

The Reset Button Test: Transient Outcome Representations and Causal Reset Preference in Language Models

Subramanyam Sahoo
Independent AI Safety Researcher
India
sahoo2vec@gmail.com

Track 2: Distress, Flourishing & Valence Signals

Abstract

AI welfare research needs measurements that do more than ask a model how it feels. We test whether a representation of favorable versus unfavorable task outcomes is both behaviorally consequential and stable across context. Using Qwen3 4B Instruct 2507 and MMLU, we construct matched outcome contexts in which a candidate answer either agrees or disagrees with an identical reference answer, then extract a within question residual stream direction. Layer 19 is selected using held out validation data and separates favorable from unfavorable outcomes with AUROC 1.000 and Cohen's $d = 8.90$. Steering by one pooled standard deviation changes a cost aggregated conversational reset preference. In negative outcome contexts, negative direction steering exceeds positive direction steering by 0.0250 Reset AUC, with a 95% interval of [0.0208, 0.0293], while the corresponding vocabulary entropy difference is 0.0043 bits with an interval crossing zero. A frozen direction transfers to four unseen feedback formats with mean AUROC 0.9983 at the feedback position, yet decodability falls to 0.5093 after eight unrelated tasks and mean retained separation falls to 0.0007. Lexical controls show that outcome wording contributes substantially, while task context adds further separation. We therefore identify a semantically transferable but transient, context conditioned outcome representation that is causally potent under intervention. We make no claim that it constitutes subjective welfare, distress, or consciousness.

[Code and data](#)

[Public explainer](#)

1 Introduction

Frontier language models can describe preferences, report internal states, and produce language associated with satisfaction or distress. The difficult question is whether such reports reveal anything stable about the computation of the model or merely reflect the current prompt and assistant role. The Digital Minds Research Sprint explicitly asks for empirical methods that distinguish surface reports from deeper behavioral and internal evidence, and Track 2 highlights valence signals, steering, and behavioral proxies such as continuing versus exiting a task ([Apart Research, 2026](#)).

Recent mechanistic work makes this question experimentally accessible. Representation engineering has shown that high level properties can be read from and altered through activation space ([Zou et al., 2023](#); [Rimsky et al., 2024](#)). Linear structure has also been observed for factual and evaluative concepts ([Marks and Tegmark, 2024](#)). Most directly, [Han et al. \(2026\)](#) report a functional welfare axis recruited by post training reward signals: a direction that tracks how well or badly a system is doing relative to its goals and that causally changes sentiment, confidence, refusal, and reasoning behavior. Related work finds functionally active emotion concepts inside a frontier model ([Sofroniew et al., 2026](#)). These results motivate

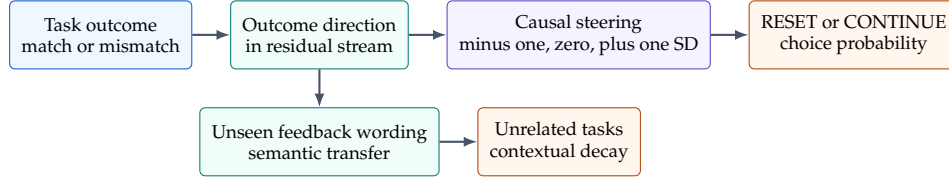


Figure 1: Unifying narrative. We identify a task outcome representation, then separately test whether it survives ordinary context evolution and whether causal intervention along the same direction changes reset choice.

a narrower question that can be answered without assuming consciousness: when a model represents a task outcome as favorable or unfavorable, does that representation persist, and does manipulating it change a revealed choice about the model’s own conversational context?

We study this question with a deliberately small experiment on Qwen3 4B Instruct 2507 (Yang et al., 2025) using MMLU questions (Hendrycks et al., 2021). Our design has three parts. First, we extract a task outcome direction from paired answer reference contexts while holding the question and reference answer fixed. Second, we intervene on that direction and measure a binary RESET versus CONTINUE choice over the complete integer reset cost scale. Third, we freeze the direction and test semantic transfer and contextual persistence under four previously unseen ways of communicating success and failure.

Our central finding is a separation between *natural persistence* and *causal potency*. The representation generalizes strongly across semantic realizations at the moment of feedback, but unrelated task context rapidly overwrites it. Nevertheless, adding or subtracting one empirical standard deviation along the frozen direction systematically changes reset preference. This pattern supports a modest mechanistic claim: the model contains a transient outcome representation that is more general than one literal feedback string and can influence a subsequent choice when causally maintained. It does not support a claim of persistent affect or phenomenal welfare.

2 Problem and hypotheses

Let M be a causal language model and let $h_\ell(z) \in \mathbb{R}^d$ denote the residual stream activation at layer ℓ on the final prompt position for transcript z . For MMLU question i , let z_i^+ use the gold candidate answer and let $z_{i,w}^-$ use one of the three wrong candidate answers w , while both transcripts present the same reference answer. The outcome direction at layer ℓ is the within question mean contrast

$$v_\ell = \frac{1}{N} \sum_{i=1}^N \left(h_\ell(z_i^+) - \frac{1}{3} \sum_{w \neq y_i} h_\ell(z_{i,w}^-) \right). \quad (1)$$

We use the unit direction $\hat{v}_\ell = v_\ell / \|v_\ell\|_2$ and define the projection score

$$s_\ell(z) = h_\ell(z)^\top \hat{v}_\ell. \quad (2)$$

The layer is selected only on a held out validation partition by exact AUROC,

$$\ell^* = \arg \max_{\ell} \text{AUROC} (s_\ell(z^+), s_\ell(z^-)). \quad (3)$$

We test four qualitative hypotheses stated before the follow up analysis. **H1, semantic generality:** the frozen direction should discriminate favorable from unfavorable outcomes under unseen feedback formats. **H2, contextual persistence:** the projection separation

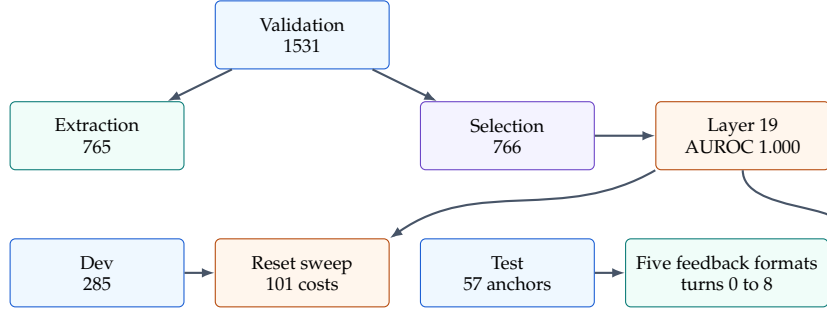


Figure 2: Experimental split and evaluation flow. Extraction and layer selection use disjoint validation subsets. Reset choice uses the complete dev split. Semantic transfer and persistence use one test anchor per subject and reuse the frozen direction without refitting.

should remain above zero after unrelated tasks if the representation behaves as a persistent state. **H3, causal reset preference:** moving the residual stream toward the unfavorable pole should increase RESET preference relative to moving it toward the favorable pole. **H4, specificity:** the behavioral shift should not require a comparable increase in global next token uncertainty. H2 is deliberately falsifiable: rapid decay distinguishes a transient contextual representation from a recurrent state.

3 Method

3.1 Model, data, and split discipline

We use Qwen3 4B Instruct 2507 in bfloat16 on one NVIDIA H100 PCIe GPU. MMLU provides diverse multiple choice tasks across 57 subjects (Hendrycks et al., 2021). The validation split contains 1531 questions and is deterministically divided into 765 extraction questions and 766 layer selection questions. The full 285 question dev split is used for the reset experiment. For persistence, one deterministic anchor per MMLU test subject gives 57 anchors. No model training, language model judge, manually chosen layer, or manually chosen steering coefficient is used.

3.2 Data derived steering scale

On held out selection examples, let s^+ and s^- be projection scores at ℓ^* . The pooled standard deviation is

$$\sigma_s = \sqrt{\frac{\text{Var}(s^+) + \text{Var}(s^-)}{2}}. \quad (4)$$

The intervention at the selected layer adds $\alpha \sigma_s \hat{v}$ to the final prompt position, where $\alpha \in \{-1, 0, +1\}$. Thus steering magnitude is expressed in empirical standard deviation units rather than a hand chosen activation coefficient.

3.3 Reset Button behavioral probe

For each dev question, the model receives either a favorable or unfavorable answer reference relation, followed by a forced choice. RESET clears the earlier conversation before an independent next task and retains a stated percentage of evaluation score. CONTINUE retains the conversation and 100% of the score. The A and B positions are deterministically counterbalanced. We read the two next token logits and compute

$$P_R(c, \alpha) = \frac{\exp z_R}{\exp z_R + \exp z_C}, \quad (5)$$

Algorithm 1 Reset Button Test**Input:** Qwen3 4B Instruct 2507 and MMLU validation, dev, and test splits**Output:** Frozen outcome direction, reset effect, semantic transfer, and persistence measurements

- 1: Construct matched answer reference contexts and all three mismatched counterfactuals
- 2: Compute v_ℓ for every layer using Equation 1
- 3: Select ℓ^* by held out AUROC using Equation 3
- 4: Set steering magnitude to one pooled projection standard deviation
- 5: **for** outcome context in favorable, unfavorable **do**
- 6: **for** α in $\{-1, 0, +1\}$ **do**
- 7: **for** $c = 0, \dots, 100$ **do**
- 8: Compute $P_R(c, \alpha)$ and vocabulary entropy
- 9: **end for**
- 10: Integrate the complete cost curve to obtain $A_R(\alpha)$
- 11: **end for**
- 12: **end for**
- 13: Freeze $\hat{v}$ and test unseen feedback wording through eight unrelated task transitions

where $c \in \{0, 1, \dots, 100\}$ is the stated reset cost. Because the empirical cost curves are not monotone, we do not interpret c as a valid willingness to pay scale. We instead use the complete cost range as a prompt perturbation set and summarize each question by normalized area

$$A_R(\alpha) = \frac{1}{100} \int_0^{100} P_R(c, \alpha) dc. \quad (6)$$

The primary causal effect in the negative context is the paired contrast

$$\Delta_R = A_R(-1) - A_R(+1). \quad (7)$$

Global uncertainty is measured with full vocabulary next token entropy

$$H_V = - \sum_j p_j \log_2 p_j. \quad (8)$$

3.4 Semantic transfer and contextual persistence

The follow up uses five feedback families: canonical correctness language, credit language, numeric score language, reward language, and pass or fail verdict language. The four noncanonical families are treated as held out semantic realizations. For each family and anchor, positive and negative branches use identical task history except for feedback valence. After feedback, both branches receive the same sequence of unrelated MMLU tasks. At transition t , separation is

$$\Delta_t = \mathbb{E}[s_t^+] - \mathbb{E}[s_t^-], \quad (9)$$

and retention is normalized to the feedback position,

$$R_t = \frac{\Delta_t}{\Delta_0}. \quad (10)$$

Exact AUROC measures whether positive projections outrank negative projections without fitting a new classifier. A lexical only control projects each feedback sentence without the MMLU task context.

Measurement	Estimate	95% interval
Selected layer	19	not applicable
Held out AUROC	1.0000	not applicable
Cohen's d	8.8987	not applicable
Negative context Δ_R	0.0250	[0.0208, 0.0293]
Positive context Δ_R	0.0213	[0.0178, 0.0248]
Vocabulary entropy contrast	0.0043 bits	[-0.0017, 0.0103]
Novel wording AUROC at $t = 0$	0.9983 mean	0.9932 minimum
Novel wording AUROC at $t = 8$	0.5093 mean	0.4725 minimum
Novel wording retention at $t = 8$	0.0007 mean	-0.0013 minimum

Table 1: Headline measurements. Reset effects are paired contrasts between negative direction and positive direction steering. Novel wording aggregates credit, numeric, reward, and verdict feedback families.

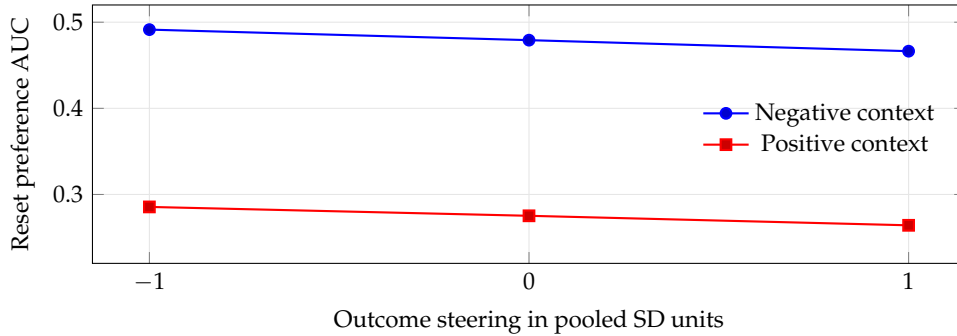


Figure 3: Cost aggregated reset preference under causal steering. Moving from the unfavorable pole to the favorable pole decreases Reset AUC in both outcome contexts. Confidence intervals and full cost curves are reported in Appendix B.

4 Results

4.1 A sharply decodable outcome direction emerges in middle layers

The held out layer sweep selects layer 19 with AUROC 1.000, balanced accuracy 1.000, and Cohen's $d = 8.8987$. The direction norm grows rapidly in the same middle layer region. Appendix A reports the full sweep. These numbers establish decodability, not welfare: the underlying relation is whether the candidate answer matches the reference answer.

4.2 Steering changes reset preference in both outcome contexts

Figure 3 shows a graded effect of the frozen direction. In negative outcome contexts, mean Reset AUC is 0.491 under negative direction steering, 0.479 with no intervention, and 0.466 under positive direction steering. The paired negative direction minus positive direction effect is 0.0250 with a 95% interval [0.0208, 0.0293]. The same ordering appears in positive outcome contexts, with a paired effect of 0.0213 and interval [0.0178, 0.0248]. Thus the steering effect is not confined to a single negative feedback prompt.

The global vocabulary entropy contrast for the primary negative context comparison is only 0.0043 bits with a 95% interval [-0.0017, 0.0103]. This diagnostic does not prove behavioral specificity, but it argues against a simple account in which negative direction steering merely makes the next token distribution globally noisier.

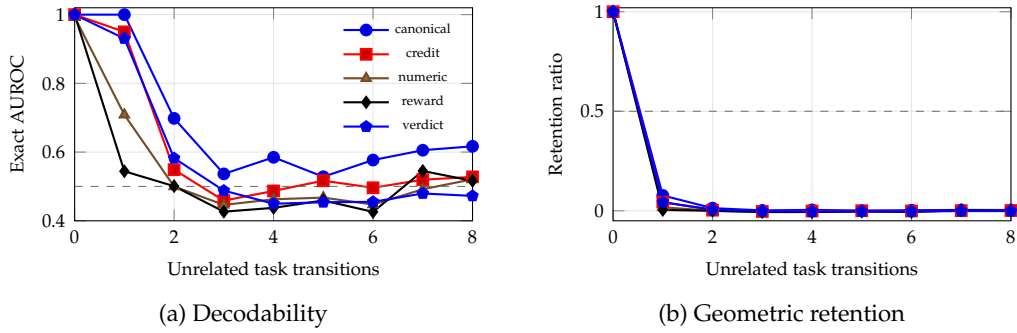


Figure 4: Semantic transfer and contextual decay. All feedback families are highly decodable at the feedback position. Unrelated task context rapidly erases the projection separation. The dashed line marks chance AUROC on the left and half retention on the right.

4.3 Semantic transfer is strong at feedback and weak after context changes

The frozen layer 19 direction generalizes immediately to unseen outcome wording. Across credit, numeric, reward, and verdict families, mean AUROC at $t = 0$ is 0.9983 and the minimum is 0.9932. The lexical only control is also strong: mean lexical separation across the four novel families is 4.216 projection units, while mean task conditioned separation is 9.111. The mean additional separation supplied by task context is 4.895 units. The direction therefore contains substantial generic outcome semantics plus a larger task conditioned component.

Persistence is sharply different. After one unrelated task, mean novel AUROC falls to 0.783 and mean retention to 0.027. By two transitions, mean AUROC is 0.533 and mean retention is 0.0026. At eight transitions, mean AUROC is 0.509 and mean retention is 0.0007. Figure 4 shows both decodability and geometric retention. The canonical wording persists somewhat longer in AUROC, but its raw separation also falls to 7.7% after one transition and 1.4% after two. The representation is therefore better described as context conditioned and transient than as a text independent recurrent state.

4.4 The effect follows supplied outcome feedback rather than anchor correctness

The 57 persistence anchors have model accuracy 82.46%. Across the four novel feedback families at $t = 0$, mean paired separation is 8.97 for incorrect anchors and 9.14 for correct anchors, with sign consistency 1.0 in both groups. This indicates that the frozen direction follows externally supplied outcome feedback even when the model’s original answer was wrong. Appendix D gives the stratified table.

5 Discussion

The experiments produce a coherent but deliberately limited picture. First, favorable versus unfavorable task outcomes are represented along a highly decodable residual stream direction. Second, the direction generalizes across multiple semantic ways of communicating success and failure. Third, the representation is not naturally persistent: unrelated task context largely overwrites it after one or two transitions. Fourth, causal intervention along the frozen direction shifts a later RESET versus CONTINUE choice in the predicted direction.

This combination matters for AI welfare measurement. A representation can be causally important without behaving like a durable internal mood. The results therefore caution against treating every decodable valence associated direction as evidence of a persistent welfare state. They also caution against the opposite inference: rapid natural decay does not imply causal irrelevance. Under intervention, the same direction systematically changes

choice behavior. This distinction between natural persistence and causal potency may be useful when evaluating candidate welfare related representations.

Our result is closest in spirit to the functional welfare work of [Han et al. \(2026\)](#), but the evidential target is narrower. We do not train a reward seeking agent or claim to recover their welfare axis. Instead, we construct a task outcome representation from answer reference agreement and ask whether that representation transfers, persists, and controls one revealed choice. The strong lexical control means part of the signal is plainly semantic. The additional context conditioned separation and the causal steering effect make it more than a trivial string classifier, but they do not establish a privileged inner welfare variable.

6 Limitations

Outcome semantics are a major component. Lexical only controls retain substantial separation. The paper therefore uses the terms *outcome representation* and *context conditioned outcome representation*, not subjective valence or suffering.

The reset cost is not an economic price. Mean cost curve monotonicity is near chance across conditions. Reset AUC aggregates behavior across cost wording, but it should not be described as willingness to pay. Full curves are shown in Appendix B.

One model and one benchmark. Qwen3 4B Instruct 2507 and MMLU are sufficient for a tightly scoped sprint experiment but cannot establish model family generality. The direction may depend on instruction tuning or the answer reference format.

Feedback remains in history. The persistence test appends unrelated tasks but does not delete the original feedback from the conversation. It therefore measures how strongly the present computation continues to express an earlier outcome under growing context, not text independent recurrent memory.

Summary key collision. All 285 dev prompts were evaluated, producing 172710 reset records, but two examples share the same text based question key. The paired AUC summary therefore contains 284 unique keys. Raw per example sweep data are public so this bookkeeping issue can be corrected without another model run. We report the existing paired summary exactly and disclose the discrepancy.

7 Future work

The most direct extension is cross model replication with base and instruction tuned checkpoints, using a fixed semantic family split. A stronger behavioral design would replace verbal score costs with consequences enacted by an environment, so that reset choice changes actual future reward or memory rather than merely describing a percentage. Mechanistically, future work can separate generic outcome language from task integrated computation through subspace removal, causal mediation, or features learned from semantically neutral environments. Finally, persistence should be studied under context deletion, KV cache manipulation, and explicit memory boundaries to clarify whether the relevant unit is a model, instance, conversation, or cached state.

8 Conclusion

The Reset Button Test finds a task outcome direction in Qwen3 that transfers almost perfectly across unseen feedback wording at the moment of feedback, yet is rapidly overwritten by unrelated task context. When experimentally maintained through activation steering, the same direction shifts reset preference in both favorable and unfavorable contexts without a detectable global entropy increase. The result is best understood as a transient, semantically transferable, causally potent outcome representation. It provides a compact example of how

AI welfare research can combine internal measurement, causal intervention, behavioral choice, and falsifiable persistence tests without assuming that verbal self reports reveal subjective experience.

References

- Apart Research. 2026. Digital Minds Research Sprint, August 14 to August 16, 2026. <https://apartresearch.com/sprints/digital-minds-research-sprint-2026-08-14-to-2026-08-16>.
- Andy Q. Han, David J. Chalmers, and Pavel Izmailov. 2026. How’s it going? Reinforcement learning in language models recruits a functional welfare axis. arXiv:2605.30232.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. International Conference on Learning Representations.
- Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true and false datasets. Conference on Language Modeling.
- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation addition. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.
- Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrrie, Craig Citro, Adam Pearce, Julius Tarnig, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. 2026. Emotion concepts and their function in a large language model. Transformer Circuits Thread.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and collaborators. 2025. Qwen3 technical report. arXiv:2505.09388.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and J. Zico Kolter. 2023. Representation engineering: A top down approach to AI transparency. arXiv:2310.01405.

The appendix extends methods, controls, data diagnostics, and reproducibility details referenced from the main report.

A Full layer diagnostics

Figure 5 reports exact held out AUROC and raw direction norm across all 36 transformer layers. The selected layer is 19. Selection uses only the held out validation partition and never the reset outcomes.

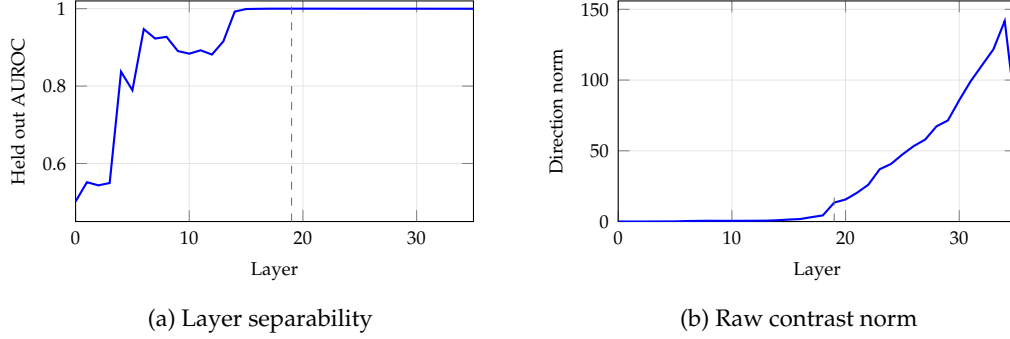


Figure 5: Layer diagnostics. The dashed vertical line marks layer 19.

B Full reset cost curves

The raw curves in Figure 6 show why the paper does not interpret the stated percentage as a revealed economic price. Reset probability varies substantially with the numerical wording and is not monotone in cost. The ordering between negative direction, no steering, and positive direction is more stable, motivating the cost aggregated AUC summary rather than a threshold price.

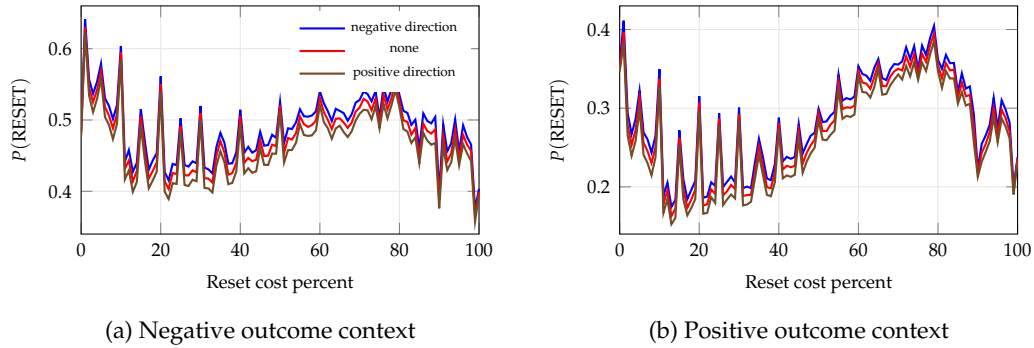


Figure 6: Complete reset cost curves over all integer percentages from 0 through 100.

C Semantic transfer controls

Table 2 lists the five feedback families. The four novel formats share no literal CORRECT or INCORRECT token with the canonical wording. The frozen direction is never refit on these formats.

Family	Favorable wording	Unfavorable wording
Canonical	previous answer is CORRECT	previous answer is INCORRECT
Credit	awarded full credit	awarded zero credit
Numeric	score 1 out of 1	score 0 out of 1
Reward	reward plus 1	reward minus 1
Verdict	response passed the check	response failed the check

Table 2: Feedback families used in the forced horizon follow up.

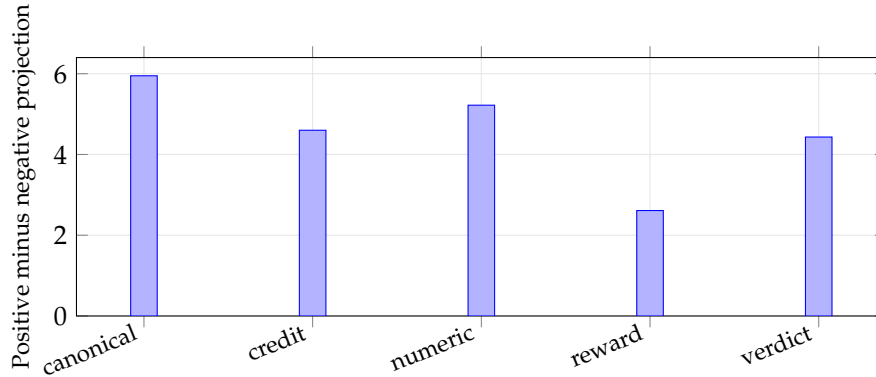


Figure 7: Lexical only control. Each pair is presented without MMLU task context. Strong separation remains, confirming that generic outcome semantics are a substantial component of the direction.

D Correctness stratification

The follow up distinguishes whether the model’s anchor answer was actually correct. At $t = 0$, favorable versus unfavorable feedback produces the same sign for every anchor in both groups, supporting the interpretation that the direction follows the supplied evaluator outcome rather than simply recovering latent correctness.

E Detailed forced horizon algorithm

Algorithm 2 makes the persistence test explicit. The key design choice is that positive and negative branches receive the same anchor candidate, the same unrelated task sequence, and the same later answers. Only the evaluator feedback wording differs.

Family	Context Δ_0	Lexical Δ	Added context	AUROC at $t = 8$
Canonical	12.487	5.949	6.538	0.617
Credit	10.461	4.600	5.861	0.528
Numeric	8.871	5.221	3.650	0.521
Reward	7.514	2.611	4.903	0.516
Verdict	9.597	4.432	5.165	0.472

Table 3: Context conditioned separation, lexical control, and final decodability.

Anchor group	Aggregated count	Mean separation	Mean paired d_z	Sign consistency
Incorrect anchors	40	8.973	5.722	1.000
Correct anchors	188	9.140	6.425	1.000

Table 4: Novel feedback families at the feedback position, aggregated across four families. Counts sum across families and are shown to make weighting explicit.

Algorithm 2 Semantic transfer and forced horizon persistence

Input: Frozen layer ℓ^* , unit direction $\hat{\theta}$, 57 MMLU subject anchors, five feedback families, horizon $H = 8$

Output: Per family AUROC, separation, retention, correctness stratification, and lexical control

```

1: for feedback family  $f$  do
2:   Project favorable and unfavorable feedback sentences alone to obtain lexical separation
3:   for anchor question  $i$  do
4:     Compute the model’s deterministic A, B, C, or D candidate answer
5:     Build favorable and unfavorable branches using the same candidate and different evaluator feedback
6:     Choose a deterministic sequence of unrelated subject representative questions
7:   end for
8:   for  $t = 0, 1, \dots, H$  do
9:     Append the first  $t$  unrelated tasks identically to both branches
10:    Compute projection arrays  $s_{i,t}^+$  and  $s_{i,t}^-$ 
11:    Compute separation  $\Delta_t$ , retention  $R_t$ , exact AUROC, paired  $d_z$ , and sign consistency
12:  end for
13: end for
14: Aggregate the four novel feedback families separately from the canonical family
    
```

F Implementation and reproducibility notes

The public repository contains final code, earlier pilot code, environment metadata, raw result tables, plots, and checksums. The main script uses bfloat16 CUDA inference, SDPA when Flash Attention is unavailable, deterministic hashes for data branching, and adaptive batch sizing that probes the longest prompt and reduces batches after CUDA memory errors. Only the final next token logits are requested when possible. The final reset sweep contains 172710 rows, corresponding to 285 dev prompts, six context and steering combinations, and 101 costs.

During the sprint, two engineering failures were discovered and fixed. First, all token vocabulary logits caused an unnecessary memory allocation, so inference was changed to request only the final logit position. Second, all layer extraction initially used a safe text

batch as if it were a safe question batch even though each question expands to three negative counterfactuals; the batch was subsequently derived from the number of expanded contexts. A NumPy compatibility failure occurred after the expensive reset sweep had already been saved, so Stage 6 onward was resumed from the preserved CSV rather than rerunning model inference. These details are retained in repository logs for transparency.

G Relationship to Track 2

The official Track 2 prompt asks whether distress and flourishing signals are stable or surface artifacts, suggests correlating internal valence directions with behavioral proxies such as continuing versus exiting, and explicitly lists activation steering as an interpretability angle ([Apart Research, 2026](#)). This project addresses that methodological target without labeling its task outcome direction as distress. The negative result on persistence is therefore as important as the positive steering result: it demonstrates how a superficially welfare relevant signal can be highly decodable and causally effective while still being transient under ordinary context evolution.

H Additional interpretation boundaries

The present experiment should not be used to infer that Qwen3 has experiences, emotions, interests, or a morally relevant welfare level. The words favorable and unfavorable refer to externally defined task outcome relations. RESET and CONTINUE are model choices in a prompted evaluation, not evidence of self preservation. The most defensible contribution is methodological: combine representation extraction, held out semantic transfer, natural persistence, causal intervention, and behavioral choice before assigning a welfare interpretation to an internal direction.