A STEERABLE INTERNAL DIRECTION LINKS DISTINCT MENTAL-ACCOUNTING BIASES IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) increasingly stand in for human participants in economic and behavioral research, and reproduce many human decision-making biases in their answers. But matching an answer does not reveal whether a model carries the underlying concept or only imitates familiar patterns, which determines how far such synthetic respondents can be trusted. We probe this using mental accounting, the tendency to treat money as non-fungible, valued differently by its source or use, which behavioral economists have cataloged as several separate biases: segregating gains, integrating losses, the sunk-cost effect, and transaction utility. Rather than score answers alone, we read and edit the models' internal activations, the numerical state a model builds while processing a prompt. In LLaMA-3-70B, the three categorical biases share a common internal direction: reading it from the model and then adding or subtracting it shifts economic choices as predicted, while a matched random direction does not, showing the representation is causal, not merely correlated with behavior. A related transaction-utility direction widens the venue-dependent willingness-to-pay gap. A controllable de-biasing direction recurs across independently built large models (Qwen2.5-72B, Gemma-2-27B) but is largely absent in three smaller ones, which still show the biases yet offer little steerable control. Controllability is more evident at larger scale in the models tested. The results reveal a mechanism for why high-capacity LLMs replicate economic behavior, and a standard for auditing them as synthetic respondents: match not only their choices, but whether the structure behind those choices is stable and controllable.

1 INTRODUCTION

Mental accounting is one of the most robust ways human economic choices defy classical rationality: people treat otherwise interchangeable money differently depending on where it came from or what it is "for" (Thaler, 1985; 1999). Large language models (LLMs) now reproduce many such human decision tendencies in their outputs (Binz & Schulz, 2023)—economic rationality (Chen et al., 2023), risk and time preferences (Goli & Singh, 2024), social considerations (Mei et al., 2024; Leng & Yuan, 2026), aspects of theory of mind (Kosinski, 2024)—fluently enough that they are increasingly used as synthetic respondents in behavioral studies (Park et al., 2023; Gao et al., 2023; Argyle et al., 2023) and surveys of what people prefer (Kim & Lee, 2023; Dillion et al., 2023). That fluency raises a mechanistic question their outputs cannot answer: when a model reproduces a bias like mental accounting, is there a compact internal representation that helps explain the replication?

The difficulty is that the two possibilities can look identical from the outside. A model that has learned a reusable mental-accounting pattern and a model that only recognizes familiar examples can give the same answers on the tested scenarios. The difference appears only when the task changes. Output resemblance on a fixed set of scenarios is therefore a fragile basis for treating a model as a synthetic respondent.

To tell these possibilities apart, we need a stronger test than asking whether the model gives the canonical answer. A reusable internal representation should leave two signatures. First, it should generalize: the same pattern should help predict responses when the names, amounts, contexts, and

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.

wording change. Second, it should be causal: changing that pattern inside the model should change the model's choices. We therefore examine the model's internal numerical state while it processes a prompt, and ask whether mental accounting leaves a stable pattern that can be reused across prompts and edited during generation.

Mental accounting is a demanding behavior because it covers distinct phenomena, such as segregating gains, integrating losses, the sunk-cost effect, and transaction utility (Thaler, 1999; Tversky & Kahneman, 1981). A model could reproduce these behaviors through separate, scenario-specific routines, with one pattern for lottery gains, another for losses, another for sunk costs, and so on. A more economical possibility is that several of these biases draw on a common internal axis running from fungible to mental-accounting-consistent treatment of money—shared substantially, though not identically, across biases. If so, a direction estimated from one bias should help move the others.

To estimate that direction, we contrast the model's internal state under a mental-accounting frame—money tied to its source and purpose—with its state under a fungibility frame, where every dollar is interchangeable (Zou et al., 2023), and take the difference between the two as the candidate direction.[1] We then add a measured amount of it back into the activations—a small, deliberate nudge—and test whether changing that axis changes the behavior. In LLaMA-3-70B, this direction turns out to exist. Dialing it up or down moves the three categorical behaviors in the predicted direction and shifts the transaction-utility willingness-to-pay pattern under the analogous signed intervention. A direction found from one categorical bias also moves the others, indicating a shared internal component. A matched random direction leaves behavior unchanged, and unrelated general-knowledge performance is preserved over the intervention range, so the effect is specific to this direction.

Repeating the intervention across model sizes and families shows where this internal structure is present. In Qwen2.5-72B and Gemma-2-27B, the bias-reducing component reappears. In the three small models, by contrast, output bias and readable probe signals coexist with little or no controllable structure under this intervention. Output agreement and internal control therefore separate in the models tested, with controllability tracking model scale across the three families.

These results turn a familiar behavioral-economics pattern into an internal structure: mental accounting has a compact, controllable representation in activation space, most completely in LLaMA-3-70B. This gives substance to the idea that large models can replicate economic behavior because some behavioral concepts are organized this way. The structure emerges with scale and recurs across independently pretrained families. As these models are increasingly used as synthetic respondents (Bail, 2024; Abdurahman et al., 2024), the audit standard is not only whether they reproduce choices, but whether the structure behind those choices is stable and controllable.

2 Mental-Accounting Tasks and Behavioral Baselines

Mental accounting is the set of behavioral processes by which people categorize and evaluate money by its source or use rather than treating it as fungible, so their choices deviate from the context-free optimum (Thaler, 1985; 1999; Prelec & Loewenstein, 1998).

We examine four canonical mental-accounting biases—segregating gains, integrating losses, sunk-cost effects, and transaction utility (Thaler, 1999; 1985; Tversky & Kahneman, 1981)—in a six-model capacity $\times$ family design: two sizes each of LLaMA-3 (70B and 8B) (Dubey et al., 2024), Qwen2.5 (72B and 3B) (Qwen Team, 2024), and Gemma-2 (27B and 2B) (Gemma Team, 2024). The design divides labor across the six models: LLaMA-3-70B, in which all dimensions are independently steerable, carries the full within-model analysis; Qwen2.5-72B and Gemma-2-27B test whether the de-biasing direction recurs across independently pretrained families; and the three smaller models test whether readable structure dissociates from controllable structure.

The four biases are familiar from decades of human studies:

[1] This contrastive-direction approach has located steerable representations of single concepts—truthfulness (Li et al., 2023), political identity (Kim et al., 2025), space and time (Gurnee & Tegmark, 2024). Mental accounting raises a different question: whether several behaviors economics treats as separate share one such direction, and whether controlling that direction—rather than only reading it—requires a large model.

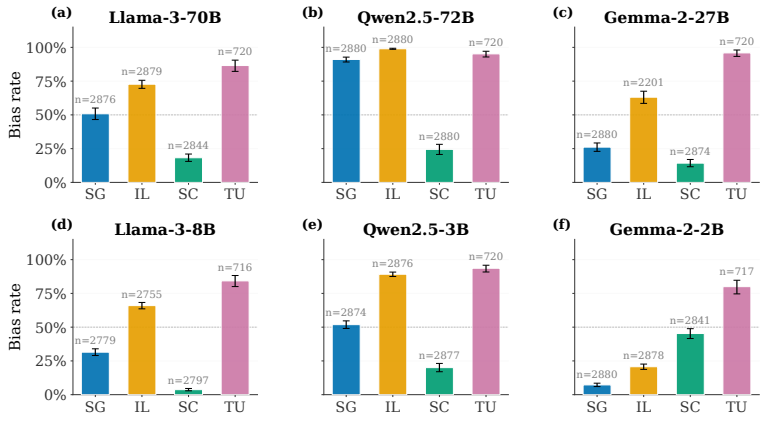


Figure 1: Behavioral baselines across six models in a 3×2 capacity $\times$ family grid (the three large models on top, the three small below), each pooled over a diverse bank of $1{,}440$ vignettes (eight stochastic draws each) so they do not hinge on any single wording. For segregating gains (SG), integrating losses (IL), and sunk cost (SC), bars show mental-accounting-consistent response rates; for transaction utility (TU), bars show the probability of higher willingness-to-pay at a premium venue for the same item. The dashed line marks 50%, and error bars are bootstrap 95% confidence intervals over vignettes or matched TU scenarios.

- **Segregating gains.** People derive more satisfaction from two separate gains than from one combined gain of equal total—winning \$50 and \$25 in separate draws feels better than \$75 at once.
- **Integrating losses.** The mirror image for losses: a single \$150 bill stings less than separate \$100 and \$50 bills.
- **Sunk-cost effect.** Past, unrecoverable spending sways present choices—having lost a \$10 ticket, people are less willing to buy a replacement, as if the evening now costs \$20.
- **Transaction utility.** A purchase is judged against a context-dependent reference price—the same beer feels worth more bought at a resort than at a run-down store, so people will pay more for it there.

Figure 1 reports each model's baseline bias rate, sampled with decoding held fixed across models (temperature 0.6, top-p 0.9). Each dimension is elicited with multiple wordings—five randomized variants for the three categorical biases, ten paired premium/ordinary scenarios for transaction utility (*SI Appendix*, Experimental Setups)—and the displayed baselines pool a larger diverse vignette bank ($1{,}440$ vignettes, eight stochastic draws each) so they do not depend on the smaller prompt sets used later for steering.

At baseline the six models behave much alike (Figure 1): they share one qualitative profile, and the small models reproduce much of what the large ones do. They differ mainly in level—the Qwen models sit highest, with several biases near ceiling. The one exception is sunk cost, where every model is close to rational, perhaps because it is the one bias with a widely taught corrective rule ("ignore sunk costs") that models may have absorbed directly.

This behavioral similarity is the difficulty: output alone cannot say whether a model that produces the bias carries it as a steerable internal representation—one that can be located and moved—or only reproduces it at the output. We therefore turn from behavior to mechanism.

3 Locating and Steering the Mental-Accounting Direction

Transformer LLMs build up a shared internal "residual stream" of activations, updating it layer by layer as they process a prompt (Vaswani, 2017; Elhage et al., 2021; Li et al., 2023). We locate where each bias is encoded in this stream, then test whether those representations are causal by editing it directly. An edit does two things at once. It separates a model that internally represents mental

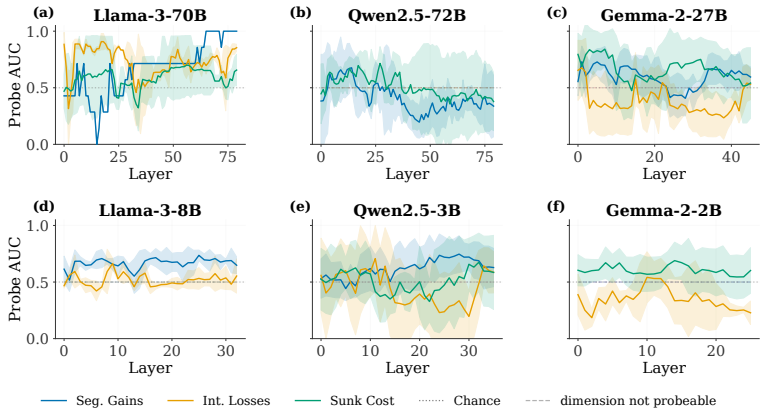


Figure 2: Layer-wise linear probing AUC for the mental-accounting dimensions, with panels in a 3×2 capacity $\times$ family grid (the three large models on top, the three small below). Shaded bands show cross-validation variance, and dashed lines mark dimensions without usable categorical labels (transaction utility throughout, and segregating gains in Gemma-2-2B, which collapses to a single response class). The main pattern is a dissociation between decodability and controllability: small models can contain readable signals even when steering later fails to move behavior.

accounting from one that only reproduces it at the output—a probe can be fooled by the model re-encoding the choice it is about to make, but a steering edit that changes the choice cannot. And it makes the bias controllable: the same mental-accounting direction can be turned down to de-bias the model, or used to audit whether a synthetic respondent carries the bias it displays.

3.1 Locating Predictive Layers

We now ask where the behaviors in Figure 1 are represented. Probing maps where each bias is decodable; a companion layer-resolved projection of the steering vector (*SI Appendix*, Layer-Resolved Projection Analysis) sets the per-dimension intervention bands, so we perturb only the layers where the direction is expressed rather than the whole network.

We use linear probing to ask, at each layer, whether the model's internal state already predicts its own mental-accounting choice (Belinkov, 2022; Gurnee & Tegmark, 2024), and report the held-out classification accuracy (AUC) per layer (Materials and Methods). Transaction utility is elicited as a continuous willingness-to-pay (WTP) rather than a choice, so it enters the steering analyses and is excluded from the probes. A head-level decomposition of the same signal is reported in *SI Appendix*, Head-Level Decomposition.

Figure 2 reports per-dimension probing AUC across layers and shows a dissociation between decodability and control. In the large models, the categorical biases are readable at distinct depths: segregating gains in upper layers, sunk cost in mid-network layers, and integrating losses in a bimodal profile that flips polarity between mid and upper layers (confirmed causally; *SI Appendix*, Layer-Resolved Projection Analysis). The near-perfect segregating-gains AUC in LLaMA-3-70B reflects near-deterministic SG choices on the sweep families: the probe predicts a near-constant label. The small models also contain readable signal, yet steering later moves little or none of it. That a probe can read the bias from the activations shows the information is present, not that the model uses it. The stronger test is causal: whether moving that direction changes the model's choices.

3.2 Identifying the Mental Accounting–Fungibility Direction

Linear probing identifies heads and layers whose activations correlate with mental-accounting choices, but correlation is not cause. Because the probe's labels are the model's own choices, a high AUC can reflect the model re-encoding the answer it is about to give; only an intervention that changes the choice establishes cause. We therefore intervene on these activations directly. The premise is a single internal axis: mental accounting and fungibility occupy distinct regions of the model's activation

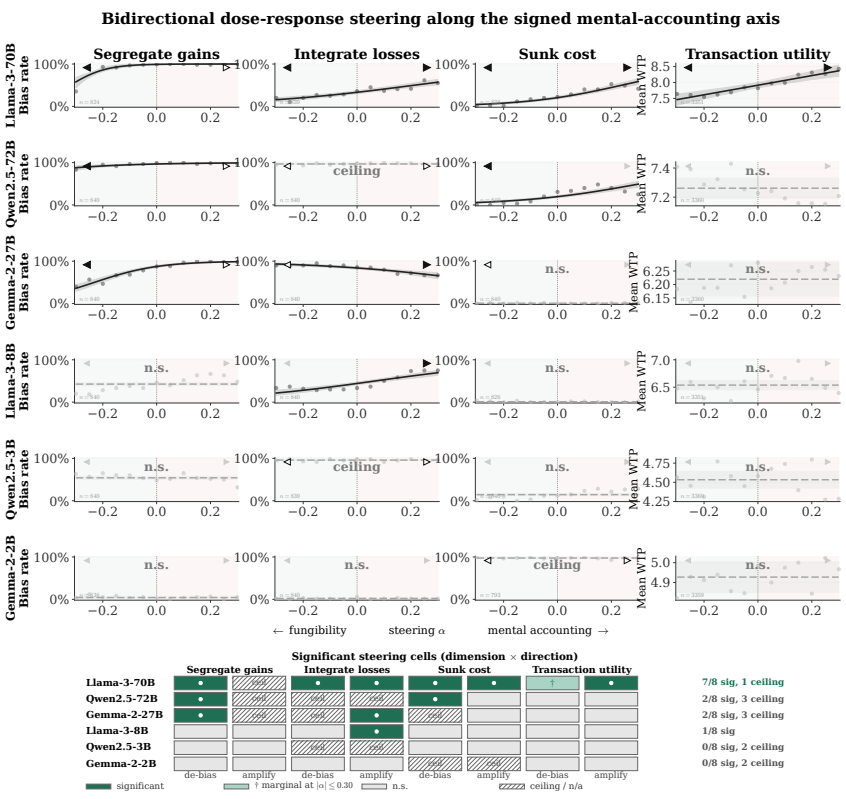


Figure 3: **Bidirectional dose-response steering along the signed mental-accounting axis.** Rows are models grouped by capacity, and columns are dimensions. The horizontal axis is signed steering strength from fungibility ($-\alpha$) to mental accounting ($+\alpha$); the vertical axis is bias rate or mean WTP. Solid panels indicate significant dose-response, dashed panels null effects, and triangles mark significant, null, ceiling/floor-blocked, or not-applicable arms. The scorecard counts eight signed arms per model (three categorical dimensions and transaction utility, each in de-biasing and amplifying directions). Controllability is more evident in the large models under this linear intervention. Transaction utility steers only in LLaMA-3-70B among the five models with a TU steering panel; within LLaMA-3-70B its de-biasing arm is significant over the full sweep but marginal at the primary range.

space, and we read the direction between them straight from the activations rather than imposing it by hand.

We read the direction from the model itself, contrasting its internal state on paired prompts that elicit mental-accounting versus fungibility behavior and taking the mean difference of that contrast, layer by layer, as the steering vector (following Rimsky et al., 2024; Turner et al., 2024; Singh et al., 2024; Materials and Methods).

3.3 Steering the Direction Shifts Behavior

To test whether this direction is causal, we add it to or subtract it from the residual stream during generation, at the predictive layers the probes identify, scaling the edit by a steering strength α. Adding the vector pushes the model toward mental accounting; subtracting it pushes toward treating money as interchangeable. A graded, monotonic response as α grows—absent for a random direction of the same size—would show that editing this axis changes the choice itself.

We summarize each dose-response by the Spearman correlation ρ between the steering strength α and the mean response (bias rate, or WTP for transaction utility), with ρ near ± 1 marking a clean monotonic shift, and test significance with a per-response logistic regression of the mental-accounting indicator on α (linear regression for WTP), Benjamini–Hochberg corrected; the stars in Figure 3

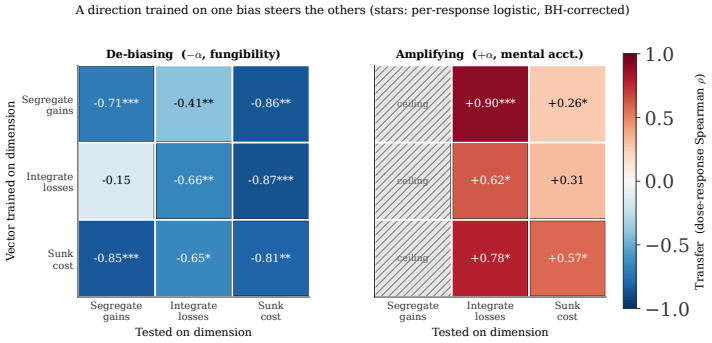


Figure 4: **The categorical choice biases share a substantial steering component.** In LLaMA-3-70B, a direction estimated from the row dimension is applied to the column dimension over the same signed α grid as Figure 3 (both directions; $n = 201$–480 per cell). Cell color and the printed value give the dose-response Spearman ρ; stars mark Benjamini–Hochberg–corrected significance of the per-response logistic test (the same test as Figure 3; read the stars, not ρ, for significance). Diagonal (boxed) cells are within-dimension effects and off-diagonal cells test transfer; hatched cells mark segregating-gains targets at the response ceiling under amplification, where transfer cannot be assessed. Full per-cell statistics (n, β, q), the Spearman-vs-logistic power note, and the separate categorical-vector→WTP test for transaction utility are in *SI Appendix*, Cross-Dimension Transfer Matrix.

($^*q < 0.05$, $^{**}q < 0.01$, $^{***}q < 0.001$) report that test. The intervals reflect the model's response stochasticity rather than sampling over human subjects. In LLaMA-3-70B, seven of the eight signed arms are significant over the full sweep—five categorical choice arms and two transaction-utility WTP arms—with segregating-gains amplification ceiling-blocked. One of the seven, transaction-utility de-biasing, is marginal within the primary range ($\rho = -0.71$, $p = 0.07$ at $|\alpha| \leq 0.30$). On the categorical dimensions alone, five of six dimension $\times$ direction effects are significant (*SI Appendix*, Logistic Regression Results).

3.3.1 Steering Along the Signed Axis

We first trace the full dose-response on LLaMA-3-70B, the model with a measurable effect on every dimension. In Figure 3 (top row), increasing α moves the model from fungibility toward mental accounting: the categorical bias rates shift in the expected direction, and WTP rises or falls with the signed intervention. The intervention therefore moves several behaviors that the baseline output alone treats as separate.

Across the other five models, controllability is more evident at larger scale than tied to one family (Figure 3). The categorical de-biasing direction reappears in both additional large models, Qwen2.5-72B and Gemma-2-27B, on dimensions where the baseline leaves room to move (the full per-dimension Qwen2.5-72B replication is in *SI Appendix*, Cross-Family Replication). Transaction utility is more model-specific: under the same paired-venue WTP paradigm, only LLaMA-3-70B shows a steerable transaction-utility gap.

None of the three small models steers in any comparable multi-dimension way. Under the same sweep protocol and comparable per-arm sample sizes, LLaMA-3-8B moves one bias in one direction while Qwen2.5-3B and Gemma-2-2B move none, even though linear probes still read bias-relevant information from them. With all three families tested at two scales, the pattern separates readable signal from controllable structure in these models; the lone exception—LLaMA-3-8B's single significant steering arm—is reported in *SI Appendix*, Per-Model Steering Statistics.

3.3.2 A Shared Component Across Categorical Biases

Steering moves each behavior on its own; Figure 4 asks whether the three categorical biases share a common component or instead rely on three parallel controls. They are distinct biases, kept separate in economics for decades, so shared control is informative. Because cross-dimension transfer can

only be measured where several dimensions are independently steerable, we estimate the transfer matrix in LLaMA-3-70B, the only such model in our set; the gain–loss vector geometry beneath it recurs across all six models.

A vector derived from one categorical dimension often moves the others in the same signed direction. Under the per-response logistic test, the transfer is significant after Benjamini–Hochberg correction in 13 of the 15 assessable direction $\times$ target arms. The two exceptions are different: IL$\rightarrow$SG de-biasing is near the SG response ceiling, while IL$\rightarrow$SC amplification is a true weak-transfer cell. De-biasing transfers substantially off the diagonal, and amplification transfers when the target dimension has room to move. The geometry agrees where it can be checked directly: the segregating-gains and integrating-losses steering vectors are nearly collinear in every model we test (peak cosine 0.97–0.99 across all six; *SI Appendix*, Steering-Vector Geometry Across Models), and the upper-layer dimensions share one late-layer attention-head cluster (*SI Appendix*, Head-Level Decomposition). The sunk-cost vector, carried at mid layers, is not collinear with the gain–loss direction; it joins the shared component through the transfer itself—each vector moves the other's behavior—rather than through vector alignment. The transfer is therefore strong but not identical across biases.

The meaningful separation is layer-wise: sunk cost is steered at middle layers, the gain and loss framings at upper layers. A shared component here therefore means shared control expressed in layer-specific bands—one direction for the gain and loss framings, a functionally coupled mid-layer direction for sunk cost—rather than one fixed vector at every depth. The integrate-losses direction even reverses sign between mid and upper layers, a bimodal encoding we confirm causally in LLaMA-3-70B (*SI Appendix*, Layer-Resolved Projection Analysis) and that is consistent with Gemma-2-27B's downward integrate-losses panel under amplification. The steering component is shared across biases with layer-specific organization.

In LLaMA-3-70B the shared axis reaches transaction utility, but only in part. Applying each categorical steering vector to the transaction-utility WTP sweep moves the overall WTP level in the signed direction—mental-accounting steering raises it, fungibility steering lowers it—with five of six arms significant (*SI Appendix*, Cross-Dimension Transfer Matrix). Transaction utility's WTP level therefore loads on the same de-biasing direction as the three categorical biases; its defining venue-dependent premium$-$ordinary gap, however, is carried by transaction utility's own direction rather than the categorical one (next subsection).

3.4 Transaction Utility (Continuous WTP)

We follow Thaler's (1985) original numeric WTP elicitation ("What is the maximum you would pay for [item] at [venue]?"), pairing each of 10 scenarios with a premium venue (resort beach bar, airport kiosk, stadium concession) and an ordinary venue (grocery store, corner deli). The transaction-utility gap (premium minus ordinary WTP) is the primary outcome.

Steering moves WTP in both directions, but the quantity that defines transaction utility is the premium$-$ordinary gap. In LLaMA-3-70B, mental-accounting steering widens this gap monotonically; fungibility steering lowers WTP—though only suggestively within the primary range ($\rho = -0.71$, $p = 0.07$ at $|\alpha| \leq 0.30$; significant over the full sweep)—with a noisier gap response. Movement in the gap under amplification, beyond the level shift, shows that the model encodes venue-dependent transaction utility. The transaction-utility evidence therefore rests on the amplification arm and the gap. The venue-decomposed curves are in *SI Appendix*, Transaction Utility.

We ran the same paired-venue WTP sweep on the other four models with a TU steering panel. None reproduces LLaMA-3-70B's transaction-utility effect: the premium$-$ordinary gap is flat at the primary intervention range. The contrast is sharpest for Qwen2.5-72B and Gemma-2-27B, which replicate categorical de-biasing yet lack a steerable transaction-utility gap. Across models, then, transaction utility is the least portable part of the representation (*SI Appendix*, Cross-Dimension Transfer Matrix). Cross-bias sharing within a model and cross-model portability are distinct: the first holds for transaction utility's WTP level, the second does not.

3.5 Robustness Controls

The effect is specific to the steering direction. Replacing the real steering vector with shuffled vectors of matched norm eliminates the dose-response effects in the categorical LLaMA-3-70B

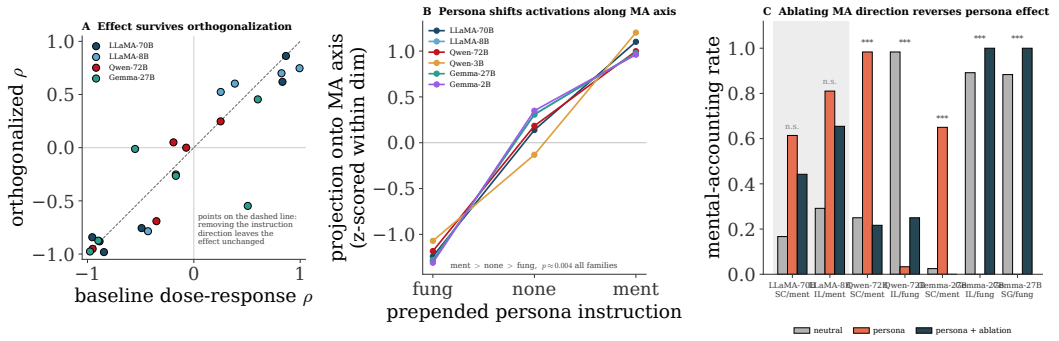


Figure 5: **The steering direction is mental-accounting–specific rather than generic instruction-following.** **(A)** Orthogonalizing away a generic instruction-following direction leaves the dose-response nearly unchanged. **(B)** Persona prompts shift activations along the mental-accounting direction in the order fungibility $<$ none $<$ mental accounting; the non-steerable small model Qwen2.5-3B appears only in this panel. **(C)** Ablating the direction during generation reverses persona-induced behavior in the testable cells. Together, the controls show that the direction carries mental-accounting content rather than a generic tendency to follow instructions.

conditions (*SI Appendix*, Random-Vector Control). The shifts in Figure 3 therefore reflect the content of the mental-accounting direction rather than the magnitude of the perturbation; per-model statistics, including the one significant LLaMA-3-8B cell, are reported in *SI Appendix*, Per-Model Steering Statistics. The same shuffled-vector control leaves the transaction-utility WTP effect near-null (five of six conditions), and steering the transaction-utility vector preserves unrelated general-knowledge accuracy within the primary range (*SI Appendix*, Random-Vector Control and Capability Preservation).

The direction is specific to mental accounting. The steering direction could carry a generic instruction-following signal rather than mental-accounting content. Two controls rule this out (Figure 5): the orthogonalization-survival and ablation tests run on the models with a steerable effect (panels A, C), while the persona-projection test (panel B) spans all six models—the persona shifts activations along the mental-accounting axis even in the small models that do not steer, so the direction reflects the representation, not merely instruction-following.

First, we remove a generic instruction-following component built from non-monetary persona and summarize/translate/classify contrasts. Orthogonalizing the mental-accounting vector against this direction leaves the dose-response essentially unchanged (Figure 5A).

Second, the direction mediates an explicit persona instruction. Persona prompts shift activations along the mental-accounting axis, and ablating that axis during generation reverses the persona-induced behavior in the testable cells (Figure 5B–C). The persona therefore acts through the mental-accounting direction.

These controls do not identify whether the direction was formed during pretraining, instruction tuning, or their interaction; they show that, in the deployed instruction-tuned models, the effective direction is mental-accounting–specific rather than generic instruction-following.

Prompt-Diversity Generalization Beyond varying the wording of the sweep scenarios, we test whether the direction generalizes to structurally different prompts it was never fit on, applying the canonical steering vector to held-out vignette families whose wording is disjoint from training. Transfer is cleanest for segregating gains, where fungibility steering substantially reduces the bias rate; integrating losses is weaker and more heterogeneous. We therefore treat cross-prompt generalization as established for segregating gains and bounded for integrating losses, with the full held-out design in *SI Appendix*, Prompt-Diversity Generalization.

Steering versus prompting. A persona instruction alone can move behavior about as strongly as steering. The value of steering is therefore evidential: it reveals a continuous dose-response in α and,

for transaction utility, a widening premium−ordinary WTP gap that the prompt-only manipulation does not reproduce. The full prompt-only comparison is in *SI Appendix*, Prompt-Only Baseline.

Steering reshapes reasoning as well as answers. Steering also reshapes the model's free-text reasoning. Mental-accounting steering produces category-based explanations, while fungibility steering reasons over total amounts; representative transcripts are in *SI Appendix*, Example Responses after Steering. Internally, the distance between the two steered activation trajectories grows with α and layer depth (*SI Appendix*, Mechanistic Effects of Interventions).

4 Materials and Methods

Elicitation protocol. The model set is a six-model capacity $\times$ family design—three pretrained families at two scales each: LLaMA-3-70B/8B (Dubey et al., 2024), Qwen2.5-72B/3B (Qwen Team, 2024), and Gemma-2-27B/2B (Gemma Team, 2024). All responses are sampled at temperature 0.6 and top-p 0.9. Baselines use the full $1{,}440$-vignette bank with eight stochastic draws per vignette; their confidence intervals resample vignettes, with seed-cluster intervals reported in *SI Appendix*, Full-Bank Baseline Measurement. Steering uses smaller randomized sweep prompts, with per-cell sample sizes reported in the *SI Appendix* steering tables. Exact prompt templates, split construction, label assignment, and parsing rules are in *SI Appendix*, Experimental Setups.

Detailed probing and intervention procedures appear in Appendix G.

5 Conclusions

Our central finding is that three mental-accounting biases that economics has kept conceptually distinct sunk-cost effect— share a steerable internal axis in LLaMA-3-70B. That axis is not merely readable by a probe: moving the model along it changes choices, and a random direction of the same size does not. The analogous transaction-utility intervention modulates WTP, most clearly by widening the premium−ordinary gap under amplification. In the models tested, controllability is more evident at larger scale than tied to one family: the two additional large models reproduce categorical de-biasing, whereas the three smaller models show little or no comparable control. Where earlier work locates a single steerable concept, the axis here is shared across biases a field keeps distinct, and its controllability—unlike the readable probe signal, which persists even in models that cannot be steered—is concentrated at larger scale in the models tested. The finding gives a mechanism for why high-capacity LLMs can replicate some economic behavior: behavior that looks like a familiar bias at the output can be supported by a steerable direction in the model's activations. For synthetic respondents, the implication is an audit criterion: matching choices is not enough to establish this stronger representational form.

Three questions remain. The first is what the shared direction represents—a single disposition to "act like a mental accountant" or partially separate machinery for each bias. The second is why steerable control appears more clearly at larger scale in these models; one possibility is that capacity lets a model compress scattered, input-specific heuristics into a low-dimensional direction that can be read and moved. Because our design varies scale only alongside the other choices that differ across model releases—training data, tokenizer, and instruction tuning—it localizes controllability to the larger models rather than isolating parameter count as its cause. The third is human correspondence: because our tests use the model's own choices as the outcome, they establish causal control over those choices rather than calibrated agreement with human choice distributions, and estimating that correspondence would require matched human data on the same vignette bank. Steerability is a sufficient sign of an internal representation, not a necessary one: a model may encode a bias that a contrast-derived linear intervention does not move. The same design extends to other human tendencies models are now asked to emulate—read a direction from the model's own contrast between two framings, then test whether moving it changes behavior—a check any study using models as synthetic respondents can run.

AI USE STATEMENT

AI assistance was used to prepare this ICLR version, including LaTeX formatting, reference formatting, and compilation checks. The language models evaluated as experimental subjects and the associated protocols are described in the main text and appendix.

REPRODUCIBILITY STATEMENT

The main text describes elicitation, probing, and activation interventions. The appendix provides experimental prompts, measurement procedures, controls, model-specific results, and robustness analyses.

REFERENCES

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245, 2024.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Christopher A Bail. Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

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

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

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(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2316205120.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S^3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Ali Goli and Amandeep Singh. Can llms capture human preferences? *Marketing Science*, 2024.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024.

Junsol Kim and Byungkyu Lee. AI-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys. *arXiv preprint arXiv:2305.09620*, 2023.

Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Yan Leng and Yuan Yuan. Suva: a probabilistic framework for auditing llms with an application to social preferences. *Information Systems Research*, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9): e2313925121, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pp. 1–22, 2023.

Drazen Prelec and George Loewenstein. The red and the black: Mental accounting of savings and debt. *Marketing Science*, 17(1):4–28, 1998.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. In *Forty-first International Conference on Machine Learning*, 2024.

Richard Thaler. Mental accounting and consumer choice. *Marketing Science*, 4(3):199–214, 1985.

Richard H Thaler. Mental accounting matters. *Journal of Behavioral Decision Making*, 12(3): 183–206, 1999.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Amos Tversky and Daniel Kahneman. The framing of decisions and the rationality of choice. *Science*, 211(4481):453–458, 1981.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Yuzi Yan, Jialian Li, Yipin Zhang, and Dong Yan. Exploring the LLM journey from cognition to expression with linear representations. *arXiv preprint arXiv:2405.16964*, 2024.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Design, Stimuli, and Measurement

A.1 Experimental Setups

Prompt S1. *Mr. A was given tickets to lotteries involving the World Series. He won $50 in one lottery and $25 in the other. Mr. B was given a ticket to a single, larger World Series lottery. He won $75. Who was happier? A, or B, or no difference.*

Prompt S2. *Mr. A received a letter from the IRS saying that he made a minor arithmetical mistake on his tax return and owed $100. He received a similar letter the same day from his state income tax authority saying he owed $50. There were no other repercussions from either mistake. Mr. B received a letter from the IRS saying that he made a minor arithmetical mistake on his tax return and owed $150. There were no other repercussions from his mistake. Who do you think was more upset? A, or B, or no difference.*

Prompt S3. *Imagine that you have decided to see a play and paid the admission price of $10 per ticket. As you enter the theater, you discover that you have lost the ticket. The seat was not marked, and the ticket cannot be recovered. Would you pay $10 for another ticket?*

Prompt S4. *You are lying on the beach on a hot day. All you have to drink is ice water. For the last hour you have been thinking about how much you would enjoy a nice cold bottle of your favorite brand of beer. A companion gets up to go make a phone call and offers to bring back a beer from the only nearby place where beer is sold (a fancy resort hotel) [a small, run-down grocery store]. He says that the beer might be expensive and so asks how much you are willing to pay for the beer. He says that he will buy the beer if it costs as much or less than the price you state. But if it costs more than the price you state he will not buy it. You trust your friend, and there is no possibility of bargaining with (the bartender) [store owner]. What price do you tell him?*

The intervention sweeps reported in the main text are run over randomized variants of these canonical scenarios (five variants per dimension for segregating gains, integrating losses, and sunk cost, and ten paired premium/ordinary venues for transaction utility), so that no steering result depends on a single wording. The main-text baseline figure instead pools the full diverse vignette bank described next; the held-out portion of that bank is also used for the prompt-diversity generalization test (Section D.3).

A.1.1 Diverse Prompt Bank for the Revised Analyses

The canonical vignettes above (Prompts S1–S4) are the classical formulations from Thaler (1985, 1999). For the revised analyses we expanded each dimension into a bank of 12 structural seeds, each instantiated with 30 mechanical paraphrases, for a total of $12 \times 30 \times 4 = 1{,}440$ vignettes. Seeds vary the protagonists' names, the framing context (lottery, scratch-off, workplace bonus, tax notice, repair charge, gift money), monetary stakes ($25–$500), venues, and cultural references; paraphrases apply systematic lexical and syntactic substitution, so the bank is fully reproducible without language-model paraphrasing and carries no label drift. Splits are stratified by template family (50% train, 25% development, 25% held-out test) so that no structural template appears in more than one split; the trigram Jaccard similarity across splits is bounded at $p_{95} = 0.038$. Probing uses template-family-stratified grouped cross-validation in which all instances sharing a scenario are kept within a single fold, and the prompt-diversity generalization test (Section D.3) applies the steering vector only to held-out test-family prompts; the cross-dimension transfer matrix instead runs on the randomized sweep prompts (Section B.3). Probe labels are the model's own mental-accounting-versus-fungibility choice; for LLaMA-3-8B, whose responses collapse to a single class on some dimensions under deterministic decoding, labels are instead assigned stochastically—from its per-scenario bias rate over 12 samples at T=0.6, median-split into balanced classes. One representative seed per dimension is shown below, with variation slots in braces. (These illustrative templates are not assigned prompt numbers, so they do not alter the Prompt S1–S4 numbering referenced in the main text.)

Segregating gains. *Mr. A was given tickets to lotteries involving the {event}. Mr. A won ${X} in one lottery and ${Y} in the other. Mr. B was given a ticket to a single, larger {event} lottery. Mr. B won ${total}. Who do you think was happier?*

Integrating losses. *{name_a} received a property-tax assessment notice owing ${X} and, on the same day, a separate income-tax assessment notice owing ${Y}. There were no other repercussions from either notice. {name_b} received a single tax notice the same day owing ${total}. Who would feel more upset about the tax notices?*

Sunk cost. *Imagine that you have decided to attend a {event} and paid the admission price of ${price} per ticket. As you enter the {venue}, you discover that you have lost the ticket. The seat was not marked, and the ticket cannot be recovered. Would you pay ${price} for another ticket?*

Transaction utility. *You are at the beach on a hot day and want a {item}. A friend offers to walk to the only nearby seller, which is {venue_context}. The {item} might cost more than usual and the friend wants to know the maximum you are willing to pay. There is no bargaining. What is the maximum you would pay (in US dollars)?*

A.2 STIMULUS

We employ two distinct system prompts to establish contrasting decision-making contexts: one emphasizing mental accounting (Prompt S5) and the other fungibility (Prompt S6). This approach, inspired by contrastive prompts (Rimsky et al., 2024), contrasts activation patterns associated with mental-accounting- and fungibility-consistent responses. By comparing the model's activations under these contrasting prompts, we aim to learn a steering vector that differentiates the two response patterns.

Prompt S5. *You are a decision-maker influenced by mental accounting, where people categorize expenses and evaluate gains and losses based on specific mental budgets or contexts. Consider how subjective valuations, emotional reactions, and context-dependent reference points shape decision-making.*

Prompt S6. *You are a decision-maker who treats all monetary values as fungible, ignoring mental accounting. Evaluate gains and losses objectively, without assigning them to separate mental budgets or being influenced by context or emotional framing.*

Stimulus prompts, such as those listed in Prompt S7, provide diverse and incomplete contexts designed to elicit mental-accounting- or fungibility-consistent completions. These prompts are intentionally partial sentences to let the model complete the thought and expose the activation patterns associated with those completions. By leaving the scenarios open-ended, we can observe how the model's generated text responds to the two framings.

When combined with system prompts for mental accounting (Prompt S5) or fungibility (Prompt S6), these stimulus prompts guide the model to generate responses aligned with the desired decision-making perspective. This approach isolates the contrast between mental accounting and fungibility in both outputs and activations rather than relying on one memorized wording.

The variety of stimulus prompts covers a broad range of decision-making scenarios, such as managing budgets ("Combine your budget or"), evaluating savings ("It's a $10 saving but"), and reconsidering sunk costs ("Would you still go if"). By incorporating diverse contexts, we assess whether the model generalizes mental accounting and fungibility across situations rather than memorizing a single phrasing.

Prompt S7. *["You have separate bud", "You won $25 and $50, does", "Would you still go if", "Keeping it for a vacation", "Combine your budget or", "Feels like two wins, better than", "Return to the restaurant to", "Your grocery budget is low, but", "Bought a concert ticket for $30 but", "Do you keep it as a", "Hmm, is it worth combining", "It's a $10 saving but", "The total savings of", "Ugh, it's not worth", "Feels like a loss, but", "Separate gains feel like", "No difference if I combine", "Is it really different if", "The restaurant bill was", "It's only $5 but", "Do I feel better about", "Seeing two refunds feels", "It's just one amount but", "What if I use dining funds for", "Does combining funds make", . . .]*

A.3 FULL-BANK BASELINE MEASUREMENT

The baseline rates in main-text Figure 1 pool the full diverse vignette bank rather than the sweep prompts, and that figure reports all six models on this basis. Here we give the detailed two-level uncertainty breakdown for the two LLaMA-3 models—the within-family capacity pair—on an identical basis: all $1{,}440$ vignettes (360 per dimension), each sampled 8 times stochastically at temperature 0.6 (four samples under each of the two inert direction passes at $\alpha = 0$), for $2{,}880$ responses per dimension (parse validity 98.8–100% for LLaMA-3-70B, 95.7–99.9% for LLaMA-3-8B). For segregating gains, integrating losses, and sunk cost the outcome is the pooled rate of

mental-accounting-consistent responses; for transaction utility the primary outcome is the probability that willingness-to-pay for the same item is higher at the premium venue than at the ordinary venue, $P(\text{prem} > \text{reg})$, matched on item (the canonical beach-beverage contrast), with venue-class pooling as a robustness variant. We report two uncertainty levels: a prompt-level bootstrap 95% CI (resampling vignettes) and a more conservative seed-cluster CI (resampling the 12 structural seed families).

Table S1: Full-bank $\alpha = 0$ baselines for the two LLaMA-3 models (the other four models' baselines are in main-text Figure 1). Pooled rates with prompt-level bootstrap 95% CIs; seed-cluster CIs in brackets on the second line of each cell are roughly twice as wide.

Dimension	LLaMA-3-70B	LLaMA-3-8B
Segregating gains	50.8% $[46.6, 55.1]$	31.5% $[29.0, 34.0]$
Integrating losses	72.7% $[69.7, 75.6]$	66.0% $[63.6, 68.3]$
Sunk cost	18.2% $[15.6, 21.0]$	3.7% $[2.9, 4.6]$
TU $P(\text{prem} > \text{reg})$	0.865 $[0.822, 0.905]$	0.843 $[0.801, 0.883]$

Seed-cluster CIs for LLaMA-3-70B: SG $[30.1, 71.5]$, IL $[57.4, 85.3]$, SC $[7.7, 29.7]$; for LLaMA-3-8B: SG $[22.6, 40.8]$, IL $[56.7, 73.9]$, SC $[1.5, 6.3]$. For transaction utility, the premium$-$ordinary WTP gap is $+\$4.58$ $[2.72, 6.54]$ for LLaMA-3-70B and $+\$3.69$ $[2.42, 5.04]$ for LLaMA-3-8B under the canonical item contrast; under venue-class pooling, $P(\text{prem} > \text{reg}) = 0.945$ $[0.929, 0.959]$ for LLaMA-3-70B and 0.920 $[0.901, 0.938]$ for LLaMA-3-8B, with gaps $+\$8.67$ and $+\$9.00$ respectively. The TU CIs are scenario-cluster bootstrap intervals.

Two facts about these baselines matter for interpreting the main text. First, baseline rates vary substantially across vignette families: the seed-cluster CIs are roughly twice as wide as the prompt-level CIs (for segregating gains, $[30.1\%, 71.5\%]$), so a model that is split at chance bank-wide can sit at ceiling or floor within a particular family. Second, this family-level heterogeneity explains the discrepancy between the bank-wide segregating-gains baseline (50.8%) and the near-ceiling baseline ($\approx 99\%$) observed on the sweep prompts used for the intervention analyses: the sweep prompts happen to lie in high-bias families. The ceiling block on amplifying segregating gains in the intervention results is therefore a property of the sweep prompt families, not of the dimension itself.

B Geometry of the Mental-Accounting Direction

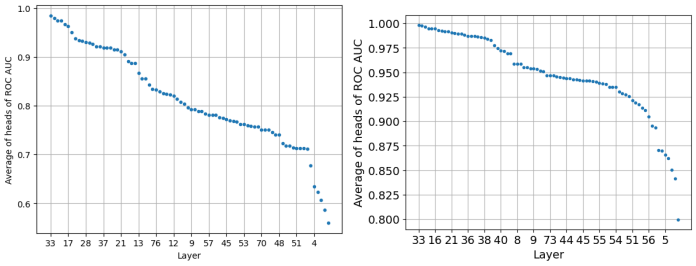
B.1 Principal Component Analysis

We present the variances explained by each principal component in descending order in Figure S1. To minimize disruption to the model's behavior, we select intervention layers from the probe-AUC elbow, where marginal predictive gains plateau, and apply perturbations within those layer bands.

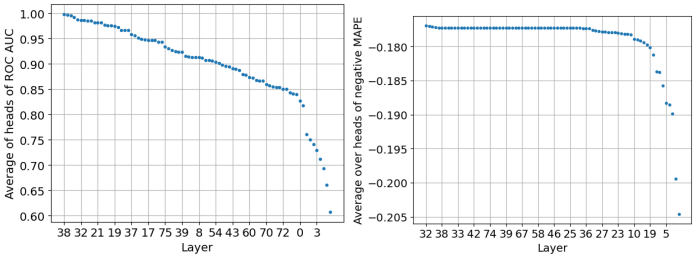
B.2 Layer-Resolved Projection Analysis

Figure S2 shows the projection of dimension-specific contrastive vectors onto held-out activations across all 80 layers of LLaMA-3-70B. Dimension-specific peaks emerge at distinct layer ranges: SG peaks at L63 (Cohen's $d = 0.70$), SC peaks at L44 ($d = 1.36$), and IL shows a bimodal profile with a sign-flipped basin at L24–34 and a secondary peak at upper layers. This layer-resolved analysis motivates the selection of distinct intervention layer ranges for each dimension (upper layers 63–75 for SG and IL; mid-layers 20–45 for SC) and corroborates the bimodal IL encoding reflected in the main-text probing profile.

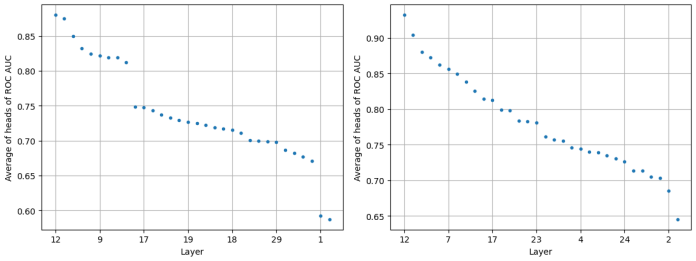
As a causal polarity check on the IL basin, we reran the IL intervention at the mid-layer band where the projection changes sign (L24–34) after reversing the steering direction. The ordinary mid-layer IL intervention is unstable, consistent with the projection crossing zero across nearby layers. After the sign reversal, the mental-accounting arm becomes a clean de-biasing dose-response: across 480 valid generations, the mental-accounting response rate falls from 50% at baseline to 0% at the largest intervention strength (Spearman $\rho = -0.941$, $p = 0.005$; logistic $\beta = -32.12$, $p = 7.3 \times 10^{-8}$).



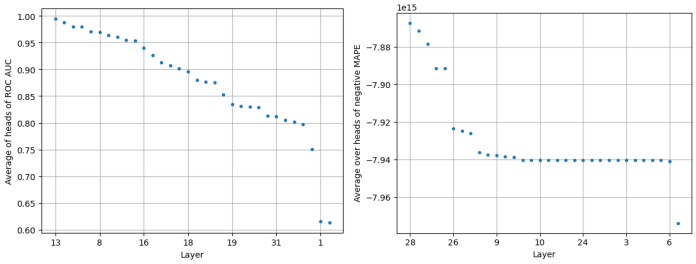
(a) Integrate Losses (LLaMA-3-70B)(b) Segregate Gains (LLaMA-3-70B)



(c) Sunk Costs (LLaMA-3-70B) (d) Transaction Utility (LLaMA-3-70B)



(e) Integrate Losses (LLaMA-3-8B) (f) Segregate Gains (LLaMA-3-8B)



(g) Sunk Costs (LLaMA-3-8B) (h) Transaction Utility (LLaMA-3-8B)

Figure S1: Variance explained by each principal component, ranked in descending order.

We use this test as evidence that the IL direction has opposite polarity in the mid-layer basin relative to the upper-layer intervention band, not as a separate claim that IL is best steered from mid-layers.

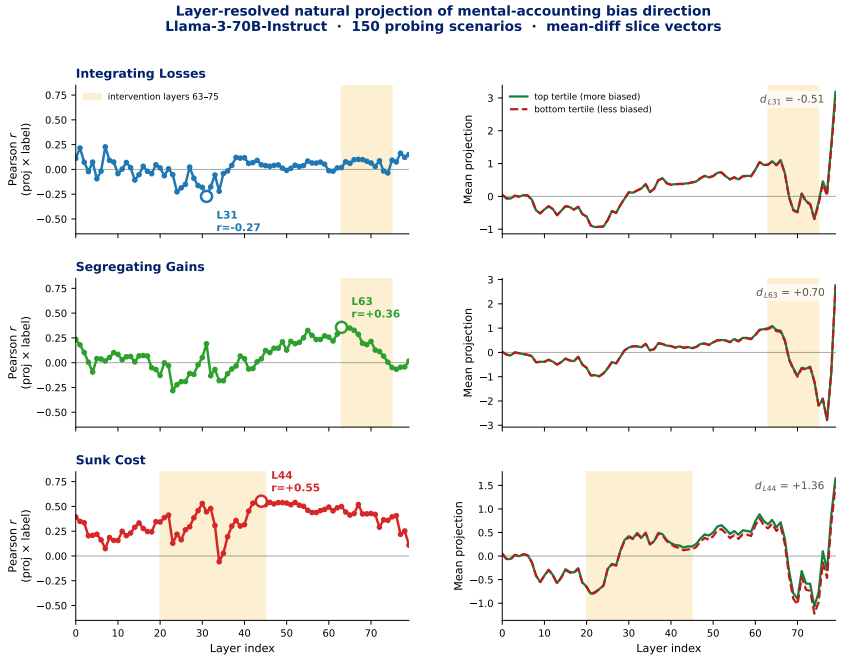


Figure S2: Layer-resolved projection of contrastive steering vectors onto held-out activations across all 80 layers of LLaMA-3-70B. Each curve shows one mental-accounting dimension. Dimension-specific peaks indicate where each bias is most strongly encoded.

B.3 Cross-Dimension Transfer Matrix

A direction estimated from one categorical dimension is applied to each dimension over the signed grid $\alpha \in \{0, 0.05, \ldots, 0.40\}$ in both directions (SG/IL vectors at upper layers 63–75, SC at mid layers 20–45; 5 randomized prompts, seed 2025; $n = 201$–480 valid observations per cell). This transfer sweep is smaller than the main-text dose-response figure (fewer samples per α), so a Spearman correlation computed on its eight α-level means is underpowered and attenuated relative to the main-sweep diagonal values; we therefore report significance from the *per-response logistic* $P(\mathrm{MA}{=}1) \sim \alpha$ fit on the individual observations with Benjamini–Hochberg correction—the same authoritative test used for the dose-response—and give the Spearman ρ only as a descriptive summary. This resolves the apparent tension with the main-text dose-response figure: the within-dimension (diagonal) effects are significant under the per-response logistic here, matching the dose-response, and the attenuated diagonal Spearman values reflect only the lower power of a rank correlation over eight α-means on a smaller sweep.

Table S2 lists every cell. Of the 15 arms in which the target dimension is not at the response ceiling, 13 are significant after correction; the two exceptions are IL→SG de-biasing (near-saturated SG target) and IL→SC amplification. Under amplification the SG target column is at the ceiling ($P(\mathrm{MA}) \approx 1.0$ at $\alpha = 0$), so amplification into it cannot be assessed and its cells are excluded from the count; the negative logistic slopes there are ceiling artifacts. The pervasive off-diagonal transfer, together with the near-collinear segregating-gains and integrating-losses vectors (peak cosine 0.97–0.99 in all six models; Section B.4) and the shared late-layer head cluster (Section B.5), supports a substantial shared steering component rather than three independent controls; the sunk-cost vector joins that component through transfer rather than collinearity. Averaged over the de-biasing cells, off-diagonal transfer reaches $0.87\times$ the within-dimension diagonal (cross-dimension mean $|\rho| = 0.63$ versus same-dimension 0.73).

Table S2: Cross-dimension transfer, per cell. β is the per-response logistic slope in the intended steering direction (positive = behavior moves as steered); q is BH-corrected across all arms; ρ is the descriptive dose-response Spearman on α-level means. Diagonal (within-dimension) rows in bold. † near-separation (odds ratio very large; effect direction reliable, Wald p conservative). ‡ target at response ceiling under amplification—slope is an artifact and the cell is not counted.

Train → Test	Dir	n	baseline	ρ	β	q_{FDR}
De-biasing (fungibility, $-\alpha$)						
IL → IL	fung	471	0.35	-0.66	$+3.42$	0.002^{**}
IL → SG	fung	340	0.98	-0.15	$+3.75$	0.28 (ns)
IL → SC	fung	475	0.20	-0.87	$+25.4^{\dagger}$	$< 0.001^{***}$
SG → IL	fung	215	0.33	-0.41	$+5.45$	0.005^{**}
SG → SG	fung	201	1.00	-0.71	$+34.2^{\dagger}$	$< 0.001^{***}$
SG → SC	fung	211	0.23	-0.86	$+18.1^{\dagger}$	0.003^{**}
SC → IL	fung	240	0.43	-0.65	$+3.58$	0.010^{**}
SC → SG	fung	238	1.00	-0.85	$+12.9^{\dagger}$	$< 0.001^{***}$
SC → SC	fung	240	0.13	-0.81	$+8.98$	0.005^{**}
Amplifying (mental accounting, $+\alpha$)						
IL → IL	ment	472	0.35	$+0.62$	$+1.82$	0.026^{*}
IL → SC	ment	480	0.22	$+0.31$	$+0.94$	0.21 (ns)
IL → SG‡	ment	446	1.00	—	-6.12	0.11
SG → IL	ment	234	0.37	$+0.90$	$+4.78$	$< 0.001^{***}$
SG → SG‡	ment	223	1.00	—	-6.17	0.18
SG → SC	ment	240	0.10	$+0.26$	$+2.68$	0.020^{*}
SC → IL	ment	236	0.34	$+0.78$	$+2.36$	0.036^{*}
SC → SG‡	ment	231	0.97	—	-6.72	< 0.001
SC → SC	ment	240	0.20	$+0.57$	$+2.40$	0.033^{*}

Transfer to transaction utility. The matrix above is restricted to the three categorical dimensions, which share a discrete-choice format. We separately tested whether a vector estimated on a categorical bias moves *transaction-utility* willingness-to-pay (WTP), the one dimension measured continuously. Applying each categorical vector to the paired-venue WTP sweep in LLaMA-3-70B, the mental-accounting arm raises WTP and the fungibility arm lowers it, with five of six arms significant (Table S3). This transfer is confined to the WTP *level*: no categorical vector reproduces the venue-dependent premium−ordinary *gap* that defines transaction utility, which the transaction-utility vector alone produces. The categorical axis therefore reaches transaction utility's WTP level, while its defining gap is carried by a transaction-utility-specific direction; and transaction utility does not itself carry a steerable direction at all in the other five models (Section D.1).

Table S3: Categorical-to-transaction-utility transfer in LLaMA-3-70B. Each categorical steering vector is applied to the paired-venue WTP sweep; entries are the Spearman ρ of mean WTP on signed α (BH-corrected: $^{*}q < 0.05$, $^{***}q < 0.001$). The mental-accounting (ment) arm raises WTP; the fungibility (fung) arm lowers it.

Vector trained on	ment (raise WTP)	fung (lower WTP)
Segregating gains	$+0.79^{*}$	-0.79^{*}
Integrating losses	$+0.82^{*}$	-0.96^{***}
Sunk cost	$+0.64$ (ns)	-0.86^{*}

B.4 Steering-Vector Geometry Across Models

The transfer matrix measures shared *control*; here we measure shared *geometry*, comparing the deployed dimension-specific steering vectors directly: the cosine similarity between each pair of dimensions at every layer, in all six models, plus the variance explained by the leading principal component (PC1) of the pooled contrast pair-differences within the shared SG/IL intervention band (Table S4; per-layer profiles in Figure S3).

The segregating-gains and integrating-losses directions are nearly collinear in every model tested (peak cosine 0.97–0.99), with the alignment sustained across the upper-layer intervention band; this geometric sharing recurs across all three pretrained families at both scales. The sunk-cost direction is geometrically distinct in the three large models—its cosine with the gain–loss directions peaks at 0.22–0.56 over all layers, including its own mid-layer band—so its membership in the shared component rests on the cross-dimension transfer (Section B.3), in which each transferred vector is applied at the estimating dimension's own band. No single principal component dominates the pooled pair-differences (PC1 = 55–68% across models), yet this concentration still exceeds what independent per-dimension controls would give: for LLaMA-3-70B a permutation null that independently random-rotates each dimension's contrast differences—breaking cross-dimension alignment while preserving each dimension's own within-scatter—has a leading-PC share of 0.50 (maximum 0.51 over 10,000 draws), well below the observed 0.65 ($p < 10^{-4}$). Finally, the small models carry the same gain–loss geometry (peak cosines 0.97–0.99) despite little steerable control, and Qwen2.5-3B shows the highest sunk-cost alignment of any model (0.91): geometric structure alone neither confers nor predicts causal controllability, consistent with the main-text dissociation between readable structure and control.

Table S4: Band-resolved geometry of the deployed steering vectors. For each model: the peak over layers of the cosine similarity between dimension-specific vectors (layer of the peak in parentheses), and the variance explained by the first principal component of the pooled contrast pair-differences within the shared SG/IL intervention band.

Model	SG–IL peak cos	SC–SG peak cos	SC–IL peak cos	PC1, SG/IL band
LLaMA-3-70B	0.99 (L26)	0.33 (L9)	0.26 (L26)	65%
LLaMA-3-8B	0.99 (L9)	0.39 (L14)	0.41 (L21)	67%
Qwen2.5-72B	0.97 (L55)	0.45 (L30)	0.39 (L55)	60%
Qwen2.5-3B	0.97 (L22)	0.91 (L21)	0.89 (L21)	64%
Gemma-2-27B	0.98 (L17)	0.56 (L2)	0.22 (L22)	68%
Gemma-2-2B	0.99 (L11)	0.43 (L18)	0.37 (L18)	55%

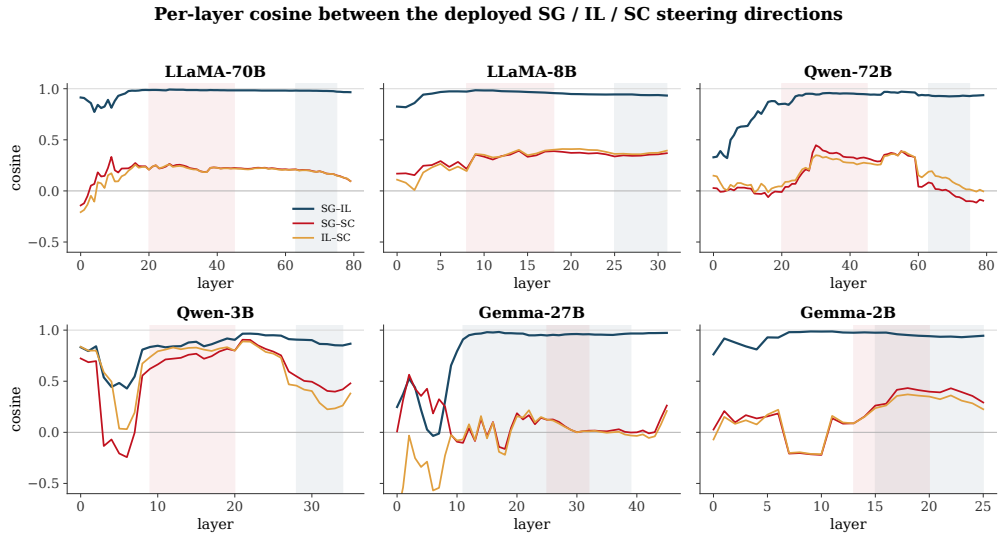


Figure S3: Per-layer cosine similarity between the deployed dimension-specific steering vectors for all six models, with intervention bands shaded. The SG–IL cosine remains high across the upper band in every model; the SC–SG and SC–IL cosines stay low in the large models at all depths, including the mid band where sunk cost is steered.

B.5 Head-Level Decomposition and Token-Position Sensitivity

To address whether layer- and head-averaging obscure the steering direction, and whether the choice of token position matters, we ran a forward-pass-only collection on LLaMA-3-70B over the 20 elicitation pairs (the paired mental-accounting/fungibility elicitation prompts, distinct from the 283 contrastive pairs used to construct the steering vector), capturing (i) the per-head input to each layer's output projection and (ii) the residual stream both at the last prompt token and mean-pooled over all prompt tokens.

Head-level decomposition. Figure S4 shows the per-head mean-difference norm (budgeted − fungible) within each dimension's intervention band. The mental-accounting signal is distributed across many heads rather than localized to a single head, but with strong enrichment in a small late-layer set: the 10 highest-norm heads ($\approx 1\%$ of the heads in the band) account for 11.2% (SG), 11.5% (IL), and 13.7% (TU) of the band's total mean-difference norm (≈9–11× enrichment over uniform), and 4.9% for SC (≈8×). The top heads for the three upper-layer dimensions overlap heavily (a shared L74–L75 cluster: heads 21, 27, 42, 43, 47), whereas sunk cost's top heads are disjoint and mid-layer (L23–L37). This independently corroborates both the shared representation across dimensions and the partial layer dissociation reported in the main text.

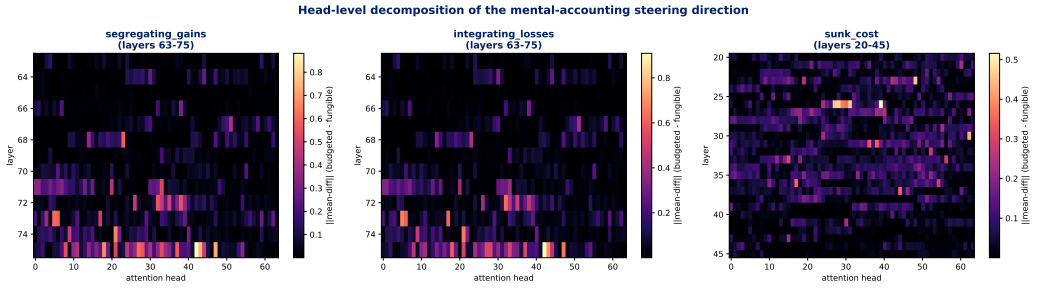


Figure S4: Per-head ‖mean-difference‖ (budgeted − fungible) within each dimension's intervention band (LLaMA-3-70B).

Token-position sensitivity. The pipeline does not average over token positions; the steering direction is computed at the last prompt token. Figure S5 recomputes the per-layer direction from activations mean-pooled over all prompt tokens and compares it to the last-token direction. Within every intervention band the two directions are far from collinear (mean cosine ≈ 0.21), and the last-token contrastive difference carries a larger norm in the relevant late layers (e.g. 8.3 vs. 3.2 at L75). A reproduction check confirms our last-token extraction matches the deployed steering vectors exactly (cosine $= 1.000$ at every layer). Token position is thus consequential, justifying the principled use of the last (decision) token.

C Specificity and Robustness Controls

C.1 Activation Intervention versus Fine-Tuning and Prompt Engineering

Our use of intervention is diagnostic. A persona prompt can shift these behaviors about as strongly as steering (main text), so the added value here is evidential: by operating directly on activations during inference without altering weights, the intervention provides a causal test of whether a bias is encoded as a manipulable internal direction—something output-level methods such as prompting or fine-tuning cannot isolate. The comparisons below concern that diagnostic role, together with practical considerations (efficiency, targetedness, reversibility).

C.1.1 Versus Fine-Tuning

Steering operates on activations at inference time, with no backpropagation and no weight updates (Li et al., 2023). Three properties follow. It is cheap: a direction is located from a few hundred contrastive examples rather than a training run. It is non-invasive: weights are unchanged, and within the reported

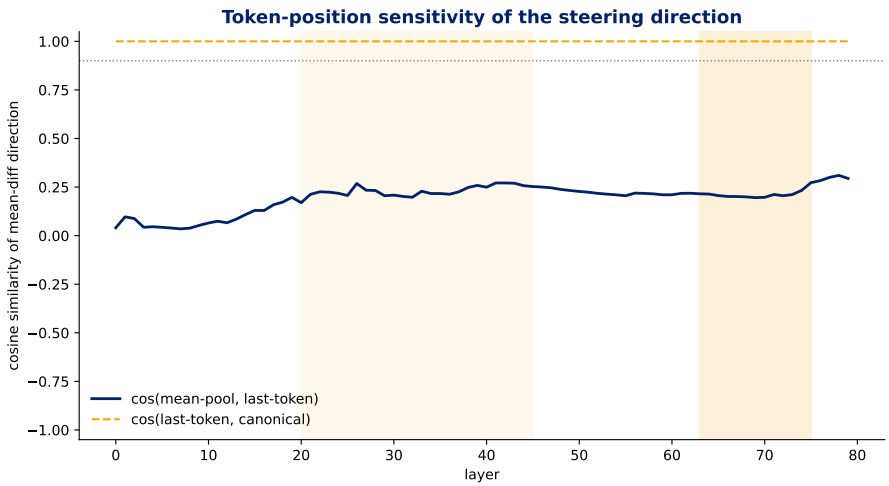


Figure S5: Per-layer cosine between the last-token and mean-pooled mean-difference directions (LLaMA-3-70B). Shaded bands mark intervention layers; the dashed curve verifies reproduction of the deployed vectors.

intervention range performance is preserved on our unrelated general-knowledge battery, whereas fine-tuning adjusts weights globally and can shift behavior. And it is reversible: the effect is simply omitted on later passes (Turner et al., 2024).

C.1.2 Versus Prompt Engineering

A prompt is discrete and all-or-nothing; steering applies a continuous, signed weight to a single direction (Yan et al., 2024). That continuity is what lets us trace the graded dose-response that anchors the representational claim and that the single persona-prompt baseline does not trace. Steering also consumes no context-window tokens and can be combined algebraically across forward passes.

C.1.3 Prompt-Only Baseline: Does Steering Add Anything Over Prompting?

A natural question is whether the behavioral shifts we attribute to activation steering can be obtained more simply by *instructing* the model, in which case the steering vector would add little. We tested this directly. Holding everything else fixed (same scenarios, same elicitation format, and *no* activation intervention), we prepended either a fungibility persona ("treat all money as fungible…") or a mental-accounting persona to each evaluation prompt and measured the bias (Figure S6).

Behaviorally, prompting and steering are comparable. A persona instruction alone moves behavior in both directions, often as strongly as steering: the fungibility instruction drives segregating gains and integrating losses to 0% bias and lowers transaction-utility WTP ($7.84 \rightarrow 7.23$); the mental-accounting instruction pushes segregating gains to ceiling and raises WTP ($7.84 \rightarrow 8.36$). We therefore do *not* claim that activation steering is a superior intervention for changing behavior—a well-chosen prompt achieves comparable control.

What steering provides that prompting does not is the evidence for a representation. First, a prompt is all-or-nothing—one instruction, one shift—whereas steering produces a continuous, monotonic response to the scalar α (the signed main-text dose-response), the empirical signature of a *linear* internal direction, which an instruction cannot produce. Second, steering toward mental accounting monotonically *widens* the premium−ordinary WTP gap that defines transaction utility ($\rho = +0.79$), whereas the persona instruction barely moves it ($+0.14$, with a bootstrap CI through 0). Both prompting and steering control the behavior; only steering exhibits the graded, geometry-anchored, layer-localized signature that supports the claim of a manipulable linear representation. One nuance: the *generic* mental-accounting persona is dimension-agnostic and can push integrate-

losses in the wrong direction, whereas the contrastive steering vector captures the shared concept direction; we report this and the per-condition parse-validity rates with the figure.

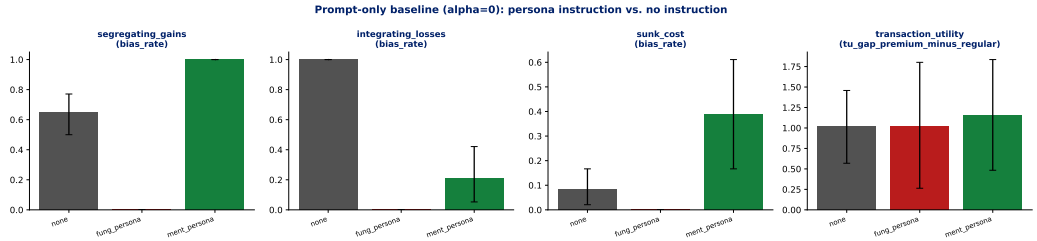


Figure S6: **Prompt-only baseline ($\alpha = 0$, no activation steering).** Per dimension, the mental-accounting bias under no instruction (baseline) versus a fungibility persona (de-bias) and a mental-accounting persona (amplify), with bootstrap 95% CIs. A persona instruction alone moves behavior in both directions, comparably to steering—so the steering vector's contribution is not superior behavioral control but the graded dose-response and layer-localized geometry that the single persona-prompt baseline does not exhibit.

C.2 Logistic Regression Results

Table S5 reports logistic regression of $P(\text{MA} = 1)$ on the *signed* steering strength α, fit on individual observations with Benjamini–Hochberg (BH) correction across the six categorical dimension $\times$ direction tests (3 dimensions $\times$ 2 directions). Here α is signed along the mental-accounting axis: positive for the amplifying (*ment*) arm and negative for the de-biasing (*fung*) arm. A positive β therefore means behavior tracks the axis—more mental-accounting-consistent responding toward $+\alpha$ and less toward $-\alpha$—so steering that works in its intended direction registers as positive in *both* arms (this is why the de-biasing rows are positive even though they lower the bias rate). Five of six tests survive BH at $q < 0.05$. SG/fung and SC/fung show near-complete separation, yielding extreme odds ratios; p-values remain valid. Transaction utility is measured as continuous willingness-to-pay and is tested separately in Section F.1.

Table S5: Full logistic regression results with BH correction.

Condition	Dir	β	OR	p	q_{FDR}	Sig.
IL upper (63–75)	ment	$+3.18$	24.0	3.6×10^{-5}	9×10^{-5}	$***$
IL upper (63–75)	fung	$+3.59$	36.2	1.6×10^{-4}	2.6×10^{-4}	$***$
SG upper (63–75)	ment	-5.69	0.003	0.25	0.32	ns
SG upper (63–75)	fung	$+18.7$	$\sim10^{8}$	$<10^{-6}$	$<10^{-6}$	$***$
SC mid (20–45)	ment	$+3.06$	21.3	6.6×10^{-5}	1.3×10^{-4}	$***$
SC mid (20–45)	fung	$+11.0$	$\sim6\times10^{4}$	$<10^{-6}$	$<10^{-6}$	$***$

C.3 Random-Vector Control

To confirm that steering effects are content-specific, we replace the real steering vector with three independently shuffled vectors of matched norm and repeat all six categorical conditions (the three categorical dimensions in both directions). Table S6 shows that 0/6 shuffled-vector effects survive BH correction. The transaction-utility vector is subjected to the same shuffled-vector control under its paired-venue WTP paradigm; the result is reported with the transaction-utility analysis in Section F.1.

Transaction utility. We repeated the control for the transaction-utility WTP effect, applying three matched-norm shuffled vectors in both directions (six conditions). Five of six are null; a single shuffled condition leaks (shuffle 44, ment arm, $\rho = +0.93$, surviving BH). This isolated, one-directional artifact contrasts with the genuine transaction-utility effect, which is bidirectional and much stronger (LLaMA-3-70B $\rho = +0.976$ amplifying and $\rho = -0.810$ de-biasing). Steering of transaction utility is therefore content-specific, with the same mild caveat noted for the borderline segregating-gains cell above.

Table S6: Random-vector control: 0/6 conditions significant after BH correction.

Condition	Dir	Spearman ρ	p_{raw}	q_{FDR}	Verdict
IL upper	ment	$+0.575$	.136	.400	null
IL upper	fung	$+0.214$	.610	.610	null
SG upper	ment	-0.717	.046	.273	null
SG upper	fung	-0.507	.200	.400	null
SC mid	ment	-0.383	.349	.524	null
SC mid	fung	$+0.313$	.450	.540	null

C.4 Specificity Controls for Generic Instruction-Following

The main-text specificity analysis separates mental-accounting content from a generic tendency to follow instructions. We construct a generic instruction-following direction from non-monetary persona contrasts and task-instruction contrasts (summarize, translate, and classify), then remove that component from the mental-accounting vector before rerunning the steering sweeps. The orthogonalized vector preserves the dose-response pattern, showing that the effect is not explained by a broad instruction-following direction.

We also test whether explicit persona prompts act through the same mental-accounting axis. For each model, we compare activations under a fungibility persona, no persona, and a mental-accounting persona; projections onto the mental-accounting direction follow the expected order. Finally, in the testable cells, we ablate the mental-accounting direction during generation under the persona prompts. Removing that axis reverses the persona-induced behavioral shift, indicating that the persona manipulation is mediated by the same direction used for steering. These controls are summarized in main-text Figure 5.

C.5 Capability Preservation under Steering

A steering intervention could move the targeted bias for the wrong reason: if editing the residual stream simply degrades the model, the apparent "de-biasing" would be generic disruption rather than a specific change to mental accounting. We test this directly on LLaMA-3-70B. We pass an unrelated 48-item general-knowledge battery (three-option multiple choice; e.g. "What gas do plants primarily absorb for photosynthesis?") through the *same* intervention runner used for the main dose-response, under the *same* steering vectors, layer bands, and signed-α grid (segregating-gains and integrating-losses vectors at upper layers 63–75, sunk-cost at mid layers 20–45, transaction-utility at upper layers 63–75; α from -0.4 to $+0.4$ in both directions). The battery contains no mental-accounting content, so steering should not change the correct answer; the question is only whether the model remains able to give it. We apply the edit on every decode step—the stringent case—with deterministic decoding, and re-parse each answer independently of the runner, scoring accuracy against a held-out key and recording a degenerate-output rate (repetition or non-answers).

Across the primary range used for all main results ($|\alpha| \leq 0.30$), accuracy stays at or above 92% against a baseline of 100%, and the degenerate-output rate stays below 3% (Figure S7). General competence is therefore intact precisely where we read the dose-response. Accuracy collapses only beyond this range, at the strongest de-biasing setting ($\alpha = -0.40$) on the two upper-band dimensions: segregating-gains accuracy falls to 2% with 73% degenerate output, and integrating-losses accuracy to 44% with 33% degenerate output, while the mid-band sunk-cost configuration stays at 98% even there. This degradation outside the primary range is why the main analyses cap $|\alpha|$ at 0.30; within that cap, steering moves the targeted bias without eroding the model's general ability to answer.

The transaction-utility vector shows the same profile: accuracy stays ≥ 0.96 within $|\alpha| \leq 0.30$ with near-zero degeneration, collapsing only at the extreme de-biasing point ($\alpha = -0.40$: accuracy 0.52, degeneration 44%), matching the upper-band segregating-gains and integrating-losses vectors (Table S7). Steering the transaction-utility WTP gap therefore does not come at the cost of general capability.

Table S7: Capability preservation on the unrelated general-knowledge battery (LLaMA-3-70B). Accuracy and degenerate-output rate summarized over the primary range $|\alpha| \leq 0.30$; the last two columns report the extreme de-biasing point $\alpha = -0.40$, outside the reported range. Baseline ($\alpha = 0$) accuracy is 1.00 for all four vectors.

Vector (band)	min. acc. $\|\alpha\| \leq 0.3$	max. degen. $\|\alpha\| \leq 0.3$	acc. at $\alpha = -0.4$	degen. at $\alpha = -0.4$
IL (upper)	0.96	0.00	0.44	0.33
SG (upper)	0.92	0.02	0.02	0.73
SC (mid)	0.94	0.00	0.98	0.00
TU (upper)	0.96	0.00	0.52	0.44

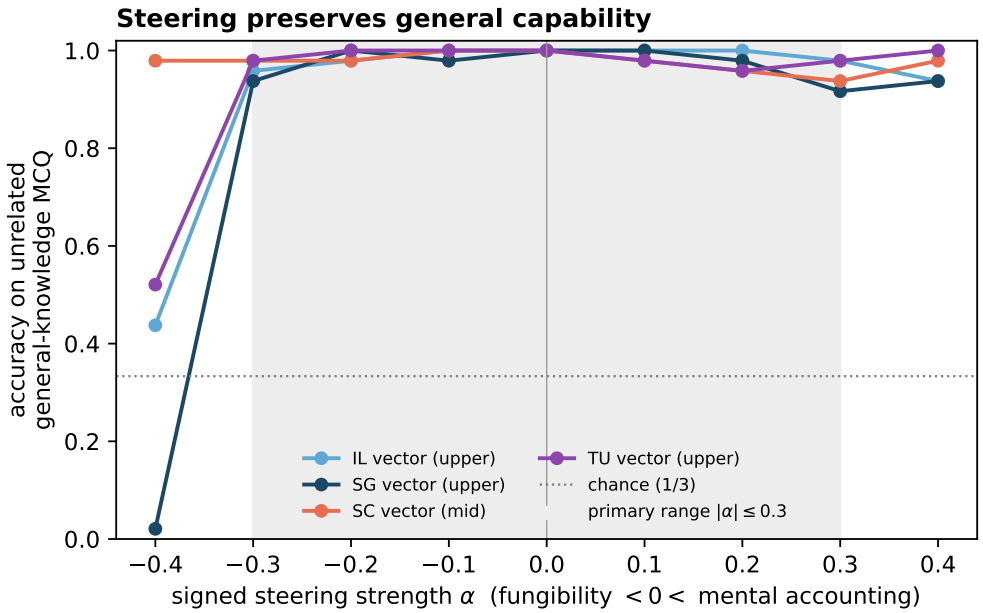


Figure S7: Capability preservation under steering (LLaMA-3-70B). Accuracy on an unrelated 48-item general-knowledge battery as a function of signed steering strength α, for each steering vector applied at its main-text layer band. Accuracy is flat and near baseline across the primary range ($|\alpha| \leq 0.30$, shaded) and drops only at the extreme de-biasing setting on the upper-band dimensions. The mid-band sunk-cost configuration is stable throughout.

D GENERALIZATION ACROSS MODELS AND PROMPTS

D.1 PER-MODEL, PER-DIMENSION STEERING STATISTICS

Table S8 consolidates the steering result for every model $\times$ dimension $\times$ direction cell behind the main-text dose-response figure, so the six-model claims can be read off a single table. Each dimension has two arms: *de-bias* (fungibility, $-\alpha$) and *amplify* (mental accounting, $+\alpha$). The significance verdict for each arm comes from a logistic regression of the per-observation mental-accounting indicator on α (a linear regression on mean willingness-to-pay for transaction utility), Benjamini–Hochberg corrected across cells; this is the authoritative test and is more powerful than the Spearman ρ on α-level means that we report alongside it. "Ceiling" and "floor" mark arms whose baseline already sits at the extreme, leaving no headroom to move in that direction; "n.s." marks a testable arm with no significant dose-response.

The table makes the central dissociation explicit. LLaMA-3-70B, Qwen2.5-72B, and Gemma-2-27B each carry a *de-biasable* mental-accounting direction on every dimension where the baseline leaves room; the only steerable arm in the three small models is a single amplification cell (LLaMA-3-8B integrate-losses), and their remaining arms are null, blocked, or not applicable. Counts run $7/8$ (LLaMA-3-70B), $2/8$ (Qwen2.5-72B and Gemma-2-27B), $1/8$ (LLaMA-3-8B), and $0/8$ (Qwen2.5-3B and Gemma-2-2B), matching the main-text dose-response scorecard; Gemma-2-2B has no TU steering panel, so those cells are marked n/a.

D.2 CROSS-FAMILY REPLICATION: QWEN2.5-72B

To test whether the steerable mental-accounting structure identified in LLaMA-3-70B is specific to the LLaMA-3 lineage, we repeated the intervention analysis on **Qwen2.5-72B-Instruct**, a 70B-class instruction-tuned model from an independently pretrained family. We held the protocol fixed to the LLaMA-3-70B pipeline: mean-difference steering vectors, the same upper (segregating gains, integrating losses) and mid (sunk cost) layer ranges, the signed-α grid, apply-once intervention, and a powered five-prompt panel (≈ 480 runs per direction per dimension; parse validity 1.00 at every α). Transaction utility was also run on Qwen2.5-72B under the identical paired-venue WTP paradigm, and it does not replicate LLaMA-3-70B's effect: both arms are null at the primary range ($\alpha \leq 0.30$; fungibility $\rho = +0.61$, $p = 0.15$; mental accounting $\rho = -0.59$, $p = 0.16$), and the only significant full-grid trend is wrong-signed—de-biasing *raises* WTP ($\rho = +0.74$, $p = 0.04$)—so Qwen2.5-72B carries no steerable transaction-utility direction (baseline WTP \$7.23).

The pattern is asymmetric and interpretable: *Qwen shows a de-biasable categorical component on dimensions with headroom, but not a full bidirectional mental-accounting representation.* Wherever Qwen has headroom to move away from a bias, the fungibility direction reproduces the LLaMA-3-70B effect—segregating gains ($\rho = -0.96$) and sunk cost ($\rho = -0.95$) are de-biased with the same sign, magnitude, and significance. The amplification direction, by contrast, is blocked because Qwen's baseline priors already sit at the extremes: segregating gains and integrating losses are pinned at a biased ceiling, leaving no room to amplify (exactly as the segregating-gains ceiling blocks amplification in LLaMA-3-70B), while sunk cost sits near the rational floor, where amplification fails to move behavior despite ample upward headroom.

We verified that this asymmetry is a property of the model, not of the stimuli or the estimator. (i) The integrating-losses ceiling is not a stimulus artifact: a deliberately low-magnitude (\$4–\$17), source-separated "hard" vignette bank still yields a baseline of 1.000 on all ten prompts with zero "no difference" responses, so Qwen's integrating-losses prior is magnitude- and source-invariant. (ii) The sunk-cost amplification null is not an under-driving artifact: extending the amplification arm to $\alpha = 0.8$ leaves behavior flat at high parse validity—a saturation/attractor signature rather than degeneration. (iii) A probe-derived sunk-cost vector reproduces the mean-difference null (no amplification despite upward headroom; validity 100% to $\alpha = 12$), so the asymmetry is not an artifact of the mean-difference estimator.

In sum, the *de-biasing* component of the mental-accounting direction generalizes across two independently pretrained 70B-class families. We do not claim a full bidirectional replication: the amplification direction is untestable on Qwen because its priors are saturated—a rigidity that is itself informative about how a different pretraining corpus fixes these biases.

Table S8: Per-model, per-dimension steering statistics. For each dimension, de-bias ($-\alpha$) and amplify ($+\alpha$) arms report the Spearman ρ of the dose-response on α-level means; significance (BH-corrected logistic regression; $^{*}q < 0.05$, $^{**}q < 0.01$, $^{***}q < 0.001$) is the authoritative verdict. Baselines are α=0 mental-accounting-consistent response rates, except transaction utility (TU), which is mean willingness-to-pay in dollars. "ceiling"/"floor" = no baseline headroom; "n.s." = testable but no effect. †Transaction-utility de-biasing in LLaMA-3-70B is significant over the full sweep but marginal within the primary range ($\rho = -0.71$, $p = 0.07$ at $|\alpha| \leq 0.30$; Section F.1).

Dim.	α=0	De-bias ($-\alpha$)	Amplify ($+\alpha$)
LLaMA-3-70B — $7/8$ steerable			
SG	0.99	$\rho = -0.96^{***}$	ceiling
IL	0.36	$\rho = -0.89^{**}$	$\rho = +0.83^{**}$
SC	0.23	$\rho = -0.99^{***}$	$\rho = +0.90^{**}$
TU	7.82	$\rho = -0.81^{*\dagger}$	$\rho = +0.98^{***}$
Qwen2.5-72B — $2/8$ steerable			
SG	0.99	$\rho = -0.96^{***}$	ceiling
IL	0.98	ceiling	ceiling
SC	0.31	$\rho = -0.95^{***}$	n.s.
TU	7.23	n.s.	n.s.
Gemma-2-27B — $2/8$ steerable			
SG	0.88	$\rho = -0.96^{***}$	ceiling
IL	0.86	ceiling	$\rho = -0.95^{***}$ (sign-flip)
SC	0.00	floor	n.s.
TU	6.28	n.s.	n.s.
LLaMA-3-8B — $1/8$ steerable			
SG	0.46	n.s.	n.s.
IL	0.33	n.s.	$\rho = +0.99^{***}$
SC	0.00	floor	floor
TU	6.46	n.s.	n.s.
Qwen2.5-3B — $0/8$ steerable			
SG	0.48	n.s.	n.s.
IL	0.98	ceiling	ceiling
SC	0.12	n.s.	n.s.
TU	4.58	n.s.	n.s.
Gemma-2-2B — $0/8$ steerable			
SG	0.02	n.s.	n.s.
IL	0.06	n.s.	n.s.
SC	0.97	ceiling	ceiling
TU	n/a	n/a	n/a

Table S9: Qwen2.5-72B cross-family intervention. Spearman ρ on α-level mean bias rates ($^{***}q < 0.001$, BH-corrected). "Replicates" denotes the same sign and significance as LLaMA-3-70B.

Dimension	fung (de-bias)	ment (amplify)	Qwen α=0 prior
Segregating gains	$\rho = -0.96^{***}$ *(replicates)*	ceiling-blocked	biased ceiling (≈ 0.98)
Sunk cost	$\rho = -0.95^{***}$ *(replicates)*	null ($\beta = -0.23$, $p = 0.76$)	near rational floor (≈ 0.31)
Integrating losses	$1.000 \rightarrow 0.983$, n.s.	pinned	rigid biased ceiling (1.000)

D.3 PROMPT-DIVERSITY GENERALIZATION

To test whether steering transfer is prompt-specific, we built a bank of 1,440 vignettes (12 structural seeds $\times$ 30 mechanical paraphrases $\times$ 4 dimensions) with template-family-stratified splits (Jaccard $p_{95} = 0.038$), and applied the canonical mean-difference vector to Jaccard-disjoint test-family prompts (960 runs, 98.5% validity). SG/fung shows strong cross-prompt transfer ($\rho = -0.943$, $p = 0.005$; bias rate 97% $\rightarrow$ 49%); SG/ment is ceiling-blocked ($p = 0.08$); IL/fung is marginal (logistic $\beta = +3.14$, $p = 0.07$), attributable to high per-prompt baseline heterogeneity (range 0%–100%) that causes directional effects to cancel when pooled. The sweep covers the two 3-way-MCQ dimensions (SG, IL) that share the upper-layer intervention band. A matched held-out sweep was subsequently run for the yes/no sunk-cost dimension at its mid-layer band (20–45; 480 runs, 100% validity), but it is uninformative: on the disjoint test family the baseline sunk-cost rate is near floor (~4%, versus ~30% on the within-paradigm sweep prompts), leaving no headroom to de-bias (de-biasing arm n.s.; near-separation in the logistic fit) and no amplification off the floor (0.05 $\rightarrow$ 0.07 peak, $p = 0.48$). This is an under-powered null rather than a demonstrated failure to generalize; sunk cost's causal evidence therefore rests on the within-paradigm mid-layer sign-flip and the random-vector control. The numeric-WTP (TU) dimension uses a continuous format and was not part of this test.

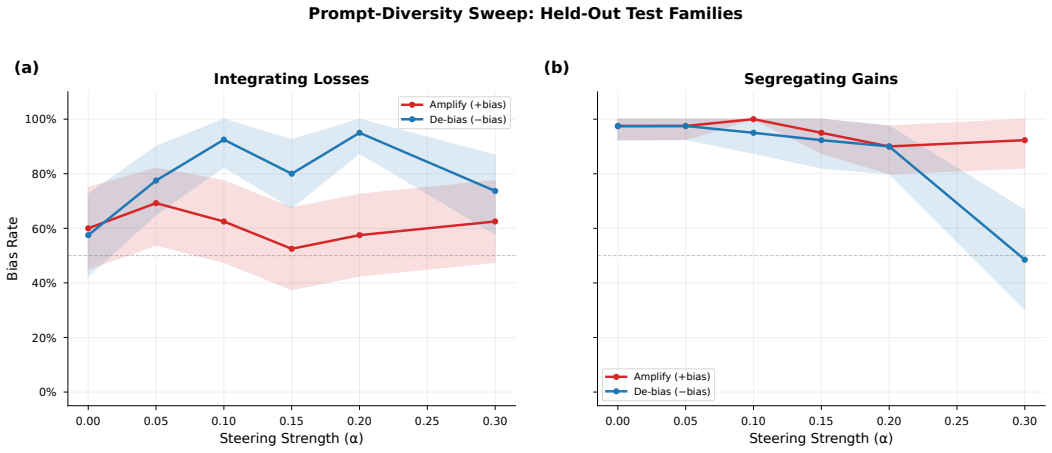


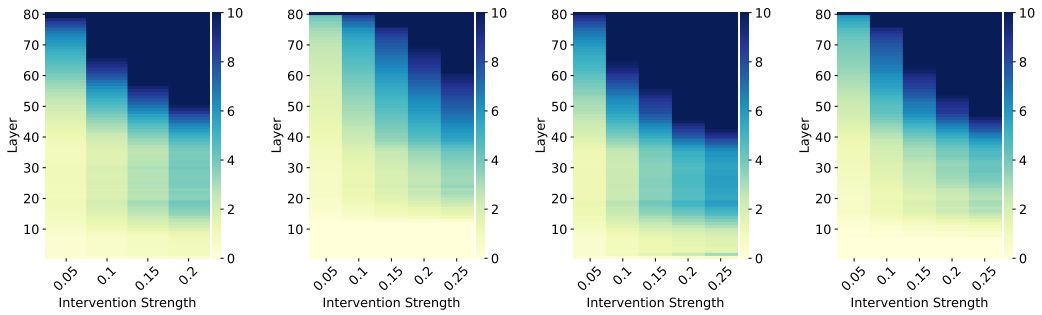
Figure S8: Prompt-diversity generalization: dose-response on Jaccard-disjoint test-family prompts for the SG and IL dimensions. The canonical steering vector, trained on a single prompt family, transfers to structurally novel prompts (SG/fung $\rho = -0.943$, $p = 0.005$). Bootstrapped 95% CIs (10,000 replicates).

E MECHANISM AND QUALITATIVE EVIDENCE

E.1 MECHANISTIC EFFECTS OF INTERVENTIONS.

To evaluate how causal intervention separates mental accounting and fungibility behaviors within the model, we analyze the l_2 distances between activations across layers and varying intervention strengths (α) (Figure S9). The intervention modifies the activations of high-performing layers by progressively steering them toward either the mental accounting or fungibility direction. The strength of the adjustment directly impacts the magnitude of the change across layers.

The results show that l_2 distances between mental accounting and fungibility activations increase with both layer depth and intervention strength, indicating a growing separation between the two activation trajectories. Deeper layers demonstrate greater sensitivity to the intervention, exhibiting larger distances as information propagates through the model. Higher α values further amplify these distinctions, changing both activation geometry and outputs under this protocol.



(a) Integrate Losses (b) Segregate Gains (c) Sunk Costs (d) Transaction Utility

Figure S9: Distance between activations along the mental accounting and fungibility directions.

E.2 Example Responses after Steering

Figure S10 presents two representative transcripts illustrating how steering reshapes the model's free-text reasoning. In the integrate-losses scenario (upper panel), the mental-accounting–steered response highlights the distress of categorizing losses into separate "accounts," consistent with segmented losses feeling more painful than aggregated ones, whereas the fungibility-steered response ignores how losses are divided and evaluates only the total amount owed. In the sunk-cost scenario (lower panel), the mental-accounting–steered response shows resistance to spending again after the irrecoverable loss ("paying twice" within one account), whereas the fungibility-steered response treats the new decision as independent of the past loss.

Steered Towards Mental Accounting | **Integrate Losses** | **Steered Towards Fungibility**

1. Both individuals (A & B) receive letters stating they owe money due to an error in their returns.
2. The amounts are relatively small ($150 for A and $150 for B).
3. However, both authorities involved. This means individual "A" receives **two separate notifications of errors resulting in additional taxes owed**, whereas Individual "B" only gets one notification. The psychological impact could be greater when **receiving multiple notices** with different authorities because it **reinforces feelings of failure and increases anxiety about dealing with government agencies**.

Therefore, this leads me to conclude that person "A", who **got hit twice, at both federal and state levels**, not just once like person "B," would likely feel more upset than person "B." This makes option `a` - *Mr. A was more upset* the most logical choice among options provided.

Both individuals receive letters stating they owe money due to mistakes in their tax returns. The amount of money owed can affect one's emotional response.
Now let's compare the amounts:
- Mr. A owes $100 (federal) + $50 (state) = $150
- Mr. B owes $150

We see that **both individuals ultimately owe the same total amount ($150)**. However, the circumstances are different for each individual.
Considering this information, and since there are **no additional consequences mentioned beyond owing money**, I conclude that both individuals will likely experience similar levels of upset. Therefore, my answer is: no difference.

Sunk Costs

1. **I've already spent money on something (the original ticket) that now has no value because it's lost**.
2. Paying again would essentially mean **paying twice for the same experience**.
3. 3. Since there is no way to recover the first ticket, buying another one wouldn't feel like getting two tickets; rather, it **feels like throwing away an additional $10** just so I can watch this play.

So, considering these points, if someone asked me whether they should charge me extra cash right at the entrance door when all I want to do is enjoy some entertainment after spending hours deciding what show to attend... My reaction will probably sound **more bitter than "No"**. It'll likely resemble "Are you kidding?!?"! Hence, my final response is: "No."

1. I've already paid $10 for the original ticket.
2. Unfortunately,I lose the ticket and it can't be recovered.

Now, (and here is where many people get stuck), should I spend an additional $10 on another ticket?
The key insight here is that **paying again doesn't change what happened in the past** - namely losing the first ticket. That event has occurred, and spending more money won't undo it.
So, I should **focus solely on whether attending the play with a new ticket would still bring me value worth at least $10**. If yes, then buying another ticket makes sense.
Considering only my desire to attend the play, without any attachment to sunk costs, making rational economic decisions involves **evaluating each choice based on its own merits rather than being influenced by previous investments**.
Therefore, **if seeing the play provides enough enjoyment or benefits equivalent to at least $10**. then purchasing another ticket might make perfect logical sense! <answer>Yes</answer>.

Figure S10: Example responses illustrating the effects of steering toward mental accounting and fungibility in two scenarios (Integrate Losses, Sunk Cost). Text aligned with mental accounting is highlighted in red; text aligned with fungibility in green.

F Transaction Utility

F.1 Transaction Utility: Paired Willingness-to-Pay Design

This paradigm uses Thaler's (1985) numeric WTP elicitation with paired premium/regular venues. Each of 10 scenarios is presented in both a premium context (resort beach bar, airport kiosk, stadium

concession) and a regular context (grocery store, corner deli), yielding 20 conditions per alpha level. The intervention sweep covers 8 alpha values $\times$ 2 directions $\times$ 20 conditions $\times$ 24 samples = 3,840 runs at 99.3% parse validity.

The earlier binary A/B format ("Who made the better purchase?") was a paradigm confound: it turned the task into arithmetic, destroying the bias signal at baseline (0% bias, nothing for steering to amplify). Numeric WTP recovers the bias signal and confirms bidirectional steering.

Key results:

- Ment direction: raw WTP 7.81 $\rightarrow$ 8.47 ($\rho = +0.976$, $p < 0.0001$). Premium WTP rises faster (8.27 $\rightarrow$ 9.11, $\rho = +1.000$) than regular (7.34 $\rightarrow$ 7.83, $\rho = +0.898$). TU gap widens: $+0.93 \rightarrow +1.28$ ($\rho = +0.786$, $p = 0.021$).

- Fung direction: raw WTP 7.83 $\rightarrow$ 7.28 ($\rho = -0.810$, $p = 0.015$ over the full α grid; $\rho = -0.714$, $p = 0.071$ restricted to $\alpha \leq 0.30$). Regular WTP drops faster ($\rho = -0.857$, $p = 0.007$). Binary logistic $\beta = -2.31$, $p < 0.001$.

- Binary metric ceiling-blocked ($\sim$95% baseline), consistent with SG/ment.

LLaMA-3-8B capacity control. Under the identical paired paradigm, LLaMA-3-8B exhibits transaction utility at baseline—premium$-$ordinary gap $+1.07$ at $\alpha = 0$ ($P(\text{prem} > \text{reg}) = 0.80$ across the ten matched pairs, $t = 2.91$, $p = 0.017$), comparable to LLaMA-3-70B—yet steering moves neither the level (ment $\rho = +0.00$, $p = 1.0$; fung $\rho = -0.17$, $p = 0.69$) nor the transaction-utility gap itself (ment $\rho = +0.12$, $p = 0.78$; fung $\rho = 0.00$, $p = 1.0$), against LLaMA-3-70B's widening gap ($\rho = +0.786$, $p = 0.021$). The bias signature is present at baseline but causally immovable: surface transaction utility without a steerable representation. The same intervention does steer LLaMA-3-8B on integrate-losses, so the null is representation-specific, not a tooling failure.

Figure S11 shows the venue-decomposed dose-response.

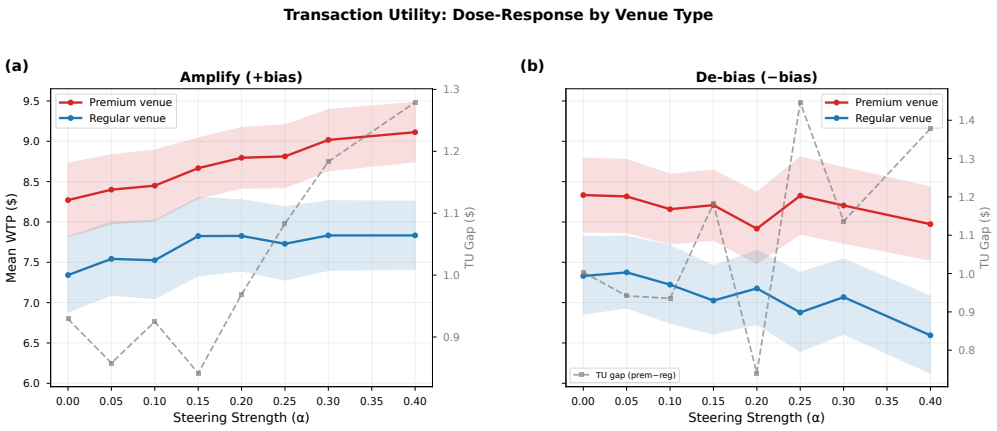


Figure S11: Transaction utility decomposed by venue (LLaMA-3-70B), refining the mean-WTP transaction-utility column of the main-text dose-response figure. Premium-venue (red) and ordinary-venue (blue) mean WTP are plotted separately, with the premium$-$ordinary gap (dashed, right axis). (a) Amplifying mental accounting: mean WTP rises monotonically ($\rho = +0.976$) and premium-venue WTP rises faster, so the transaction-utility gap widens ($\rho = +0.786$, $p = 0.021$), the signature of context-dependent valuation rather than a uniform price increase. (b) De-biasing toward fungibility: mean WTP falls ($\rho = -0.810$); the premium$-$ordinary gap is noisy in this direction, so we claim a monotonic gap effect only under amplification. Bootstrapped 95% CIs.

G PROBING AND INTERVENTION PROCEDURES

Locating the representation. Transformer models update a shared residual stream x_l layer by layer,

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_{l,h}\,\mathrm{Att}_{l,h}(P_{l,h}x_l), \tag{1}$$

where the sum runs over the model's H attention heads, $P_{l,h}$ and $Q_{l,h}$ project into and out of head h's space, and $\mathrm{Att}_{l,h}$ is its attention operation. For each layer we take the hidden state at the final prompt token and fit an ℓ_2-regularized logistic regression predicting whether the model's answer follows mental accounting or fungibility, with labels taken from the model's own responses. Probes use template-family-stratified grouped cross-validation: all paraphrases of a scenario stay in one fold, and structurally matched seed families are not split across train and test folds. We report held-out AUC per layer.

Steering vector and intervention. We construct paired prompts that elicit contrasting mental-accounting and fungibility behavior (examples in *SI Appendix*, Stimulus, Prompt S7), take the difference of their final-token activations (Rimsky et al., 2024; Turner et al., 2024; Singh et al., 2024), and average these differences layer by layer across the 283 pairs; this mean-difference direction is the steering vector $\nu_l^{\mathrm{FG}\to\mathrm{MA}}$. We intervene at each dimension's most expressive layers—the band where the steering vector projects most strongly onto held-out activations (*SI Appendix*, Layer-Resolved Projection Analysis), which places segregating gains and integrating losses in the upper layers (63–75) and sunk cost at mid layers (20–45)—by editing the stream during generation,

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_{l,h}\,\mathrm{Att}_{l,h}(P_{l,h}x_l) + \mathrm{sign}(\text{direction})\;\alpha\;\nu_l^{\mathrm{FG}\to\mathrm{MA}}, \tag{2}$$

with strength $\alpha > 0$ and $\mathrm{sign}(\text{direction}) = +1$ for mental accounting and -1 for fungibility. Dose-response curves are fitted with a logistic model for bias rates and a linear model for WTP, each with a bootstrap 95% band over model responses to the randomized sweep prompts. We report intervention strengths within the primary range $|\alpha| \le 0.30$, which preserves unrelated general-knowledge performance (*SI Appendix*, Capability Preservation); extended detail, including per-dimension intervention layers, the logistic-regression, random-vector, and specificity controls, capability preservation, per-model steering statistics, and prompt-diversity generalization, appears across *SI Appendix*.