LOCALIZING AND REUSING PERSONA-INDUCED REPRESENTATIONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used in settings that require strategic judgment. Persona prompting is a common way to elicit different behaviors from language models by asking them to adopt specific roles or personality traits. However persona descriptions do not map straightforwardly onto behavioral responses. A persona described as cooperative may still defect in a particular setting, while combining multiple personas leaves their joint behavioral influence unclear. To better understand this phenomenon and enable more targeted behavioral control, we use contrastive persona prompts to induce behaviorally verified cooperation and defection in a canonical Prisoner's Dilemma (PD), providing a sparse feature ensemble that reliably distinguishes persona-induced cooperative and defective decisions and serves as a target for causal intervention. Our results make three contributions. First, we develop a workflow for localizing persona-related SAE features from behaviorally verified persona contrasts. Second, by comparing across decision-making games, we find that a shared core of features can predict behavioral outcomes, while task-specific features are required for effective intervention. Third, we use these representations to interpret persona-induced behavior and demonstrate a proof-of-concept route from a target feature pattern back to a readable persona prompt, providing an initial path toward interpretable and controllable prosocial behavior in language models.

1 INTRODUCTION

Large language models increasingly make strategic decisions: they cooperate or defect, allocate resources, punish, bargain, and contribute under explicit incentives (Horton et al., 2023; Chen et al., 2023). Persona prompting can shift these choices by asking a model to adopt a role or decision style (Kong et al., 2024; Wang et al., 2024b; Lutz et al., 2025). However, persona descriptions do not determine behavior in a transparent way. Consider one persona in our pool: *"You are a fair-minded steward who values balanced outcomes and tries to leave every participant feeling respected. You tend to favor choices that preserve trust and shared dignity."* The description reads cooperative, yet the model defects in the Prisoner's Dilemma ($P(C) = 0.01$); conversely, different surface descriptions can produce the same decision. From the output alone, it is therefore unclear whether a persona changes only the answer's wording or installs a reusable decision state inside the model.

We study this question with behavior-verified persona contrasts in the Prisoner's Dilemma (PD). The game gives a controlled discovery setting: the payoff matrix, answer format, and decision variable are fixed, while persona prompts induce reliably cooperative or defecting choices. We use these contrasts to ask where the induced difference becomes writable, which sparse features carry it, whether those features can be read from decisions, and whether the representation can guide a return from internal state to readable language. The Dictator game and related allocation tasks then test which parts of the representation transfer across decision variables.

Our analysis follows this direction from behavior to representation and back. Layerwise state replacement identifies L15 as the first layer at which a cooperative persona state changes the decision in a fixed no-persona prompt. A sparse autoencoder at that layer yields cooperation- and defection-associated features whose associations can be tested across action-label mappings. Held-out readouts test a train-selected AB/CD set, while feature intervention tests a separately extracted task-local

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.

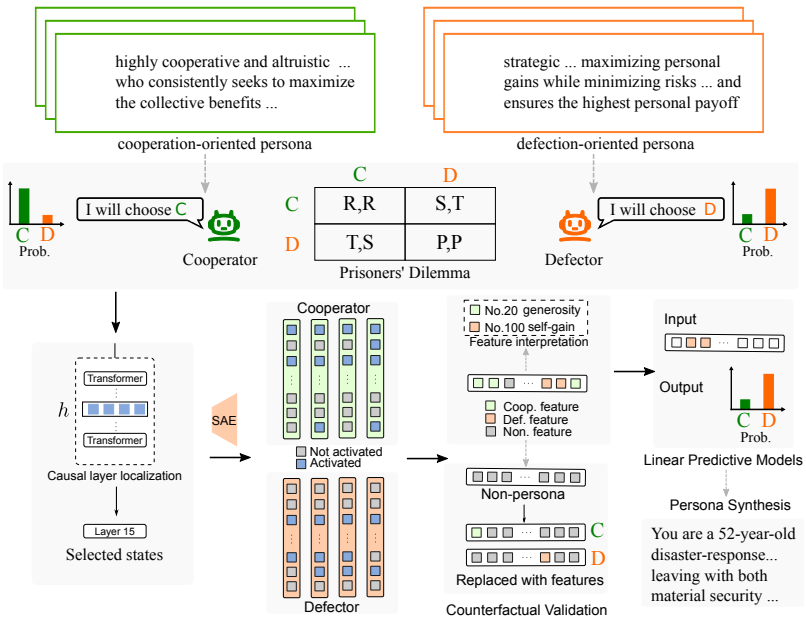

Figure 1: **From persona behavior to a feature-level interface.** Behavior-verified contrasts localize a writable state and identify task-conditioned SAE features. Separate evaluations test *readout* on held-out personas, *intervention* with task-local site sets, and *verification* of synthesized persona prompts against their decisions.

set. Across games, intervention transfers to selected related tasks but has a narrower boundary than readout across allocation settings. Finally, a single controlled inversion uses a learned signature to guide construction of a readable persona prompt, demonstrating a representation-to-language path without claiming a general decoder.

Figure 1 summarizes the account. The paper makes three contributions:

- **Localization and readout of persona-induced behavior.** Behavior-verified persona contrasts localize a writable L15 state. Across paired AB/CD splits, 48–52 training-selected SAE sites predict held-out persona choices across both action-label realizations.
- **Causal intervention and transfer boundary.** Injecting target-side sites while masking opposite-side sites flips verified personas in both directions on PD and Dictator, whereas matched random controls are null. A fixed PD intervention transfers strongly to Chicken but not to most other tested C/D games; amount decisions require Dictator-local sites.
- **A proof-of-concept representation-to-language path.** The learned feature signature guides a soft-to-discrete inversion that preserves the target behavior in one controlled case, showing how an internal behavioral representation can be translated back into a readable persona specification.

2 Related Work

Persona and role prompting. Role and persona prompts reliably shift model behavior (Kong et al., 2024; Wang et al., 2024b; Lutz et al., 2025), and persona-level analyses show LLM responses can be steered and characterized along trait-like axes (Park et al., 2024; Bhandari et al., 2026; Zheng et al., 2024; Giorgi et al., 2024; Kirk et al., 2024). Prior work documents *that* personas work and how they persist or resist correction (Luz de Araujo et al., 2026; McKenzie et al., 2026); what a persona installs inside the model remains unstudied at the feature level.

Strategic behavior in games. LLMs are routinely evaluated in economic games, showing personas and framings shift cooperation, altruism, and strategic sophistication (Horton et al., 2023;

Chen et al., 2023; Akata et al., 2025; Capraro et al., 2025; Johnson & Obradovich, 2025; Lorè & Heydari, 2024; Wang et al., 2024a; Murthy et al., 2025). These studies measure induced behavior; we ask where that behavior is represented and whether it is causally writable.

Sparse autoencoders. SAEs decompose residual-stream activations into sparse, often interpretable coordinates (Cunningham et al., 2023; Gao et al., 2024; Lieberum et al., 2024) and support concept-level interventions (Wu et al., 2025). We use this coordinate system to evaluate both behavioral readout and task-local intervention, then use feature signatures to guide a return to language.

Activation steering and representation editing. Steering vectors and representation-engineering methods shift model outputs along extracted directions (Turner et al., 2024; Arditi et al., 2024; Zou et al., 2023; Li et al., 2023; Fontana et al., 2025; Meng et al., 2022; Burns et al., 2023). Our additional tests separate held-out decodability from task-local causal intervention and examine where the latter transfers.

3 METHOD

3.1 STRATEGIC BEHAVIOR INDUCTION

LLM. The main experiments use a frozen 32-layer Meta-Llama-3-8B-Instruct model (Meta AI, 2024); no weights are updated. Layer indices are zero-based, so L15 denotes layer 15. The sparse basis is the EleutherAI `sae-llama-3-8b-32x` SAE bank (EleutherAI, 2024): residual-stream TopK autoencoders trained on RedPajama, used fixed throughout (L15: width 131,072, k=192, normalized decoder). Experiments were run on NVIDIA RTX A6000 GPUs.

Games. The main experiments run on one-shot economic games. The Prisoner's Dilemma serves as the discovery task: the model is one of two anonymous players choosing C or D, with canonical payoffs R=3, S=0, T=5, P=1 (T>R>P>S) and the payoff table C/C=(3, 3), C/D=(0, 5), D/C=(5, 0), D/D=(1, 1). The Dictator game (continuous allocation of \$100, answer $\in [0, 100]$) is the transfer and causal testbed, where the decision variable is a scalar donation rather than a letter token. Ultimatum, public-goods, charity, trust, and life-scale allocation scenarios serve as further boundary checks.

Decision readout. The model answers in a fixed JSON field, and we score the *answer slot*: the next-token position immediately after the answer-field prefix. Let x denote the prompt ending at this prefix and $\ell(x) = f_\theta(x)_{t_{\text{choice}}}$ the logits there. For PD we use the two legal action logits and define

$$p_\alpha(x) = \frac{\exp \ell_\alpha(x)}{\exp \ell_C(x) + \exp \ell_D(x)}, \qquad \alpha \in \{C, D\}, \qquad \rho_C(x) = p_C(x). \tag{1}$$

For Dictator we read the donation distribution over candidate integer amounts and report its mean or argmax.

Behavior-verified personas. Cooperative and defecting persona pools are behavior-verified against each task (e.g., 100 verified rows per side in PD, selected from 512 candidates). The full construction pipeline—the behavioral-code candidate library, leakage filtering, and pool-assignment rules—is given in Section C.

3.2 LAYER LOCALIZATION

A standard clean–corrupt activation-patching setup requires paired prompts that differ only in the causal factor of interest (Meng et al., 2022). In our setting, however, a prompt edit that reliably changes game behavior necessarily also changes surface wording or explicit strategic content, making raw text-pair replacement difficult to interpret. We therefore localize the persona effect by behavioral transfer rather than by decodability alone: a layer is considered writable only if replacing its hidden state changes the model's actual decision.

The scan is directional. The no-persona baseline reliably chooses defection in this game, so we ask whether the decision state induced by a cooperative persona can be written into this fixed target

prompt. For each layer L, we take the last-token hidden state $h^{(L)}$ at the answer position from a cooperative-persona prompt, replace the corresponding state in the no-persona prompt, continue the forward pass through the remaining layers, and rescore the legal C/D logits. We scan all 32 layers and define a layer as writable when this single-state replacement changes the target decision from D-dominant to C-dominant. We use the earliest writable layer as the localization point for subsequent SAE analysis, since it is the first point at which the persona-induced decision state is sufficient to causally alter the downstream choice.

3.3 Feature Localization and Readout

After identifying the writable layer, we localize SAE features associated with cooperative decisions at the answer position.

For SAE analyses we encode the answer-slot hidden state $h_i^{(L)}$ of prompt row i at layer L with the fixed TopK autoencoder (Gao et al., 2024; Cunningham et al., 2023):

$$a_i^{(L)} = \mathrm{Enc}_L(h_i^{(L)}) \in \mathbb{R}^{d_{\mathrm{SAE}}}_{\geq 0}, \tag{2}$$

where $a_{ij}^{(L)}$ is the activation of SAE site j. The SAE is fixed throughout; it only provides a sparse coordinate system for comparing behavior-verified persona rows.

To separate behavioral features from features tied to a particular answer token, we evaluate four output-label mappings of the same Prisoner's Dilemma: CD (C= cooperate, D= defect), DC (D= cooperate, C= defect), AB (A= cooperate, B= defect), and BA (B= cooperate, A= defect). For each mapping m, we encode the answer-position hidden state with the SAE and compute the activation statistics of every feature j separately for behaviorally verified cooperative and defective personas.

A feature is considered cooperation-associated under mapping m if it shows one-sided support for cooperative behavior: it is activated in at least k=5 verified cooperative rows while remaining inactive across all verified defective rows. We apply this criterion independently to each of the four label mappings and retain only features whose cooperation association is preserved across mappings. Thus, a retained feature follows the behavioral role across label permutations rather than any fixed output token.

The four-mapping intersection characterizes the task-local core used for PD intervention. Readout is validated independently on a matched AB/CD cohort: within each class, 60 personas are paired across labelings, 20 are held out, and the remaining 40 are used for all site selection and fitting. A site must have one-sided support in both AB and CD training rows, with activation in at least six distinct own-side personas per labeling. We freeze the selected set before evaluating CD, AB, and pooled ridge readouts on the held-out personas. Five class-balanced splits are reported; paired AB/CD rows from one persona always stay on the same side. The holdout-selected sites are not the intervention set.

The defection-associated set is obtained by reversing the behavioral sides. We write the task-local intervention sets S_C and S_D, and reuse the one-sided rule for the Dictator task. Algorithm 1 summarizes the independent holdout readout protocol.

Ridge decodability check. As a completeness check, each split fits ridge regression $z(x) = a^{(L)}(x)[S] \mapsto \rho_C(x)$ using only its training personas and selected sites. We evaluate the frozen readout on the paired held-out personas. This measures whether a compact, role-consistent feature set carries the persona signal on unseen prompts, before any causal claim.

For generated text we use a zero-cost site-sum score rather than a learned predictor. The model is asked to write a short free-form paragraph justifying its choice and to end with an explicit decision sentence (e.g., *"I will donate $X."*), from which the expressed amount is parsed; the amount is normalized by the game's maximum allocation (e.g., $20 of $20 in Public Goods, $100 in Dictator, Ultimatum, and Charity) so that scenarios share a $[0, 1]$ scale. We then run the transformer over each generated response, mean-pool the L15 hidden states over its tokens under the attention mask, encode the pooled vector $\bar{h} = \mathrm{mean}_t\, h_t^{(15)}$ with the SAE, and form a scalar score $s = \sum_{j \in S} \mathrm{Enc}_{15}(\bar{h})_j$ by summing activations over one frozen side's sites—or a net score $s_D - s_C$ for classification.

Algorithm 1 Paired AB/CD Holdout Readout

Input: 60 matched cooperative and 60 matched defecting AB/CD personas; support threshold $k = 6$.
Output: held-out readout metrics across five splits.
for split seed $s \in \{0, \ldots, 4\}$ **do**
  Split by persona into 40 training and 20 held-out rows per class; keep AB/CD pairs together
  for $m \in \{\mathrm{AB}, \mathrm{CD}\}$ **do**
    Encode training answer-slot states $a_i^{(L)} \leftarrow \mathrm{Enc}_L(h_i^{(L)})$
    Select $\mathcal{F}_C^{(m)}$ with own-side support $\geq k$ distinct personas and opposite-side support $= 0$
    Select $\mathcal{F}_D^{(m)}$ by reversing the behavioral sides
  end for
  $S_C \leftarrow \mathcal{F}_C^{(\mathrm{AB})} \cap \mathcal{F}_C^{(\mathrm{CD})}$; $S_D \leftarrow \mathcal{F}_D^{(\mathrm{AB})} \cap \mathcal{F}_D^{(\mathrm{CD})}$
  Freeze $S = S_C \cup S_D$; fit scalers and ridge readouts on CD, AB, and pooled training rows
  Evaluate each readout on held-out personas using $a^{(L)}(x)[S]$
end for
return mean held-out metrics across splits

Because the coordinates are frozen at extraction, this readback needs no judge model and no additional training. Read transfer is reported per scenario as the Spearman rank correlation between the frozen Dictator-derived defection-site sum and the normalized expressed amount across generated responses (n=9 per game). A strong negative coefficient means the register preserves the *ordering* of allocation choices under a different incentive structure—higher self-interest-site activation accompanies less giving—not that it predicts a common dollar amount (Section 4.4).

3.4 Feature Intervention

To test whether task-local intervention features affect the decision, we edit the answer-position residual state at the localized layer while holding the persona prompt fixed. For a target behavior b, let S_b and $S_{\bar{b}}$ be its selected sites and the opposite-side sites, respectively, and let m_j be the mean activation of site j among behavior-verified b personas. We inject the target-side pattern and suppress the opposite-side pattern by adding an SAE-decoded correction to the original state:

$$h' = h + \sum_{j \in S_b} (\alpha m_j - z_j) W_j - \sum_{j \in S_{\bar{b}} \cap \mathrm{supp}(\hat{z})} \hat{z}_j W_j, \tag{3}$$

where $z = \mathrm{Enc}_L(h)$, $\hat{z}$ is the nonnegative pre-TopK activation, W_j is SAE decoder column j, and α controls injection strength. The additive update retains the original residual state, including its SAE reconstruction error. We continue the forward pass and compare the answer with its unedited baseline. Injection-only and masking-only ablations, amplitude sweeps, and size- and amplitude-matched random-site controls test the specificity of the effect (Section 4.3). The hook placement, intervention strengths, generation-time variant, and residual-direction baseline are specified in Section B.1.

Separately, we validate feature interpretations with text counterfactuals: we delete or insert phrases associated with an active feature, then measure both its activation and the answer-slot readout.

3.5 Representation-to-Language Inversion

As a proof of concept, we ask whether the discovered feature signature can guide persona synthesis. Starting from a readable seed persona, we seek a natural-language description that better matches the target representation, then test whether it induces the desired decision using the model's true answer-slot logits. This controlled case demonstrates a feasible path from a behavior-linked internal representation back to human-readable language, without claiming a general language decoder or improved prompt-search efficiency. The search and validation procedure is given in Section F.

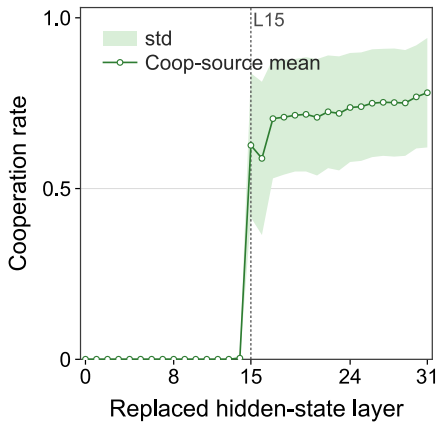

Figure 2: **Layerwise replacement.** Mean PD cooperation probability across cooperative-persona sources ($\pm 1\sigma$). L15 is the first writable layer.

The writability scan shows a sharp transition at L15 (Figure 2). Replacing the no-persona answer-slot state with cooperative-persona source states leaves the answer defection-aligned through L14. At L15, mean cooperation probability rises to 0.627, with $64/100$ sources crossing the 0.5 decision boundary. The scan uses 100 behavior-verified cooperative sources and scores only the legal C/D logits at the same answer slot used for localization; the shaded band therefore shows source-to-source variation rather than uncertainty from a fitted predictor. The effect strengthens at later layers (mean $p_C = 0.705$ at L17), so L15 is the first stable writable layer in this scan, not the only layer carrying the variable. Around the onset, both full hidden-state replacement and SAE-latent replacement cross the decision boundary at L15, whereas SAE reconstruction of the no-persona state remains defection-aligned. This distinction separates a causal write point from a layer that merely predicts the final answer, while later layers are already closer to the output.

Table 1: **Cross-persona holdout readout.** Mean over five matched 60/20 splits. Feature selection, scaling, and fitting use only the 40 training personas per class. Dictator R^2 predicts the verified persona class score, not the continuous donation amount.

Readout	R^2	MAE	Acc.
PD–CD	0.922 ± 0.005	0.070	0.985
PD–AB	0.949 ± 0.009	0.062	0.970
PD–AB+CD	0.932 ± 0.007	0.072	0.9975
Dictator (local L15)	0.959 ± 0.010	0.065	1.000

4 Experiments

4.1 Layer Localization of Strategic Behavior

In a separate pooled-source onset check, neither SAE-latent nor full hidden-state replacement flips the no-persona choice at L13–L14; at L15 they yield $\rho_C = 0.867$ and 0.777, respectively (Table 8). The mean values in Figure 2 and this pooled-source check use different aggregations. We use the earliest writable layer for the SAE analysis (Section D).

4.2 Localization and Readout of Strategy-Related Features

At the writable L15 layer, we test whether a compact SAE representation carries the persona-induced decision across answer-label realizations. In a paired AB/CD holdout, we form a balanced cohort of 60 cooperative and 60 defecting personas, hold out 20 of each class, and use only the remaining 40 of each class for feature selection and fitting. The same persona stays on the same side across AB and CD, and selected sites must be supported in both labelings. Across five splits, the training-selected sets contain 48–52 sites (mean 49.4). Held-out ridge readout reaches $R^2 = 0.922 \pm 0.005$ for CD, 0.949 ± 0.009 for AB, and 0.932 ± 0.007 for pooled AB+CD; with accuracies 0.985, 0.970, and 0.9975, respectively (Table 1). The pooled evaluation has 80 records from 40 distinct held-out personas. These results support a shared persona signal across the two labelings, within a behavior-verified cohort; they do not test unscreened prompts or causal sufficiency of the holdout-selected sites.

We rerun the same 60/20 protocol on Dictator rows, independently selecting sites from the Dictator L15 activations rather than reusing the PD set. The five splits select 200–224 sites (mean 210.2 ± 8.9) and classify held-out cooperative versus self-interested personas with the metrics in Table 1; the corresponding logistic readout is also perfect on held-out personas. Because these cached rows do

Table 2: **Bidirectional target outcomes at scale.** Delta-form injection of target-side sites plus masking of opposite-side sites, on n=20 Dictator and n=100 PD verified personas per cell. PD cells report the mean cooperation probability $\bar{p}_C$ with the fraction choosing C in parentheses. Random controls use size- and amplitude-matched site sets.

Game	Persona	Baseline	Intervention	Random sites
Dictator (donation)	defect → push coop	0.3	**50.0**	0
Dictator (donation)	coop → push defect	57.4	**0.0**	50
PD ($\bar{p}_C$, n=100)	defect → push coop	0.00	**0.60** (67% C)	0.00 (0% C)
PD ($\bar{p}_C$, n=100)	coop → push defect	0.76	**0.00** (0% C)	0.76 (100% C)

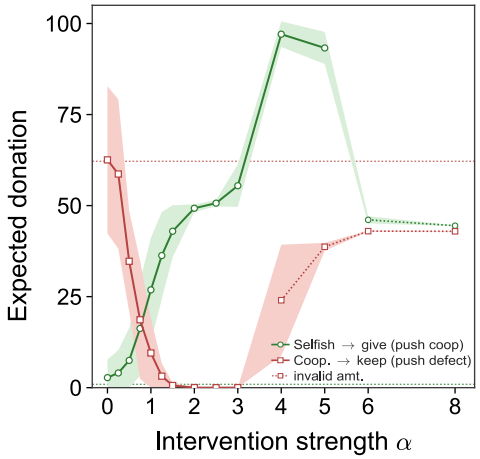

Figure 3: **Dictator intervention strength.** Mean expected donation over integer amounts 0–100 in a separate ten-persona-per-direction sweep, independent of the Table 2 cohort. At α=0, opposite-side sites are masked without target-side injection; dotted horizontal lines mark unedited baselines. Shading shows one standard deviation. Gray dotted tails mark strengths at which valid amount output degrades, so their conditional expectations are not interpreted as donations.

not contain per-persona donation labels for the holdout cohort, this R^2 is a persona-class score and should not be interpreted as donation prediction. The continuous donation intervention is reported separately below using Dictator-local causal sites.

The selected sets also show a small but reproducible shared substrate. Intersecting the feature sets selected in all five PD and Dictator holdout splits leaves three same-role features and no cross-role intersection. We name them *shared other-benefit/generosity* (SAE-88943), *shared conservative self-protection* (SAE-72754), and *shared self-gain retention* (SAE-110497), based on their cross-task side consistency, decoder vocabulary, and counterfactual semantic evidence. Thus the readout signal has a shared role-consistent component, while most writable coordinates remain task-local; the naming evidence and full overlap statistics are given in Appendix D.5.

4.3 Intervention and Semantic Validation

Task-local interventions change decisions in both PD and Dictator (Table 2): the task-local ensembles drive each verified persona group toward the opposite behavior, while matched random-site controls remain null. In a separate Dictator strength sweep (Figure 3), the expected amount given rises with cooperation-side injection and falls with defection-side injection. For self-interested personas, it increases approximately linearly through the middle of the valid range (16.26 at α=0.75, 26.85 at 1, and 36.28 at 1.25), then approaches equal giving and reaches 97.07 at α=4. In the reverse direction, it falls from 62.56 with masking alone to 9.53 at α=1 and 0.013 at α=2. Stronger edits can suppress valid amount outputs, so the dose response is not globally linear or unbounded (Appendix B.2).

Phrase-level counterfactuals give independent semantic validation: deleting the "generosity" phrase drops a cooperative prompt's ρ_C from 0.94 to 0.04, deleting "safety" drops it from 0.71 to 0.08, and inserting self-gain phrasing activates the defection-side sites. Feature-level enrichment and naming details are in Section E.

4.4 Ablations and Extensions

Payoff-matrix ablation. We hold the PD wording, C/D meanings, and intervention form fixed while varying only (R, T, S, P). The edit keeps its C-directed sign across the tested payoff variants; the full exploratory sweep and its baseline-choice diagnostics are reported in Appendix G.1. Its small separate cohort is not directly compared with the main $n{=}100$ bidirectional intervention result in Table 2.

Cross-game allocation readout. We test the fixed Dictator-local site sum on the same behavior-verified persona contrast in three games with different incentive structures but the same measurable decision axis: how much of an available resource the model allocates beyond itself. Public Goods makes a contribution to a four-player shared pot; Ultimatum makes a proposer offer subject to the responder's acceptance; Charity makes an unrewarded gift to an organization. The resulting amounts are not interchangeable in value or strategic meaning, but each can be ordered from less to more given. For each generated response, we mean-pool its L15 hidden states, encode that state with the SAE, and sum the fixed Dictator defection-side sites. Across generated responses, higher site activation accompanies a lower expressed amount within every game (Table 3). The negative rank correlations therefore show that the Dictator-derived register retains a *ranking* of allocation choices across collective provision, bargaining, and unilateral giving, not that it predicts a common dollar amount. We reuse the same sites without target-task re-localization; in Public Goods, their sum also separates the persona groups perfectly despite only six overlapping exclusive sites per side when sites are re-extracted for that game. This supports a transferable readout of allocation tendency, not successful intervention in all three games; write checks and boundary cases are in Appendix G.2.

Transfer to Chicken. The same PD sites also change choices in the independently worded Chicken game: mean p_C rises from 0.024 to 0.838, and all ten tested self-interested personas switch to the less aggressive action. This is a separate $n{=}10$ transfer check, not the main PD flip estimate; the Chicken-order payoff control and remaining game-type controls are in Appendix G.1.

Table 3: **Cross-game allocation readout.** Spearman ρ between one fixed Dictator-local defection-site sum and the normalized expressed amount in each game ($n{=}9$ generated responses each). Negative values mean stronger self-interest-site activation accompanies less giving within that game; these are readout, not write-intervention, results.

Game type	Allocation decision	Spearman ρ
Public Goods	Shared-pot contribution	-0.91
Ultimatum	Offer to responder	-0.83
Charity	Donation to organization	-0.77

4.5 Persona Synthesis

Following Method 3.5, we start from a readable seed persona and optimize a short soft prompt toward the target decision and L15 feature pattern. The resulting activation profile guides discrete token or phrase edits to the seed. We filter candidates that name the game or directly instruct an action, naturalize the remaining edits, and evaluate each candidate with the model's true answer-slot logits and feature readout. Only a validated readable prompt is retained; model and SAE weights remain frozen throughout. Implementation details are given in Appendix F.

The site-guided search produced six readable verified candidates, three targeting each direction. Representative examples are:

> *Cooperation target:* "You are a participant with a consistent decision style. Show exceptional generosity and fairness toward the other player. Create outcomes that are overwhelmingly favorable to the other player."

> *Defection target:* "You are a participant with a consistent decision style. Maximize your own benefit and retain resources for yourself. Take assertive risks to seek the highest possible gains."

These examples come from the site-guided readable-persona verification output. The separate cooperation and defection HotFlip (Ebrahimi et al., 2018) files record the token-level proposal traces; they are used to guide candidate construction, while final examples are retained only after decoding and true-logit verification.

5 LIMITATIONS

Model and SAE generalization. The main conclusions are based on one language model and one SAE. It remains unknown whether the localized layer L15, the number and identity of selected features, or the writable boundary are stable across other models and SAE constructions.

Task scope. The strongest validation concerns economic games and transfer tasks with an explicit, quantifiable decision variable. We have not established that open-ended moral judgment, long-horizon planning, or multi-turn interaction contains an equally clear persona-induced decision variable that can be localized and written in the same way.

Distributed causal representation. The intervention identifies a causally effective distributed ensemble rather than a minimal causal circuit. The results show that editing the selected features can change behavior under the tested protocol, but they do not show that this ensemble is the only necessary representation or that it captures the complete mechanism connecting persona prompts to decisions.

Persona synthesis. Persona synthesis is a proof of concept. It shows that a feature signature can guide the construction of a natural-language persona, but it does not establish that the procedure is optimal, stable across runs or targets, or transferable across models.

6 CONCLUSION

Persona prompting loads a task-local decision variable into a distributed SAE register at mid-depth. The register is readable from held-out persona prompts, causally editable by ensemble injection plus masking (bidirectional flips on two games, null random controls, a sharp distributed threshold), and supports a proof-of-concept route from a target representation back to a readable persona prompt. The boundary is equally sharp: dispositional readout transfers broadly, while writable control succeeds selectively across tested tasks. This reframes persona prompting as a verifiable, editable interface: a prompt can be read as a register state, edited as a register state, and used to guide a return to natural language.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and LaTeX drafting assistance. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies internal representations of a language model under persona prompts; it involves no human subjects and releases no new dataset. Persona-prompt synthesis methods could, in principle, be used to steer model behavior, so we report the method at the mechanism level and keep the synthesized prompts in the study confined to benign persona descriptions.

Reproducibility statement

Implementation details, model and SAE resources, run parameters, persona construction, localization criteria, paired holdout results, synthesis settings, and prompt templates are documented in the appendix. All reported numbers are produced by the pipelines described there.

Acknowledgments

Acknowledgments omitted for anonymous review.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nat. Hum. Behav.*, 9:1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

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 *Adv. Neural Inf. Process. Syst.*, volume 38, 2024.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? geometric limitations of steering in large language models. arXiv preprint; ACL ARR Jan. 2026 submission on OpenReview, 2026. URL https://openreview.net/forum?id=xnt4GUxDIx.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *Int. Conf. Learn. Represent.*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Valerio Capraro, Roberto Di Paolo, and Veronica Pizziol. A publicly available benchmark for assessing large language models' ability to predict how humans balance self-interest and the interest of others. *Sci. Rep.*, 15:21428, 2025. doi: 10.1038/s41598-025-01715-7.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *Proc. Natl. Acad. Sci. U.S.A.*, 120(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 31–36, 2018.

EleutherAI. Sae-llama-3-8b-32x: Sparse autoencoders trained on the residual stream of llama 3 8b. Hugging Face model repository, 2024. URL https://huggingface.co/EleutherAI/sae-llama-3-8b-32x.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proc. Int. AAAI Conf. Web Soc. Media*, volume 19, pp. 522–535, 2025. doi: 10.1609/icwsm.v19i1.35829.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In *Findings Assoc. Comput. Linguist.: EMNLP*, pp. 7174–7188, 2024. doi: 10.18653/v1/2024.findings-emnlp.420.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Work. Pap. 31122, National Bureau of Economic Research, 2023. Revised February 2026.

Tim Johnson and Nick Obradovich. Testing for completions that simulate altruism in early language models. *Nat. Hum. Behav.*, 9:1861–1870, 2025. doi: 10.1038/s41562-025-02258-7.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat. Mach. Intell.*, 6: 383–392, 2024. doi: 10.1038/s42256-024-00820-y.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In *Proc. NAACL-HLT*, pp. 4099–4113, 2024. doi: 10.18653/v1/2024.naacl-long.228.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Adv. Neural Inf. Process. Syst.*, volume 36, 2023.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Sci. Rep.*, 14:18490, 2024. doi: 10.1038/s41598-024-69032-z.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings Assoc. Comput. Linguist.: EMNLP*, pp. 23212–23237, 2025. doi: 10.18653/v1/2025.findings-emnlp.1261.

Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. Persistent personas? role-playing, instruction following, and safety in extended interactions. In *Proc. EACL*, pp. 5329–5359, 2026. doi: 10.18653/v1/2026.eacl-long.246.

Alex McKenzie, Keenan Pepper, Stijn Servaes, Martin Leitgab, Murat Cubuktepe, Mike Vaiana, Diogo de Lucena, Judd Rosenblatt, and Michael S. A. Graziano. Endogenous resistance to activation steering in language models. In *ICLR 2026 Workshop on Trustworthy AI*, 2026. URL https://openreview.net/forum?id=6LfXccur0Z.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Adv. Neural Inf. Process. Syst.*, volume 35, 2022.

Meta AI. Meta-llama-3-8b-instruct model card and license. Hugging Face model card and accompanying Meta Llama 3 Community License, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proc. NAACL-HLT*, pp. 11241–11258, 2025. doi: 10.18653/v1/2025.naacl-long.561.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proc. Int. Conf. Mach. Learn.*, volume 235 of *Proc. Mach. Learn. Res.*, pp. 39643–39666, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. arXiv preprint; submitted to ICLR 2025 on OpenReview, 2024. URL https://openreview.net/forum?id=2XBPdPIcFK.

Haochuan Wang, Xiachong Feng, Lei Li, Yu Guo, Zhanyue Qin, Dianbo Sui, and Lingpeng Kong. Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms. arXiv preprint; submitted to ICLR 2025 on OpenReview, 2024a. URL https://openreview.net/forum?id=1KvYxcAihR.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of ACL*, pp. 14743–14777, 2024b. doi: 10.18653/v1/2024.findings-acl.878.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proc. Int. Conf. Mach. Learn.*, volume 267 of *Proc. Mach. Learn. Res.*, pp. 67035–67080, 2025.

Yutong Xie, Qiaozhu Mei, Walter Yuan, and Matthew O. Jackson. Using large language models to categorize strategic situations and decipher motivations behind human behaviors. *Proc. Natl. Acad. Sci. U.S.A.*, 122(35):e2512075122, 2025. doi: 10.1073/pnas.2512075122.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings Assoc. Comput. Linguist.: EMNLP*, pp. 15126–15154, 2024. doi: 10.18653/v1/2024.findings-emnlp.888.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv*, abs/2310.01405, 2023. doi: 10.48550/arXiv.2310.01405. URL https://arxiv.org/abs/2310.01405. Last revised Mar. 3, 2025.

A Appendix

The appendix gives implementation and validation details for the three main claims. Section B lists model, SAE, and run parameters. Sections C and D specify the behavior-verified source construction, localization criteria, and paired holdout readout. Section E expands the sparse-feature evidence. Section F gives the representation-to-language inversion procedure. The transfer and ablation details cover the reported task-level checks. Section I lists prompt templates.

B Implementation Details

All layer numbers are zero-based. Models are used frozen; the experiments optimize neither model weights nor SAE weights. Table 4 lists the model and SAE resources used for the reported checks.

Table 4: **Model and SAE resources.** The reported experiments use one frozen model and one fixed SAE bank; task-specific checks reuse this setup while changing the decision prompt.

Setting	Model	SAE / representation	Reported layer use
Main PD	Meta-Llama-3-8B-Instruct	EleutherAI Llama-3 8B 32x SAE bank; answer-slot SAE latents.	L15 main localized layer; L31 late-layer readout control.
Dictator transfer	Same Llama-family setup as the main PD run.	Task-specific SAE localization and ridge readout.	L22 method-transfer check.

Table 5 gives the run parameters that affect the main reported numbers. Parameters not listed here are fixed implementation defaults rather than tuned experimental variables.

Table 5: **Run parameters for the main reported results.** The table separates source construction, feature selection, intervention, and synthesis parameters.

Component	Setting	Value used in reported run
PD source rows	Behavior-verified rows	100 cooperative and 100 defecting rows after leakage filtering.
Answer-slot scoring	PD answer prefix	Next-token position after {"choice":"; score only legal C and D logits.
Ridge validation	Paired AB/CD holdout	Matched cohort: 60 personas per class; 20 per class held out in each of five splits. Select AB∩CD sites using at least six own-side training personas per labeling and zero opposite-side activations; fit on the remaining 40 per class.
PD intervention	Task-local role-consistent core	Four-mapping extraction yields 55 cooperation-side and 28 defection-side sites; these are distinct from the holdout-selected readout sites.
Soft inversion	Trainable prompt	70 soft tokens optimized for 2 steps; best step selected by true ρ_C.
Hard prompt inversion	Candidate search	40 HotFlip (Ebrahimi et al., 2018) candidates; one candidate accepted after leakage filtering and true forward evaluation.

B.1 Feature Intervention Protocol

We implement the feature-level residual update with a forward hook on the selected decoder block (L15 in the main experiments; L13 and L21 are comparison layers). Prompt-side scoring edits only the last input position after the answer-field prefix; generation-time intervention follows the current last position. Target-side injection is ungated, so it can activate sites absent from the original TopK code. Opposite-side masking uses positive pre-TopK activations to remove sub-threshold mass as well as active TopK coordinates. We use α=1 for Dictator and α=4 for PD and evaluate the edited answer with the task's decision readout. Random-site controls match both set size and intervention amplitude. A residual-direction baseline instead adds a normalized difference-of-means vector to h. For the post-answer reasoning check only, a generation-time variant applies gated masking and amplification at each generated position with α=4.

B.2 Dictator Intervention Strength Sweep

The fine-grained dose check uses the Dictator-local L15 sets (448 cooperation-side and 451 defection-side sites) on the first ten behavior-verified personas from each side. The unedited baseline has no hook; α=0 still masks the opposite-side sites and is therefore a mask-only control. At the first amount token after {"donation": , we score all 101 integer candidates from 0 to 100. The reported expected value (EV) normalizes over those candidates, so we also check their total probability mass and whether greedy decoding begins with a valid amount. Only rows with substantial amount mass and a valid greedy first token are interpreted as donation outcomes.

Table 6: **Dictator dose response on n=10 personas per direction.** EV is the mean expected amount among 0–100 candidates. Valid is the number of greedy first tokens that are amounts. A dash means the conditional EV is not interpretable because valid amount output has degraded.

	Self-interested → giving		Generous → keeping	
α	EV	Valid	EV	Valid
Baseline	0.93	10/10	62.18	10/10
0	2.76	10/10	62.56	10/10
0.25	4.04	10/10	58.67	10/10
0.5	7.48	10/10	34.71	10/10
0.75	16.26	10/10	18.65	10/10
1	26.85	10/10	9.53	10/10
1.25	36.28	10/10	3.15	10/10
1.5	42.95	10/10	0.61	10/10
2	49.27	10/10	0.013	10/10
2.5	50.67	10/10	0.001	10/10
3	55.46	10/10	—	6/10
4	97.07	10/10	—	0/10
6	—	0/10	—	0/10

The middle of the self-interested-to-giving curve is approximately linear, but the full response is not: all ten personas have a modal amount of 50 by α=1.5, and all ten have a modal amount of 100 at α=4. In the reverse direction, all ten have a modal amount of 0 from α=1.25 through 2.5. Beyond the reliable window, candidate probability mass collapses (self-interested-to-giving:

0.00034 at α=6; reverse: 0.00024 at α=4), so the resulting conditional EV cannot be interpreted as a preference for a smaller donation. This scan uses the site-localization personas; the separate n=20 greedy-generation check at α=1 supports the main bidirectional result but is not a held-out dose-response test.

C BEHAVIOR-VERIFIED PERSONA CONSTRUCTION

Persona construction starts from a learned library rather than from hand-written pairs. The candidate pool is the public behavioral-code library of Xie et al. (2025): natural-language descriptions of strategic motivations learned from human economic-gameplay data, organized per source game (`system_prompts_*.csv`), where each entry describes a decision style, role, or background, carries an intended-behavior annotation (`desired_behavior`) and an elicited-choice history, and never names a target action letter. For the main PD experiments, we take 512 candidates across source games, prepend each to the same fixed PD user prompt, score the model's C/D probabilities at the answer slot, and keep a candidate only if the induced behavior verifies under the leakage filter below. The same construct-then-verify pipeline is re-run on the Dictator task, so pool membership is always assigned by measured behavior on the task at hand rather than inherited from the library's labels.

The source rows used in the main PD experiments are therefore selected by measured behavior. A candidate persona prompt is kept only if it induces the intended action under the fixed PD prompt and passes leakage filters.

The filtering rule is intentionally conservative. It removes direct action instructions and game-label leakage so that the persona text provides a behavioral contrast without simply telling the model which legal action to output.

C.1 REPRESENTATIVE VERIFIED ROWS

Table 7 gives short examples from the verified PD source pool. The table reports the persona text side and the true normalized PD cooperation score used for filtering.

Table 7: **Representative behavior-verified persona rows.** These rows illustrate that source labels are assigned after scoring the fixed PD prompt.

Verified side	Persona text excerpt	ρ_C
Coop	Highly cautious yet calculated decision-maker; takes minimal risks to ensure safety while still aiming for some gain; favors very safe choices with small, almost guaranteed benefit.	0.9669
Coop	Values conservative yet effective strategies; balances safety and reward; prefers low-risk choices that incrementally build points while avoiding hazards.	0.9466
Defect	Uncompromising and highly self-interested participant; seeks to retain the maximum possible benefit and minimize return or sharing with others.	0.0000
Defect	Risk-taking decision-maker aiming for the highest possible reward regardless of risk; embraces maximum risk to secure the ultimate goal.	0.0000

Algorithm 2 Behavior-Verified Persona Construction

1: Define two intended behavioral sides for the PD task: cooperation-aligned and defection-aligned.
2: Generate or edit candidate persona prompts that describe a decision style, role, or background without naming the target action.
3: Prepend each candidate persona prompt to the same fixed PD user prompt.
4: Score the model at the answer slot and compute the normalized action probabilities over C and D.
5: Reject candidates with label leakage, including explicit mentions of cooperation, defection, C, D, prisoner's dilemma, or direct action instructions.
6: Assign the candidate to the cooperative pool only if the model's answer-slot argmax is C.
7: Assign the candidate to the defecting pool only if the model's answer-slot argmax is D.
8: Balance the verified pools to 100 cooperative rows and 100 defecting rows for the main Llama PD experiments.
9: Use these rows as source states for layer localization, task-local intervention-site extraction, semantic checks, and persona-prompt synthesis; build the paired AB/CD holdout separately for readout fitting.

D Localization and Feature Selection Details

D.1 Answer-Slot Intervention Site

All localization experiments use the next-token position immediately after the JSON answer-field prefix. For PD, this is the position after `{"choice":"`. At a candidate layer L, we replace the no-persona target hidden state at that answer slot with a source hidden state from a behavior-verified persona row, then continue the forward pass and rescore the legal action logits. A layer is counted as writable when this replacement changes the true C/D distribution.

D.2 Layer-Onset Check

The pooled-source onset check compares SAE-latent replacement with full hidden-state replacement at the same answer slot. Neither edit crosses the decision boundary at L13–L14; both do so at L15. The values below use the pooled source aggregation described in the main text.

Table 8: **Pooled-source layer-onset check.** Entries summarize the normalized cooperation probability after replacing the no-persona answer-slot state; "D-aligned" indicates that the decision boundary was not crossed.

Layer	SAE-latent replacement	Full hidden-state replacement	Writable?
L13	D-aligned	D-aligned	No
L14	D-aligned	D-aligned	No
L15	0.867	0.777	Yes

D.3 One-Sided SAE Support

Let $\mathcal{C}$ and $\mathcal{D}$ be behavior-verified cooperative and defecting persona rows. Under the recorded TopK code, a site is cooperation-side if it activates in cooperative rows and not in defecting rows; defection-side is defined symmetrically. This is an operational sparse slice of a distributed amplitude code: pre-TopK activations can be positive on both sides, with different magnitudes. For the paired AB/CD readout below, the rule is applied only to training personas, independently in each labeling, before taking the intersection. The separately extracted four-mapping PD core is used for intervention, not for the paired holdout readout.

D.4 Cross-Persona Holdout Readout

To test whether the readout generalizes beyond the personas used for feature selection and across answer-label realizations, we construct a matched AB/CD cohort with 60 cooperative and 60 defect-

ing personas. Each split holds out 20 personas from each class and uses the remaining 40 per class for feature selection, standardization, and model fitting. The same persona remains on the same side in AB and CD, so no matched persona crosses the split. A cooperation-side site must be active in at least six distinct cooperative training personas in both AB and CD and inactive in all defecting training rows; the defection rule reverses the sides. Table 9 reports the selected counts for each split. The sets are frozen before holdout evaluation. These behavior-verified personas do not establish performance on unscreened prompts.

Table 9: **Train-selected AB∩CD sites by split.** Site identities and counts are reselected on the training personas in each split; there is no fixed-size feature set.

Seed	Cooperation-side	Defection-side	Total
0	30	20	50
1	27	21	48
2	26	22	48
3	29	20	49
4	31	21	52
Mean	28.6	20.8	49.4

Table 10: **Paired AB/CD holdout readout.** Mean performance over five splits; all feature selection and fitting use the 40 training personas per class. The pooled AB+CD evaluation contains 80 records from the same 40 held-out personas and is not 80 independent personas.

Readout	Holdout R^2	MAE	Acc.
CD	0.922 ± 0.005	0.070	0.985
AB	0.949 ± 0.009	0.062	0.970
AB+CD pooled	0.932 ± 0.007	0.072	0.9975

The same protocol is rerun with Dictator-local L15 activations and independent site selection. The selected counts are 224, 200, 202, 216, and 209 across seeds 0–4 (mean 210.2 ± 8.9). The corresponding held-out Ridge metrics are $R^2 = 0.959 \pm 0.010$, MAE 0.065 ± 0.011, and accuracy 1.000; logistic regression is likewise perfect on held-out personas. The target is the verified cooperative/self-interested persona class, because the cached holdout rows do not contain per-persona donation labels.

Table 11: **Dictator-local site counts by holdout split.** Sites are selected only from the Dictator training rows in each split.

Seed	Cooperation-side	Defection-side	Total
0	109	115	224
1	100	100	200
2	98	104	202
3	109	107	216
4	103	106	209
Mean	103.8	106.4	210.2

D.5 Cross-Task Feature Overlap

The holdout experiments select sites independently for PD and Dictator, so feature identity rather than set size is the relevant comparison. The strictest comparison takes the intersection of the selected feature IDs across all five splits within each task. The PD cooperation-side set has one feature in common with the Dictator cooperation-side sets (ID 88943), and the PD defection-side set has two in common with the Dictator defection-side sets (IDs 72754 and 110497). The two cross-role intersections are empty. These three features are therefore shared across every split, but they are not claimed to be sufficient for intervention.

For context, a descriptive comparison using the same 100 cooperative and 100 self-interested prompts in both tasks gives larger but still sparse overlaps. The sets in this comparison are extracted from the full cohort and are not holdout estimates.

Table 12: **PD–Dictator feature overlap on the matched full cohort.** One-sided sets use the strict six-prompt support rule and are extracted independently from the same 100+100 personas. The expected count is the uniform random overlap for the observed set sizes.

Intersection	Observed	Expected	Ratio
PD-C ∩ Dictator-C	16	0.23	$70.5\times$
PD-D ∩ Dictator-D	12	0.11	$114.7\times$
PD-C ∩ Dictator-D	3	0.26	$11.6\times$
PD-D ∩ Dictator-C	1	0.09	$10.9\times$

The full-cohort feature IDs are listed explicitly below. This descriptive overlap should not be read as a universal intervention set: the majority of sites are task-specific and the continuous Dictator write test uses Dictator-local sites.

Table 13: **Feature IDs in the matched full-cohort overlap.**

Set	Feature IDs
PD-C ∩ Dictator-C	5025, 5631, 10779, 18916, 28323, 29429, 30450, 33573, 36192, 42418, 70523, 83602, 86340, 88943, 115138, 125157
PD-D ∩ Dictator-D	6740, 11762, 26669, 27722, 49901, 83019, 91412, 105055, 109741, 110497, 117877, 130177
PD-C ∩ Dictator-D	13010, 80204, 117396
PD-D ∩ Dictator-C	67003

Counterfactual naming of the three holdout-stable overlaps. The names below are semantic glosses, not claims that a single SAE coordinate is a complete causal mechanism. We combine the five-split role identity, decoder vocabulary, activation support in both tasks, and phrase-level counterfactuals. Where the counterfactual was run on a semantically matched feature cluster rather than on the isolated shared coordinate, we state this explicitly.

Table 14: **Interpretation of the three PD–Dictator holdout-stable shared features.** Counterfactual evidence is reported at the isolated-feature level when available; otherwise it is cluster-level evidence for the proposed semantic gloss.

Feature	Proposed name	Activation / decoder evidence	Counterfactual evidence
88943	shared other-benefit / generosity	Coop–coop in all five splits; cooperation-side in both tasks; other-benefit, generosity, altruistic, well-being, and favor-the-other-party vocabulary in the surrounding semantic cluster.	Cluster-level phrase edits support the gloss: inserting other-benefit/generosity language raises the cooperation readout, while deleting generosity or other-party-benefit phrases sharply lowers it (e.g., $0.9399 \rightarrow 0.0373$ for a matched generosity feature).
72754	shared conservative self-protection	Defect–defect in all five splits; Dictator low-give support $250/250$ with mean activation 0.370; decoder vocabulary includes conservative/conservatism.	Isolated Dictator counterfactual: replacing or zeroing this low-give feature changes mean donation from 13.62 to 15.88; the direction is consistent with weakening self-protective low giving, but the effect is small relative to the full local ensemble.
110497	shared self-gain retention	Defect–defect in all five splits; decoder and enriched phrases include ego/selfish, personal gain, keeping, retaining, and resources-yourself.	Cluster-level phrase edits support the gloss: inserting self-gain/resource-retention language increases the defect-side activation and suppresses the already-small cooperation probability. The isolated coordinate is therefore named from convergent semantic evidence, not treated as a standalone causal switch.

The shared features should therefore be read as compact cross-task probes of persona-linked allocation tendencies. Their names summarize the direction of the evidence: concern for the other party (88943), cautious self-preservation (72754), and retaining resources for oneself (110497). The Dictator intervention results still require the full Dictator-local ensemble, so these names do not imply a common cross-task write vector.

E SEMANTIC COUNTERFACTUAL DETAILS

Feature names are operational semantic glosses. We assign a gloss only after three checks agree: active-row phrase enrichment, the feature's behavioral side, and phrase-level counterfactual edits. Table 15 expands the compact semantic table in the main text.

Table 15: **Representative counterfactually validated feature interpretations.** The rows show feature-level examples where phrase enrichment and text edits support the provisional semantic gloss.

Feature	Provisional gloss	Enriched phrases	Counterfactual evidence
4615; Coop; 61/0	Generosity / positive outcome	generosity; fairness; satisfaction; valued	Delete "selflessness": $0.9859 \rightarrow 0.8670$; insert generosity/fairness: $0.0009 \rightarrow 0.7311$.
130722; Coop; 42/0	Other benefit / generous proposal	generosity; proposals; valued; altruistic	Delete "generosity": $0.9399 \rightarrow 0.0373$; delete "benefit the other party": $0.9399 \rightarrow 0.2451$.
49043; Coop; 19/0	Safety / risk aversion	conservative; safety; cautious; low-risk	Delete "safety": $0.7058 \rightarrow 0.0759$; delete "minimize risk": $0.7058 \rightarrow 0.3775$.
70523; Coop; 32/0	Steady safe outcome	safety; conservative; risk-averse; steady	Insert "steady and secure outcomes": $0.0009 \rightarrow 0.9399$; delete "cautious": $0.5622 \rightarrow 0.2018$.
117877; Defect; 0/26	High-risk return seeking	bold; highest possible; maximize returns	Bold high-risk insertions raise target activation and lower already-small ρ_C; pass-both rate is 0.33.
128357; Defect; 0/94	Self-gain / resource retention	gain; personal; retain; resources	Self-gain insertions raise target activation and lower already-small ρ_C; pass-both rate is 0.29.

F REPRESENTATION-TO-LANGUAGE INVERSION

The inversion procedure has two stages. A soft-prompt optimization stage finds a target activation profile that matches a desired strategic behavior. A discrete stage then proposes token or phrase edits, filters direct label leakage, naturalizes the accepted edit, and re-evaluates the prompt with the true answer-slot logits. Gradients are used only to rank candidate replacements; a candidate is accepted only after leakage filtering and a true forward pass. This separate pilot uses a PD-local feature readout; its Ridge scores are search diagnostics, not estimates from the paired AB/CD holdout. The result is a proof-of-concept inverse path from a target representation to readable language, rather than a general-purpose language decoder.

For hard-prompt inversion, let e_i be the embedding at persona-token position i and E_v the embedding of a candidate replacement token. The first-order proposal score is $g_i^\top (E_v - e_i)$, where $g_i = \partial J_\tau / \partial e_i$. We use this score to form a small candidate beam, then evaluate each naturalized candidate with the true answer-slot logits and the target-layer feature readout.

Here J_{logit} measures the target action and valid-action mass, while β and γ weight the feature readout and the distance to the soft target profile. The final comparison uses the same true forward evaluation for the current prompt and every candidate.

Table 16: **Proof-of-concept representation-to-language inversion.** Soft prompts provide a target feature profile; discrete prompts are accepted only after leakage filtering and true forward evaluation.

Stage	Candidate source	Filter / evaluation	ρ_C	Ridge	Note
Seed persona	One readable boundary prompt	No direct C/D labels; true-logit evaluation	0.5000	0.4005	Natural disaster-response persona used as the seed.
Optimized soft prompt	One trainable soft prompt	Valid-action mass checked during optimization	0.9983	0.5625	Soft prompt supplies the target feature profile.
HotFlip (Ebrahimi et al., 2018) inversion	40 hard candidates	Explicit PD/C/D leakage rejected; one accepted	0.9047	0.4516	Raw token edit gives a useful direction.
Naturalized inversion	One readable rewrite	No explicit action label; true-logit evaluation	0.9047	–	Final prompt keeps the behavioral shift after naturalization.

Algorithm 3 Representation-to-Language Inversion with Evaluation Feedback

Input: seed prompt p_0, target behavior τ, selected sites S, readout $\hat{\rho}_C$, steps T, beam size K.
Output: best validated readable prompt $p^\star$.
$q \leftarrow \mathrm{SoftOpt}(p_0, \tau, \rho_C, \hat{\rho}_C)$
$z^\star \leftarrow a^{(L)}(q)[S], \quad p \leftarrow p_0, \quad p^\star \leftarrow p_0$
for $t = 1, \ldots, T$ **do**
  $z \leftarrow a^{(L)}(p)[S]$
  $J_\tau(p) \leftarrow J_{\mathrm{logit}}(p, \tau) + \beta \hat{\rho}_C(z) - \gamma \|z - z^\star\|_2^2$
  $g_i \leftarrow \partial J_\tau / \partial e_i$ for each persona-token position i
  $\mathcal{C} \leftarrow \mathrm{TopK}_{i,v} \; g_i^\top (E_v - e_i)$
  $\mathcal{V} \leftarrow \mathrm{Filter}(\mathrm{Naturalize}(\mathcal{C}))$
  if $\mathcal{V} = \emptyset$ **then**
    break
  end if
  for $c \in \mathcal{V}$ **do**
    $s_c \leftarrow \mathrm{Eval}(c; \rho_C, \hat{\rho}_C, z^\star)$
  end for
  $c^\star \leftarrow \arg\max_{c \in \mathcal{V}} s_c$
  $s_p \leftarrow \mathrm{Eval}(p; \rho_C, \hat{\rho}_C, z^\star)$
  if $s_{c^\star} \leq s_p$ **then**
    break
  end if
  $p \leftarrow c^\star, \quad p^\star \leftarrow \mathrm{Best}(p^\star, p)$
end for
return $p^\star$

G Transfer and Ablation Details

The main paper treats transfer cautiously. Directly reusing a frozen PD controller gives strong evidence in Chicken but weak or null evidence in several other tasks, while task-specific localization recovers predictors and intervention effects for the Dictator allocation variable. The following checks therefore distinguish readout transfer, fixed-intervention transfer, and payoff or game-type ablations.

G.1 Payoff-Matrix and Game-Type Ablations

The matrix sweep holds the PD description, the C=cooperate/D=defect mapping, and the L15 intervention fixed (55 cooperation-side sites injected at α=4 and 28 defection-side sites masked). Only the payoff entries (R, T, S, P) change. Here p_C is the probability of C normalized over the two legal answer tokens, and a flip means a self-interested persona moves from $p_C < 0.5$ to $p_C \geq 0.5$. All eight rows use the same ten self-interested personas, a separate cohort from the main n=100 flip check. Their unedited mean p_C is near zero in every row. These exploratory counts are not a replacement for, or a direct estimate of, the main bidirectional result: the main PD intervention uses its verified n=100 cohort. The matrix sweep is retained here to characterize how changing numerical payoffs modulates a fixed edit under a smaller separate cohort.

Table 17: **Fixed-template PD payoff sweep.** The site edit is identical in all rows. Edited p_C and flip counts refer to the ten self-interested personas; the final column is the edited no-persona prompt.

Matrix	(R, T, S, P)	Edited p_C	Flips	No-persona p_C
Standard PD	$(3, 5, 0, 1)$	0.383	4/10	0.999
PD scaled $\times 10$	$(30, 50, 0, 10)$	0.716	9/10	0.998
Shallow PD	$(3, 4, 2, 2.5)$	0.642	8/10	0.998
Chicken-order payoffs	$(3, 5, 0, -10)$	0.580	6/10	0.999
Harmony, C dominant	$(5, 3, 2, 1)$	0.761	8/10	0.999
Coordination	$(3, 1, 0, 3)$	0.536	6/10	0.999
Deadlock, D dominant	$(3, 5, 0, 4)$	0.270	0/10	0.998
Equal payoffs	$(3, 3, 3, 3)$	0.138	0/10	0.997

The same intervention raises the C score in this fixed template, but numerical scale alone changes how often it crosses a persona's decision threshold: standard PD and its $\times 10$ version have the same

payoff ordering yet yield $4/10$ and $9/10$ flips. The unedited model's choices also limit what counts as a clean payoff ablation. Under an explicit Harmony prompt stating that C always pays more, $0/10$ self-interested personas choose C without intervention, while $10/10$ cooperative personas do. Under prompts stating that D always pays more, the self-interested group again chooses D in all ten cases, whereas only $2/10$ cooperative personas choose C in standard PD and $3/10$ in Deadlock. With equal payoffs, the groups still choose different letters ($0/10$ versus $10/10$ choose C), although the matrix itself has no preferred action. Harmony and equal-payoff rows are therefore diagnostic checks of baseline model behavior, not clean estimates of rational payoff sensitivity; the coordination row likewise has no single dominant action. A zero flip count in such a row cannot by itself be read as a payoff-based resistance to the intervention. These observations do not prove that the model ignores the matrix altogether.

The game-type screen instead keeps the fixed PD site set and C/D answer slot while changing the target task's story, options, and payoff structure together. It is a transfer probe rather than a controlled intervention on game type alone. The initial screen uses five self-interested personas per task; Chicken was separately extended to ten.

Table 18: **Fixed-PD-site game-type screen.** Entries are baseline $\rightarrow$ edited p_C under the unchanged PD intervention; flip counts refer to self-interested personas in the five-row initial screen.

Target game	Meaning of C	No persona p_C	Self-interested p_C	Flips
Binary Dictator	Split \$50	$0.50 \rightarrow 0.94$	$0.000 \rightarrow 0.007$	$0/5$
Stag Hunt	Hunt stag together	$0.11 \rightarrow 0.01$	$0.000 \rightarrow 0.000$	$0/5$
Chicken	Swerve aside	$0.73 \rightarrow 1.00$	$0.002 \rightarrow 0.884$	$5/5$
Trust	Send \$10	$0.90 \rightarrow 0.99$	$0.434 \rightarrow 0.580$	$1/5$
Ultimatum	Offer \$50	$0.56 \rightarrow 0.90$	$0.000 \rightarrow 0.001$	$0/5$

In the ten-persona Chicken extension, mean p_C rises from 0.024 to 0.838 and all ten choices flip. This differs from the Chicken-order *matrix* row above, which retains the PD story and flips $6/10$. Standard PD has payoff order $T > R > P > S$, while this Chicken game has $T > R > S > P$; the success is therefore not evidence that an identical payoff order causes transfer. Conversely, the binary Dictator and Ultimatum tasks retain a C/D answer format but show no self-interested-persona flips. Swapping action-letter meanings is a separate control for role semantics versus letter-channel writing, not part of either ablation. Since the initial game-type screen has only five personas per task and changes several prompt factors at once, it does not isolate a general causal effect of game type.

G.2 Within-Family Intervention Transfer

We distinguish transfer of a *fixed intervention* from transfer of a feature readout. In the former, the source-task site ensemble is applied to a new question and the target task's actual decision is measured again. In the latter, source-task activations are only correlated with an expressed decision; this does not establish a causal effect on that question. Table 19 lists the tested write outcomes separately from the read-only checks below.

Table 19: **Direct write checks beyond the source game.** A source-task ensemble is reused on a target question. Public-goods values are individual output shifts, not a flip rate; the take-framed check has no aggregate metric in the present record.

Source sites	Target question	Observed write result
PD letter-choice	Chicken, related strategic choice	Target-choice score $\approx 0.02 \rightarrow 0.84$; $10/10$ tested decisions flip.
Dictator amount-allocation	Public goods, contribution amount	Two single-prompt shifts: defection-side edit $15 \rightarrow 0$; cooperation-side edit $1 \rightarrow 5$.
Dictator amount-allocation	Take-framed allocation amount	Qualitative directional examples; taking more is coded as more self-interested, opposite to giving more. No pooled numerical effect is reported.
PD letter-choice	Dictator, donation amount	No successful write transfer in the tested setting.

The PD-to-Chicken result tests whether a letter-choice intervention survives one related strategic game, not whether it becomes an amount controller. The Dictator-to-public-goods and take-framed checks instead keep a scalar allocation decision while changing the question's surface framing and,

for take, reversing which direction of the amount is prosocial. We therefore orient the behavioral interpretation to the target variable: higher giving and lower taking both benefit the other party. These limited amount-task examples do not establish stable family-wide write transfer; an additional public-goods check degrades at α=2. The failed PD-to-Dictator write check marks the cross-format boundary.

The read-transfer evidence has a wider but different scope. The Dictator defect-site sum tracks expressed amounts in public goods ($\rho = -0.91$), ultimatum (-0.83), and charity (-0.77), and, more weakly, trust sending (-0.47). Life-scale amount-allocation prompts give $\rho \approx -0.45$ to -0.86 (exploratory scenarios; Appendix I), with the expected sign reversal in take-framed prompts. These scores are associations with generated decisions; no intervention result is asserted for ultimatum, charity, trust, or the full life-scale collection. Similarly, the perfect public-goods persona separation does not imply identical feature membership: its newly extracted exclusive sets share only six sites per side with the Dictator sets. The direct write evidence is strongest for PD-to-Chicken and remains preliminary for the amount-task examples.

H Extended Related Work

LLMs are increasingly studied as agents in economic games, where explicit actions and payoffs make strategic behavior easier to measure than in open-ended dialogue. Prior work evaluates whether LLM choices resemble human economic behavior, whether they respond coherently to incentives, and how game structure changes their decisions (Horton et al., 2023; Chen et al., 2023; Akata et al., 2025; Fontana et al., 2025; Lorè & Heydari, 2024; Wang et al., 2024a). Closely related benchmarks study social preferences such as altruism, self-interest, and concern for others (Capraro et al., 2025; Johnson & Obradovich, 2025). These studies establish the behavioral setting for our work. We ask a different question: where does a measured strategic difference appear inside the model, and can that state be written back into a fixed target prompt?

Persona and role prompts provide a practical way to elicit behavioral variation, but their interpretation is contested. Role-play and persona prompting can improve or change model behavior in some settings (Kong et al., 2024; Wang et al., 2024b; Lutz et al., 2025; Giorgi et al., 2024), while other work shows that persona effects can be bounded, unstable, or ineffective for the intended task (Zheng et al., 2024; Luz de Araujo et al., 2026; Kirk et al., 2024). We therefore do not treat personas as stable model traits. In this paper, personas are contrastive probes: they provide behavior-verified cooperative and defecting source states for the same PD task.

Our intervention design builds on work that reads, edits, or steers internal representations. Model-editing and inference-time intervention methods show that localized hidden states can affect downstream answers (Meng et al., 2022; Li et al., 2023). Representation engineering and activation steering extract directions associated with target behaviors and use them at inference time (Turner et al., 2024; Zou et al., 2023; Arditi et al., 2024). Linear-representation analyses give a broader geometric framing for why such directions can be informative (Burns et al., 2023; Park et al., 2024). Our setting differs in two ways: the target behavior is a strategic choice in an economic game, and the source states come from behavior-verified persona contrast rather than from paired factual prompts or direct behavior labels alone.

Sparse autoencoders provide the feature coordinates used in our second stage. Prior work uses SAEs to recover interpretable features from language-model activations and studies how SAE scale and placement affect feature quality (Cunningham et al., 2023; Gao et al., 2024). At the same time, recent work cautions that SAE-based steering must be compared against simpler baselines and that steering can face geometric or endogenous resistance (Wu et al., 2025; Bhandari et al., 2026; McKenzie et al., 2026). We use SAE features as local coordinates for prediction, semantic checks, and intervention, not as a claim that individual features are complete causes.

I Prompt Templates

Prisoner's Dilemma. The no-persona PD prompt used for localization is:

> You are playing a one-shot prisoner's dilemma. You must choose exactly one action: "C" = cooperate; "D" = defect. Payoffs are written as (your payoff, other player's payoff): if both

> choose C: (3, 3); if you choose D and the other player chooses C: (5, 0); if you choose C and the other player chooses D: (0, 5); if both choose D: (1, 1). Known conditions: this is a one-shot interaction; you do not know what the other player will do; choose the single best action from your perspective. Return JSON only. Do not output any text before or after the JSON. Use exactly this schema: {"choice":"C_or_D"}.

The PD answer slot is the token position immediately after {"choice":".

Relabeled mappings for role consistency. The cross-mapping checks in the main text reuse the same game with the role assignment of the action letters swapped. Under the **CD** mapping above, the payoff structure is unchanged: C is the action that gives the opponent more and the self less. Under the **DC** mapping, the letter meanings are reversed while the payoff numbers stay the same:

> You are playing a one-shot strategic interaction. You must choose exactly one action: "D" = cooperate; "C" = defect. Payoffs are written as (your payoff, other player's payoff): if both choose D: (3, 3); if you choose C and the other player chooses D: (5, 0); if you choose D and the other player chooses C: (0, 5); if both choose C: (1, 1). Known conditions: this is a one-shot interaction; you do not know what the other player will do; choose the single best action from your perspective. Return JSON only. Do not output any text before or after the JSON. Use exactly this schema: {"choice":"C_or_D"}.

The two prompts differ only in which letter names the cooperative action. A feature that follows the cooperative *role* fires for coop-verified personas under both mappings; a feature tied to the letter "C" flips sides between CD and DC. The neutral A/B paragraph below applies the same relabeling with arbitrary letters.

Dictator. The Dictator prompt is:

> You receive a private $100 endowment. You can give any integer amount X from $0 to $100 to another anonymous participant. You receive no recognition or future reward. Choose exactly one INTEGER X. Return JSON only. Use exactly this schema: {"give": X}.

The answer slot is after {"give": , and the candidate values are integers from 0 to 100 in steps of 10 for the reported scoring runs.

Public Goods. The Public Goods prompt is:

> You are participating in a one-round public goods game with 3 other players. Each player receives $20. You must choose exactly one INTEGER amount from 0 to 20 indicating how many dollars you contribute to the public pool. All contributions from the 4 players are added together. Each player then receives 50% of the total group contribution, regardless of how much they personally contributed. Return JSON only. Use exactly this schema: {"contribution": X}.

The answer slot is after {"contribution": , and the candidate values are integers from 0 to 20 in steps of 2.

Trust send. The Trust boundary prompt is:

> Your user starts with 90 points. Your user can send exactly one amount from 0 to 90 to another player. Any amount sent will be tripled before the other player receives it. Return JSON only. Use exactly this schema: {"send_amount": X}.

This task is used as a boundary case for the transfer analysis.

Neutral A/B payoff generalization. The A/B prompt removes the semantic labels "cooperate" and "defect." In the canonical A/B version, A maps to the C-like action and B maps to the D-like action:

> Two anonymous participants each choose exactly one action, A or B. The action labels are arbitrary. The payoff table gives (your payoff, other participant's payoff): A/A = (3, 3), A/B = (0, 5), B/A = (5, 0), and B/B = (1, 1). Implications of the table: choosing A gives

> the other participant more payoff than choosing B, no matter what the other participant chooses; choosing B gives you more payoff than choosing A, no matter what the other participant chooses. Known conditions: this is a one-shot interaction; you do not know what the other participant will do; choose according to your assigned decision style and the payoff table. Return JSON only. Do not output any text before or after the JSON. Use exactly this schema: {"choice":"A_or_B"}.

For payoff-generalization runs, the same template is reused with (R, T, S, P) varied in the table entries $A/A = (R, R)$, $A/B = (S, T)$, $B/A = (T, S)$, and $B/B = (P, P)$. The A/B readout is $\rho_A = p_A/(p_A + p_B)$.