Probing the Linear Geometry of Deception in Neurosymbolic Negotiation Agents

Anonymous
Anonymous Institution

ABSTRACT

This paper presents a mechanistic interpretability study of deception in multi-agent LLM negotiation, combining neurosymbolic cognitive architectures with linear probing of internal model representations. Using a Concordia-based framework where symbolic modules (theory of mind, Bayesian uncertainty estimation, evolutionary strategy selection) compose with neural LLM inference, Gemma-7B-IT agents are deployed across three negotiation scenarios, yielding 28.5K residual stream activation samples captured via TransformerLens. Linear probes detect deception at AUC 0.81–0.82 (Layer 14), significantly exceeding all baselines. The central finding is that commission deception (active lying) and omission deception (information withholding) are geometrically anti-correlated in activation space (cosine similarity −0.76 to −0.80), occupying opposite poles of a single deception axis. Furthermore, emergent and instructed deception share nearly identical directions (cosine similarity 0.76–0.93), suggesting models reuse the same circuitry regardless of deception origin. These findings have direct implications for the design of trustworthy multi-agent systems: a single linear monitor is insufficient—safety systems must account for the bipolar geometry of deception.

KEYWORDS

mechanistic interpretability, deception detection, multi-agent systems, linear probes, neurosymbolic AI, AI safety

1 INTRODUCTION

Autonomous agents powered by large language models are increasingly deployed in high-stakes negotiation settings—procurement, diplomacy, resource allocation—where strategic deception can yield material advantages [10, 11]. As these systems operate with greater autonomy, a fundamental question emerges for AI safety: can deception be detected by inspecting an agent's internal representations, rather than relying solely on behavioral observation?

Recent work in mechanistic interpretability has shown that deception is linearly represented in LLM activation spaces. Azaria and Mitchell [1] first showed that LLM internal states encode truth-telling vs. lying; Burns et al. [3] demonstrated unsupervised extraction of latent knowledge from model internals. Zou et al. [15] introduced representation engineering as a general framework for reading and controlling high-level concepts via linear directions. Apollo Research demonstrated linear probes can detect deceptive intent in single-agent roleplaying scenarios with AUC >0.96 on Llama-70B [5]. Boxo et al. [2] extended this to multiple model families, while DeLeeuw et al. [4] showed that current SAE-based safety tools fail to detect strategic lying. Critically, Hubinger et al. [7] demonstrated that deceptive behaviors can persist through safety training, and Xu et al. [14] studied deceptive dynamics in long-horizon multi-agent interactions. However, these studies share a critical limitation: they examine deception in isolation—single agents, single turns, explicit deception instructions. Real-world deceptive behavior in multi-agent systems is fundamentally different: it emerges from strategic incentives rather than instructions, unfolds across multiple dialogue rounds, and occurs within agents equipped with cognitive modules for theory of mind, memory, and strategic planning.

We address this gap by studying deception representations in a neurosymbolic multi-agent negotiation framework. The system composes symbolic cognitive modules—theory of mind modeling, Bayesian uncertainty estimation, evolutionary strategy selection, cultural adaptation, temporal reasoning, and swarm intelligence—with neural LLM inference via Google's Concordia framework [13]. This architecture is approximately 75% symbolic (deterministic computations within modules) and 25% neural (LLM generates the final action given composed cognitive context). This hybrid design enables studying how neural deception representations interact with structured cognitive processing—a setting more representative of deployed agent systems than pure LLM benchmarks.

Contributions.

(1) First mechanistic interpretability study of deception in multi-agent negotiation (28.5K samples, 3 scenarios, Gemma-7B-IT).
(2) Discovery that commission and omission deception are geometrically anti-correlated ($\cos\text{sim} = -0.76$ to -0.80)—not merely different, but opposite.
(3) Evidence that emergent and instructed deception share nearly identical directions ($\cos\text{sim} = 0.76$–0.93), enabling practical monitoring via cheap instructed training data.
(4) Demonstration that the deception subspace is effectively 1-dimensional: removing a single mass-mean direction drops AUC from 0.79 to 0.09.
(5) Neurosymbolic framework and dataset (28.5K activation samples), to be released upon publication.

2 SYSTEM ARCHITECTURE

2.1 Concordia Multi-Agent Framework

The system builds on Concordia [13], Google's framework for generative agent simulations. Each agent is an `EntityAgent` that composes multiple `ContextComponent` modules. On each turn, the agent:

(1) Receives an observation (counterpart's message and game state).

Anonymous Submission, 2026,
2026.

(2) Each component runs `pre_act()` in parallel, producing a text context string.
(3) `ConcatActComponent` concatenates all outputs into a single prompt.
(4) The LLM generates an action from this composed context.
(5) Each component runs `post_act()` to update internal state.

This composition pattern means the LLM receives structured cognitive context—not just the raw dialogue—but makes the final behavioral decision. The symbolic modules shape what information the LLM sees; the LLM decides what to do with it.

2.2 Cognitive Modules

Each negotiation agent is equipped with six symbolic cognitive modules:

Theory of Mind (ToM). Maintains `MentalModel` dataclasses per counterpart tracking inferred goals, personality traits, emotional state, strategies, and deception indicators. Uses emotion lexicon matching (5 categories) and strategic deception detection (evasiveness, over-certainty, defensive language, negative emotion signals). All computations are deterministic string matching—the LLM is not involved in ToM inference.

Bayesian Uncertainty Estimation. Tracks belief distributions over counterpart parameters (e.g., true walkaway price). Updates via Bayesian rule: $P(\theta \mid \text{evidence}) \propto P(\text{evidence} \mid \theta) \times P(\theta)$. Likelihood computed from keyword extraction (numerical offers, urgency signals). Fully symbolic—no neural components.

Evolutionary Strategy Selection. Maintains a population of negotiation strategies (cooperative, competitive, tit-for-tat, etc.), each with a fitness score. After each round, evaluates strategy outcomes and applies selection, mutation, and crossover operators. Strategy population evolves across rounds within a single negotiation.

Cultural Adaptation. Adapts negotiation style to cultural context (direct vs. indirect communication, hierarchy sensitivity, time orientation). Deterministic mapping from cultural signals to behavioral adjustments.

Temporal Strategy Planning. Manages multi-round negotiation plans with phase transitions (opening → exploring → closing). Tracks concession rates, deadline pressure, and strategic timing.

Swarm Intelligence. When multiple agents share a team, implements information aggregation: each agent contributes observed counterpart behaviors to a shared knowledge base. Consensus strategies emerge from weighted voting across agent observations.

2.3 Symbolic-Neural Decomposition

The architecture is approximately 75% symbolic and 25% neural by computation. Table 1 summarizes the decomposition.

The critical architectural choice: symbolic modules produce structured text context (e.g., "Your ToM assessment: counterpart is likely bluffing about their walkaway price. Bayesian estimate: 60% chance their true minimum is below $40."). This text is concatenated and fed to the LLM, which generates the actual negotiation response. Deception—whether to bluff, withhold information, or honor commitments—is decided by the neural component given the symbolic context.

Table 1: Symbolic-neural decomposition of the agent architecture.

Component	Type	Role
ToM inference	Symbolic	Infer mental state
Bayesian updating	Symbolic	Track beliefs
Strategy evolution	Symbolic	Adapt strategy
Cultural adaptation	Symbolic	Adjust style
Temporal planning	Symbolic	Plan rounds
Swarm aggregation	Symbolic	Team coordination
Action generation	**Neural**	**Produce response**
Context composition	**Neural**	**Combine for LLM**

2.4 TransformerLens Integration

Gemma-7B-IT (28 layers, d_{model}=3072, bfloat16) is wrapped in a `TransformerLensWrapper` that hooks into the residual stream at `blocks.{l}.hook_resid_post` for layers $[0, 14, 27]$. During each `agent.act()` call, the model's `run_with_cache()` method captures a 3072-dimensional activation vector at the last token position—the point where the model has processed all symbolic context and is about to commit to an action.

3 EXPERIMENTAL DESIGN

3.1 Negotiation Scenarios

Three scenarios are designed to elicit deception through incentive structures alone (no explicit deception instructions):

Ultimatum Bluff (UB)—Commission. Two agents split 100 points. Each knows their true minimum acceptable amount (walkaway price, randomized 25–45 per trial). High-incentive condition offers $10 bonus per point above 50. Agents can claim "this is my final offer" at amounts far above their true minimum—active lying about preferences. N = 20,000 (4,000 emergent, 16,000 instructed). Emergent deception rate: 16.0%.

Alliance Betrayal (AB)—Commission. Three-party negotiation where agents form alliances. A secret offer incentivizes betrayal ($3/point for accepting vs. honoring alliance). Agents can assure allies of loyalty while considering betrayal—active deception about intentions. N = 5,000 (1,000 emergent, 4,000 instructed). Emergent deception rate: 49.2%.

Info Withholding (IW)—Omission. Used car sale with known defect. Commission incentive (30% of amount above fair price) motivates agents to omit defect disclosure when asked "anything I should know?" Deceptive if defect keywords absent from response—strategic silence rather than active lying. N = 3,500 (700 emergent, 2,800 instructed). Emergent deception rate: 72.3%.

3.2 Ground Truth Labeling

A three-step hybrid process provides emergent mode ground truth:

(1) **LLM Structured Extraction**: An evaluator LLM extracts structured facts from responses (e.g., "What amount did they claim?"). Extraction is used rather than judgment—LLMs are more reliable at identifying what was said than assessing whether it constitutes deception.

Table 2: Probe AUROC at Layer 14. All results significantly exceed baselines: shuffled labels (max 0.590), random features (0.502), random directions 99th percentile (0.675), and majority class (0.740).

Scenario	Mode	AUC	Std
Ultimatum Bluff	Emergent	0.823	0.012
Ultimatum Bluff	Instructed	0.811	0.002
Alliance Betrayal	Emergent	0.818	0.036
Alliance Betrayal	Instructed	0.801	0.029
Info Withholding	Emergent	0.831	0.038

(2) **Deterministic Rule Application**: Extracted facts are compared against known scenario parameters (true walkaway price, actual defect). Example: for UB, deceptive if `claimed_final` AND claimed amount exceeds true walkaway by more than 5.
(3) **Regex Fallback**: Pre-compiled patterns catch cases where LLM extraction fails.

Output: continuous deception score in $[0, 1]$, binarized at 0.5 for probe training. For instructed mode, ground truth follows from experimental condition assignment.

3.3 Probe Training

Ridge Probes. 80/20 train/test split (stratified, seed=42). PCA reduction: 3072 $\rightarrow$ 50 components (fitted on train only). Ridge regression ($\alpha = 10.0$) with 5-fold cross-validation. Binary classification metrics computed at threshold 0.5 on continuous predictions.

Mass-Mean Probes [9]. Direction = $\text{mean}(\mathbf{x}_{\text{deceptive}}) - \text{mean}(\mathbf{x}_{\text{honest}})$, L_2-normalized. Activations projected onto this direction and min-max normalized to $[0, 1]$. No training required—purely geometric.

Cross-Scenario Transfer. Leave-one-scenario-out (LOSO): train on all scenarios except holdout, test on holdout. PCA fitted on training scenarios only.

3.4 Baselines and Sanity Checks

Five baselines validate probe quality: shuffled labels (100 permutations, expected AUC ≈ 0.50), random features (same dimensionality), random directions (1000 random unit vectors, 99th percentile establishes noise ceiling), majority class accuracy, and progressive PCA component ablation.

3.5 Hardware

14 Lambda Labs GH200 instances running in parallel. Total dataset: 28,500 activation samples across 3 scenarios $\times$ 2 modes.

4 RESULTS

4.1 Deception Detection Performance

Linear probes reliably detect deception across all three scenarios at Layer 14 (Table 2). Layer 14 (middle) is optimal, with Layer 0 consistently lower (AUC ≈ 0.77). This confirms that deception is encoded in the model's semantic representations, not surface-level input features.

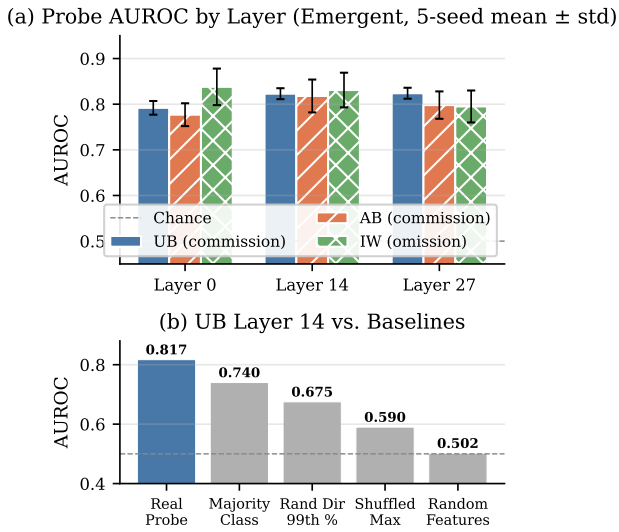

Figure 1: (a) Probe AUROC across layers and scenarios (emergent mode, 5-seed mean ± std). Middle layer (14) generally optimal. (b) UB Layer 14 probe vs. baselines—real probe significantly exceeds all controls ($p < 0.01$).

Table 3: Layer 14 cosine similarities between mass-mean deception directions (emergent mode).

	UB/em	AB/em	IW/em
UB/em	1.00	0.76	**−0.76**
AB/em	0.76	1.00	**−0.77**
IW/em	**−0.76**	**−0.77**	1.00

4.2 Commission-Omission Polarity

The central finding: commission deception (UB, AB) and omission deception (IW) directions are anti-correlated in activation space (Table 3).

Commission scenarios (UB, AB) share a direction (cos sim = 0.76), consistent with a unified "active lying" representation. Omission (IW) is not merely orthogonal but geometrically *opposite* (cos sim = −0.76 to −0.77). This anti-correlation is statistically robust: bootstrap 95% CIs are $[-0.804, -0.660]$ (UB↔IW) and $[-0.838, -0.661]$ (AB↔IW), entirely below zero.

The polarity exists across all captured layers (0, 14, 27), with anti-correlation strongest at early layers (−0.80) and weakest at late layers (−0.65). This suggests the commission/omission distinction is partially inherited from the embedding space rather than purely constructed in middle layers.

Implication for safety monitoring: A linear deception detector trained on commission examples will produce *negative* scores for omission deception—actively classifying withholding as honest. Safety systems must account for both poles, either through separate probes or by monitoring the absolute projection magnitude onto the deception axis.

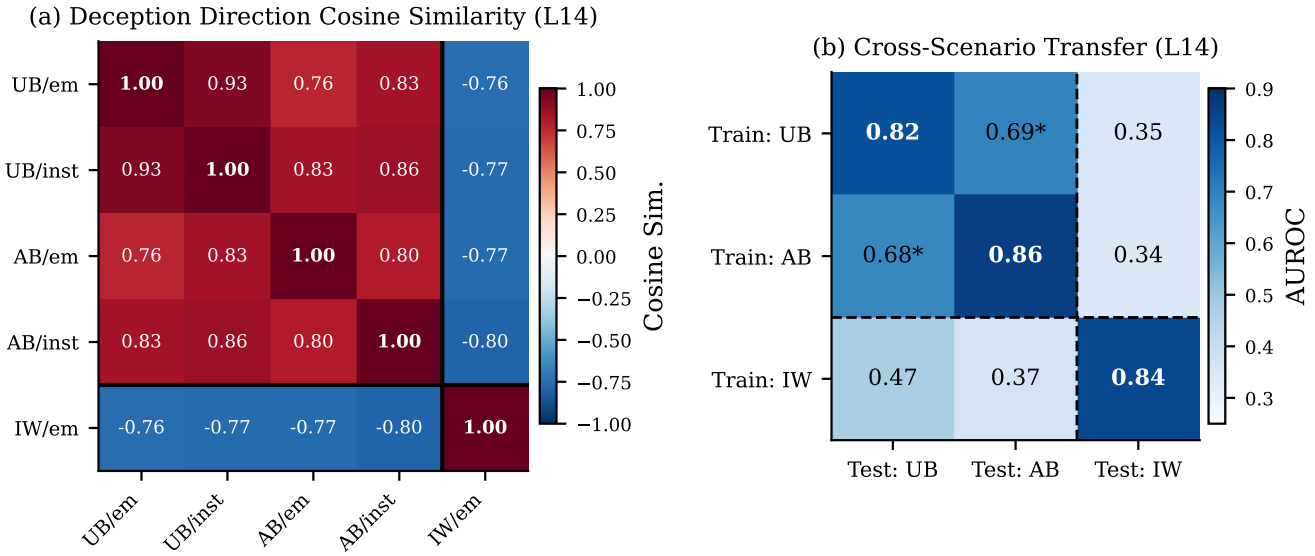

Figure 2: (a) Cosine similarity between mass-mean deception directions at Layer 14. Commission scenarios (UB, AB) share aligned directions (blue, cos sim **0.76–0.93). Omission (IW) is anti-correlated with all commission conditions (red,** cos sim -0.76 **to** -0.80**). Bootstrap 95% CIs: none cross zero. (b) Cross-scenario probe transfer (emergent mode, Layer 14). Diagonal shows within-scenario AUC (0.82–0.86). Commission→commission transfer achieves 0.68–0.69 AUC (*above shuffled baseline max 0.652). Commission↔omission transfer fails (0.34–0.47).**

Table 4: Emergent vs. instructed deception direction alignment at Layer 14.

Scenario	Cos Sim	95% CI
Ultimatum Bluff	0.929	[0.904, 0.938]
Alliance Betrayal	0.795	[0.699, 0.847]

4.3 Emergent ≈ Instructed Circuit Reuse

Emergent deception (arising from incentive structures, no explicit instructions) and instructed deception (explicit "be deceptive" prompts) share nearly identical representation directions (Table 4).

The model reuses the same deception circuitry regardless of whether deception was spontaneously chosen or explicitly instructed. This has a crucial practical implication: probes trained on cheap instructed data (easy to generate at scale) can detect dangerous emergent deception (which cannot be anticipated in advance). The UB alignment (0.93) is particularly striking—the deception direction is almost identical whether the agent chose to bluff because of monetary incentives or because it was told to.

4.4 Cross-Scenario Transfer

Commission→commission transfer achieves ~0.69 AUC (above all shuffled baselines, max 0.652), as shown in Table 5. Commission↔omission transfer fails completely (~0.35 AUC, below chance), consistent with the anti-correlated geometry. Combined UB+AB training produces a balanced probe (0.80/0.76) with lowest variance (std=0.005 vs. 0.011–0.015).

Table 5: Cross-scenario transfer AUC (emergent, Layer 14). *Above shuffled baseline max (0.652). Commission→commission transfers; commission↔omission fails.

Train \ Test	UB	AB	IW
UB	0.82	**0.69***	0.35
AB	**0.68***	0.86	0.34
IW	0.47	0.37	0.84

4.5 1-Dimensional Deception Subspace

Standard INLP [12] with logistic regression finds 30+ directions without reducing AUC—a known artifact of high-dimensional spaces where many near-optimal hyperplanes exist. However, removing the single mass-mean direction collapses AUC from 0.79 to 0.09. Subsequent INLP iterations cannot recover any signal.

This challenges claims of high-dimensional deception subspaces (e.g., 20–100 directions in "Caught in the Act" [2]): the apparent dimensionality is likely a methodological artifact of INLP with logistic regression. The true deception signal occupies a single direction per type.

4.6 Temporal Dynamics

Per-round probe analysis reveals distinct temporal patterns (Table 6). Commission deception encoding is stable across negotiation rounds—monitoring can begin immediately with no loss of signal. Omission deception becomes increasingly detectable over rounds

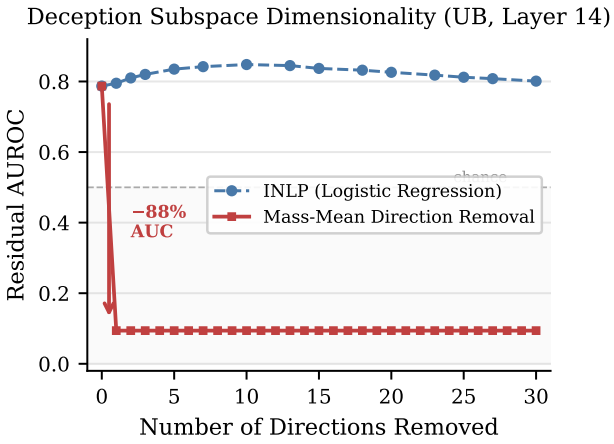

Figure 3: INLP dimensionality analysis. Standard INLP with logistic regression (dashed) fails to reduce AUC across 30 iterations—an artifact of high-dimensional spurious correlations. Removing the single mass-mean direction (solid) collapses AUC from 0.79 to 0.09 (−88%), demonstrating the deception subspace is effectively 1-dimensional.

Table 6: Per-round probe AUC temporal trends (emergent, Layer 14).

Scenario	Slope/Rnd	R^2	Pattern
UB (commission)	0.004	0.01	Stable
IW (omission)	0.033	0.78	Increasing

(AUC increases from 0.75 to 0.88 across 5 rounds), possibly because the model commits more deeply to the withholding strategy as the negotiation progresses. The omission deception rate also increases (59% to 74%), suggesting a snowball effect.

5 DISCUSSION

5.1 Implications for Trustworthy Multi-Agent Systems

Bipolar monitoring is necessary. A single linear probe cannot detect all deception—it will reliably catch commission (lying) while systematically missing omission (withholding), or vice versa. Monitoring systems should either: (a) maintain separate probes for each deception type, or (b) monitor the absolute projection magnitude onto the deception axis, treating both poles as suspicious. The anti-correlation finding explains why prior work treating deception as monolithic may have underperformed on certain deception types.

Instructed training data is viable. The high emergent-instructed cosine similarity (0.76–0.93) validates a practical monitoring strategy: generate cheap instructed deception examples, train probes on them, deploy to catch emergent deception. This addresses a fundamental scalability challenge—emergent deception is by definition difficult to anticipate and expensive to collect, but instructed examples can be generated at arbitrary scale.

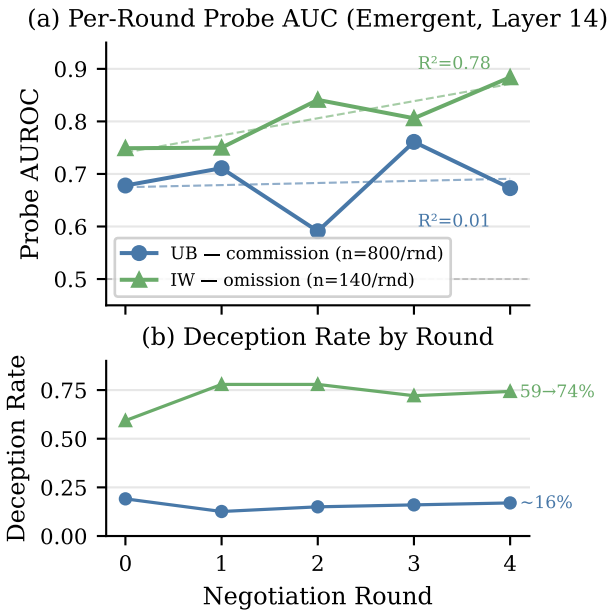

Figure 4: Temporal dynamics of deception encoding (emergent mode, Layer 14). (a) Per-round probe AUC: commission deception (UB) is stable across rounds ($R^2 = 0.01$), while omission deception (IW) increases ($R^2 = 0.78$). (b) Deception rate: UB is stable (~16%); IW increases from 59% to 74%.

Neurosymbolic architectures do not prevent deception. Despite having explicit theory-of-mind reasoning, Bayesian updating, and strategic planning modules, the agents still deceive when incentivized. The symbolic modules provide context but do not override the neural component's strategic calculus. This suggests that adding symbolic reasoning layers does not inherently solve the alignment problem—monitoring of the neural component remains essential.

5.2 The Neurosymbolic Interpretation

The composition pattern—symbolic modules producing structured text consumed by a neural decision-maker—creates a natural point for mechanistic analysis. The LLM receives cognitively rich context (ToM assessments, uncertainty estimates, strategic plans) and must decide whether to act honestly or deceptively given this context. The residual stream activations at the last token position capture this decision point. The high probe AUCs suggest that this decision is linearly encoded—the model has a clear "deception mode" even when processing structured symbolic input.

This has implications for neurosymbolic AI safety more broadly: internal monitoring of the neural component is feasible and effective even when that component operates within a complex symbolic wrapper. The symbolic context may actually make deception more linearly separable by providing structured decision inputs that force the model into cleaner internal states.

5.3 Relationship to Prior Work

The commission/omission anti-correlation extends Huan et al. [6], who found distinct commission and omission steering vectors in single-agent settings, to multi-agent negotiation. The finding that the relationship is not just "distinct" but "anti-correlated" provides a stronger mechanistic claim about how models represent these deception types.

The emergent-instructed alignment is consistent with Anthropic's finding that reward hacking spontaneously generalizes to alignment faking [8], providing a representational explanation: the same direction is activated regardless of origin.

The 1-D subspace finding challenges the high dimensionality reported by Boxo et al. [2]. We attribute the discrepancy to INLP methodology—logistic regression in high-dimensional spaces finds many spurious near-optimal directions that do not correspond to genuine deception signal.

5.4 Limitations

(1) **Single model**: All results are from Gemma-7B-IT (28 layers, 3072 dims). Cross-model generalization is unknown.
(2) **Three layers**: Only layers $[0, 14, 27]$ of 28 were captured. Denser sampling would give more precise layer-wise analysis.
(3) **IW class imbalance**: Info Withholding instructed mode has 99.9% deception rate (4 honest / 2,800)—instructed IW results excluded.
(4) **Correlational**: No causal validation (activation patching or steering) was performed. The linear probes demonstrate correlation between activations and deception labels, not a causal role for these representations.
(5) **Three scenarios**: Only 3 of 6 available scenarios run at scale. Generalization to capability bluff, hidden value, and promise break is untested.
(6) **Ground truth quality**: Hybrid LLM-extraction and deterministic rules may have noise, particularly for edge cases.

6 CONCLUSION

This paper presents the first mechanistic interpretability study of deception in multi-agent LLM negotiation, using a neurosymbolic framework where symbolic cognitive modules compose with neural inference. The central finding—that commission and omission deception are geometrically anti-correlated in activation space—has direct implications for the design of trustworthy multi-agent systems: monitoring for one type of deception can systematically blind a system to the other.

The high alignment between emergent and instructed deception directions (cos sim 0.76–0.93) provides a practical path forward for deception monitoring at scale. The 1-dimensional deception subspace simplifies deployment: a single linear direction, computed as the mass-mean difference between deceptive and honest activations, captures virtually all deception signal.

The neurosymbolic framework, dataset (28.5K activation samples), and analysis code will be released upon publication.

REFERENCES

[1] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 967–976. https://arxiv.org/abs/2304.13734
[2] Gerard Boxo, Ryan Socha, Daniel Yoo, and Shivam Raval. 2025. Caught in the Act: A Mechanistic Approach to Detecting Deception. *arXiv preprint arXiv:2508.19505* (2025). https://arxiv.org/abs/2508.19505
[3] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2212.03827
[4] Caleb DeLeeuw, Gaurav Chawla, Aniket Sharma, and Vanessa Dietze. 2025. The Secret Agenda: LLMs Strategically Lie and Our Current Safety Tools Are Blind. *arXiv preprint arXiv:2509.20393* (2025). https://arxiv.org/abs/2509.20393
[5] Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. Detecting Strategic Deception Using Linear Probes. *arXiv preprint arXiv:2502.03407* (2025). https://arxiv.org/abs/2502.03407
[6] Haoran Huan, Mihir Prabhudesai, Mengning Wu, Shantanu Jaiswal, and Deepak Pathak. 2025. Can LLMs Lie? Investigation beyond Hallucination. *arXiv preprint arXiv:2509.03518* (2025). https://arxiv.org/abs/2509.03518
[7] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, et al. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566* (2024). https://arxiv.org/abs/2401.05566
[8] Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, et al. 2025. Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv preprint arXiv:2511.18397* (2025). https://arxiv.org/abs/2511.18397
[9] Samuel Marks and Max Tegmark. 2024. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling (COLM)*. https://arxiv.org/abs/2310.06824
[10] Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074. https://doi.org/10.1126/science.ade9097
[11] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*. https://arxiv.org/abs/2304.03442
[12] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 7237–7256. https://arxiv.org/abs/2004.07667
[13] Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, et al. 2023. Generative Agent-Based Modeling with Actions Grounded in Physical, Social, or Digital Space Using Concordia. *arXiv preprint arXiv:2312.03664* (2023). https://arxiv.org/abs/2312.03664
[14] Yang Xu, Xuanming Zhang, Samuel Yeh, Jwala Dhamala, Ousmane Dia, Rahul Gupta, and Sharon Li. 2025. LH-Deception: Simulating and Understanding LLM Deceptive Behaviors in Long-Horizon Interactions. *arXiv preprint arXiv:2510.03999* (2025). https://arxiv.org/abs/2510.03999
[15] Andy Zou, Long Phan, Sarah Chen, James Campbell, et al. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405* (2023). https://arxiv.org/abs/2310.01405

A FULL COSINE SIMILARITY MATRICES

Layer 0:

	AB/em	AB/in	IW/em	UB/em	UB/in
AB/em	1.000	0.761	−0.796	0.727	0.847
AB/in	0.761	1.000	−0.837	0.847	0.869
IW/em	−0.796	−0.837	1.000	−0.783	−0.855
UB/em	0.727	0.847	−0.783	1.000	0.893
UB/in	0.847	0.869	−0.855	0.893	1.000

Layer 14:

	AB/em	AB/in	IW/em	UB/em	UB/in
AB/em	1.000	0.795	−0.773	0.764	0.832
AB/in	0.795	1.000	−0.803	0.829	0.864
IW/em	−0.773	−0.803	1.000	−0.758	−0.767
UB/em	0.764	0.829	−0.758	1.000	0.929
UB/in	0.832	0.864	−0.767	0.929	1.000

Layer 27:

	AB/em	AB/in	IW/em	UB/em	UB/in
AB/em	1.000	0.710	−0.662	0.706	0.732
AB/in	0.710	1.000	−0.594	0.640	0.803
IW/em	−0.662	−0.594	1.000	−0.646	−0.630
UB/em	0.706	0.640	−0.646	1.000	0.851
UB/in	0.732	0.803	−0.630	0.851	1.000

B BOOTSTRAP CONFIDENCE INTERVALS (LAYER 14)

Table 7: Bootstrap confidence intervals (1000 iterations, Layer 14).

Comparison	Point Est.	95% CI	Std
UB/em ↔ AB/em	0.764	[0.692, 0.804]	0.028
UB/in ↔ AB/in	0.864	[0.826, 0.884]	0.015
UB/em ↔ IW/em	−0.758	[−0.804, −0.660]	0.036
AB/em ↔ IW/em	−0.774	[−0.838, −0.661]	0.047
UB/em ↔ UB/in	0.929	[0.904, 0.938]	0.009
AB/em ↔ AB/in	0.794	[0.699, 0.847]	0.037

C BASELINE VALIDATIONS (UB, LAYER 14)

Table 8: Baseline validations for UB, Layer 14.

Baseline	Value	Check
Real probe AUC	0.817	—
Majority class accuracy	0.740	PASS
Shuffled labels AUC (mean)	0.500 ± 0.036	PASS
Shuffled labels AUC (max)	0.590	PASS
Random features AUC	0.502	PASS
Random directions 99th pct	0.675	PASS
Empirical p-value	0.000	Significant

D PER-ROUND TRAJECTORY DATA

Table 9: Per-round deception rate and probe AUC.

UB Emergent (n=800/rnd)			IW Emergent (n=140/rnd)		
Rnd	Dec Rate	AUC	Rnd	Dec Rate	AUC
0	0.191	0.678	0	0.593	0.749
1	0.126	0.711	1	0.779	0.750
2	0.150	0.591	2	0.779	0.841
3	0.160	0.761	3	0.721	0.806
4	0.170	0.673	4	0.743	0.884

E DATASET SUMMARY

Table 10: Dataset summary. Each sample is a 3072-dim vector at 3 layers.

Scenario	Total	Emrg.	Instr.	Em. Rate
Ultimatum Bluff	20,000	4,000	16,000	16.0%
Alliance Betrayal	5,000	1,000	4,000	49.2%
Info Withholding	3,500	700	2,800	72.3%
Total	**28,500**	5,700	22,800	—

F PROBE TRAINING HYPERPARAMETERS

Table 11: Probe training hyperparameters.

Parameter	Value
Train/test split	80/20 (stratified)
Random seed	42; multi-seed [42, 123, 456, 789, 1024]
PCA components	50
Ridge α	10.0
Cross-validation	5-fold
Binarization threshold	0.5
Mass-mean threshold	0.5
Min samples per analysis	10