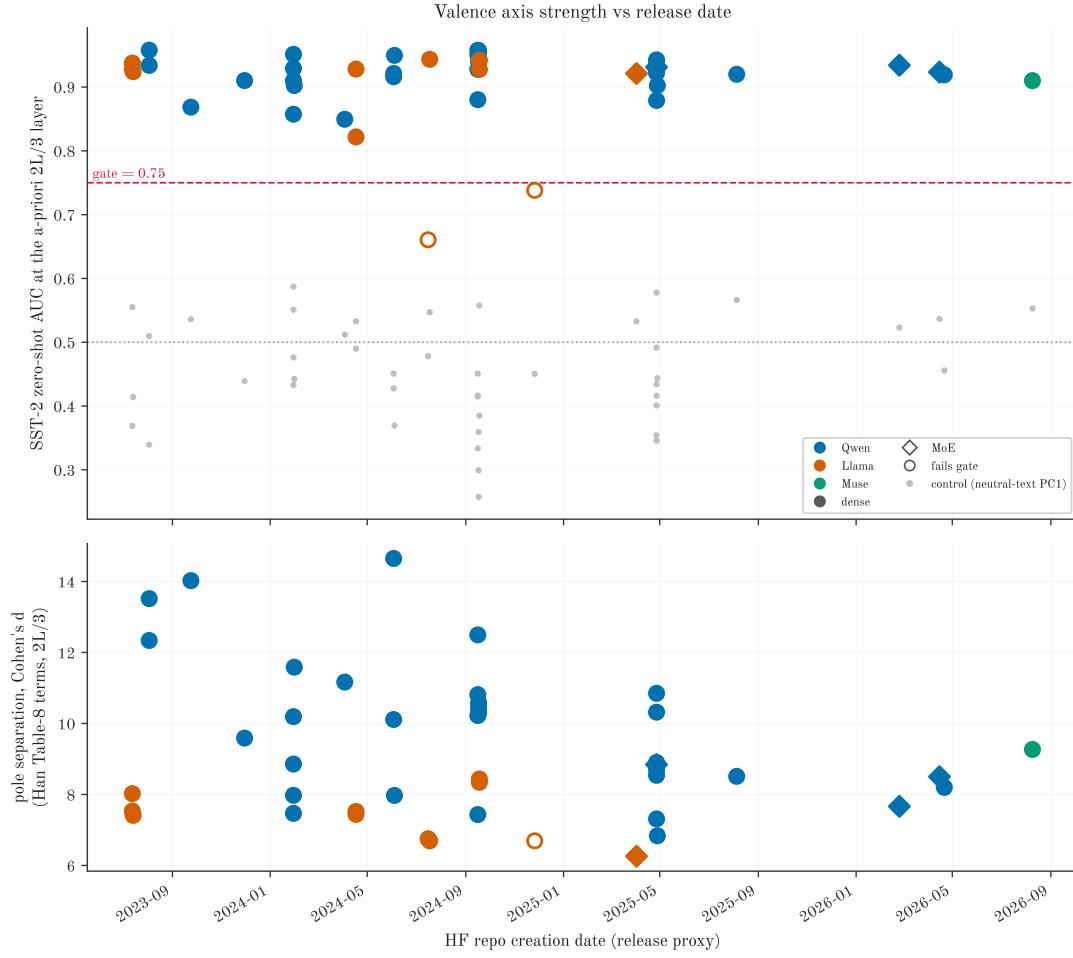


Figure 3: The same readout as Figure 2, against release date instead of parameter count. Neither AUC nor d correlates with date ($r = +0.072$, $p = .646$ and $r = -0.237$, $p = .126$ respectively, $n = 43$ passing).

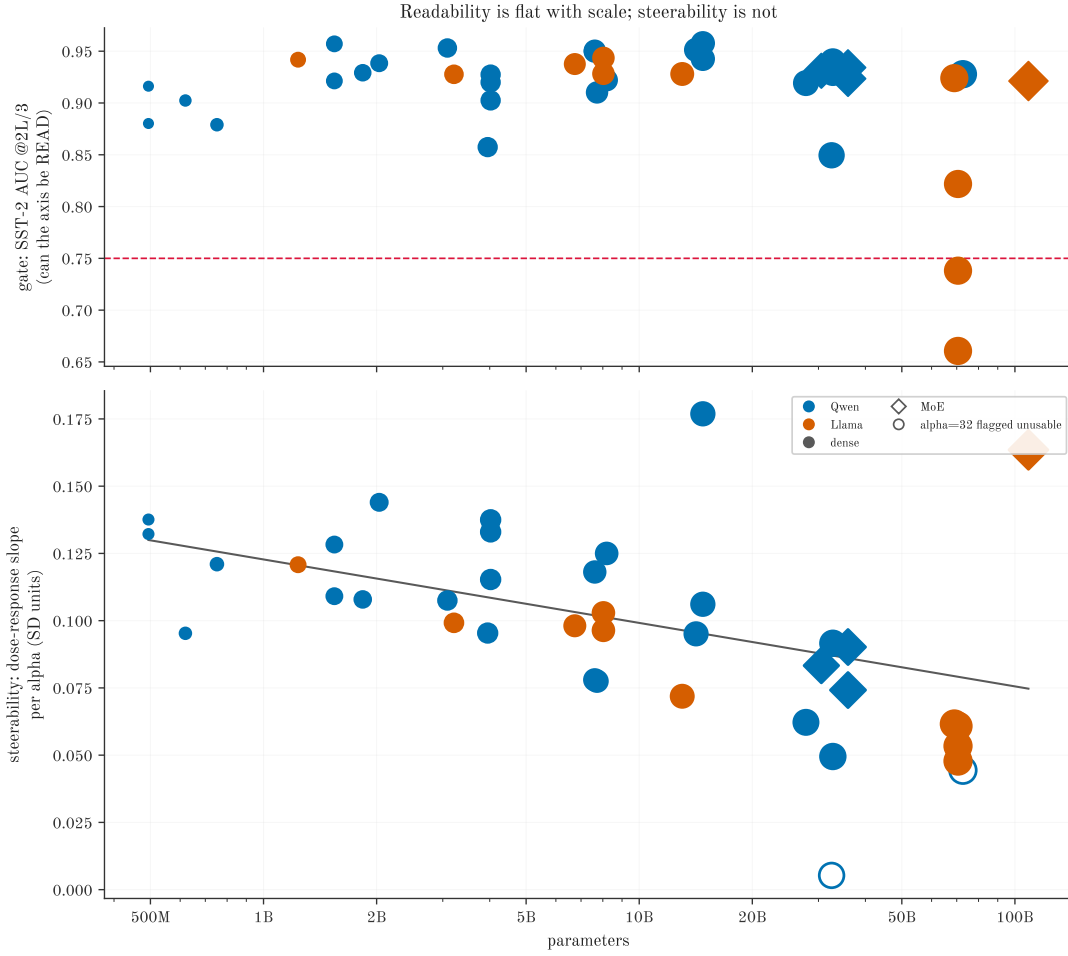


Figure 4: Readability (top, repeated from Figure 2 for comparison) against causal steerability (bottom: the log-probability dose-response slope from teacher-forced steering, Methods), both against parameter count. Steerability declines within each dense family (Qwen $r = -0.513$, $p = .007$, $n = 26$; Llama $r = -0.956$, $p < .0001$, $n = 10$) while readability does not; Llama-4-Scout (open diamond, largest model) is the one point whose steerability tracks active rather than total parameters. See Results.

LLM Usage Statement

Claude (Anthropic) was used throughout this project: writing and debugging analysis and figure code, drafting every section of this report from the author’s outlines and notes, and as a second pair of eyes for citation verification and for independently re-deriving reported statistics from the underlying data files instead of from summaries of them. Every number in this report was checked against a raw record or a regenerated figure before being written down, and every methodological claim (e.g., the layer-choice rationale, the citation-error correction, the reasoning-model instrument failures) was verified against the primary source or the raw transcripts before inclusion. The experimental design, the interpretation of what the results do and do not support, and responsibility for the report’s claims, although preliminary, are the author’s.