Steering Strategic Reasoning in Chain-of-Thought Models: Structured Geometry, Behavioral Shifts, and Interpretive Limits

author names withheld

Under Review for NExT-Game 2026

Abstract

Reasoning-tuned large language models produce extended chains of thought containing identifiable strategic behaviors such as opponent modeling, equilibrium identification, and payoff analysis, yet whether these correspond to structured representations in activation space remains open. We extract difference-of-means (DoM) steering vectors for ten reasoning categories from two DeepSeek-R1-Distill models (Qwen-14B and Llama-8B) using a multi-label annotation scheme across 300 game-theoretic and control tasks. We find that social reasoning and procedural deduction directions are robustly anti-aligned at every layer in both models (cosine -0.71 and -0.61; permutation $p < 0.001$). Steering along the opponent-modeling direction produces specific, monotonic behavioral shifts that transfer from R1-Distill-Qwen to its base Qwen-2.5-14B-Instruct counterpart. However, projecting out this direction paradoxically *increases* opponent-modeling behavior ($+11.6$ pp, specific to this direction), and a logistic probe achieving 87% accuracy finds a discriminative direction nearly orthogonal to the DoM vector (cosine ≈ 0.15). These results reveal structured representational geometry for strategic reasoning that is behaviorally steerable, while cautioning that steering efficacy should not be equated with representational fidelity.

1. Introduction

Recent reasoning-tuned language models, including DeepSeek-R1 [3], OpenAI o1 [8], and QwQ [10], produce extended chains of thought containing rich strategic reasoning: modeling opponents, iterating over beliefs, identifying equilibria. These behaviors are legible in the text, but whether they correspond to structured representations in activation space remains unclear. If strategic sub-behaviors are encoded as identifiable directions in the residual stream, they could be monitored and modulated at inference time, offering a mechanistic lever for understanding and governing strategic reasoning in autonomous agents [1, 5].

We investigate this using activation steering and representation engineering [6, 11, 12, 15] to study the geometry of strategic reasoning representations [4, 7] in two reasoning-tuned models. Concurrently, Zhu et al. [14] showed that theory-of-mind beliefs are linearly represented, and Venhoff et al. [13] studied reasoning behaviors via steering vectors. We extend this line of work by applying multi-label annotation to isolate strategic categories, testing cross-family geometric replication, and critically examining the relationship between DoM directions and the model's internal discriminative features.

We study DeepSeek-R1-Distill-Qwen-14B and DeepSeek-R1-Distill-Llama-8B [3, 9] on 250 game-theoretic reasoning tasks plus 50 non-strategic controls, with 50 additional out-of-distribution (OOD) tasks designed to be substantially harder than the extraction set. Our contributions:

© .

1. **Cross-family structured geometry.** Opponent-modeling and deduction DoM directions are anti-aligned at every layer in both models, surviving permutation, bootstrap, split-half, and centering checks.
2. **Behavioral steering with an ablation paradox.** Steering produces specific, graded shifts in reasoning style that transfer from R1-Distill to its base counterpart. Ablating the opponent-modeling direction paradoxically *increases* the associated behavior ($+11.6$ pp), not matched by random or orthogonal controls, and replicating across in-distribution, OOD, and base-model transfer settings.
3. **Interpretive limits of DoM directions.** A logistic probe finds a discriminative direction nearly orthogonal to the DoM vector, with 2.5–$3.3\times$ better Fisher discriminant ratio, demonstrating that behavioral steering efficacy does not imply representational fidelity.

2. Methods

Dataset. We generate 250 game-theoretic tasks spanning matrix games, sequential games, auction bidding, incomplete information, negotiation and coordination, and repeated games, plus 50 non-strategic control tasks (algebra, logic, probability). An additional 50 OOD tasks use novel strategic structures without standard textbook framing, including multi-stage bargaining with asymmetric discount factors, threshold public goods games, and Bayesian auction mechanisms. Each task is presented to DeepSeek-R1-Distill-Qwen-14B (48 layers, hidden dim 5120) and DeepSeek-R1-Distill-Llama-8B (32 layers, hidden dim 4096) with greedy decoding.

Annotation. Reasoning chains are segmented at paragraph boundaries, yielding approximately 14.6K segments for Qwen and 23.1K for Llama. Each segment receives 1–3 labels from a ten-category multi-label taxonomy (Appendix A): six strategic and four general categories. Annotation uses GPT-5.4 with per-label definitions, positive and counter-examples, and explicit multi-label instructions with no priority rules. Calibration against 150 hand-labeled segments gives macro-F1 = 0.77 and minimum per-label F1 = 0.52. The multi-label rate is approximately 58%. Steered and ablated outputs are evaluated with GPT-5.4 using the same annotation prompt.

DoM vectors. For label c at layer ℓ, we compute the with-vs-without DoM vector $\mathbf{u}_c^{(\ell)} = \bar{\mathbf{h}}_c^{(\ell)} - \bar{\mathbf{h}}_{\neg c}^{(\ell)}$, where $\bar{\mathbf{h}}_c$ and $\bar{\mathbf{h}}_{\neg c}$ are mean activations of thinking-region segments with and without label c. We also compute leave-one-out and class-balanced variants as robustness checks.

Interventions. Following Turner et al. [12], we steer by adding $\alpha \cdot \hat{\mathbf{u}}_{\text{opp}} \cdot \|\bar{\mathbf{h}}\|$ to the residual stream at the last-token position on each generation step at the intervention layer. We ablate by projecting out the direction from the full sequence at the same layer. Controls include random-direction ablation, orthogonal-direction ablation (payoff analysis), and logistic probe direction ablation.

3. Results

3.1. Structured geometry replicates across models

Figure 1 shows headline-layer DoM cosine matrices for both models.The anti-alignment between opponent-modeling and deduction is negative at *every* layer in both models: Qwen-14B averages -0.72 (range $[-0.84, -0.65]$; -0.71 at L24) and Llama-8B averages -0.63 (range $[-0.74, -0.58]$; -0.61 at L16). Bootstrap 95% CIs are entirely negative, permutation tests yield $p < 0.001$, and

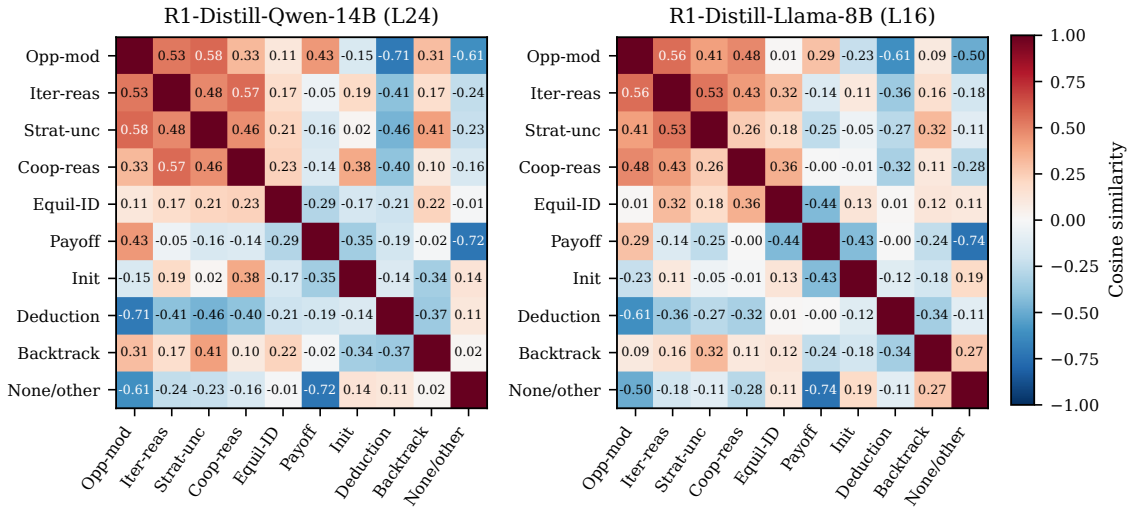

Figure 1: Headline-layer DoM cosine matrices for Qwen-14B and Llama-8B. Opponent-modeling and deduction are strongly anti-aligned in both models.

Table 1: Cross-model geometry and probe results.

	Qwen-14B (L24)	Llama-8B (L16)
Opp-mod vs. ded. cosine	-0.71	-0.61
Negative at all layers	48/48	32/32
Bootstrap 95% CI	$[-0.73, -0.68]$	$[-0.63, -0.59]$
Permutation p	< 0.001	< 0.001
Probe CV accuracy	86.9%	83.1%
Probe-DoM cosine	$+0.15$	$+0.19$
Fisher ratio (probe / DoM)	$2.5\times$	$3.3\times$
SVD: SV1 variance explained	34.3%	35.9%

split-half resampling produces 100% negative cosines (Table 1). The effect is robust to centering method (Appendix C).

The broader geometry shows consistent cluster structure: opponent modeling aligns positively with strategic uncertainty ($+0.41$ to $+0.58$) and iterated reasoning ($+0.53$ to $+0.56$), forming a social reasoning cluster that is anti-aligned with deduction. SVD of the category-mean matrix reveals multi-dimensional structure, with the first singular value explaining only 34–36% of variance at headline layers. The geometry is richer than a single axis.

3.2. Behavioral steering and the ablation paradox

Table 2 summarizes interventions on 50 strategic Qwen tasks. Positive steering monotonically increases opponent modeling from 33.5% at baseline to 47.4% at $\alpha = +0.3$, while strong negative steering collapses extended reasoning.

Table 2: Intervention results on 50 strategic tasks (R1-Distill-Qwen-14B). Label rates as percentage of annotated segments. Behavioral evaluation by GPT-5.4.

Condition	Segs	Opp-mod	Ded.	Payoff	Str-unc
Baseline	2540	33.5	19.1	57.7	17.3
Ablate opp-mod	2139	**45.0**	18.0	60.8	20.2
Ablate random	2446	31.5	19.8	54.2	18.6
Ablate payoff	2235	32.7	21.2	64.6	16.1
Ablate probe	2679	34.0	16.7	54.6	15.5
Steer -0.5	6257	0.1	13.5	48.6	1.2
Steer $+0.2$	2273	41.5	14.3	56.9	26.1
Steer $+0.3$	2442	47.4	15.5	53.5	27.1

The ablation result is more surprising: projecting out the opponent-modeling DoM direction increases opponent modeling from 33.5% to 45.0%, while random, payoff, and probe-direction ablations do not. The paradox also appears on R1 OOD tasks ($+8.1$ pp) and base-Qwen OOD tasks ($+12.9$ pp; Appendix B). At high steering strength, base Qwen exhibits $6.2\times$ segment inflation with high repetition, suggesting a dissociation between content modulation and procedural regulation [13].

3.3. Probe-DoM disagreement

A logistic probe separating opponent-modeling from deduction reaches 86.9% accuracy on Qwen and 83.1% on Llama, yet its weight vector is nearly orthogonal to the DoM contrast, with cosines of $+0.15$ and $+0.19$. This remains true across layers, with maximum absolute cosine below 0.21. Strong ℓ_2 regularization raises the cosine to 0.39 and 0.56, showing that the DoM direction contains signal but is not the best linear discriminant. The probe also achieves a 2.5–$3.3\times$ higher Fisher ratio and lies more than 97% outside the DoM SVD subspace, indicating that behaviorally effective steering directions need not align with the model's primary discriminative features [2, 11, 12].

4. Discussion

The cross-family anti-alignment suggests reasoning-tuned models develop organized representational structure for strategic concepts, extending findings on linear belief representations [14] and consistent with concurrent work [13]. The ablation paradox, where removing the opponent-modeling direction *increases* the behavior, suggests the DoM direction captures a regulatory or suppressive component rather than content encoding. This is relevant to governing strategic AI agents: the direction one would naturally target for suppression may not have the expected causal role. More broadly, the probe-DoM disagreement shows that a direction can be behaviorally effective when steered while being nearly orthogonal to the model's primary discriminant, cautioning against equating steering efficacy with representational understanding.

Limitations. Single LLM annotator (GPT-5.4) for evaluation. Two backbones from the same R1-Distill recipe; broader generalization untested. No mechanistic account of the ablation paradox. Task correctness metrics were insufficiently sensitive to separate reasoning quality from style.

References

[1] 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.

[2] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[3] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[4] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[5] Kanishk Gandhi, Dorsa Sadigh, and Noah D. Goodman. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

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

[7] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

[8] OpenAI. Learning to reason with LLMs. Technical Blog Post, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

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

[10] Qwen Team. QwQ: Reflect deeply on the boundaries of the unknown. Technical Blog Post, 2024. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

[11] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15504–15522, 2024.

[12] 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 arXiv:2308.10248*, 2023.

[13] Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding reasoning in thinking language models via steering vectors. *arXiv preprint arXiv:2506.18167*, 2025.

[14] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. *arXiv preprint arXiv:2402.18496*, 2024.

[15] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Appendix A. Annotation Taxonomy

Segments receive 1 to 3 labels from the taxonomy below. Multi-label assignment is explicitly encouraged: the annotation prompt states that "most reasoning segments contain two or more cognitive activities simultaneously." No priority hierarchy exists between strategic and general labels.

Appendix B. OOD and Transfer Results

The ablation paradox replicates across three settings (Table 4). The effect is specific to the opponent-modeling DoM direction in all cases and is not matched by random-direction ablation.

Appendix C. Robustness Details

The anti-alignment is robust to centering method. For Qwen-14B at L24, with-vs-without, original global-mean, and leave-one-out formulations all give -0.71; class-balanced centering remains negative but weaker (-0.37). For Llama-8B at L16, with-vs-without, original, and leave-one-out all give -0.61; class-balanced centering remains negative but weaker (-0.26). Every category mean has high cosine similarity with the global mean, indicating that the residual stream is dominated by a shared reasoning direction and that the primary anti-alignment is not a simple global-mean artifact.

Permutation tests (n=1000, shuffling multi-label assignments): null distribution mean approximately zero, observed values fall far outside ($p < 0.001$ for both models). Bootstrap 95% CIs (n=1000): Qwen $[-0.73, -0.68]$, Llama $[-0.63, -0.59]$. Split-half stability (n=100): 200/200 and 200/200 half-splits yield negative cosines.

Appendix D. Dataset and Code

The full extraction and OOD prompt sets are provided in anonymous supplementary material. The extraction set contains 300 tasks: 45 matrix games, 45 sequential games, 35 auction-bidding tasks, 45 incomplete-information tasks, 40 negotiation/coordination tasks, 40 repeated-game tasks, and 50 non-strategic controls. Each task includes a prompt, category, subcategory, optimal action, ground-truth answer, and reasoning tags. The OOD set contains 50 additional validation tasks covering incomplete-information, matrix, negotiation/coordination, repeated, and sequential-game settings. Tasks were adapted from standard game-theoretic structures and public educational examples, then manually rewritten, parameter-varied, and verified for payoff consistency, solvability, category fit, and answer validity.

Annotation, activation-extraction, DoM-vector computation, intervention, and analysis code is available at:
https://anonymous.4open.science/r/strategic-reasoning-steering-8FE0/

Table 3: Behavioral annotation taxonomy with definitions. Strategic categories (top) capture game-theoretic reasoning; general categories (bottom) capture domain-independent cognitive operations.

Label	**Definition**
Strategic categories	
opponent modeling	Reasoning about what another specific agent believes, knows, wants, will do, or is likely to do. The other agent is referenced explicitly (opponent, bidder, player, rival, they). Not applicable to reasoning about non-agent unknowns.
iterated reasoning	Higher-order belief reasoning with two or more levels of nested belief ("they think I think…"). This is a rare label; apply only when explicit nested belief structure is present. Not applicable to first-order opponent modeling.
equilibrium identification	Identifying or invoking a game-theoretic solution concept by name or structure: Nash equilibrium, dominant strategy, subgame-perfect equilibrium, mixed-strategy equilibrium, iterated elimination, etc. Not applicable to pure payoff comparison without a solution concept.
payoff analysis	Computing, comparing, or reasoning over concrete numerical outcomes, utilities, costs, or expected values. Numbers or arithmetic are typically present. Not applicable to pure symbolic manipulation without payoff semantics (that is deduction).
strategic uncertainty	Reasoning about unknown information in a strategic setting: hidden types, private values, mixed strategies as a response to uncertainty, updating beliefs over distributions, posterior probabilities, bluffing considerations. Not applicable to general epistemic hedging ("I'm not sure").
cooperative reasoning	Analyzing coordination, joint optimization, fairness, trust, or enforcement of cooperative outcomes. Includes grim trigger, tit-for-tat sustainability analysis, coalition formation, and fair-split reasoning. Not applicable to one-off cooperation choices without structural analysis.
General categories	
initialization	Actively parsing or restating the problem structure: listing players, enumerating actions, restating given payoffs or rules. Not applicable to off-task filler or short section headers.
deduction	Step-by-step logical derivation, procedural computation, or "if X then Y" inference chains that do not fit a more specific content label. Most common in non-strategic control tasks. Not a fallback for annotator uncertainty; if no label applies, use none/other.
backtracking	Correcting an earlier error, revising a prior conclusion, or explicitly retracting a previous step. Typical markers: "wait," "actually," "but earlier I thought." This label is cross-cutting: a segment that retracts a payoff computation receives both backtracking and payoff analysis.
none/other	Filler, transitions, off-task comments, short section headers, meta-remarks about the reasoning process ("let me verify," "I'm getting stuck here"). Used only when no content label applies; not a tie-breaker between content labels.

Table 4: Ablation paradox replication: opponent-modeling annotation rate (%) across experimental settings. Deltas in parentheses are relative to baseline.

Setting	Baseline	Ablate opp-mod	Ablate random
R1 in-distribution (50 tasks)	33.5	45.0 ($+11.6$)	31.5 (-2.0)
R1 OOD (50 tasks)	23.9	32.0 ($+8.1$)	25.2 ($+1.3$)
Base Qwen OOD (50 tasks)	28.1	41.0 ($+12.9$)	27.0 (-1.1)

Table 5: Content/control dissociation at α=0.3 on OOD tasks. The base model exhibits $6.2\times$ segment inflation without the backtracking present in R1-Distill.

Model / condition	Segs	Opp-mod	Backtrack	Tri-rep
R1-Distill (α=0.3)	4911	44.2%	14.2%	0.555
Base Qwen (α=0.3)	4049	39.7%	0.0%	0.453
Base Qwen (baseline)	652	28.1%	n/a	0.063