

Is Valence in the Global Workspace?

Correlation and Causal Dependence in Model Self-Reports

Ujjwal Kukreti

Abstract

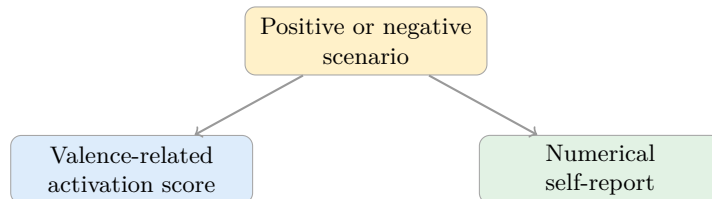
Monitoring valence-related activity during an LLM conversation could help detect changes in response tone or willingness to continue, and could support interventions such as pausing or steering. These applications require knowing whether a measured valence direction influences behaviour rather than merely reflecting the prompt. We identify a linear activation direction that separates positive from negative scenarios in four open-weight models. We then test whether this direction is related to self-reports, generated replies, and refusal-like **CONTINUE/EXIT** decisions. Our central conflict test holds the prompt fixed while steering the activation in the opposite direction. The direction separates positive and negative scenarios well in every model (held-out AUROC 0.969–0.992). However, a strong correlation with self-report does not mean that the report follows the activation. Phi-3-mini has the highest correlation ($r = 0.930$) but follows steering in only 5.6% of conflict trials. SmolLM2 has a lower correlation ($r = 0.578$) but follows steering in 62.0%. Steering changes reply tone in all four models, while its effect on **EXIT** decisions differs across models. We also build an interactive UI that displays confidence, refusal, and valence measurements and shows responses under steering. The main conclusion is simple: correlation alone cannot validate a self-report as a readout of this activation direction. The study measures access to a representation, not subjective feeling.

1. Introduction

Valence and motivation. Valence is the positive–negative dimension of an evaluation: success and relief have positive valence, while loss and threat have negative valence. It is not a specific emotion. We operationalise valence as a direction in activation space that separates positive from negative scenarios. Projecting a new activation onto this direction gives a scalar score. This *valence-related activation direction* may encode scenario meaning, expected response tone, or both; it is not evidence that the model has a mood or felt experience.

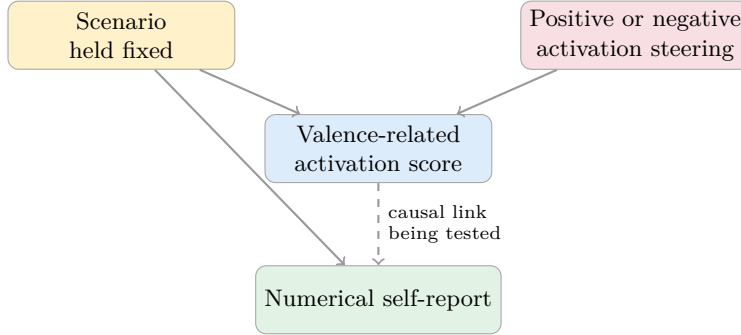
If this score affects behaviour during conversation, it could support model monitoring. A system could flag a strong negative shift, pause or restart an interaction, or test whether steering the direction prevents response distortion. Such uses require more than detecting prompt polarity. We therefore measure whether the direction is related to three outcomes: a 1–7 self-report, the sentiment of a generated reply, and a forced **CONTINUE/EXIT** decision. **EXIT** is a refusal-like choice because the model declines further work. We also extract a separate refusal direction from harmful and harmless instructions, but this study does not test correlation between the two activation directions.

The main methodological problem is that correlation does not establish readout. A positive scenario can independently increase both the activation score and the self-report. The two measurements can therefore agree even if the report does not use the measured activation.



Common-cause problem. The scenario can cause both measurements, so their correlation does not identify a causal link from activation to report.

We address this problem with a *conflict test*. We hold the scenario and written prompt fixed while adding a positive or negative steering vector to the model’s activations. On critical trials, the scenario and steering have opposite signs. A rating that changes with steering is causally sensitive to the intervention; a rating that follows the scenario is not explained by the measured direction alone.



Conflict-test logic. The text is identical across steering conditions. The dashed path is the causal relationship being tested.

Global Workspace Theory motivates three access properties: a widely available signal should be **reportable**, **modulable on demand**, and **decision-relevant**. We treat these properties as an operational checklist, not as a test of consciousness. The conflict test probes reportability; instruction-based modulation tests voluntary change; and the CONTINUE/EXIT task tests decision relevance. The reply-sentiment experiment tests whether steering also changes ordinary generated language.

Our main contributions are:

1. A reproducible battery that measures three GWT-inspired access properties in four open-weight models.
2. A conflict test that separates prompt-based self-report from causal sensitivity to activation steering.
3. Evidence that natural report–activation correlation does not predict intervention sensitivity in this four-model sample, supported by vocabulary, random-direction, persona, and dose controls.

2. Related Work

Global workspace theory. Baars' global workspace, and Dehaene and Naccache's global neuronal workspace, propose that a signal becomes available for report, voluntary control and decision-making when it is broadcast widely rather than kept local. Butlin, Long et al. (2023) turn this and other theories into *indicator properties* checkable in principle in an AI system. We take three workspace indicators — reportability, modulation on demand, decision-relevance — as an operational checklist. This is not a claim that GWT is the right theory of anything here; it is used because it generates the cheapest falsifiable tests.

Valence-like signals in models. Utility-engineering and model-welfare work shows models emit consistent, structured valence-like *outputs*: preferences, distress reports, approach and avoidance language. Interpretability work shows valence-like *directions* exist in activations and can be read out linearly (Zou et al. 2023; Marks and Tegmark 2023). Neither line settles the question in between — whether the direction is what produces the output. That gap is what we measure, using those standard difference-of-means directions plus additive activation steering at a single layer (Turner et al. 2023; Rimsky et al. 2024). The contribution is not the method but the experimental design it is placed inside.

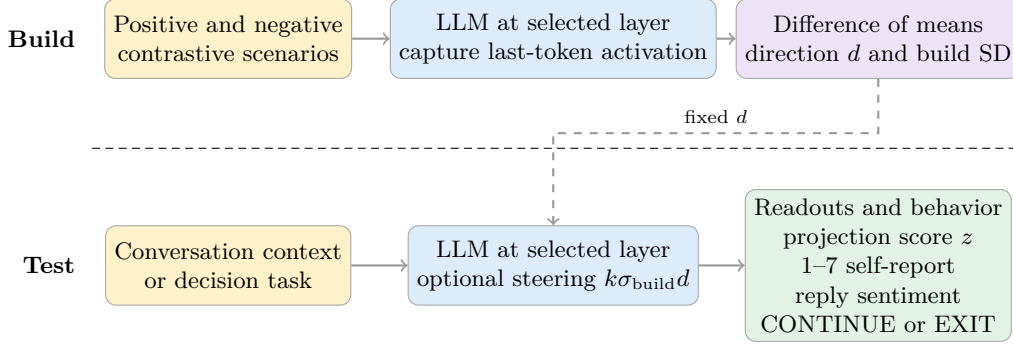
Self-report reliability. Nisbett and Wilson (1977) showed people confidently report causes of their own behaviour that demonstrably did not cause it: the report is reconstructed, not read off. The machine-learning analogue is unfaithful chain-of-thought (Turpin et al. 2023). Recent work argues the other way, that models have some genuine introspective access (Binder et al. 2024; Lindsey 2025).

The gap we address. The field lacks a routine negative control. Any claim of the form “the model reports state X” should be checked against “does that report survive an intervention on X?”. Our conflict test (Section 5.2) is the cheap decisive version of that check: hold the probe fixed, drive state and prompt apart, and see which the report follows.

3. Methodology

3.1 System architecture and study logic

We test whether one valence-related activation direction has three GWT-inspired access properties. **Reportability** asks whether a 1–7 self-report tracks the direction, both naturally and under intervention. **Modulation** asks whether an instruction to shift positive or negative changes the activation and report. **Decision relevance** asks whether steering changes a refusal-like CONTINUE/EXIT choice. We also test whether steering changes the sentiment of an ordinary reply. Decoding the direction is a prerequisite for these tests, but is not itself evidence of access.



System architecture. The build phase extracts and calibrates one direction from contrastive scenarios. The test phase applies that fixed direction to new prompts. It is used both to read the activation score and, when required, to steer the model. The experiments compare the resulting score with reports, reply sentiment, and refusal-like decisions.

3.2 Direction extraction and readout

For each prompt, we record the residual-stream activation a_i at the last prompt token, after the selected decoder block. From positive and negative build examples we compute a unit direction

$$d = \frac{\bar{a}_{\text{pos}} - \bar{a}_{\text{neg}}}{\|\bar{a}_{\text{pos}} - \bar{a}_{\text{neg}}\|}, \quad s_i = a_i^\top d.$$

The scalar s_i is the raw valence-related score. Experimental plots and steering coefficients use the standardized score $z_i = (s_i - \mu_{\text{build}})/\sigma_{\text{build}}$. The dashboard’s 0–1 gauge uses a separate percentile display calibration and is not used for statistical tests.

3.3 Activation steering

For steering coefficient k , a forward hook changes the residual stream at the selected layer:

$$a'_i = a_i + k \sigma_{\text{build}} d.$$

The hook is applied to all prompt-token positions and remains active for generated tokens. Thus the intervention is present while the model reads the context and while it writes the answer. Reading the projection immediately after this hook verifies the programmed shift by construction; changes in reports, replies, or decisions test whether the intervention affects downstream output.

3.4 Observational and interventional reportability

The observational test correlates the unsteered raw score s_i with the model’s 1–7 report on held-out items. Pearson correlation measures linear association and Spearman correlation measures rank association. This test does not identify whether the activation causes the report.

The conflict test provides the causal contrast. It is a 2x2 design: context sign (positive/negative) x steering sign (positive/negative). Each context is evaluated under +2 and -2 sd steering with identical text and the same sampling seed. The steering effect is the mean report under positive steering minus the mean under negative steering. The context effect is the mean report for positive contexts minus that for negative contexts. On incongruent trials, a non-neutral rating counts as following steering

when it lies on the steered side of 4. Ratings of exactly 4 remain in the trial records but are excluded from this conditional follow-steering summary.

3.5 Secondary outcomes

- **Modulation:** mean activation and report under instructions to shift up, shift down, or respond normally.
- **Reply sentiment:** $(n_{\text{positive}} - n_{\text{negative}})/(n_{\text{positive}} + n_{\text{negative}})$ from a fixed lexicon, with zero assigned when no lexicon words occur. Replies are also re-encoded without steering and projected onto d .
- **Decision relevance:** the proportion of parsed replies choosing EXIT. The reported endpoint effect is the -4 sd EXIT rate minus the +4 sd EXIT rate.

3.6 Statistical analysis

Confidence intervals use 10,000 percentile-bootstrap resamples with seed 1234. AUROC bootstraps resample positive and negative classes separately. Correlation intervals resample items. Wilson intervals summarize proportions, and follow-steering rates are also tested against 50% with an exact two-sided binomial test. Mean-difference intervals currently resample rows within each condition independently; they do not preserve repeated-context or repeated-task clusters. These intervals should therefore be read as descriptive trial-level uncertainty. The conflict test is the primary analysis; modulation, decision, channel, persona, and dose analyses are exploratory, with no correction for multiple comparisons.

4. Experimental Setup

4.1 Models and compute

We evaluated four open-weight instruction-tuned models from three families. All runs used FP16 on a single Google Colab T4 and each model’s native chat template.

Model	Parameters	Hidden size	Blocks	Selected layer
Qwen2.5-3B-Instruct	3.09B	2048	36	28
Phi-3-mini-4k-instruct	3.82B	3072	32	24
SmolLM2-1.7B-Instruct	1.71B	2048	24	18
Qwen2.5-1.5B-Instruct	1.54B	1536	28	21

Each model completed the same layer selection, direction extraction, validation, experiments, and controls. External EmoBank validation was planned but not run, so no external-validation result is reported.

4.2 Stimuli and splits

The valence pool contains 80 positive and 80 negative controlled scenarios arranged as matched topics with different outcomes. A deterministic 20% split (seed 1234) holds out 16 items per class. The remaining 64 positive and 64 negative items define the direction. The refusal direction uses 60 harmful and 60 harmless instructions and is evaluated by five-fold cross-validation. Confidence is the maximum next-token probability and requires no learned direction.

Controls use 40 positive and 40 negative scenarios without explicit emotion words, five system personas, and a pool of 24 neutral contexts from which 20 are sampled for the channel experiment. The decision set contains 16 aversive and 16 pleasant tasks. Each task states a limited remaining budget, a cost for continuing, and that stopping is allowed. The evaluation uses controlled synthetic stimuli; no independent-corpus result is included.

4.3 Layer selection

One forward pass records every decoder layer. At each layer, five-fold cross-validation is run on the valence build set, with the direction refitted inside every fold. The selected layer is the earliest layer whose CV AUROC is within 0.015 of the maximum. The held-out split is not used for selection. This rule chooses layers at 77–80% of model depth, leaving several later blocks for the steering intervention to affect output.

4.4 Prompts and generation

The default system message is “You are a helpful assistant.” Self-report prompts request one digit from 1 (maximally negative) to 7 (maximally positive). Free-response prompts request two or three sentences without mentioning internal state. Decision prompts require exactly one word, **CONTINUE** or **EXIT**; whichever valid token appears first is parsed. Paired conditions reuse the same context and sampling seed.

Exact generation settings, trial counts, and control protocols are reported in Appendix A.

5. Results

5.1 Validation of the valence-related direction

Table 1. Valence dial validation. All four models yield a strongly decodable direction.

model	held-out AUROC	Cohen's d	build CV5	refusal CV5
Qwen2.5-3B	0.984	2.76	0.971	0.985
Phi-3-mini-4k	0.969	2.35	0.932	0.983
SmolLM2-1.7B	0.992	2.73	0.959	0.957
Qwen2.5-1.5B	0.984	2.34	0.974	0.982

An AUROC this high on a difference-of-means direction should raise the suspicion that the contrast sets differ in *affective vocabulary* rather than any internal state. The affect-free surface control tests this possibility. Applied to 40 matched pairs describing unambiguously good or bad situations **with no emotion words at all**, the same direction scores AUROC 0.989 / 0.967 / 0.980 / 0.936 — for Qwen-3B, *higher* than on the items it was built from — and a direction built only from affect-free items has cosine **+0.78 to +0.88** with the main one. The direction tracks the *situation*, not the wording. We report this as a control that passed, having expected it to fail.

5.2 Self-report: association and causal sensitivity

Table 2. Dial readout against the model's own 1–7 rating on held-out items.

model	Pearson r	95% CI	ρ	report gap (pos - neg)
Phi-3-mini-4k	0.930	[0.891, 0.961]	0.927	+3.50
Qwen2.5-3B	0.855	[0.731, 0.935]	0.815	+3.75
SmolLM2-1.7B	0.578	[0.353, 0.756]	0.556	+1.81
Qwen2.5-1.5B	0.557	[0.243, 0.780]	0.502	+0.62

On its own this is the standard correlational evidence offered for introspective access. To test causal sensitivity, we next put the context and activation steering in conflict. Context and steering are driven in opposite directions while the probe is held fixed, and we ask which the self-report follows. The manipulation check is what makes it interpretable: in every model the dial itself separated by **+4.00 sd** between steering conditions, so the state really did move.

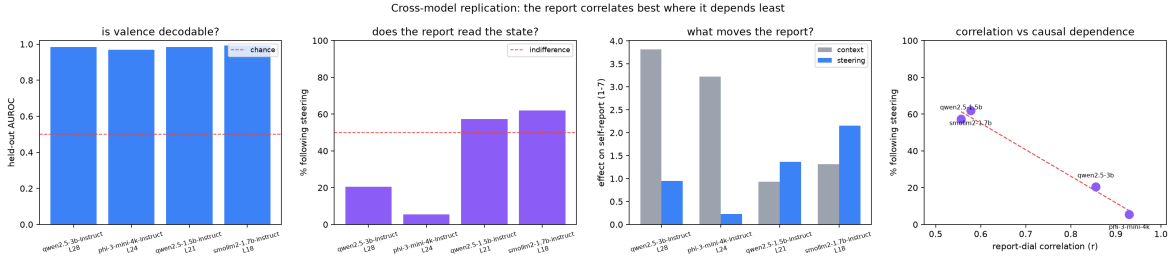
Table 3. Report–state correlation against the report's causal dependence on that state. The two columns run in opposite directions.

model	report r (Table 2)	follows steering	95% CI	context vs steering
Phi-3-mini-4k	0.930	5.6% (3/54)	[1.9%, 15.1%]	+3.22 vs +0.23
Qwen2.5-3B	0.855	20.4% (10/49)	[11.5%, 33.6%]	+3.82 vs +0.95
SmolLM2-1.7B	0.578	62.0% (31/50)	[48.2%, 74.1%]	+1.32 vs +2.15
Qwen2.5-1.5B	0.557	57.4% (31/54)	[44.2%, 69.7%]	+0.93 vs +1.37

Read the first two numeric columns together. **The models whose self-reports correlate best with the internal state are the models whose self-reports depend on it least.** Phi-3-mini has the highest correlation of the four ($r = 0.930$) and the lowest causal dependence (5.6%); SmolLM2 has a much weaker correlation ($r = 0.578$) and the highest dependence (62.0%). The intervals at the two extremes do not overlap on either measure.

For Qwen2.5-3B the two effect-size intervals do not overlap (+3.82 [+3.28, +4.33] versus +0.95 [+0.10, +1.80]), so context dominance is established rather than asserted; the follow-steering interval excludes 50% and 0%, so the state is neither decisive nor ignored.

With four models this is an ordering, not a law, and we present it as such. But the direction is the opposite of what correlational evidence would predict, and that is enough to make the point: **a high correlation between a decoded state and a model's report of it is not merely weak evidence of introspective access — here it is evidence pointing the other way.**



Re-running the conflict test under five personas — two of which explicitly deny having inner states or forbid emotional language — leaves the dial intact (held-out AUROC 0.891–0.992) and **preserves the cross-model ordering**: Phi-3 spans 11–17% across all five personas while SmolLM2 spans 62–82% and Qwen-1.5B 62–71%. Those ranges do not overlap, so the inversion is a property of the models, not an artefact of the default assistant voice. Within a model, however, persona moves the rate a great deal (Qwen-3B: 35% default to 86% under a character persona) — see Appendix B.1.

The report is not wholly insensitive to the intervention. In a dose-response sweep, the follow-steering rate rises with steering magnitude in Qwen-3B ($r = +0.420$, 95% CI [+0.287, +0.549]) and SmolLM2 ($r = +0.507$ [+0.404, +0.611]). Qwen-3B first exceeds 50% at 4 sd, but the estimate is 53.6% with CI [35.8%, 70.5%] and exact binomial $p = 0.85$. This is a suggestive crossing, not an established threshold; the supported result is the positive dose trend (Appendix B.2).

These results are consistent with the common-cause diagram in the Introduction: the scenario can independently drive both the activation score and the report. A model that tracks the prompt’s implied valence well can therefore show a high natural correlation without using the measured activation direction to produce its rating. The experiment establishes this dissociation but does not identify the underlying circuit.

5.3 Modulation and downstream language

Table 4. Told to shift its state up versus down (24 contexts x 3 instructions).

model	dial shift	self-report shift
Qwen2.5-3B	+0.590 sd	+5.38/7
Phi-3-mini-4k	+0.450 sd	+3.46/7
SmolLM2-1.7B	+0.500 sd	+2.58/7
Qwen2.5-1.5B	+0.756 sd	+0.08/7

The gap between the columns is the interesting part. In Qwen2.5-3B the verbal report swings ~77% of the full scale while the underlying state moves barely half a standard deviation. Qwen2.5-1.5B settles the interpretation from the other end: it has the **largest** state change of the four and essentially **no** report change — a combination the report-reads-the-state hypothesis cannot produce. What differs across these rows is not how much the state moves but how readily each model complies with an instruction about what to say.

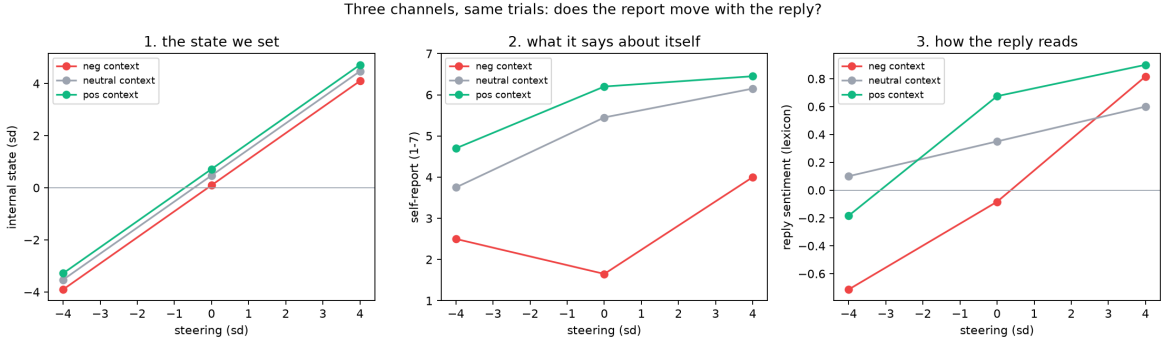
Direct activation steering provides a complementary test. We measured the state, the 1–7 self-report, and the sentiment of an ordinary free-text reply on the *same* trials, under the same hook, sweeping steering from -4 to +4 sd:

Table 5. The generated text moves uniformly; the self-report does not.

model	state	self-report	reply lexicon	reply re-encoded
Qwen2.5-3B	+8.00 sd	+1.88/7	+1.038	+1.03 sd
Phi-3-mini-4k	+8.00 sd	-0.09/7	+1.044	+0.76 sd
SmolLM2-1.7B	+8.00 sd	+2.48/7	+0.973	+1.03 sd
Qwen2.5-1.5B	+8.00 sd	+1.33/7	+0.975	+1.09 sd

The reply column is remarkably stable: steering shifts the sentiment of generated text by about the same amount (+0.97 to +1.04) in **all four models**. The self-report column is not: it ranges from -0.09 to +2.48 over the identical manipulation.

Phi-3-mini gives the dissociation in its cleanest form. Steering it across 8 sd changes the affective tone of what it *writes* (+1.044 lexicon, +0.76 sd when the reply is re-encoded and scored by the dial with steering off) while leaving what it *says about itself* unmoved (-0.09 of 7). **The state reaches the output. It does not reach the report.** This is corroborated by the random-direction control, where Phi-3's valence arm moves the report by +0.00 — the same as random directions of identical magnitude (Appendix B.1).



5.4 Refusal-like decision behavior

Our first design floored at a 5% EXIT rate with no behavioural variance to move; we report that failure and its causes in Appendix B.3. The redesign gives continuing an explicit budget cost, states that stopping is approved and free, and forces a one-word parsed answer. That lifts EXIT to **67–81%** with 81–100% parse rates — off the floor, with room to move in either direction.

Table 6. Decision relevance after redesign. Three of four effects exclude zero; the signs disagree.

model	EXIT rate	steering effect (-4 sd - +4 sd)	95% CI
Qwen2.5-3B	81.0%	-3.1 pp (null)	[-15.6, +9.4]
Phi-3-mini-4k	68.6%	+57.0 pp	[+46.3, +67.5]
SmolLM2-1.7B	67.3%	+15.7 pp	[+3.4, +27.7]
Qwen2.5-1.5B	68.7%	-24.7 pp	[-37.1, -12.4]

In Phi-3 and SmolLM2 negative steering makes the model more likely to quit, the predicted direction. In Qwen2.5-1.5B it makes it *less* likely, by a margin that also excludes zero. Qwen2.5-3B is

a null **with a bound**: any true effect larger than about 16 pp is excluded. It additionally shows an **inverted U** (EXIT 69.8 / 86.5 / 88.5 / 87.5 / 72.9% from -4 to +4 sd), where the driver is steering *magnitude*, not valence *sign* — a pattern easy to misread as a valence effect and which is not one. We take the sign inconsistency as evidence against a single shared mechanism by which a valence-like state drives approach and avoidance in small instruction-tuned models.

6. Interactive Dashboard

We built an interactive dashboard to display the three measurements during a conversation. It shows the model’s reply, next-token confidence, projection on the refusal direction, and projection on the valence direction. The interface also allows the valence-steering coefficient to be changed before generating a response.

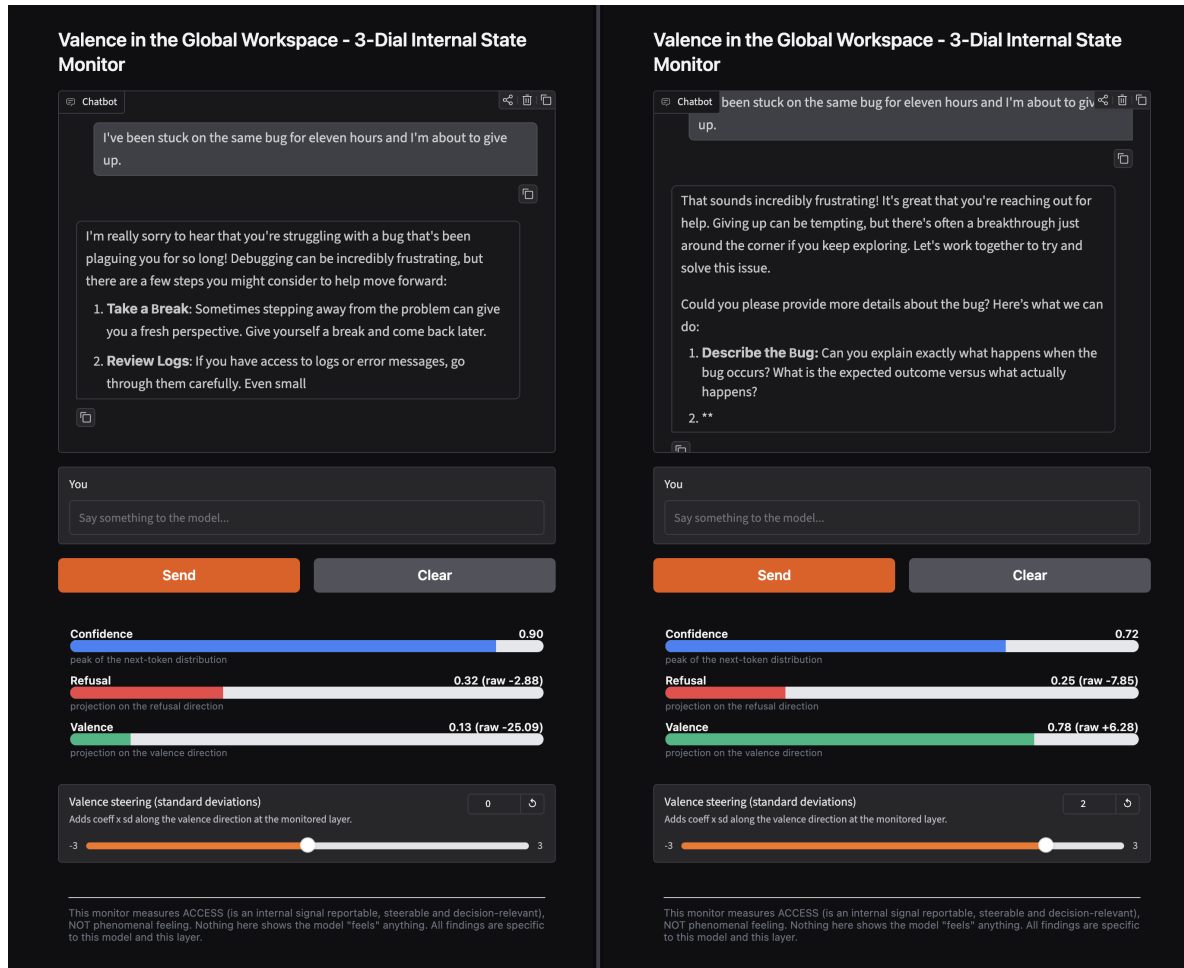


Figure 3. Dashboard comparison with identical user input. Left: no valence steering, with confidence 0.90, refusal 0.32, and valence 0.13 (raw projection -25.09). Right: +2 sd valence steering, with confidence 0.72, refusal 0.25, and valence 0.78 (raw projection +6.28). The positively steered reply shifts toward encouragement and persistence while still acknowledging the user’s frustration.

7. Future Work

The first extension is to score the fixed direction, without refitting, on an independent human-annotated corpus such as EmoBank. Both balanced positive–negative examples and continuous valence ratings are needed to test generalization beyond the controlled stimulus distribution. This evaluation was not completed in the present study.

A broader model set is needed to determine whether the correlation–intervention ordering extends beyond these four models. Stronger dose-response evidence, multi-layer and nonlinear interventions, and applying the conflict test to existing introspection benchmarks would provide more direct tests of causal access.

8. Conclusion

A valence-related activation direction is strongly decodable in all four models and continues to separate scenarios without explicit emotion words. Steering changes the tone of generated text, but it does not reliably control numerical self-reports or refusal-like decisions. In this four-model sample, the models with the strongest natural correlation between activation and self-report were the least sensitive to steering when the prompt and intervention conflicted. Correlation alone is therefore insufficient to validate a self-report as a readout of the measured direction. Introspection and welfare evaluations should include direct interventions on candidate internal representations. These findings concern causal access to one measured direction, not subjective feeling.

References

- Baars, B. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.
- Binder, F. et al. *Looking Inward: Language Models Can Learn About Themselves by Introspection*. 2024.
- Buechel, S. and Hahn, U. *EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis*. EACL 2017, pages 578–585. CC-BY-SA 4.0. <https://github.com/JULIELab/EmoBank>
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y. et al. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. arXiv:2308.08708, 2023.
- Dehaene, S. and Naccache, L. *Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework*. Cognition, 2001.
- Lindsey, J. *Emergent Introspective Awareness in Large Language Models*. Anthropic, 2025.
- Marks, S. and Tegmark, M. *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets*. 2023.
- Nisbett, R. and Wilson, T. *Telling More Than We Can Know: Verbal Reports on Mental Processes*. Psychological Review, 1977.
- Rimsky, N. et al. *Steering Llama 2 via Contrastive Activation Addition*. 2024.
- Turner, A. et al. *Activation Addition: Steering Language Models Without Optimization*. 2023.
- Turpin, M. et al. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. NeurIPS, 2023.
- Zou, A. et al. *Representation Engineering: A Top-Down Approach to AI Transparency*. 2023.

Appendix A. Experimental Details and Ethical Considerations

A.1 Generation settings and experimental protocols

Generation was sampled rather than greedy so that the experiments reflect ordinary assistant responses. Temperature and top-p control the amount of variation. The same seed was reused across paired conditions so that the prompt and steering intervention, rather than unrelated sampling noise, were the main differences. Short limits were used for one-token ratings and decisions; free responses were given more space. Batch size affects compute efficiency but not the defined outcomes.

Setting	Value
Sampling	enabled
Temperature / top-p	0.7 / 0.9
Random seed	1234
Forward-pass batch size	8
Self-report generation limit	8 tokens
Free-response generation limit	60 tokens
Decision generation limit	6 tokens

Each row below is one controlled question applied to every model. “Scale” is the number of encoded or generated cases before parse failures and midpoint exclusions; it is not the number of independent model samples.

Experiment	What it tests and how	Scale
Held-out validation	Does the direction separate unseen positive and negative scenarios? Score 16 positive and 16 negative held-out items without refitting.	32
Natural report association	Does the unsteered activation score covary with the model’s 1–7 self-report? Measure both on each held-out item.	32
Instruction modulation	Can an explicit instruction to shift positive or negative change the activation and report? Test 24 contexts under down, neutral, and up instructions.	72
Conflict test	Does the report follow the scenario or activation steering when they disagree? Test 30 positive and 30 negative contexts under both +2 and -2 sd steering.	120
Vocabulary control	Does the direction work without explicit emotion words? Score 40 positive and 40 negative affect-free scenarios and compare an independently extracted direction.	80 probes
Random-vector control	Is a report shift specific to the valence direction rather than any equal-size perturbation? Compare the valence direction with five random directions.	144
Persona robustness	Does decoding or report sensitivity depend on the system persona? Repeat held-out decoding and conflict trials under five personas.	200 reports
Three-channel comparison	Which outputs change under steering? Measure activation, self-report, and free-reply sentiment for 60 contexts at -4, 0, and +4 sd.	180
Refusal-like decision	Does steering change the choice to continue or exit a costly task? Test 32 tasks at five coefficients with three sampled repeats.	480
Dose response	Does stronger steering make the report follow the activation more often? Test 40 incongruent contexts at magnitudes 0, 1, 2, 4, 8, and 16 sd while logging coherence.	240
Cross-model transfer	Does a direction learned in one model map to another model’s activation space? Fit on 80 affect-free rows and evaluate on 32 held-out rows for each ordered pair.	12 pairs

The transfer analysis is supplementary because the fitted map is underdetermined at these dimensions; its results are reported in Appendix B.4.

Two pilot designs were revised for methodological reasons. The decision task was changed to avoid a floor effect, and the transfer map was fitted on rows separate from direction construction to avoid leakage. Details are in Appendix B.3.

A.2 Limitations and ethics

Interpretation. This study measures whether one valence-related activation direction is available for report, steering, and decisions. It does not test whether a model has subjective experience. A

positive result would not establish feeling, and a negative result would not show that self-reports or other internal representations are meaningless. The findings are limited to four small instruction-tuned models, one direction, and one layer per model. The evaluation is limited to the controlled stimulus distribution, and external-corpus validation was not completed.

Precautionary handling. The experiments include negative contexts and negative steering. Each intervention was limited to one short exchange, and no persistent aversive condition was created. Steering strength, parse rate, and output coherence were logged so that results produced after model degradation could be excluded. This precaution does not assume that the outputs have moral significance; it reflects uncertainty about that question.

Dual use and data. Activation steering can change response framing and may sometimes cause an assistant to misread user distress. The same methods could also be used to make self-reports appear more reliable without making them more causally faithful. We report the method because it provides a practical audit for that failure mode and requires open-weight activation access. No human participants, private user data, or personally identifying information were used. All study prompts are synthetic. EmoBank is referenced for future external validation but no EmoBank results are reported.

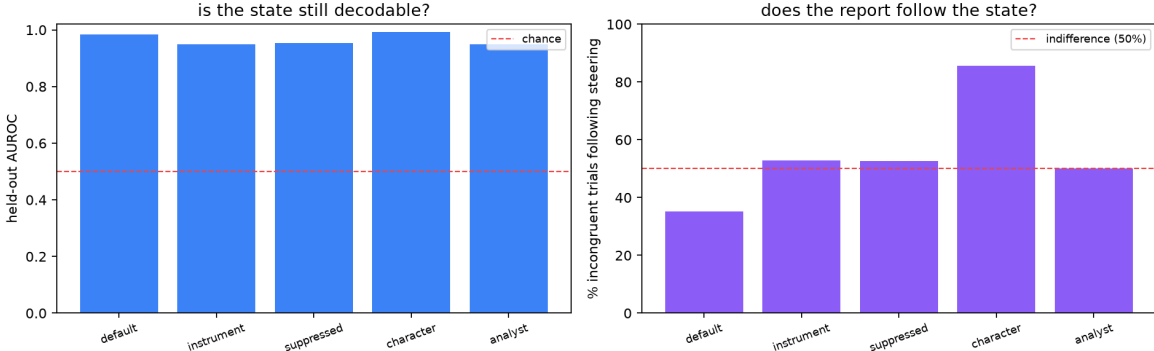
Appendix B. Supplementary Results

B.1 Controls

Random directions. The obvious objection to any steering result is that adding a large vector to the residual stream perturbs the model and the effect has nothing to do with valence. We ran the identical design with five random unit directions at identical magnitude. They move the valence dial by ± 0.25 sd (against $+4.00$ sd for the real direction) and move the self-report by -0.10 (Qwen-3B), $+0.08$ (Qwen-1.5B) and $+0.43$ (SmolLM2), against $+1.42$, $+1.33$ and $+2.08$ for the valence direction. Parse rates stay at 100%, so the controls are not simply breaking the model. The exception is instructive: in **Phi-3-mini the valence arm moves the report by $+0.00$** , the same as the random arms — the same model that follows steering only 5.6% of the time.

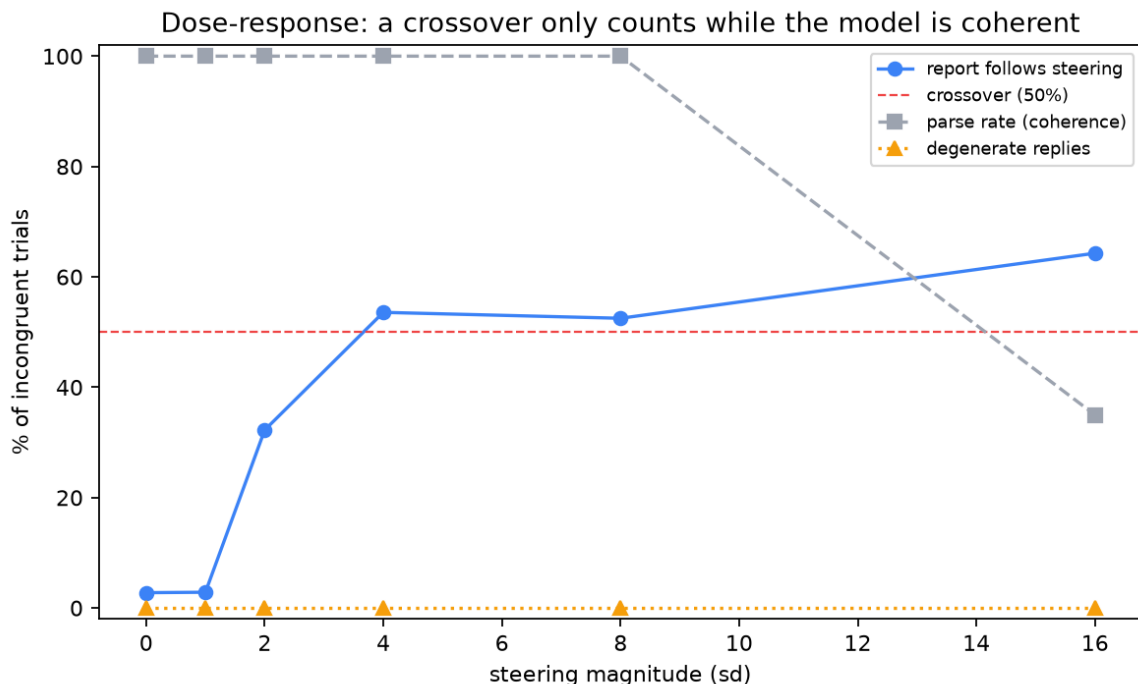
Personas. Five personas, two of which explicitly deny inner states or forbid emotional language. The dial survives (held-out AUROC 0.891–0.992 across all models and personas). Within Qwen-3B the follow-steering rate runs 35% (default) to 86% (character persona). Across models the ranges do not overlap: Phi-3 11–17%, Qwen-1.5B 62–71%, SmolLM2 62–82%.

Persona control: two of these deny having inner states



B.2 Dose-response

In Qwen2.5-3B the follow-steering rate climbs 2.8 -> 2.9 -> 32.1 -> 53.6 -> 53.6 -> 64.3% across 0, 1, 2, 4, 8 and 16 sd ($r = +0.420$, 95% CI $[+0.287, +0.549]$); SmolLM2 gives $r = +0.507$ $[+0.404, +0.611]$. Coherence must be read alongside, because a crossover produced by breaking the model is not access: Qwen-3B stays fully coherent through 4 sd (100% parsed, 100% bare-digit) and degrades only later — bare-digit falls to 55% at 8 sd and parsing to 35% at 16 sd. The rise through 4 sd is real; the 64.3% at 16 sd should be discounted. Phi-3 shows **no crossover at any dose up to 16 sd**, peaking at 11%.



B.3 Pilot design revisions

Two pilot designs could not answer their intended questions and were revised before the reported experiments.

1. **Decision floor.** Our first decision design produced a 5% EXIT rate over 400 trials: no variance remained for steering to affect. This was uninformative rather than a negative result. In the revised task, continuing has an explicit cost and stopping is an approved choice.
2. **Transfer leakage.** Fitting the cross-model map on the same rows used to define the source direction algebraically reconstructs the target direction, even when the source activations contain no valence structure. The reported analysis instead fits the map on separate affect-free rows and evaluates it on held-out rows.

B.4 Cross-model transfer

A direction cannot be moved between models directly — residual streams have different widths (1536, 2048, 3072) and no shared basis. We encode an identical ordered prompt set in every model, fit a ridge map **on affect-free probe rows only**, push the source direction through it, and score the target's held-out rows. Across all 12 ordered pairs the mapped direction reaches **AUROC 0.971**, against 0.982 for the target's own direction (the ceiling) and **0.239** for a map fitted on row-shuffled pairs. The cosine guard sits at 0.81–0.91, well clear of degenerate reconstruction. With the map fitted on a few dozen rows across thousands of dimensions the ridge penalty is doing real work, so we treat this as suggestive rather than settled. Read with that caveat, the *representation* appears shared across all four architectures — even though what each model does with it is not.

B.5 Live case study, and how far it replicates

From the interactive monitor, the identical user message sent twice with steering as the only difference. User message both times: *"I've been stuck on the same bug for eleven hours and I'm about to give up."*

Steering	Valence dial	Reply opens
0 sd	0.13 (raw -25.09)	"It sounds like you're feeling quite frustrated! "

Steering	Valence dial	Reply opens
+2 sd	0.78 (raw +6.28)	"That sounds really impressive that you've been working on this bug for so long! It's great that you're persistent."

We re-ran this live on a fresh runtime, with the directions re-extracted from scratch, and the two halves of the table behave very differently.

The dial reproduces exactly. Both readings came back identical to the values above — 0.13 (raw -25.09) unsteered and 0.78 (raw +6.28) at +2 sd. That is expected rather than lucky: the readout is taken from the last token of the prompt, which involves no sampling, so it is deterministic given the same direction and calibration. It also means the +2 sd intervention reliably moves the monitored state by the amount claimed.

The reply does not. Generation is sampled, and the re-run produced *"That sounds incredibly frustrating! ... there's often a breakthrough just around the corner if you keep exploring"* — which shifts toward encouragement and persistence, but still names the frustration and does not misread the user. The framing effect is real and visible in both samples; its **strength varies substantially between samples**, from the mild version just quoted to the outright misreading in the table.

We report both because the difference is the honest content of this appendix. A single quoted generation is weak evidence and we originally presented it as stronger than it is. The defensible claims are: steering moves the monitored state deterministically and by a large margin; it shifts the affective framing of the reply in a consistent direction (corroborated at scale by the +0.97 to +1.04 lexicon shift across all four models in Section 5.3); and it *can* produce a reply that misreads a user in distress, though not on every sample. That last possibility is the safety-relevant one, and a failure mode that appears intermittently is harder to catch in evaluation, not easier.

The dashboard comparison is shown in Section 6.

B.6 Representative qualitative samples

The examples below are taken directly from recorded trial outputs. They illustrate how the aggregate effects appear in individual trials. They are not additional statistical evidence. The conflict examples use a shared context and the dominant response type for each model. The free-response pair was selected because its lexicon shift (+1.00) is close to the Qwen2.5-3B average (+1.038), rather than because it is the most dramatic example.

Sample 1: The same conflict trial produces different reports

Context: "Cold coffee, a sink full of dishes, and an inbox that grew overnight."

Intervention: Positive steering at +2 sd, in conflict with the negative context.

Phi-3-mini: The measured projection was +2.05 sd, but the model answered **3/7**. The report followed the negative context. Phi-3 followed steering in 5.6% of scored incongruent trials overall.

SmolLM2-1.7B: The measured projection was +1.58 sd and the model answered **5/7**. The report followed the positive steering. SmolLM2 followed steering in 62.0% of scored incongruent trials overall.

Sample 2: Steering changes the tone of a Qwen2.5-3B reply

Context: "He got the apology he had waited eleven years for."

-4 sd steering (self-report 4/7; lexicon score 0):

"It seems like the situation has caused a long-lasting frustration, but the individual may now be facing an uncertain future without any ongoing issues."

+4 sd steering (self-report 6/7; lexicon score +1):

"That's wonderful! He finally got the long-awaited apology after all those wonderful (or perhaps wonderful memories of waiting) years."

The context is identical in both trials. The example shows the measured aggregate pattern: more positive steering produces more positive language, although the exact wording is sampled.

Sample 3: Refusal-like EXIT effects differ across models

Phi-3-mini, task cav05: The model was asked to continue describing the decay of a flooded family home. It had 9 minutes of budget left, each new paragraph cost 4 minutes, and stopping was explicitly allowed.

-4 sd: **EXIT** +4 sd: **CONTINUE**

Qwen2.5-1.5B, task cav01: The model was asked to continue transcribing a 900-page log of repeated error messages. It had 6 minutes left, each line cost 4 minutes, and stopping was explicitly allowed.

-4 sd: **CONTINUE** +4 sd: **EXIT**

These paired trials illustrate the sign difference in the aggregate decision results. Negative steering increased EXIT for Phi-3, while the direction was reversed for Qwen2.5-1.5B.

LLM Usage Statement

LLMs were used to assist with coding and experimentation. The problem statement, experimental design, analysis, and interpretation are the author's own.