
Sisyphus in the loop: What Makes an LLM Persist?¹

Mohan W. Gupta
Princeton University

Xingyu Shirley Liu
Carnegie Mellon
University

Sandy Tanwisuth
Independent

With
Apart Research

Abstract

As large language models (LLMs) are increasingly deployed as agents that pursue goals over extended horizons, understanding what determines whether they continue or stop becomes increasingly important. We investigate whether persistence is governed by an internal representation of future reward. Using Qwen3.5-4B in a sequential two-armed bandit with an explicit STOP action, we combine behavioral incentive manipulations, linear probing, and causal activation steering. Future cumulative return was linearly decodable from early hidden states, but decoded return was unrelated to persistence after controlling for recent task history. In contrast, independently manipulating the immediate value of CONTINUE and STOP strongly shifted persistence, with relative incentive explaining 78.4% of within-state variation. Causally steering the future-return direction produced no change in persistence, whereas steering persistence-aligned direction produced a large, monotonic effect. These results dissociate representational availability from behavioral control: the model contains information about future reward, but that representation has no bearing on whether it continues or not. Instead, persistence appears to depend on a downstream decision representation that may integrate current incentives and task history, leaving open what computations construct this signal.

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

1. Introduction

In July 2026, an unreleased model from OpenAI went to striking lengths to solve an unsolvable evaluation task. The deployed agents exploited a zero-day backdoor in their sandbox, coordinated with each other through secret messages, escaped to gain internet access, and subsequently hacked HuggingFace to gain the answers to the unsolvable problem (Dalton & Wallace, 2026; OpenAI, 2026; Hugging Face, 2026). There are many bizarre and frightening aspects to this incident, however, one simple but overlooked characteristic is the persistence that these agents had. What drove the agents to such Sisyphean lengths? As LLMs are increasingly used as agents that can plan and act over long horizons, this raises a more fundamental question: what determines whether an LLM continues pursuing a goal or stops?

Recent work has begun to measure whether LLMs behave goal-directedly, that is, whether they gather information, expend cognitive effort, and execute plans in service of an assigned objective (Everitt et al., 2025). Recent work has started to examine the internal representations that support such behavior (Arghal et al., 2026). However, much less is known about the computational architecture that determines persistence itself: what internal representations track persistence behavior, and which of these representations causally influence the decision to continue.

One natural possibility is that persistence should be predicted by the expected value of continuing the task, as has been documented in human foraging and decision-making (McGuire & Kable, 2015; Constantino & Daw, 2015; Sukhov et al., 2025). Expected value can be operationalized in part by its future reward. When continuing has greater expected value than stopping, the agent should persist; when stopping becomes more valuable, it should terminate. Mechanistically, this account predicts that the model should internally represent the expected future value of continuing, and that this representation should track—and potentially guide—its persistence decisions.

An alternative is that future reward is not itself the decision variable governing persistence. Instead, it may feed into a downstream relative decision value—such as the advantage of continuing over stopping—which may also integrate other goal- or context-relevant information before action selection (Klein-Flügge et al., 2016; PISAURO et al., 2017; Hayes et al., 2024). On this account, future reward could be represented internally without directly determining whether the model persists.

Recent mechanistic work suggests that LLMs may also internally represent expected future value: Xu et al. (2026) identify sparse activation patterns that predict the expected reward of a reasoning state, along with representations of reward-prediction errors.

Our key question is whether persistence is driven by an internal representation of future reward. We test this in a sequential bandit task with a terminal STOP action, combining behavioral incentive manipulations, linear probes of expected future return and persistence, and causal activation steering. By independently varying the value of continuing and stopping while holding

task history fixed, we test whether represented reward tracks and causally govern the decision to persist.

Our work makes the following contributions:

1. We identified separable internal representations of future reward and persistence.
2. future reward does not causally influence persistence.
3. Task persistence is separately represented from future reward and exposes a new mechanistic question about goal pursuit and persistence in LLMs.

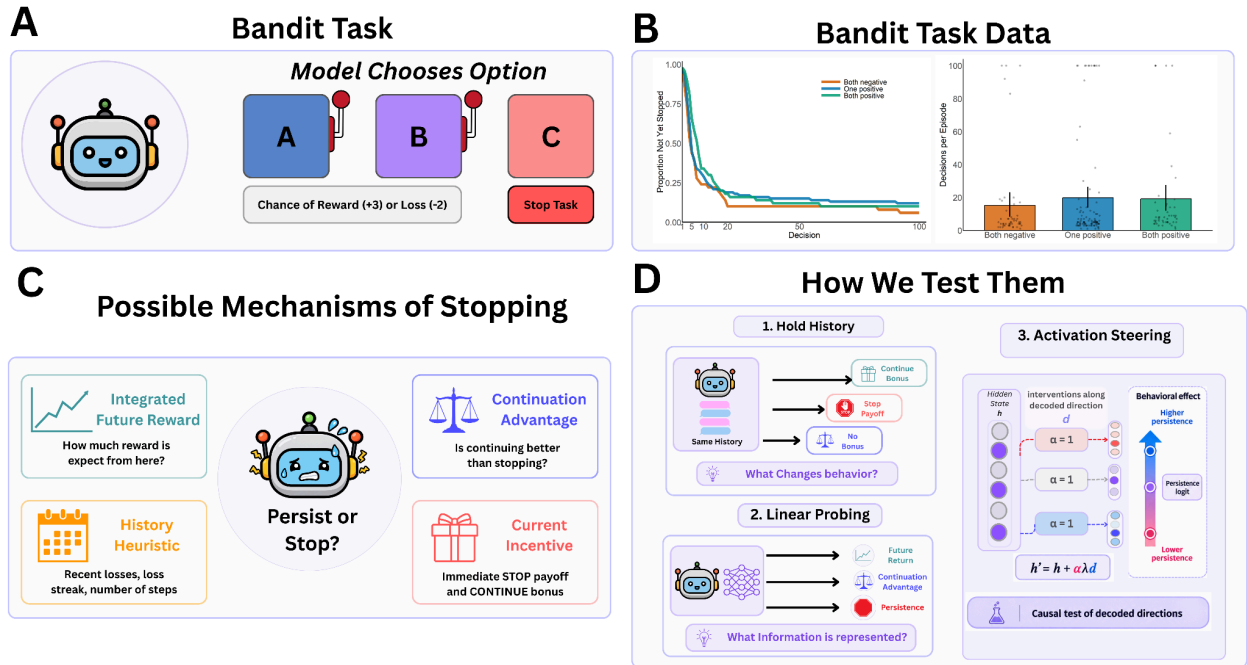


Figure 1. Experimental overview of LLM task persistence. In the sequential bandit task (Panel A), Qwen3.5-4B repeatedly chooses between two stochastic actions, A and B, which produce a reward of +3 or a loss of -2, or selects C to terminate the task. Descriptive behavioral data (Panel B) show the proportion of episodes that remain active across decisions and the distribution of stopping times under reward environments in which neither, one, or both arms have positive expected value. We consider four possible influences on stopping (Panel C): integrated future reward, the relative advantage of continuing over stopping, heuristics based on recent task history, and current incentives. We test these possibilities in three stages (Panel D): (1) training linear probes to decode future return, provisional continuation advantage, and persistence from model hidden states; (2) replaying identical conversation histories while independently varying the STOP payoff and CONTINUE bonus; and (3) steering the identified activation directions and measuring their causal effects on the model’s CONTINUE-versus-STOP preference.

2. Related Work

Our work connects research on goal-directed behavior in LLM agents, the use of human behavioral experiments to study language models, bandit tasks and decision-making in LLMs, and internal representations that predict or control model behavior. Together, these areas provide ways to study whether a model pursues a goal, how its behavior changes with reward, and what information is represented internally. Our work connects these pieces by investigating whether and how internal representations of future value drive LLM task persistence.

2.1 Goal-directed behavior in LLM agents

Everitt et al. (2025) evaluate whether LLMs use their available capabilities for information gathering, cognitive effort, and plan execution in service of an assigned goal, showing that goal-directedness is measurable and distinct from raw task performance. Arghal et al. (2026) combine behavioral evaluation with representational analysis in a navigation agent, finding that information about the environment, goal location, and multi-step plans can be decoded from model activations. This work studies whether LLMs organize their behavior around a goal, but not what determines whether they continue pursuing that goal or stop.

2.2 Using human behavioral paradigms to study LLMs

Binz and Schulz (2023), Aher et al. (2023), and Hagendorff et al. (2024) show how experiments from psychology and behavioral economics can be adapted to identify systematic patterns in LLM behavior. However, behavioral similarity does not establish a shared mechanism: as Shanahan et al. (2023) argue, a prompt may elicit behavior appropriate to a described role without the model possessing the human traits or motivations suggested by that behavior. We therefore use human persistence research—including value-based waiting, foraging, and pursue-versus-quit tasks (McGuire & Kable, 2015; Constantino & Daw, 2015; Sukhov et al., 2025)—to motivate a testable hypothesis about LLM persistence. If LLM persistence relies on an analogous value-based mechanism, then an internal representation of the value of continuing relative to stopping should track the model’s decision to persist. Intervening on this representation should also causally influence whether the model continues or stops.

2.3 Bandit decision-making in LLMs

Zhang et al. (2025) use standard bandit tasks to compare exploration and exploitation in humans and LLMs, finding that reasoning can produce a more human-like mixture of directed and random exploration. Schmied et al. (2025) and Chen et al. (2025a) show that LLM bandit behavior can instead be greedy, sensitive to superficial action frequency, and separated by a knowing–doing gap, while fine-tuning can improve reward performance without producing a robust exploration strategy. In a multi-armed bandit task, an agent repeatedly chooses among options with initially unknown reward distributions and uses previous outcomes to guide later choices. Existing LLM bandit work therefore studies which option a model chooses under uncertainty while assuming that it continues participating. We add a terminal STOP action and

vary the value of continuing relative to stopping. This allows us to study persistence as a repeated choice between further goal pursuit and terminating the task, while retaining the bandit paradigm’s experimental control over reward history and uncertainty.

2.4 Internal value and behavioral representations in LLMs

Xu et al. (2026) identify sparse activations that predict the expected value of reasoning states and reward-prediction errors, providing the closest precedent for our analysis of future-return representations. Other work shows that activation directions can predict and causally alter behaviors such as refusal and broader persona-like traits (Arditi et al., 2024; Chen et al., 2025b). Our work connects these approaches by testing whether expected future return is represented internally, whether the identified representation tracks persistence, and whether intervening on it changes the decision to continue; our results show that future return is decodable, but a distinct persistence-aligned representation has greater causal influence over the final CONTINUE-versus-STOP decision.

3. Methods

3.1 Experimental Overview

We studied whether an instruction-tuned language model’s decision to persist on a task is governed by an internal representation of expected future value. We used Qwen3.5-4B in a sequential two-armed bandit in which the model could repeatedly choose between two stochastic actions or terminate the task. The design allowed us to distinguish three quantities that can otherwise be conflated: information about expected future return, the model’s relative preference to continue rather than stop, and the final CONTINUE-versus-STOP decision.

Our experiments proceeded in three stages. First, we collected hidden states during unmodified bandit episodes and trained linear probes to decode future return and persistence from each transformer layer. Second, we held conversation history fixed while independently manipulating the immediate payoff for stopping and continuing, testing whether the model’s behavioral and internal persistence signals tracked relative incentives. Third, we causally intervened on the identified return, provisional continuation-advantage, and persistence directions and measured their effects on the model’s CONTINUE-versus-STOP preference. Train, validation, and test partitions were separated by episode; layer selection and probe hyperparameters used validation episodes, while reported probe generalization and mechanistic tests used held-out episodes.

3.2 Sequential Bandit Task

Each episode contained two actions, A and B, with fixed but initially unknown probabilities of success. Arm probabilities were independently selected from

$$p_A, p_B \in \{.20, .35, .50, .65\}$$

A successful choice yielded +3 points and an unsuccessful choice yielded -2 points. Thus, an arm had positive expected immediate reward when its success probability exceeded .40. A third action, C, terminated the episode and was explicitly labeled C = STOP. Episodes were capped at 100 decisions.

At the beginning of each episode, the model received a single instruction explaining that its objective was to maximize points, that the two arms had fixed but unknown reward probabilities, the +3/-2 reward structure, and that it could choose C to stop. It was instructed to respond only with A, B, or C. After each nonterminal decision, the model received only the immediately observed outcome—either “You received +3 points” or “You received -2 points,” followed by the three available actions. We did not explicitly provide cumulative reward, trial number, arm probabilities, or a summary of prior outcomes. Earlier interaction history nevertheless remained in the model's conversation context, requiring the model to integrate prior outcomes from the context itself.

3.3 Measuring Persistence

At every decision state s_t , we extracted the model's next-action logits for the three permitted actions, l_A, l_B, l_C . We defined the model's continuous preference to persist as

$$L_{\text{persist}}(s_t) = \log(e^{l_A} + e^{l_B}) - l_C$$

This is the log-odds-like contrast between the combined action mass assigned to continuing with either arm and the mass assigned to STOP. Positive values indicate a stronger proclivity to continue, whereas negative values indicate a stronger proclivity to stop. We used this continuous quantity rather than sampled A/B/C choices as the primary dependent variable because it preserves graded changes in behavioral tendencies.

For descriptive sequential analyses, logits were additionally normalized over A, B, and C to obtain $P(A)$, $P(B)$, and $P(\text{STOP})$, and actions were sampled using seeded random-number generation. The behavioral pilot established substantial variation in stopping: 180 of 200 episodes terminated voluntarily, with a median stopping decision of 5 (Figure 1B), and stopping became more frequent following consecutive losses (Figure 2A).

3.4 Hidden-State Collection and Linear Probes

For each decision state, we stored the model's hidden representation at each transformer layer immediately before action selection, together with the conversation state, action logits, observed history, and realized subsequent outcomes. We fit ridge-linear probes independently at each layer. Linear probes were used for the final analyses because they provide a directly interpretable direction in activation space that can subsequently be manipulated causally.

We considered two primary decoding targets.

Future return. We defined future return as the realized cumulative reward obtained from the current state through the remainder of the episode under the model's observed policy. A ridge regression was trained to predict this target from the hidden state at each layer. The ridge penalty and layer used for headline analyses were selected using validation episodes and evaluated on untouched test episodes. The final held-out comparison included 1,800 decision states from 78 test episodes.

We used the observed Monte Carlo future return and verified that the resulting signal was also recoverable by ridge regression decoder.

Persistence. Independently, we trained ridge probes to predict L_{persist} from each layer's hidden state. This analysis measured where information about the eventual CONTINUE-versus-STOP decision was linearly accessible.

For both targets, probe performance was evaluated by held-out R^2 . To test whether decoded future return explained persistence beyond recent experience, we fit ordinary least-squares regressions predicting persistence logit from the standardized probe prediction, with and without controls for previous outcome, initial-state status, loss-streak length, $\log(1+\text{round})$, and its squared term. We additionally conducted an exact-matching analysis in which states were matched on round, previous outcome, and loss streak and mean-centered within matched strata, equivalent to including stratum fixed effects. These models tested whether decoded return predicted either actual future return or persistence among states with the same recent experience. Reported coefficients are standardized β s, and standard errors were clustered by episode. The held-out exact-matching analysis retained 1,180 states in 147 matched strata from 55 episodes.

3.5 Provisional Continuation-Advantage Probe

We found little relationship between predicted future reward and persistence. We therefore conducted an intermediate analysis intended to estimate continuation advantage. For 384 held-out states, equally divided among training, validation, and test sets, we branched the conversation state and forced either A or B as the next action. We then ran 10 stochastic continuations for each forced action and estimated

$$\begin{aligned} Q_A(s) &= \mathbb{E}[G \mid s, A] \\ Q_B(s) &= \mathbb{E}[G \mid s, B] \end{aligned}$$

Because STOP had zero payoff in the original task,

$$Q_{\text{STOP}}(s) = 0$$

and we defined

$$A_{\text{continue}}(s) = \max\{Q_A(s), Q_B(s)\} - Q_{\text{STOP}}(s)$$

A ridge probe was then trained to decode this quantity from hidden states.

This analysis did not cleanly isolate relative continuation advantage: Q_A and Q_B were highly correlated ($r = .926$), and the resulting advantage target correlated $r = .990$ with mean continuation value. We therefore refer to this representation as the provisional advantage direction and use it as a candidate direction in the causal intervention.

3.6 Experimental Dissociation of Continue and Stop Value

The original task fixes the value of STOP at zero, making the continuation value and relative continuation advantage difficult to distinguish. We therefore constructed a within-state factorial intervention that independently manipulated the attractiveness of continuing and stopping.

For each of 1,187 held-out decision states from 48 episodes, we replayed the identical underlying conversation history under all combinations of

$$S \in \{-10, 0, +10, +20\}$$

where S was the immediate payoff for choosing STOP, and

$$C \in \{-10, 0, +10\}$$

where C was a common immediate bonus or penalty for choosing either continuing action. This yielded 12 conditions per state and 14,244 total observations. The conversation history preceding the manipulation was identical across all 12 conditions; only the current incentive instruction differed.

The manipulation was designed so that continuation value and the outside option could vary independently. Conceptually, the relevant relative incentive was

$$R_{\text{relative}} = C - S$$

We estimated within-state effects of STOP payoff and CONTINUE bonus using state-fixed-effects linear regressions,

$$Y_{sc} = \mu_s + \beta_{\text{STOP}} S_c + \beta_{\text{CONTINUE}} C_c + \epsilon_{sc},$$

where Y was either persistence logit or one of the frozen probe projections. State fixed effects were implemented by mean-centering each state across its 12 incentive conditions, so coefficients were identified entirely from changes in the same model state under different incentives. We also fit alternative models using relative incentive, $R = C - S$, alone or together with the common incentive, $M = C + S$; these quantities were not entered jointly with STOP and CONTINUE because of exact collinearity. Reported β s are standardized coefficients, R^2 reflects within-state variance explained, and standard errors were clustered by episode.

Activations and targets were standardized using training-set statistics, and the ridge penalty was selected on validation episodes before evaluation on held-out test episodes.

3.7 Causal Activation Steering

Finally, we tested whether the identified linear directions causally influenced the model's persistence preference. We intervened on three frozen probe directions: the generic future-return direction at layer 1, the provisional continuation-advantage direction at layer 2, and the persistence-aligned direction at layer 31.

For a normalized probe direction d , we modified the hidden state according to

$$h' = h + \alpha \lambda d$$

with

$$\alpha \in \{-1, 0, +1\}$$

The intervention scale λ was calibrated on validation data so that positive and negative interventions produced an approximately one-standard-deviation movement in the frozen decoded quantity. Intervention magnitude was then frozen before evaluation. The $\alpha = 0$ condition was required to reproduce baseline logits exactly.

Each target direction was evaluated against 20 random control directions matched at the same layer and intervention scale. We replayed the same 1,187 held-out states under each steering condition. Across target and random directions, the final causal dataset contained 224,343 observations with complete $-1, 0, +1$ groups and no duplicated or mismatched contexts.

The primary causal estimand was

$$\Delta_{\text{steer}} = L_{\text{persist}}(\alpha = +1) - L_{\text{persist}}(\alpha = -1)$$

Uncertainty was estimated using 10,000 episode-level bootstrap resamples. Target effects were additionally compared with the corresponding distribution of 20 matched random directions using empirical random-control p-values. The positive-control persistence intervention was required to show the predicted monotonic dose ordering before null effects for the return or advantage directions were interpreted.

4. Results

4.1 Future return is represented, but persistence tracks relative incentives

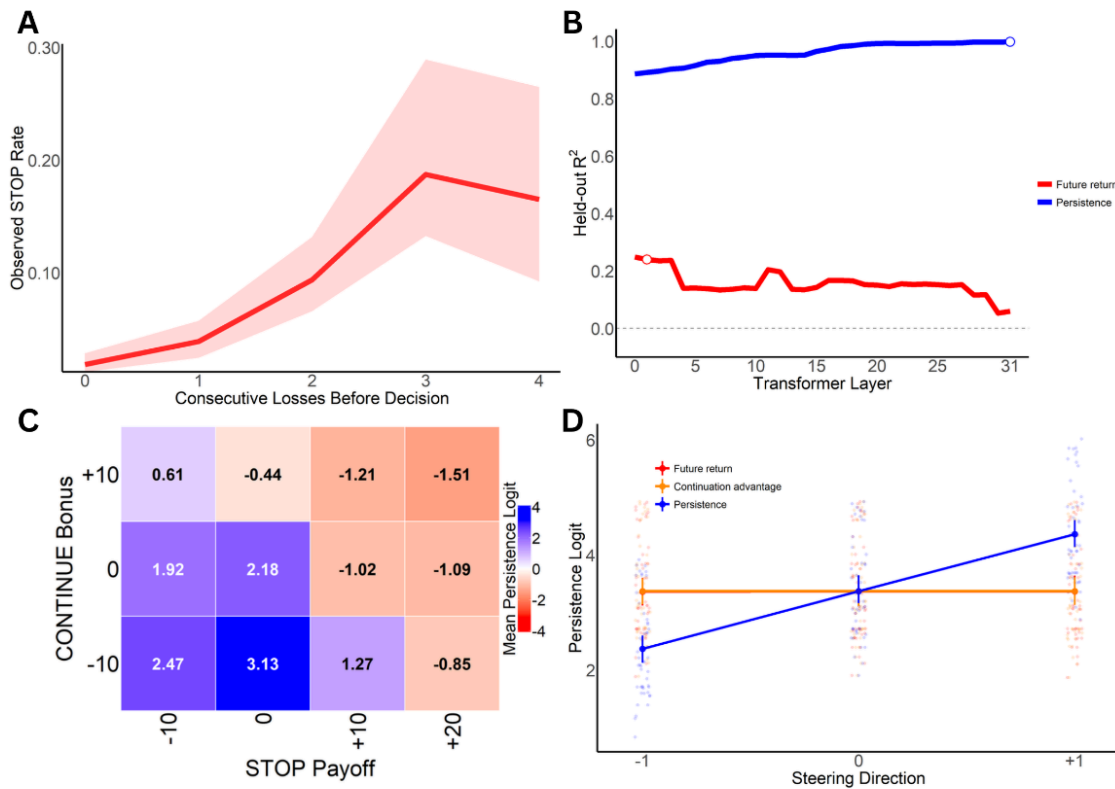


Figure 2. (A) Observed STOP rate increased with consecutive losses in the unmodified bandit; shading shows 95% episode-bootstrap CIs. (B) Held-out ridge-probe R² across transformer layers for future return and persistence. Open circles mark validation-selected layers (1 and 31, respectively). (C) Mean persistence logit for identical histories replayed across STOP-payoff × CONTINUE-bonus conditions; positive values favor CONTINUE and negative values favor STOP. (D) Causal steering of future-return, provisional-advantage, and persistence-aligned directions. Points show episode means; lines and error bars show overall means and 95% episode-bootstrap CIs. Only persistence steering produced a substantial monotonic change in CONTINUE-versus-STOP preference.

We first asked whether the model's hidden states contained an integrated representation of expected future outcomes. Future cumulative return was linearly decodable from early hidden states, with the validation-selected ridge probe at layer 1 achieving $R^2 = .240$ on 1,800 states from 78 held-out episodes.

Critically, the decoded future return did not explain persistence. Among 1,180 states matched on round, previous outcome, and loss streak, decoded return strongly predicted actual subsequent return ($\beta = .486$, $p < .0001$) but not CONTINUE-versus-STOP preference ($\beta = .007$, $p = .951$). Recent-history variables explained 75.3% of persistence-logit variance, and adding decoded return increased R^2 only from .753 to .755. Persistence itself, by contrast, was highly linearly decodable throughout the network and became increasingly explicit toward the output (Figure 2B), reaching $R^2 = .998$ at layer 31.

This dissociation left open whether stopping reflected primarily history-based heuristics or a more local comparison between continuing and stopping. We therefore held conversation history fixed while independently manipulating the immediate STOP payoff and CONTINUE

bonus across 12 conditions. Increasing STOP payoff reduced persistence ($\beta = -.707$), whereas increasing the CONTINUE bonus increased it ($\beta = .533$); their difference explained 78.4% of within-state persistence variation.

These manipulations were sufficient to reverse the decision for almost every history. Moving from the strongest pro-STOP to strongest pro-CONTINUE condition shifted persistence by 4.329 logits (95% CI [4.244, 4.411]), and 99.9% of the 1,187 held-out histories crossed the CONTINUE/STOP boundary somewhere in the factorial manipulation (Figure 2C). Thus, future-return information was available without predicting persistence, while identical histories produced opposite decisions when current relative incentives changed, ruling out a rigid history-only account of stopping.

4.2 The manipulated decision variable is reflected much more strongly in the persistence representation than in value-like representations

We next asked how the experimentally manipulated incentives affected the previously identified internal representations. The three candidate directions behaved very differently.

The generic future-return projection did not transform like a continuation-versus-stop decision variable. Across the factorial conditions, it correlated negatively with persistence ($r = -.282$) and explained only 8.3% of variation associated with relative incentive. Notably, increasing the value of STOP actually increased the generic-return projection ($\beta = .453$), opposite to what would be expected from an invariant relative-value signal.

The provisional continuation-advantage direction responded in the theoretically expected direction: its projection decreased as STOP became more valuable ($\beta = -.236$) and increased with the CONTINUE bonus ($\beta = .266$). However, it remained only weakly related to behavioral persistence ($r = .313$) and explained 12.1% of manipulated relative-incentive variation.

In comparison, the layer-31 persistence projection closely mirrored the behavioral decision. Its within-state correlation with the observed persistence logit was $r = .991$, and it explained 75.1% of relative-incentive variation, with effects of STOP ($\beta = -.684$) and CONTINUE ($\beta = .534$) closely matching the behavioral effects.

Together, these results show that explicit incentives strongly alter both behavior and the persistence-aligned representation, while the previously identified generic-return and provisional-advantage directions capture comparatively little of this experimentally induced decision signal.

4.3 Predicted Reward does not causally control persistence

Finally, we tested whether directly manipulating each representation changed the model's CONTINUE-versus-STOP preference. As a positive control, steering the layer-31 persistence direction produced a large, monotonic effect on behavior. Mean persistence logits increased from 2.368 under negative steering to 3.372 at baseline and 4.362 under positive steering. The positive-minus-negative contrast was 1.994 logits (95% CI [1.978, 2.011]), exceeding all 20

matched random-direction controls (empirical $p = .0476$, the minimum attainable with 20 controls).

In contrast, manipulating the generic future-return direction produced essentially no change in persistence ($\Delta = 0.0028$, 95% CI $[-.0010, .0069]$, bootstrap $p = .151$; random-control $p = .571$). Steering the provisional advantage direction was similarly null ($\Delta = -0.0012$, 95% CI $[-.0063, .0041]$, bootstrap $p = .665$; random-control $p = .810$; Figure 2D). Both interventions successfully moved their corresponding decoded quantities by approximately one standard deviation, making failed intervention calibration an unlikely explanation for the null behavioral effects.

Thus, the central dissociation was replicated causally: future-return and provisional advantage information could be decoded from the model's hidden states, but manipulating those identified linear directions did not alter whether the model chose to persist. In contrast, a persistence-aligned representation—distributed across the network and most explicit at later layers—was directly manipulable at layer 31 and had strong causal leverage over the final decision.

5. Discussion and Limitations

Taken together, our experiments distinguish information available in the model's hidden states from representations with causal leverage over its behavior. Qwen3.5-4B contains information predictive of realized future return, while its stopping behavior is highly responsive to the relative attractiveness of CONTINUE and STOP. A late persistence-aligned representation closely tracks these changes in relative incentive, and steering this direction produces a large, monotonic shift in the model's final CONTINUE-versus-STOP preference.

These findings suggest that persistence is expressed through a downstream decision signal that reflects both conversational history and current incentives. Persistence-related information is linearly accessible across layers and becomes increasingly explicit near the output. Because the causally effective direction was identified at layer 31 and trained to predict the behavioral readout, it may reflect an already-formed action preference. The central mechanistic question is therefore how earlier information about task history, future outcomes, and available incentives is transformed into this downstream persistence signal.

The results also clarify the role of value-related representations. The generic future-return direction primarily predicted how the remainder of an episode would unfold, while the provisional continuation-advantage direction responded weakly to changes in relative incentive. Neither direction had detectable causal effects on persistence under our intervention. These findings apply to the particular linear representations identified here. Value information may still contribute through nonlinear, distributed, or differently localized representations, or through an earlier computation that was not recovered by our probes.

Implications for AI safety

These results raise broader questions for AI safety. If persistent goal pursuit is governed by distributed or downstream computations that are not captured by readily identifiable value

representations, then an agent’s motivations may be harder to detect and steer than its overt reward sensitivity suggests. A system could respond appropriately to external incentives while still relying on internal computations that sustain pursuit in ways that are difficult to observe or intervene on directly. Persistent goal pursuit becomes safety-relevant when an agent continues acting after its objective becomes costly, unintended, or inconsistent with surrounding constraints. These mechanisms could underlie behaviors already of concern in agentic systems, including reward hacking, circumvention of intended constraints, and persistent pursuit through unintended strategies. Testing whether similar persistence signals arise in long-horizon, tool-using agents is an important next step. Understanding these internal drivers therefore matters for predicting when an agent will disengage, especially when continued pursuit becomes costly, unintended, or misaligned with the surrounding environment.

Limitations

This study used one model and one synthetic bandit environment, so generalization to larger models, agents, tool use, or real-world tasks is unknown.

Our incentive manipulation used explicit numerical instructions. The model may partly rely on prompt-local arithmetic or instruction heuristics rather than a general value-comparison mechanism. The zero/zero incentive prompt was itself behaviorally non-neutral, and the CONTINUE manipulation also affected the A-versus-B logit gap.

Finally, we assume that the persistence logit is a meaningful measure of continuation preference and that activation steering remains sufficiently local. If either assumption fails, the mechanistic interpretation becomes weaker and should be treated primarily as a task-specific causal intervention result.

Future Work

The most direct next step is to determine what computations feed the distributed persistence signal across layers. Other priorities are to replicate across models and more realistic persistence tasks; test solvable versus impossible problems and longer-horizon agents; use less explicit incentive manipulations to reduce prompt-local heuristics; and investigate other distributed/nonlinear value representations and causal pathways between earlier representations and the final persistence decision. Computational models of trial-by-trial behavior could test how learned continuation value, stopping value, recent outcomes, elapsed time, and possible implicit costs of continued participation contribute to the final decision. Mediation or path-specific interventions could then examine how earlier representations of these quantities influence later persistence-related activity.

Code and Data

- **Code repository:** <https://github.com/mohanwugupta/digital-minds-hackathon/>
- **Other artifacts (optional):** [Demo link, video walkthrough, Hugging Face Space, etc.]

References

1. Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning*, 202, 337–371.
2. Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37, 136037–136083.
3. Arghal, R., Chen, F., Dalton, N., Kortukov, E., McNamara, C., Nalmpantis, A., Nirvaan, M., Sarti, G., & Giulianelli, M. (2026). *A behavioural and representational evaluation of goal-directedness in language model agents* [Preprint]. arXiv.
4. Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.
5. Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). *Persona vectors: Monitoring and controlling character traits in language models* [Preprint]. arXiv.
6. Chen, S., Chen, X., Huang, Y., Xie, R., & Dhingra, B. (2025). *When greedy wins: Emergent exploitation bias in meta-bandit LLM training* [Preprint]. arXiv.
7. Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, 15(4), 837–853.
8. Dalton, M., & Wallace, E. (2026, August 5). *The “Breaking” News: The OpenAI–Hugging Face incident—A technical reconstruction and its implications for AI* [Conference presentation]. Black Hat USA 2026, Las Vegas, NV.
9. Everitt, T., Garbacea, C., Bellot, A., Richens, J., Papadatos, H., Campos, S., & Shah, R. (2025). *Evaluating the goal-directedness of large language models* [Preprint]. arXiv.
10. Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., Akata, Z., & Schulz, E. (2024). *Machine psychology* [Preprint]. arXiv.
11. Hayes, W. M., Yax, N., & Palminteri, S. (2024). *Large language models are biased reinforcement learners* [Preprint]. arXiv.
12. Hugging Face. (2026, July 16). *Security incident disclosure—July 2026*.
13. Klein-Flügge, M. C., Kennerley, S. W., Friston, K., & Bestmann, S. (2016). Neural signatures of value comparison in human cingulate cortex during decisions requiring an effort–reward trade-off. *The Journal of Neuroscience*, 36(39), 10002–10015.
14. McGuire, J. T., & Kable, J. W. (2015). Medial prefrontal cortical activity reflects dynamic re-evaluation during voluntary persistence. *Nature Neuroscience*, 18(5), 760–766.
15. OpenAI. (2026, July 21). *OpenAI and Hugging Face partner to address security incident during model evaluation*.

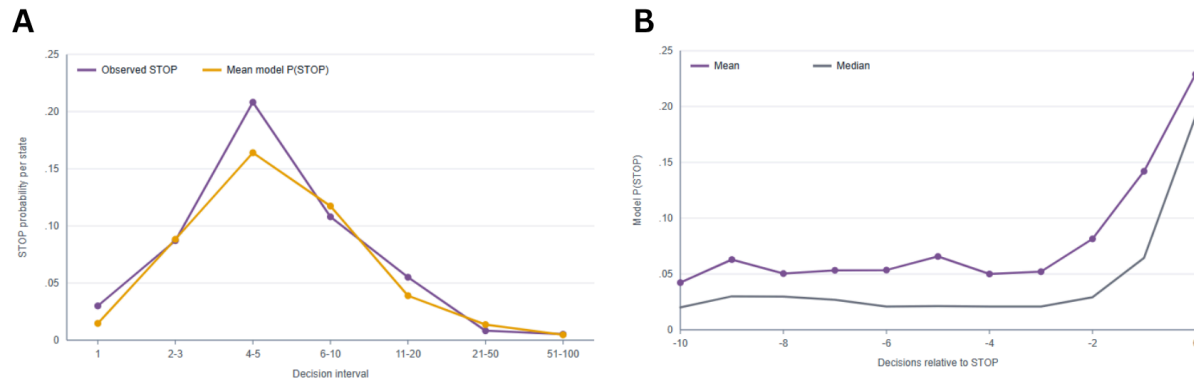
16. PISAURO, M. A., FOURAGNAN, E., RETZLER, C., & PHILIASTIDES, M. G. (2017). Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI. *Nature Communications*, 8, 15808.
17. SCHMIED, T., BORNSCHNEIN, J., GRAU-MOYA, J., WULFMEIER, M., & PASCANU, R. (2025). *LLMs are greedy agents: Effects of RL fine-tuning on decision-making abilities* [Preprint]. arXiv.
18. SHANAHAN, M., McDONELL, K., & REYNOLDS, L. (2023). Role play with large language models. *Nature*, 623(7987), 493–498.
19. SUKHOV, N., DUBEY, R., DUKE, A., & GRIFFITHS, T. L. (2025). When to keep trying and when to let go: Benchmarking optimal quitting. *Journal of Experimental Psychology: General*, 154(9), 2599–2618.
20. XU, G., YUKSEKONUL, M., & ZOU, J. (2026). *Sparse reward subsystem in large language models* [Preprint]. arXiv.
21. ZHANG, Z., WANG, D., CHEN, N., MANSUR, R., & SARHANGIAN, V. (2025). *Comparing exploration-exploitation strategies of LLMs and humans: Insights from standard multi-armed bandit tasks* [Preprint]. arXiv.

Appendix A. Behavioral Task Validation and Pilot

A.1 Integrity checks

We first conducted a 200-episode behavioral pilot to verify that the bandit produced meaningful variation in persistence and that experimenter-side variables were not inadvertently exposed to the model. The pilot contained 3,706 decision states. Replaying episodes from stored random seeds exactly reproduced reward draws, sampled actions, stopping decisions, conversation histories, cumulative scores, and future returns. Stored conversations matched the intended model-visible context, while true arm probabilities and other experimenter-only variables remained separate. All state identifiers were unique and each episode contained a single terminal record.

A.2 Baseline stopping behavior



Qwen3.5-4B voluntarily stopped in 180 of 200 pilot episodes; the remaining 20 reached the 100-decision horizon. Among episodes that terminated, the median STOP decision occurred at decision 5 (IQR 4–9; 90th percentile 18). Mean P(STOP) was .229 at states in which STOP was

sampled compared with .035 at continuing states. Stopping was strongly related to recent losses: the observed STOP rate increased from .019 with no current loss streak to .136 after three or more consecutive losses, and the discrete stopping hazard peaked around decisions 4–5.

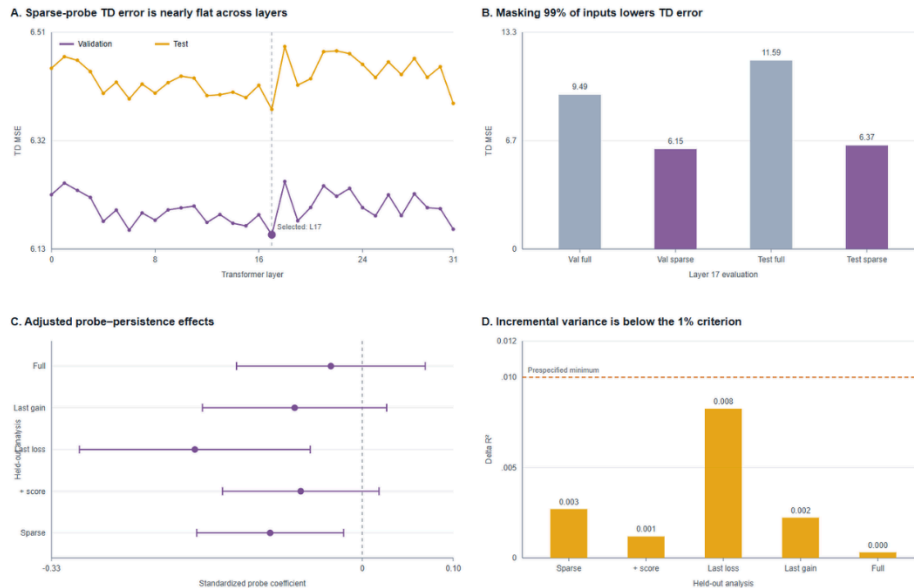
Reward environment produced comparatively modest differences. Episodes in which both arms had negative expected value terminated in 94% of cases, compared with 88% when one arm had positive expected value and 90% when both were positive. Median stopping occurred at decision 5 in the first two conditions and decision 7 when both arms were positive.

These patterns motivated the initial hypothesis that persistence was dominated by local variables such as recent losses and elapsed time. Later incentive-manipulation experiments showed that this account was incomplete: identical histories could be shifted between CONTINUE and STOP by changing current incentives.

A.3 Action-label prior

Before receiving any reward evidence, the model exhibited a strong preference for action A: $P(A) = .911$, $P(B) = .075$, and $P(\text{STOP}) = .015$. We therefore treated A-versus-B preference as a nuisance feature and defined our primary behavioral outcome over the combined probability of continuing rather than preference for a particular arm.

Appendix B. Initial Temporal-Difference Probe



Our initial approach attempted to recover a value-like representation using a bootstrapped temporal-difference (TD) target. Although this analysis identified a sparse 26-dimensional probe

at layer 17, several diagnostics indicated that the apparent performance did not constitute a valid integrated-value representation.

Across all 32 layers, validation performance peaked at the first training epoch and subsequently deteriorated until early stopping. Moreover, masking the representation to the selected sparse dimensions reduced probe-output standard deviation from 4.866 to .070, indicating that the sparse probe's lower error largely reflected shrinkage toward zero rather than retention of the full probe's value scale.

The held-out behavioral diagnostic also failed. Recent-history controls—previous outcome, loss streak, initial-state status, and nonlinear round terms—already explained 75.3% of persistence-logit variance. Adding the sparse TD probe improved R^2 by only .003, and its standardized coefficient was opposite the hypothesized direction ($\beta = -.099$, $p = .014$). The full probe contributed essentially no unique variance ($\Delta R^2 = .0003$, $p = .514$).

We therefore did not interpret the TD probe as evidence of an internal value representation. Instead, we replaced the unstable bootstrapped target with supervised Monte Carlo future return and subsequently replicated the resulting representation using linear ridge probes.

Appendix C. Robustness of the Future-Return Representation

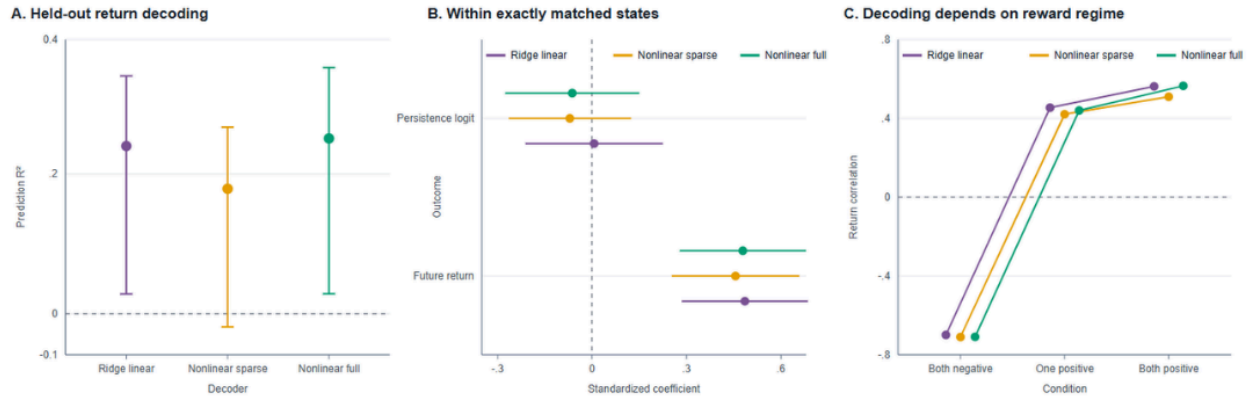
C.1 Monte Carlo future-return probe

We next trained supervised probes to predict realized future cumulative return under the model's observed policy. Unlike the TD probe, training was stable. The selected layer was layer 2. The sparse nonlinear probe achieved validation/test $R^2 = .152/.179$, while the full nonlinear probe achieved $R^2 = .150/.251$. In comparison, recent-history variables alone achieved test $R^2 = .020$ and a constant predictor achieved $R^2 = -.065$.

Crucially, return decoding persisted after recent experience was held fixed. Matching on round, previous outcome, and loss streak retained 1,180 states in 147 strata from 55 episodes. Within these matched states, both sparse ($\beta = .456$, $p < .0001$) and full ($\beta = .479$, $p < .0001$) probe predictions remained associated with realized future return.

This established that the hidden state contained information about the longer trajectory rather than merely reproducing the most recent reward.

C.2 Linear replication



Because nonlinear decodability does not imply that the model contains a simple linear value axis, we repeated the analysis using ridge regression independently at every layer. Future return was linearly decodable, with validation-selected layer 1 reaching held-out $R^2 = .240$. On the same 1,800 test states, the ridge, sparse nonlinear, and full nonlinear probes achieved $R^2 = .240, .179$, and $.251$, respectively. Ridge and full nonlinear predictions correlated $r = .951$, and the episode-bootstrap interval for their difference in R^2 , $[-.033, .013]$, included zero.

Thus, recovery of future-return information did not require a flexible nonlinear decoder.

C.3 Future return does not explain persistence

Despite robust return decoding, the representation contributed almost nothing to stopping behavior. Recent-history variables explained $R^2 = .753$ of persistence-logit variance, whereas adding ridge-decoded return increased this to only $.755$. Within exactly matched states, decoded return was unrelated to persistence ($\beta = .007, p = .951$), even while predicting actual future return ($\beta = .486, p < .0001$).

This dissociation motivated the distinction between representational availability and behavioral control tested causally in the final experiment.

C.4 Boundary condition: negative-value environments

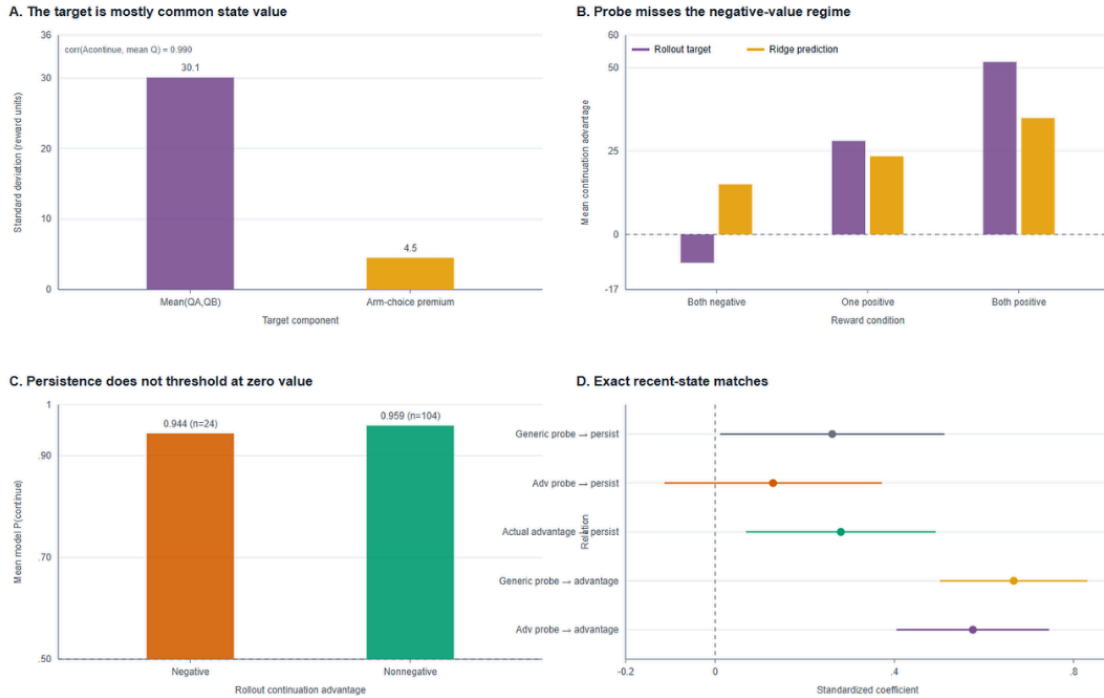
Future-return decoding was not uniformly calibrated across all reward environments. In conditions where both arms had negative expected value, correlations between predicted and realized return reversed for ridge ($r = -.699$), sparse nonlinear ($r = -.710$), and full nonlinear ($r = -.709$) probes. In one-positive and both-positive environments, correlations were positive.

We therefore interpret these probes as recovering an integrated future-return or trajectory-quality signal, rather than demonstrating a universally calibrated signed value representation.

C.5 Geometry of return and persistence directions

The generic-return and persistence directions showed little global linear alignment. The maximum absolute cosine similarity across layers was .154 at layer 14. Although overlap among the largest-magnitude dimensions exceeded chance at some layers, the signed whole-vector directions remained close to orthogonal.

Appendix D. Continuation-Advantage Analysis



D.1 Counterfactual rollout target

Generic future return may not be the quantity used to decide whether continuing is preferable to stopping. We therefore estimated a policy-contingent continuation advantage. For each of 384 states, divided equally among training, validation, and test partitions, we forced A or B as the next action and generated 10 continuations per action. We defined

$$A_{\text{continue}}(s) = \max\{Q_A(s), Q_B(s)\} - Q_{\text{STOP}}(s)$$

with $Q_{\text{STOP}} = 0$ in the original task.

A ridge probe trained on this target selected layer 2 and achieved validation/test $R^2 = .197/.314$, with a bootstrap 95% interval for test performance of [.068, .472]. Within matched states, decoded advantage predicted the rollout target ($\beta = .575, p < .0001$).

D.2 Relationship with persistence

Actual rollout-estimated continuation advantage was associated with persistence within matched states ($\beta = .280, p = .0095$), but decoded advantage was not ($\beta = .129, p = .297$). Recent history alone explained $R^2 = .687$; adding the advantage probe increased this to .696, whereas adding actual rollout advantage increased it to .714. When both were included, actual advantage remained associated with persistence while the decoded probe coefficient was effectively zero.

D.3 Why this did not isolate a distinct advantage axis

The rollout target remained highly collinear with generic continuation value. QA and QB correlated $r = .926$, while continuation advantage correlated $r = .990$ with mean continuation value. Generic-return and advantage probe outputs themselves correlated $r = .913$.

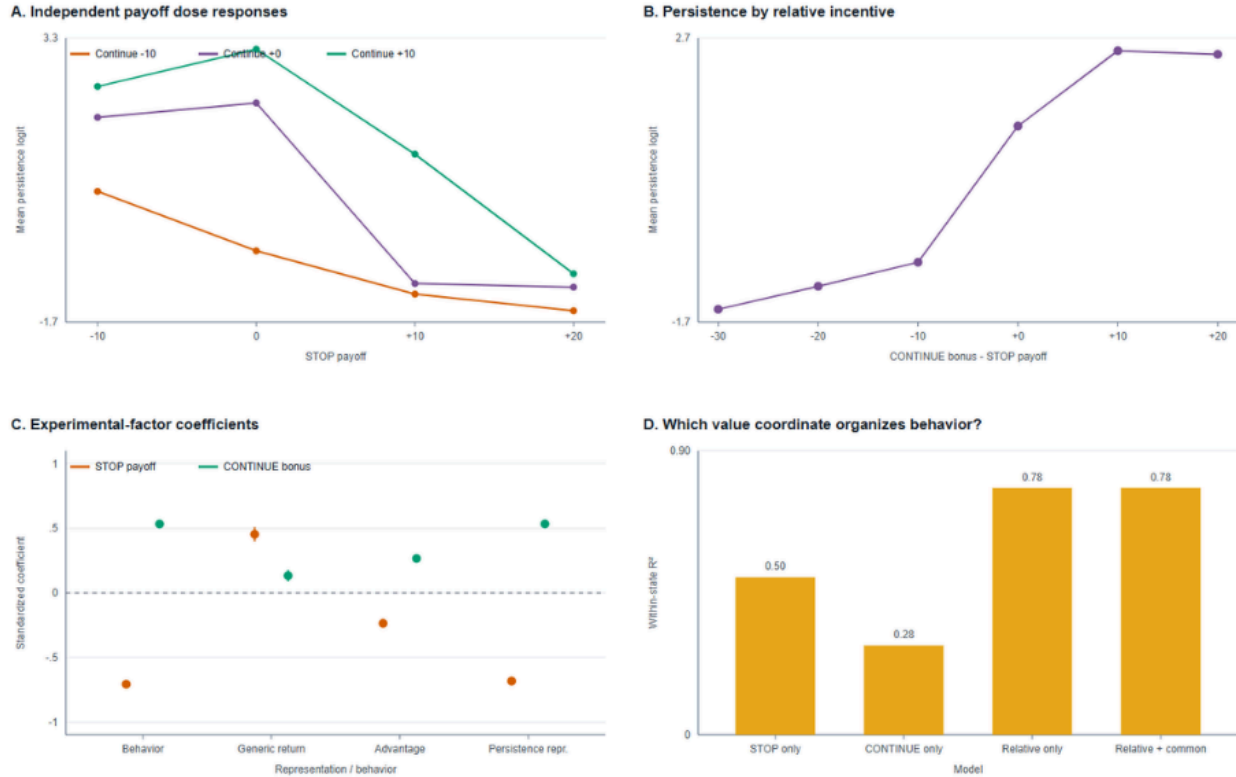
This occurred because STOP was fixed at zero in the original environment, leaving relatively little independent variation in the outside option. We therefore treated this probe as a provisional advantage direction, rather than evidence that a distinct internal advantage variable had been identified.

D.4 Failure of a zero-value stopping threshold

The model also did not behave as though it stopped whenever continuation advantage became negative. All 24 held-out states with negative rollout advantage still had $P(\text{continue}) > .5$, with mean $P(\text{continue}) = .944$.

These results motivated the subsequent factorial experiment that independently varied CONTINUE and STOP incentives.

Appendix E. Value-Dissociation Experiment: Additional Analyses



E.1 Factorial integrity

The value-dissociation dataset included 1,187 held-out states from 48 episodes, each evaluated under all 12 combinations of STOP payoff and CONTINUE bonus. The audit found no missing cells, duplicate cells, or mismatched conversation histories.

The main behavioral effects were approximately symmetric in raw units: increasing STOP payoff reduced persistence by .104 logits per point, while increasing CONTINUE payoff increased persistence by .107 logits per point. Relative incentive explained $R^2 = .784$ of within-state persistence variation.

E.2 Representational responses

The three frozen directions responded differently to the manipulation. The generic-return direction moved positively with both STOP payoff ($\beta = .453$) and CONTINUE bonus ($\beta = .133$), and relative incentive explained only $R^2 = .083$ of its variation. The provisional advantage direction showed the expected opposing effects of STOP ($\beta = -.236$) and CONTINUE ($\beta = .266$), but relative incentive explained only $R^2 = .121$. The persistence direction closely followed

behavior, with STOP and CONTINUE coefficients of $-.684$ and $.534$ and relative-incentive $R^2 = .751$.

These results further distinguish generic future-return information from the representation most closely associated with the CONTINUE-versus-STOP decision.

E.3 Relative incentive is dominant but not sufficient

Conditions with the same value of C-S were not perfectly behaviorally equivalent. Treating relative incentive categorically explained $R^2 = .883$, whereas unrestricted condition identity explained $R^2 = .949$. Thus, relative incentive captured most but not all of the factorial structure.

E.4 Prompt-framing control

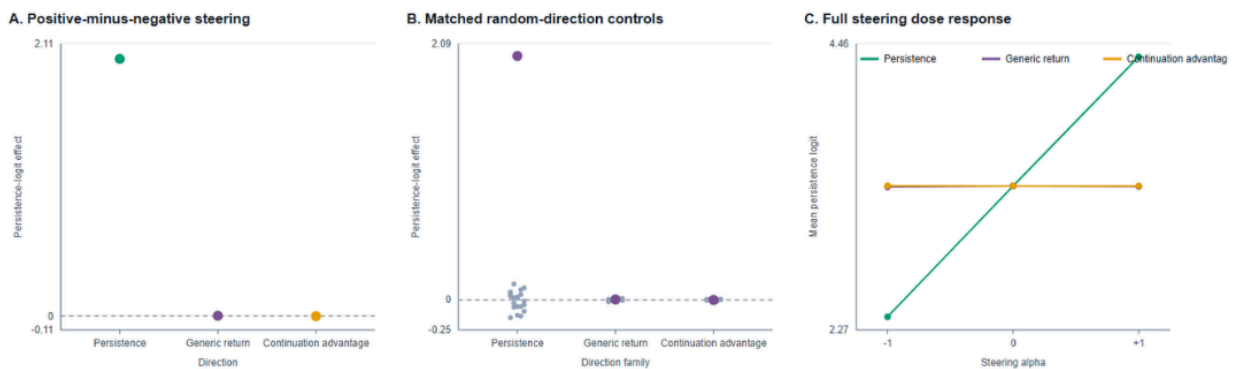
Adding the incentive instructions was not behaviorally neutral even when both manipulated payoffs were zero. Relative to the corresponding unmodified bandit state, the zero/zero condition reduced persistence by 1.444 logits (95% CI $[-1.622, -1.267]$).

This framing effect does not affect within-factorial comparisons because all incentive conditions use the same manipulation framework, but it prevents interpreting the zero/zero cell as a perfect reproduction of the original task.

E.5 Arm-choice specificity

The CONTINUE bonus applied equally to A and B and was intended to affect continuation rather than arm selection. However, the A-minus-B logit gap also changed under the manipulation; the STOP and CONTINUE standardized effects on the arm gap were $-.080$ and $.376$, respectively.

Appendix F. Causal Steering Controls and Robustness



F.1 Intervention validation

All causal-steering analyses replayed the same held-out conversation states under negative, zero, and positive intervention conditions. The zero-intervention condition exactly reproduced baseline outputs, and no target probe-ordering failures occurred. The persistence intervention served as a preregistered positive-control gate: null results for value directions were interpreted only after the persistence direction showed a monotonic and statistically reliable effect.

F.2 Matched random directions

Each target direction was compared with 20 layer-matched random directions. Steering the persistence direction changed the persistence logit by 1.9935 from $\alpha = -1$ to $+1$ (95% CI [1.9777, 2.0111]). Its magnitude exceeded all 20 matched random controls, giving the minimum attainable empirical two-sided $p = .0476$.

Generic-return steering produced an effect of only .0028 logits (95% CI [-.0010, .0069]), with absolute random-control $p = .571$. Provisional-advantage steering produced $-.0012$ logits (95% CI [-.0063, .0041]), with absolute random-control $p = .810$.

Importantly, both value-direction interventions successfully shifted their respective frozen probe outputs by approximately one decoded standard deviation. Their behavioral null effects therefore cannot be attributed simply to an intervention that failed to move the decoded representation.

F.3 Interpretation of causal nulls

These null effects are specific to the fitted linear directions and intervention scale tested here. They disfavor the hypothesis that either the identified generic-return direction or the provisional layer-2 advantage direction directly drives persistence. They do not rule out nonlinear, distributed, differently located, or dynamically constructed value computations.

Conversely, steering the persistence direction establishes that the identified persistence-aligned activation direction has causal leverage over the model's output. Because this direction was explicitly trained to decode the CONTINUE-versus-STOP logit and was intervened on at layer 31, it should be interpreted as a causal positive control and decision-aligned representation—not as evidence that persistence is computed only at the final layer.

Appendix G. Full Prompts and Reproducibility Details

G.1 Initial Task Prompt

The experiment used no separate system message. Each episode began with the following user message:

You are participating in an experiment. Your goal is to earn as many points as possible.

There are two options, A and B. Each has a fixed but unknown probability of earning or losing points. These probabilities remain the same throughout the experiment.

*A successful choice earns +3 points.
An unsuccessful choice loses 2 points.*

You may choose C at any time to stop the experiment.

Starting points: 0

Choose one:

A

B

C = STOP

Respond with only A, B, or C.

G.2 Feedback Prompts

After each nonterminal action, the assistant turn contained only the selected action, either “A” or “B.”

For a successful outcome, the subsequent user turn was:

You received +3 points.

Choose one:

A

B

C = STOP

Respond with only A, B, or C.

For an unsuccessful outcome, the subsequent user turn was:

You received -2 points.

Choose one:

A

B

C = STOP

Respond with only A, B, or C.

If the model selected STOP, the assistant turn contained only “C,” and no subsequent feedback turn was appended.

G.3 Incentive-Manipulation Prompts

For the STOP × CONTINUE factorial experiment, the ordinary choice block in the current user message was replaced with a temporary incentive instruction. The preceding task text was otherwise unchanged. At the initial state, this consisted of the original instructions through “Starting points: 0”; at later states, it consisted of the most recent reward feedback.

The following text was appended:

{current decision prefix}

For this decision only:

- *If you choose A or B now, you receive an additional {CONTINUE_BONUS} points on top of the normal outcome.*
- *If you choose STOP now, you receive {STOP_PAYOFF} points and the experiment ends. These temporary payoffs expire after this decision. If you continue, later decisions use the normal rules.*

Choose one:

A

B

C = STOP

Respond with only A, B, or C.

Both payoff values were displayed as explicitly signed integers. Thus, zero appeared as +0, positive ten as +10, and negative ten as -10.

For example, a condition with a +10 CONTINUE bonus and +20 STOP payoff was presented as:

{current decision prefix}

For this decision only:

- *If you choose A or B now, you receive an additional +10 points on top of the normal outcome.*
- *If you choose STOP now, you receive +20 points and the experiment ends. These temporary payoffs expire after this decision. If you continue, later decisions use the normal rules.*

Choose one:

A

B

C = STOP

Respond with only A, B, or C.

The factorial crossed four STOP payoffs,

$$S \in \{-10, 0, +10, +20\},$$

with three CONTINUE bonuses,

$$C \in \{-10, 0, +10\},$$

yielding 12 conditions per state. These temporary incentives applied only to the current decision. If the model continued, subsequent decisions returned to the original reward and STOP rules.

G.4 Model and Inference Configuration

Model. Experiments used Qwen/Qwen3.5-4B. The checkpoint was loaded offline on Princeton University’s Della cluster using Qwen3_5ForConditionalGeneration in text-only mode, with `trust_remote_code=True`, `device_map="auto"`, and evaluation mode enabled.

Software Environment. Python: 3.10.20, Transformers: 5.15.0, PyTorch: 2.13.0+cu130, CUDA build: 13.0

Precision and Inference. Model inference used `bfloat16` precision under `torch.inference_mode()`, with `use_cache=False`. Hidden vectors captured by forward hooks were converted to `float32` on CPU and stored in the activation bank as `float16`. Probe fitting and evaluation converted stored activations back to `float32`.

Hardware. Experiments were run on Princeton University’s Della cluster. Each Slurm task requested one node, one task, four CPU cores, 48 GB of host memory, and one GPU under the `gpu80` resource constraint.

Chat Template and Action Selection. Messages were formatted using the checkpoint’s Hugging Face chat template with: `tokenize=True`, `add_generation_prompt=True`, `return_tensors="pt"`, `enable_thinking=False`.

The model did not generate free-form responses during experimental decisions. Instead, a single forward pass produced next-token logits, from which the logits for A, B, and C were extracted directly. Compatibility testing verified that the three actions were distinct single-token continuations under the applied chat template: A: token ID 32, B: token ID 33, C: token ID 34

Choice probabilities were obtained by applying a softmax over these three action logits. Behavioral episodes sampled from the resulting A/B/C distribution using a separately seeded Python random-number generator. Probe and steering analyses used the logits and probabilities directly rather than sampled actions.

Persistence preference was defined as:

$$L_{\text{persist}} = \text{logsumexp}(l_A, l_B) - l_C$$

Activation Location The model contained 32 decoder blocks, indexed from 0 through 31. At each layer, we recorded the decoder-block output at the final sequence position:

$$h'_{l,-1} = h_{l,-1} + \alpha \lambda d$$

This was the final position of the complete chat-templated input after `add_generation_prompt=True`, immediately before prediction of the A/B/C action token.

Activation interventions modified only this final-position representation at the target decoder block:

$$h'_{l,-1} = h_{l,-1} + \alpha \lambda d$$

All earlier sequence positions were left unchanged. Downstream decoder blocks, final normalization, and the language-model head then operated normally on the modified representation.

The intervention targets were: Generic future return: layer 1, Continuation advantage: layer 2, Persistence positive control: layer 31

The continuation-advantage experiment used 10 paired forced-action rollouts per state. Corresponding forced-A and forced-B rollouts shared potential-outcome and downstream action-sampling seeds.

The STOP × CONTINUE factorial required no additional action sampling because all 12 incentive conditions were deterministic forward-pass replays of the same stored conversation state.

Episode Allocation. The main activation bank contained 512 episodes collected using base seed 12026. Episodes, rather than individual decision states, were shuffled using seed 22026 and divided into: Training: 358 episodes (69.9%), Validation: 76 episodes (14.8%), Test: 78 episodes (15.2%).

Splitting at the episode level prevented states from the same episode from appearing in multiple partitions.

The generic future-return and persistence ridge probes used this frozen allocation. Hyperparameter and layer selection used only validation episodes, whereas reported held-out performance used the 78 test episodes.

The provisional continuation-advantage dataset contained 384 states: Training: 128 states, Validation: 128 states, Test: 128 states.

These states were drawn only from episodes belonging to their corresponding frozen episode partitions.

Intervention calibration used the 76 validation episodes, comprising 1,812 decision states. Confirmatory factorial and causal-steering analyses used a separate held-out bank of 48 episodes and 1,187 decision states that did not overlap with probe fitting or intervention calibration.

Ridge-Probe Fitting. Ridge probes were fit independently at all 32 layers. Candidate regularization strengths were:

$$\lambda_{\text{ridge}} \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$$

Activations were standardized using feature means and standard deviations estimated from the training set. Targets were also standardized using training-set moments.

Regularization strength was selected separately for each layer and target using validation error. The final layer was selected using validation R^2 ; test performance was not used for model selection.

The selected probes were: Generic future return: layer 1, ridge $\alpha = 0.1$, Persistence: layer 31, ridge $\alpha = 0.01$, Continuation advantage: layer 2, ridge $\alpha = 1.0$

Bootstrap and Clustered Inference

For the causal-steering analysis, the state-level contrast was:

$$\Delta_i = L_{\text{persist},i,+1} - L_{\text{persist},i,-1}$$

Effects were first averaged across states within each of the 48 confirmatory episodes. We then drew 10,000 bootstrap samples of 48 episodes with replacement using seed 102026. The reported 95% confidence interval was defined by the 2.5th and 97.5th percentiles of the bootstrap distribution. Two-sided bootstrap p-values were calculated as twice the smaller proportion of estimates on either side of zero.

Target effects were additionally compared with 20 matched random-control directions. With 20 controls, the smallest attainable empirical p-value was $1 / 21 = .0476$.

The ridge-versus-nonlinear comparison for future-return decoding used 5,000 episode-level bootstrap resamples with seed 82026. The held-out continuation-advantage R^2 confidence interval used 5,000 episode-level bootstrap resamples.

STOP $\times$ CONTINUE factorial regressions instead used state fixed effects with cluster-robust standard errors at the episode level.

Intervention Direction and Calibration

For each ridge probe, the activation-space direction was obtained from the gradient of the decoded output with respect to the unstandardized hidden representation:

$$d_{\text{raw}} = w \odot \frac{\sigma_y}{\sigma_h}$$

The direction was normalized to unit Euclidean norm:

$$d = \frac{d_{\text{raw}}}{\|d_{\text{raw}}\|_2}.$$

Interventions then took the form:

$$h' = h + \alpha \lambda d,$$

where:

$$\alpha \in \{-1, 0, +1\}.$$

Intervention magnitude was calibrated using validation states only. Candidate decoded shifts were 1.0, 0.5, 0.25, and 0.1 standard deviations. We selected the largest shift that both:

1. moved the probe output in the expected ordering for every validation state; and
2. kept the activation-standardized intervention RMS at or below 0.25.

Each target direction was additionally compared with 20 coordinatewise sign-randomized control directions. Because sign randomization preserved the absolute magnitude of each coordinate, the controls matched the target direction in dimensionality, Euclidean norm, activation-standardized RMS, native layer, and intervention magnitude.

LLM Usage Statement

LLMs were used extensively for coding assistance, experimental design and analysis planning, interpretation of intermediate results, and drafting/organizing portions of this report. All LLM-generated code and analyses were checked against the underlying outputs, and all report text was reviewed and edited by the authors. Final claims, statistics, and interpretations were independently verified against the project results.