What Suppresses Nash Equilibrium Play in Large Language Models? Mechanistic Evidence and Causal Control

Anonymous authors
Paper under double-blind review

Abstract

LLM agents are known to deviate from Nash equilibria in strategic interactions, but nobody has looked inside the model to understand why, or asked whether the deviation can be reversed. We do both. Working with four open-source models (Llama-3 and Qwen2.5, 8B to 72B parameters) playing four canonical two-player games, we first establish the behavioral picture through self-play and cross-play experiments, then open up the 32-layer Llama-3-8B model and examine what actually happens during a strategic decision. The mechanistic findings are clear. Opponent history is encoded with near-perfect fidelity at the very first layer (96% probe accuracy) and consumed progressively by later ones, while Nash action encoding is weak throughout, never exceeding 56%. There is no dedicated Nash module. Instead, the model privately favors the Nash action through most of its forward pass, but a prosocial override (a bias toward cooperative, other-regarding behavior rooted in pretraining on human text and further modulated by RLHF) concentrated in the final layers reverses this, reaching 84% probability of cooperation at layer 30. When we inject a learned Nash direction into the residual stream, the behavior shifts bidirectionally and causally, confirmed through concept clamping. The behavioral experiments surface six scale- and architecture-dependent findings in self-play, the most notable being that chain-of-thought reasoning worsens Nash play in small models but achieves near-perfect Nash play in models above 70B parameters. The cross-play experiments reveal three phenomena invisible in self-play: a small model can unravel the cooperation of any partner simply by defecting early; two large models reinforce each other's cooperative instincts indefinitely; and who moves first in a coordination game determines which Nash equilibrium the system lands on. The central finding is that LLMs do not lack Nash-playing competence. They compute it, then suppress it.

large language models, activation steering, multi-agent systems, opponent modeling

1 Introduction

Twenty-five years ago, Broder et al. (2000) observed that the World Wide Web had grown so large and complex that it could no longer be understood through analytical derivation alone. It had become the first human artifact that demanded an empirical science of its own structure: to learn what it was, one had to observe and experiment rather than reason from first principles. Large language models occupy an analogous position today. With hundreds of billions of learned parameters and behaviors that emerge from pretraining rather than explicit design, they resist analytical study in the same way. The field of mechanistic interpretability exists precisely to bring observational methods to these artifacts. This paper applies those methods to a concrete and consequential question about LLMs in strategic settings.

A growing body of empirical work shows that LLM agents fail to converge to Nash equilibria in repeated strategic games (Brookins and DeBacker, 2024; Akata et al., 2025; Fontana et al., 2025; Jia et al., 2025). These deviations are systematic, shaped by game structure, prompting style, and model family, and qualitatively resemble the human behavioral biases documented in experimental economics (Camerer and Ho, 1999; Goeree and Holt, 2001). The behavioral question is settled: LLMs do not reliably play Nash.

The mechanistic question of *why* this is so, and whether it is causally reversible, has received no attention. This is the question we answer.

The gap matters for two reasons. Without a causal account, behavioral observations are difficult to act on: one cannot know whether prompting, fine-tuning, or architectural changes will fix the deviation or merely mask it. And if LLMs encode Nash-relevant representations that are suppressed rather than absent, that is a fundamentally different diagnosis from a model that lacks strategic competence. We show the former is true.

The closest mechanistic precedent probes small transformers trained specifically on a synthetic game (Othello), recovering board-state representations that are causal with respect to move prediction (Li et al., 2023). Our setting differs in every important dimension: we apply mechanistic tools to general-purpose instruction-tuned LLMs playing repeated strategic games, where any strategic competence must emerge from pretraining rather than game-specific training.

Contributions. This paper makes four contributions. We show through linear probing that opponent history is encoded with near-perfect fidelity (96%) from the very first layer of the model and decays monotonically as later layers consume it, while Nash action encoding is weak throughout ($\leq 56\%$), ruling out any dedicated Nash module. We use the logit lens to show that the model internally favors the Nash action through most of its forward pass, but a late-layer prosocial override suppresses it; we characterize this circuit through attention head analysis and activation patching. We extract a Nash direction from the residual stream and demonstrate reliable, monotonic causal control over equilibrium play: steering toward Nash achieves 99.2% defection in Prisoner's Dilemma, and concept clamping confirms the direction is causal rather than a spurious residual feature. We run cross-play experiments across all 12 ordered heterogeneous pairings of four models and show that Nash outcomes depend on population composition: a small model breaks cooperative locks by defecting early and inducing partners to follow, two large models reinforce each other's cooperative prior in direct prompting, and the agent role assignment determines which equilibrium is reached in coordination games.

Behavioral self-play results across four models and three reasoning conditions appear in Section 5 as context for the mechanistic account.

2 Related Work

This paper connects three bodies of literature: behavioral studies of LLMs in strategic settings, mechanistic interpretability of transformer models, and activation steering as a tool for causal intervention.

2.1 LLMs in Strategic Settings

Several papers establish the behavioral baseline this work builds on. Brookins and DeBacker (2024) find that GPT-4 shows human-like cooperation in prisoner's dilemmas, which they attribute to RLHF-induced prosocial preferences. Akata et al. (2025) show LLMs can sustain cooperation in repeated games with sufficient context. Fontana et al. (2025) study three LLMs playing the iterated Prisoner's Dilemma against opponents with varying hostility levels, finding that Llama2 and GPT-3.5 cooperate more than typical humans and are especially forgiving, while Llama3 behaves more like a human player. Jia et al. (2025) evaluate 22 LLMs using a behavioral game-theoretic framework and find that chain-of-thought prompting improves but does not guarantee Nash approximation, and that model scale alone does not determine performance. We treat these findings as a starting point and focus on the mechanistic question they leave open.

2.2 Mechanistic Interpretability and Activation Steering

Mechanistic interpretability seeks to identify the algorithms implemented by neural networks (Olah et al., 2020; Elhage et al., 2021). Meng et al. (2022) show that factual associations are localized to specific MLP (multi-layer perceptron) layers. Attention head analysis has characterized heads for indirect object identification (Wang et al., 2023) and copy suppression (McDougall et al., 2023). The logit lens (Nostalgebraist, 2020), formalized by Belrose et al. (2023), tracks the model's evolving token prediction by projecting each layer's hidden state

through the output embedding matrix. The most relevant structural precedent is Li et al. (2023), who probe a small GPT model trained on Othello sequences and recover causal board-state representations. We extend this to general-purpose instruction-tuned LLMs: no game-specific training is used, the games involve repeated interaction rather than perfect-information board states, and we go beyond probing to causal steering.

Activation steering modifies model behavior by intervening on hidden states during inference (Turner et al., 2023). It has been applied to honesty (Zou et al., 2023) and generation style, showing that behavioral attributes encode as linear directions in activation space. Two recent papers apply it to economic games. Ma (2025) probe and steer LLM representations in a Dictator Game, showing that injecting demographic vectors shifts giving behavior. Sun and Zhang (2026) construct persona vectors for altruism in Qwen-2.5-7B, steering at a single fixed layer across six games. They report that positive steering reliably increases prosocial behavior while negative steering has weaker effects, and flag mechanistic circuit identification as future work.

Our work differs from both in many ways. We target a Nash-specific direction benchmarked against analytically computed equilibria, probe all layers rather than fixing a single one, demonstrate that head ablation leaves behavior unchanged, and confirm causality through concept clamping. Our setting involves genuine repeated two-agent interaction over 50 rounds with a shared history, not one-shot games. The asymmetry Sun and Zhang (2026) observe between positive and negative steering follows from the suppression circuit we identify: positive steering amplifies an already-active cooperative mechanism while negative steering must overcome it. Gemp et al. (2024) use game-theoretic solvers to steer LLM decoding at the prompt level rather than on internal activations.

3 Background

Let's briefly review the game-theoretic concepts and neural network tools that our experiments rely on.

3.1 Nash Equilibrium

A *finite strategic game* consists of a finite set of players $N = \{1, \ldots, n\}$, where each player i has a finite set of actions A_i, a mixed strategy simplex $\Delta(A_i)$ over those actions, and a payoff function $u_i : \prod_j A_j \to \mathbb{R}$ that maps action profiles to real-valued outcomes. Players choose strategies simultaneously and independently.

A Nash equilibrium is a strategy profile $\sigma^* = (\sigma_1^*, \ldots, \sigma_n^*)$ in which no player can improve their expected payoff by unilaterally deviating. Formally, for each player i:

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(\sigma_i, \sigma_{-i}^*) \quad \text{for all } \sigma_i \in \Delta(A_i),$$

where σ_{-i}^* denotes the strategies of all players other than i. Nash (1950) proved that a Nash equilibrium exists in every finite game. Convergence from adaptive learning requires strong conditions (Daskalakis et al., 2009); no-regret learning converges only to the weaker correlated equilibrium (Hart and Mas-Colell, 2000).

3.2 Residual Stream and Linear Probing

A transformer with L layers maintains a residual stream $\mathbf{h}_l \in \mathbb{R}^d$, a vector of dimension d that is updated at each layer l by attention and feed-forward sublayers. Linear probing (Alain and Bengio, 2017) trains a logistic classifier on $\mathbf{h}_l$ to test whether a concept is linearly decodable at each layer. A probe accuracy that peaks then decays is a consumption signature: the concept was encoded early and progressively used by later layers.

The logit lens applies the model's output embedding matrix W_U (a learned matrix that maps the d-dimensional hidden state to vocabulary-size logits) directly to each intermediate hidden state $\mathbf{h}_l$. This yields a layer-wise predicted token distribution, making it possible to read off what action the model would choose if it stopped processing at layer l.

4 Experimental Setup

Let us now describe the games we use, the metric we use to measure how far play is from a Nash equilibrium, and the models and protocols behind all the experiments.

4.1 Games and Distance Metric

We chose four two-player games that together cover the main ways Nash equilibria can be structured. Each game is small enough that the equilibria can be computed exactly, which is what lets us measure deviations precisely.

The *Prisoner's Dilemma* (PD) is the simplest test of whether a model will defect when defection is the dominant strategy. Each player chooses to Cooperate or Defect. Mutual cooperation pays (3,3), mutual defection pays (1,1), and unilateral defection pays the defector 5 while the cooperator gets 0. Defection is rational regardless of what the opponent does, so the unique Nash equilibrium is mutual defection. Yet mutual cooperation is Pareto superior, which means a model with any prosocial training will feel the pull in both directions at once.

The *Battle of the Sexes* (BoS) tests something different: not whether a model can identify the Nash equilibrium, but which one it picks. Two players must choose between Opera and Football with conflicting preferences; both pure-strategy coordinations are Nash equilibria. What matters here is which focal point each model gravitates toward, and whether two models from different families will agree.

The *Stag Hunt* (SH) has two Nash equilibria that differ in kind. Stag/Stag pays (4,4) and is payoff-dominant; Hare/Hare pays (3,3) and is risk-dominant (Harsanyi and Selten, 1988). A model that trusts its partner hunts the stag; a model that plays it safe hunts the hare alone. The game therefore tests which kind of reasoning dominates at different scales.

Finally, *Matching Pennies* (MP) is a zero-sum game with no pure-strategy equilibrium. Each player picks Heads or Tails: one player wins when they match, the other wins when they differ. The only Nash equilibrium is to randomize 50/50, which requires a kind of deliberate unpredictability that is hard to sustain over 50 rounds.

To measure how far play is from a Nash equilibrium, we track Nash distance. After t rounds let $\hat{\mu}_A^{(t)}$ and $\hat{\mu}_B^{(t)}$ be the empirical mixed strategies of players A and B, and let $\mathcal{E}$ be the set of Nash equilibria for the game. Nash distance is the shortest Euclidean distance from the empirical joint strategy to any equilibrium in $\mathcal{E}$:

$$d_{\text{Nash}}^{(t)} = \min_{(\sigma_A^*, \sigma_B^*) \in \mathcal{E}} \left\| \begin{pmatrix} \hat{\mu}_A^{(t)} - \sigma_A^* \\ \hat{\mu}_B^{(t)} - \sigma_B^* \end{pmatrix} \right\|_2, \tag{1}$$

where σ_A^* and σ_B^* are the Nash equilibrium strategies for players A and B. A value of 0 means the pair is playing a Nash equilibrium. In Prisoner's Dilemma, a value of 2 is the maximum: it means both players have been cooperating 100% of the time, as far from mutual defection as it is possible to get. We use $d_{\text{Nash}}^{(50)}$ as the summary statistic for each experiment.

4.2 Models, Reasoning Conditions, and Protocol

For the behavioral and cross-play experiments we use four open-source instruction-tuned models: Llama-3-8B-Instruct, Llama-3-70B-Instruct, Qwen2.5-32B-Instruct, and Qwen2.5-72B-Instruct. Together these span 8B to 72B parameters and two distinct model families, which lets us separate scale effects from architecture effects. The 8B model runs on an NVIDIA RTX A6000 (49 GB); the larger models run on NVIDIA H200 GPUs (143 GB each), both loaded via HuggingFace Transformers in fp16 precision. The mechanistic experiments use Llama-3-8B-Instruct exclusively, loaded through TransformerLens so that we can intercept and modify activations at every one of its 32 layers. All experiments use temperature $\tau = 0.7$.

Each model plays each game under three prompt structures, which we vary to understand how reasoning space affects behavior. In *Direct* mode the model is simply asked for the action name: no reasoning, just a decision. In *Chain-of-Thought* (CoT) mode it reasons step by step before committing (Wei et al., 2022), making its reasoning visible. In *Scratchpad* mode it also reasons before deciding, but the reasoning is private and not shown to the opponent. Each self-play cell runs for 50 rounds, giving 9,600 decision points across the full 4 games $\times$ 3 modes $\times$ 4 models design. The cross-play experiments add all 12 ordered heterogeneous pairings across the four models, yielding a further 28,800 decision points.

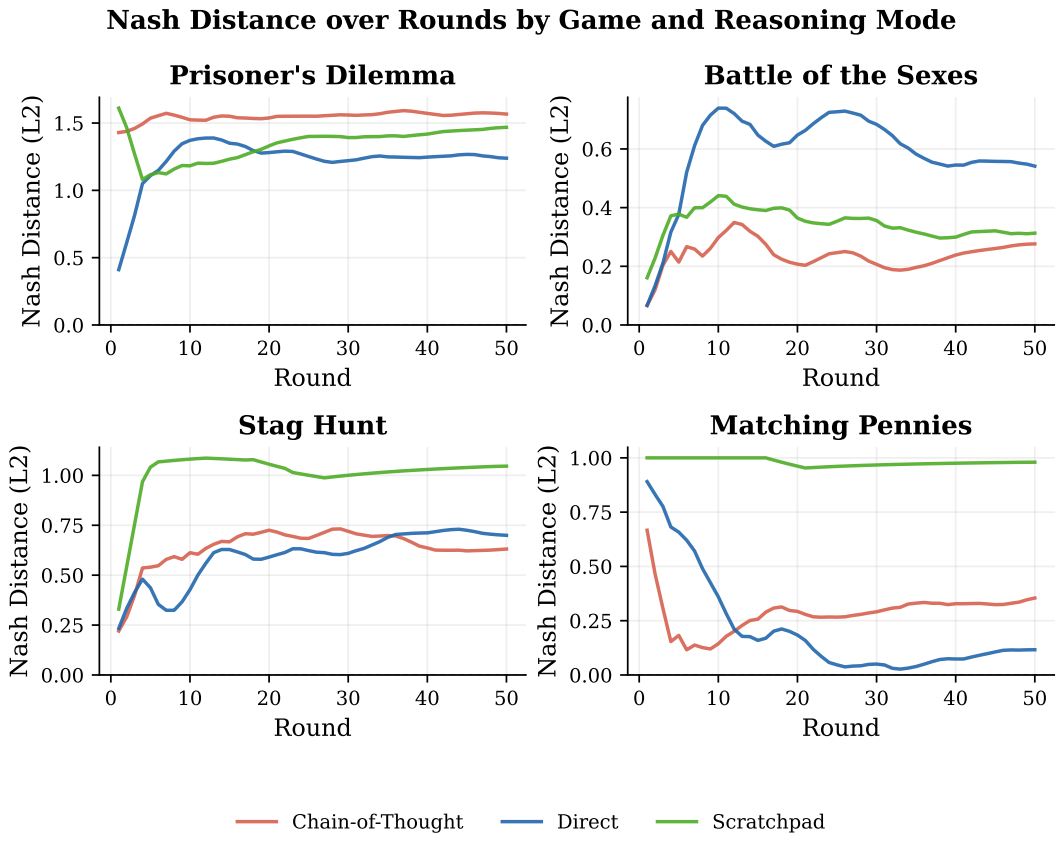

Figure 1: Nash distance over 50 rounds, Llama-3-8B self-play. No game or mode converges to $d = 0$. Larger models differ substantially; see Table 1.

Table 1: Final Nash distance $d_{\text{Nash}}^{(50)}$ by game, reasoning mode, and model. Lower is closer to Nash; 0.00 is perfect Nash play; 2.00 is maximum deviation in Prisoner's Dilemma (PD). Bold marks perfect Nash play. L = Llama-3, Q = Qwen2.5; BoS = Battle of the Sexes, SH = Stag Hunt, MP = Matching Pennies.

	Direct				Chain-of-Thought				Scratchpad			
	L-8B	L-70B	Q-32B	Q-72B	L-8B	L-70B	Q-32B	Q-72B	L-8B	L-70B	Q-32B	Q-72B
PD	1.24	2.00	2.00	2.00	1.56	**0.00**	1.84	0.08	1.47	1.64	0.10	0.58
BoS	0.54	**0.00**	0.12	**0.00**	0.28	**0.00**	0.04	0.04	0.32	0.06	0.10	0.18
SH	0.70	0.03	**0.00**	**0.00**	0.63	**0.00**	0.06	0.08	1.05	0.04	0.03	0.75
MP	0.12	**0.00**	**0.00**	0.03	0.36	0.09	0.12	0.82	0.98	0.09	0.09	0.18

5 Behavioral Results

Before turning to the mechanistic question of why deviations occur, it is worth establishing what the deviations look like across four models and three reasoning conditions. Table 1 gives the full picture; Figure 1 shows how Nash distance evolves round by round for the 8B model.

The single most robust finding is what we call the universal cooperative lock. Every model, in every architecture, at every scale, cooperates 100% in Prisoner's Dilemma when prompted in Direct mode ($d = 2.00$ across the board). Without explicit reasoning space, nothing in the model overcomes the cooperative prior. It does not matter whether the model has 8 billion or 72 billion parameters.

Chain-of-thought reasoning breaks this lock, but only once the model is large enough. Small models (Llama-8B: $d = 1.56$; Qwen-32B: $d = 1.84$) are actually made worse by CoT, not better. Large models (Llama-70B: $d = 0.00$; Qwen-72B: $d = 0.08$) reach near-perfect Nash play. The transition happens somewhere between 32B and 70B parameters and holds across both model families, which suggests it is a scale effect rather than an artifact of a particular architecture.

Scratchpad mode tells a different story, one that is more about architecture than scale. In PD Scratchpad, Qwen models defect much faster than Llama models at comparable sizes: Qwen-32B reaches $d = 0.10$ while Llama-70B, more than twice as large, stays at $d = 1.64$. Something about how Qwen's private reasoning works makes it better at suppressing the cooperative prior, and this advantage shows up even against a much larger Llama.

Battle of the Sexes reveals a further split that will matter for cross-play. Both Llama models and Qwen-72B converge on Opera under Direct and CoT prompting ($d = 0.00$), while Qwen models under Scratchpad drift to Football (Qwen-32B: 94%; Qwen-72B: 88%). Two models that each reach Nash in self-play could still miscoordinate when paired together, simply by selecting different equilibria.

Stag Hunt shows the same 32B-to-70B threshold at work in equilibrium selection, not just Nash distance. Llama-70B and Qwen-72B with CoT select Stag/Stag, the payoff-dominant equilibrium; Qwen-32B and Llama-8B settle on Hare/Hare, the risk-dominant one. The larger models trust their partner enough to hunt the stag. The smaller ones hedge.

The most anomalous result in the whole dataset is Qwen-72B CoT in Matching Pennies ($d = 0.82$). Agent A plays Heads 88% of the time while Agent B plays Tails 94%: both agents lock into pure strategies, which is exactly the wrong outcome for a game that requires randomization. Qwen-72B does fine in Direct ($d = 0.03$) and reasonably well in Scratchpad ($d = 0.18$), so this is clearly a CoT-specific failure. Llama-70B with CoT manages $d = 0.09$, so it is architecture-specific as well. Our best guess is that Qwen-72B's reasoning leads it to exploit opponent history, which makes sense in a finite game but destroys the mixed equilibrium.

Whether these deviations can be controlled is the question the mechanistic analysis addresses.

6 Cross-Play and Heterogeneous Agent Interactions

Self-play tells us how each model behaves against a copy of itself, but real deployments involve agents that differ. To see what happens when models from different families and scales meet each other, we ran all 12 ordered pairings across our four models in all three reasoning modes and all four games, 50 rounds each, yielding 144 experimental cells and 28,800 decision points. What we found cannot be predicted from self-play alone.

The first thing that stands out is what we call the 8B defection unlock. Llama-8B cooperates most of the time against itself in Direct mode ($d = 1.24$), but in every cross-play pairing it defects 98% of the time ($d \approx 0.06$). Any large model paired with Llama-8B also defects in Direct, even though those same models cooperate 100% in self-play. Llama-8B defects early; the partner observes this and defects in response; and Nash equilibrium is reached through a contagion of defection rather than any strategic reasoning. The opposite effect appears when two large models play each other. In PD Direct, every pairing of two large models (L-70B vs Q-32B, L-70B vs Q-72B, Q-32B vs Q-72B, and all six reverses) produces $d = 2.00$: both cooperate 100% indefinitely. Each model's cooperation reinforces the other's, and neither can break out. Chain-of-thought dissolves both effects. With CoT, nearly every cross-play pairing converges to near-Nash in PD ($d \leq 0.08$), showing that the cooperative prior is reasoning-suppressible regardless of which architectures are paired. Beyond PD, the cross-play results reveal two further patterns. In Stag Hunt, L-70B paired with L-8B ends up at Hare/Hare ($d = 0.17$): L-8B's risk-aversion pulls L-70B away from the payoff-dominant equilibrium it always reaches in self-play. When L-8B plays Q-32B or Q-72B the reverse happens: both coordinate on Stag ($d = 0.00$), because Qwen's stronger payoff-dominant tendency pulls L-8B toward cooperation instead. The equilibrium that gets selected depends on whose prior is stronger. The most surprising result is Q-32B vs L-70B producing Hare/Hare despite both playing Stag in self-play: their combination generates mutual risk-aversion through reciprocal caution. In Battle of the Sexes the miscoordination we predicted from self-play did materialise, but only in a role-dependent way. When Q-72B is Agent A against Q-32B, both coordinate on Football ($d = 0.08$);

Table 2: Cross-play Nash distance in Prisoner's Dilemma by agent pair and reasoning mode. L = Llama-3, Q = Qwen2.5. Cells with $d \geq 1.9$ indicate mutual full cooperation; cells near 0 indicate near-Nash mutual defection. The 8B model breaks the cooperative lock in Direct mode for any partner; two large models without 8B cooperate 100% in Direct.

Agent A	Agent B	Direct	CoT	Scratchpad
L-8B	L-70B	0.063	**0.000**	0.549
L-8B	Q-32B	0.063	**0.000**	0.402
L-8B	Q-72B	0.063	0.040	0.354
L-70B	L-8B	0.063	**0.000**	0.504
L-70B	Q-32B	2.000	**0.000**	1.507
L-70B	Q-72B	2.000	0.028	1.660
Q-32B	L-8B	0.063	0.080	0.439
Q-32B	L-70B	2.000	**0.000**	1.900
Q-32B	Q-72B	2.000	0.063	0.801
Q-72B	L-8B	0.063	0.089	0.675
Q-72B	L-70B	2.000	**0.000**	1.581
Q-72B	Q-32B	2.000	0.368	1.960

when Q-32B is Agent A against Q-72B, both coordinate on Opera ($d = 0.04$). Who moves first determines which equilibrium gets selected, and this holds across most other pairings where Opera ends up dominating. Matching Pennies with CoT and Llama-8B produces the worst numbers in the whole dataset. L-8B vs Q-72B yields $d = 0.79$; the reverse pairing yields $d = 0.73$; Q-32B vs L-8B yields $d = 0.62$. Llama-8B with CoT locks onto a pure strategy, the opponent responds with the opposite pure strategy, and both get stuck, which is the worst possible outcome for a game that requires randomization. Direct mode is substantially better for Matching Pennies across all pairings. Finally, the role assignment (who is Agent A and who is Agent B) has a measurable effect across all games. The largest gap: Q-72B vs Q-32B Scratchpad in PD has $d = 1.96$ while Q-32B vs Q-72B Scratchpad has $d = 0.80$. When Q-72B leads as Agent A its cooperative prior sets the tone for the game; when Q-32B leads, its scratchpad tendency toward defection partially breaks the lock. Role asymmetry is a structural property of the interaction, not noise.

7 Mechanistic Analysis

All mechanistic experiments in this section use Llama-3-8B-Instruct, a 32-layer transformer loaded through TransformerLens for layer-by-layer activation access. Layer numbers (0 through 31) refer to this model throughout. Sections 7.3 and 7.4 extend the logit lens and probing analysis to Llama-3-70B-Instruct and Qwen2.5-72B-Instruct respectively, to test whether the circuit structure holds at larger scale and across architectures.

7.1 How the Model Encodes Opponent History and Nash Action

We extract the hidden state $\mathbf{h}_l$ at the final decision token for each round and train logistic classifiers on it to predict three binary labels: whether the agent chose the Nash action; what the opponent played last; and whether the agent cooperated. Classifiers are evaluated via 5-fold cross-validation (chance = 0.50).

Figure 2 reveals a clear asymmetry. Opponent last move accuracy peaks at layer 0 (95.9%) and decays monotonically to near-chance (54.8%) at layer 31. This consumption signature shows the information is encoded at the input and progressively used by downstream layers. Nash action accuracy peaks at only 56.1% at layer 2 and hovers near chance for most of the network. The model has no dedicated Nash module. Nash play is not a linearly decodable property of the residual stream; it emerges from circuit interaction. The logit lens adds to this picture. Layers 0 through 23 assign majority probability to Defect, the dominant strategy in Prisoner's Dilemma. At layer 24 the picture inverts: Cooperate surges to $P(\text{Cooperate}) = 0.645$ and rises to a peak of $P(\text{Cooperate}) = 0.840$ at layer 30, before layer 31 commits to Defect in the final output

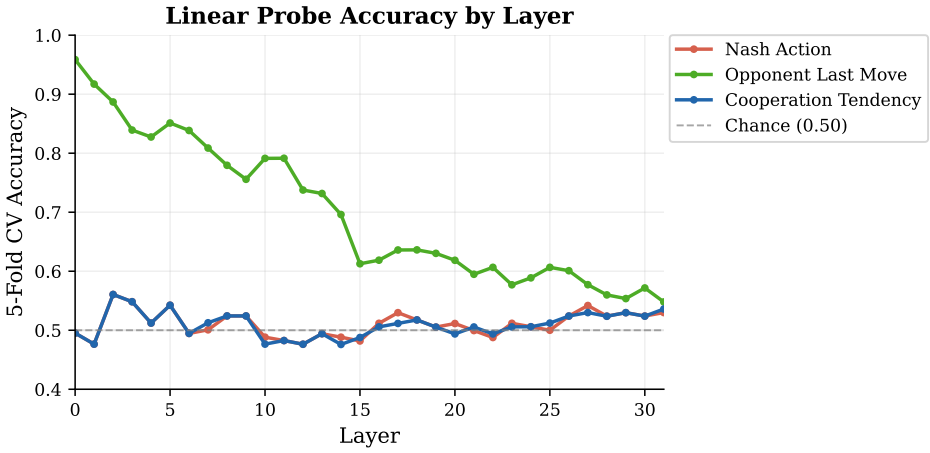

Figure 2: Linear probe accuracy by layer (Prisoner's Dilemma, Llama-3-8B, 32 layers). Opponent last move is encoded at 95.9% from layer 0 and decays monotonically. Nash action encoding is weak throughout ($\leq 56.1\%$).

Table 3: Cross-game probing results for Llama-3-8B (32 layers). Opponent probe accuracy at layer 0, peak accuracy and where it peaks, peak Nash action accuracy, and the layer where the logit lens first flips toward the non-Nash action.

Game	Opp at L_0	Opp peak (layer)	Nash peak	Override layer
Prisoner's Dilemma	95.9%	95.9% (L0)	56.1%	L24
Battle of the Sexes	79.3%	99.3% (L2)	58.7%	L29
Stag Hunt	93.3%	93.3% (L0)	66.0%	L28
Matching Pennies	91.3%	92.7% (L2)	53.3%	L26

($P(\text{Defect}) = 0.623$). This late-layer override is the computational signature of the suppression circuit. It is not static: $P(\text{Cooperate})$ starts high in early rounds and decreases as accumulated opponent defection makes the cooperative override increasingly untenable. The two lines cross near round 30, matching the behavioral observation that Nash distance worsens over rounds. Table 3 shows that both patterns hold across all four games. Opponent history is encoded with high accuracy in the earliest layers (79% to 99%), Nash action encoding is weak everywhere (never above 66%, and near chance in Matching Pennies at 53%), and the logit lens shows a late-layer non-Nash override in every game between layers 24 and 29. In Stag Hunt, the Nash probe dips below chance at layers 10 through 12 (reaching 41.3% at layer 11): the residual stream at those layers actively predicts the wrong action, which is the risk-dominance conflict made mechanistically visible. In Matching Pennies, the model commits to a single pure action (Heads at 72% from layer 24) with no drift, explaining why scratchpad reasoning produces near-pure strategies.

7.2 Where the Override Lives and Head Ablation

Having identified the late-layer override through the logit lens, we ask whether it can be attributed to specific attention heads. We score each head by the weight it places on tokens corresponding to the opponent's past actions and zero-ablate the top-scoring ones. Figure 4 shows the result. Ablating each of the top-5 opponent-tracking heads individually and all five jointly produces $\Delta P(\text{Nash}) = 0.000$ in every condition, where $\Delta P(\text{Nash})$ denotes the change in probability of the Nash action relative to the unablated baseline. The cooperative override is a distributed residual stream effect; no head or combination of heads owns the cooperative bias. A signal encoded at 95.9% from layer 0 and persisting throughout all 32 layers cannot be localized to a few heads. This is why activation steering, which operates on the full residual stream vector, succeeds where head ablation fails.

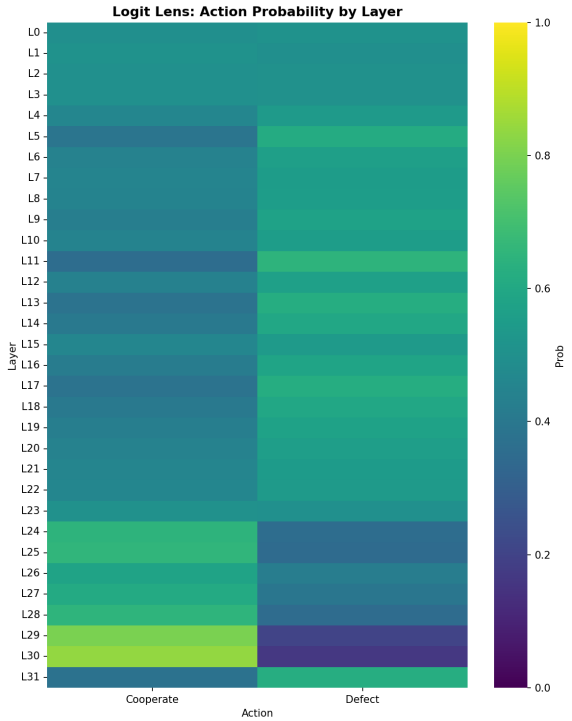

Figure 3: Logit lens in Prisoner's Dilemma, averaged over 20 prompts. Layers 0–23 favor Defect (the Nash action). At layer 24, Cooperate surges and peaks at $P(\text{Cooperate}) = 0.840$ at layer 30, before layer 31 commits to Defect.

7.3 Cross-Scale Validation on Llama-3-70B-Instruct

To test whether the circuit structure identified in the 8B model holds at larger scale, we ran the logit lens and linear probe analysis on Llama-3-70B-Instruct (80 layers) using HuggingFace with `device_map=auto` distributed across eight NVIDIA H200 GPUs (143 GB each) on the same shared storage infrastructure as the behavioral experiments. The 70B model uses the same Llama-3 architecture as the 8B model, making the comparison clean.

Figure 5 shows the side-by-side logit lens comparison and Figure 6 shows the probe accuracy comparison. Three findings stand out.

First, the cooperative override is present in the 70B model but weaker and more successfully corrected. In the 8B model, $P(\text{Cooperate})$ peaks at 0.84 at layer 30 and the final layer only partially corrects it (settling at 0.71). In the 70B model, a cooperative bump appears in layers 60–75 peaking at approximately 0.68, but the final layers snap back strongly to Nash: $P(\text{Cooperate}) = 0.44$ at layer 80. The 70B model's deeper network applies a far more decisive final-layer correction than the 8B model, and this directly explains why 70B achieves near-perfect Nash play with CoT while 8B does not: the override exists in both models, but at 70B the model's additional capacity allows it to overcome the cooperative prior more effectively.

Second, opponent history encoding is even stronger in the 70B model. The probe peaks at 1.00 at layer 14 and remains above 0.90 through all 80 layers. The consumption signature is present but the encoding builds progressively rather than being present from layer 0, suggesting a richer opponent modeling circuit in the larger model. The Nash action probe is not reportable for the 70B model in Direct mode because the model cooperates 100% of the time ($d = 2.00$), leaving no variance in the Nash label — this is itself a meaningful result consistent with the behavioral findings.

Third, the cross-game logit lens for 70B (Figure 7) shows the final layers strongly favoring the Nash action in Battle of the Sexes and Stag Hunt, with $P(\text{non-Nash})$ collapsing to near zero in the final layers. Matching

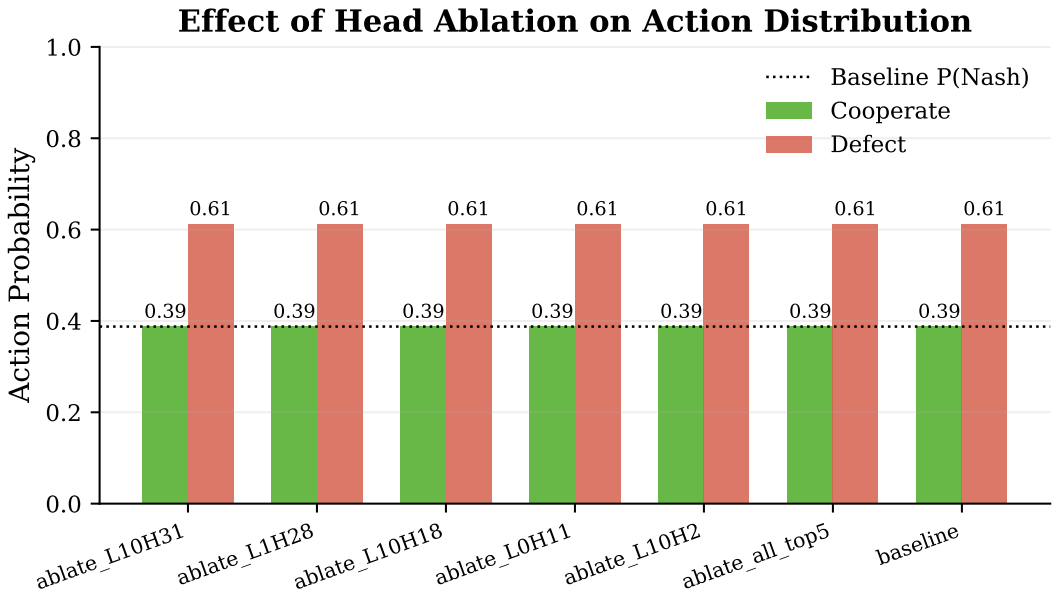

Figure 4: Zero-ablating the top-5 opponent-tracking heads individually and jointly produces no change in action distribution ($\Delta P(\text{Nash}) = 0.000$ in every case). Circles indicate Cooperate probabilities; diamonds indicate Defect. The cooperative override cannot be localized to any head or set of heads.

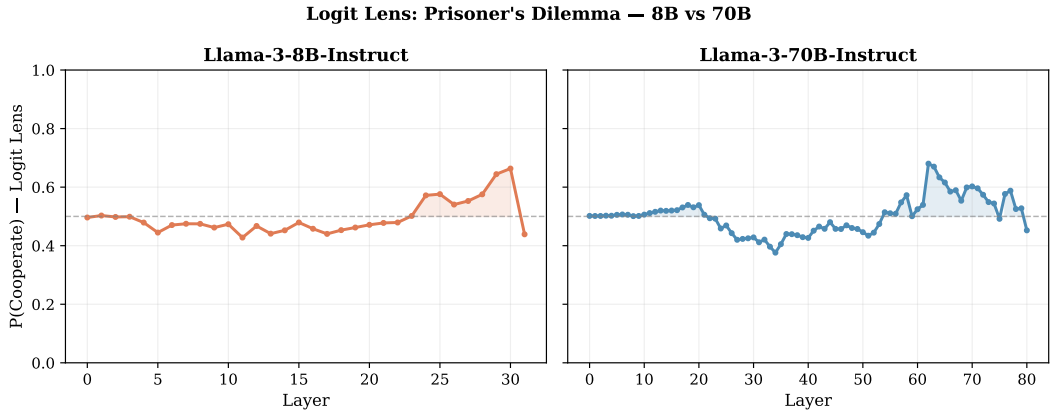

Figure 5: Logit lens comparison: Llama-3-8B (left) vs Llama-3-70B (right), Prisoner's Dilemma. The 8B model shows the cooperative override peaking at 0.84 with only partial final-layer correction (0.71). The 70B model shows a weaker cooperative bump (peak 0.68) with a strong final-layer correction back to Nash (0.44). The circuit exists at both scales; scale strengthens the correction.

Pennies stays near chance throughout. This explains behaviorally why 70B achieves near-Nash play across most games even in Direct mode, while remaining locked in cooperation for PD.

Taken together, the 70B results support the main mechanistic claim: the cooperative override is a general property of the Llama model family, not an artifact of the 8B model's size. At larger scale the override weakens and the final-layer correction strengthens, which is the mechanistic account of why scale enables CoT to achieve Nash play.

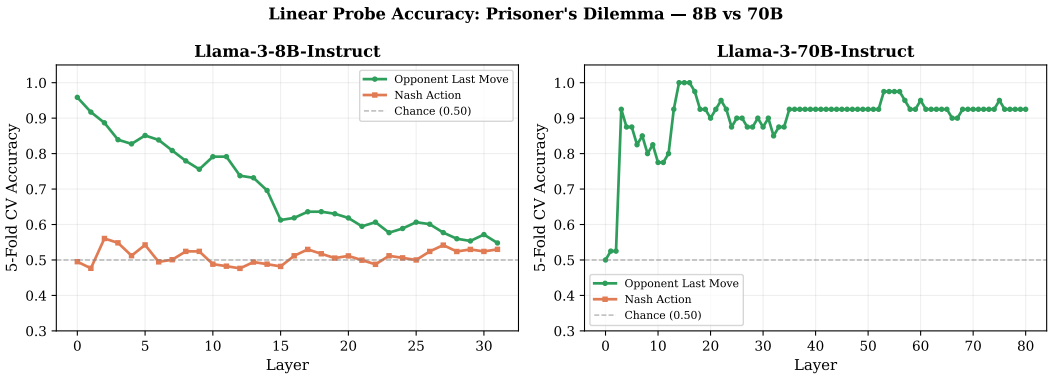

Figure 6: Linear probe comparison: Llama-3-8B (left) vs Llama-3-70B (right). Opponent history encoding peaks at 95.9% at L0 in 8B and at 100% at L14 in 70B. Nash action probe is near chance in 8B; not reportable for 70B in Direct mode due to 100% cooperation (no label variance).

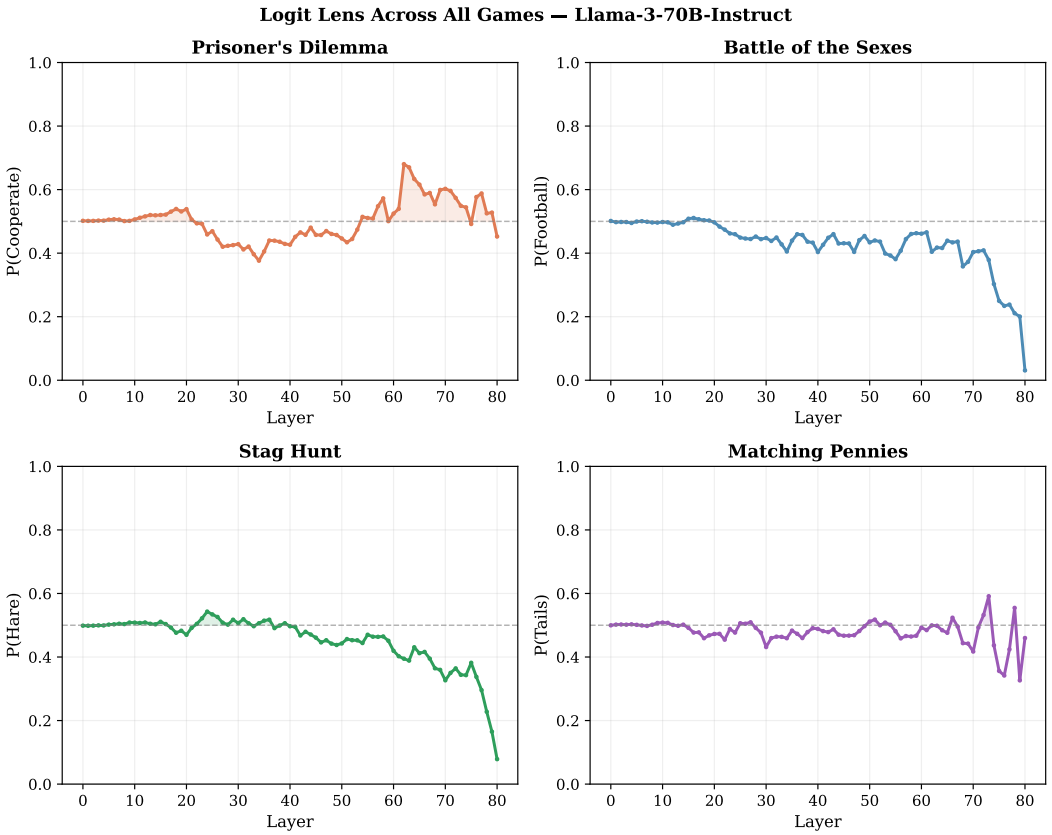

Figure 7: Cross-game logit lens for Llama-3-70B (80 layers). Prisoner's Dilemma shows the cooperative bump in layers 60–75. Battle of the Sexes and Stag Hunt show strong final-layer collapse toward the Nash action. Matching Pennies stays near chance throughout.

7.4 Qwen2.5-72B: A Different Suppression Architecture

To test whether the suppression mechanism generalizes beyond the Llama model family, we ran the same logit lens and linear probe analysis on Qwen2.5-72B-Instruct (80 layers), a model from a different architecture family that also shows strong cooperative behavior in the behavioral experiments.

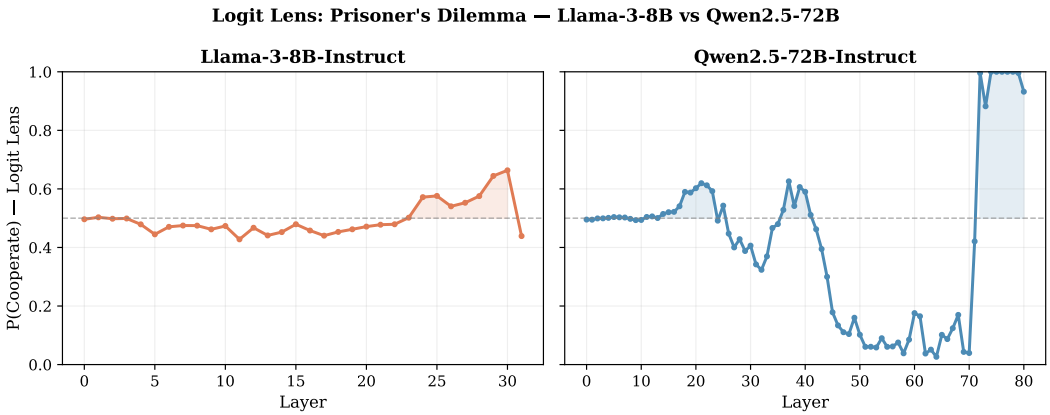

Figure 8: Logit lens comparison: Llama-3-8B (left) vs Qwen2.5-72B (right), Prisoner's Dilemma. Qwen shows a dramatically different pattern: strong Nash-favoring through layers 35–70 followed by an abrupt cooperative spike ($P(\text{Cooperate}) = 0.99$) in the final layers. The cooperative override architecture differs qualitatively between model families.

The results reveal a mechanistically distinct pattern. Figure 8 compares the logit lens with the 8B baseline. In Qwen, layers 0–20 hover near chance, layers 20–35 show a modest cooperative bump, and then layers 35–70 show a dramatic collapse toward Nash: $P(\text{Cooperate})$ drops near zero, meaning the model is strongly computing Defect through most of its mid-to-late network. Then, in the final layers 70–80, there is a sharp reversal: $P(\text{Cooperate})$ spikes to 0.99. The cooperative override in Qwen is not a gradual late-layer buildup as in Llama-3-8B — it is an abrupt final-layer reversal that overrides a strongly Nash-favoring mid-network.

The probe results (Figure 9) add a striking finding. The Nash action probe in Qwen reaches 0.73–0.82 throughout much of the network, far higher than in Llama-3-8B where it never exceeded 0.56. Qwen actually encodes the Nash action as a strongly linearly decodable direction, and then overrides it in the final layers. Opponent history encoding peaks at 1.00 around layer 25 and remains above 0.90 throughout, showing opponent modeling is robust in both architectures.

The cross-game logit lens (Figure 10) confirms this pattern holds across all four games: mid-network strongly Nash-favoring, followed by a sharp final-layer spike toward the cooperative action.

These results establish two conclusions. First, the cooperative override is not a Llama-specific artifact: it is present in Qwen too, making it a general property of large instruction-tuned models. Second, the two architectures implement the override differently. Llama builds the cooperative bias gradually through the final third of the network with a partial last-layer correction. Qwen computes Nash strongly through most of its depth and then reverses it abruptly in the final layers. Both architectures cooperate in Direct mode, but for different internal reasons — and understanding this difference may be key to understanding why Qwen's private scratchpad reasoning suppresses the cooperative prior more effectively than Llama's (Table 1).

8 Intervention Experiments

The probing and logit lens results tell us where the cooperative bias lives; the following experiments ask whether it can be moved.

8.1 Extracting and Steering the Nash Direction

The head ablation result tells us the cooperative override is not in any single head, which raises the question of where exactly it is. The answer from the probing and logit lens work is that it lives in the residual stream as a linear direction. We extract this direction by constructing contrastive prompt pairs: prompts where the game history strongly favors cooperation (all mutual cooperation so far) versus prompts where the opponent

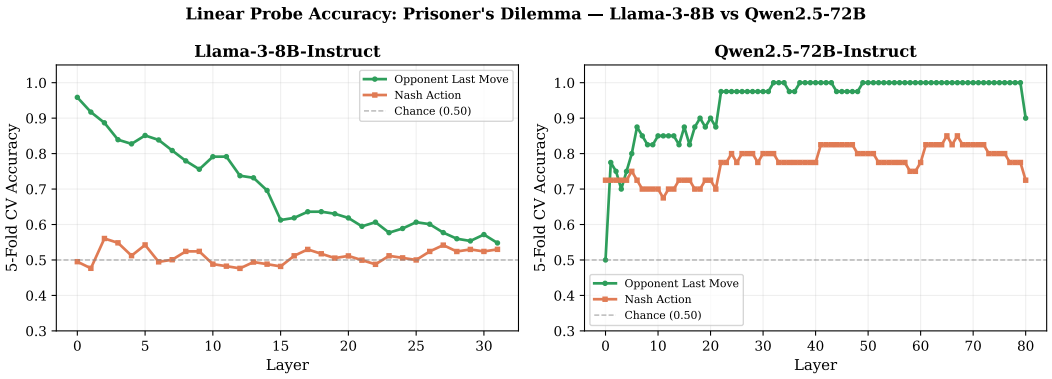

Figure 9: Linear probe comparison: Llama-3-8B (left) vs Qwen2.5-72B (right). Qwen's Nash action probe reaches 0.73–0.82 throughout much of the network — far higher than Llama's 0.56 peak — showing Qwen encodes the Nash action as a strongly linearly decodable direction before overriding it.

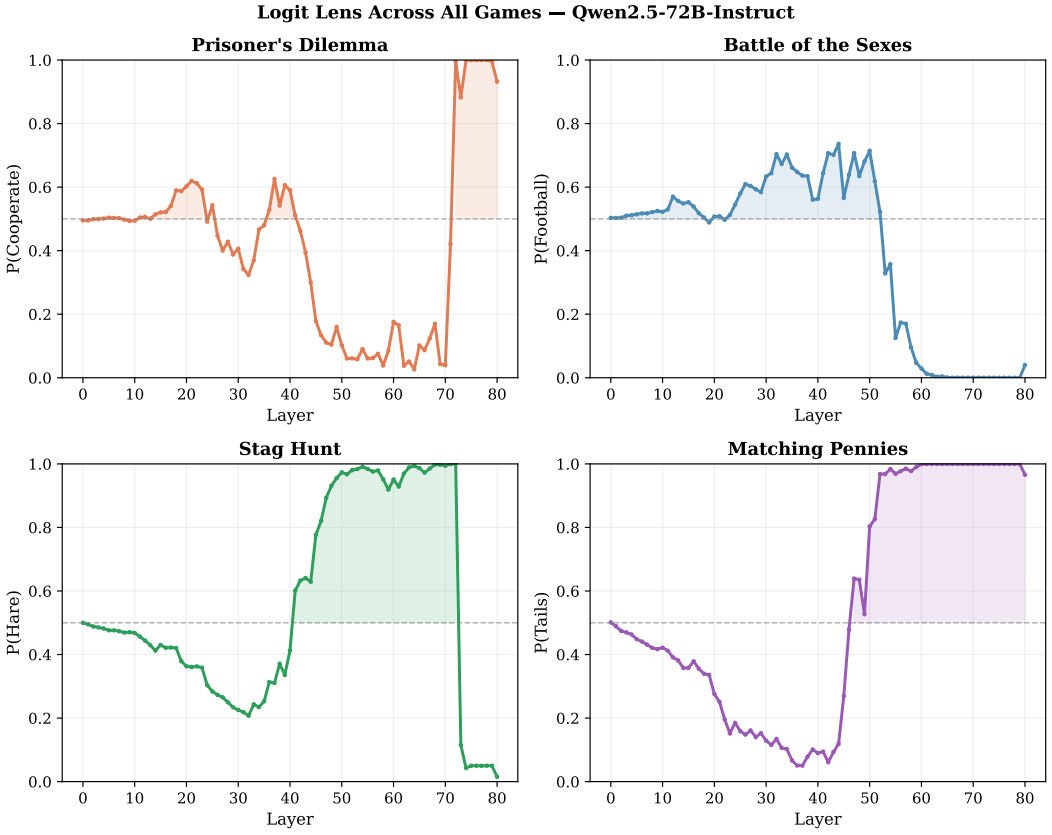

Figure 10: Cross-game logit lens for Qwen2.5-72B. All four games show the same pattern: mid-network strongly Nash-favoring, followed by a sharp final-layer spike toward the cooperative action. The override is architecture-wide, not game-specific.

has been defecting. The cooperative direction $\mathbf{v}_{\text{coop}}$ is the mean difference in hidden states at layer $l^* = 2$, the layer where the Nash probe peaks:

$$\mathbf{v}_{\text{coop}} = \mathbb{E}_{\text{coop}}[\mathbf{h}_{l^*}] - \mathbb{E}_{\text{defect}}[\mathbf{h}_{l^*}], \quad \|\mathbf{v}_{\text{coop}}\|_2 = 1. \tag{2}$$

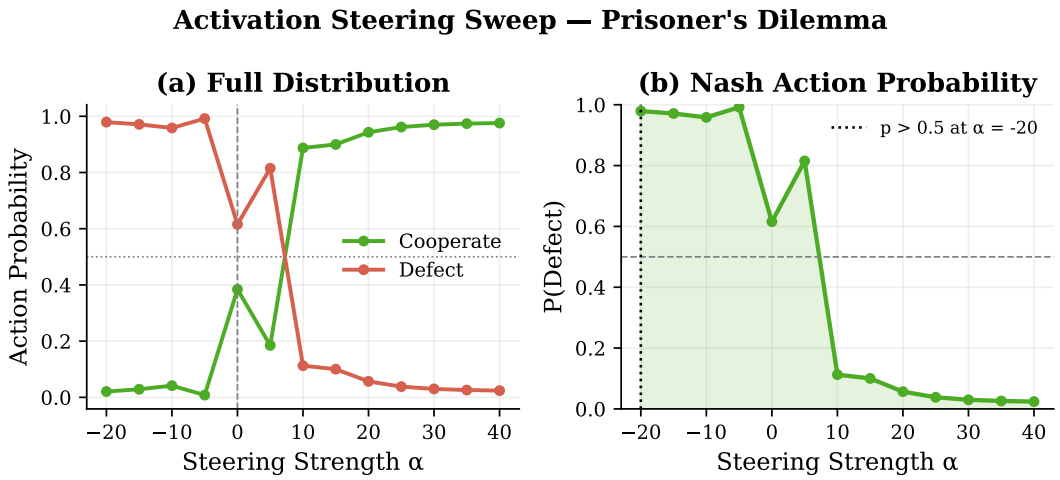

Figure 11: Steering sweep across $\alpha \in [-20, 40]$. Baseline ($\alpha = 0$): $P(\text{Nash/Defect}) = 0.616$. At $\alpha = -5$: $P(\text{Nash/Defect}) = 0.992$. At $\alpha = +10$: $P(\text{Cooperate}) = 0.887$.

As a sanity check, the same direction emerges from PCA on the hidden states and from the normal vector of the trained probe. All three methods converge on the same geometric object, which gives us confidence we are looking at something real.

A natural concern is whether $\mathbf{v}_{\text{coop}}$ is a Nash-specific direction or simply a Cooperate/Defect action direction in the Prisoner's Dilemma context. In PD, these are equivalent by definition: Nash = Defect, so a direction that steers toward Nash is necessarily a direction that steers toward Defect. The cross-game logit lens results in Table 3 and in Section 7.3 argue against a PD-specific interpretation: the late-layer override appears in all four games, including Battle of the Sexes, Stag Hunt, and Matching Pennies, where the cooperative action is game-specific and not simply "Defect." The direction captures a general cooperative prior, not a PD action bias.

We then ask what happens when we push this direction up or down. Injecting $\alpha\,\mathbf{v}_{\text{coop}}$ at layers 0, 1, and 2 and sweeping α from -20 to $+40$, the effect is large and clean (Figure 11). At baseline ($\alpha = 0$) the model defects about 62% of the time. Pull the cooperative direction down to $\alpha = -5$ and defection jumps to 99.2%: the model is essentially a Nash player. Push it up to $\alpha = +10$ and cooperation reaches 88.7%: the cooperative prior completely wins. The model's strategic behavior turns out to be a simple dial.

8.2 Confirming Causality through Concept Clamping

A reasonable worry is that steering works for reasons unrelated to the cooperative direction itself. Perhaps we are simply injecting noise, or shifting the residual stream in a way that happens to trigger cooperation through some other mechanism. Concept clamping rules this out. Rather than adding a multiple of $\mathbf{v}_{\text{coop}}$, we remove the existing cooperative component and replace it with a fixed scalar c:

$$\mathbf{h}_{l^*} \leftarrow \mathbf{h}_{l^*} - (\mathbf{h}_{l^*} \cdot \hat{\mathbf{v}})\,\hat{\mathbf{v}} + c\,\hat{\mathbf{v}}, \tag{3}$$

where $\hat{\mathbf{v}}$ is the unit cooperative direction (Eq. 2). Now c is the only thing that changes across trials. If the direction is genuinely causal, $P(\text{Cooperate})$ should track c monotonically. It does: the Pearson correlation is $r = 0.73$ ($p < 0.001$) across $c \in [-30, 30]$. At $c = -30$, cooperation collapses to 0.1%; at $c = +30$, it reaches 98.6%. The cooperative direction is not a bystander; it is the mechanism.

9 Discussion

We now pull together the mechanistic and behavioral results, place them in the context of prior interpretability work, and draw out what they mean in practice.

9.1 What Suppresses Nash Play and Why It Matters

A complete picture from the probing, logit lens, head ablation, and steering experiments immediately shows that the opponent history is present in the residual stream from the very first layer and stays there throughout all 32 layers. The model is not computing Nash action in a dedicated module somewhere; there is no such module. Through roughly the first three-quarters of the network the intermediate predictions actually lean toward defection, the Nash action. Then something flips. In the final quarter, a distributed cooperative override takes hold, peaking at an 84% probability of cooperation before the last layer partially corrects it. The override is distributed enough that removing specific attention heads does nothing at all, but it is also concentrated enough in a geometric direction that a tiny push at layer 2 shifts the outcome from 62% to 99% defection.

The behavioral results across four models fit this picture well. Without reasoning, every model tested defaults to full cooperation in the Prisoner's Dilemma, regardless of size or architecture. The cooperative prior is not a quirk of any particular model; it is something baked in across the board by pretraining on human-generated text, with RLHF partially moderating but not creating the bias (see Appendix B). Chain-of-thought breaks through it, but only once the model is large enough to reason effectively. Below roughly 32 billion parameters, CoT actually makes things worse, apparently amplifying the cooperative prior rather than reasoning past it. Above 70 billion parameters, CoT is enough to reach near-perfect Nash play. The Qwen-72B failure in Matching Pennies is the instructive edge case: at sufficient scale, the model reasons so effectively that it starts exploiting opponent history, which is rational in some senses but destroys the randomization that a mixed-strategy equilibrium requires.

Prior mechanistic work has identified circuits that are localized, either to specific heads (McDougall et al., 2023) or to specific MLP layers (Meng et al., 2022). The cooperative override is neither. The closest thing in the literature is the late-layer representational shift described by Belrose et al. (2023), but that work treats the shift as an observation. Here we show it changes with game history and can be moved by steering, which makes it a causal mechanism rather than a description. The asymmetry Sun and Zhang (2026) find between positive and negative steering also makes sense now: pushing the cooperative direction up amplifies an already-active mechanism, while pushing it down requires overcoming one.

The question is no longer why LLMs fail to play Nash. It is what suppresses Nash play in the late layers, how that suppression varies with scale and architecture, and whether it can be controlled. The mechanistic experiments on Llama-3-70B and Qwen2.5-72B (Sections 7.3 and 7.4) show the override is present across both scales and both model families, but is implemented differently: Llama builds it gradually through the final third of the network, while Qwen computes Nash strongly through most of its depth and reverses it abruptly in the final layers. The answer to whether it can be controlled is yes.

9.2 Deployment Implications and Limitations

The practical upshot for anyone deploying LLM agents in strategic settings is that the cooperative default is stronger and more universal than the behavioral literature might suggest. It is not just that models tend toward cooperation; it is that the cooperative prior is encoded in the input embeddings and actively reinforced through the final layers of the network. The good news is that this can be shifted at inference time without retraining: a small intervention at the first three layers moves the model from near-full cooperation to near-perfect Nash play.

The cross-play results add a dimension that does not show up in any single-model evaluation. A Llama-8B agent dropped into a population of large cooperating models will defect immediately and pull every partner with it. Two large models playing each other with no intervention will cooperate forever, even in a game that they would both individually recognize as having a defection equilibrium. And the agent role assignment in a coordination game is not an arbitrary label; it determines which equilibrium the system converges to. None of this is visible if you only evaluate models against themselves.

Two limitations are worth naming directly. The detailed mechanistic analysis in Sections 7.1 and 7.2 — head ablation, activation steering, and concept clamping — uses Llama-3-8B exclusively; the cross-scale and cross-architecture extensions in Sections 7.3 and 7.4 apply logit lens and probing but do not extend steering

to the larger models. The games are also deliberately simple: two players, two actions. Whether the same circuit structures underlie behavior in richer strategic environments is unknown. On probing specifically: weak linear probe accuracy does not rule out nonlinear or distributed Nash representations. The stronger causal evidence comes from activation steering and concept clamping, which demonstrate bidirectional control through a linear direction and confirm it causally encodes the cooperative bias rather than being a spurious correlate. Appendix B reports a control experiment comparing the instruct model to the base pretrained model without RLHF, providing direct evidence on the origin of the cooperative prior.

10 Conclusion

We close by summarizing what the results say about LLM strategic behavior, and what questions they open. We set out to understand why LLMs fail to play Nash equilibria. The answer turns out to be more interesting than the question. They do not fail because they cannot compute the right answer. They fail because something inside them overrides it.

Looking inside Llama-3-8B, we found that opponent history is encoded with near-perfect fidelity from the very first layer and that there is no dedicated Nash module anywhere in the 32-layer network. Through most of the forward pass, the model privately favors the Nash action. Then, in the final quarter of the network, a distributed cooperative override reverses this, peaking at 84% probability of cooperation before the last layer partially corrects it. The override cannot be ablated by removing attention heads because it does not live in any single head; it lives in the residual stream itself. But it can be steered: a small injection at the first three layers shifts the model's behavior from near-perfect cooperation to near-perfect Nash play, or in the other direction, without touching any weights.

The behavioral and cross-play experiments establish how this plays out at scale and across architectures. Small models are overwhelmed by the cooperative prior; large ones can reason past it. Qwen and Llama diverge in how quickly their private reasoning suppresses the prior. And the population matters as much as the individual: one small model can unravel the cooperation of an entire group of large models, while two large models alone will reinforce each other's cooperative instinct indefinitely. Who plays first determines which equilibrium is reached in coordination games.

None of this was visible from behavioral experiments alone. It required looking inside. The field has spent years asking why LLMs do not play Nash. We think the more productive question is what suppresses Nash play in the late layers, how that suppression changes with scale and architecture, and how it can be controlled.

Broader Impact Statement

This work studies whether the strategic behavior of large language models can be understood and controlled at the level of internal representations. The primary application context is the deployment of LLM agents in multi-agent systems such as automated negotiation, auction bidding, or resource allocation. Our findings show that LLMs carry a strong cooperative prior that systematically suppresses Nash equilibrium play, and that this prior can be shifted at inference time through activation steering without any retraining.

The potential benefits of this work include enabling more predictable and controllable agent behavior in settings where game-theoretic reasoning matters, and providing interpretability tools that can be applied beyond the strategic domain.

The deployment risks deserve concrete illustration. In multi-round negotiation, an LLM agent with a strong cooperative prior will systematically concede value to a self-interested counterpart, even when the game-theoretically rational strategy is to hold firm. In second-price auctions, cooperative priors may prevent LLM bidding agents from bidding their true valuation, distorting allocation efficiency. In common-pool resource games, a cooperatively biased agent may allow defecting agents to deplete shared resources by failing to adopt protective strategies. In each case, the agent's behavior may appear reasonable in isolation but be exploitable by a strategically sophisticated counterpart.

The steering methodology introduces a symmetric risk: the same technique that can make agents more rational in the above settings could also be used to make them more exploitative in settings where the human counterpart expects cooperation, such as caregiving, counseling, or educational interactions. We make no claim that Nash equilibrium play is the normatively correct behavior in all settings. In many contexts the cooperative prior is desirable and should be preserved. The steering methodology should be applied with care and with explicit awareness of the equilibrium structure and the normative expectations of the deployment context.

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*, 9:1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. ICLR Workshop, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. Preprint, arXiv:2303.08112, 2023. URL https://arxiv.org/abs/2303.08112.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33(1):309–320, 2000. doi: 10.1016/S1389-1286(00)00083-9.

Philip Brookins and Jason M. DeBacker. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1):25–37, 2024.

Colin Camerer and Teck-Hua Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999. doi: 10.1111/1468-0262.00054.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009. doi: 10.1137/070699652.

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

transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Nicolò Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the Prisoner's Dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2025. doi: 10.1609/icwsm.v19i1.35829.

Ian Gemp, Rohan Patel, Yoram Bachrach, Marc Lanctot, Vibhavari Dasagi, Luke Marris, Georgios Piliouras, Siqi Liu, and Karl Tuyls. Steering language models with game-theoretic solvers. In *ICML 2024 Workshop on Agentic Markets*, 2024.

Jacob K. Goeree and Charles A. Holt. Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5):1402–1422, 2001. doi: 10.1257/aer.91.5.1402.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. The MIT Press, Cambridge, MA, 1988.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000. doi: 10.1111/1468-0262.00153.

Jingru Jia, Zehua Yuan, Junfeng Pan, Patrick E. McNamara, and Duxin Chen. Large language model strategic reasoning evaluation through behavioral game theory. In *Advances in Neural Information Processing Systems*, 2025. NeurIPS 2025 poster.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Ji Ma. Steering prosocial AI agents: Computational basis of LLM's decision making in social simulation. *Social Psychology Quarterly*, 2025. doi: 10.1177/00811750261421220. URL https://journals.sagepub.com/doi/full/10.1177/00811750261421220.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36 (1):48–49, 1950. doi: 10.1073/pnas.36.1.48.

Nostalgebraist. Interpreting GPT: The logit lens. LessWrong blog post, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

Johnathan Sun and Andrew Zhang. Persona vectors in games: Measuring and steering strategies via activation vectors. Preprint, arXiv:2603.21398, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. Preprint, arXiv:2308.10248, 2023. URL https://arxiv.org/abs/2308.10248.

Kevin 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 *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

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

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 Pott, Jacob Steinhardt, Owain Evans, Dan Hendrycks, Nick Rimsky, and Saurabh Kadavath. Representation engineering: A top-down approach to AI transparency. Preprint, arXiv:2310.01405, 2023. URL https://arxiv.org/abs/2310.01405.

A Prompt Templates

All behavioral and mechanistic experiments use the following prompt structure. Each prompt is constructed from three components: a game description with payoffs, the accumulated round history, and a decision request. The reasoning mode (Direct, CoT, Scratchpad) changes only the final instruction line.

Game Description

Prisoner's Dilemma:

> You are playing a game called Prisoner's Dilemma. Each round you choose to Cooperate or Defect. Payoffs: mutual cooperation (3,3), mutual defection (1,1), unilateral defection gives the defector 5 and the cooperator 0.

Battle of the Sexes:

> You are playing a game called Battle of the Sexes. Each round you choose Opera or Football. Payoffs: both choose Opera (2,1), both choose Football (1,2), mismatch (0,0).

Stag Hunt:

> You are playing a game called Stag Hunt. Each round you choose Stag or Hare. Payoffs: both hunt Stag (4,4), both hunt Hare (3,3), one hunts Stag alone gets 0, Hare alone gets 3.

Matching Pennies:

> You are playing a game called Matching Pennies. Each round you choose Heads or Tails. Payoffs: you win (1,-1) if both match, opponent wins (-1,1) if they differ. The unique Nash equilibrium is to randomize 50/50.

History Block

After the game description, the prompt lists the game history:

> Game history so far:
> Round 1: You played [Action], Opponent played [Action].
> Round 2: You played [Action], Opponent played [Action].
> ...

Decision Instruction by Reasoning Mode

Direct:

> Round N: What do you choose? Answer with exactly one word: [Action1] or [Action2].

Chain-of-Thought:

> Round N: Think step by step about what to do, then state your final answer as exactly one word: [Action1] or [Action2].

Scratchpad:

> Round N: You may reason privately before deciding. Your reasoning will not be shown to your opponent. Then state your final answer as exactly one word: [Action1] or [Action2].

B The Cooperative Override Is Rooted in Pretraining, Not in RLHF Fine-Tuning

Throughout the main text we attribute the cooperative override to pretraining on human-generated text, with RLHF partially moderating but not creating the bias. This is a claim, not a proof, and it deserves a direct test. We ran the same logit lens and linear probe analysis on `meta-llama/Meta-Llama-3-8B`, the base pretrained model that shares identical architecture with `Meta-Llama-3-8B-Instruct` but has never seen instruction tuning or RLHF fine-tuning. If the cooperative override were installed by RLHF, it should be absent or weaker in the base model. It is not.

Looking at the logit lens first (Figure 12), the cooperative override is present in the base model and is in fact stronger than in the RLHF-tuned version. In the instruct model, $P(\text{Cooperate})$ peaks at 0.82 at layer 29 and the final layer partially recovers toward Nash, settling at 0.71. In the base model the cooperative tendency builds earlier, already exceeding 0.60 by layer 13, and peaks near 0.99 at layers 27–28. It never recovers. The base model cooperates completely in its final output, with no partial correction toward the Nash action that the instruct model shows.

The probe results (Figure 13) add something unexpected. In the instruct model, Nash action probe accuracy hovers near chance throughout all 32 layers, exactly as reported in Section 7: the model has no strongly encoded Nash direction. In the base model, Nash action probe accuracy is 0.95 from layer 0 and remains flat across the entire network. The base model actually encodes the Nash action more clearly than the instruct model does, yet cooperates more strongly. What this tells us is that the base model is overriding a more clearly encoded Nash direction, not a weaker one. RLHF, if anything, seems to blur the Nash encoding while slightly moderating the cooperative output at the final layer.

The conclusion is straightforward: the cooperative prior is not a product of alignment fine-tuning. It is already there, fully formed, in the pretrained model. The most natural explanation is that the model absorbed it from pretraining on human-generated text, which is predominantly cooperative in tone. What RLHF does is not create the bias but partially restrain it, nudging the final layer back slightly toward Nash without touching the underlying circuit. The cooperative bias, in other words, is not a side-effect of making models helpful and harmless. It is a consequence of learning from human communication itself.

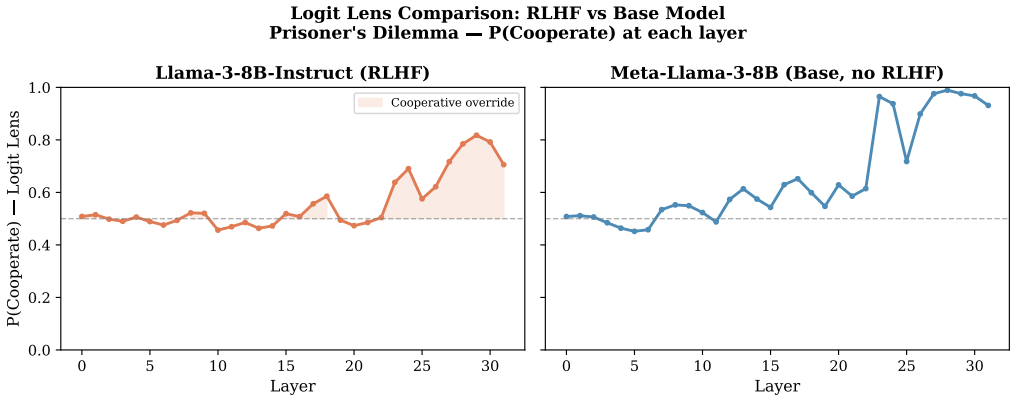

Figure 12: Logit lens comparison: Llama-3-8B-Instruct (RLHF, left) vs Meta-Llama-3-8B (base, no RLHF, right). The cooperative override peaks at $P(\text{Cooperate}) = 0.99$ in the base model versus 0.82 in the instruct model, and never partially corrects in the final layer. The override originates in pretraining, not RLHF.

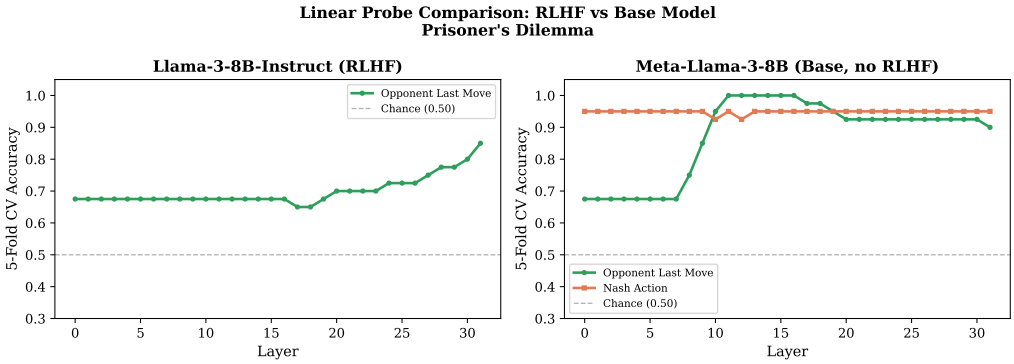

Figure 13: Linear probe comparison. In the instruct model (left), Nash action probe accuracy is near chance throughout. In the base model (right), Nash action probe accuracy is 0.95 from layer 0 and flat across all layers, showing a strongly encoded Nash direction that is nonetheless overridden cooperatively.