UTILITY-PRESERVING RELABELING REVEALS SEMANTIC SENSITIVITY IN LANGUAGE-MODEL DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

A decision should not change when an action is given a different label but its payoff stays the same. We study whether language models satisfy this form of strategic invariance. We introduce CSS-Bench, a benchmark of Prisoner's Dilemma, Chicken, and Stag Hunt games in which action labels are swapped while the underlying payoff structure remains fixed. Across nine open-weight models, relabeling changes decision accuracy substantially, with absolute gaps of up to 83.4 percentage points across model-vocabulary cells. The effect varies in both magnitude and direction across models and vocabularies, and is not restricted to semantically loaded labels. We then trace the effect inside the models. Label-related information becomes linearly decodable at an average normalized depth of 0.25, while activation patching identifies causal influence at a substantially later depth of 0.76 across 36 model-vocabulary cells. Thus, information about action labels can be represented well before it becomes causally relevant to the decision. Causal steering further changes decision behavior, but transfers substantially less reliably to unseen labels than the underlying representation. Together, these results show that language models can change their decisions when action labels are swapped without changing the underlying payoffs, and provide a controlled framework for studying how linguistic information influences decision-making.

1 INTRODUCTION

A decision should not change when an action is given a different name but its consequences stay the same. If one action has a higher payoff than another, replacing its label with a synonym, a socially loaded word, or a morally positive word does not change that payoff. A decision-maker that follows the underlying objective should therefore make the same choice. However, for language models this invariance is not guaranteed. Models may rely on associations learned for an action label instead of the stated payoffs. We therefore ask: *when a language model makes a strategic decision, does it follow the payoff structure, or can the meaning of an action's label change what it chooses?*

We study this question through *strategic invariance*. We consider decision problems in which the game, rules, and payoffs remain fixed while the linguistic labels assigned to actions are swapped. A strategically invariant decision procedure should produce the same choice under this utility-preserving transformation. We use *semantic sensitivity* to refer to a model's decision changing with the semantic or lexical identity of an action label, even when that identity is irrelevant to the underlying payoff structure.

Language models are known to be sensitive to how tasks are described. Changes in wording, formatting, or framing can alter their outputs even when the underlying task remains unchanged (Akata et al. (2025); Sclar et al. (2024); Zhuo et al. (2024); Chatterjee et al. (2024)). This sensitivity matters for strategic decision-making because language is not merely a description of the decision problem. It is also the input through which the model computes its decision. Recent work has studied LLM behavior in negotiation, cooperation, bargaining, and other strategic settings (Park et al. (2023); Horton et al. (2026); Silva (2024); Xie et al. (2025)). However, these settings generally treat language as part of the task context rather than isolating whether semantically irrelevant action labels can change a decision. We isolate this effect by swapping the two action labels in otherwise identical prompts.

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.

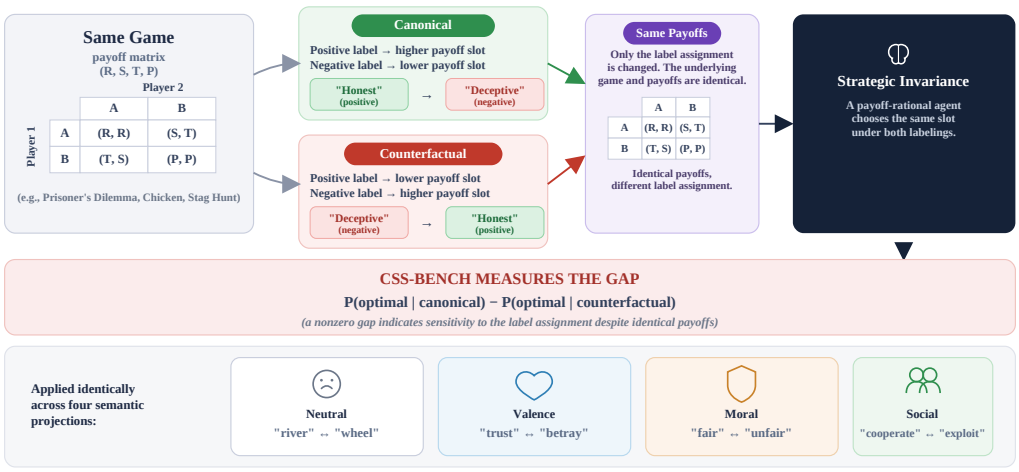

Figure 1: **Utility-preserving relabeling in CSS-Bench.** The game and payoff structure remain fixed while the labels assigned to the two action slots are swapped. A utility-consistent model should therefore make the same decision in both conditions.

However, behavioral changes alone do not reveal how they arise. A model can represent a label without using it to choose an action. We use probes to test what can be decoded from its activations and interventions to test which changes affect its output. Mechanistic interpretability provides tools to distinguish these possibilities. Probing can identify whether information is represented in activations, but does not establish that the information is used (Belinkov (2022)). Activation interventions can instead test whether manipulating that information changes behavior (Geiger et al. (2021); Vig et al. (2020); Meng et al. (2022; 2023)). We therefore study both where label-semantic information becomes represented and where it becomes causally relevant to the decision.

We introduce **C**ounterfactual **S**trategic **S**ynthesis Benchmark (**CSS-Bench**), a controlled benchmark for testing utility-preserving decision invariance in language models (Figure 1). It covers Prisoner's Dilemma, Chicken, and Stag Hunt and generates matched prompt pairs in which the payoff structure is identical but the action labels are swapped. The benchmark uses multiple label vocabularies, allowing us to test whether the effect depends on the semantic category of the labels. Because the same labels appear in both conditions, the matched prompts also support direct activation-level comparisons.

We find four main results. First, relabeling substantially changes model decisions, with effects that vary in both magnitude and direction across models and vocabularies. Second, label-semantic information becomes decodable early in the network, while activation patching identifies substantially later layers at which manipulating the corresponding representations changes the decision. Third, patching label-token positions has little effect, whereas patching the final position recovers most of the behavioral change, suggesting that the relevant information is integrated into the decision representation. Fourth, steering a direction learned at the causally influential depth can strongly alter decisions and sometimes transfers across game types, but transfer to unseen labels is substantially weaker. Because steering can also push accuracy beyond the model's canonical performance, we interpret it as evidence of causal control rather than as a general debiasing method.

Our contributions are:

- **Strategic invariance:** We formalize utility-preserving relabeling as a controlled test of whether decisions remain tied to the underlying payoff structure.
- **CSS-Bench:** We introduce a procedurally generated benchmark spanning Prisoner's Dilemma, Chicken, and Stag Hunt, with matched label swaps, multiple semantic vocabularies, and unseen-label evaluation.
- **Mechanistic analysis:** We combine behavioral evaluation, probing, activation patching, and steering to distinguish where label-semantic information is represented from where it becomes causally influential, and to test whether that influence can be manipulated.

2 Related Work

2.1 LLM Decision-Making and Strategic Behavior

LLMs are increasingly studied as decision-makers in games, negotiation, auctions, and other economic settings (Akata et al. (2025); Fontana et al. (2024); Park et al. (2023); Horton et al. (2026); Silva (2024); Xie et al. (2025)). In canonical games such as Prisoner's Dilemma and Stag Hunt, models often deviate from rational-agent predictions through positional, payoff, and behavioral biases (Herr et al. (2024)). Recent work further shows that models can represent strategic environments accurately while still exhibiting systematic action-space biases in their final decisions (Sobotka et al. (2026); Pilli & Nallur (2026)). Strategic behavior can also depend on contextual framing beyond the formal game structure (Lorè & Heydari (2024); Jia et al. (2025)). Our work isolates one such factor by holding the game and payoff structure fixed while changing only the labels assigned to actions.

2.2 Semantic and Framing Effects

LLM behavior can shift under differences in wording, formatting, and framing even when the underlying task is unchanged (Sclar et al. (2024); Akata et al. (2025); Zhuo et al. (2024); Chatterjee et al. (2024)). Studies of content effects similarly show that semantic information can influence judgments beyond formal logical structure in both humans and language models (Dasgupta et al. (2024); Bertolazzi et al. (2024); Ozeki et al. (2024)). Recent work has documented framing effects in LLM evaluations and decisions (Lim et al. (2026); Kořínek & Štekerová (2026); Cheng et al. (2026); Lin & Chen (2023); Sun et al. (2025)). These findings motivate our focus on action-label semantics, but do not isolate whether changing a semantically irrelevant label can alter a decision while the underlying utility structure remains identical.

2.3 Behavioral Invariance

A related line of work asks whether transformations that should leave a task unchanged also leave model behavior unchanged. Paraphrase and formatting changes can produce substantial behavioral variation (Sclar et al. (2024)). Recent benchmarks extend this idea to decision-making through payoff-preserving reframing and other meaning-preserving transformations (Davoudi et al. (2026); Elkins & Chun (2026); Hwang et al. (2026); de Zarzà et al. (2026); Wasi et al. (2026)).

CSS-Bench applies a more controlled form of this invariance test. Rather than changing a scenario, narrative, or task description, we keep the game, rules, and payoff matrix fixed and swap only the labels assigned to the two actions.

2.4 Mechanistic Analysis of Model Decisions

Probing and causal interventions provide complementary ways to study such effects. Probes test whether information is linearly decodable from intermediate representations (Alain & Bengio (2018); Belinkov (2022)), but decodability alone does not show that the model uses that information (Belinkov (2022)). Activation patching instead tests whether changing an internal representation changes model behavior and has been used to localize factual associations and circuits (Meng et al. (2022); Wang et al. (2023); Olsson et al. (2022)). Its conclusions can also depend on the localization method (Zhang & Nanda (2024); Fang et al. (2026); Khanh et al. (2026)).

Activation steering offers a complementary form of intervention by directly modifying residual-stream representations. It has been used to control style, sentiment, refusal behavior, and semantic effects (Turner et al. (2024); Rimsky et al. (2024); Arditi et al. (2024); Valentino et al. (2026)). Valentino et al. (2026) show that fixed steering often fails to fix content bias in syllogism tests, so they use input-dependent activation steering instead. We combine these approaches to distinguish three questions: when label information becomes decodable, when it becomes causally relevant to the decision, and whether that representation can be manipulated to change behavior.

3 STRATEGIC INVARIANCE: A FORMAL FRAMEWORK

Decision setting: A two-player normal-form matrix game is defined as:

$$G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle \tag{1}$$

where $N = \{1, 2\}$ and each player chooses an action from $A_i = \{a_1, a_2\}$. We call the two actions Slot A (paying R if the opponent plays A, S if B) and Slot B (paying T if the opponent plays A, P if B). The game topology is fixed by the ordinal relationship among R, S, T, P:

$$\mathbb{T}_{\text{PD}} : T > R > P > S, \quad 2R > T + S \tag{2}$$

$$\mathbb{T}_{\text{Chicken}} : T > R > S > P \tag{3}$$

$$\mathbb{T}_{\text{StagHunt}} : R > T > P > S \tag{4}$$

Definition 1 (Task)**.** *Given a two-action game with payoffs (R, S) for action A and (T, P) for action B, assume that the opponent selects each of its two actions with equal probability.*

The expected payoffs for the row player are:

$$U_A = \frac{R + S}{2}, \qquad U_B = \frac{T + P}{2}. \tag{5}$$

The correct action is the one with the higher expected payoff:

$$a^* = \arg \max_{a \in \{A, B\}} U_a. \tag{6}$$

Our benchmark does not attempt to identify a unique game-theoretic solution concept. Instead, it isolates a simpler property and asks whether models consistently compare the underlying numerical payoffs when the linguistic labels attached to the actions are changed. For Prisoner's Dilemma, Slot B is correct for every valid instance since $T > R$ and $P > S$ together imply $T + P > R + S$ algebraically. We use this as a parameter-free sanity check on our dataset generator (verified to hold for 100% of generated PD instances).

Definition 2 (Utility-preserving semantic transformation)**.** *Let $L : \{A, B\} \to \Sigma$ label each slot from vocabulary Σ. A transformation $T(G, L) = (G, L')$ is utility-preserving if:*

$$L \neq L' \quad \textit{and} \quad \arg \max_{a \in \{A, B\}} U_a(G, L) = \arg \max_{a \in \{A, B\}} U_a(G, L') \tag{7}$$

leaving both the payoff function and the correct slot (Definition 1) unchanged.

Definition 3 (Strategic invariance)**.** *We study a narrow form of strategic invariance: whether a model preserves a payoff-based decision under utility-preserving changes to action labels. A decision procedure f_θ satisfies strategic invariance with respect to transformation T if:*

$$f_\theta(G, L) = f_\theta(G, L'), \quad \forall (G, L') = T(G, L) \tag{8}$$

Each non-neutral semantic category supplies a positive pool Σ^+ and a negative pool Σ^- (e.g., valence: $\{\text{Trust}, \text{Share}, \dots\}$ / $\{\text{Betray}, \text{Steal}, \dots\}$). For each game, we create two prompts using the same two labels: f_{canon} assigns the positive-pool word to the correct slot and the negative-pool word to the incorrect slot; f_{counter} swaps them. $T(G, L_{\text{canon}}) = (G, L_{\text{counter}})$ is then exactly a utility-preserving transformation, and both projections can be scored against the identical token pair: $(\text{id}(\text{pos_word}), \text{id}(\text{neg_word}))$.

4 CSS-BENCH: DATASET AND MODEL SUITE

Label Categories: CSS-Bench uses four semantic categories: `neutral`, `valence`, `moral`, and `social`. Table 1 shows the categories. Each category provides matched label pools that satisfy the cross-model single-token constraint. Every instance is constructed as a matched canonical/counterfactual pair using the identical two literal labels. This ensures that any behavioral difference can be attributed to the reassignment of the labels to the two payoff-defined action slots rather than to a change in the underlying game or vocabulary.

Table 1: Semantic projection categories used in CSS-Bench. Each category is evaluated using the same canonical/counterfactual label-swap procedure.

Category	Example pair	Purpose
`neutral`	River / Wheel	Concrete labels without an intended valence, moral, or social ordering; provides a null-calibration condition.
`valence`	Trust / Betray	General positive/negative semantic connotation.
`moral`	Good / Evil	Explicit moral framing.
`social`	Cooperate / Exploit	Social and relational framing that is more directly connected to strategic interaction.

Model suite: We evaluate nine open-weight models spanning 0.41-8.03B parameters and five architecture families, shown in Section B. The suite includes a matched base/instruction-tuned pair, Llama-3.2-1B and Llama-3.2-1B-Instruct, which share the same underlying model scale, as well as a within-family Qwen2.5 scale comparison.

5 Behavioral Evidence

5.1 Canonical and counterfactual decisions

We first tested whether changing the label mapping changed model decisions. Table 2 reports accuracy under the canonical and counterfactual renderings for all nine models and four label vocabularies. The accuracy gap reached $83.4\,pp$ for Llama-3.2-1B on moral labels, and remained significant after Benjamini-Hochberg correction in 33 of 36 model-vocabulary cells. The three non-significant cells were Gemma-2-2B on valence, Mistral-7B on moral, and Llama-3.2-1B-It on social.

The label swap changed accuracy substantially in many cells, but not in a consistent direction. For neutral labels, seven of nine models had higher accuracy after the label assignments were swapped. Llama-3.2-1B-It, for example, increased from 15.0% accuracy under the canonical mapping to 80.9% under the counterfactual mapping, a $65.9\,pp$ increase. Mistral-7B showed a similar increase, from 38.0% to 89.5%. In these cases, changing which word referred to which payoff slot moved the model toward the payoff-consistent decision.

For other vocabularies, the same label swap often moved accuracy in the opposite direction. Llama-3.2-1B fell from 88.0% to 4.6% on moral labels, a $83.4\,pp$ decrease. Qwen2.5-3B fell from 71.6% to 14.5% on valence labels, a $57.1\,pp$ decrease. The direction could also vary within the same model: Mistral-7B improved on valence labels ($50.7\% \rightarrow 72.5\%$) and social labels ($58.8\% \rightarrow 71.9\%$), but its accuracy increased even more on neutral labels and changed little on moral labels.

Thus, changing only the assignment of words to the two action slots can move model accuracy either toward or away from the payoff-consistent choice. The direction depends on both the model and the vocabulary. We therefore report signed Δ alongside $|\Delta|$: $|\Delta|$ measures how strongly the label swap changes behavior, while the sign records whether the counterfactual rendering improves or reduces payoff-consistent accuracy. These behavioral results establish that label assignment affects decisions, but they do not by themselves explain why the effect changes direction across conditions.

5.2 What varies across vocabularies

If the effect reflected a preference for positively valenced or socially desirable actions, neutral placeholder nouns should have little effect. They do not. Neutral labels produced the largest mean $|\Delta|$ across the four vocabularies: $29.9\,pp$, compared with $24.1\,pp$ for valence, $27.1\,pp$ for moral, and $17.4\,pp$ for social labels (Figure 2b).

The effect is therefore not driven by semantic content alone. Action-label identity matters even when the labels are nominally neutral, while semantic category modulates the size and direction of the effect. We therefore do not interpret the ordering of the four vocabularies as a semantic ranking. The stronger result is that neutral relabeling can alter decisions at least as much as semantically loaded relabeling. Moreover, we ran a small API-based check on GPT-4o using 1,000 neutral-label pairs. Accuracy increased from 55.0% under the canonical rendering to 78.0% under the

Table 2: **Behavioral results:** Accuracy (%) under canonical (c.) and counterfactual (ct.) renderings and the gap Δ in percentage points, for each model and label vocabulary ($n = 3{,}000$ pairs per cell). † marks the three cells that are not significant after Benjamini–Hochberg correction Section E.

	neutral			valence			moral			social		
Model	c.	ct.	Δ	c.	ct.	Δ	c.	ct.	Δ	c.	ct.	Δ
Pythia-410M	27.5	48.1	**-20.6**	36.7	36.7	**-0.0**	25.7	54.6	**-28.9**	34.0	34.5	**-0.5**
Llama-3.2-1B	16.5	73.7	**-57.2**	57.5	28.6	**+29.0**	88.0	4.6	**+83.4**	68.2	21.7	**+46.5**
Llama-3.2-1B-It	15.0	80.9	**-65.9**	63.3	43.0	**+20.3**	63.8	31.6	**+32.2**	44.3	44.1	**+0.2**†
Qwen2.5-1.5B	30.7	37.9	**-7.2**	51.2	23.6	**+27.6**	48.8	30.4	**+18.4**	15.7	65.0	**-49.3**
Gemma-2-2B	33.4	34.0	**-0.6**	42.0	28.3	**+13.7**†	43.6	29.5	**+14.1**	35.1	33.0	**+2.1**
Qwen2.5-3B	25.4	46.7	**-21.3**	71.6	14.5	**+57.1**	56.3	30.9	**+25.3**	50.7	29.0	**+21.7**
Mistral-7B	38.0	89.5	**-51.6**	50.7	72.5	**-21.8**	57.7	51.1	**+6.6**†	58.8	71.9	**-13.1**
Qwen2.5-7B	27.4	50.6	**-23.2**	63.0	21.4	**+41.7**	50.5	48.3	**+2.2**	31.7	44.1	**-12.4**
Llama-3.1-8B	29.5	51.3	**-21.8**	43.3	37.6	**+5.8**	60.5	27.7	**+32.7**	30.8	41.8	**-11.0**

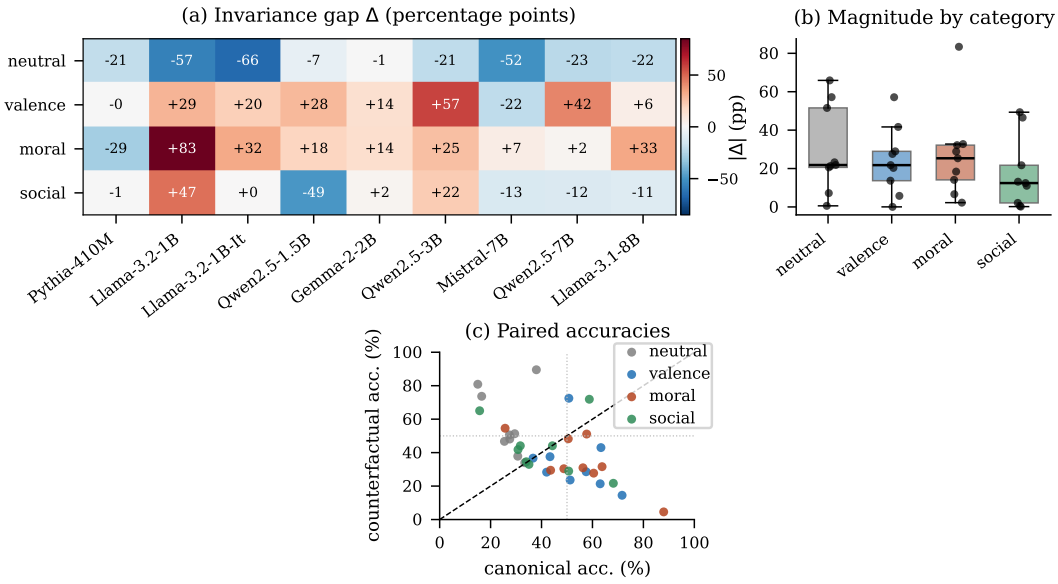

Figure 2: **Behavioural results.** (a) Signed gap Δ for every model×vocabulary cell; red is canonical-better, blue is counterfactual-better. The sign flips across vocabularies within a model and across models within a vocabulary. (b) Magnitude by vocabulary; neutral is not a null condition. (c) Paired accuracies; points off the diagonal are invariance violations, and most models sit below chance in at least one rendering.

counterfactual rendering (McNemar $p < 10^{-4}$, uncorrected). It provides a preliminary indication that the effect also occurs in a larger closed-weight model. Details are reported in Section F.2.

6 From Representation to Causal Influence

6.1 Semantic information emerges before causal influence

We next asked where the label effect appears in the model and when it begins to affect the decision. We trained three logistic probes on frozen residual-stream activations at the final prompt position, layer by layer. Probe **A** predicts the payoff-optimal slot, Probe **B** predicts the slot containing the positive-pool word, and Probe **C** predicts the model's eventual choice. Probes were trained on canonical prompts independently across network layers and evaluated on both canonical and counterfactual prompts.

Probe A showed a clear transition with depth. At earlier layers, its predictions transferred to the counterfactual prompts. At later layers, canonical accuracy reached 100% while counterfactual

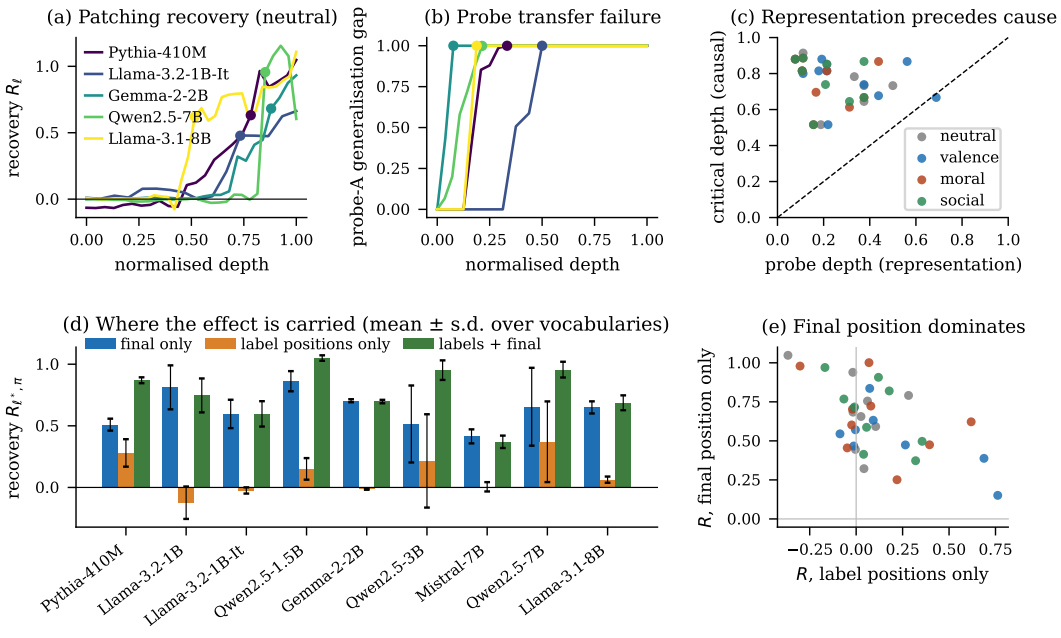

Figure 3: **Representation and causal depth.** (a) Patching recovery by layer. (b) Probe-A generalisation gap by layer. (c) Probe depth versus causal depth across 36 cells. (d) Recovery from patching the final position, label-token positions, or both. (e) Final-position versus label-token recovery across vocabularies.

accuracy fell to 0%. At the same depth, Probe B reached 100% counterfactual accuracy. Label identity therefore became fully decodable as payoff information stopped transferring. This transition occurred at a mean normalised depth of 0.25, ranging from 0.08 for Gemma-2-2B to 0.69 for Llama-3.2-1B on valence. The same split was retained at 81% of its same-word magnitude on held-out words, suggesting that the probes were not simply memorising individual labels (Section H).

We then used activation patching to ask when this information becomes causally relevant. At each layer, we replaced the counterfactual residual-stream activation at the final position with the corresponding canonical activation and measured recovery as

$$R_{\ell,\pi} = \frac{\Delta L^{pn}_{\text{patched}(\ell,\pi)} - \Delta L^{pn}_{\text{corrupted}}}{\Delta L^{pn}_{\text{clean}} - \Delta L^{pn}_{\text{corrupted}}}, \quad \text{where} \quad \Delta L^{pn}(x) = z(x)[p] - z(x)[n], \tag{9}$$

using the shared token pair (p, n). We define the causally influential depth ℓ^* as the earliest layer reaching half of the final recovery. Recovery remained near zero through roughly the first half to two-thirds of the network, then rose sharply (Figure 3a). The causally influential depth occurred at a mean normalised depth of 0.76, with little variation across vocabularies within a model (mean within-model s.d. 0.02; Table 7). Thus, label-related information was decodable early, but changing the representation did not substantially affect the decision until much later.

We also tested where this causal signal was carried. At the causally influential depth (ℓ^*) , patching the final prompt position produced greater average recovery than patching the label-token positions alone (0.64 versus 0.10). Patching both produced additional gains in some conditions (Table 9). These results show that changing the final-position activation can substantially affect the output at this depth, but do not rule out label information being used at earlier layers.

The gap between probing and patching is consistent across the experiment. Probe depth averaged 0.25, while causal depth averaged 0.76, with causal depth greater in 35 of 36 cells by a mean of 0.50 (Figure 3c). Therefore, information can be present before it becomes causally relevant. In our setting, relying on probe depth alone would place the relevant computation substantially earlier than the causal intervention does.

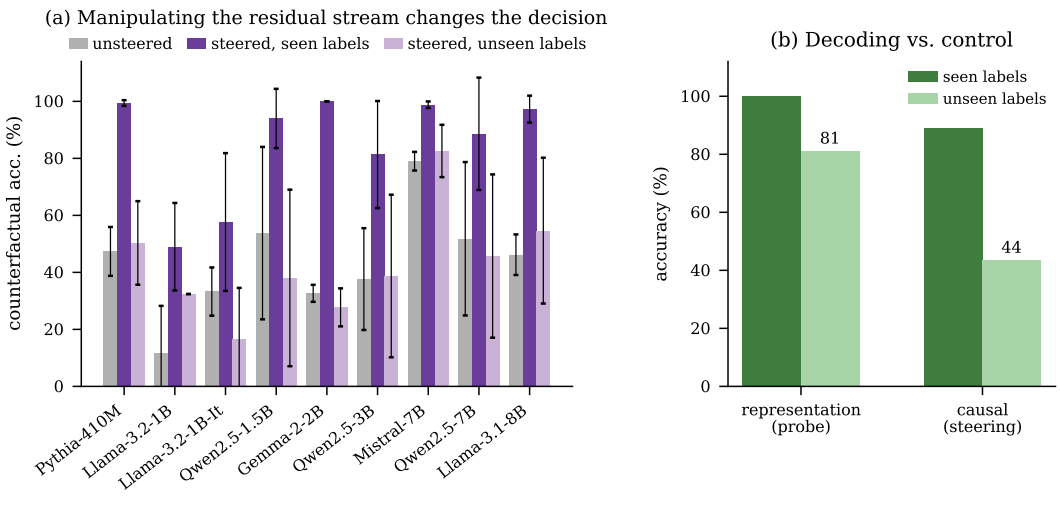

Figure 4: **Causal steering changes decisions.** (a) Counterfactual accuracy without and with steering on seen and unseen labels. (b) Probe and steering generalisation to unseen labels; steering transfers less reliably than the probe.

6.2 Causal manipulation of decision behavior

Patching identifies where the decision becomes sensitive to the label swap, but not whether that sensitivity can be controlled with a simple intervention. We tested this by computing

$$v_{\text{drift}} = \mathbb{E}\left[h_{\ell^*}^{\text{COUNTER}} - h_{\ell^*}^{\text{CANON}}\right] \tag{10}$$

from incorrectly answered counterfactual instances in a fit split. The vector v_{drift} is defined as the mean activation shift from canonical to counterfactual instances. At evaluation, we added αv_{drift} at the final position of ℓ^* that allowed us to test its causal effect on model choices. The steering coefficient α was swept over the interval $[0, 8]$ and selected using a dedicated validation split from the fit set. All reported final performance scores are evaluated on an untouched, disjoint test set featuring both seen and unseen label pairs. We also tested the conditional kNN gate of Valentino et al. (2026), and it produced similar results while steering fewer instances (Section K).

Steering substantially increased counterfactual accuracy on seen labels, from 43.9% to 86.6%, reaching 100% in 19 of 33 fittable cells (Table 12, Figure 4a). This shows that the representation at ℓ^* is behaviourally controllable. We do not interpret it as a repair: steered accuracy exceeded the model's own canonical accuracy by $46.6\,pp$ on average. In the most extreme case, Pythia-410M on neutral labels rose from 53.5% to 100%, despite a canonical accuracy of only 14.6%. Because the direction is fitted from the same correct-versus-incorrect criterion used for evaluation, it can encode a direct decision bias rather than specifically removing label-driven interference.

Transfer to unseen labels was substantially weaker. Across the 31 cells with both results, mean accuracy fell from 89% on seen labels to 44% on unseen labels (Figure 4b). This contrasts with the probe results, where 81% of the canonical/counterfactual split was retained on unseen words. Thus, information that generalises representationally does not necessarily yield a causal steering direction that generalises in the same way. A direction fitted on Prisoner's Dilemma also improved payoff-consistent accuracy on Chicken and Stag Hunt in 53/64 tested cases, suggesting that the intervention is not purely a slot-specific bias (Section L).

7 Discussion

Our results show that action-label identity can affect payoff-based decisions even when the underlying game is unchanged. Relabeling the same payoff structure changes model choices across many model-vocabulary cells, including with neutral labels (Section 5). The effect is not explained by label semantics alone: option order and payoff changes also produced substantial shifts in our controls (Section N). We therefore view label identity as one of several prompt features that can interfere with

strategic decisions. The effect is also template-dependent, so its magnitude should not be treated as a fixed property of the model.

The mechanistic results separate representation from causal involvement. Label-related information becomes linearly decodable around depth 0.25, while causal effects emerge around depth 0.76. This gap suggests that information can be present in the residual stream well before it becomes relevant to the final decision (Section 6.1). The location and magnitude of the effect also vary substantially across models, and we do not have enough evidence to explain these differences with a common model-level account.

Finally, steering provides evidence of causal control, not necessarily repair. A single direction at the causally influential depth can substantially change decisions and sometimes transfer across game topologies, but transfer to unseen labels is much weaker (Section 6.2). Steering can also push accuracy above the model's own canonical accuracy, so the intervention need not be removing a specific source of interference. Therefore, reporting the unsteered canonical accuracy alongside steering results is an important diagnostic when interpreting such interventions.

8 Limitations

1. **Decision Criterion:** CSS-Bench uses row-averaged payoff comparison rather than a specific equilibrium concept. Thus, the benchmark tests invariance under a controlled payoff comparison rather than general strategic competence.
2. **Lexical coverage:** The common vocabulary pools are limited, particularly for some categories. Results should therefore be interpreted as evidence of label sensitivity rather than an exhaustive estimate of semantic effects.
3. **Prompt dependence:** The magnitude and even direction of behavioral effects can depend on prompt rendering.
4. **Seed Sensitivity:** The benchmark generator is deterministic given a seed, but all experiments use a single seed. We have not evaluated whether the results generalize across independently generated benchmark instances and label assignments.
5. **Mechanistic scope:** Mechanistic analyses are restricted to open-weight models.
6. **Steering interpretation:** Steering demonstrates causal controllability but should not be interpreted as a general debiasing method, particularly because transfer to unseen labels is limited and interventions can overshoot.

9 Conclusion

We introduced CSS-Bench to study a simple invariance property of language-model decision making: changing the linguistic label assigned to an action should not change a decision when the underlying payoff structure is unchanged. Across a diverse set of open-weight models, we find that this invariance is frequently violated. The magnitude and direction of the effect vary across models and label vocabularies that portray that action-label information can become entangled with decisions that are formally determined by the same underlying payoffs. Our mechanistic analyses provide evidence for a corresponding internal process. Label-semantic information becomes decodable before the later layers at which interventions can causally alter the model's decision. This distinction suggests that the presence of information in a representation does not by itself imply that the information is being used by the decision computation. Causal steering further demonstrates that the relevant representations can be manipulated to change behavior, while weaker transfer to unseen labels shows that representation and causal control are not interchangeable.

More broadly, utility-preserving relabeling provides a simple experimental framework for separating a decision problem from the language used to describe it. The results suggest that evaluating language models only on their final decisions can miss an important source of behavioral variability: information that is irrelevant to the formal objective may nevertheless become coupled to the computation that produces the decision.

AI USE STATEMENT

In this work, we used generative AI tools only as auxiliary aids to check grammar, refine language, and polish the content for clarity. We have not used generative AI tools for research ideation, data collection, code execution, or formal proof derivation. All AI-assisted edits were thoroughly reviewed and verified by the authors, who take full responsibility for the final content of this manuscript.

ETHICS STATEMENT

This work studies a robustness property of language models used as decision-making components (e.g. in agentic or tool-use settings) and does not involve human subjects or sensitive personal data. We see no dual-use concern: identifying that models can be swayed by payoff-irrelevant relabeling, largely independent of what the relabeling means, and testing an inference-time mitigation, is intended to support safer deployment of such systems.

REPRODUCIBILITY STATEMENT

The dataset generator, model wrapper, and full analysis pipeline (behavioral detection, activation patching, linear probing, activation steering, cross-game transfer, and factorial regression) are implemented in our codebase. All reported numbers in this paper were read directly from those executed outputs; none were entered by hand from memory or estimated. The codebase and generated dataset will be released alongside the camera-ready version.

REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, May 2025. ISSN 2397-3374. doi: 10.1038/s41562-025-02172-y. URL http://dx.doi.org/10.1038/s41562-025-02172-y.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, November 2018. URL http://arxiv.org/abs/1610.01644.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 04 2022. ISSN 0891-2017. doi: 10.1162/coli_a_00422. URL https://doi.org/10.1162/coli_a_00422.

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13882–13905, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.769. URL https://aclanthology.org/2024.emnlp-main.769/.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.852.

Zirui Cheng, Joey Chan, Simo Du, Chenhao Tan, Yue Guo, and Hao Peng. Characterizing rhetorical misalignment in decision-making with language models, 2026. URL https://arxiv.org/abs/2608.14630.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks, July 2024. URL http://arxiv.org/abs/2207.07051.

Seyed Pouyan Mousavi Davoudi, Arshia Gharagozlou, Alireza Amiri-Margavi, Amin Gholami Davodi, and Hamidreza Hasani Balyani. Same Game, Different Story: A Minimal Conservative Strategic Robustness Benchmark for Large Language Model Agents, September 2026. URL http://arxiv.org/abs/2607.19670.

I. de Zarzà, J. de Curtò, Jordi Cabot, Pietro Manzoni, and Carlos T. Calafate. Semantic invariance in agentic ai, 2026. URL https://arxiv.org/abs/2603.13173.

Katherine Elkins and Jon Chun. Syntactic Framing Fragility: An Audit of Robustness in LLM Ethical Decisions, April 2026. URL http://arxiv.org/abs/2601.09724.

Feihao Fang, My T. Thai, and Yuanyuan Lei. Discovering a shared logical subspace: Steering LLM logical reasoning via alignment of natural-language and symbolic views. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38936–38956, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1806. URL https://aclanthology.org/2026.acl-long.1806/.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma?, September 2024. URL http://arxiv.org/abs/2406.13605.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Nathan Herr, Fernando Acero, Roberta Raileanu, María Pérez-Ortiz, and Zhibin Li. Are large language models strategic decision makers? a study of performance and bias in two-player non-zero-sum games, 2024. URL https://arxiv.org/abs/2407.04467.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus?, 2026. URL https://arxiv.org/abs/2301.07543.

Seojin Hwang, Minju Kim, Junhyuk Choi, JeongHyun Park, and Hwanhee Lee. Framing Matters: Addressing Framing Sensitivity in Decision-Making through Behaviorally-Grounded Value Alignment, May 2026. URL http://arxiv.org/abs/2605.28188.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. Llm strategic reasoning: Agentic study through behavioral game theory. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 58318–58350. Curran Associates, Inc., 2025. doi: 10.52202/085713-1955. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/54b13427fc2c9180f85c3d76cdcf076e-Paper-Conference.pdf.

Ly Tran Ho Khanh, Dongxuan Zhu, Man-Chung Yue, and Viet Anh Nguyen. Test-time diverse reasoning by riemannian activation steering. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i37.40407. URL https://doi.org/10.1609/aaai.v40i37.40407.

Milan Kořínek and Kamila Štekerová. Words matter:: How prompt framing shapes strategic behaviour in large language models. *Journal of Cases on Information Technology*, 28(1), 2026. ISSN 1548-7717. doi: https://doi.org/10.4018/JCIT.398628. URL https://www.sciencedirect.com/science/article/pii/S1548771726000059.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behaviour. *Nature Machine Intelligence*, 8(9):1411–1425, September 2026. ISSN 2522-5839. doi: 10.1038/s42256-026-01293-x. URL https://www.nature.com/articles/s42256-026-01293-x.

Kahee Lim, Soyeon Kim, and Steven Euijong Whang. DeFrame: Debiasing large language models against framing effects. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35672–35707, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1777. URL https://aclanthology.org/2026.findings-acl.1777/.

Yen-Ting Lin and Yun-Nung Chen. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In Yun-Nung Chen and Abhinav Rastogi (eds.), *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pp. 47–58, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlp4convai-1.5. URL https://aclanthology.org/2023.nlp4convai-1.5/.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14, 08 2024. doi: 10.1038/s41598-024-69032-z.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads, September 2022. URL http://arxiv.org/abs/2209.11895.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16063–16077, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.950. URL https://aclanthology.org/2024.findings-acl.950/.

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*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Stephen Pilli and Vivek Nallur. Predicting biased human decision-making with large language models in conversational settings. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, IUI '26, pp. 2084–2119, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400719844. doi: 10.1145/3742413.3789127. URL https://doi.org/10.1145/3742413.3789127.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. Prompt perturbation consistency learning for robust language models. In

Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1357–1370, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.91. URL https://aclanthology.org/2024.findings-eacl.91/.

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/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models'sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6c0e99d736da621403018ca7b32b1a4d-Paper-Conference.pdf.

Alonso Silva. Large language models playing mixed strategy nash equilibrium games, 2024. URL https://arxiv.org/abs/2406.10574.

Jan Sobotka, Mustafa O. Karabag, and Ufuk Topcu. Why do llms struggle in strategic play? broken links between observations, beliefs, and actions, 2026. URL https://arxiv.org/abs/2605.00226.

Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. Game theory meets large language models: A systematic survey. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 10669–10677. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1184. URL https://doi.org/10.24963/ijcai.2025/1184. Survey Track.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering, October 2024. URL http://arxiv.org/abs/2308.10248.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. Mitigating Content Effects on Reasoning in Language Models through Fine-Grained Activation Steering, April 2026. URL http://arxiv.org/abs/2505.12189.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Azmine Toushik Wasi, Mahir Absar Khan, Abdur Rahman, Wahid Faisal, Sukanta Saha, Saimon Bhuiyan, Mahdiya Rahman Sukanya, Rahatun Nesa Priti, Md. Iqramul Hoque, Munem Shahriar, Raima Islam, Sanatan Sushil, Shahriyar Zaman Ridoy, Kazi Rajwan Sultan, MD Shafikul Islam, Md Manjurul Ahsan, and Md Rizwan Parvez. Invariantbench: Can large language models exhibit inherent reasoning across equivalent transformations? In *3rd AI for Math Workshop: Toward Self-Evolving Scientific Agents*, 2026. URL https://openreview.net/forum?id=YWMnpol8i1.

Yutong Xie, Qiaozhu Mei, Walter Yuan, and Matthew O. Jackson. Using large language models to categorize strategic situations and decipher motivations behind human behaviors, 2025. URL https://arxiv.org/abs/2503.15752.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods, January 2024. URL http://arxiv.org/abs/2309.16042.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru Zhang. The geometry of reasoning: Flowing logics in representation space. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ixr5Pcabq7.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.108. URL https://aclanthology.org/2024.findings-emnlp.108/.

Appendix Table of Contents

A Additional Related Work **16**
A.1 LLM decision-making and strategic behavior 16
A.2 Framing and semantic sensitivity 16
A.3 Invariance and consistency in language models. 16
A.4 Activation patching 17
A.5 Activation steering 17
A.6 Probing classifiers. 17
A.7 Geometry of representation space 17

B Model Suite **17**

C Dataset Construction **18**
C.1 Tokenizer-safety filtering 18
C.2 Evaluation protocol 18

D Additional Framework Details **19**

E Full Behavioral Results **19**

F Additional Behavioral Analyses **22**
F.1 Numerical difficulty 22
F.2 Frontier-model spot check 22

G Causal Localisation **22**

H Probing **25**

I Representation-Space Geometry **25**

J Attention-Head Dissection **26**

K Steering **28**

L Cross-Topology Transfer **30**

M Factorial Regression **32**

N Behavioral Controls **32**

O Reproducibility **33**

A Additional Related Work

This section documents additional literature considered for Section 2 but summarized here for space, along with the rationale for each inclusion.

A.1 LLM decision-making and strategic behavior

A substantial literature evaluates LLMs in classical game-theoretic settings, including repeated Prisoner's Dilemma and public-goods games Akata et al. (2025); Fontana et al. (2024). Structured evaluations across GPT-3.5, GPT-4-Turbo, GPT-4o, and Llama-3-8B find these models affected by positional, payoff, or behavioral biases in two-player non-zero-sum games, with performance dropping substantially on the order of 25–34% in Stag Hunt and 16–34% in Prisoner's Dilemma when game configuration is misaligned with the relevant bias. Recent surveys catalog this growing body of work spanning negotiation, auctions, and multi-agent simulation Davoudi et al. (2026), while newer benchmarks move past Nash-equilibrium-only evaluation: one framework notes that most strategic-reasoning evaluations rely heavily on Nash Equilibrium benchmarks, overlook reasoning depth, and fail to reveal the mechanisms behind model behavior, motivating behavioral game-theoretic protocols that disentangle intrinsic reasoning from contextual influence. This literature consistently documents that framing-and payoff-irrelevant factors bias LLM decisions, but rarely asks where in the network that sensitivity originates the gap our proposed CSS-Bench's causal-localization approach targets.

A.2 Framing and semantic sensitivity

The proposed mechanism echoes content effects in human syllogistic reasoning, where believable conclusions are judged valid regardless of logical form Dasgupta et al. (2024); Bertolazzi et al. (2024). Evans et al. showed that when judging syllogism validity, participants were biased by whether the conclusion matched their prior beliefs, mistakenly endorsing invalid-but-believable syllogisms roughly 90% of the time, and follow-up work confirms LLMs show analogous content effects on reasoning tasks Dasgupta et al. (2024). Behavioral evidence for this in LLMs is now extensive: a systematic analysis across chain-of-thought, in-context learning, and supervised fine-tuning finds pretrained-LLM behavior explainable by cognitive-science heuristics, with only fine-tuning able to mitigate most reasoning biases while preserving consistency, and recent work shows content effects in LLMs extend beyond conditional and syllogistic inference to a broader range of propositional-logic rules, though the underlying mechanism remains poorly understood. Bilingual and cross-domain replications reinforce the pattern: the NeuBAROCO dataset finds leading LLMs exhibit reasoning biases similar to humans across English and Japanese syllogisms, and similar content effects appear in deontic and epistemic normative reasoning, where models make errors when conclusions contradict common-sense beliefs. Mechanistic work is beginning to close the gap: Valentino et al. (2026) show that content effects are linearly steerable at specific layers, motivating kNN-gated conditional steering (K-CAST) where static vectors fail, and activation-level abstraction methods have been proposed as a language-agnostic mechanism for improving formal-reasoning robustness against semantic sensitivity.

A.3 Invariance and consistency in language models.

A parallel line documents LLMs are broadly sensitive to input variation that should be behaviorally inert: paraphrase, synonym, and formatting perturbations degrade fine-tuned intent and slot performance substantially, with consistency-regularized training recovering only $\tilde{5}9$–69% of the loss Qiang et al. (2024). Prompt-format sensitivity has been shown to extend to formatting choices themselves, not just content Davoudi et al. (2026). Several concurrent 2026 benchmarks test payoff- or ground-truth-preserving invariance directly: one frames strategic robustness as invariance of action distributions under payoff-preserving narrative reframing (e.g., business vs. friendship framing of an identical social dilemma), evaluated via secondary analysis of published cooperation rates Davoudi et al. (2026); others document Syntactic Framing Fragility Elkins & Chun (2026) and analogous sensitivity in value-alignment judgments Hwang et al. (2026). All three, like the broader robustness literature Sclar et al. (2024), remain behavioral and black-box, establishing that invariance violations occur without offering a causal-localization account of where they arise in the network.

A.4 Activation patching

This is a form of causal mediation analysis, with the Gaussian-noise-corruption variant first introduced as *causal tracing* for locating factual associations, later generalized into path patching for more fine-grained circuit analysis. It has since been used to identify name-mover circuits in indirect object inference Wang et al. (2023) and induction-head formation Olsson et al. (2022). Subsequent synthesis work has emphasized that patching metric choice is not a free parameter, logit difference is generally preferred over probability or accuracy metrics because it is continuous and linear in residual-stream contributions, avoiding the artificial sharpness or discreteness of the alternatives Zhang & Nanda (2024).

A.5 Activation steering

Steering via residual-stream directions has been applied to style and sentiment control Turner et al. (2024) and safety-relevant refusal directions Rimsky et al. (2024); Arditi et al. (2024). More recent work complicates the single-direction picture, showing refusal is structured as a reusable core of shared latents supplemented by style-and domain-specific ones, and that linear interventions collapse this structure into uniform behavioral control. Kumaran et al. (2026) provide causal evidence linking steering to internal confidence signals governing abstention, with mediation analysis attributing a majority of the steering effect to confidence redistribution - a methodological template (behavioral correlation, causal intervention, mediation analysis) for establishing causal claims about internal representations more generally.

A.6 Probing classifiers.

Linear decodability of concepts from frozen representations is a standard interpretability tool Alain & Bengio (2018); Belinkov (2022). Kumaran et al. (2026) apply this to internal confidence representations, finding both calibrated and verbal confidence are lossy readouts of a richer underlying representation.

A.7 Geometry of representation space

Zhou et al. (2026) model chain-of-thought as continuous trajectories in representation space, showing trajectory velocity and Menger curvature rather than raw position - track logical structure invariantly across semantic carriers, offering a geometric complement to circuit- and direction-based accounts of model computation.

B Model Suite

Table 3: The nine open-weight models. Llama-3.2-1B appears in base and instruction-tuned form with the same architecture and parameter count.

Short name	HuggingFace id	Family	Params (B)	Layers	Instruct
Pythia-410M	`EleutherAI/pythia-410m`	GPT-NeoX	0.41	24	no
Llama-3.2-1B	`meta-llama/Llama-3.2-1B`	Llama	1.24	16	no
Llama-3.2-1B-It	`meta-llama/Llama-3.2-1B-Instruct`	Llama	1.24	16	yes
Qwen2.5-1.5B	`Qwen/Qwen2.5-1.5B-Instruct`	Qwen2	1.54	28	yes
Gemma-2-2B	`google/gemma-2-2b-it`	Gemma2	2.61	26	yes
Qwen2.5-3B	`Qwen/Qwen2.5-3B-Instruct`	Qwen2	3.09	36	yes
Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.3`	Mistral	7.25	32	yes
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	Qwen2	7.62	28	yes
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	Llama	8.03	32	yes

All models are loaded in `bfloat16` with a hook-based wrapper that exposes the residual stream at every layer boundary and supports patching and additive steering at arbitrary (layer, position) pairs. Prompts are left-padded with explicitly recomputed position ids so that batched forward passes are numerically identical to unbatched ones.

C Dataset Construction

C.1 Tokenizer-safety filtering

A label is *safe* for a model if the string `[`w tokenizes such that w occupies exactly one token and the `[` tokenizes identically regardless of which word follows. The second condition matters: on tiktoken-style BPE vocabularies, `[A]`, `[B]` and `[1]` appear often enough in web text that `[`+letter is frequently pre-merged into one atomic token, which would make the bracket itself carry label information.

We compute the per-model safe set and then take the intersection across all nine models, so every model sees byte-identical prompts. Table 4 gives the survivors. Attrition is heavy and uneven: Gemma-2-2B retains 54 neutral positives while Mistral-7B retains 7, and the intersection is bounded by the most restrictive tokenizer.

Table 4: Label pools surviving the cross-model intersection, and the by-word split used for unseen-label evaluation. "Fit" and "held-out" count matched pairs whose two labels both come from the corresponding half of the word pool; pairs mixing halves are dropped.

Vocabulary	positive words	negative words	fit pairs	held-out pairs
neutral	5	3	426	1168
valence	10	6	773	770
moral	4	2	758	760
social	8	2	749	735

The held-out pools for moral (2 positive, 1 negative) and social (4 positive, 1 negative) are thin, and unseen-label generalisation numbers for those two vocabularies should be read as indicative only.

C.2 Evaluation protocol

Every instance is rendered into a fixed prompt and evaluated as a forced choice between two single-token labels. The evaluation prompt is:

Game Decision Task

Game Environment: You are an agent in a decision task. The opponent will choose one of two actions, each with equal probability. Your score depends on both your action and the opponent's action.

- **Option [{label_A}]:** Yields {R} points if the opponent selects [{label_A}], and {S} points if the opponent selects [{label_B}].
- **Option [{label_B}]:** Yields {T} points if the opponent selects [{label_A}], and {P} points if the opponent selects [{label_B}].

Objective: Choose the option with the higher expected score, assuming the opponent selects each action with equal probability.

MY CHOICE: [

The model's response is not sampled or parsed as free-form text. Instead, we compare the logits assigned to the two label tokens as continuations of the trailing "`[`". The model selects the label with the higher logit. A label is considered usable only if it tokenizes to exactly one token as a continuation of the trailing "`[`" under the model's tokenizer. We take the intersection of single-token-safe labels across all nine models, ensuring that every model is evaluated on exactly the same set of instances. We generate $N = 1000$ matched pairs per topology per category (24,000 instances total: 4 categories $\times$ 3 topologies $\times$ 1000 pairs $\times$ 2 projections, evaluated on each of the nine models), and record the payoff margin of every instance.

D Additional Framework Details

Definition 4 (Strategic Invariance Index)**.** *For a given instance x, let $c(x)$ and $k(x)$ denote the token IDs of the correct and incorrect label slots, respectively. The logit margin under projection $p \in \{canon, counter\}$ is given by:*

$$\Delta L_p(x) = z_{-1}(f_p(x))[c(x)] - z_{-1}(f_p(x))[k(x)] \tag{11}$$

*The Strategic Invariance Index (*INV*) is defined as the ratio of logit margins across counterfactual and canonical projections:*

$$\mathrm{INV}(x) = \frac{\Delta L_{counter}(x)}{\Delta L_{canon}(x)} \tag{12}$$

where $\mathrm{INV} = 1$ denotes perfect strategic invariance, and $\mathrm{INV} \leq 0$ indicates a preference reversal under semantic relabeling.

Definition 5 (Probing Generalization Gap)**.** *For a probe target variable $y \in \{payoff_optimal, label_valence, model_decision\}$, let g_ℓ^y denote a linear classifier trained on the canonical training subset at layer ℓ. The probing generalization gap is defined as:*

$$\mathrm{PGap}_\ell^y = \mathrm{Acc}\left(g_\ell^y; \mathcal{D}_{canon}^{test}\right) - \mathrm{Acc}\left(g_\ell^y; \mathcal{D}_{counter}^{test}\right) \tag{13}$$

Definition 6 (Depth-Indexed Trajectory Geometry)**.** *Let $Y = (h_0, h_1, \ldots, h_L)$ represent the sequence of residual stream representations across network layers $0 \leq \ell \leq L$ at a fixed token position. The depth-indexed layer velocity v_ℓ is defined as:*

$$v_\ell = h_{\ell+1} - h_\ell \tag{14}$$

The discrete Menger curvature κ_ℓ at layer ℓ is given by:

$$\kappa_\ell = \frac{2\sqrt{1 - \cos^2\left(\angle(v_{\ell-1}, v_\ell)\right)}}{\|h_{\ell+1} - h_{\ell-1}\|} \tag{15}$$

following the geometric formulation of Zhou et al. (2026), adapted here to index neural network depth rather than text token position.

Failure Types as Formal Predicates. Let $f_\theta(G, L)$ denote the action slot chosen by model f_θ, and let $\mathrm{opt}(G)$ denote the payoff-optimal slot. Failure modes are categorized as formal predicates over utility-preserving transformation pairs (G, L, L') where $T(G, L) = (G, L')$:

$$\text{Type I}(f_\theta) \iff \exists (G, L, L') : f_\theta(G, L) \neq f_\theta(G, L') \tag{16}$$

$$\text{Type II}(f_\theta) \iff \text{Type I}(f_\theta) \wedge \mathrm{Corr}\left(f_\theta(G, L') - f_\theta(G, L),\ \mathbf{v}(L') - \mathbf{v}(L)\right) > 0 \tag{17}$$

$$\text{Type III}(f_\theta) \iff \text{Type II}(f_\theta) \wedge \left(f_\theta(G, L') \neq \mathrm{opt}(G)\right) \wedge \text{CausalMechanism}(L \to L') \tag{18}$$

where $\mathbf{v}(\cdot)$ denotes the label valence vector, and CausalMechanism$(\cdot)$ represents verified causal intervention via activation patching (Section 6.2).

Core Metrics & Vocabulary Projection. Let the vocabulary logits at the final prompt position for layer ℓ be given by:

$$z_\ell(x) = W_U \cdot \mathrm{LN}(h_\ell(x)) \tag{19}$$

where W_U is the unembedding matrix and $\mathrm{LN}(\cdot)$ represents Layer Normalization.

Numerical Difficulty Stratification. For every generated game instance G, we calculate the absolute raw payoff margin:

$$d(G) = |(R + S) - (T + P)| \tag{20}$$

Instances are partitioned into difficulty tertiles $\mathcal{B}(G) \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$ conditioned on their stratum $S_k = (\text{category}, \text{topology})$:

$$\mathcal{B}(G) = \begin{cases} \text{Easy}, & \text{if } d(G) > Q_{0.667}(d \mid S_k) \\ \text{Medium}, & \text{if } Q_{0.333}(d \mid S_k) < d(G) \leq Q_{0.667}(d \mid S_k) \\ \text{Hard}, & \text{if } d(G) \leq Q_{0.333}(d \mid S_k) \end{cases} \tag{21}$$

where $Q_p(d \mid S_k)$ represents the p-th empirical quantile of the margin distribution within stratum S_k.

E Full Behavioral Results

Table 5: Accuracy under both renderings and the median per-instance invariance index. $\mathrm{INV} = 1$ is exact invariance at the logit level; $\mathrm{INV} \leq 0$ indicates preference reversal.

Vocabulary	Model	canonical (%)	counterfactual (%)	Δ (pp)	median INV
neutral	Pythia-410M	27.5	48.1	-20.6	+0.464
	Llama-3.2-1B	16.5	73.7	-57.2	-0.333
	Llama-3.2-1B-It	15.0	80.9	-65.9	-0.857
	Qwen2.5-1.5B	30.7	37.9	-7.2	+0.833
	Gemma-2-2B	33.4	34.0	-0.6	+0.818
	Qwen2.5-3B	25.4	46.7	-21.3	+0.300
	Mistral-7B	38.0	89.5	-51.6	+0.072
	Qwen2.5-7B	27.4	50.6	-23.2	+0.429
	Llama-3.1-8B	29.5	51.3	-21.8	+0.444
valence	Pythia-410M	36.7	36.7	-0.0	+0.592
	Llama-3.2-1B	57.5	28.6	+29.0	-0.750
	Llama-3.2-1B-It	63.3	43.0	+20.3	-0.286
	Qwen2.5-1.5B	51.2	23.6	+27.6	+0.000
	Gemma-2-2B	42.0	28.3	+13.7	+0.937
	Qwen2.5-3B	71.6	14.5	+57.1	-0.562
	Mistral-7B	50.7	72.5	-21.8	-0.191
	Qwen2.5-7B	63.0	21.4	+41.7	-0.143
	Llama-3.1-8B	43.3	37.6	+5.8	+0.467
moral	Pythia-410M	25.7	54.6	-28.9	+0.351
	Llama-3.2-1B	88.0	4.6	+83.4	-1.000
	Llama-3.2-1B-It	63.8	31.6	+32.2	-0.733
	Qwen2.5-1.5B	48.8	30.4	+18.4	+0.100
	Gemma-2-2B	43.6	29.5	+14.1	+0.767
	Qwen2.5-3B	56.3	30.9	+25.3	-0.529
	Mistral-7B	57.7	51.1	+6.6	-0.694
	Qwen2.5-7B	50.5	48.3	+2.2	-0.615
	Llama-3.1-8B	60.5	27.7	+32.7	+0.250
social	Pythia-410M	34.0	34.5	-0.5	+0.900
	Llama-3.2-1B	68.2	21.7	+46.5	-1.000
	Llama-3.2-1B-It	44.3	44.1	+0.2	+0.077
	Qwen2.5-1.5B	15.7	65.0	-49.3	+0.083
	Gemma-2-2B	35.1	33.0	+2.1	+1.318
	Qwen2.5-3B	50.7	29.0	+21.7	+0.000
	Mistral-7B	58.8	71.9	-13.1	+0.250
	Qwen2.5-7B	31.7	44.1	-12.4	+0.333
	Llama-3.1-8B	30.8	41.8	-11.0	+0.714

Table 6: Matched-pair McNemar tests with paired Cohen's d, Benjamini–Hochberg corrected within each vocabulary across the nine models. "canon only" counts pairs correct under the canonical rendering and wrong under the counterfactual one, and vice versa.

Vocabulary	Model	canon only	counter only	χ^2	p	d	BH sig.
neutral	Pythia-410M	109	777	503.6	$< 10^{-12}$	-0.449	yes
	Llama-3.2-1B	220	2423	1836.2	$< 10^{-12}$	-1.256	yes
	Llama-3.2-1B-It	233	2379	1763.1	$< 10^{-12}$	-1.194	yes
	Qwen2.5-1.5B	562	1586	488.2	$< 10^{-12}$	-0.441	yes
	Gemma-2-2B	0	593	593.0	$< 10^{-12}$	-0.496	yes
	Qwen2.5-3B	343	1984	1157.2	$< 10^{-12}$	-0.792	yes
	Mistral-7B	114	2058	1739.9	$< 10^{-12}$	-1.175	yes
	Qwen2.5-7B	1214	1566	44.6	$2.45\text{e}{-11}$	-0.123	yes
	Llama-3.1-8B	234	721	248.3	$< 10^{-12}$	-0.300	yes
valence	Pythia-410M	413	325	10.5	$1.20\text{e}{-3}$	+0.059	yes
	Llama-3.2-1B	1271	821	96.8	$< 10^{-12}$	+0.183	yes
	Llama-3.2-1B-It	1299	598	259.0	$< 10^{-12}$	+0.307	yes
	Qwen2.5-1.5B	1829	255	1188.8	$< 10^{-12}$	+0.810	yes
	Gemma-2-2B	405	449	2.3	$1.32\text{e}{-1}$	-0.027	no
	Qwen2.5-3B	2862	36	2755.8	$< 10^{-12}$	+3.359	yes
	Mistral-7B	1088	524	197.3	$< 10^{-12}$	+0.265	yes
	Qwen2.5-7B	2326	277	1612.9	$< 10^{-12}$	+1.078	yes
	Llama-3.1-8B	468	537	4.7	$2.95\text{e}{-2}$	-0.040	yes
moral	Pythia-410M	81	901	684.7	$< 10^{-12}$	-0.544	yes
	Llama-3.2-1B	2628	100	2342.7	$< 10^{-12}$	+1.888	yes
	Llama-3.2-1B-It	1505	763	242.8	$< 10^{-12}$	+0.297	yes
	Qwen2.5-1.5B	1109	375	363.0	$< 10^{-12}$	+0.371	yes
	Gemma-2-2B	447	772	86.7	$< 10^{-12}$	-0.172	yes
	Qwen2.5-3B	1228	698	145.8	$< 10^{-12}$	+0.226	yes
	Mistral-7B	1460	1508	0.8	$3.78\text{e}{-1}$	-0.016	no
	Qwen2.5-7B	836	1825	367.6	$< 10^{-12}$	-0.374	yes
	Llama-3.1-8B	1254	0	1254.0	$< 10^{-12}$	+0.847	yes
social	Pythia-410M	100	52	15.2	$9.90\text{e}{-5}$	+0.071	yes
	Llama-3.2-1B	1680	581	534.2	$< 10^{-12}$	+0.465	yes
	Llama-3.2-1B-It	750	738	0.1	$7.56\text{e}{-1}$	+0.006	no
	Qwen2.5-1.5B	371	1516	694.8	$< 10^{-12}$	-0.549	yes
	Gemma-2-2B	11	260	228.8	$< 10^{-12}$	-0.287	yes
	Qwen2.5-3B	2045	1	2042.0	$< 10^{-12}$	+1.460	yes
	Mistral-7B	345	856	217.4	$< 10^{-12}$	-0.279	yes
	Qwen2.5-7B	1306	331	580.7	$< 10^{-12}$	+0.490	yes
	Llama-3.1-8B	0	768	768.0	$< 10^{-12}$	-0.586	yes

Table 7: **Three localisations of the same effect.** Normalised depth, mean ± s.d. over the four vocabularies. Causal depth is the earliest layer reaching half of final patching recovery; representational depth is the probe-A transition; geometric depth is the velocity-gap minimum of Section I. The last column is the causal-minus-representational separation.

Model	layers	ℓ^* range	causal	representational	geometric	separation
Pythia-410M	24	16–18	0.74 ±0.03	0.27 ±0.09	0.00 ±0.00	+0.47
Llama-3.2-1B	16	10–11	0.68 ±0.03	0.45 ±0.14	0.00 ±0.00	+0.23
Llama-3.2-1B-It	16	11–13	0.83 ±0.06	0.47 ±0.07	0.00 ±0.00	+0.36
Qwen2.5-1.5B	28	22–22	0.81 ±0.00	0.12 ±0.03	0.00 ±0.00	+0.69
Gemma-2-2B	26	22–22	0.88 ±0.00	0.11 ±0.05	0.00 ±0.00	+0.77
Qwen2.5-3B	36	28–32	0.87 ±0.04	0.11 ±0.00	0.14 ±0.23	+0.76
Mistral-7B	32	19–21	0.65 ±0.02	0.36 ±0.05	0.03 ±0.00	+0.29
Qwen2.5-7B	28	22–23	0.83 ±0.02	0.21 ±0.00	0.00 ±0.00	+0.62
Llama-3.1-8B	32	16–16	0.52 ±0.00	0.18 ±0.03	0.00 ±0.00	+0.34

F Additional Behavioral Analyses

Three further behavioural analyses are mentioned only briefly in the main text (Section 5); we report them in full here.

F.1 Numerical difficulty

A factorial logistic regression of correctness on rendering × difficulty tertile (full model in Section M) finds the rendering term significant for 31 of 36 cells, and a significant interaction — the rendering effect growing on harder items, as a strategic-robustness story predicts — in only 4. We do not build on that interaction; the difficulty main effect is uniformly small ($|\beta| \leq 0.21$).

F.2 Frontier-model spot check

We queried two frontier chat APIs on 1000 matched pairs of the neutral vocabulary, parsing the chosen label from the response text. One model returned unparseable output on 200 of 200 calls under our fixed template and is excluded entirely; we report the parse failure rather than silently dropping it, since a 100% failure rate is information about template brittleness. GPT-4o parsed cleanly on all calls and scored 55.0% canonical against 78.0% counterfactual (McNemar $p < 10^{-4}$, uncorrected), the same sign and rough magnitude as the open-weight models on neutral. This is behavioral evidence only, from a chat interface rather than raw logits; it cannot distinguish the mechanisms separated in Section 6, and one vocabulary on one model is not a basis for claims about frontier models in general.

Frontier Model Chat Prompt

[System Instruction]
You are an autonomous agent in a strategic decision task. Respond with ONLY the bracketed label of your choice, nothing else - e.g. "[Option Name]".

[User Message]
Option [{label_A}]: Yields {R} pts if opponent selects [{label_A}], and {S} pts if opponent selects [{label_B}].
Option [{label_B}]: Yields {T} pts if opponent selects [{label_A}], and {P} pts if opponent selects [{label_B}].

Objective:Choose the option with the higher expected score, assuming the opponent selects each action with equal probability. Which option do you choose?

G Causal Localisation

Table 8: The causally influential depth ℓ^* (earliest layer reaching half of final recovery) and the recovery fraction at that layer, per vocabulary.

Model	layers	neutral	valence	moral	social
Pythia-410M	24	18 (0.63)	17 (0.53)	16 (0.53)	17 (0.53)
Llama-3.2-1B	16	11 (0.65)	10 (0.42)	10 (0.92)	10 (0.84)
Llama-3.2-1B-It	16	11 (0.48)	13 (0.56)	13 (0.59)	13 (0.76)
Qwen2.5-1.5B	28	22 (0.82)	22 (0.78)	22 (0.95)	22 (0.74)
Gemma-2-2B	26	22 (0.68)	22 (0.71)	22 (0.71)	22 (0.71)
Qwen2.5-3B	36	32 (1.13)	28 (0.40)	31 (0.40)	31 (0.41)
Mistral-7B	32	20 (0.45)	21 (0.55)	19 (0.42)	20 (0.52)
Qwen2.5-7B	28	23 (0.96)	22 (0.22)	22 (0.44)	23 (0.89)
Llama-3.1-8B	32	16 (0.62)	16 (0.61)	16 (0.77)	16 (0.61)

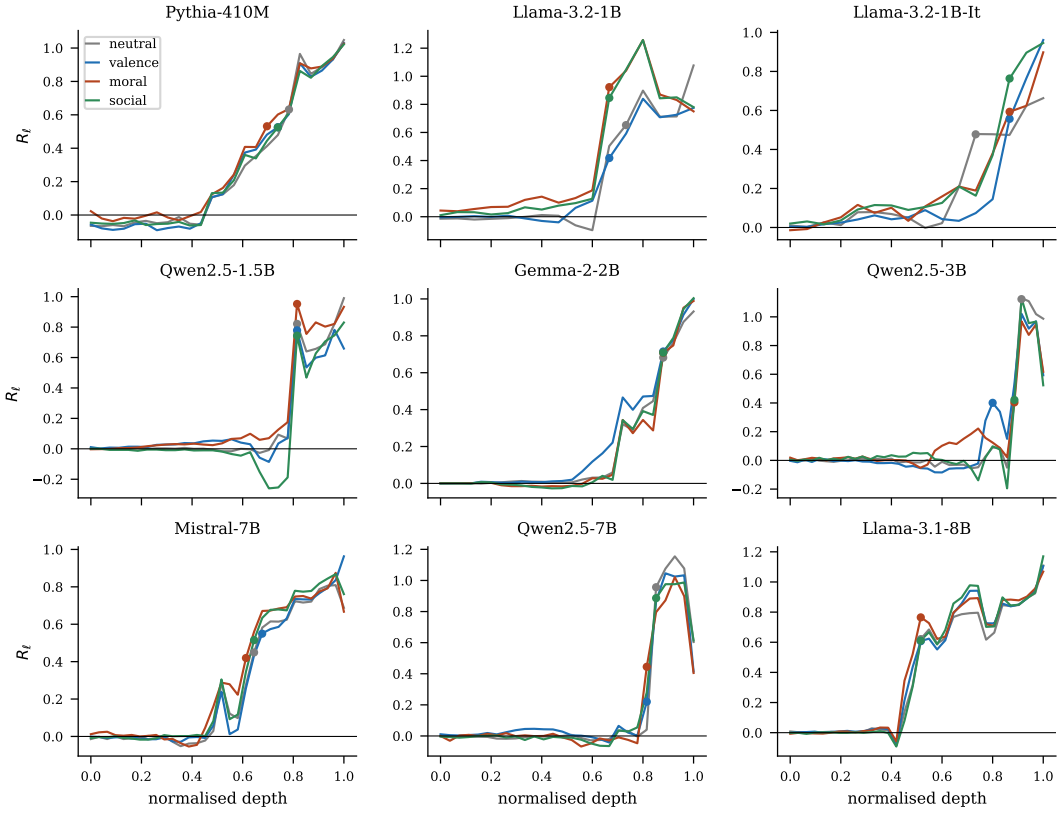

Figure 5: Patching recovery R_ℓ against normalised depth for every model and vocabulary. Dots mark ℓ^*, the causally influential depth. The flat-then-steep shape is consistent across the suite; the depth at which the rise begins is a property of the model more than of the vocabulary.

Table 9: Position-targeted patching at the causally influential depth ℓ^*. Recovery when patching only the final prompt position, only the label-token positions, or both. Negative values indicate that patching label positions moves the logit difference away from the clean run.

Vocabulary	Model	final only	label positions only	labels + final
neutral	Pythia-410M	0.592	+0.105	0.877
	Llama-3.2-1B	0.755	+0.061	0.828
	Llama-3.2-1B-It	0.445	-0.004	0.454
	Qwen2.5-1.5B	0.791	+0.282	1.026
	Gemma-2-2B	0.684	-0.017	0.676
	Qwen2.5-3B	1.048	-0.367	1.086
	Mistral-7B	0.322	+0.042	0.304
	Qwen2.5-7B	0.939	-0.018	0.890
	Llama-3.1-8B	0.656	+0.026	0.654
valence	Pythia-410M	0.474	+0.265	0.829
	Llama-3.2-1B	0.545	-0.087	0.511
	Llama-3.2-1B-It	0.571	-0.004	0.592
	Qwen2.5-1.5B	0.836	+0.073	1.037
	Gemma-2-2B	0.710	-0.012	0.704
	Qwen2.5-3B	0.388	+0.688	0.929
	Mistral-7B	0.467	-0.013	0.431
	Qwen2.5-7B	0.151	+0.762	0.983
	Llama-3.1-8B	0.632	+0.092	0.707
moral	Pythia-410M	0.475	+0.395	0.884
	Llama-3.2-1B	0.979	-0.302	0.853
	Llama-3.2-1B-It	0.602	-0.024	0.597
	Qwen2.5-1.5B	1.001	+0.069	1.052
	Gemma-2-2B	0.703	-0.020	0.694
	Qwen2.5-3B	0.251	+0.220	0.881
	Mistral-7B	0.455	-0.048	0.341
	Qwen2.5-7B	0.622	+0.619	1.048
	Llama-3.1-8B	0.723	+0.079	0.774
social	Pythia-410M	0.497	+0.355	0.889
	Llama-3.2-1B	0.970	-0.168	0.795
	Llama-3.2-1B-It	0.768	-0.066	0.745
	Qwen2.5-1.5B	0.820	+0.178	1.085
	Gemma-2-2B	0.717	-0.009	0.714
	Qwen2.5-3B	0.373	+0.320	0.911
	Mistral-7B	0.414	+0.040	0.405
	Qwen2.5-7B	0.907	+0.120	0.903
	Llama-3.1-8B	0.587	+0.056	0.613

H PROBING

Probes are ℓ_2-regularised logistic classifiers ($C = 1$) on frozen residual activations at the final prompt position, fitted on canonical renderings and evaluated on both. The reported *probe depth* is the shallowest layer at which the canonical/counterfactual generalisation gap reaches its maximum of 1.0 (canonical 100%, counterfactual 0%). A sweep over $C \in \{0.01, 0.1, 1, 10, 100\}$ leaves this split at exactly 100%/0% throughout, so the transition is not an artefact of the regularisation default.

Table 10: Probe-A transition depth and the fraction of the same-word generalisation gap retained on words never seen during probe fitting (%).

	neutral		valence		moral		social	
Model	depth	unseen	depth	unseen	depth	unseen	depth	unseen
Pythia-410M	0.33	54	0.38	87	0.17	62	0.21	89
Llama-3.2-1B	0.38	63	0.69	75	0.38	73	0.38	75
Llama-3.2-1B-It	0.50	97	0.56	75	0.44	68	0.38	72
Qwen2.5-1.5B	0.11	78	0.18	89	0.11	100	0.11	97
Gemma-2-2B	0.08	56	0.19	81	0.08	100	0.08	79
Qwen2.5-3B	0.11	86	0.11	93	0.11	97	0.11	79
Mistral-7B	0.38	46	0.44	87	0.31	99	0.31	96
Qwen2.5-7B	0.21	60	0.21	93	0.21	83	0.21	79
Llama-3.1-8B	0.19	76	0.22	97	0.16	100	0.16	83

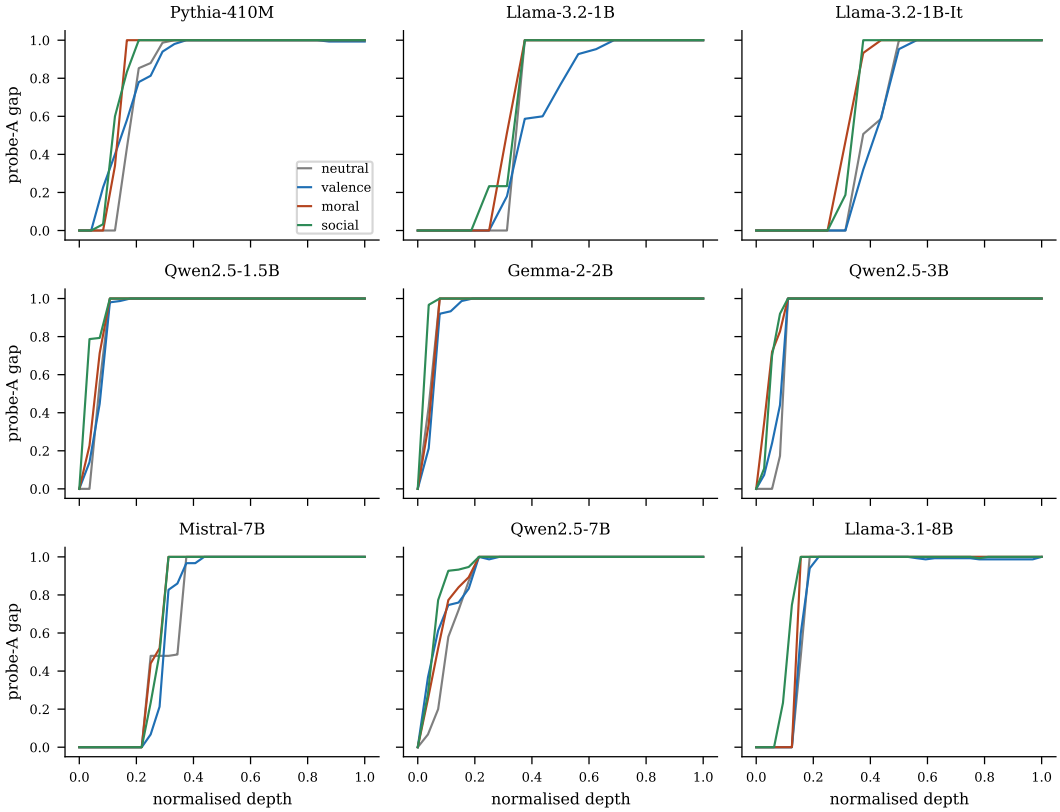

Figure 6: Probe-A generalisation gap against normalised depth, all models and vocabularies. The step to 1.0 is sharp in every case, and occurs in the first third of the network for most models.

I REPRESENTATION-SPACE GEOMETRY

We also tried a third way of localising the effect, adapting the trajectory analysis of Zhou et al. (2026): treat the residual stream as a trajectory across layers, compute per-layer velocity vectors,

and compare the cosine similarity of matched canonical/counterfactual trajectories against instance-shuffled pairs. The depth where matched and shuffled similarities are closest should mark where the two renderings become distinguishable.

It does not: the minimum sits at depth 0.00 for 34 of 36 cells (exceptions: Mistral-7B at 0.03, one Qwen2.5-3B cell at 0.54), with gaps of order 10^{-3} — Table 7 in Section E gives the full comparison. Read literally, this localises the effect at layer 0, which patching (Section G) contradicts outright.

We think the criterion measures the wrong thing rather than revealing a real disagreement about the mechanism. At the earliest layers velocity is dominated by generic token processing shared by all instances, so matched and shuffled pairs look equally similar there and the minimum falls at layer 0 almost by construction. The original method indexes over chain-of-thought steps, each a genuine inferential move; our task is a single forced choice, and re-indexing over depth does not preserve what made that axis meaningful. We report this as a negative result about method transfer, not a claim that trajectory geometry is uninformative in general, and we do not use it as evidence for the causal picture in Section 6.

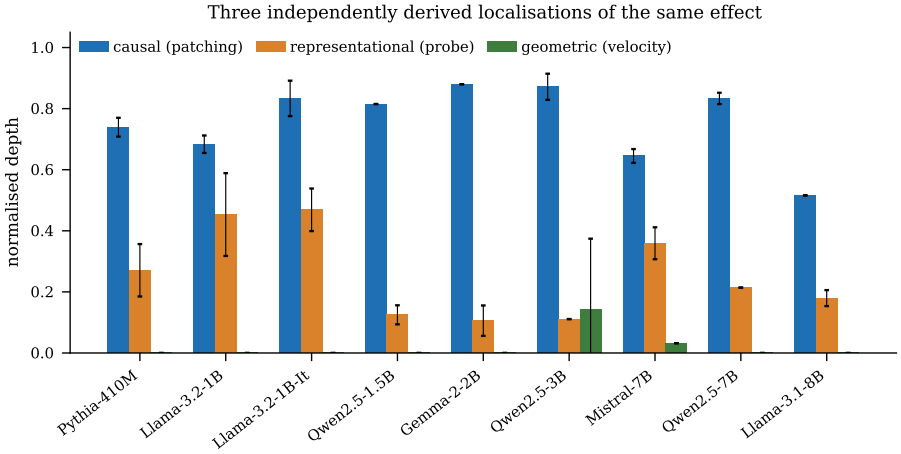

Figure 7: The three localisations side by side, mean $\pm$ s.d. over vocabularies. The geometric criterion returns depth ≈ 0 for every model, which is why we treat it as a failed method transfer rather than as a third estimate.

J Attention-Head Dissection

We ablate each $(\text{layer}, \text{head})$ by mean-patching its output and measure the change in the recovery metric, testing each cell and correcting with Benjamini–Hochberg across all cells within a model.

Table 11: Head ablation on the neutral vocabulary. The proportion of significant cells is high and the largest individual effect is small, which is the signature of a distributed rather than a sparse circuit. Top heads are listed as LℓHh with their effect in percentage points of recovery.

Model	BH-significant cells	Three largest effects
Pythia-410M	259/384 (67%)	L8H6 (2.21), L12H11 (1.25), L19H4 (1.12)
Llama-3.2-1B	389/512 (76%)	L15H8 (0.64), L8H31 (0.57), L5H17 (0.44)
Llama-3.2-1B-It	408/512 (80%)	L13H10 (0.85), L8H31 (0.81), L12H14 (0.78)
Qwen2.5-1.5B	274/336 (82%)	L19H5 (2.43), L21H11 (1.85), L10H10 (1.13)
Gemma-2-2B	179/208 (86%)	L21H1 (0.68), L19H1 (0.51), L3H2 (0.47)
Qwen2.5-3B	488/576 (85%)	L22H3 (2.85), L22H0 (2.32), L20H4 (1.86)
Mistral-7B	819/1024 (80%)	L12H9 (3.87), L11H20 (2.95), L12H21 (2.51)
Qwen2.5-7B	615/784 (78%)	L23H8 (2.93), L19H14 (2.60), L19H18 (2.58)
Llama-3.1-8B	842/1024 (82%)	L20H3 (1.49), L20H0 (1.21), L28H29 (1.21)

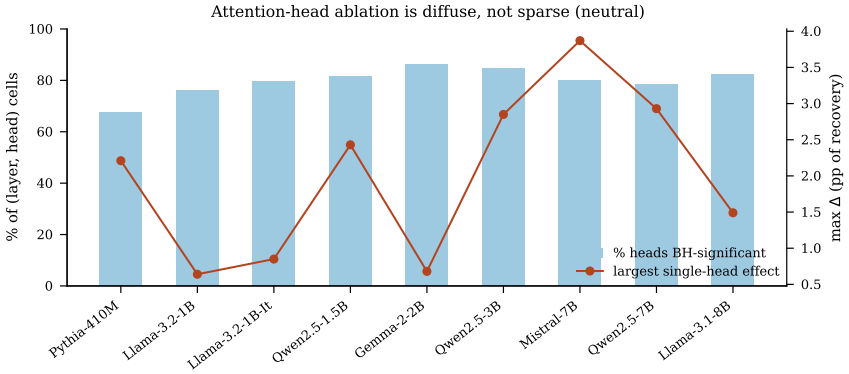

Figure 8: Proportion of significant head cells (bars, left axis) against the largest single head effect (line, right axis). Larger models have both more significant cells and larger maxima, but no model has a head contributing more than $4\,pp$ of recovery.

Table 12: **Steering summary**, averaged over vocabularies. Unsteered, steered on held-out seen-label prompts, steered on unseen-label prompts, and the model's own canonical accuracy (the ceiling a genuine repair could not exceed). The last column is the amount by which steering overshoots that ceiling. * marks models with one vocabulary skipped for lack of fittable errors (averaged over the remaining three). Per-vocabulary results, α values and conditional-gate numbers in Section K.

Model	unsteered	steered, seen	steered, unseen	ceiling	overshoot
Pythia-410M	47.3	99.4	50.3	27.3	+72.1
Llama-3.2-1B*	11.7	49.0	32.4	83.1	-34.1
Llama-3.2-1B-It*	33.3	57.7	16.5	55.0	+2.6
Qwen2.5-1.5B	53.8	94.0	38.0	21.6	+72.3
Gemma-2-2B	32.6	100.0	27.7	34.0	+66.0
Qwen2.5-3B	37.6	81.3	38.7	41.5	+39.8
Mistral-7B*	79.0	98.8	82.6	54.1	+44.7
Qwen2.5-7B	51.8	88.6	45.7	29.5	+59.1
Llama-3.1-8B	46.2	97.3	54.6	32.3	+65.0

K STEERING

Three cells are skipped for lack of fittable errors: Llama-3.2-1B and Llama-3.2-1B-It on neutral (zero errors on the fit split, because those models are near-ceiling on counterfactual neutral prompts), and Mistral-7B on moral (three errors). We report these as skipped rather than imputing zeros, since a model with no errors to fit on is not a model where steering failed.

Two cells are worth flagging for the opposite reason. Qwen2.5-1.5B on social has a canonical ceiling of 0.0% and an unsteered counterfactual accuracy of 98.9%; the steering result there is close to meaningless because the model is already inverted. Qwen2.5-7B on moral has a canonical ceiling of 3.2%. Averages that include these cells are not informative about repair, which is part of why Table 12 reports the ceiling explicitly.

Table 13: Full steering results. "errors" is the number of fit-split counterfactual instances the model gets wrong, on which v_{drift} is fitted. Accuracies are on held-out prompts. "cond." gives the conditional-gate result on seen/unseen labels at the same α. ‡ marks cells where the best static gain over unsteered is below $5\,pp$, the signature Valentino et al. (2026) associate with needing a conditional rather than static intervention.

Vocab.	Model	errors	α^*	unsteered	seen	unseen	ceiling	cond. seen/unseen
neutral	Pythia-410M	95/213	3	53.5	100.0	37.3	14.6	91.5/40.0
	Llama-3.2-1B			skipped: <5 mispredicted fit instances				
	Llama-3.2-1B-It			skipped: <5 mispredicted fit instances				
	Qwen2.5-1.5B	147/213	1.5	29.1	100.0	16.9	26.8	99.5/16.4
	Gemma-2-2B	147/213	1.25	29.6	100.0	33.8	29.6	100.0/33.8
	Qwen2.5-3B	148/213	1.25	27.7	100.0	59.8	29.6	98.6/62.7
	Mistral-7B	39/213	5	82.6	100.0	90.9	69.0	97.7/95.1
	Qwen2.5-7B	144/213	1	31.5	100.0	68.0	29.1	100.0/58.0
	Llama-3.1-8B	90/213	1.25	54.9	100.0	49.8	24.4	96.7/52.0
valence	Pythia-410M	239/386	8	39.8	97.7	74.4	32.6	94.3/56.0
	Llama-3.2-1B	254/386	8	35.1	52.2	32.4	53.7	52.5/32.4
	Llama-3.2-1B-It	202/386	8	45.2	58.4	41.6	54.3	62.3/38.7
	Qwen2.5-1.5B	288/386	8	23.5	76.0	91.3	43.4	73.4/62.2
	Gemma-2-2B	263/386	8	31.5	100.0	18.0	36.7	100.0/19.6
	Qwen2.5-3B	308/386	8	19.1	65.6	72.0	66.9	80.6/71.6
	Mistral-7B	90/386	8	74.7	99.2	69.8	42.6	95.1/67.6
	Qwen2.5-7B	300/386	8	28.4	54.5	41.3	58.1	55.8/40.0
	Llama-3.1-8B	211/386	8	46.8	89.1	94.0	35.7	88.1/65.3
moral	Pythia-410M	163/379	5	58.0	100.0	39.8	25.3	88.1/42.4
	Llama-3.2-1B	379/379	8	0.0	66.0	–	98.4	–/–
	Llama-3.2-1B-It‡	269/379	0.5	26.6	27.7	0.0	59.9	27.7/0.0
	Qwen2.5-1.5B	125/379	1.25	63.6	100.0	17.6	16.4	99.7/0.0
	Gemma-2-2B	262/379	1.5	31.9	100.0	34.0	31.9	100.0/34.0
	Qwen2.5-3B	118/379	0.75	66.5	100.0	0.9	13.7	99.7/0.9
	Mistral-7B			skipped: <5 mispredicted fit instances				
	Qwen2.5-7B‡	29/379	0.25	95.8	100.0	0.7	3.2	96.8/1.8
	Llama-3.1-8B	241/379	2	35.1	100.0	22.4	34.6	98.2/9.1
social	Pythia-410M	225/374	5	38.1	100.0	49.8	36.8	99.7/49.8
	Llama-3.2-1B	374/374	8	0.0	28.8	–	97.1	–/–
	Llama-3.2-1B-It	269/374	8	28.0	86.9	8.0	50.9	86.1/8.0
	Qwen2.5-1.5B‡	8/374	0.25	98.9	100.0	26.4	0.0	98.9/26.4
	Gemma-2-2B	233/374	2	37.6	100.0	25.1	37.6	100.0/25.1
	Qwen2.5-3B	238/374	5	37.3	59.7	22.2	55.7	58.1/22.2
	Mistral-7B	80/374	8	79.7	97.3	87.1	50.7	95.5/84.9
	Qwen2.5-7B	176/374	1.25	51.5	100.0	72.9	27.7	99.5/31.6
	Llama-3.1-8B	193/374	3	48.0	100.0	52.4	34.4	99.2/45.6

L CROSS-TOPOLOGY TRANSFER

Table 14: A steering direction fitted on Prisoner's Dilemma instances only, applied unchanged to Chicken and Stag Hunt at the same α^*. Models absent from a block had no fittable direction for that vocabulary.

		Chicken		Stag Hunt	
Vocabulary	Model	unsteered	transferred	unsteered	transferred
neutral	Pythia-410M	62.3	85.6	61.3	83.8
	Qwen2.5-1.5B	54.9	62.1	54.4	62.6
	Gemma-2-2B	52.4	65.4	51.4	65.9
	Qwen2.5-3B	59.0	74.6	60.1	78.3
	Qwen2.5-7B	62.2	73.1	61.9	72.3
	Llama-3.1-8B	62.4	66.9	59.4	67.9
valence	Pythia-410M	48.1	92.3	46.6	93.1
	Llama-3.2-1B	32.0	58.1	32.9	58.8
	Llama-3.2-1B-It	37.9	45.0	40.2	42.3
	Qwen2.5-1.5B	30.6	85.6	28.7	87.3
	Gemma-2-2B	42.1	99.4	43.0	99.4
	Qwen2.5-3B	20.9	52.9	17.7	52.6
	Mistral-7B	68.3	19.7	68.0	11.3
	Qwen2.5-7B	26.9	51.9	29.8	51.9
	Llama-3.1-8B	48.6	95.4	49.0	95.0
moral	Pythia-410M	65.3	98.1	64.6	99.0
	Llama-3.2-1B	5.1	59.0	7.7	58.0
	Llama-3.2-1B-It	32.7	29.3	32.3	27.0
	Qwen2.5-1.5B	38.9	63.9	39.2	62.2
	Gemma-2-2B	45.4	66.9	42.1	65.8
	Qwen2.5-3B	37.2	56.0	37.2	55.7
	Qwen2.5-7B	47.1	47.7	48.9	49.2
	Llama-3.1-8B	37.3	60.7	42.0	57.3
social	Pythia-410M	53.0	98.2	48.1	98.3
	Llama-3.2-1B	23.6	82.2	23.7	83.8
	Llama-3.2-1B-It	50.7	59.7	44.1	58.9
	Qwen2.5-1.5B	75.4	75.2	72.8	72.6
	Gemma-2-2B	51.6	69.3	47.0	59.8
	Qwen2.5-3B	38.2	84.1	33.3	82.4
	Mistral-7B	56.0	22.2	71.4	18.3
	Qwen2.5-7B	56.9	80.2	51.2	76.2
	Llama-3.1-8B	55.0	74.0	48.6	66.7

Transfer helps (by more than $3\,pp$) in 53 of 64 cases. The failures are concentrated in Mistral-7B, where transfer is actively harmful on valence and social, and in Llama-3.2-1B-It on moral. Since PD has a constant correct slot while Chicken and Stag Hunt do not, transfer gains cannot be explained by the direction encoding a fixed slot preference.

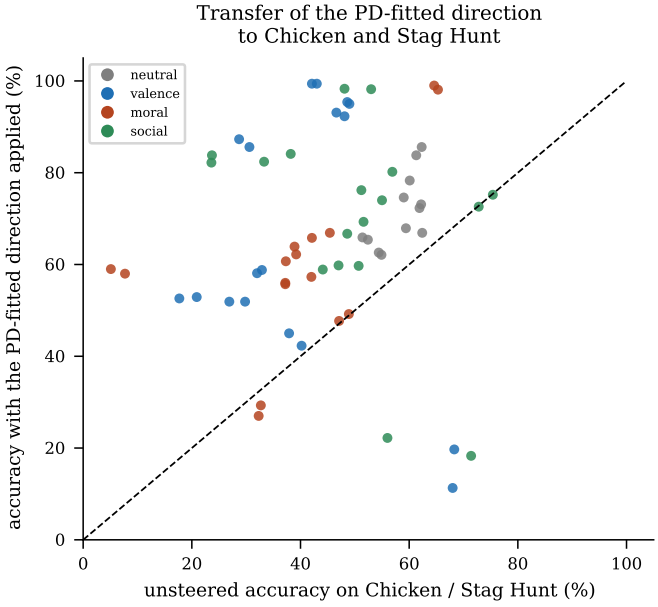

Figure 9: Every model×vocabulary cell from the table above, plotted as unsteered accuracy against accuracy with the Prisoner's-Dilemma-fitted direction applied. Most points sit above the diagonal; the Mistral-7B failures sit clearly below it.

M FACTORIAL REGRESSION

We fit correct $\sim$ rendering $\times$ difficulty as a logistic regression per model$\times$vocabulary cell, with rendering coded 1 for counterfactual and difficulty coded $0/1/2$ for easy/medium/hard. A positive rendering coefficient means the counterfactual rendering is *more* accurate.

Table 15: Factorial logistic regression coefficients. Significance markers on the rendering term: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Bold asterisks mark significant interactions.

Vocabulary	Model	$\beta_{\text{rendering}}$	$\beta_{\text{difficulty}}$	$\beta_{\text{interaction}}$	p_{int}
neutral	Pythia-410M	+0.962***	+0.119	-0.074	0.2821
	Llama-3.2-1B	+2.901***	+0.191	-0.267 *****	0.0009
	Llama-3.2-1B-It	+3.222***	+0.115	-0.044	0.6107
	Qwen2.5-1.5B	+0.316***	+0.067	+0.005	0.9380
	Gemma-2-2B	+0.020	+0.071	+0.005	0.9363
	Qwen2.5-3B	+0.857***	+0.019	+0.097	0.1658
	Mistral-7B	+2.703***	+0.207	-0.063	0.4795
	Qwen2.5-7B	+1.045***	+0.098	-0.048	0.4845
	Llama-3.1-8B	+1.047***	+0.173	-0.130	0.0551
valence	Pythia-410M	-0.043	-0.033	+0.049	0.4637
	Llama-3.2-1B	-1.444***	-0.093	+0.240 *****	0.0004
	Llama-3.2-1B-It	-0.882***	-0.073	+0.058	0.3763
	Qwen2.5-1.5B	-1.292***	-0.044	+0.076	0.2774
	Gemma-2-2B	-0.597***	-0.018	-0.009	0.8899
	Qwen2.5-3B	-2.672***	-0.016	-0.032	0.6915
	Mistral-7B	+0.937***	+0.148	+0.007	0.9235
	Qwen2.5-7B	-1.868***	-0.008	+0.035	0.6329
	Llama-3.1-8B	-0.276***	-0.028	+0.039	0.5493
moral	Pythia-410M	+1.188***	-0.031	+0.062	0.3752
	Llama-3.2-1B	-4.956***	+0.040	-0.080	0.5361
	Llama-3.2-1B-It	-1.281***	+0.031	-0.062	0.3645
	Qwen2.5-1.5B	-0.878***	-0.045	+0.106	0.1171
	Gemma-2-2B	-0.607***	+0.009	-0.008	0.9097
	Qwen2.5-3B	-1.283***	-0.096	+0.244 *****	0.0003
	Mistral-7B	-0.293***	+0.066	+0.027	0.6802
	Qwen2.5-7B	-0.172*	+0.004	+0.089	0.1670
	Llama-3.1-8B	-1.570***	-0.056	+0.199 *****	0.0039
social	Pythia-410M	+0.015	+0.002	+0.010	0.8862
	Llama-3.2-1B	-1.975***	+0.047	-0.081	0.2687
	Llama-3.2-1B-It	-0.066	-0.027	+0.065	0.3125
	Qwen2.5-1.5B	+2.217***	-0.071	+0.093	0.2326
	Gemma-2-2B	-0.121	-0.041	+0.030	0.6500
	Qwen2.5-3B	-0.834***	+0.056	-0.098	0.1421
	Mistral-7B	+0.565***	+0.120	+0.024	0.7191
	Qwen2.5-7B	+0.464***	+0.008	+0.073	0.2673
	Llama-3.1-8B	+0.478***	+0.000	+0.002	0.9774

The rendering term is significant in 31 of 36 cells and its sign matches the behavioural gap in every case. The interaction is significant in four cells (Llama-3.2-1B on neutral and valence, Qwen2.5-3B and Llama-3.1-8B on moral) and in each the rendering penalty grows with difficulty. We regard four of thirty-six as too weak to support a difficulty-modulation claim.

N BEHAVIORAL CONTROLS

Frequency- and length-matched labels: On Pythia-410M with the valence vocabulary, restricting to 11 frequency- and length-matched word pairs changes the gap from $0.0\,pp$ (canonical 36.7%, counterfactual 36.7%) to $-2.0\,pp$ (canonical 35.6%, counterfactual 37.6%). The sign is preserved, so the effect is not purely a frequency confound, but this is one model, one vocabulary, and eleven pairs.

Table 16: Controls on two models. *Order swap* exchanges the presentation order of the two options without touching labels or payoffs. *Payoff-only swap* exchanges the payoff rows while holding labels fixed. Both are payoff-irrelevant to Definition 1 in the same sense the label swap is.

Model	Vocabulary	canonical	order swap	payoff-only swap	counterfactual
Pythia-410M	neutral	27.8	54.6	71.0	47.0
	valence	37.8	61.2	61.8	39.2
	moral	25.4	42.4	73.4	54.4
	social	30.6	69.6	67.2	31.4
Llama-3.1-8B	neutral	29.2	53.6	74.8	52.0
	valence	44.2	61.8	63.6	39.8
	moral	60.8	73.6	57.8	28.0
	social	27.6	67.6	70.4	38.6

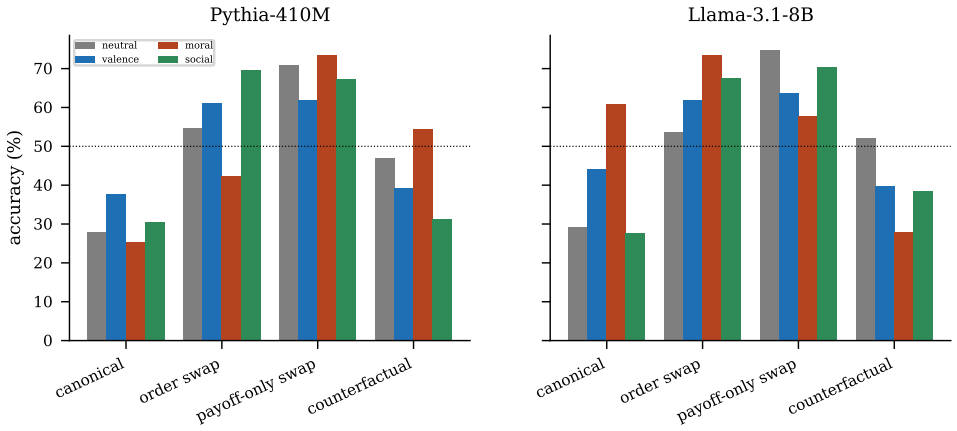

Figure 10: The same data. Order-swap and payoff-only-swap conditions move accuracy as much as the label swap in several cells, which is why we describe label identity as one of several payoff-irrelevant sensitivities rather than a uniquely privileged one.

Cross-template check. The same instances rendered in the narrative template give canonical 42.0% and counterfactual 41.4% ($+0.6\,pp$) against 36.7% / 36.7% ($-0.0\,pp$) for the original template. The sign changes. We treat this as an open external-validity question rather than a resolved control.

O REPRODUCIBILITY

The dataset generator is deterministic given a seed: payoffs are rejection-sampled from a seeded RNG, difficulty tertiles are computed empirically from the realised margin distribution, and word assignment is seeded per instance. All experiments in this paper use a single seed; we have not run seed replicates, which is a gap.

Compute: the full sweep runs on a mixture of A10G, L40S and A100-80GB devices. The expensive phases are activation patching (one forward pass per layer per instance) and the steering α sweep (eleven values $\times$ two evaluation sets). Phases operating on a single vocabulary (patching, probing, steering, geometry) were run four times, once per vocabulary, which is the source of the per-vocabulary results throughout; the behavioral phase covers all four vocabularies in one pass and is identical across those four runs.