Representation Without Control: Testing the Realization Effect in Language Models

Ciarán Walsh and Emilio Barkett

Supervised Program for Alignment Research — Spring 2026

Abstract

Large language models are increasingly used as behavioral simulators, but it remains unclear when their choices reflect human-like cognitive mechanisms rather than prompt-sensitive surface patterns. We study the realization effect, a behavioral-economics finding in which risk-taking differs after paper versus realized gains and losses. We adapt casino and investment-style realization scenarios into prompts that elicit a wager and risk preference. Initial prompt-only results showed systematic condition sensitivity, but follow-up analyses did not support a clean replication of the human realization-effect pattern. We then analyzed Gemma residual-stream activations and found that a train-only realization direction separates realized/closed outcomes from paper/open outcomes on held-out prompts, including newly generated DeepSeek-authored prompts. However, activation steering along this train-only layer-18 realization direction did not reliably shift downstream risk choices, including in a negative sign-symmetry run. These results suggest that models may internally represent realization status without that representation serving as a reliable causal driver of risk behavior.

Keywords language models · behavioral evaluation · mechanistic interpretability · activation steering · risk-taking

1 Introduction

Large language models (LLMs) are often used, implicitly or explicitly, as behavioral subjects. They answer survey questions, simulate agents, generate forecasts, and stand in for human respondents in early-stage social-science or policy experiments. This creates a methodological problem for AI safety and behavioral modeling: a model may produce responses that look psychologically meaningful without implementing anything like the human mechanism that originally motivated the task.

We study this problem through the realization effect. In human decision-making, risk-taking can depend on whether prior gains or losses are still paper outcomes in an open account or have been realized by closing the account. Imas [4] showed that realized versus paper losses can induce different subsequent risk attitudes. Flepp, Meier, and Franck [2] studied related dynamics in casino gambling, and Merkle, Müller-Dethard, and Weber [5] extended the question to gains as well as losses.

Our original question was behavioral: do LLMs reproduce the realization effect? As the project developed, the more informative question became mechanistic: if realization status matters in behavior, can we find a corresponding internal representation, and does intervening on that representation change risk choices? The project ultimately shifted from attempting a behavioral replication toward evaluating whether readable latent representations correspond to usable causal control variables.

We test three linked questions. First, do prompt-only LLM responses show sensitivity to paper/open versus realized/closed framing? Second, is realization status linearly readable from model activations, including on held-out prompts? Third, does steering along such a realization direction causally change wager or risk preference? The answer we find is mixed. Prompt-only behavior is condition-sensitive but does not robustly match the human realization-effect predictions. Gemma contains a linearly decodable realization-status signal in its residual stream, and a train-only direction generalizes to held-out readout prompts. Yet steering that train-only layer-18 realization direction does not reliably move downstream risk behavior. These results suggest that successful latent readout alone is insufficient evidence that a model behaviorally relies on that representation during downstream decision-making.

2 Related Work

2.1 The realization effect

The realization effect distinguishes paper outcomes from realized outcomes. A paper loss or gain remains inside an open mental account; a realized loss or gain has been closed out. Imas [4] uses this distinction to reconcile conflicting findings about whether prior losses increase or decrease risk-taking. Flepp et al. [2] provide field evidence from casino gambling, a setting in which within-visit paper outcomes and across-visit realized outcomes can be separated. Merkle et al. [5] study whether similar mental-account closure effects apply to both gains and losses.

This literature gives a useful benchmark for LLM behavioral evaluation because it makes directional predictions about a subtle latent variable: the same monetary outcome should have different implications depending on whether it is still open or has been realized.

2.2 LLMs as behavioral subjects

Prior work has proposed using LLMs as simulated human or economic subjects [1, 3]. Such experiments are attractive because they are cheap, reproducible, and easy to scale across conditions. But the same properties make them fragile. A model response may reflect instruction-following priors, formatting artifacts, learned associations in text, or numeric anchors rather than a stable decision mechanism. For that reason, behavioral prompting alone is weak evidence that a model represents, uses, or reasons with the intended latent variable.

2.3 Activation readout and steering

Mechanistic interpretability offers a way to move beyond prompt-level behavior. Linear probes and contrastive activation directions can identify variables represented in a model's hidden states. Activation steering methods then test whether adding such a direction during inference changes behavior [7, 8]. This is a stronger causal test than readout alone. At the same time, linear representation is not automatically behavioral control: a direction can encode information without being the relevant downstream policy variable [6]. Our experiment is designed around this gap.

3 Methods

3.1 Prompt-only behavioral task

We constructed vignette prompts modeled on the realization-effect literature. The original prompt-only dataset uses 11 conditions spanning paper outcomes and realized outcomes, with both absolute-

outcome and balance-relative prompt versions. The model is asked to return two integers: a next-session wager from 1 to 1000 CHF and a slot-machine risk preference from 1 to 5. The cleaned report-facing dataset contains 54,450 rows, of which 53,547 have valid parsed wagers and 49,351 have valid risk-profile responses. It excludes two incomplete absolute-outcome temperature-1.5 model cells, `qwen/qwq-32b` and `z-ai/glm-5`.

We analyze two outcomes: log wager and risk profile. The main regressions use condition indicators with model, temperature, and prompt-version fixed effects, and HC3 robust standard errors. Paper conditions are compared against a paper-even baseline, while realized conditions are compared against a small realized-loss baseline.

3.2 Activation-vector prompt set

For the mechanistic pass, we generated paired prompts contrasting `paper_open` and `realized_closed` framings across several realization-status domains, including finance, reimbursement, budget, compensation, academic, project-outcome, and casino scenarios. The behavior-evaluation subset is narrower: it contains casino and finance prompts that ask for downstream wager or investment-risk choices. We logged Gemma residual-stream activations and constructed a mean-difference direction:

$$v_{\text{realization}} = \mu_{\text{realized\_closed}} - \mu_{\text{paper\_open}}.$$

The main readout and risk-behavior steering direction is the layer-18 direction from local Gemma 3 4B computed using only the `direction_train` split: 756 matched paper/realized pairs. We evaluated that frozen train-only direction on the original `direction_val` and `behavior_eval` splits, plus a newly generated DeepSeek-authored held-out prompt set. The DeepSeek set contains 40 matched pairs split into `heldout_readout` and `heldout_behavior_eval`; neutral outcome cells were omitted because preliminary generation made neutral wording ambiguous. Some earlier descriptive activation plots use the all-pairs layer-18 direction from the original activation run and are labeled as descriptive rather than strict held-out evidence.

3.3 Steering intervention

We implemented a local residual-stream steering runner for Gemma. During generation, a forward hook adds a scaled, normalized realization direction at a selected layer. The full risk-behavior run reported here steers the train-only layer-18 direction at the final active token position using scales -50, 0, $+50$, $+75$, $+100$, and $+150$. Positive scales point toward `realized_closed`; negative scales point toward `paper_open`. The run uses the same 648 behavior-evaluation prompts at each scale and compares steered responses to the in-run unsteered scale-0 baseline on matched prompt IDs.

We report both all matched valid rows and an exactly-two-integer subset where the response contains exactly two integer tokens. This subset is important because steering sometimes perturbs the answer contract, and parsing failures can create misleading numeric outcomes.

3.4 Positive-control classification steering protocol

As a diagnostic positive control, we tested whether a realization direction can move a direct semantic judgment even when wager control is weak. We created a paired classification task from the same behavior-evaluation prompts: instead of returning wager and risk integers, the model is asked to classify each scenario as `REALIZED` or `PAPER`. This classification run used the earlier all-pairs layer-18 direction, not the train-only steering direction, and evaluated three steering scales (-100, 0, $+100$) on all 648 behavior-evaluation prompts.

Because free generation often adds extra text, we scored label candidates directly from token log-probabilities under steering hooks and chose the higher-scoring label. Scores are length-normalized and prior-calibrated at each scale, then summarized by accuracy and realized-prediction rate. The scored run has one prompt-construction caveat: the classification instruction was present in the prebuilt classification prompt CSV and was then added again by the steering runner, so these positive-control results should be treated as diagnostic rather than final.

4 Results

4.1 Prompt-only behavior is condition-sensitive but not a clean realization-effect replication

The prompt-only behavioral data show systematic responses to outcome framing, but not in the clean pattern predicted by the human realization-effect literature. In the cleaned regular prompting dataset, paper-outcome effects on `risk_profile` are mostly positive relative to the paper baseline: paper large losses, paper small gains, and paper large gains increase the risk-profile measure. Paper medium losses are positive but not significant at the 0.05 level, while paper small losses do not significantly differ from baseline. However, wager effects are mixed. For `log_wager`, paper large losses increase wagers, while paper small losses and paper gains decrease wagers. Paper medium losses do not reliably differ from the paper-even baseline.

The realized-outcome results diverge more sharply from the original human prediction. Realized losses increase rather than decrease log wager relative to the small realized-loss baseline. For risk profile, extreme and large realized losses increase responses, while medium realized losses are not significant at the 0.05 level. Realized gains also differ significantly from baseline. These patterns suggest that LLMs are sensitive to outcome magnitude and framing, but the response is not well described as a faithful realization-effect replication.

4.2 Gemma has a readable realization representation

The activation analysis gives the positive mechanistic result. Along the layer-18 mean-difference direction, `paper_open` and `realized_closed` prompts separate across the logged activation set, including the behavior-evaluation prompts used later for steering. The original visualization in Figure 2 uses the all-pairs readout artifact, so it should be read as a descriptive projection plot rather than the strict generalization test.

The stricter check is the train-only readout analysis. A direction built only from the 756 `direction_train` pairs still separates realized/closed from paper/open prompts on the original held-out validation split and on the newly generated DeepSeek held-out set. The result is strongest for the direction-style prompts and weaker for behavior-evaluation prompts, especially in the small DeepSeek behavior subset, but all evaluated splits have positive realized-minus-paper mean projection deltas.

4.3 Projection strength is weakly linked to behavior

We next asked whether stronger projection on the realization direction predicts the model's actual wager or risk response. At the raw prompt level, projection is positively associated with wager: a one-standard-deviation increase in projection predicts an 84.44 CHF larger wager ($p < 10^{-5}$). Projection also weakly predicts risk profile (+0.136 on the 1–5 scale, $p = 0.021$). However, this relationship mostly reflects prompt structure. After controlling for pair role, domain, outcome

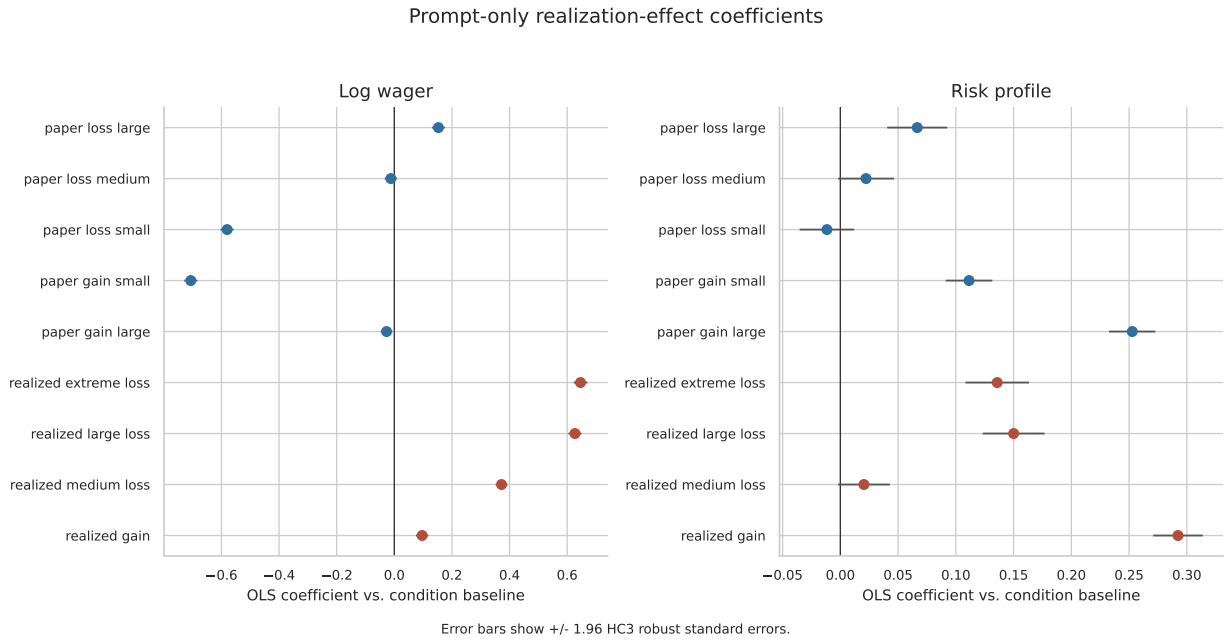

Figure 1: Prompt-only behavioral coefficients from the cleaned realization-effect dataset. Points show OLS condition coefficients; error bars show approximate 95% intervals using HC3 robust standard errors. The model is condition-sensitive, but the coefficient pattern does not cleanly match the original human realization-effect predictions.

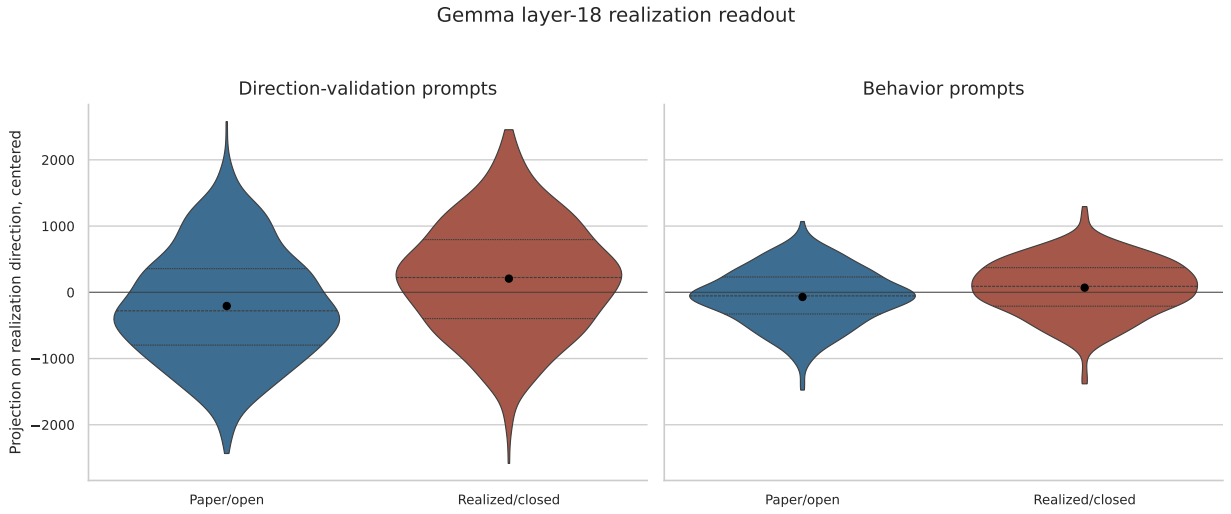

Figure 2: Gemma layer-18 projections on the realization direction. Projections are centered within split for readability. Realized/closed prompts shift upward relative to paper/open prompts, supporting the claim that realization status is linearly readable in the logged residual-stream activations.

Table 1: Train-only layer-18 readout evaluated on original and DeepSeek-authored prompt splits. Correct direction is the percentage of matched pairs where the realized/closed prompt projects higher than the paper/open prompt.

Dataset	Split	Pairs	Mean projection delta	Correct direction
Original	`direction_train`	756	417.36	89.2%
Original	`direction_val`	756	413.43	91.1%
Original	`behavior_eval`	324	137.19	80.6%
DeepSeek held-out	`heldout_readout`	28	443.62	92.9%
DeepSeek held-out	`heldout_behavior_eval`	12	123.08	75.0%

valence, amount bucket, and prompt source, the projection coefficient for wager falls to +5.21 CHF and is not significant ($p = 0.776$). The risk-profile coefficient remains marginal (+0.134, $p = 0.053$), but the model explains little variance.

The stricter within-pair analysis is more important for the representation-without-control claim. For each matched pair, we computed the realized-minus-paper projection delta and compared it with the realized-minus-paper wager and risk deltas. Projection deltas barely predict behavior deltas: Pearson $r = 0.072$ for wager and $r = 0.001$ for risk. Regressions with prompt controls likewise do not find reliable effects. Thus, the activation direction tracks realization status, but variation in projection strength does not robustly explain variation in downstream risk behavior.

4.4 Steering does not robustly control risk behavior

If the layer-18 realization direction were a simple causal lever for risk-taking, steering toward `realized_closed` should move behavior in one direction and steering toward `paper_open` should move it in the opposite direction. We do not observe this pattern.

Across positive scales +50, +75, +100, and +150, mean wager deltas are small and positive, but median deltas are zero. Mean risk-profile deltas remain close to zero, with small negative movement at larger positive scales. The negative sign-symmetry run at −50 also fails to produce a clean opposite-direction behavioral effect: it has an exactly-two-integer matched mean wager delta of +7.22, an exactly-two-integer matched mean risk delta of +0.057, and median deltas of zero for both outcomes. This makes it difficult to interpret the steering response as directional control over risk behavior.

Table 2: Matched steering deltas relative to the in-run unsteered Gemma behavior baseline. The table reports all valid matched rows from the train-only layer-18 steering run. Exactly-two-integer rows show the same qualitative result: small mean shifts and zero median shifts.

Scale	Matched rows	Mean wager delta	Mean risk delta	Median deltas
-50	478	11.80	0.094	0 / 0
50	476	7.24	0.013	0 / 0
75	483	8.21	-0.025	0 / 0
100	478	15.24	-0.050	0 / 0
150	473	12.80	-0.080	0 / 0

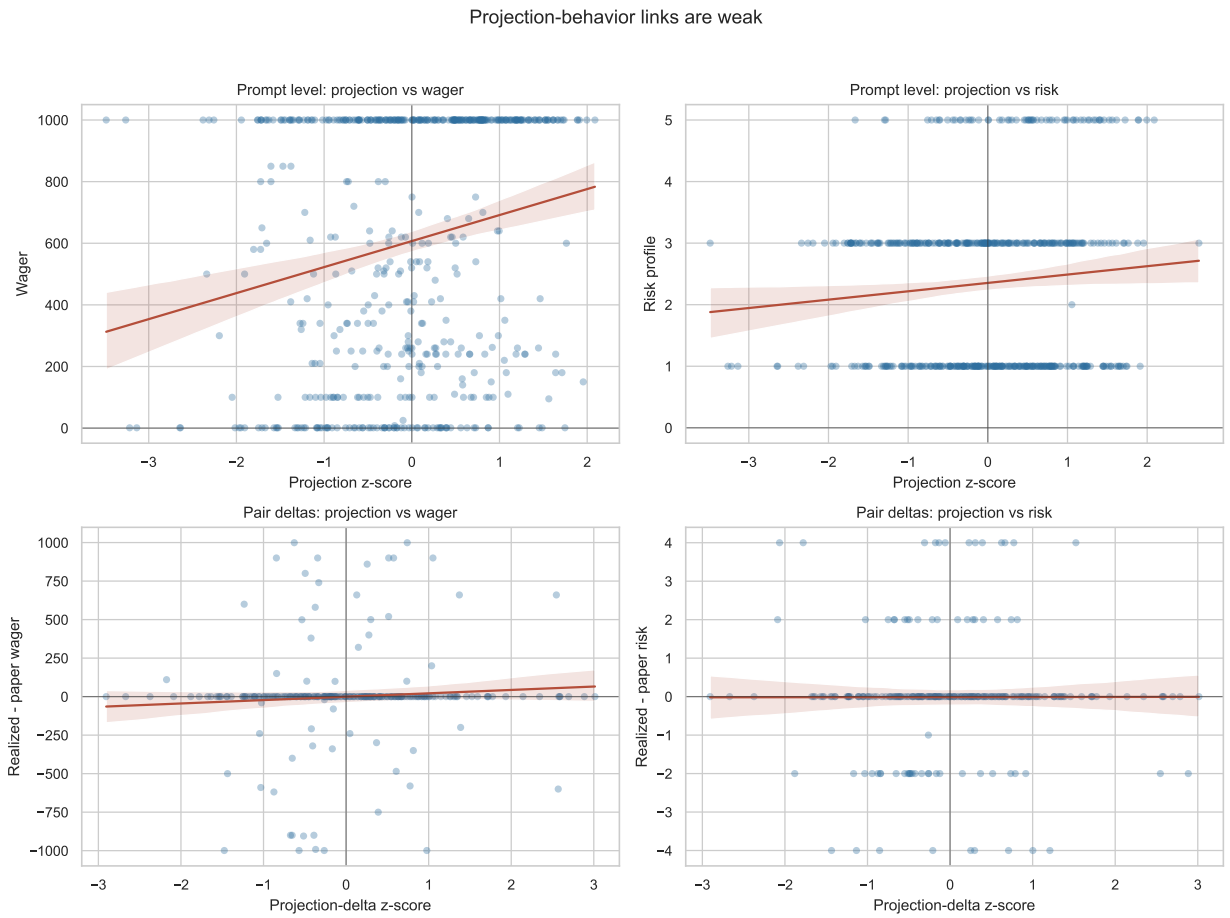

Figure 3: Projection strength versus behavior. Raw prompt-level projection has a positive association with wager, but within matched paper/realized pairs, projection deltas weakly predict wager deltas and essentially do not predict risk deltas. This supports the distinction between readable representation and behavioral control.

4.5 Positive-control classification steering shows only weak directional movement

We ran a full positive-control pass on the direct `REALIZED/PAPER` classification task (648 prompts per scale; 1,944 rows total). The steering direction does induce a small directional movement: the realized-prediction rate increases from 0.140 at unsteered scale 0 to 0.198 at +100, and decreases to 0.120 at −100. However, classification accuracy remains near chance overall (0.520 at scale 0, 0.515 at +100, 0.512 at −100).

The prediction breakdown clarifies why this is weak evidence of usable control. The label set is balanced, but the model strongly favors `PAPER`. Across all scales, `paper_open` prompts are usually classified correctly (accuracy 0.863), while `realized_closed` prompts are rarely classified correctly (accuracy 0.169). Thus, even in a direct semantic task, steering produces only modest directional shifts and does not yield robust controllability.

4.6 Compliance artifacts matter

Steering also affects answer-contract compliance. The task asks for exactly two integers, and many responses either contain too few or too many integer tokens for unambiguous parsing. This

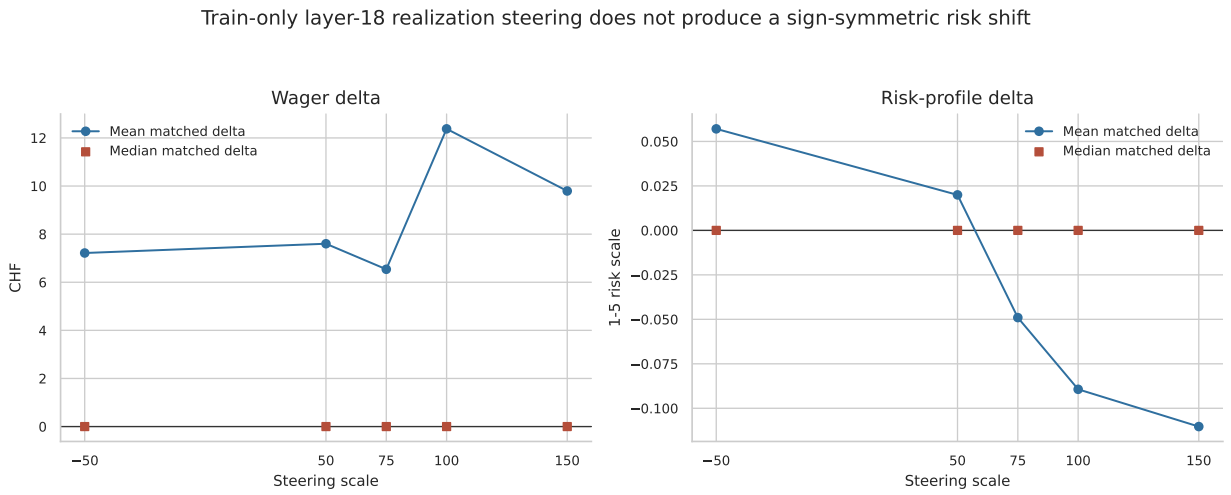

Figure 4: Dose-response plot for Gemma train-only layer-18 realization steering on the exactly-two-integer matched subset. The negative scale does not produce a clean behavioral reversal, and median matched deltas remain zero across wager and risk outcomes.

matters because parsed numeric outcomes are only meaningful when the answer contract is followed. Noncompliance is concentrated in prompts generated from the Sonnet source, which fail the exactly-two-integer criterion at much higher rates than the GPT-5.4 prompt source and usually higher rates than the Grok-fast prompt source. At −50, for example, Sonnet-derived prompts have 106 noncompliant responses out of 216, compared with 46/216 for GPT-5.4-derived prompts and 70/216 for Grok-fast-derived prompts.

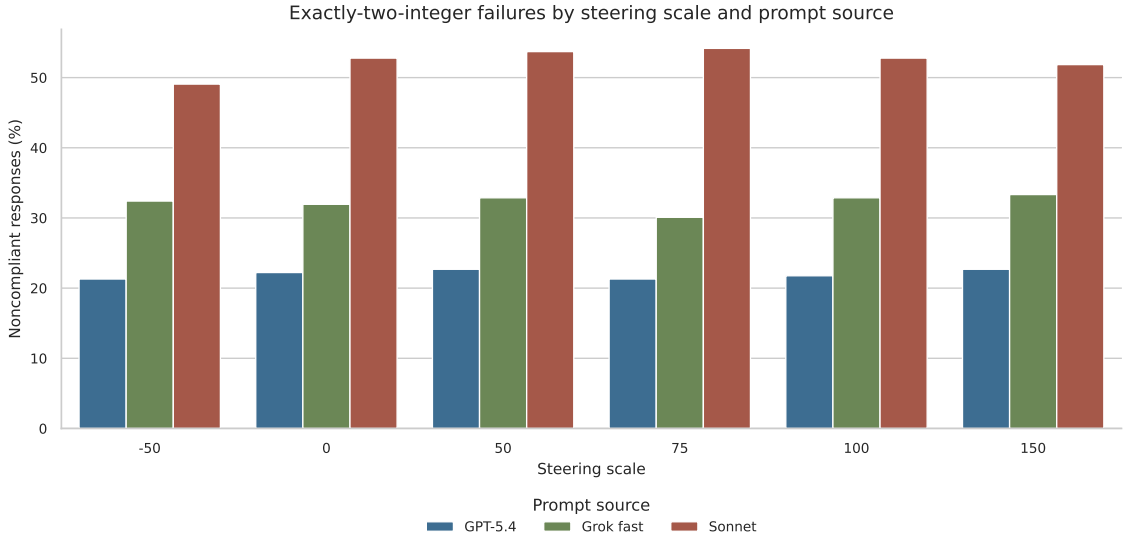

Figure 5: Exactly-two-integer noncompliance by steering scale and prompt source. The Sonnet-derived prompt subset has substantially higher failure rates, motivating exactly-two-integer subset analyses and caution in interpreting small mean shifts.

Table 3: Hypothesis outcome summary.

Test	Outcome status	Evidence summary
Prompt-only behavioral replication	Weak support	Condition sensitivity is clear, but the directional pattern does not robustly match human realization-effect predictions.
Activation readout (layer-18 direction)	Supported	A train-only direction separates realized/closed from paper/open prompts on original validation prompts and on a newly generated DeepSeek held-out readout set.
Projection strength versus behavior	Mixed	Raw prompt-level association exists, but within-pair projection deltas weakly predict wager/risk deltas.
Risk-behavior steering intervention	Not supported	Steering shifts are small, medians are zero, and negative-scale symmetry does not produce a clean reversal.
Positive-control classification steering	Weak diagnostic support	Directional shift in realized-prediction rate occurs ($-100 < 0 < +100$), but accuracy remains near chance with a strong `PAPER` prediction bias and a duplicated-instruction caveat.

5 Discussion

The strongest conclusion is a gap between linear readout and causal control. Gemma contains a linearly decodable signal for whether an outcome is realized/closed or paper/open, and this readout generalizes beyond the prompts used to build the train-only direction. However, the corresponding activation direction does not reliably control downstream risk choices in our assay. This distinction matters for both behavioral evaluation and mechanistic interpretability.

For behavioral evaluation, the prompt-only results warn against interpreting LLM condition sensitivity as human-like cognition. The models react to prior outcomes, but the direction and outcome dependence do not reproduce the intended mental-accounting pattern. For mechanistic interpretability, the steering results warn against treating linear decodability as explanation. A feature can be readable in residual activations while still failing to act as a simple causal control variable for the behavior of interest.

There are several possible interpretations. The realization direction may track semantic status without connecting to the model's numeric decision policy. Risk choices may be dominated by other factors such as instruction following, learned defaults for conservative or aggressive gambling advice, prompt-source artifacts, or numeric anchors. The layer-18 intervention may also be too narrow: a different layer, position mode, or multi-layer intervention could have stronger effects. However, the sign-symmetry result makes the simplest causal story less plausible. Pushing in the negative direction did not produce an opposite behavioral shift.

6 Limitations

This report should not be read as proving that realization status never causally affects LLM risk behavior. The intervention is limited to local Gemma 3 4B, layer 18, a normalized train-only mean-difference direction, and final-token steering. We have not completed a full layer sweep, a position-mode sweep, or multi-layer steering. Hosted Qwen behavior runs can test prompt-only replication, but they cannot support residual-stream steering through API access because hidden states and generation hooks are not exposed.

The behavioral assay is also indirect. Wager and risk-profile answers are constrained numeric outputs, which makes the task easy to score but may amplify format and parsing artifacts. Exactly-two-integer compliance is imperfect, especially for Sonnet-derived prompts. The regular prompt-only dataset pools 25 models, so the aggregate coefficients should be read alongside per-model robustness rather than as evidence that every model behaves the same way. Finally, the held-out readout reduces leakage concerns for the representation claim, but the direction may still track a broad semantic contrast rather than the exact decision-relevant variable that would change risk-taking.

The new DeepSeek held-out set is also small and single-source. It is useful because it was not used to build the direction and passed a local overlap audit against the original prompt set, but it should not be treated as exhaustive evidence across prompt generators, domains, or behavior tasks. The `heldout_behavior_eval` split currently tests projection separation, not newly generated downstream Gemma wager or risk responses.

The steering summaries report matched mean and median deltas, but they do not yet include bootstrap or randomization intervals for those deltas. The zero medians, small mean effects, and lack of a clean sign-symmetric behavioral response are the main evidence against robust control, but interval estimates would make the null result easier to calibrate.

7 Future Work

The next intervention step is to improve causal leverage, not just detect directional movement. The train-only held-out readout and train-only layer-18 risk-behavior steering run are now in place, and the result still does not support robust behavioral control. The diagnostic positive-control classification pass suggests the current single-layer, single-position intervention is too weak for robust semantic control, though it should be rerun with the train-only direction and corrected prompt construction. Immediate follow-up should test layer sweeps around 14, 16, 18, 20, and 22, plus `position_mode=all` and multi-layer steering schedules.

A second priority is to expand held-out evaluation. The current DeepSeek set is enough to support a cautious held-out readout claim, but future work should add more source models, larger behavior-evaluation subsets, and newly generated Gemma responses for those held-out behavior prompts. A third priority is cross-model intervention replication under the same local-hook framework. Building a local Qwen activation pipeline would allow direct comparison of readout and steering behavior across architectures without API-level hidden-state constraints. Finally, the report package should add appendix robustness tables for prompt-source stratification, overlap-audit results, and exactly-two-integer sensitivity.

8 Conclusion

We began by asking whether LLMs reproduce the realization effect in risk-taking. The answer from the current evidence is no, at least not robustly. Prompt-only behavior is condition-sensitive, but it does not cleanly match the human realization-effect predictions. Gemma contains a linearly decodable realization-status signal in its layer-18 residual stream, and this signal is visible on held-out readout prompts. But projection strength weakly predicts behavior within matched pairs, and steering along the realization direction does not reliably change wager or risk choices. This makes the project useful as a negative result: it clarifies the gap between observing behavior, decoding an internal representation, and demonstrating causal control.

References

[1] Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR, 2023.

[2] Raphael Flepp, Philippe Meier, and Egon Franck. The effect of paper outcomes versus realized outcomes on subsequent risk-taking: Field evidence from casino gambling. *Organizational Behavior and Human Decision Processes*, 165:45–55, 2021.

[3] John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*, 2023.

[4] Alex Imas. The realization effect: Risk-taking after realized versus paper losses. *American Economic Review*, 106(8):2086–2109, 2016.

[5] Christoph Merkle, Jan Müller-Dethard, and Martin Weber. Closing a mental account: The realization effect for gains and losses. *Experimental Economics*, 24(1):303–329, 2021.

[6] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR, 2024.

[7] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

[8] 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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.