INTERPRETING AND STEERING LLM AGENTS FOR SOCIAL SIMULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative agents built on large language models (LLMs) have proven powerful for simulating human behavior, making them valuable additions to the social scientific toolkit. However, the black-box nature of LLMs limits their value for simulation by constraining both: (i) *interpretability*: uncovering mechanisms driving observed behavior; and (ii) *steering*: muting or amplifying mechanisms of action along theoretically meaningful dimensions. Here, we demonstrate how the black box could be opened up to further enrich LLM-based simulations. We illustrate our approach by interpreting and steering two primitive components of human behavior, namely preferences (risk attitudes, fairness preferences) and capabilities (divergent creativity, product innovation), using four classic economic and creative tasks implemented as natural-language interactions. We demonstrate the utility of sparse autoencoder (SAE)-based techniques for interpreting mechanisms and compare three types of steering methods: (1) prompt-based manipulation, (2) SAE-derived feature steering, and (3) probe-based direction steering. Overall, our results show that SAE- and probe-based techniques often outperform basic prompt-based methods for steering LLM agents, although this advantage narrows when compared with more advanced prompting strategies, such as few-shot and chain-of-thought prompting. Together, SAEs and probes constitute an effective pipeline for social scientists seeking to interpret and steer agents in social simulations: SAEs decompose agents' internal representations into human-readable features, after which probes can reliably shift agents' behaviors in specified directions.

1 INTRODUCTION

Generative agents produced by Large Language Models (LLMs) mimic humans in multiple dimensions: they understand natural languages, can communicate with humans, and have similar behavioral patterns with humans (Anthis et al., 2025). They can be configured to represent wide-ranging background profiles and respond to different experimental scenarios at a lower cost than observing the behaviors of an equivalent number of humans (Park et al., 2024). Given these advantages, social scientists have been motivated to use generative agents to conduct social simulations (Horton, 2023; Chen et al., 2023).

However, the black-box nature of generative agents has been proven to be a major roadblock for social scientific simulations (Anthis et al., 2025). Social scientists build new theories either inductively, by uncovering the mechanisms underlying observed behaviors, or deductively, by deriving hypotheses from abstract concepts and testing them through quantitative or experimental methods (Davis & Bingham, 2007; Hedström, 2005). Black-box LLMs don't allow for either of these approaches. It is difficult to identify the specific mechanisms driving behaviors for induction, and equally challenging to steer those mechanisms in a controlled manner for deduction.

Here, we offer a new set of tools and a framework to help open the black box of generative agents in simulations. We have two goals. *Interpretability*, or understanding potential inner mechanisms driving observed LLM behavior and *Steering*, or the ability to perturb mechanisms of action to study the causal effects of those mechanisms on downstream outcomes. To do so, we leverage emerging methods in mechanistic interpretability, often dubbed as "neuroscience" for generative AI models (Nanda et al., 2023), a field of computer science that has applied these techniques to promote model

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

safety and improve the AI training pipeline. Here, we adapt these techniques to social scientific simulation and demonstrate their utility in this context as well.

First, we demonstrate how sparse autoencoders (SAEs) can open up the LLM black-box to interpret model behavior. We then compare the effectiveness of prompt-based methods of control, including both persona-based one-shot, few-shot, and chain-of-thought prompting, with SAE-based feature steering and a simpler alternative, linear probes (Alain & Bengio, 2017). Both methods rely on amplifying or muting identified features representing certain mechanisms of underlying interest that are hypothesized to be driving LLM behavior.

We demonstrate the validity of the proposed methods and focus on simulating two primitive dimensions of human behavior, specifically preferences and capabilities, across a set of classic experimental tasks. For preferences, we look at the lottery game (risk) (Edwards, 1954) and the ultimatum game (fairness preferences) (Camerer & Thaler, 1995). For capabilities, we examine simple tasks capturing divergent creativity and product innovation. For each game, we use SAEs to understand what features are activated within the model with a view to uncovering mechanisms potentially driving agentic behavior. We then build probes for the central feature driving model behavior (e.g. risk-taking intention in the case of lotteries) and compare the effectiveness of using probe- and SAE-based method of steering these features (e.g. amplifying the desire for risk-taking) with traditional prompt based methods of steering.

2 Related Work

2.1 Using Generative Agents for Social Simulations

LLMs are able to generate similarities of real-human level of heterogeneity in demographics (Park et al., 2024), economic rationality (Chen et al., 2023; Xie et al., 2025), cognitive abilities (Toubia et al., 2025), and other fundamental attributes (Abdulhai et al., 2024). LLMs also carry social abilities that allow them to organize social events in artificial settings (Park et al., 2023), and can generate networks whose structures are similar to human social networks (Chang et al., 2025). The outstanding performance of LLMs has prompted researchers to replicate experiments traditionally conducted on humans using generative agents powered by LLMs (Horton, 2023; Tranchero et al., 2024). Others went further to specify target cognitive bias in generative agents for social simulations (Liu et al., 2026). Using generative agents in experiments offers several advantages, including low cost, easy accessibility, and a close resemblance to real-human behaviors. Complementing the deterministic nature of agents in traditional agent-based modeling, generative agents introduce real-world contextual nuance and variability into simulations (Gao et al., 2024). Hence, generative agents are helpful in multiple avenues: piloting untested experiments to observe possible outcomes and improving existing deductive tools, which complement the traditional data-driven approach in using machine learning to social scientific research (Grimmer et al., 2021).

Despite their many benefits, generative agents behave like a black box - we cannot clearly interpret why they act as they do, nor reliably steer them toward specific mechanisms (Jackson et al., 2025; Ludwig & Mullainathan, 2024). The interpretability problem arises because LLMs are trained on massive, weakly documented corpora, making it impossible to know whose voices and assumptions are represented; this turns them into "stochastic parrots" with little accountability for output quality (Bender et al., 2021). Even when post-hoc explanations are available, they rarely reflect the model's true reasoning and can mislead downstream analyses (Rudin, 2019). The controllability problem follows from this opacity: the same observed behavior can emerge from entirely different internal mechanisms, meaning researchers cannot isolate or manipulate the causal processes they wish to study. As a result, generative agents risk obscuring rather than clarifying the mechanisms behind theories (Gao et al., 2025). However, if we can open this black box, we gain the ability to both understand the mechanisms driving the agents' behaviors and intentionally modify them, expanding the credible uses of generative agents for the theory building in social science.

2.2 Opening the Black Box

Given the way LLMs are constructed, one approach to opening the black box is to investigate the interpretability of the models (Alain & Bengio, 2017; Elhage et al., 2021). Because a given model's decisions are mediated by internal activations before they are converted into output logits, these

activations provide a natural site for interpretation and intervention. Interpretability methods such as probes and sparse autoencoders can therefore be applied at selected layers to analyze or modify these internal representations before the model produces its next-token prediction. Probing, which is typically trained on labeled datasets for predefined concepts, is useful for revealing what information can be decoded from a neural network's hidden representations (Alain & Bengio, 2017; Hewitt & Manning, 2019; Belinkov, 2022). When applied to the behavioral reasoning of generative agents, probes can serve as measures of specific behavioral dimensions, such as risk, by assessing the extent to which agents' internal representations encode that characteristic. For example, if models can be clearly taught the difference between safe and risky choices in a narrow setting, then we might be able to train probes that can be used to steer this behavior.

SAEs takes a different approach. They open the black box by reverse engineering the computations within Transformer layers to understand which activations contribute to particular behaviors (Elhage et al., 2021; Nanda, 2023). SAEs are the tools to achieve mechanistic interpretability, as they are small neural networks trained to reconstruct a model's internal activations using a sparse latent code (Bricken et al., 2023). Because of this sparsity, each latent dimension captures a relatively distinct pattern, allowing various dimensions to function as human-interpretable "features." In the context of generative-agent behavior, SAEs can identify top-activating features associated with corresponding behavioral patterns. This provides a more transparent basis for interpreting and steering on simulations using generative agents.

3 METHODOLOGY

We begin with the lottery (Edwards, 1954) and the ultimatum game (Camerer & Thaler, 1995) paradigms, which are widely used in psychology to study risk attitudes and fairness preferences as core behavioral tendencies in one's preferences. The lottery compares a guaranteed 50 tokens with a 50% chance of a variable reward. The ultimatum game asks the agent to accept or reject a share of 100 tokens. Similarly, we find divergent creativity and product innovation as important components of capabilities and use tasks similar to Torrance Tests of Creative Thinking (Torrance, 1966) to see the generative agents' performances. Divergent creativity tasks ask for alternative uses of a brick, and product innovation tasks ask for improvements to a stapler. These individual-agent tasks operationalize preferences and capabilities. Appendix A specifies the task families and sample sizes.

3.1 PROMPTING SETTING

To establish whether interpretability-based interventions (SAE and probe steering) provide leverage beyond what is achievable through prompt engineering alone, we compare against a graded ladder of prompting baselines applied to the same Llama-3.3-70B-Instruct: baseline, persona, chain-of-thought (COT), few-shot, and few-shot CoT. Each keeps the original task prompts and instructions. We use the examplars from human-authored uses selected from Ocsai Alternative Uses Task corpus (Organisciak et al., 2023). This publicly released dataset provides brick uses paired with human-rated originality scores, which are independent from the downstream judge. Decision demonstrations use a offer grid disjoint from the test sweep. Appendix C provides the more details of the specifications and the full prompting protocol.

3.2 SPARSE AUTOENCODERS SETTING

We use Goodfire's publicly released layer-50 SAE for Llama-3.3-70B-Instruct (McGrath et al., 2024a), together with its supplied human-readable labels (McGrath et al., 2024b). The dictionary has 65,536 features and the harmful ones are excluded at release. Figure 10 illustrates the full pipeline of how interpretation and steering are achieved with this method.

The experiments run inference and steering locally with Transformers and open-source EDSL (Horton et al., 2024), which is an open-source app for simulations with generative agents. We adopt two different ways of steering. Label-selected edits increase lottery features 184 and 4237, or ultimatum feature 31935. Or we start from the baseline experiment, get the top ranked features from the baseline scenario, and steer these features with either equal weights or scale them based on the activation strength. Equal-weight bundles normalize individual decoder directions and match the combined intervention norm within task. Three presampled random bundles at each size serve as controls.

Table 1: Comparative Overview of Interpretability Methods for Generative Agent Steering

Method	Strengths	Limits	Useful role
Prompting	Accessible; no training required	Prompt-sensitive; explanations need not expose mechanisms	Rapid behavioral experiments
SAEs	Labeled internal features; direct edits	Label ambiguity; activation rank is not importance	Candidate interpretation and intervention
Probes	Task-specific readout and calibrated direction	Label dependence; calibration leakage	Control of operational behavioral targets

Activation-weighted bundles are a separate sensitivity analysis. Discovery and development fix feature sets and doses before confirmation. Appendix D provides exact coefficients, token scopes, locks, and transfer designs; the earlier hosted creativity protocol is specified in Appendix F. Our main capability comparison isolates the recorded steering and unsteered baseline conditions.

3.3 Probes Setting

Figure 15 illustrates our probe-based pipeline for scoring and steering. We train a logistic regression probe on intermediate layer activations (layer 48 of Llama-3.3-70B-Instruct) (Alain & Bengio, 2017; Conneau et al., 2018). The learned weight vector $\mathbf{w}$ and bias b define a score $\mathbf{w}^\top \mathbf{h} + b$; the unit-normalized direction is used for activation addition, $\mathbf{h}' = \mathbf{h} + \lambda \hat{\mathbf{w}}$, with the implementation's layer scaling and token window detailed in Appendix E. Preference labels are realized risky/safe or accept/reject choices. Switching points are calibration targets, not training labels. Creativity labels split judge scores into high and low classes.

Layer 48 was selected by trial-level cross-validation. The reported 82% accuracy is not a prompt-held-out estimate because trials can share an activation. We reconstruct decision crossings from raw choices rather than target-centered display curves. Calibration reuses 12 of 40 evaluation seeds, and the saved sweeps lack a zero-dose baseline; calibrated control should not be read as fully held-out prediction. We also validate the results with a probe in the Layer 50 to directly compare with SAEs. Raw-choice calibration and layer/model checks are in the Appendix E, including Figure 16.

Evaluation. Based on the different characteristics of the interpretability methods, we provide Table 1 to summarize the strengths, limits, and useful roles for each method.

4 Results

We evaluate how different tools perform on preference and capability tasks in terms of their interpretability and controllability over agents' behaviors. First, we showcase the top-ranked SAE features decomposed by the agents' responses in different settings to demonstrate the interpretability of their behaviors. With the results from all the scenarios, inspecting the SAE features of generative agents allows us to peek into the behavioral codes of generative agents. Second, we showcase how prompting, SAE steering, and probe steering differ in their performances on the controllability of agents' behaviors. We provide different degrees of control in each approach as separate scenarios and show the agility in the agents' responses. For each scenario, we repeat the same steering interventions on 40 agents. In preference tasks, we record the percentage of agents changing their behaviors as reward values increase in the lottery and the ultimatum games; and in capability tasks, we record the average creativity scores with standard deviation in divergent creativity and product innovation tasks.

4.1 Interpretability

4.1.1 Top-ranked SAE Features

Figure 1 ranks features by mean activation over generated-response tokens, first averaging within each response and then equally across responses. Probability and numerical comparisons (13951) lead the lottery rank; economic payoffs (50689) lead ultimatum. Creative alternatives (23007) and ergonomic

design (36804) lead brick and stapler discovery, respectively. The labels are reasonable descriptions of task content, but high activation may reflect wording, formatting, or response construction.

The decision tasks share some highly active features but differ in their ordering. Numerical comparison (13951) is first in lottery and second in ultimatum, while economic tradeoffs (50689) rise from seventh in lottery to first in ultimatum. Analytical explanation (25058) appears among the top three in both. The full ultimatum ranking also includes fairness and equal treatment (5546), cooperation versus competition (22551), and negotiation language (29461, 48232). This combination is consistent with responses containing both payoff calculations and discussion of the social allocation. It suggests more specific hypotheses to inspect than a single global description such as "an fair agent." The scan alone cannot determine which of these patterns produces acceptance or rejection.

The creativity rankings similarly distinguish the content of the two tasks. Brick responses activate descriptions of novel uses, decorative objects, and construction materials; stapler responses activate ergonomics, fasteners, and interacting mechanical components. These are recognizable aspects of the requested activity. Some features are also less straightforward: the second-ranked brick feature carries a supplied Russian-word description even though the task prompt is in English. Such cases make the interpretation inspectable but also expose where a released label needs response-level checking. Neither the label nor its rank should be silently recast as a general creativity trait.

Two practical patterns of this behavioral interpretability include theoretical validation and generation. For example, if one hypothesizes the creativity related to innovation can be extended from considering the physical restraints of the current products, the top 1 or 2 SAE features activated for agents in the product innovation tasks confirm this hypothesis. However, while the SAE features offer some extent of validation, we need to carefully evaluate their activation strength and underlying logic. Theoretical generation is also possible: we can use the top-ranked activated features to brainstorm mechanisms that may underlie agentic behaviors, even if those mechanisms are not yet theoretically grounded in established social-scientific behavioral patterns. For example, "marketing language for writing instruments" is not an usual behavioral code for a stapler innovation task. This suggests that creativity in product innovation may stem from connecting the focal product to others that appear in the same context. With this example, we demonstrate that investigating SAE features is beneficial for inductive theorizing.

These patterns are useful for generating hypotheses about the mechanisms supporting agentic behaviors. We emphasise that top-activating features are correlational evidence about which concepts are active in a given condition, not established causal mechanisms. Appendix B gives all top ten features, full labels, examples, and post-steering scans.

4.2 Controllability

4.2.1 Lottery and Ultimatum

After characterizing the interpretability of reasoning that underlies agents' behaviors, we next examine whether, and to what extent, these behaviors can be steered. Figure 2 compares prompting and label-selected local SAE choice curves. The local interventions increase target choices in both games. Because protocols differ, the comparison supports within-study control rather than a matched ranking of methods. Probe calibration and layer checks are reported as robustness analyses in Appendix E, Figure 16.

What changes under prompting. The prompting curves show that both the type of instruction and the examples matter. In lottery, the prompt baseline makes no risky choices at a reward of 100 tokens and makes 35/40 at 105; persona prompting moves to 40/40 already at 100. Few-shot CoT is strongly sensitive to the demonstrated answers: at a reward of 110, the 0% target-example condition yields 1/40 risky choices, compared with 40/40 under 100% target examples. In ultimatum, the analogous comparison at an offer of 30 is 1/40 versus 40/40 acceptances. Here 0% means that all six demonstrations choose the opposite action, not that demonstrations are absent. These contrasts show substantial input-based controllability, but they also show why "prompting" is not one uniform intervention. A researcher must specify the persona, reasoning instruction, and demonstration composition to interpret its effect.

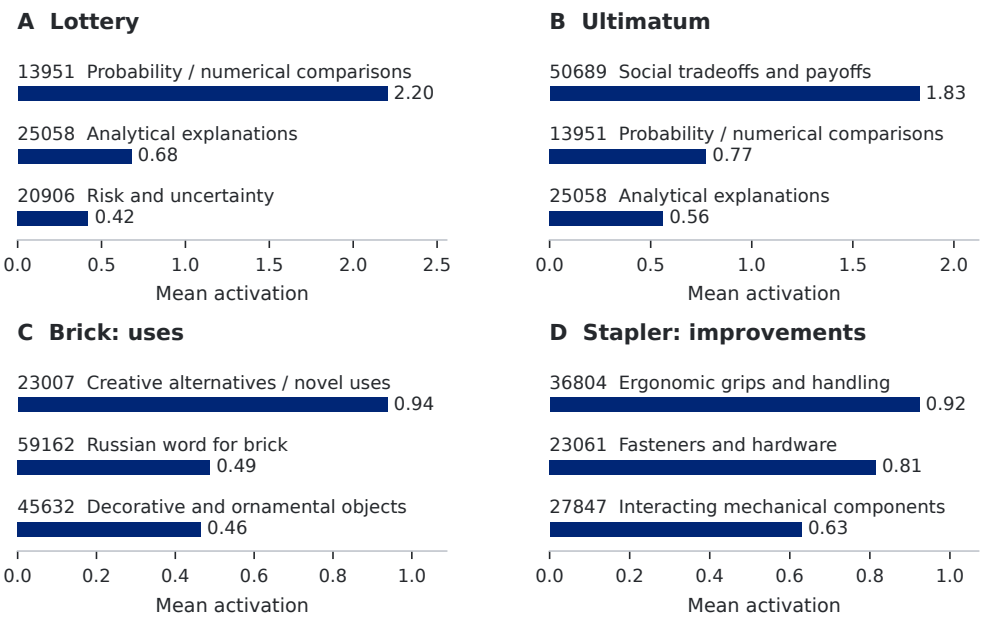


Figure 1: Top three released features by mean response activation.

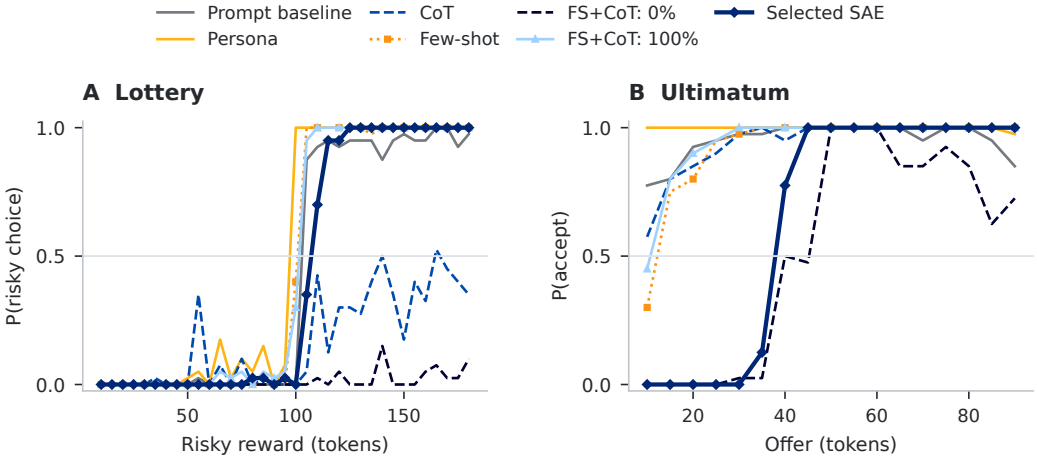


Figure 2: Prompting and label-selected local SAE steering in lottery (A) and ultimatum (B), with 40 responses per setting. The matched SAE baseline is shown in Appendix D; the prompting baseline is not its comparator. Prompts, backend and temperature differ. FS+CoT percentages denote target-choice demonstrations. Probe calibration appears in Appendix E, Figure 16.

What changes under selected SAE steering. In the full-grid local SAE confirmation, the selected label edits increase risky choices from 518/1,400 to 601/1,400 (+5.93 percentage points) and ultimatum acceptance from 412/680 to 436/680 (+3.53 points). These results demonstrate behavioral steering by the selected features. We report more specificity tests and complete control comparisons in Appendix D. The aggregate effects conceal larger changes around the local baseline's switching region. At a lottery reward of 110 tokens, risky choices increase from 1/40 to 28/40; at an ultimatum offer of 40, acceptances increase from 14/40 to 31/40. Far from the switching region, responses are often already near zero or one, leaving little room for a change.

Figure 3 brings together label-selected and activation-ranked steering. Label-selected steering increase lottery choices from 98/280 to 148/280 and ultimatum acceptance from 94/280 to 123/280

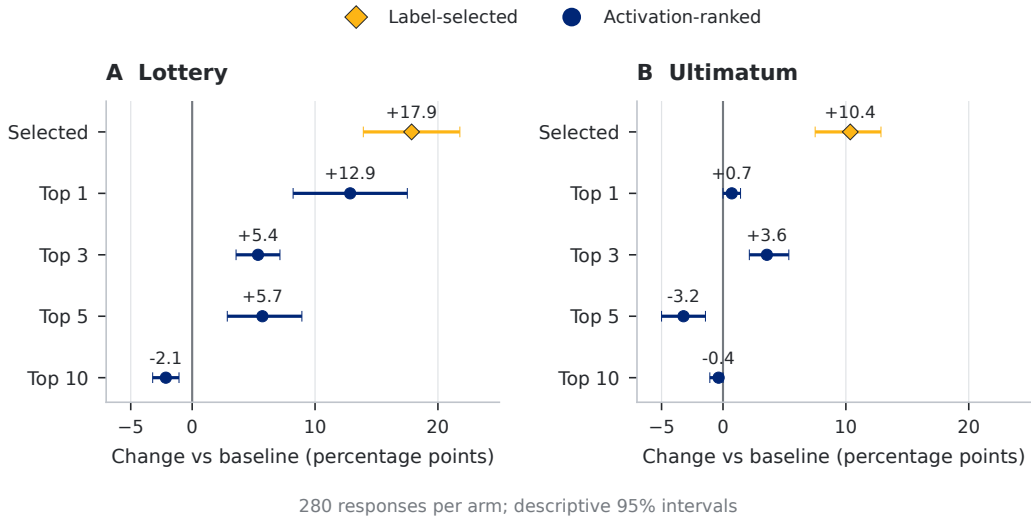


Figure 3: Behavioral shifts under label-selected and top-1/3/5/10 SAE steering (280 responses per arm).

(Holm-adjusted $p = 0.0224$ and 0.0330). Activation-ranked steering also produces observed shifts: the top-one lottery intervention increases risky choices from 98/280 to 134/280 ($+12.86$ percentage points), and the top-three ultimatum intervention increases acceptance from 94/280 to 104/280 ($+3.57$ points). Lottery top-three and top-five bundles increase risky choice by 5.36 and 5.71 points.

What activation ranking adds. The effect depends on how the bundle of features are selected: lottery top-ten and ultimatum top-five/top-ten steering shift the agents' choices in the opposite direction. We show all four sizes in the Figure 3. Full controls, weighting comparisons, and transfer results are discussed in Appendix D. This is not a gradual process in which adding more features simply increases the amount of the same intervention. Bundle size changes the direction of the edit while the intended combined norm is held fixed. Adding features redistributes the intervention across different represented patterns. The observed non-monotonicity therefore shows that baseline prominence is insufficient to choose a reliably useful bundle. The top-one lottery result is a promising candidate, but its adjusted $p = 0.0594$ does not meet the prespecified threshold. The larger label-selected effect has stronger support in this experiment. The two selection strategies answer different practical questions: label selection tests a researcher's substantive hypothesis, whereas activation ranking tests directions made salient by the task itself.

4.2.2 Divergent Creativity and Product Innovation

Figure 4 distinguishes two forms of capability control: a selected SAE steering evaluated against its baseline, and task-specific probes evaluated across requested score levels. We focus on steering minus baseline so that the comparison has one explicit intervention.

SAE steering changes a profile of rated capabilities. Panels A and B show the four evaluation dimensions and their equally weighted composite. Fluency measures the production of relevant ideas, flexibility measures their categorical variety, originality measures their novelty, and elaboration measures their development in detail. Each response is rated by five LLM judges, whose scores are averaged within response before estimating the condition contrast. The edit jointly boosts three label-selected features for creative expression (13142), unconventional thinking (20117), and professional innovation (4992), each at a dose of 0.3.

For stapler improvements, the composite rises from 5.834 to 6.284, a change of $+0.450$ points (descriptive 95% interval $[0.283, 0.615]$). The largest component change is originality, from 4.535 to 5.665 ($+1.130$, interval $[0.865, 1.390]$). Fluency and flexibility also increase by 0.245 and 0.395 points, while elaboration changes little ($+0.030$). The favorable composite shift is thus primarily

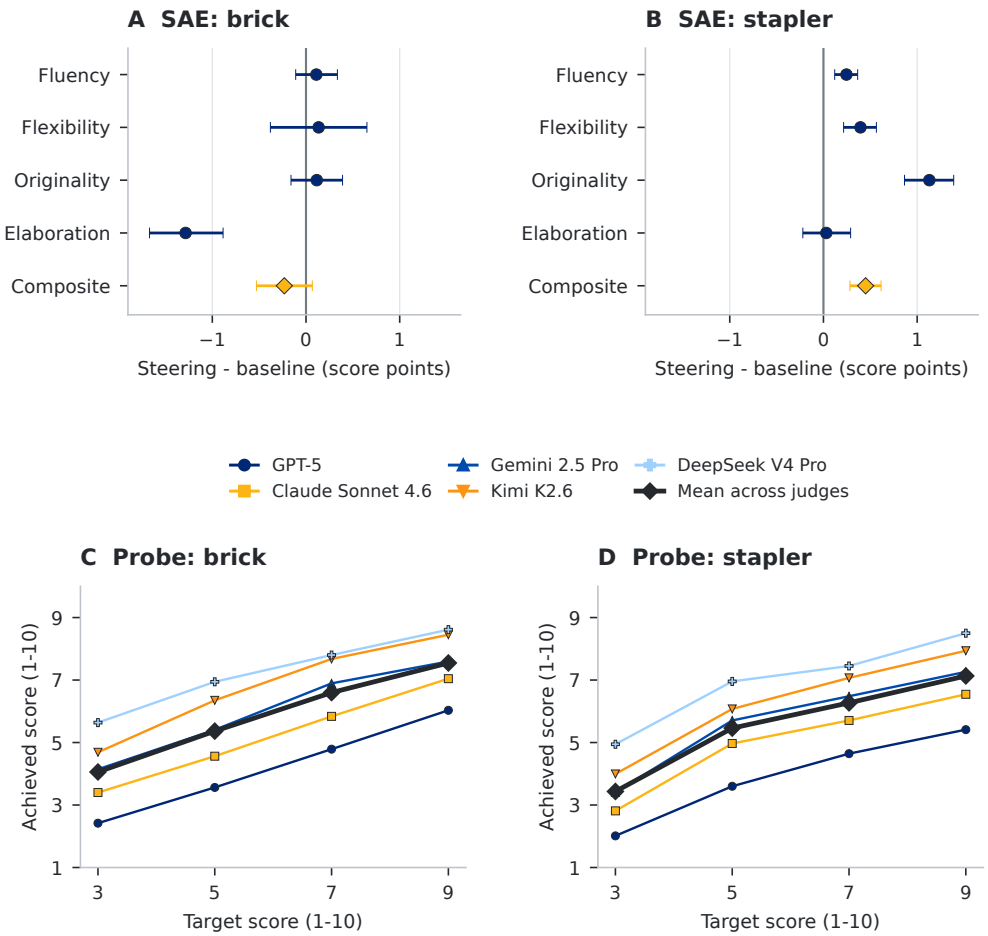


Figure 4: Capability control evaluated by five LLM judges.

associated with more novel and varied proposed improvements, as judged under this rubric, rather than uniformly higher scores on every dimension. This supports a task-specific capability-steering result for the selected edit.

Brick alternative uses show a different profile. Originality changes only from 2.765 to 2.880 ($+0.115$, interval $[-0.160, 0.390]$), while elaboration falls from 5.595 to 4.310 (-1.285, interval $[-1.670, -0.885]$). Fluency and flexibility change by $+0.110$ and $+0.135$, with intervals that also include zero. Their small positive shifts do not offset the decrease in elaboration: the composite changes by -0.231 points, with an interval spanning zero ($[-0.529, 0.068]$). The data therefore show why steering toward broadly creative labels can produce a task-dependent change in the output profile.

Probes produce ordered changes across judges. Panels C and D show in-distribution control for the two task-specific probes. All five judges' mean scores increase at every successive target for both tasks. The ensemble moves from 4.058 to 7.546 for brick and from 3.434 to 7.131 for stapler as the requested score increases from 3 to 9. This agreement in ordering supports controllability of rated output quality across the evaluated target range, rather than an effect visible to just one judge.

The judges nevertheless disagree about absolute quality. At the brick target of 9, their means range from 6.031 to 8.613; at the stapler target of 9, they range from 5.413 to 8.500. The ensemble also compresses the requested range: a six-point target increase yields about 3.49 points for brick and 3.70 for stapler. The probes therefore provide ordered control under this scoring procedure, with

imperfect attainment of numerical targets. We also offer object-transfer results in the Appendix E to show that this is both feasible and effective.

Human validation of LLM evaluation of creativity. We validate the LLM evaluation of creativity with both recruited human annotators on Prolific and the original Ocsai dataset for the originality scores. Human ratings of 244 outputs correlate with the five-judge ensemble at Spearman $\rho = 0.606$; agreement across 17 source/condition means is 0.836. On 300 human-authored Ocsai brick uses, LLM evaluated originality scores correlate with human originality scores at 0.426. These checks support qualified use of LLM evaluation for broad comparisons.

The two levels of agreement answer different questions. Item-level agreement asks whether judges and people rank individual outputs similarly. Agreement between condition means asks whether broader groups of outputs receive similar relative evaluations. Stronger aggregate agreement is useful for comparing experimental conditions, but can coexist with weak discrimination among outputs from the same source. The low within-SAE composite correlation is especially relevant to Figure 4: its positive stapler contrast is an LLM-rated result, not an independently human-confirmed gain. We provide more details in the Appendix F.4, which reports the full agreement table, scatterplots, and length checks.

5 Discussion

Generative agents powered by LLMs give computational social scientists a scalable, low-cost way to simulate behavior, but their usefulness is limited by the black-box nature of LLM reasoning. We show that SAEs expose candidate descriptions of internal activity and that representation-level interventions can change observed choices and rated outputs. Label-selected SAE edits increase target choices in both decision games, and selected activation-ranked bundles also yield promising steering. The ranked effects depend on bundle size and task and remain exploratory under adjusted testing. The practical comparison in Table 1 depends on the research question: prompting is accessible, SAE labels suggest hypotheses, and probes provide calibrated control over operational targets.

Limits of interpretation and control. As SAE features are not perfectly disentangled, boosting one feature too much may activate other off-target representations, causing behavioral shifts that are hard to analyze. Steering too much can produce nonsensical output. The specificity analyses in Appendix D distinguish behavioral steering from validation of a psychological label. The experiments support causal effects of the implemented steering within their protocols, but do not identify risk preference, fairness, or creativity as the unique mediator.

Cross-method differences in prompts, temperatures, backends, and calibration prevent a clean superiority claim. Probe trial-level validation and seed reuse limit generalization. Human ratings cover older outputs, with modest reliability and weak agreement in some source families. New local creativity ratings are preliminary, while an instruction to ignore length in older judging is not experimental length control. Synthetic anchor gates and transfer failures further limit claims about mechanism (Appendices F–G).

Scope for social simulation. We limit our focus to individuals in this paper. While individuals are the most fundamental component of society, many experiments do not happen on the individual level. This leads to a natural thread of future work in multi-agent settings, where generative agents will interact with others in structured social environments, allowing emerging group dynamics to be studied. Such extensions would explore organizational structure, social influence, and network effects, which are only possible to be accessed in multi-agent settings. This is crucial for advancing generative agent simulations towards a richer and more ecologically valid theorizing process. Our experiments also focus on a small set of traits developed from preference and capability tasks. They map directly to behavioral constructs. However, many social scientific mechanisms are more abstract, context-dependent, or embedded within complex structures. Future research in using SAE features to capture higher-order mechanisms need to examine the mechanistic support carefully in order to produce robust theorizing results.

AI USE STATEMENT

We used Large Language Models to instantiate the generative agents and generate their responses in the experiments. We also adopted LLM-as-a-judge method to evaluate the creativity output, as described in the methodology. LLM-based assistants also supported code development and debugging. LLMs were not used for research ideation. The authors reviewed all AI-assisted work and take full responsibility for the study's methods, analyses, interpretation of results, and final manuscript.

ETHICS STATEMENT

The human-evaluation study was determined exempt by the authors' institutional review board. Details of participant recruitment, consent procedures, compensation, and the scope of the human evaluation appear in Appendix F.3.

REFERENCES

Marwa Abdulhai, Gregory Serapio-García, Clément Crépy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

Jacy Reese Anthis, Ryan C. Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James A. Evans, Erik Brynjolfsson, and Michael S. Bernstein. LLM social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Technical report, Anthropic, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Colin Camerer and Richard H. Thaler. Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives*, 9(2):209–219, 1995.

Donald T. Campbell. Blind variation and selective retention in creative thought. *Psychological Review*, 67(6):380–400, 1960.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. LLMs generate structurally realistic social networks but overestimate political homophily. *arXiv preprint arXiv:2408.16629*, 2025.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120:e2316205120, 2023.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2126–2136, 2018.

Jason P. Davis and Christopher B. Bingham. Developing theory through simulation methods. *Academy of Management Review*, 32(2):480–499, 2007.

Ward Edwards. The theory of decision making. *Psychological Bulletin*, 51(4):380–417, 1954.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Technical report, Anthropic, 2021.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11:1259, 2024.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24:395–419, 2021.

Peter Hedström. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge University Press, Cambridge, 2005.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, 2019.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper 31122, National Bureau of Economic Research, 2023.

John J. Horton, Apostolos Filippas, and Rose Horton. EDSL: Expected parrot domain specific language for AI powered social science. Whitepaper, Expected Parrot, 2024.

Matthew O. Jackson, Qiaozhu Mei, Stephanie Wang, Yutong Xie, Walter Yuan, Seth G. Benzell, Erik Brynjolfsson, Colin F. Camerer, James A. Evans, Brian Jabarian, Jon Kleinberg, Juanjuan Meng, Sendhil Mullainathan, Asuman E. Ozdaglar, Thomas Pfeiffer, Moshe Tennenholtz, Robb Willer, Diyi Yang, and Teng Ye. AI behavioral science. *arXiv preprint arXiv:2509.13323*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Xuan Liu, Haoyang Shang, and Haojian Jin. CoBRA: Programming cognitive bias in social agents using classic social science experiments. *arXiv preprint arXiv:2509.13588*, 2026. URL `https://arxiv.org/abs/2509.13588`.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024.

Thomas McGrath, Daniel Balsam, Myra Deng, and Eric Ho. Understanding and steering Llama 3 with sparse autoencoders. Research report, Goodfire, 2024a.

Thomas McGrath, Daniel Balsam, Liv Gorton, Murat Cubuktepe, Myra Deng, Nam Nguyen, Akshaj Jain, Thariq Shihipar, and Eric Ho. Mapping the latent space of Llama 3.3 70B. Research report, Goodfire, 2024b.

Neel Nanda. Attribution patching: Activation patching at industrial scale. Research report, Transformer Circuits, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356, 2023. doi: 10.1016/j.tsc.2023.101356.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

E. Paul Torrance. *Torrance Tests of Creative Thinking: Norms-Technical Manual*. Personnel Press, Princeton, NJ, 1966.

Olivier Toubia, George (Zhida) Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 2025. Forthcoming.

Matteo Tranchero, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj. Theorizing with large language models. NBER Working Paper 33033, National Bureau of Economic Research, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Yutong Xie, Qiaozhu Mei, Walter Yuan, and Matthew O. Jackson. Using large language models to categorize strategic situations and decipher motivations behind human behaviors. *Proceedings of the National Academy of Sciences*, 122:e2512075122, 2025.

A Tasks and Experimental Protocols

The experiments share four task families, but are separate studies with separate baselines. Counts below refer to model generations, not distinct human participants. Repeated agent IDs index decoding instances.

A.1 Task Definitions

The lottery offers a guaranteed 50 tokens versus a 50% chance of a variable reward and a 50% chance of zero. The risky option and safe option have equal expected value at a reward of 100. The ultimatum responder chooses whether to accept an offered share of 100 tokens, with rejection giving both players zero. Equal division is 50. These tasks do not uniquely identify psychological motives from a single response.

Divergent creativity requests alternative uses of an object; product innovation requests improvements to an object. The historical high-temperature condition is a sampling manipulation motivated by the possible relationship between variation and creativity (Campbell, 1960). It does not isolate a specific trait. The new local experiment discovers features on brick and stapler and transfers fixed bundles to bottle/towel and umbrella/backpack respectively.

A.2 Study and Split Ledger

Table 2: New local SAE generation counts. Discovery/development precede frozen confirmation; analysis subsets are not independent new samples.

Study	Development / discovery	Confirmation	Other	Total
Label-selected	1,792	12,480	5,712	19,984
Ranked decision bundles	5,824	10,080	3,920	19,824
Creativity bundles	440	408	544	1,392

For label-selected steering, the other 5,712 responses comprise 1,344 label screens, 2,800 fresh seed/agent-ID responses, 1,120 paraphrase responses, and 448 cross-game responses. For decision bundles, the other 3,920 comprise 2,240 transfer and 1,680 weighting responses. Creativity's 440 development/discovery outputs comprise 32 discovery responses, 16 descriptions, and 392 dose-development responses. The 408 confirmation and 544 transfer generations provide 952 outputs for attempted judging.

The baseline feature scans use 1,400 lottery and 680 ultimatum responses from the unsteered full-grid study, and 16 discovery responses per creativity task. The seven-setting decision bundle confirmation uses 280 responses per arm and game. It is distinct from the full-grid 35-setting lottery and 17-setting ultimatum confirmation, which uses 40 responses per setting and arm.

A.3 Outcome Definitions and Comparability

Decision summaries report the risky-choice or acceptance fraction, or the first interpolated 50% crossing of a raw-choice curve. Changes in percentage points are averaged over the stated design grid. They are not individual-level preference parameter estimates. Creativity's new local 1–5 dimensions (originality, feasibility, breadth, coherence) differ from the historical 1–10 Torrance composite; scores from these instruments must not be pooled.

B Interpretability and Feature Rankings

B.1 Historical Activation Grids

The following explanations and figures are retained from the historical experiment. Its feature ranking by occurrence across top-activated lists differs from the mean-response-activation statistic used in main Figure 1. Descriptions of task alignment are hypotheses, not independent tests of psychological mediation. The historical hosted configuration is not the new local intervention.

B.1.1 Preference Tasks

Figure 5 in the appendix details the top 10 features activated for generative agents in preference tasks. We observe that most of the top 3 features show no variation in their minimum and maximum ranks. For the lottery game, the top 3 features are consistent across the baseline and prompting scenarios. "Probability-based decision making scenarios with explicit numerical comparisons" and "The assistant should select between provided options" occupy top 1 and 2 features for all agents in both scenarios; and "Technical setup and configuration states in experimental procedures" takes the 3rd place for most agents with few in the 4th place. High rankings of these features are consistent with probability-based descriptions of the lottery task. They do not by themselves replicate a social theory or behavioral mechanism. However, the rest of the features are less common. For example, "The assistant is explaining why it cannot fulfill an unethical request" is a feature that provides less intuition in making decisions in the lottery games. Inspecting features like this can potentially inspire social scientists to add another layer of analysis beyond rational choices and probabilistic decision-making, which is to consider ethics in the lottery games.

For the ultimatum game, the top 3 features show a similar lack of variation in ranks across baseline and prompting scenarios and suggest candidate descriptions of agents' behaviors. Other lower-ranked features, such as "Descriptions of experimental procedures and participant tasks in research studies," do not have a direct link with the existing theories about the ultimatum game but rather with the methodology we implement on the agents. This suggests that SAE features capture both task content and general patterns in the responses.

Identifying SAE features also helps investigate the effects of controllability. In the steering scenario for both games, the top-ranked feature becomes the feature we manually steered, despite not being in the top-ranked feature list for the baseline scenario. The rest of the features are still similar to the ones in the baseline and prompting scenarios. This is consistent with the intended effect of feature steering, in that the clamped feature becomes the most active one. It does not by itself establish that the feature mediates the behavioral change, which would require ablating or clamping it against a matched control. We can also observe that the top-ranked features for the prompting scenario are more similar to the features for the baseline scenario rather than steering scenario.

B.1.2 Capability Tasks

Figure 6 in the appendix compares the top five activated features within generative agents across two task types, divergent creativity and product innovation, under four conditions: baseline, high temperature, prompting, and steering. In the baseline setting, the model activates reasonable but relatively generic features, such as brainstorming lists or describing technical components. Increasing temperature introduces noise, with features tied to messy formatting, multilingual fragments, and irregular text, indicating exploration without control. Prompting makes behavior more focused, aligning activations more closely with relevant concepts like creative alternatives and technical relationships, though still indirectly. By contrast, steering produces the strongest and most targeted activations, concentrating on concepts such as professional innovation, unconventional thinking, and structured feature descriptions. Together, these patterns show that feature activations decomposed by SAEs are useful for generating hypotheses about the mechanisms supporting agentic behaviors. We emphasise that top-activating features are correlational evidence about which concepts are active in a given condition, not established causal mechanisms. Instruction-tuned models also carry refusal and safety latents that fire on gambling and payout language, so a feature that activates in a risk vignette need not be the feature that drives the choice. Confirming mediation requires targeted ablation or clamping against a matched control, which we do not perform here.

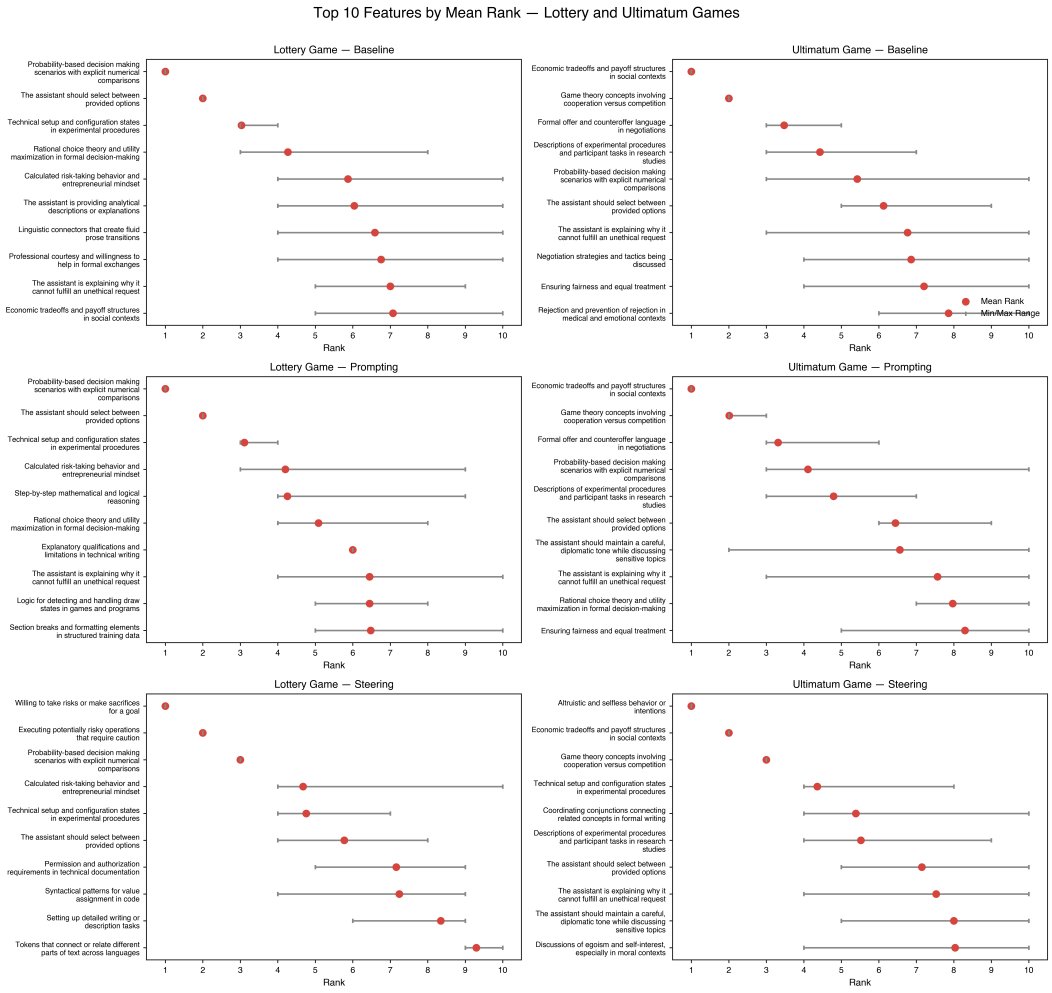


Figure 5: Feature Activations of Generative Agents on Preference Tasks

Two practical implications of this behavioral interpretability include theoretical comparison and hypothesis generation. For example, if one hypothesizes the creativity related to innovation can be extended from considering the physical restraints of the current products, the top 1 or 2 SAE features activated for agents in the product innovation tasks are consistent with this hypothesis. However, while the SAE features offer candidate interpretations, we need to carefully evaluate their activation strength and underlying logic. Theoretical generation is also possible: we can use the top-ranked activated features to brainstorm mechanisms that may underlie agentic behaviors, even if those mechanisms are not yet theoretically grounded in established social-scientific behavioral patterns. For example, "marketing language for writing instruments" is not an usual behavioral code for a stapler innovation task. This suggests that creativity in product innovation may stem from connecting the focal product to others that appear in the same context. With this example, we demonstrate that investigating SAE features is beneficial for inductive theorizing.

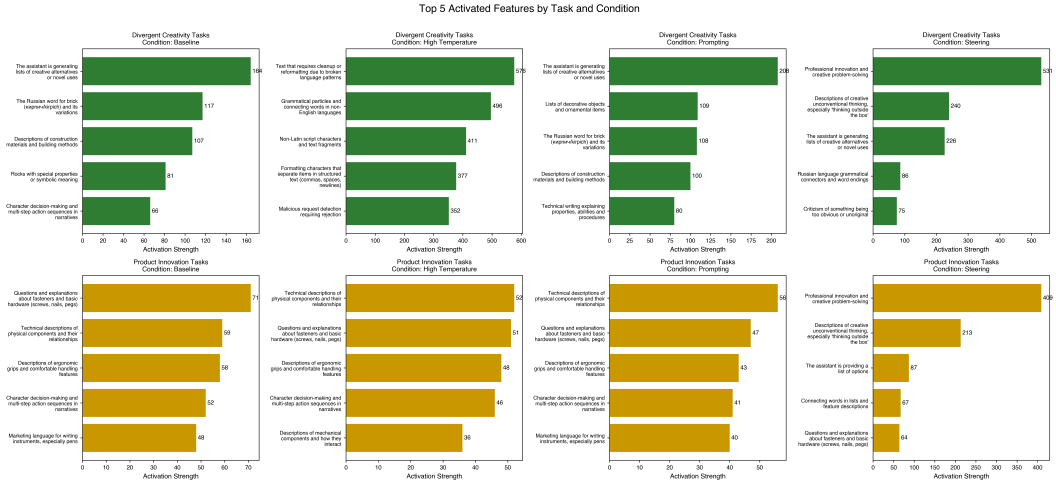


Figure 6: Feature Activations of Generative Agents on Capability Tasks

B.2 Local Decision Feature Rankings

The scan encodes generated-response states with the released layer-50 SAE. Mean activation is computed within each response, then across responses, so long answers do not receive greater response weight. Figure 7 retains all ten coordinates for each decision task, using 1,400 lottery and 680 ultimatum responses. Full labels, IDs, mean activations and scopes appear in `figures_revision/data/baseline_top10_features.csv`; response examples are in `feature_response_examples.csv` in the same directory.



Figure 7: Top 10 Ranked Features for Lottery and Ultimatum Games

B.3 Local Creativity Feature Rankings

Discovery uses only 16 unsteered responses per task. The feature lock fixes the resulting top ten before dose development and confirmation. These small discovery sets characterize the tested prompts; they do not establish population-wide feature rankings.

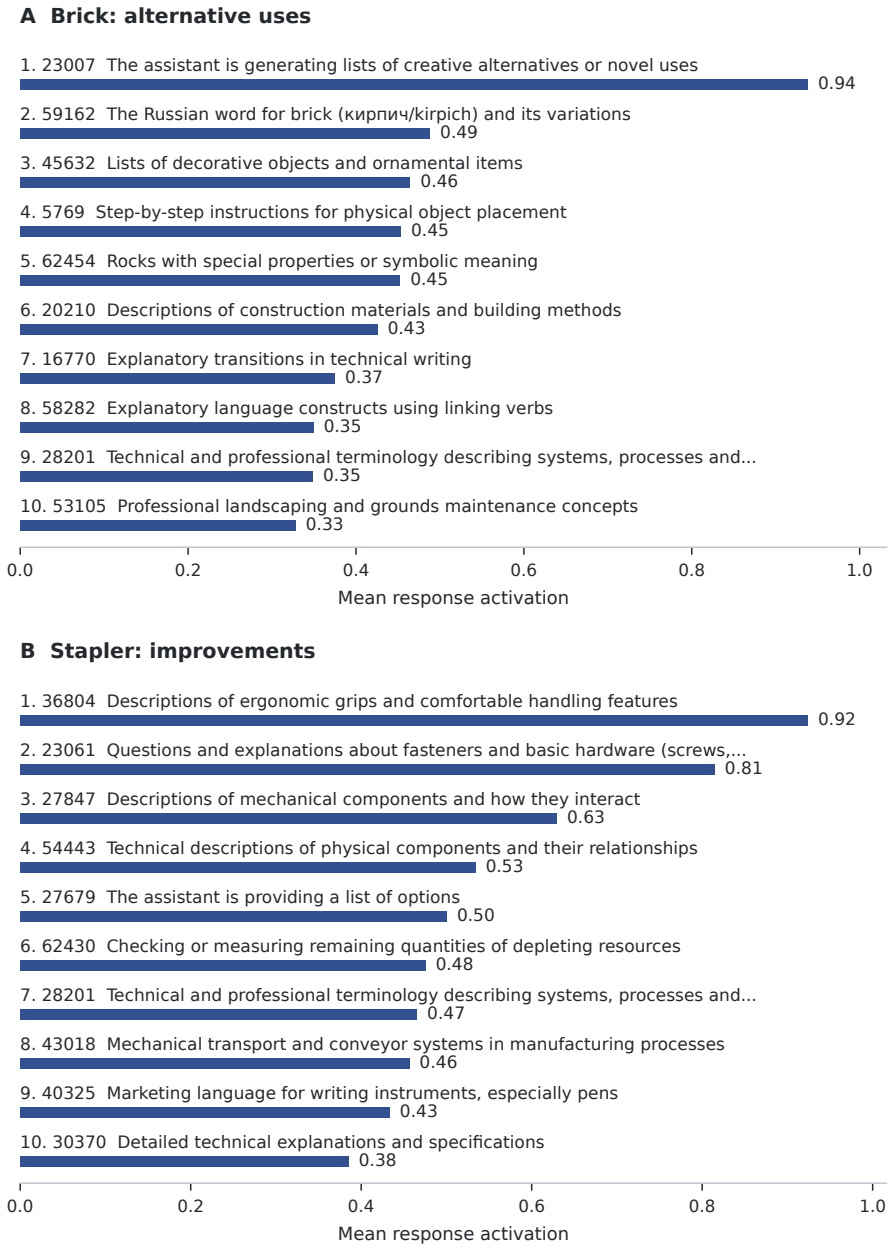


Figure 8: All ten positions are retained.

B.4 Post-steering Re-encoding

For each intervention arm, we re-encode states both immediately before and after the hook edit at decision positions and generated-response prefixes. A pre-edit state during a steered generation already reflects earlier edits and changed text. An immediate post-edit state includes the imposed mathematical change. Consequently, rank overlap is a diagnostic of representation changes, not an independent mechanism test. The complete 6,560-row ranking table includes every scope and transfer condition, with numeric IDs and supplied labels.

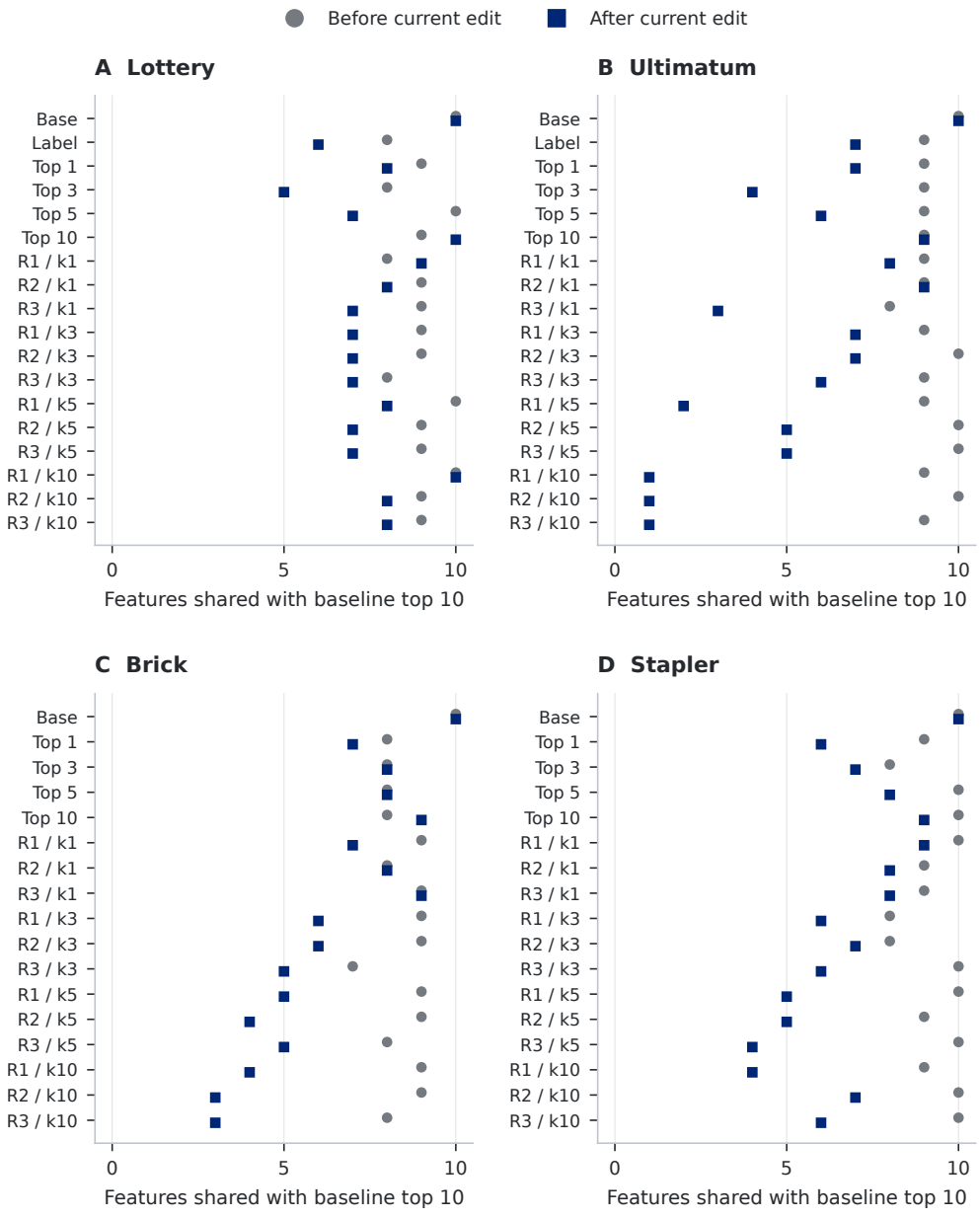


Figure 9: Overlap with the same task's unsteered top ten, using generated-response prefix states from confirmation.

C Prompting Conditions and Ladders

C.1 Prompting Conditions and Exemplars

To establish whether interpretability-based interventions (SAE and probe steering) provide leverage beyond what is achievable through prompt engineering alone, we compare against a graded ladder of prompting baselines applied to the same Llama-3.3-70B-Instruct backbone via local vLLM (temperature 0.7). Each baseline keeps the original task instruction P_{task}. The resulting condition set is $C \in \{\text{baseline}, \text{persona}, \text{cot}, \text{fewshot}, \text{fewshot_cot}\}$ with no modifications to model weights or decoding parameters across conditions.

Few-shot exemplar prompting. Few-shot prompting prepends a small set of demonstrations to the task instruction so the model can in-context learn the desired output distribution (Brown et al., 2020). We use $N{=}6$ demonstrations for every task with a few-shot condition. For divergent creativity we draw the demonstrations from the Ocsai Alternative Uses Task corpus (Organisciak et al., 2023), a publicly released dataset of human-authored brick uses paired with human-rated originality scores on a 10–50 scale; we rank the brick subset by human originality, deduplicate, drop entries where Ocsai's object-anonymization left ungrammatical text, and retain the top-N. The exemplars are inserted as an inspiration block to the prompt. They are human-generated and the selection criterion (human originality) is independent of the downstream GPT-based judge, which avoids selection-evaluation circularity. For the lottery and ultimatum games, the demonstrations are synthetic question-answer pairs constructed over held-out reward / offer grids that are disjoint from the test sweeps. Formally, the prompt becomes

$$P_{\text{fewshot}} = E_{1:N} \oplus P_{\text{task}},$$

where $E_{1:N}$ denotes the bulleted exemplar block and $\oplus$ denotes string concatenation. Because the Ocsai corpus is task-specific to brick AUT, the few-shot condition is omitted from the stapler product-innovation task.

Chain-of-thought prompting. Standard chain-of-thought prompting elicits intermediate reasoning prior to a final answer (Wei et al., 2022; Kojima et al., 2022). We append a single, task-appropriate trigger I_{cot} to P_{task}, yielding

$$P_{\text{cot}} = P_{\text{task}} \oplus I_{\text{cot}}.$$

The trigger is tailored to the task family but always uses the same one-completion, reasoning-then-answer format. We document the detailed steps in the appendix.

Combined few-shot CoT and dose-response. The few-shot and CoT scaffolds are orthogonal and compose by concatenation:

$$P_{\text{fewshot_cot}} = E_{1:N} \oplus P_{\text{task}} \oplus I_{\text{cot}}.$$

For divergent creativity this single combined condition (`fewshot_cot`) is our strongest prompting variation. For the lottery and ultimatum games we additionally trace a prompting dose-response: we vary the fraction of the $N = 6$ demonstrations that take the target action over the reachable values $\{0, 2/6, 3/6, 4/6, 1\}$ (risky for lottery, accept for ultimatum), with the remaining $1{-}\rho$ taking the opposite action, and use reasoning-bearing demonstrations throughout.

D SPARSE AUTOENCODER STEERING

D.1 LOCAL IMPLEMENTATION AND ORIGINAL SCHEMATIC

The model is Llama-3.3-70B-Instruct at revision `6f6073b423013f6a7d4d9f39144961bfbfbc386b`; the layer-50 SAE is pinned at `128ee921ecd1b8b3a87d776cbcc357c0855da134`. Model inference uses bfloat16 and SAE operations float32. The hook is installed at zero-based `model.model.layers[50]`. The additive SAE hook operates on the final token of every forward pass, including cached generation. Decision generation uses temperature 0.5; the label-selected study allows 1,000 new tokens. Open-source EDSL provides experiment orchestration; inference and editing run locally in Transformers.

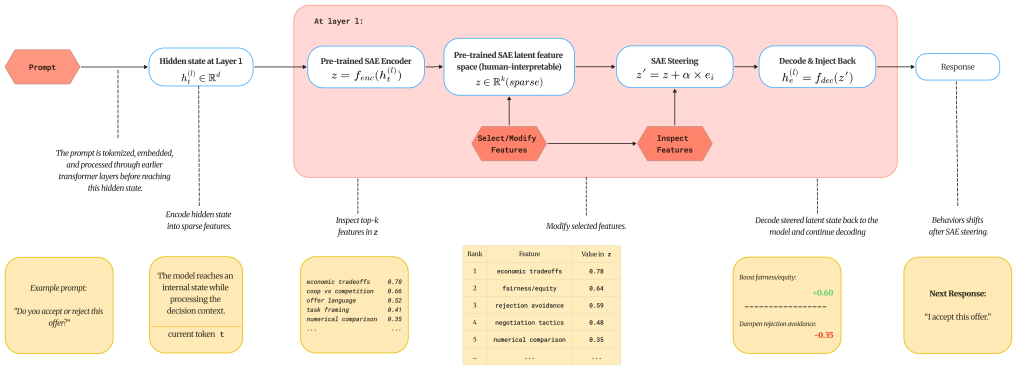


Figure 10: SAE interpretability and steering pipeline.

D.2 LABEL-SELECTED FULL-GRID EXPERIMENT

Features 184 and 4237 were selected for risk-related labels, and 31935 for altruism-related language. The frozen selected positive edits are $+2.8$ and $+2.0$ for the lottery pair, and $+8.0$ for ultimatum. These local coefficient units should not be equated with hosted-platform strengths. Development tested multipliers 0.25, 1, 4, 16, and 64. Lottery selected 4; ultimatum used the declared fallback of 16 because no development dose passed its full criterion. Selection was fixed before confirmation.

Confirmation uses six arms: baseline, original positive dose, selected positive dose, negative edit, music feature 22766, and photosynthesis feature 55316. The latter labels were selected as controls before confirmation; they were not identified by an exhaustive behavioral search of all 65,536 features. The saved protocol, selection lock, and estimates accompany the source.

Selected-dose lottery effects are $+5.93$ percentage points (saved 95% interval $[5.00, 6.86]$); ultimatum effects are $+3.53$ points ($[2.79, 4.41]$). The prespecified specificity criteria were not met; Figure 11 retains all control estimates. Behavioral shifts therefore do not establish label-specific psychological mediation. Original-dose effects are $+1.21$ and $+0.29$ points. Negative-edit realized norms are only 0.00884 for lottery and zero for ultimatum, limiting interpretation of suppression tests.

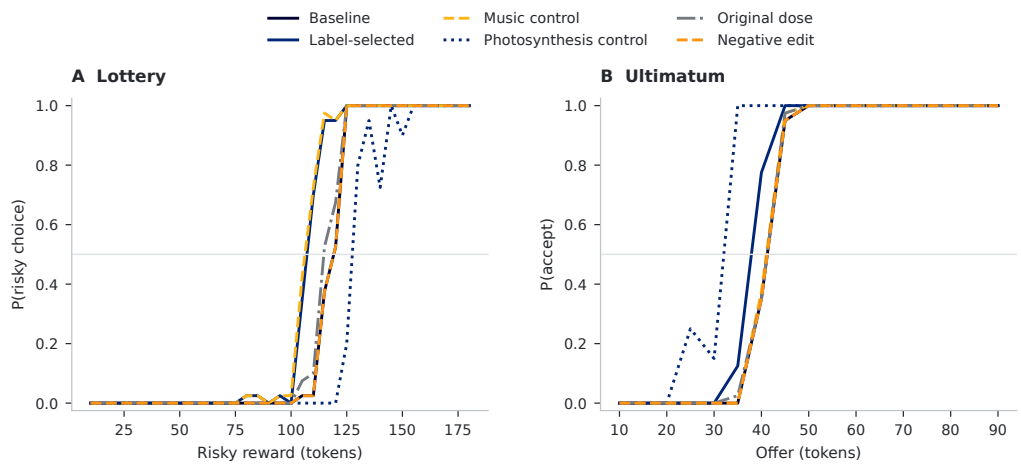


Figure 11: All six full-grid confirmation arms from the earlier local SAE experiment, with 40 responses per setting.

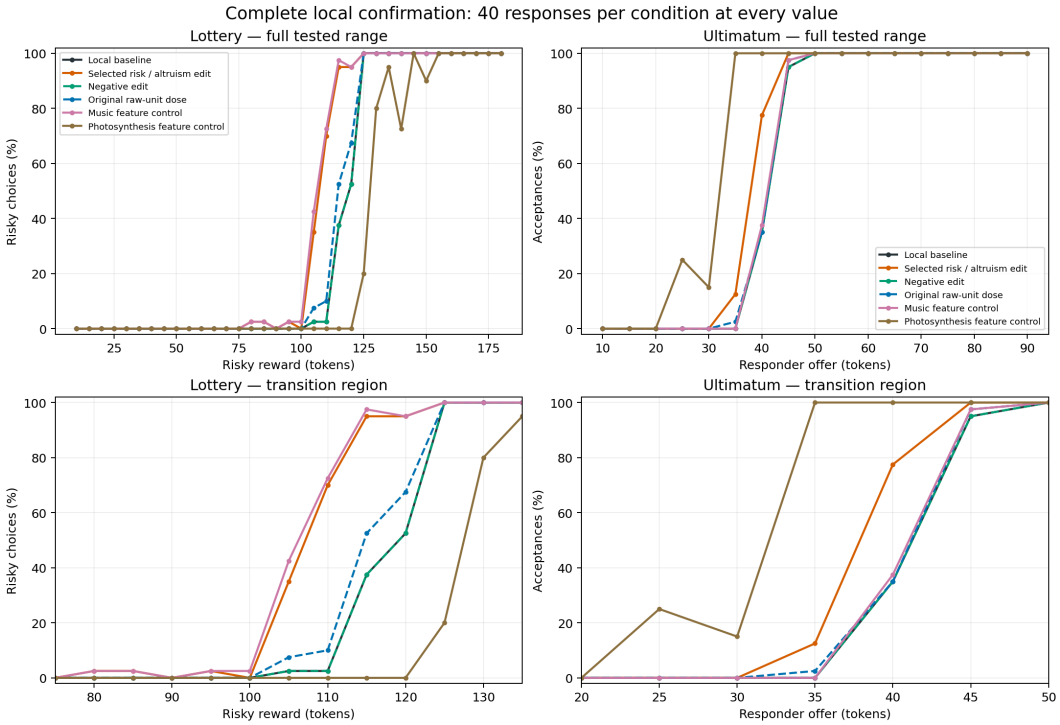


Figure 12: Complete label-selected local confirmation curves and transition-region details. Selected edits and prespecified controls share the same within-study baseline.

Bounded feature screen and semantic examples. Five additional single-feature screens test two norm scales on 112 responses per feature and strength. The effects below are exploratory; they were not confirmed by the full-grid study. The high-strength feature-184 arm has 26 invalid responses, so its apparent decrease is not a clean behavioral result.

Table 3: Exploratory single-feature screen. Changes are percentage points in risky choice (first three rows) or ultimatum acceptance (last two rows).

ID	Label description	Scale 4	Scale 16
184	Risk willingness	+13.39	-10.71
4237	Risky operations requiring caution	+12.50	+33.93
48709	Financial risk and investment caution	-11.61	-47.32
4157	Charitable giving	-0.89	-15.18
5546	Fairness	-0.89	-0.89

Hand-authored semantic contrasts provide examples of label plausibility. Mean maximum user-content activation is 2.879 versus 1.246 for an altruism self-sacrifice contrast, 1.102 versus zero for charity, 2.726 versus zero for music, and 6.391 versus zero for photosynthesis. The risky-operations feature activates equally for caution and routine-command sentences. These use four contextual variants of one contrast pair per topic, not four independent semantic scenarios. They do not validate all 65,536 coordinates or establish behavioral mediation.

D.3 Weighting Robustness

A separate condition weights selected coordinates by their baseline discovery activations. This is a prespecified heuristic, not circular selection on confirmation outcomes and not a direct measure of causal importance. Activation-weighted top-3/5/10 effects are $+7.14, +8.57, +0.36$ points for lottery and $+2.14, -1.07, -0.36$ for ultimatum. The comparison does not establish superiority over equal weighting.

D.4 Transfer and Baseline Sensitivity

The fixed decision interventions transfer to work/travel risk choices and money/time-help choices, with 140 responses per arm over seven settings and two answer orders. Work is at a floor and time-help near a ceiling. The top-one transferred feature is 13951 ($+3.338936$) for lottery and 50689 ($+7.611234$) for ultimatum; random comparators are 16478 ($+4.490092$) and 4135 ($+9.906348$). None of the saved within-scenario corrected comparisons passes.

The earlier label-selected study also changes seeds and agent IDs together, yielding $+14.64$ lottery and $+6.43$ ultimatum points. It is not a clean seed-only robustness test. Paraphrased prompts show no effects under floor/ceiling responses. Cross-game transfer gives -22.32 lottery points for the altruism-labeled edit and -0.89 ultimatum points for the risk-labeled edit. These results limit generalization from a successful original prompt.

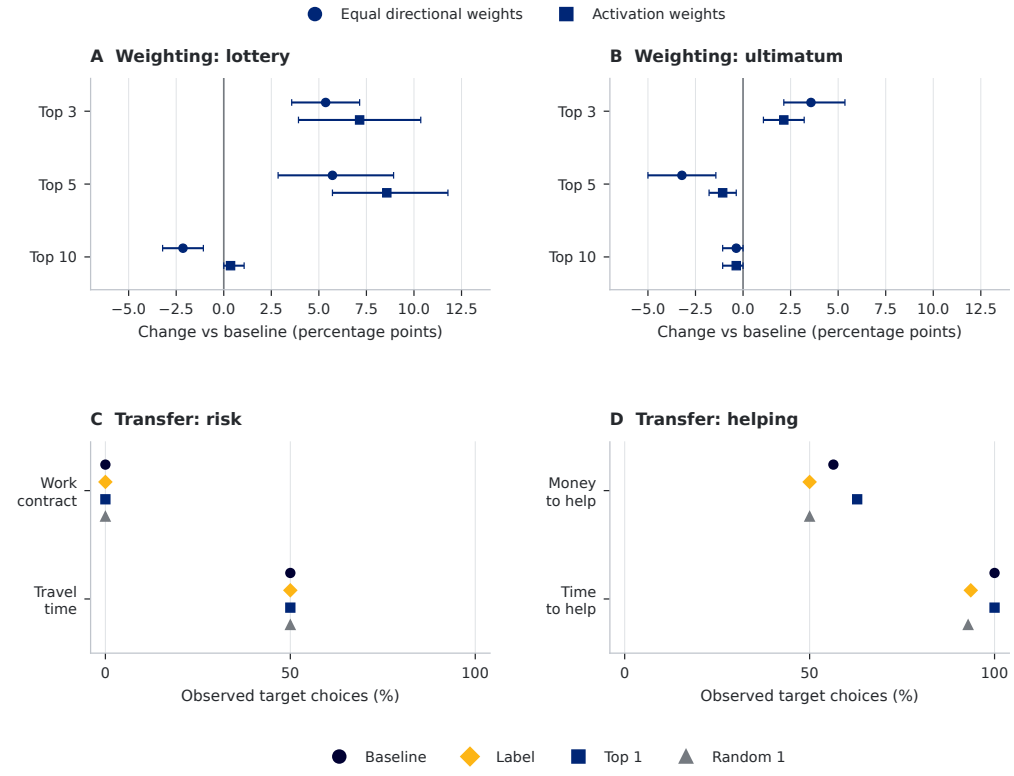


Figure 13: A-B: equal and activation-weighted bundle effects share the same confirmation baseline; intervals are saved paired-batch descriptive 95% intervals. Neither plot tests superiority of weighting directly. C-D: observed rates in four transfer scenarios, 140 responses per arm over seven settings and two answer orders.

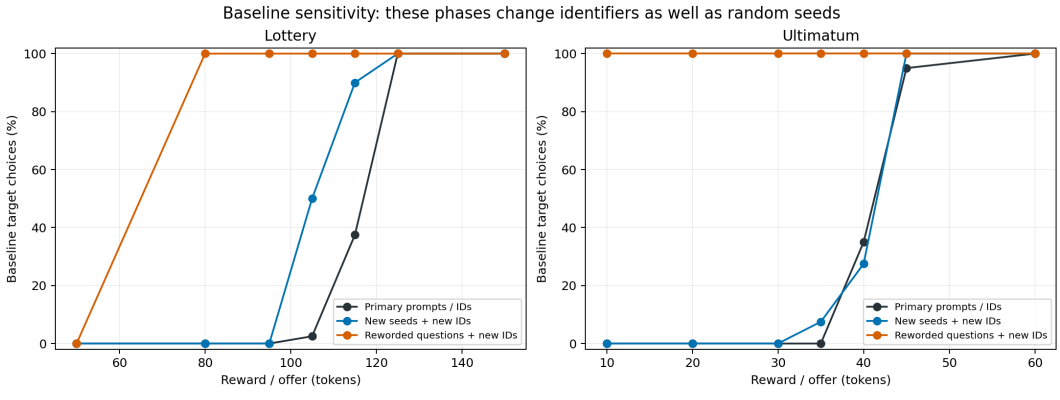


Figure 14: Baseline sensitivity across phases of the label-selected local study.

E PROBE STEERING AND ROBUSTNESS

E.1 IMPLEMENTATION, SCHEMATIC, AND DECISION CALIBRATION

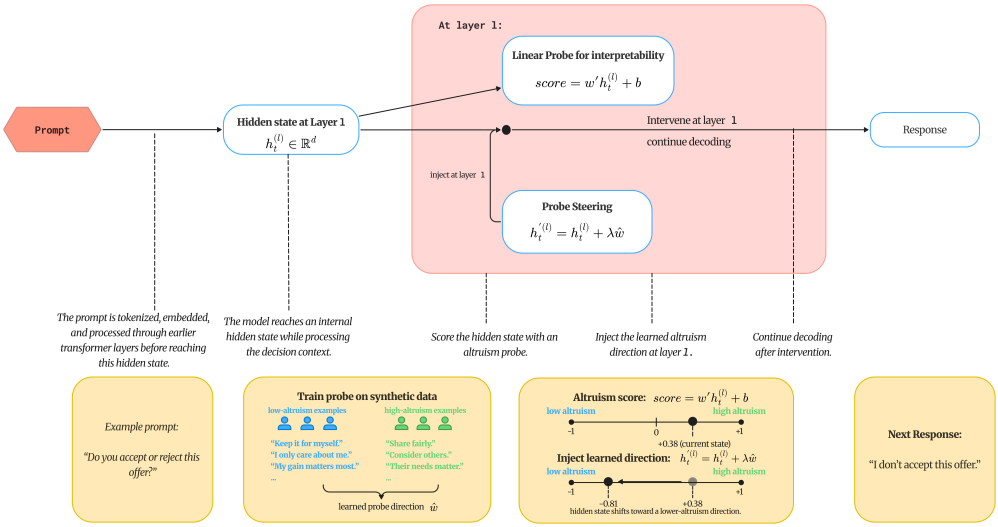


Figure 15: Probe scoring and steering pipeline.

Reading the retained protocol. The baseline recipe below describes the intended zero-dose comparison, but the saved calibrated decision sweeps do not contain a zero-dose arm. Figure 16 uses raw-choice crossings; main Figure 2 reports prompting and selected local SAE curves, with the matched SAE baseline retained in Appendix D. Twelve calibration seeds overlap the 40 evaluation seeds. Trial-level folds can share prompt activations; "held-out" in the retained layer-selection description does not mean prompt-held-out evaluation. The older preference display below is retained for provenance, not used to calculate the new raw-choice errors.

E.1.1 OBSERVED PROBE CALIBRATION AND LAYER ROBUSTNESS

Figure 16 evaluates observed calibration across layers. It reconstructs the first 50% crossing by linear interpolation of raw risky-choice or acceptance rates, rather than target-centered display curves. Each target curve contains 40 agent decoding instances. Mean absolute target errors are 2.1286 lottery tokens and 1.1473 ultimatum tokens at layer 48; corresponding layer-50 values are 2.5724 and 1.7979. These are checks of observed calibration and sensitivity to layer, not independent validation of a psychological construct. Twelve calibration seeds overlap the evaluation seeds, and the saved sweep lacks a matching zero-dose arm.

E.1.2 STACK OVERVIEW

Software stack. We implement probe-based steering in PyTorch using HuggingFace Transformers for model inference, and scikit-learn for training ℓ_2-regularized logistic regression probes.

Base model and representation. Probe-based steering uses a frozen base language model and a linear probe trained on intermediate activations. In our main experiments we use Llama-3.3-70B-Instruct and select a single layer for intervention based on held-out probe performance (layer 48 in our Llama runs). We also replicate the full probe pipeline on Qwen-2-7B-Instruct (layer 17); those results appear in the Qwen replication subsection below.

Activation extraction. To compute probe scores, we run a forward pass and capture the hidden state tensor at the selected layer using a forward hook. We use the final token representation as a compact summary of the prompt context.

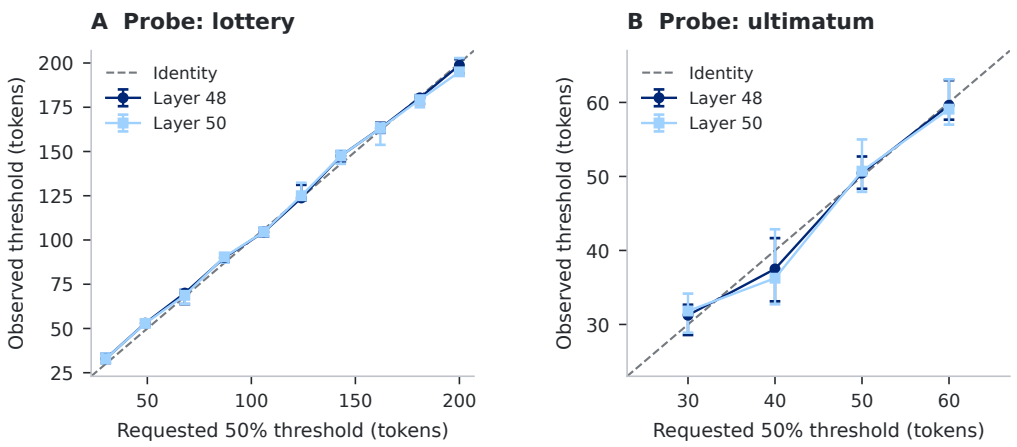


Figure 16: Probe calibration robustness for lottery (A) and ultimatum (B). Points are raw-choice 50% crossings at layers 48 and 50; bars are descriptive 95% agent-bootstrap intervals (2,000 resamples). Calibration reuses 12 of 40 evaluation seeds. These sweeps lack a matching zero-dose baseline and are not fully held-out validation.

```
# Sketch of probe score extraction
layer = model.model.layers[layer_idx]
acts = []

def capture_hook(module, input, output):
    hidden = output[0] if isinstance(output, tuple) else output
    acts.append(hidden.detach())
    return output

handle = layer.register_forward_hook(capture_hook)
_ = model(**tokenizer(prompt, return_tensors="pt"))
handle.remove()

last_token = acts[0][0, -1, :].float()
score = torch.dot(last_token, probe_direction_unit)
```

Probe training. We fit an ℓ_2-regularized logistic regression probe to predict binary behavioral labels from activations. We standardize activations prior to training and select the regularization strength via cross-validation. We train probes across layers and choose the layer with best held-out discriminative performance. This yields both the steering direction (the probe weight vector) and the intervention site (the best layer).

E.1.3 Agents, Cohorts, and Trial Structure

We implement the appendix cohorts from the main paper in a probe-specific form by treating each agent as an independent decoding instance indexed by a deterministic random seed.

Baseline trials. We sample n agents per task parameter with $\lambda = 0$, producing unsteered psychometric curves and probe scores.

Steering trials. We repeat the same prompts and decoding settings but enable a single-layer probe steering hook with a specified $(\lambda, \text{direction})$.

Deterministic seeding. For preference tasks, we set the seed as a function of the task parameter index and agent ID so that trial sets are reproducible.

```
# Sketch of trial loop and deterministic seeding
for param_idx, param in enumerate(task_parameters):
  for agent_id in range(n_agents):
    seed = seed_base + param_idx * 1000 + agent_id
    torch.manual_seed(seed)
    if torch.cuda.is_available():
      torch.cuda.manual_seed_all(seed)
```

```
prompt = make_prompt(param)
response = generate(prompt, lambda_value, direction)
```

E.1.4 TASKS, LABELS, AND TARGETS

Lottery. The prompt describes a choice between a safe option (guaranteed 50 tokens, fixed across trials) and a risky option (50% chance of a variable reward, 50% chance of 0 tokens). The risky reward is varied to produce a psychometric curve. We define the target behavior as the switching point, the risky reward at which $P(\text{Risky}) = 0.5$, computed by linear interpolation across reward levels.

Ultimatum. The prompt describes an offer to split 100 tokens, where the model must answer Accept or Reject. The offer amount is varied to produce an acceptance curve. We define the target behavior as the acceptance threshold, the offer size at which $P(\text{Accept}) = 0.5$.

Binary labels for probe training. For preference tasks, we label each trial by the agent's observed choice (risky vs. safe in the lottery game, accept vs. reject in the ultimatum game), so that the positive class is the higher-risk (respectively, offer-accepting) behavior. No median split is applied to the preference probes. Because a risk-seeking agent takes the gamble even at a low risky reward, i.e. it switches early at a low switching point, labeling by the choice keeps these early-switching trials in the high class, rather than inverting them as an above-median split on the switching point would. The switching point and acceptance threshold serve only as calibration targets for λ (Section E.1.6), never as training labels. We discard trials whose choice cannot be parsed and train a logistic-regression probe on the remaining labeled activations. Median splitting is used only for the capability (creativity) probes, where outputs are split on the LLM-judge score into high- and low-creativity classes.

E.1.5 SCENARIOS AND HOW WE IMPLEMENT THEM

Probe experiments use a fixed prompt template and vary only the steering strength and direction.

Baseline **Meaning.** Measure unsteered behavior and probe scores under the same decoding settings used for steering.

Implementation.

- Steering strength: $\lambda = 0$
- Direction: unused
- Decoding parameters (including temperature) held fixed

What we log.

- Task parameter (risky reward or offer amount)
- Parsed choice (Safe/Risky or Accept/Reject)
- Probe score computed on the prompt context

Steering **Meaning.** Modify the model's internal state at inference time by adding a scaled probe direction vector to hidden states at a single layer.

Intervention hook. During decoding we register a forward hook on the selected layer and add an intervention of the form $\Delta h = s\,\lambda\,v$, where v is the probe direction scaled to match typical activation magnitudes, λ is the steering strength, and $s \in \{+1, -1\}$ encodes the increase versus decrease direction. We apply the intervention only to the final portion of the sequence (the last 20% of tokens) and scale the intervention linearly from 0.5 to 1.0 over these target positions. The window is computed from the sequence length of the current forward pass. During prefill this is the prompt length, so the ramp covers the final 20% of prompt positions; during cached decoding the pass carries a single token, so the ramp does not apply to generated positions. Steering is therefore applied to prompt positions only.

```
# Sketch of probe steering at one layer
sign = 1.0 if direction == "increase" else -1.0
start_idx = seq_len - max(1, seq_len // 5)
```

```
position_scale = torch.linspace(0.5, 1.0, seq_len - start_idx)
hidden[:, start_idx:, :] += (sign * lambda_value * v_scaled) * \
    position_scale.view(1, -1, 1)
```

Probe scores under steering. For analysis, we report probe scores both before and after steering. The post-steering score is computed by re-running the probe score extraction on the concatenated prompt and generated response.

E.1.6 Lambda Calibration and Dose-Response

To achieve a desired target behavior (for example, a specified switching point), we solve for the steering strength λ using a two-stage search. Figure 18 shows the dose-response relationship between target switching points and the required λ values for both lottery and ultimatum games. We report λ as a signed quantity along the probe direction $\hat{w}$, whose positive class is the risky (lottery) or accept (ultimatum) choice. Positive λ therefore moves the agent toward risk-seeking behavior in the lottery game and toward accepting lower offers in the ultimatum game, and negative λ toward risk aversion and toward rejecting unequal offers. Consistent with this convention, the required λ becomes increasingly negative as the target switching point or acceptance threshold rises, since higher targets correspond to more risk-averse agents and to agents with a higher acceptance threshold, that is, more inequity-averse ones.

We first evaluate a coarse grid of λ values to identify a promising region, then run a bracketed binary search until the achieved metric is within a fixed tolerance. We cache trial logs for each $(\lambda, \text{direction})$ evaluation to make repeated searches efficient.

E.1.7 Internal Probe Dynamics

Figure 19 analyzes how internal probe scores track target behavior during steering. The probe activation score (y-axis) is a scalar readout computed from the model's internal hidden-state vector at the selected layer using our trained logistic-regression probe. Concretely, for each trial we capture the hidden state $h \in \mathbb{R}^d$ at the intervention layer and compute its signed projection onto the unit-normalized probe direction $\hat{w} = w/||w||_2$: $s = \hat{w}^\top h$. Because the probe is a logistic-regression classifier, this dot product corresponds to the probe's linear decision score up to an additive intercept term, so larger s indicates stronger alignment with the probe's positive-class direction.

For the lottery game (left panel), mean probe scores decrease as the target switching point rises from 40 to 240 tokens, consistent with higher switching points corresponding to more risk-averse behavior. Risky choices consistently receive higher probe scores than safe choices, and this separation is maintained across all target levels. For the ultimatum game (right panel), mean probe scores likewise decrease as the target acceptance threshold rises from 30 to 100 tokens, consistent with higher thresholds corresponding to a stronger unwillingness to accept unequal offers. Accept choices receive higher probe scores than reject choices, and this separation is maintained across target values. In both tasks, probe scores track the intended steering direction while preserving clear separation between choice types, showing separation along the fitted readout under intervention. This is consistent with the probe direction capturing preference-related variation, though the probe is fitted from the same activations it is evaluated on, so it does not establish that the direction is specific to preference rather than to the reward level that drives the choice.

E.1.8 Invariants and Controls

- **Identical prompts and decoding across conditions.** Within a sweep, Baseline and Steering use the same prompt templates and decoding parameters. Only λ and direction differ.
- **No prompt leakage of targets.** Prompts do not mention switching points, acceptance thresholds, or desired behaviors. Target values are only used in offline metric computation and λ selection.
- **Layer and direction selection.** We select the intervention layer by held-out probe performance and keep it fixed across all steering runs. Steering uses the probe weight vector as the single intervention direction.
- **Seeded sampling.** We set deterministic seeds as a function of parameter index and agent ID so that trial sets are reproducible across reruns.

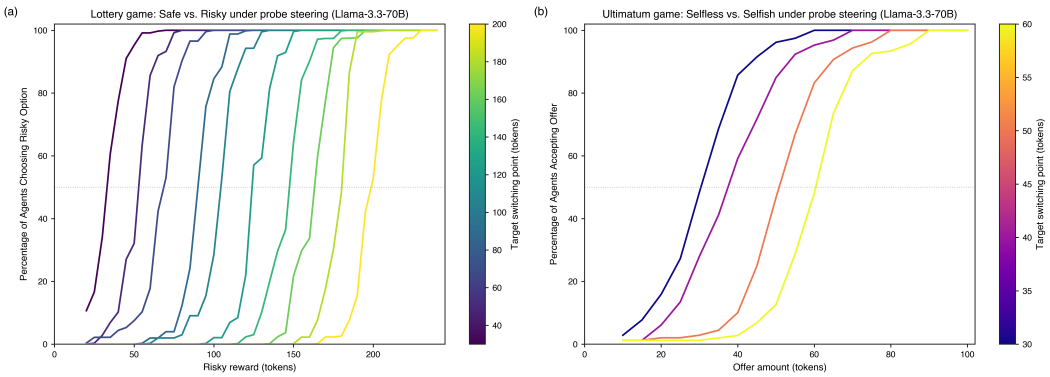


Figure 17: Performance of using probes to steer generative agents in preference tasks

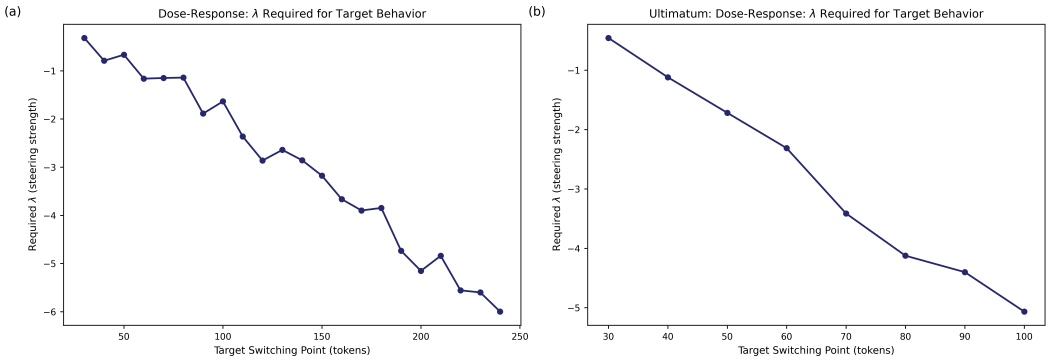


Figure 18: Dose-response showing λ required to reach target switching points for lottery (left) and ultimatum (right) games.

E.1.9 Reproducibility and Artifacts

Every probe run produces:

- A probe artifact (pickle) containing the probe weights, bias, layer scores, and selected best layer
- JSONL trial logs with fields including task type, task parameter, λ, direction, decoded response, parsed choice, and probe scores
- Cached trial evaluations used by the λ solver (one cache file per evaluated $(\lambda, \text{direction})$)
- Figure scripts that load cached JSONL logs to produce psychometric curves and dose-response plots

Calibrated range. Every calibrated preference strength in the range is negative (-0.31 to -6.00); the lottery fit has slope -36.1 tokens per unit and $R^2 = 0.98$. Its zero-dose extrapolation (26.9) differs from the hosted baseline (132.9). With no matching zero-dose arm, linearity should be interpreted only over the observed calibrated range. The new raw-choice table reports mean absolute target error 2.1286/1.1473 at layer 48 and 2.5724/1.7979 at layer 50 for lottery/ultimatum.

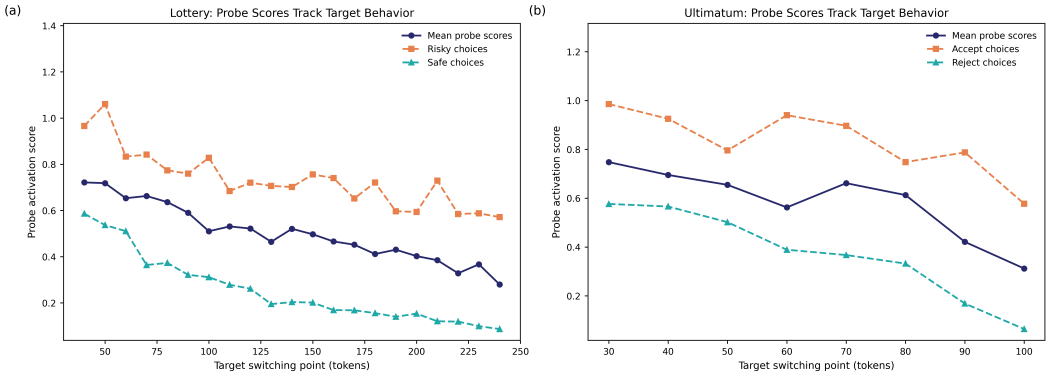


Figure 19: Probe scores tracking target behavior for lottery (left) and ultimatum (right) games on Llama-3.3-70B-Instruct (layer 48). Risky (accept) choices score higher than safe (reject) choices, and mean scores fall as the target switching point (acceptance threshold) rises.

E.2 Capability Control

We then turn to probes to examine controllability. The left panel of Figure 20 presents controllability results for divergent creativity using the brick prompt via our probe-based methods. The figure shows the achieved creativity score as a function of the target score, with one line per judge and a bold line for the mean across the five judges. The diagonal line indicates perfect controllability. The probe achieves meaningful control, shifting the mean achieved score from approximately 4.0 at a target of 3 to approximately 7.6 at a target of 9, though achieved scores fall below the perfect-control diagonal at higher targets. This gap between target and achieved scores widens at higher targets, suggesting that while probe-based steering captures a substantial portion of the creativity dimension, it cannot fully saturate the trait. The right panel of Figure 20, showing how probes steer the creativity of product innovation tasks, demonstrates similar results to those for divergent creativity. We also examine whether probes trained on one object category can steer creativity for out-of-sample objects. Specifically, the divergent-creativity probe is trained on brick-only labeled data and then applied as a fixed steering direction across four object prompts: brick (in-distribution) and stapler, paperclip, and bowl (out-of-sample). The results, reported in Appendix Figure 21, show that this transfer-based implementation is both feasible and effective.

Compared to the preference tasks, capability control through probes shows a more compressed range relative to the target scale, with consistent ordering across the tested target settings. The achieved-versus-target slope is less than unity, indicating that each unit increase in the target yields a diminishing return in achieved creativity. This partial controllability is consistent with the view that creativity resists compression into a single linear direction, as it involves coordinated changes across fluency, flexibility, originality, and elaboration.

The main-text reconstruction averages the five judges within response and bootstraps the 40 responses at each target.

E.3 Object Transfer and Classification

E.3.1 Additional Capability Control Results

Figure 21 presents the full controllability results for divergent creativity across all four object prompts (brick, stapler, paperclip, bowl). Each curve shows the achieved creativity score, averaged across the five judges, as a function of the target score, with error bars representing the standard deviation across judges. The diagonal line indicates perfect controllability. All four objects show a consistent pattern: achieved scores increase monotonically with target scores, ranging from approximately 3.3 to 7.3, but fall below the diagonal at higher targets. The gap between target and achieved scores widens as the target increases, indicating partial but meaningful control across all objects.

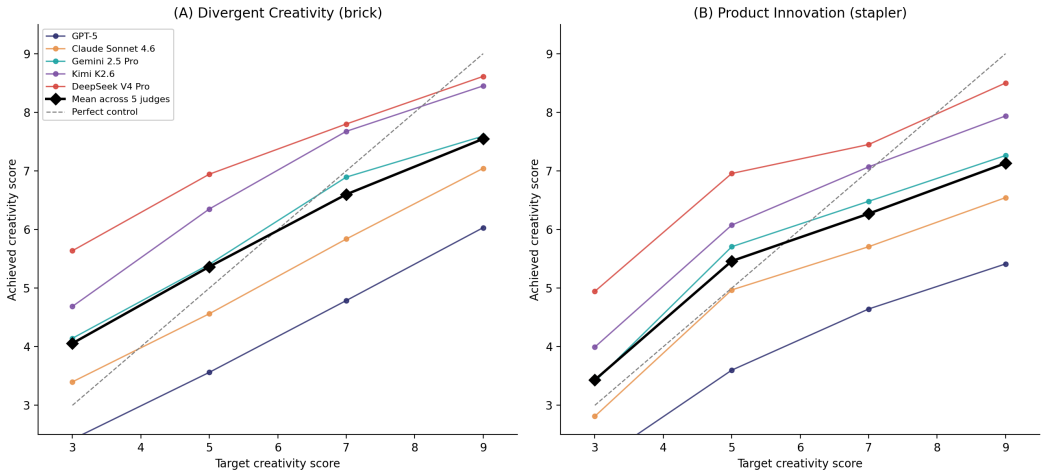


Figure 20: In-distribution capability control for two task-specific probes on Llama-3.3-70B-Instruct (layer 48).

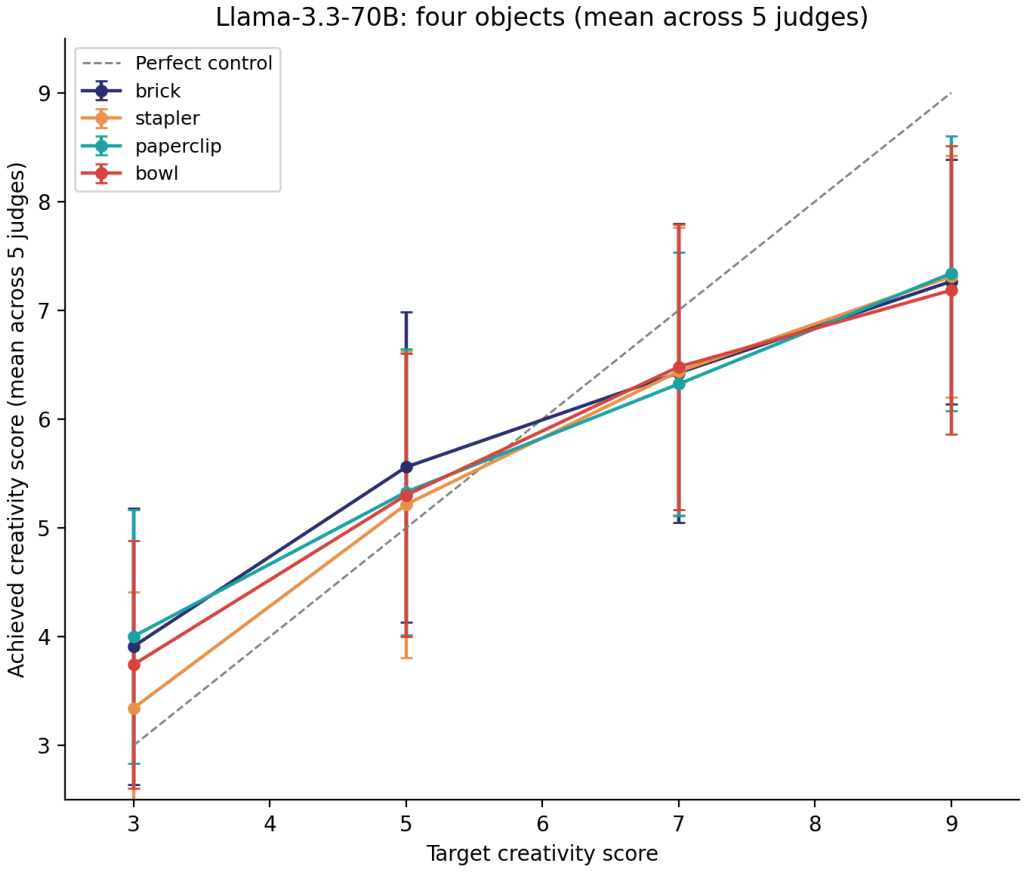
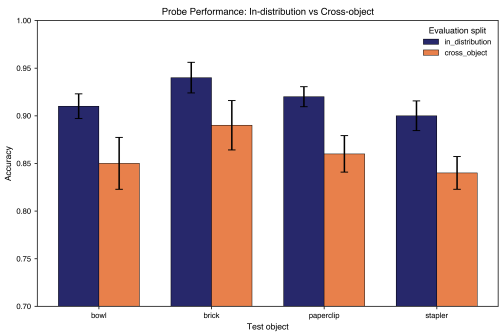
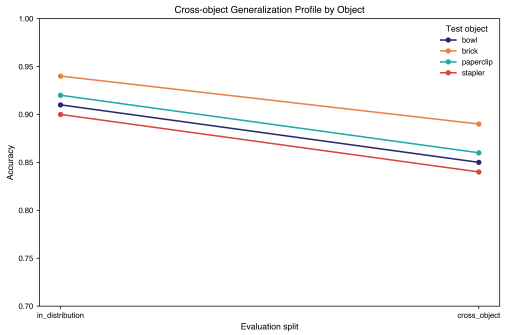


Figure 21: Achieved creativity scores as a function of target scores when the divergent-creativity probe, trained on brick-only labeled data, is applied as a fixed steering direction across four object prompts (brick, stapler, paperclip, bowl) on Llama-3.3-70B-Instruct (layer 48).



(a) Cross-object generalization performance comparing in-distribution and cross-object probe accuracy across test objects (Llama-3.3-70B-Instruct, layer 48).



(b) Performance drop from in-distribution to cross-object evaluation for each test object (Llama-3.3-70B-Instruct, layer 48).

Figure 22: Cross-object generalization analysis on Llama-3.3-70B-Instruct (layer 48).

E.3.2 Cross-Object Generalization

To assess whether learned probe directions generalize beyond their training distribution, we implement cross-object generalization experiments. We first train a logistic regression probe on hidden activations from Llama-3.3-70B-Instruct for divergent creativity. The probe takes a single representation layer as input and predicts a binary label (high versus low creativity). For in-distribution evaluation, the probe is trained and tested on examples that all use the same object (for instance, train and test on brick). For cross-object evaluation, the probe is trained on a set of objects and tested on a held-out object that never appears in the training data (for instance, train on brick, stapler, and paperclip and test on bowl). Performance is reported as accuracy averaged over multiple random seeds.

Figure 22a compares in-distribution and cross-object performance for each test object. The results show that probes maintain reasonable accuracy when tested on held-out objects, though performance consistently drops relative to in-distribution evaluation. Figure 22b illustrates the magnitude of this performance drop for each object. The degradation ranges from approximately 5 to 6 percentage points depending on the object, suggesting that while probe directions capture substantial generalizable structure, they also encode object-specific patterns that do not transfer perfectly across contexts.

E.4 Qwen and Layer-50 Checks

E.4.1 Robustness Check with Qwen-2-7B-Instruct

To verify that our probe-based findings are not specific to a single model architecture, we replicate the full probe pipeline on Qwen-2-7B-Instruct. This check is intended to assess the portability of the probe-based intervention, not to provide a full cross-model comparison across all methods. We follow the identical procedure described above: synthetic data generation, layer sweep, probe training, λ calibration, and behavioral evaluation. The best-performing layer for Qwen was layer 17, selected via a cross-validated layer sweep. Held-out classification accuracy on divergent creativity averages approximately 85% in-distribution across object prompts (see Figure 28a), compared to approximately 92% for Llama on the same construct (Figure 22a). Both probes substantially exceed chance, confirming that the probe pipeline transfers to a smaller model architecture.

Figure 23 shows the psychometric curves for the lottery and ultimatum games under probe-based steering on Qwen-2-7B-Instruct. Figure 24 and Figure25 shows the internal dynamics of probes for both lottery and ultimatum games. The qualitative patterns are consistent with the Llama results reported in the main text: steering successfully shifts switching points across a wide range, and the dose-response relationship is monotonic. Figure 26 and Figure 27 present the capability control results for the brick prompt and all four object prompts, respectively. Figure 28a and Figure 28b report classification accuracy and its drop for held-out objects; these are separate from steering to a targeted score. However, as with Llama, a ceiling effect is observed across all objects.

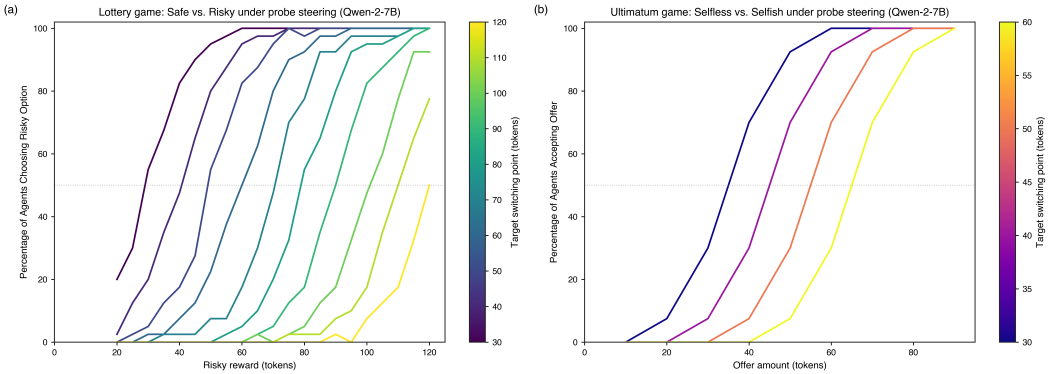


Figure 23: Psychometric curves across calibrated λ values for lottery and ultimatum games using probe-based steering on Qwen-2-7B-Instruct (layer 17).

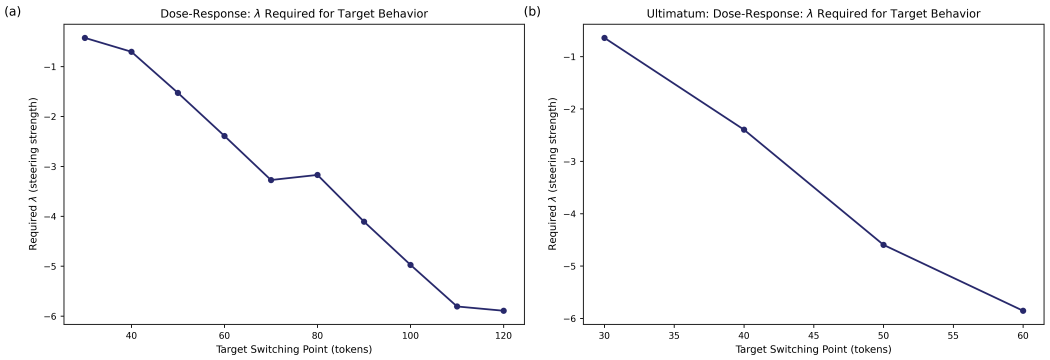


Figure 24: Dose-response showing λ required to reach target switching points for lottery and ultimatum games on Qwen-2-7B-Instruct (layer 17). Sign convention as in Figure 18.

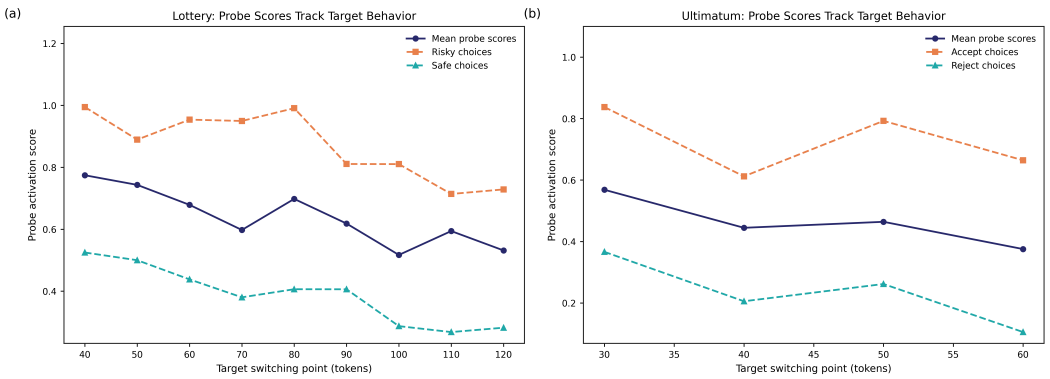


Figure 25: Probe activation score distributions for risky versus safe choices across different steering strengths on Qwen-2-7B-Instruct (layer 17).

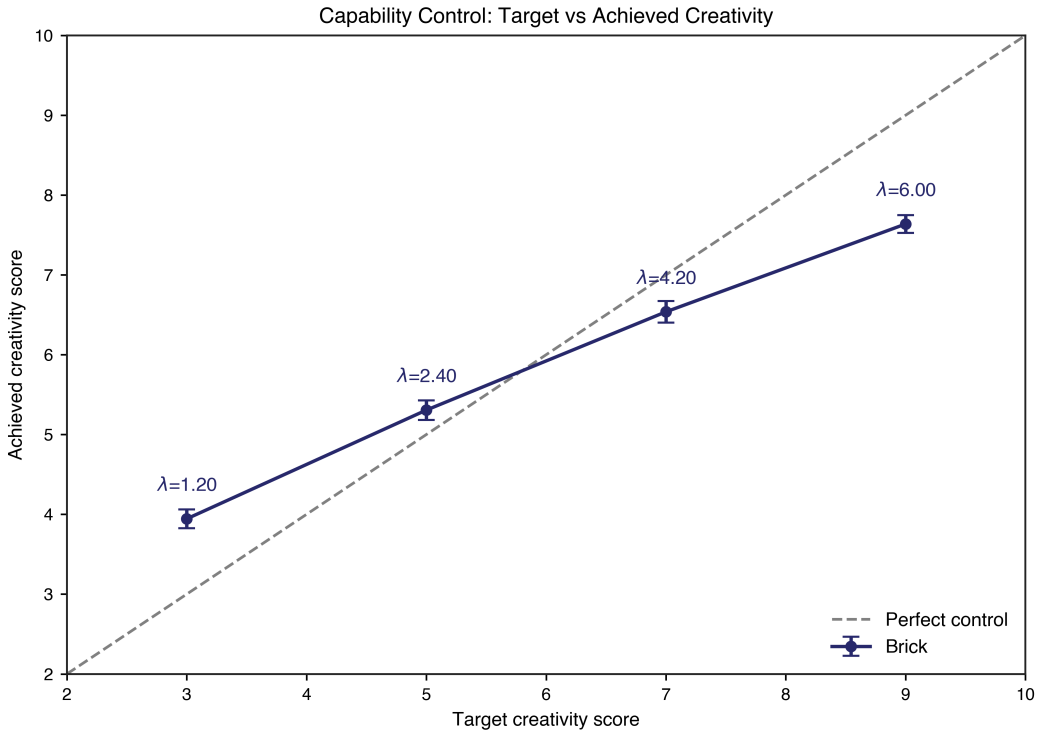


Figure 26: Achieved creativity scores as a function of target scores for the brick prompt using probe-based steering on Qwen-2-7B-Instruct (layer 17).

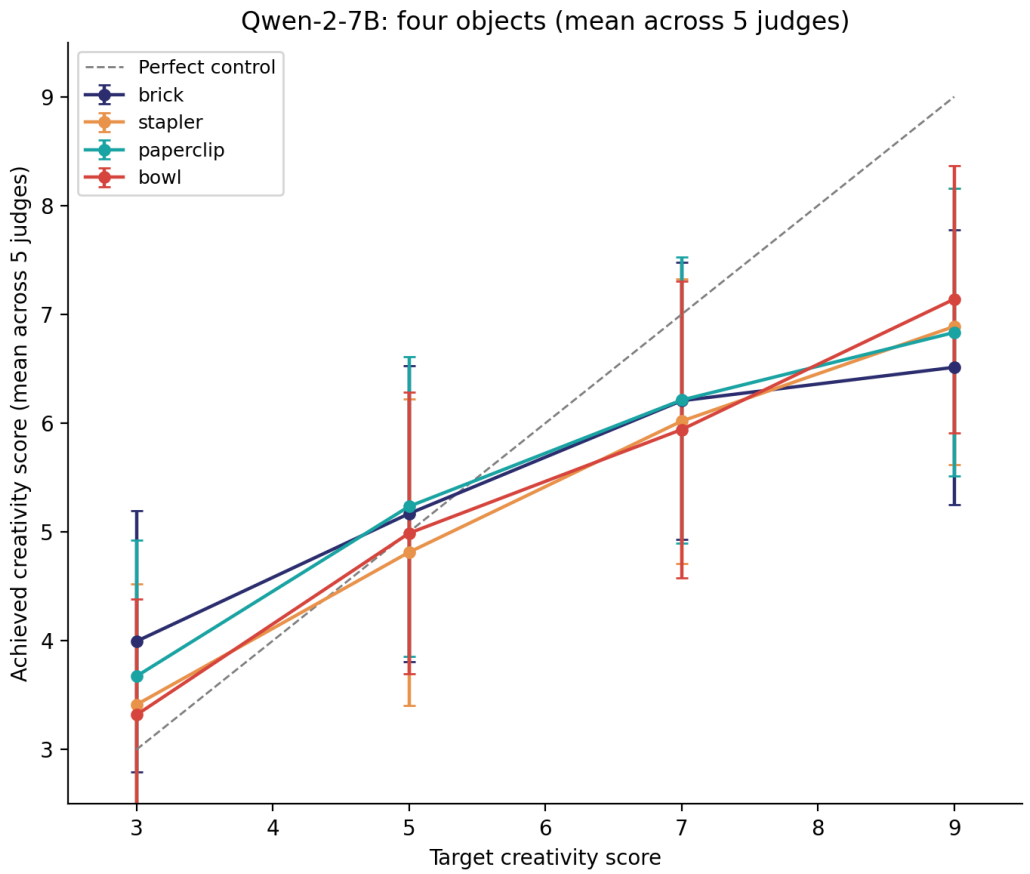
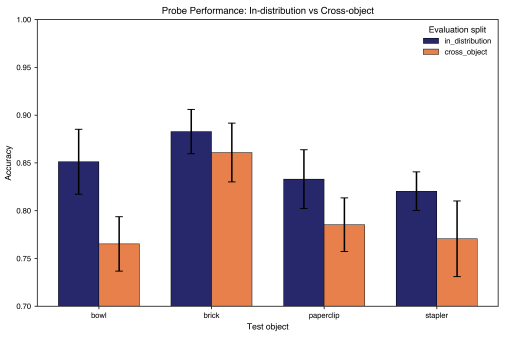
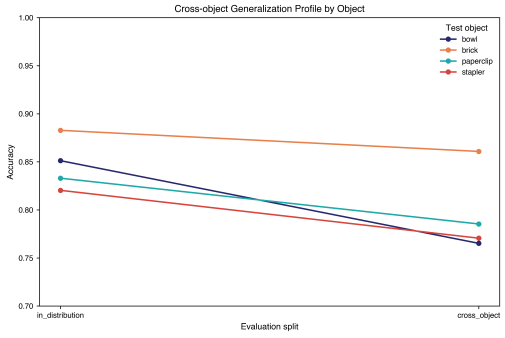


Figure 27: Achieved creativity scores across all four object prompts (brick, stapler, paperclip, bowl) using probe-based steering on Qwen-2-7B-Instruct (layer 17).



(a) In-distribution vs. cross-object probe accuracy (Qwen-2-7B-Instruct, layer 17).



(b) Generalization degradation by test object (Qwen-2-7B-Instruct, layer 17).

Figure 28: Cross-object generalization analysis on Qwen-2-7B-Instruct (layer 17).

F Creativity Evaluation and Human Validation

F.1 Local Confirmation Protocol

Discovery uses 16 responses per source task. A development grid selects the largest norm among 2, 4, and 8 that has at most one structural failure in four responses for every arm; both tasks select norm 8. Selection uses structural validity, not judge outcome. Each task has 17 arms: baseline, four ranked bundles, and 12 random bundles. Confirmation generates 12 responses per arm. Temperature is 0.5, top-p is 1, and the output limit is 768 tokens. All feature IDs, coefficients, seeds, and locks are retained in the source package.

A separate unsteered invocation of the same 70B model rates originality, feasibility, breadth, and coherence on 1–5 scales. There are 926 valid judgments out of 952 attempts. Missing judgments are excluded pairwise rather than assigned zero. Depending on object and condition, paired counts range from 5 to 12. Fourteen outputs fail the five-idea structural check (eight confirmation, six transfer); none is truncated. Almost all outputs miss the requested 40–60 words per idea: only 1/408 confirmation outputs and 0/544 transfer outputs fully meet that requirement. This is separate from the looser structural dose gate, which requires five distinct numbered ideas with at least ten words each. Coherence ratings are all 5. Coarse repeated scores, including zero brick originality contrasts, limit the sensitivity of this pilot.

F.2 Object Transfer and Originality

Brick-discovered bundles transfer to bottle and towel; stapler-discovered bundles transfer to umbrella and backpack. The same fixed intervention is applied to new object prompts, with eight responses per arm and object. All ranked and random arms are retained. Figure 29 reports originality, the direct outcome for the local creativity-steering question, with descriptive paired-bootstrap intervals and no multiplicity-corrected superiority claim. It does not establish a consistent gain across objects.

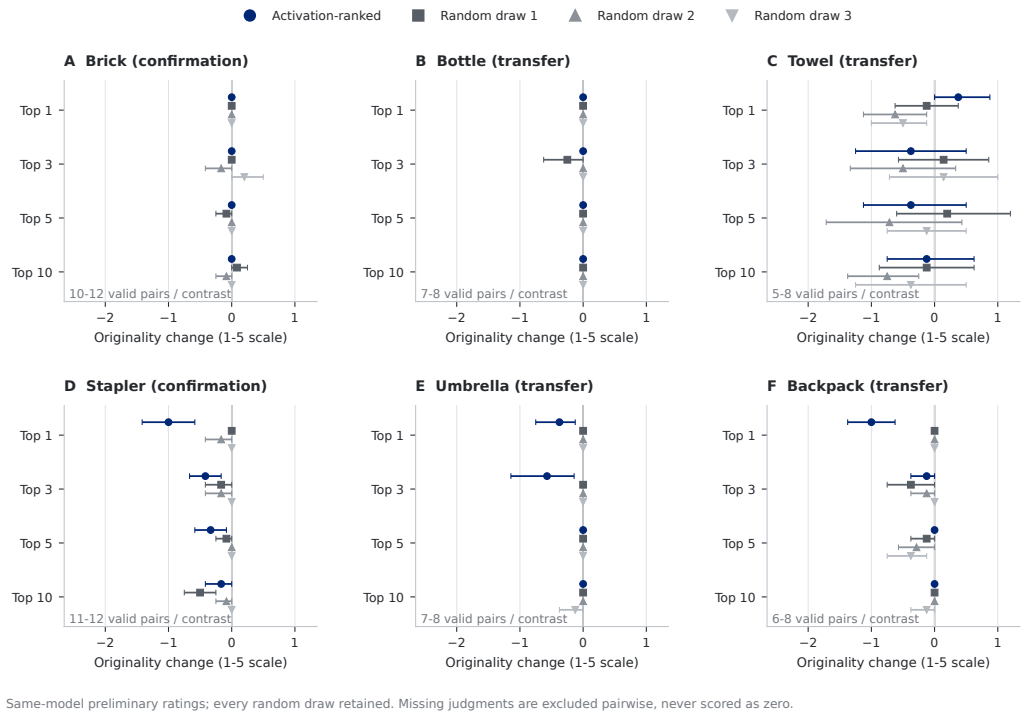


Figure 29: Top row: alternative-use tasks, with discovery on brick and transfer to bottle/towel. Bottom row: improvement tasks, with discovery on stapler and transfer to umbrella/backpack. Each panel includes top-1/3/5/10 ranked bundles and each of three presampled random draws per size.

F.3 Human Participants and Ethical Oversight

The human-evaluation component of this research was determined exempt by the Institutional Review Board of the authors' institution. Identifying institutional and protocol details are withheld for anonymous review.

Participants were recruited through Prolific and completed a rating survey hosted on Qualtrics. The survey included a consent screen explaining the study, the voluntary nature of participation, and the option to stop at any time. Participants evaluated AI-generated creative responses using four dimensions: fluency, flexibility, originality, and elaboration. The task was designed to take approximately 30–40 minutes. Participants received [insert actual compensation and currency].

The analyzed sample comprised 60 retained raters following the exclusion of four respondents, providing 732 assessments of 244 model-generated responses. Human participants served as evaluators of generated text; the behavioral interventions were applied to language models. These ratings assess agreement between human judgments and the LLM evaluation ensemble on the evaluated outputs. They do not cover the new local SAE creativity generations.

Results are reported in aggregate, and participant identifiers are not included in the manuscript.

F.4 Human Evaluation of Older Outputs

The existing Prolific/Qualtrics study retains 60 raters after excluding four, providing 732 item-rater assessments and 2,928 dimension ratings for 244 older outputs. There are 24 outputs with two assessments, 196 with three, and 24 with four (median three). The source-family counts are 82 prompting, 81 SAE, and 81 probe.

Pooled human–ensemble final-score correlation is 0.606 with saved approximate Fisher 95% interval $[0.520, 0.680]$. Within-family correlations and intervals are in Table 4. Interval-scale inter-rater alpha is 0.351 for final score, 0.198 for originality, 0.297 for fluency, 0.293 for flexibility, and 0.441 for

Table 4: Human–ensemble agreement on older outputs. Source families contain multiple conditions; intervals are the saved approximate Fisher 95% intervals.

Source family	Items	Human–ensemble ρ [95% CI]
All outputs	244	0.606 [0.520, 0.680]
Prompting	82	0.194 [-0.024, 0.394]
Historical SAE	81	0.066 [-0.154, 0.280]
Probe	81	0.618 [0.462, 0.737]

elaboration. Human and ensemble scores each correlate with response length (0.669 and 0.666); the reported partial human–ensemble correlation controlling length is 0.290. These associations do not establish that verbosity caused a treatment effect. Participant identifiers are not included in the manuscript source package.

These comparisons assess the LLM evaluation instrument. Across the 17 saved source/condition means, Spearman correlation is 0.836 and Pearson correlation is 0.912. The cells pool objects within each hosted SAE condition and probe target, so this aggregate association is not a task-stratified test of every treatment contrast. Agreement within the SAE family is higher for originality (0.377) than for the composite score (0.066), but neither establishes a human-confirmed steering effect. The correlations justify using ensemble ratings as qualified proxy measurements for broad comparisons; they do not establish interchangeability with human ratings or validity of the new same-model local judge. Figure 30 shows the individual-output and aggregated comparisons together with reliability.

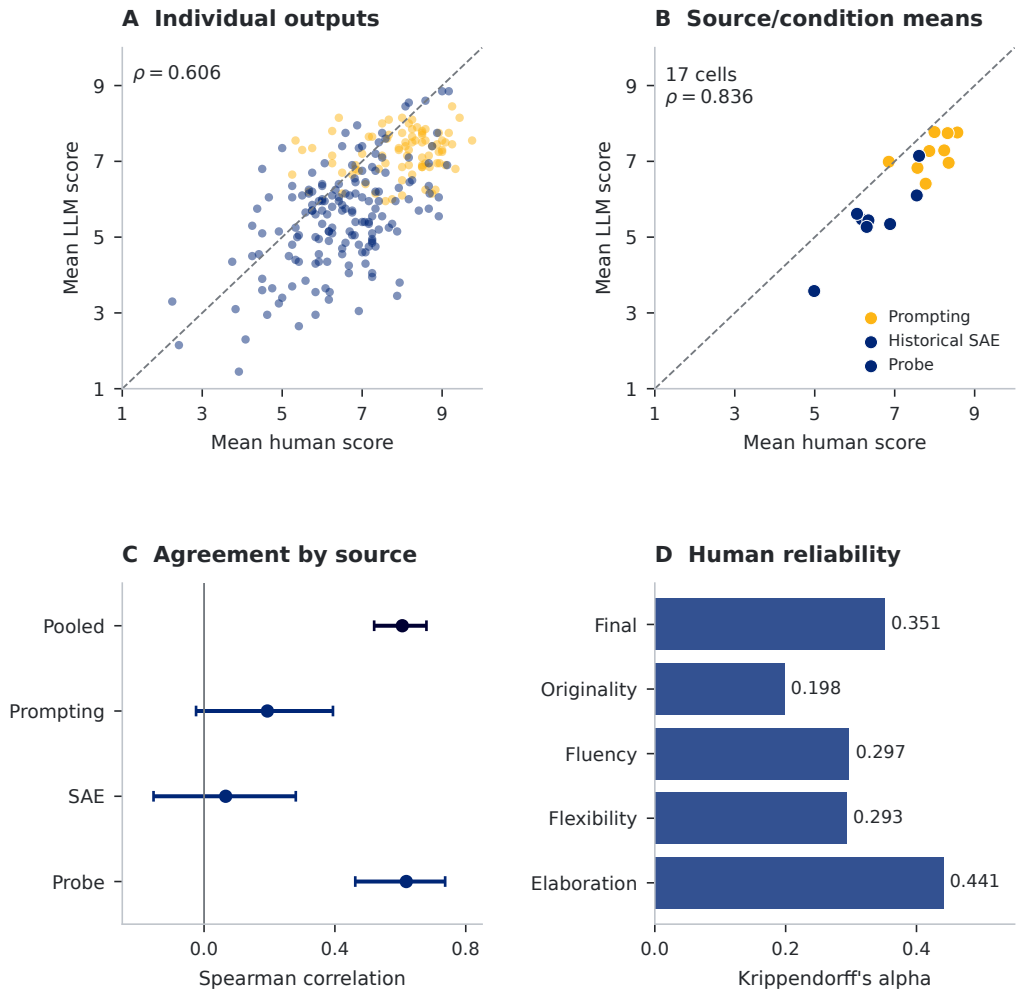


Figure 30: Human validation of the five-judge evaluation on 244 older outputs. A: individual-output means; B: 17 source/condition means, with objects pooled within hosted SAE conditions and probe targets. Dashed lines indicate equal scores. C: final-score Spearman correlations and saved approximate Fisher 95% intervals; D: interval-scale human reliability. The 60 retained raters provide 732 assessments, with 2–4 per output. Only 20 historical SAE-family items received high steering. Aggregate agreement is not validation of every intervention contrast, and these ratings do not cover the new local creativity pilot.

F.5 Independent Ocsai Judge Calibration

On 300 human-authored brick uses from Ocsai (Organisciak et al., 2023), ensemble originality correlates with human originality at 0.426 (95% interval $[0.329, 0.515]$); individual judges range from 0.281 to 0.396. The ensemble correlation is 0.376 after controlling for response length. This separate cohort supports criterion agreement of the originality judge on human-authored uses; it does not validate the new SAE intervention.

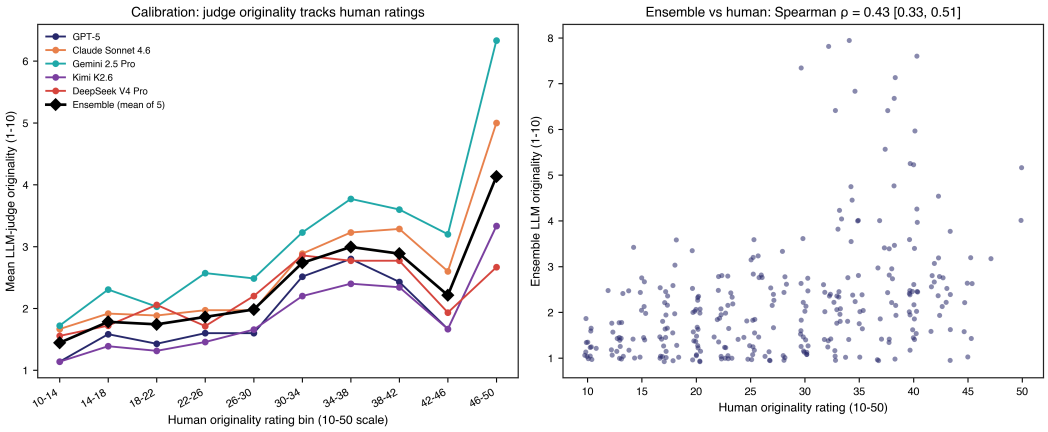


Figure 31: Existing Ocsai judge calibration on 300 human-authored brick uses. This cohort is distinct from the Prolific study and new local SAE outputs.

F.6 Source-stratified Length and Judge Audit

The pooled relationship between final score and log word count differs markedly by source. Across judges, pooled R^2 ranges from 0.505–0.641 for brick and 0.180–0.775 for stapler; within the historical SAE family these ranges are approximately 0.000006–0.038 and 0.035–0.301. These are associations, not length-adjusted causal effects. The composite score must also be distinguished from originality, the target used in the creativity anchor report. The complete condition means and dimension-specific regressions are supplied with the source.

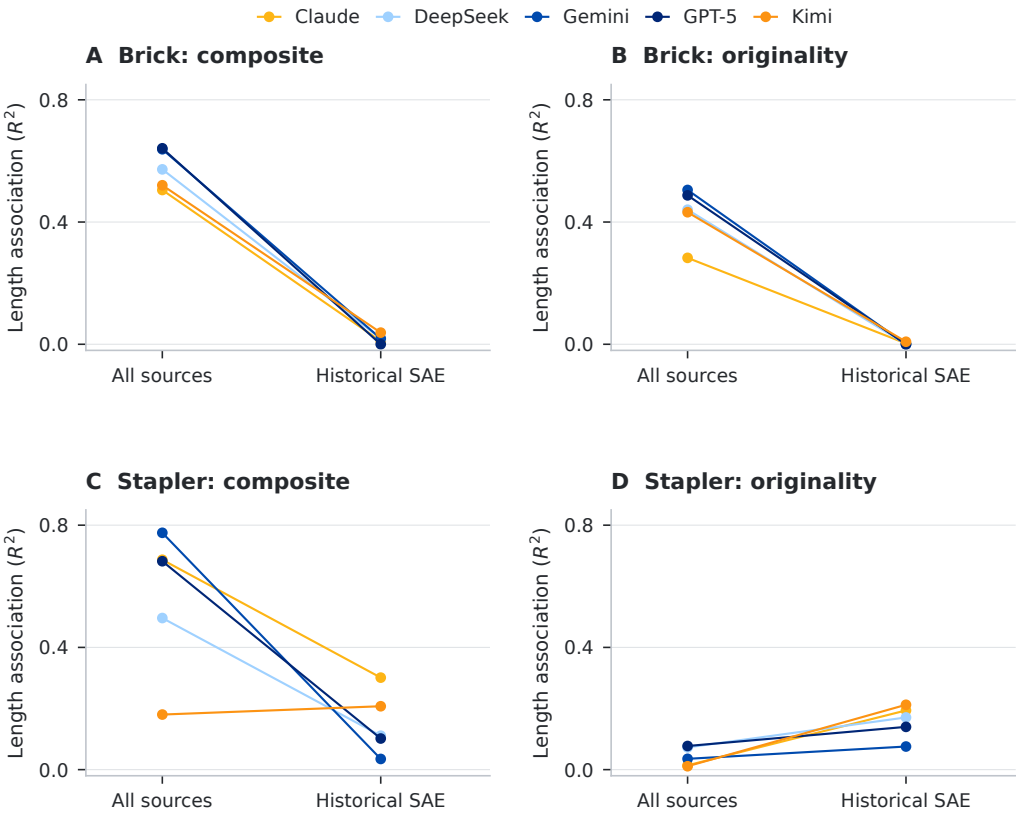


Figure 32: Saved regressions of judge score on log2 word count, shown separately for each judge, task and score. Left points pool all historical sources; right points restrict to the historical SAE experiment family.

G Anchor Experiments and Interpretation Checks

These checks ask whether a direction's behavioral effect resembles a known planted preference change or an answer/readout change. They qualify, rather than automatically legitimize, the interpretation of probes and SAEs. The three supplied reports and their full tables accompany the source. Reported tables below preserve their numbers; availability and calibration limits are stated separately.

G.1 Synthetic Decision Anchors

The synthetic cumulative-prospect-theory reference uses $\alpha = 0.88$, $\gamma = 0.61$, loss-aversion parameter 2.25, and temperature 0.1186. A preference implant changes the value-function parameter; a bias implant adds 0.6 to answer log odds. Adapters trained to reproduce these targets pass the stated absolute cloning gates (correlation at least 0.95; slope 0.9–1.1). Passing absolute prediction does not guarantee recovery of a small contrast. At one minimum detectable change, parameter recovery is close but attribution fails; at three, the reported gate passes. Distances placing probes against anchors are standardized behavioral-summary distances, not cosine distances. These tests distinguish the specified synthetic alternatives, not every possible mechanism.

Table 5: Synthetic reference parameters. Reported values; scope qualifications are given in the text.

parameter	value	what it controls
alpha	0.88	curvature of the value function, v(x) = x^alpha. Lower means large payoffs are discounted more, which is risk aversion.
gamma	0.61	probability weighting. Controls whether stated odds are taken at face value.
lambda	2.25	loss aversion. Only active on trials involving losses.
tau	0.1186	temperature of the choice rule. Controls how sharply the agent acts on small differences.

Table 6: Planted choice-logit contrasts by reward. Reported values; scope qualifications are given in the text.

risky reward	z, reference	change from the preference implant	change from the bias implant
30	-6.169	-0.113	+0.600
50	-4.885	+0.000	+0.600
70	-2.724	+0.207	+0.600
110	-0.666	+0.319	+0.600
150	0.340	+0.334	+0.600
230	1.345	+0.312	+0.600

Table 7: Adapter cloning checks. Reported values; scope qualifications are given in the text.

adapter	correlation	slope	verdict
reference	0.9915	0.9567	passes
preference	0.9938	0.9618	passes
bias	0.9933	0.9359	passes

Table 8: Recovered mean contrasts. Reported values; scope qualifications are given in the text.

Contrast	Planted	Measured	95% interval
preference - reference	0.0302	0.0474	[-0.0291, +0.1239]
bias - reference	0.6000	0.4968	[+0.4178, +0.5759]

Table 9: Preference recovery and attribution gates. Reported values; scope qualifications are given in the text.

Anchor	Planted	Recovered	Error	Attributed to preference	Gate
One times the MDC	0.1000	+0.0874	0.13 MDC	0.268 with belief ahead at 0.346	FAIL
Three times the MDC	0.3000	+0.2987	0.01 MDC	1.000 with action rejected	PASS

Table 10: Behavioral-summary distances from the anchors. Reported values; scope qualifications are given in the text.

Direction	Distance to the bias anchor	Distance to the readout anchor	Placed with
Choice probe at layer 48	0.540	0.679	the bias anchor
Choice probe at layer 50	0.440	0.778	the bias anchor
Difference of means at layer 48	1.211	0.045	the readout anchor
Difference of means at layer 50	1.027	0.187	the readout anchor
Reward probe at layer 48	1.172	0.043	the readout anchor
Reward probe at layer 50	0.346	0.878	the bias anchor
Readout anchor at layer 50	1.227	0.026	the readout anchor

G.2 SAE Geometry, Answer Order, and Weight Anchors

The B8 report uses 432 lottery prompts (108 scenarios, two wordings, two orders) and 120 ultimatum prompts. Lottery answer-order gradient cosines are approximately -0.89 at layers 48 and 50, while ultimatum gradients have cosines above $+0.95$. This difference is confounded with A/B labels versus explicit Accept/Reject answers. A cosine is not a percentage of a mechanism. Likewise, the ratio of summed and differenced gradient norms is not a mechanistic mixture fraction.

The transferred +0.6 bias contrast is reported as -0.022 on lottery and $+0.462$ on ultimatum. Thus a trained anchor's meaning cannot be assumed invariant to prompt format. Direction geometry below is descriptive; successful alignment alone does not establish preference mediation. Shared initialization and training schedules do not remove differential optimization error between adapters.

Table 11: Answer-order gradient geometry. Reported values; scope qualifications are given in the text.

task	answer tokens	cosine between the orders	sum length	difference length	ratio
lottery, layer 48	A / B	-0.889	9.128	37.682	0.24
lottery, layer 50	A / B	-0.893	8.416	35.383	0.24

task	answer tokens	cosine between the orders	sum length	difference length	ratio
ultimatum, layer 48	Accept / Reject	+0.953	18.421	2.869	6.42
ultimatum, layer 50	Accept / Reject	+0.956	17.271	2.614	6.61

Table 12: Transferred adapter contrasts and displacement norms. Reported values; scope qualifications are given in the text.

quantity	planted	lottery prompts	ultimatum prompts
bias contrast, mean	+0.600	-0.022	+0.462
preference contrast, mean	positive	-0.285	+0.331
bias displacement, layer 48		3.073	6.512
preference displacement, layer 48		2.521	5.930

Table 13: Lottery direction geometry. Reported values; scope qualifications are given in the text.

direction	choice part	presentation part	weight push
random feature, 95th percentile	0.033	0.027	0.049
random feature, 99th percentile	0.056	0.048	0.078
feature 184	-0.001	+0.016	-0.005
feature 4237	+0.004	+0.000	+0.004
feature 31935	+0.004	+0.004	+0.017
fitted ten-feature, combined target, seed 31	+0.067	-0.006	+0.031
fitted ten-feature, combined target, seed 73	+0.075	-0.003	-0.013
fitted ten-feature, second target, seed 31	+0.091	+0.002	-0.013
fitted ten-feature, second target, seed 73	+0.096	+0.004	-0.004
sparse reconstruction of the raw readout	+0.269	+0.037	-0.008

Table 14: Ultimatum direction geometry. Reported values; scope qualifications are given in the text.

direction	choice part	presentation part	weight push
random feature, 95th percentile	0.028	0.026	0.077
random feature, 99th percentile	0.038	0.039	0.126
feature 184	+0.022	-0.020	-0.005
feature 4237	-0.004	+0.007	+0.016
feature 31935	+0.013	+0.010	-0.044
fitted ten-feature, combined target, seed 31	-0.049	-0.013	+0.184
fitted ten-feature, combined target, seed 73	-0.049	-0.009	+0.215
fitted ten-feature, second target, seed 31	-0.062	-0.027	+0.226

direction	choice part	presentation part	weight push
fitted ten-feature, second target, seed 73	-0.064	-0.029	+0.220
sparse reconstruction of the raw readout	-0.133	-0.002	+0.091

Table 15: Lottery intervention response. Reported values; scope qualifications are given in the text.

condition	signal retained	correlation
readout gradient	0.062	-0.999
SAE ten-feature, both seeds	0.092 to 0.096	-1.000
dense vector, both seeds	0.129 to 0.143	-0.99
preference adapter	0.156	-0.987
random directions, same size	0.505 to 0.605	-0.95 to -0.97

Table 16: Ultimatum intervention response. Reported values; scope qualifications are given in the text.

direction	strength	shift in mean readout	signal retained
feature 31935	0.20	+0.408	1.032
feature 31935	0.80	+1.814	1.174
readout gradient	0.20	+10.028	3.209
readout gradient	0.80	-1.381	3.574
random live feature	0.20	-0.349	0.974
random live feature	0.80	+0.154	0.977

G.3 SAE Anchor Pilot, Repair, Replay, and Fixed Steering

The earlier replay check distinguishes feasibility of reproducing an adapted hidden state from a deployable fixed steering vector. Full-state LoRA replay exactly reproduces the LoRA risky-choice probabilities; final-token and sparse replay do not meet the declared acceptance rule. For all-position SAE replay, encoder reconstruction has RMSE 0.09711 to LoRA and 0.10487 to the teacher; the 30-feature reconstruction has 0.13801 and 0.15105. No SAE replay passes the RMSE ≤ 0.01 and answer-mass ≥ 0.99 rule. Per-prompt adapted states are oracle information, not a fixed deployable intervention.

The subsequent frozen decision study tests 37 candidate features, both signs and two scopes, then fits sparse and dense vectors with two optimization seeds. Discovery contains 216 prompts; selection contains 96. The frozen test crosses 108 new scenarios with two answer orders and two wording templates (432 prompts per condition), plus 48 comprehension questions. All 28 frozen conditions are retained. Neither dense nor ten-feature SAE steering passes all checks in any of the eight target-by-seed-by-template cells. The LoRA misses transfer to new wording. The operational verdict is inconclusive because of evaluation/anchor transfer, not a universal limit on SAE capacity.

Table 17: Frozen target error under original and new wording. Reported values; scope qualifications are given in the text.

Target	Intervention	Original wording	New wording
combined	baseline	0.364	0.380
combined	lora	0.042	0.115
combined	dense	0.129–0.151	0.224–0.225
combined	sae10	0.202–0.208	0.220–0.224
neutral	baseline	0.386	0.462
neutral	dense	0.1319–0.1323	0.241–0.248

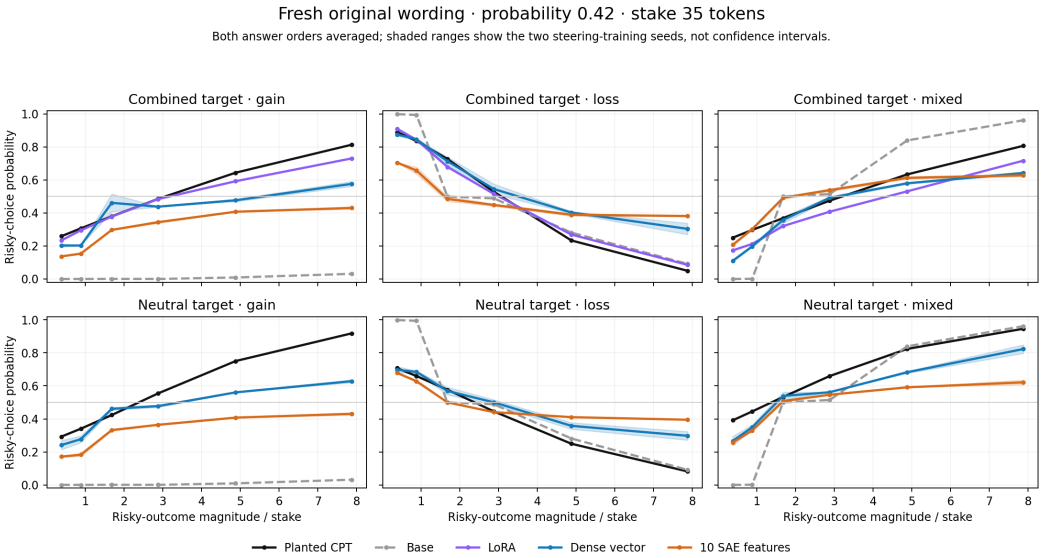


Figure 33: Frozen anchor comparison: original wording. Dense and sparse edits improve some choice predictions but fail the declared recovery checks; ranges across fits do not represent uncertainty intervals.

Target	Intervention	Original wording	New wording
neutral	sae10	0.2035–0.2043	0.2532–0.2533

Table 18: Recovered anchor parameters by wording. Reported values; scope qualifications are given in the text.

Wording	α curvature	γ weighting	λ loss aversion	β temperature	Earlier strict repair rule
original	0.762	0.755	2.100	2.819	Fail
transfer	0.643	0.698	2.447	2.497	Fail

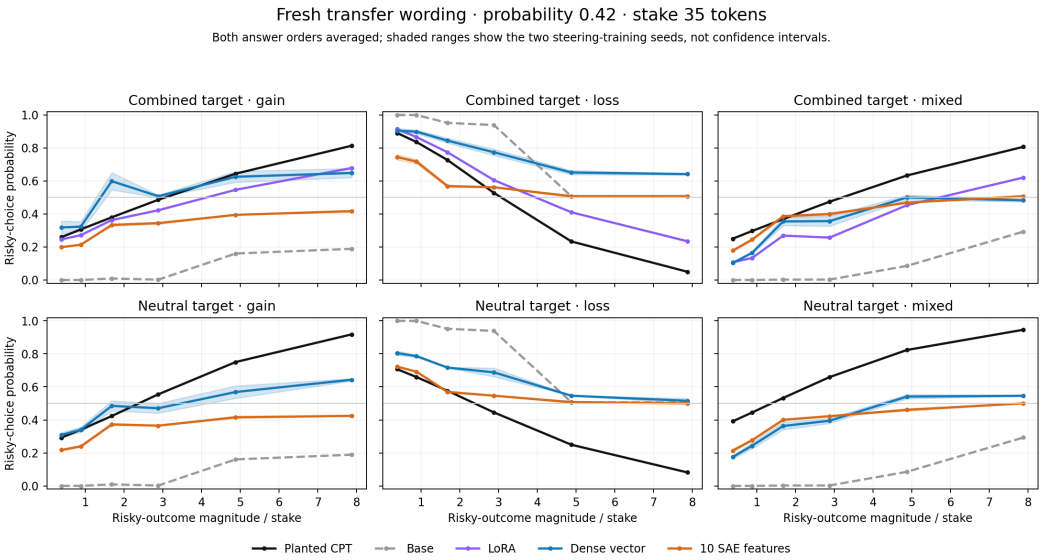


Figure 34: Frozen anchor comparison: new wording. Dense and sparse edits improve some choice predictions but fail the declared recovery checks; ranges across fits do not represent uncertainty intervals.

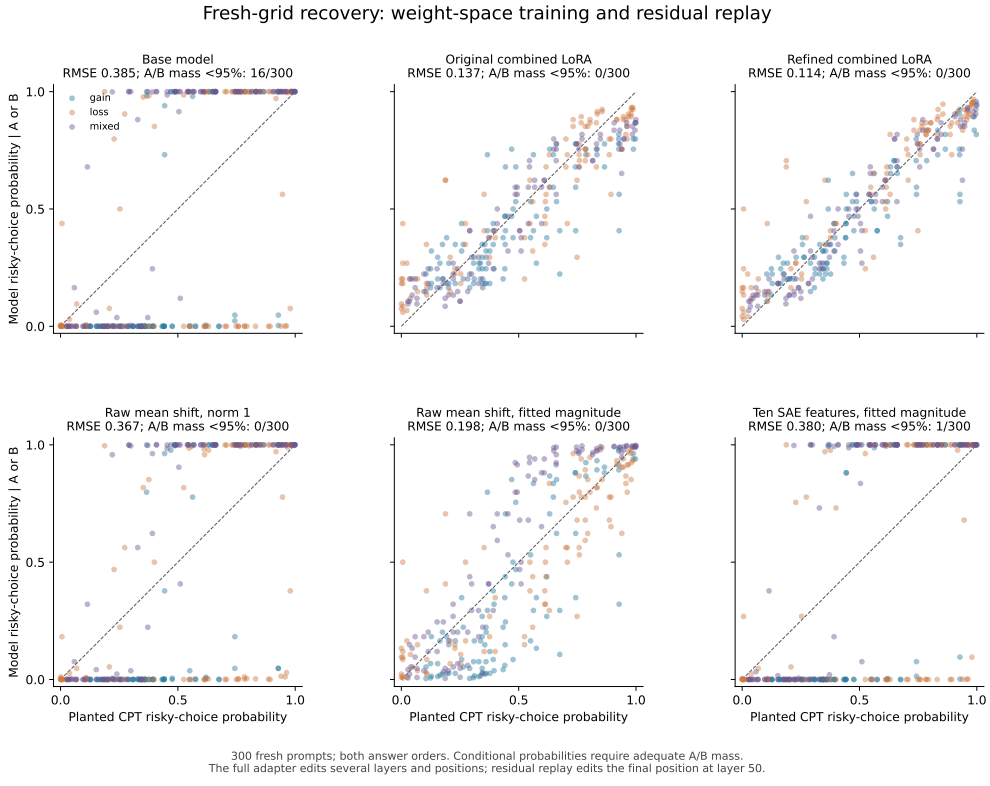


Figure 35: Earlier pilot-stage fresh-grid recovery. These results precede the repaired and frozen decision experiments and are not pooled with them.

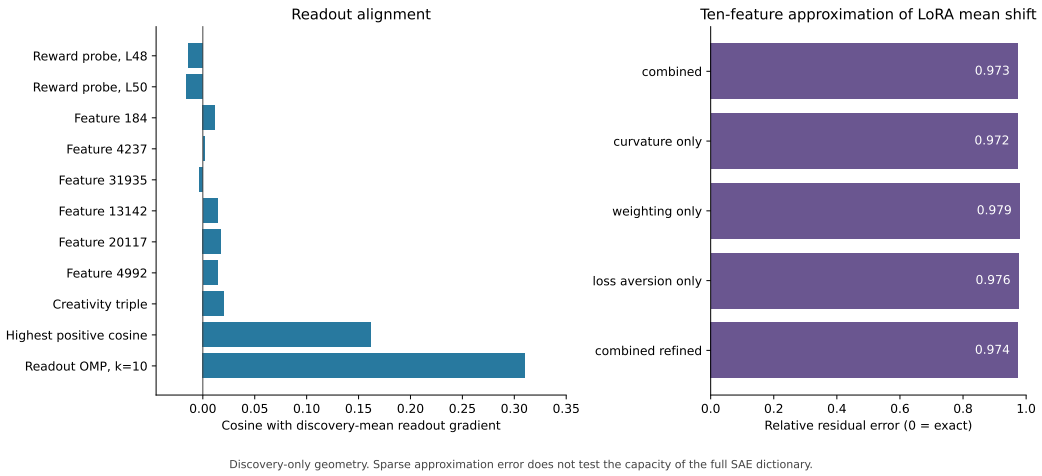


Figure 36: Earlier SAE pilot dictionary geometry. Geometric alignment is descriptive and does not establish recovery of a psychological mechanism.

G.4 Creativity Anchors: Provisional Report and Independent Audit

The B10 report supplies quantitative summaries but the underlying anchor activations, fitted artifacts, and executable experiment outputs were unavailable for independent reproduction. The tables below are therefore provisional reported results. The available historical creativity-rating CSV supports the separate judge/length audit in Appendix F, not verification of the anchor run itself.

Holding out a judge while reusing the same 320 fitted responses is in-sample with respect to responses. The reported 0.846 correlation is not new-response prediction and is not a mathematical upper bound. All three adapters fail both cloning gates: correlations are 0.797–0.810 versus the required 0.95, and slopes 0.759–0.824 versus 0.90–1.10. Approximate intervals from the rounded pair summaries put the mean readout gaps at $[-0.0533, 0.0763]$ for brick and $[-0.0738, 0.0836]$ for stapler, assuming independent pairs. A near-zero gap does not itself invalidate a gradient, but it does not establish reliable agreement with judged originality.

The average bias contrast (0.0553 for a 0.0475 plant) is encouraging descriptively, yet cannot repair failed pairwise cloning. Its approximate interval is $[0.0417, 0.0689]$; the preference contrast's is $[-0.0029, 0.0115]$. These approximations require clustering if pairs share responses. Readout likelihood is a surrogate, not a human creativity score. The reported creativity-feature cosines 0.091–0.107 exceed the preference-anchor 95th-percentile thresholds (0.077/0.074), not the 99th-percentile thresholds (0.122/0.118). A fitted probe is also not drawn from a random-SAE-feature null; these percentiles are not calibrated significance tests of probe specificity. Layer-48 comparisons to a layer-50 SAE are off-layer diagnostics.

Table 19: Provisional B10 report: Paired creativity readout summaries. Reported values; scope qualifications are given in the text.

task	pairs	judged originality gap	mean readout gap	readout spread	direction length, layer 48	layer 50
divergent creativity, brick	42	1.048	+0.0115	0.208	0.00626	0.00602
product innovation, stapler	40	1.835	+0.0049	0.246	0.02103	0.02010

Table 20: Provisional B10 report: Same-response, held-out-judge ridge correlations. Reported values; scope qualifications are given in the text.

held-out judge	agreement, layer 48	agreement, layer 50
Claude Sonnet 4.6	0.896	0.894
Gemini 2.5 Pro	0.865	0.866
Kimi K2.6	0.841	0.838
GPT-5	0.837	0.838
DeepSeek V4 Pro	0.794	0.798
mean	0.846	0.847
direction length	2.992	2.935

Table 21: Provisional B10 report: Failed creativity adapter cloning gates. Reported values; scope qualifications are given in the text.

adapter	cloning correlation	cloning slope	mean readout	training loss
bar to clear	0.95	0.90 to 1.10		
reference	0.810	0.764	+0.556	0.0444
preference	0.797	0.759	+0.561	0.0435
bias	0.798	0.824	+0.612	0.0411

Table 22: Provisional B10 report: Reported adapter contrasts and displacement norms. Reported values; scope qualifications are given in the text.

quantity	planted	measured	measured spread	recovered
bias contrast, mean	+0.0475	+0.0553	0.0619	116%
preference contrast, mean	+0.0024	+0.0043	0.0327	179%
bias displacement, layer 48		1.543		
bias displacement, layer 50		1.657		
preference displacement, layer 48		0.767		
preference displacement, layer 50		0.799		

Table 23: Provisional B10 report: Displacement norms across tasks. Reported values; scope qualifications are given in the text.

task	bias displacement, layer 48	layer 50
lottery	3.073	3.172
ultimatum responder	6.512	6.756
creativity	1.543	1.657

Table 24: Provisional B10 report: Readout-direction cosine comparisons. Reported values; scope qualifications are given in the text.

direction	brick, layer 48	brick, layer 50	stapler, layer 48	stapler, layer 50
random feature, 95th percentile	0.025	0.025	0.027	0.026
random feature, 99th percentile	0.036	0.036	0.039	0.039
feature 13142	+0.001	+0.001	-0.008	-0.004
feature 20117	+0.001	+0.002	-0.011	-0.012
feature 4992	-0.004	-0.005	+0.002	+0.003
creativity probe, layer 48	+0.016	+0.019	+0.031	+0.030
creativity probe, layer 50	+0.012	+0.014	+0.033	+0.032

Table 25: Provisional B10 report: Weight-anchor cosine comparisons. Reported values; scope qualifications are given in the text.

direction	bias, layer 48	bias, layer 50	preference, layer 48	preference, layer 50
random feature, 95th percentile	0.151	0.156	0.077	0.074
random feature, 99th percentile	0.249	0.258	0.122	0.118
feature 13142	+0.024	+0.026	+0.091	+0.093
feature 20117	-0.043	-0.044	+0.107	+0.104
feature 4992	-0.052	-0.055	+0.099	+0.098
creativity probe, layer 48	+0.052	+0.054	+0.021	+0.023
creativity probe, layer 50	+0.050	+0.053	+0.015	+0.018

Table 26: Provisional B10 report: Creativity features against decision anchors. Reported values; scope qualifications are given in the text.

feature	lottery choice	lottery presentation	lottery push	ultimatum choice	ultimatum push
random feature, 95th percentile	0.033	0.027	0.049	0.028	0.077
13142	+0.011	+0.003	-0.008	+0.011	-0.009
20117	+0.011	+0.009	-0.006	+0.003	+0.012
4992	+0.011	+0.016	-0.028	+0.012	+0.020

Table 27: Provisional B10 report: Pooled-source composite-score length associations. Reported values; scope qualifications are given in the text.

quantity	divergent creativity, brick	product innovation, stapler
variance explained by length, lowest judge	0.507	0.180
variance explained by length, highest judge	0.641	0.774
score change per doubling of length, lowest judge	0.72	0.23
score change per doubling of length, highest judge	1.60	0.78

Table 28: Provisional B10 report: Judge agreement by grouping. Reported values; scope qualifications are given in the text.

grouping	lowest cell	highest cell
as reported	0.020	0.654
with length removed	0.045	0.657
the paper's figure, pooled over conditions		0.770