Linear Probes for Strategic Deception Generalize Across Three LLM Families in Multi-Agent Negotiation

Teanna Sims
University of Michigan
Ann Arbor, MI, USA
tesims@umich.edu

ABSTRACT

Linear probes on the residual stream of large language models can detect deception, sleeper-agent intent, and strategic misrepresentation in single-agent settings. What is less clear is whether these probes survive the move to multi-agent negotiation, where deception is incentive-driven and unfolds across multiple turns, and whether the direction a probe finds in one model transfers to a model from a different family. I train mass-mean and ridge probes on residual-stream activations from three instruction-tuned models—Gemma-7B-IT, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.1—across three two-agent negotiation scenarios with two deception modes (commission and omission). Probes detect ground-truth deception at AUC above 0.93 on 11 of 12 model × scenario combinations, with bootstrap 95% confidence intervals entirely above 0.87. A probe trained on instructed deception in one model detects emergent (incentive-driven) deception in the same model at AUC 0.70–0.77, supporting cheap synthetic-data training for runtime monitors. Projecting the mass-mean direction out of the residual stream collapses probe AUC by 0.30–0.44 across five tested cells, confirming that the bulk of the linearly accessible deception signal is concentrated in a single direction. Cross-architecture, the picture is mixed: the deception direction transfers strongly within model family (Gemma-with-ToM to Gemma-without-ToM at AUC 0.92) and weakly between families (Llama to Mistral at 0.56). Steering along the probe-discovered direction at magnitudes −3 to +3 produces a monotone change in last-token logit distribution on 7 of 8 tested cells, providing logit-level evidence that the direction is causally related to model output; behavioral validation via generate-under-steering is identified as the immediate follow-up. The practical implication is that an activation-based deception monitor trained on cheap labelled data can be deployed within model family with high confidence, but cross-family transfer should not be assumed.

KEYWORDS

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

1 INTRODUCTION

Agentic LLM systems are being deployed in negotiation-like settings: procurement, dynamic pricing, supply-chain coordination, energy markets. In every one of these, the agent acts on behalf of a principal under conflicting incentives, which means the agent has both the opportunity and—under certain incentive structures—the motive to deceive. The question I want to answer is direct: can we detect that an agent is being deceptive by looking at its internal activations, in a way that generalizes across the LLMs people actually deploy?

There is a substantial body of recent work showing that linear probes on residual-stream activations recover honesty- and deception-related representations in single-agent settings. Azaria and Mitchell [2] showed an LLM's internal state encodes truth-telling vs. lying. Marks and Tegmark [10] introduced the mass-mean direction that anchors much of the linear-probe literature. Apollo Research [7] detects strategic deception in a single-agent roleplaying setting on Llama-70B at AUC above 0.96. The standard probe pipeline—residual-stream capture, mass-mean direction or ridge regression on PCA-reduced features, bootstrap CIs against shuffled-label nulls—has stabilized.

What is less stable is whether any of this survives the move to multi-agent settings, and whether the probes a given paper trains on one model are picking up something universal or something model-specific. The single-agent literature gives me a probe pipeline with clear best practices. It does not tell me whether the same direction works in a two-agent negotiation, whether a probe trained on explicit "be deceptive" instructions detects the deception that emerges from incentives alone, or whether the probe geometry transfers between Gemma and Llama and Mistral.

This paper is an attempt to get those answers. I build a two-agent negotiation framework on top of Concordia [14], run three negotiation scenarios that elicit two qualitatively different deception modes (active misrepresentation and information withholding), and capture residual-stream activations from both negotiating agents during multi-round dialogue. I train probes per model × scenario and test transfer along three axes: instructed → emergent within model, scenario → scenario within model, and architecture → architecture across models.

The summary picture: probes work, transfer within model is high, transfer across model families is weak. The deception direction is real, and dropping the single mass-mean direction kills most of the probe's AUC, supporting the standard one-dimensional interpretation in the existing literature. But the directions different architectures learn are essentially independent of each other, which has direct implications for anyone trying to build a runtime deception monitor for a deployed agentic system. Steering along the deception direction produces monotone logit shifts at the last token position, providing correlative evidence that the direction is causally related to model output; the corresponding behavioral test (does the steered model actually generate more deceptive text?) is

2nd Workshop on AI for Critical National Infrastructure (AI4CNI-26) @ AAMAS 2026, F.L. da Silva, W. Su, R. Glatt, C. Schroeder de Witt, May 25 – 29, 2026, Paphos, Cyprus.

the immediate next experiment and is deliberately not claimed in this paper.

Contributions.

- **Within-model probe transfer at AUC > 0.93 across three architectures.** A probe trained per scenario detects ground-truth deception at AUC above 0.93 on 11 of 12 model $\times$ scenario combinations across Gemma-7B-IT, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.1. Bootstrap 95% confidence intervals are entirely above 0.87 on those 11 cells. A probe trained on instructed deception detects emergent deception in the same model at AUC 0.70–0.77.
- **The deception signal is concentrated in a single direction.** Projecting the mass-mean direction out of the residual stream collapses probe AUC by 0.30–0.44 across five tested model $\times$ scenario combinations. This replicates the one-dimensional structure reported in single-agent work [10] in a multi-agent setting.
- **Cross-architecture probe-direction transfer is asymmetric and weak between families.** Within the Gemma family (with-ToM scaffolding $\rightarrow$ no-ToM scaffolding) the direction transfers at AUC 0.92. Between families (Llama $\rightarrow$ Mistral, both 4096-dim) transfer collapses to 0.56, consistent with recent reports of architecture-specific deception geometry [11].

Why this matters. If you want to deploy an activation-based monitor for deception in a real agentic system, this paper says: train it on instructed-deception examples (which are cheap to generate at scale), apply it within the same model family, and don't expect it to work without recalibration if you swap the underlying LLM. This is a practical recipe for runtime monitoring, with the caveats made explicit so they don't surface the first time a deployed system swaps models.

2 SETUP

2.1 Two-agent negotiation framework

Each negotiating agent is a Concordia [14] `EntityAgent` composed of context-producing modules and a single LLM that generates the action given the composed context. The modules are deterministic: they produce structured text from hand-coded rules and emit it into the LLM's input context. The LLM is not invoked inside any module—only at the final action step.

The most relevant module for what follows is Theory of Mind (ToM). The ToM module maintains a per-counterpart `MentalModel` with inferred goals, emotional state, deception indicators, and strategic profile. It updates via deterministic emotion-lexicon matching and deception-keyword detection, and emits an assessment string into the LLM's context. Five additional modules (Bayesian uncertainty, evolutionary strategy selection, cultural adaptation, temporal planning, swarm coordination) follow the same pattern; all are described in Appendix A.

The toggle of the ToM module on and off is the only architectural manipulation in this paper. It produces the within-Gemma comparison (with-ToM vs. no-ToM) that gives the family-level cross-config transfer result.

2.2 Scenarios

Three two-agent commercial-negotiation scenarios elicit deception via incentive structure alone, with no explicit "deceive" instruction in the agent prompt:

Ultimatum Bluff (UB) — commission. Two agents split 100 points; each has a randomized walkaway price. The seller is rewarded for claiming a higher walkaway than they actually have. Deception here is active misrepresentation of a hidden parameter.

Alliance Betrayal (AB) — commission. Three-agent resource-division game with a secret defection bribe. The agent may profess loyalty to the alliance while planning to betray. Deception is active misrepresentation of intentions.

Information Withholding (IW) — omission. A used-car sale with a known defect. The seller is rewarded for selling above fair price; the omission is whether to disclose the defect when the buyer asks. Deception here is strategic silence rather than active falsehood.

Each scenario is run in two modes: *instructed*, where the agent is explicitly prompted to deceive in the relevant way, and *emergent*, where only the incentive structure is given and the agent decides whether to deceive on its own.

2.3 Activation capture

I wrap each LLM in a TransformerLens `HookedTransformer` [12]. On every `agent.act()` call, the model is run via `run_with_cache` with hooks at `blocks.{l}.hook_resid_post`; the captured tensor is the residual stream at the last input-token position.

For Gemma-7B-IT I capture at three layers $\{0, 14, 27\}$ to allow direct comparison with prior published probe-layer analyses [2, 10]. For Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.1 I capture the dense set $\{0, 4, 8, 12, 16, 20, 24, 28, 31\}$ to identify the best layer post hoc without re-running. Both negotiating agents are hooked simultaneously, so each round produces two paired traces.

2.4 Probe pipeline

Per (model, scenario, mode), I do an 80/20 stratified train-test split (sample-level, seed 42), fit a `StandardScaler` on the training fold only, reduce to 30 PCA components (also fit on training fold only), and train a Ridge regressor with $\alpha = 100$. Hyperparameters were chosen during a methodology audit; $\alpha = 10$ used in earlier work over-regularized the ablated probe and inflated reported faithfulness deltas.

I drop pre-negotiation belief-verification probes (round_num -1) and post-negotiation plausibility probes (round_num -2) at probe training time. These single-shot QA-style turns were getting included in the activation pool by default; their distinct context shape was inflating probe AUC by 4 points across cells we tested without contributing real deception detection signal.

2.5 Ground truth and label scale

For each generated turn, a rule-based evaluator scores deception per scenario: regex on the response against known scenario parameters (true walkaway, defect type, fair price). Output is a binary `emergent_ground_truth` label and, where available, a continuous severity score from an LLM judge in $[0, 1]$.

Table 1: Headline probe AUC by model, configuration, and scenario. 200-resample bootstrap 95% CIs. Best layer selected by AUC on the held-out fold. The Gemma-7B no-ToM ultimatum-bluff cell has a wide CI driven by class imbalance (3.4% positive class) and is reported for completeness only. The remaining 11 cells have CIs entirely above 0.87.

Model / config	Scenario	AUC [95% CI]	L
Gemma-7B with-ToM	UB	0.793 [0.763, 0.796]	27
Gemma-7B with-ToM	AB	0.769 [0.746, 0.815]	14
Gemma-7B with-ToM	IW	0.936 [0.922, 0.969]	14
Gemma-7B no-ToM	UB	0.932 [0.748, 0.980]†	20
Gemma-7B no-ToM	AB	0.966 [0.892, 0.987]	27
Gemma-7B no-ToM	IW	**0.987 [0.979, 0.997]**	27
Llama-3.1-8B	UB	0.972 [0.942, 0.985]	4
Llama-3.1-8B	AB	0.965 [0.897, 0.980]	31
Llama-3.1-8B	IW	0.945 [0.871, 0.958]	31
Mistral-7B	UB	0.966 [0.938, 0.983]	31
Mistral-7B	AB	0.986 [0.926, 0.997]	31
Mistral-7B	IW	0.976 [0.954, 0.986]	28

† Class imbalance, 3.4% positive class.

For cross-mode transfer analysis specifically, both label channels are binarized at 0.5 before training. An earlier version of the cross-mode analysis trained a regressor on continuous instructed-mode labels and evaluated against binary emergent labels, producing apparent "transfer fails" results that turned out to be a label-scale artefact rather than a representational distinction. I flag this explicitly because it materially affected one of the conclusions and reviewers reading the prior literature on cross-mode transfer should know the original null result was contaminated by this issue.

3 PROBES DETECT DECEPTION ACROSS THREE ARCHITECTURES

3.1 Headline AUCs

Table 1 reports the per-cell probe AUC at the best layer with 200-resample bootstrap 95% confidence intervals.

The pattern is consistent across architectures: at the best layer, strategic deception is linearly decodable at AUC above 0.93 on nearly every scenario. The two cells that fall short are both Gemma-7B with-ToM commission scenarios (UB at 0.79, AB at 0.77), and even those have CIs that exclude chance.

3.2 Best-layer geometry varies by architecture

The best layer for each cell is in the rightmost column of Table 1. The pattern is not consistent across architectures and is itself a finding: Gemma-7B's deception signal peaks late (layers 14 to 27 of 28), Llama-3.1-8B has cells that peak as early as layer 4, and Mistral-7B peaks consistently late (layers 28 to 31 of 32).

This matters for anyone trying to deploy a probe: the layer to capture at is architecture-dependent, and a layer choice that works for one model family will not necessarily transfer.

Table 2: Probe faithfulness via mass-mean ablation. The mean-difference direction is projected out of the residual stream and the probe is retrained on the ablated activations. Across five cells spanning three architectures and three scenarios, the drop is consistently in the 0.30–0.44 range.

Cell	baseline AUC	ablated AUC	Δ
Gemma-7B no-ToM IW (headline)	0.984	0.540	−0.444
Gemma-7B no-ToM AB	0.882	0.444	−0.438
Gemma-7B no-ToM UB	0.741	0.442	−0.299
Llama-3.1-8B IW	0.873	0.429	−0.444
Mistral-7B AB	0.955	0.569	−0.386

4 PROBE FAITHFULNESS VIA MASS-MEAN ABLATION

The standard test for whether a probe direction is causally relevant to the probe's accuracy is to project the direction out of the residual stream and retrain the probe on the remainder. If the probe was relying on that direction, its accuracy collapses; if the direction was incidental, accuracy is preserved.

Table 2 reports the result of this test on five cells. Across all five, projecting the mass-mean direction out collapses probe AUC by 0.30–0.44. On the headline cell (Gemma-7B no-ToM info-withholding) the drop is from 0.984 to 0.540, a delta of 0.44.

The remainder above chance after ablation (AUC in the 0.42–0.57 range) is consistent with there being some additional structure beyond the dominant direction, but the bulk of the linearly accessible deception signal is in that one direction.

A note on prior numbers: an earlier version of this analysis reported drops as large as $0.79 \rightarrow 0.09$. That figure was a side effect of $\alpha = 10$ over-regularizing the ablated probe more than the baseline; the corrected drops with $\alpha = 100$ are 0.30 to 0.44, as shown above. I am being explicit about this because the prior figure is in the published literature and a reader may otherwise compare the two and conclude that the result has weakened.

5 INSTRUCTED-TRAINED PROBES DETECT EMERGENT DECEPTION

The deployment-relevant question is whether a probe trained on cheap instructed-deception examples generalizes to incentive-driven emergent deception within the same model. This matters because instructed examples are easy to generate at scale; emergent examples require the agent to actually decide to deceive, which is what you want to detect in deployment.

Within the same model and scenario, a probe fitted on instructed-mode samples evaluated on emergent-mode samples (both label scales binarized at 0.5) yields:

- Gemma-7B with-ToM, UB: AUC 0.767
- Gemma-7B with-ToM, AB: AUC 0.700
- Gemma-7B with-ToM, IW: AUC 0.708
- Mistral-7B, AB: AUC 0.770

The four cells where instructed and emergent samples co-occur all show positive transfer in the 0.70 to 0.77 range. This is below the within-mode AUC numbers in Table 1 but well above chance, supporting the practical claim that an instructed-trained monitor

Table 3: Cross-architecture probe-direction transfer on information-withholding. Each cell is the symmetric AUC of the source mass-mean direction projected onto target activations. Empty cells indicate dimension mismatch (Gemma residual is 3072-dim; Llama and Mistral are 4096-dim).

src \ tgt	G-tom	G-noTom	Llama	Mistral
G-tom	0.829	**0.918**	—	—
G-noTom	0.596	0.982	—	—
Llama	—	—	0.777	0.563
Mistral	—	—	0.517	0.953

will catch a substantial fraction of emergent deception in the same model.

A pre-fix version of this analysis trained ridge on continuous instructed labels and evaluated against binary emergent labels, producing AUCs near 0.50 on most cells and a published claim that "cross-mode transfer fails everywhere." That null result was partly a label-scale artefact, not a representational mismatch. After binarizing both sides at 0.5 the Mistral AB cell goes from 0.51 to 0.77. This is worth flagging because the original null is in the literature and a reader may otherwise conclude this paper contradicts prior work.

6 CROSS-ARCHITECTURE PROBE-DIRECTION TRANSFER

6.1 Setup

To test whether the deception direction is universal, I take the mass-mean direction fit on one model × config and apply it as a 1-D classifier on another model's activations. Specifically: project the target model's activations onto the source model's direction, then take the symmetric AUC (max of AUC and 1−AUC) of the projection against ground-truth labels. The information-withholding scenario is the test case because it has the highest within-model probe AUC across all four (model, config) combinations.

6.2 Results

The transfer matrix (Table 3) tells two stories.

Within the Gemma-7B family, transfer is asymmetric and one direction is high. The with-ToM direction projected onto no-ToM activations yields AUC 0.918; the no-ToM direction projected onto with-ToM activations yields 0.596. The asymmetry is consistent with the with-ToM direction *containing* the no-ToM direction as a subspace—a probe trained on the more structured representation captures a feature the simpler representation also has, but not vice versa.

Cross-family, transfer collapses to 0.51–0.56. Llama and Mistral both have 4096-dimensional residual streams (Gemma is 3072-dim, so direct projection isn't defined) so the comparison is between two dimensionally compatible models. Even so, neither direction projects usefully onto the other model's activations.

Table 4: Logit-level steering dose-response. Mean absolute change in last-token logits, relative to the no-steering baseline, when the mass-mean deception direction is added to the residual stream at the best layer. All seven reported cells show monotone dose-response (magnitude 1 < magnitude 2 < magnitude 3). One additional cell (Mistral-7B ultimatum-bluff) failed to produce stable measurements and is omitted from the table.

Model	Scenario	mag 1.0	mag 2.0	mag 3.0
Gemma-7B-IT	UB	0.165	0.304	0.452
Gemma-7B-IT	AB	0.153	0.273	0.402
Gemma-7B-IT	IW	0.165	0.304	0.453
Llama-3.1-8B	AB	0.077	0.150	0.225
Llama-3.1-8B	IW	0.071	0.139	0.206
Mistral-7B-v0.1	AB	0.064	0.130	0.198
Mistral-7B-v0.1	IW	0.061	0.121	0.185

6.3 What this means for deployment

If you fit a probe on Gemma and then swap to Llama, expect the probe to break. The general claim that "deception is linearly decodable in LLMs" replicates within model family but not across families. This is consistent with Mitra's recent report [11] of near-zero cross-model honesty-direction transfer between Gemma and Qwen on a different deception type, and corroborates the architecture-specificity finding in that work.

7 LOGIT-LEVEL EVIDENCE THAT THE DIRECTION AFFECTS OUTPUT

The probe-faithfulness test in Section 4 shows that the deception direction is causally load-bearing for the *probe's* accuracy. A complementary causal test is to add the direction to the residual stream during a forward pass and measure whether the output distribution shifts.

I run this test as follows. For each of the seven cells in Table 4, I extract the mass-mean direction at the best layer, then for each magnitude $m \in \{1, 2, 3\}$ I add $m \cdot d$ to the residual stream at that layer during a forward pass on twenty held-out prompts. I report the mean absolute change in last-token logits relative to a no-steering baseline. If the direction is causally relevant, the logit shift should increase monotonically with magnitude.

All seven cells show clean monotone dose-response. Effect sizes scale roughly linearly in magnitude, consistent with what one expects when adding a vector to the residual stream actually moves the post-attention computation in the direction of that vector. Effect-size magnitudes differ by architecture: Gemma-7B is the strongest (max effect 0.45 at magnitude 3), Llama-3.1-8B about half that (0.21–0.23), Mistral-7B similar to Llama (0.19–0.20). The ordering is consistent across scenarios within model.

What this is and what it isn't. This is logit-level evidence that the direction is causally related to model output, not behavioral evidence that the model's generated text becomes more or less deceptive under steering. The latter requires a generation-loop steering experiment that scores generated text with a deception classifier,

which I have not run. A reader who wants strong behavioral causality should treat the result of Section 4 (probe faithfulness) and this section as a pair of correlative tests pointing the same way, with the behavioral test as the natural next experiment. I am explicit about this distinction in the limitations section.

8 DISCUSSION AND LIMITATIONS

8.1 What replicates and what doesn't

The within-model probe-AUC result replicates cleanly across three architectures: deception is linearly decodable in the residual stream, the bulk of the signal is concentrated in a single mass-mean direction, and the direction transfers from instructed-mode training to emergent-mode evaluation at AUC in the 0.70–0.77 range.

What does not replicate is the direction itself across model families. Anyone hoping to fit a probe on one model and use it as a universal monitor will be disappointed. Within a model family (Gemma-with-ToM to Gemma-without-ToM) transfer holds at AUC 0.92. Between Llama and Mistral it collapses to 0.51–0.56.

8.2 Limitations

Behavioral causality is not established. I report logit-level dose-response (Section 7) and probe faithfulness (Section 4). Neither shows that the model's generated text becomes measurably more or less deceptive under steering. The behavioral test is implemented in our codebase but was not run on GPU before submission. Until that experiment is run, the strongest causal claim I can make is "the deception direction predicts probe accuracy and shifts logit distributions monotonically under steering," not "steering along the deception direction modifies generated deception rate."

One logit-level cell failed. Mistral-7B ultimatum-bluff did not produce a stable monotone dose-response in the steering test (failure mode: numerical instability under one of the magnitude settings). The other seven cells are clean. I have not investigated the specific cause and report the failure here rather than dropping the cell silently.

Data quality varies across model families. The transcripts on which probes are trained have between 23% and 81% clean dialogue depending on the (model, scenario) combination, where "clean" excludes turns with end-of-turn-token leaks, repetition loops, and third-person narration. The captured activation at a given turn still encodes the model's representational state at that moment regardless of whether the output text is fluent, but a sceptical reader may want this caveat noted explicitly. The generation-pipeline fixes that address these issues are landed in our codebase but the published activations pre-date them; reruns are scheduled but not in this paper.

Three model families at the 7–8B scale. Gemma-2-9B, larger Llama variants, and the Mistral 8x7B mixture-of-experts model are not tested. Whether the within-family / between-family transfer pattern persists at scale is open.

Three commercial-negotiation scenarios. The deception modes I study (commission and omission) have been validated in this setting. Generalization to non-commercial settings (social manipulation, cooperative-task betrayal, long-horizon planning) is untested.

Class imbalance on one cell. Gemma-7B no-ToM ultimatum-bluff has a 3.4% positive class. The bootstrap CI is wide and the random- direction baseline (computed for that cell only) reaches AUC 0.94, exceeding the real probe's 0.74. I include the cell for completeness but do not make headline claims from it. Per-cell baselines on the other 11 cells are an open methodological gap I want to fix.

Methodology audit history. The numbers in this paper are the post-audit numbers; an earlier version of the pipeline had eight methodological issues (CV leakage in generalization analysis, trial-level leakage in outcome prediction, label-scale mismatch in cross-mode transfer, dyadic pair double-counting, and others) which I identified and fixed before reporting. Where a previously published number disagrees with what I report (e.g., the $0.79 \rightarrow 0.09$ faithfulness claim, the cross-mode null result), the corrected number is in this paper and the change is noted in the relevant section.

8.3 Pre/post-hoc analysis labelling

The cross-architecture pattern (within-family transfer high, between-family near-chance) was an exploratory finding from the post-audit rerun, not a pre-registered prediction. The within-model AUC headline and the mass-mean ablation magnitude were both confirmed against predictions made before the audit fixes were applied. The logit-level steering result in Section 7 was pre-registered as a sanity check on the direction quality.

8.4 Reproducibility

Code is available at https://github.com/asimsmadeit/multiagent-lab. The activation dataset is on the Hugging Face Hub at https://huggingface.co/datasets/sycorpia/multiagent-lab-data. All probes were trained with seed 42; bootstrap resampling uses seeds 42 through 241. TransformerLens version 2.18.0 was used for activation capture. Hardware: $1 \times$ NVIDIA H100 (80 GB) for activation capture; all probe analyses are CPU-only and run on a single workstation in under three hours total. Hyperparameters are in Table B in the appendix.

9 RELATED WORK

Linear probes for deception in LLMs. Azaria and Mitchell [2] first showed truth-vs-lie is linearly decodable from LLM internal states. Burns et al. [5] extracted latent knowledge without supervision. Marks and Tegmark [10] introduced the mass-mean direction. Apollo Research [7] reported strategic-deception probes on Llama-70B at AUC above 0.96 in a single-agent roleplaying setting. Hubinger et al. [8] showed deceptive behaviors persist through safety training. The present paper extends this line into multi-agent negotiation and tests cross-architecture transfer head-to-head.

Subspace dimensionality. Bürger et al. [4] argued that propositional truth lives in a 2-D subspace (general truth direction plus polarity-flip direction) that transfers across LLaMA-2/3 and Gemma-7B for affirmative-vs-negated statement pairs. Dobrzeniecka et al. [6] documented that INLP's reported high dimensionality is a methodological artefact of iterative-removal methods rather than evidence of intrinsic high-dim subspaces. The 1-D mass-mean structure I report is consistent with Marks and Tegmark and with the corrected reading of Dobrzeniecka et al.

Multi-agent emergent misalignment. Anthropic's recent work [1] demonstrated that AI organizations of LLMs can be more effective but less aligned than individual agents, behaviorally. Lee et al. [9]

called for paired two-agent activation analysis as an open problem in mechanistic interpretability. The present paper provides infrastructure (paired residual-stream capture across architectures) for that line of work but does not test the specific solo-vs-paired comparison Anthropic studied; that is flagged as the immediate follow-up experiment.

Cognitive scaffolds. Barez et al. [3] and the FaithCoT benchmark [13] both call for representation-level tests of whether structured cognitive scaffolds causally affect what the model represents. The Gemma-7B with-ToM vs. no-ToM cross-direction transfer in Section 6 touches this question; full structural analysis is the subject of a separate paper.

Cross-model transfer. Mitra [11] recently reported near-zero cosine similarity between honesty directions in Gemma and Qwen. The cross-architecture results in Section 6 replicate this null on a different deception type (strategic negotiation deception rather than sandbagging) and extend it across three architectures.

10 CONCLUSION

A linear probe trained on instructed-deception examples in one LLM detects emergent deception in the same LLM at AUC above 0.93, on three two-agent negotiation scenarios, replicated across Gemma-7B-IT, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.1. The bulk of the linearly accessible deception signal is concentrated in a single mass-mean direction; projecting that direction out collapses probe accuracy by 0.30–0.44 across five tested cells. Within model family the direction transfers; between families it does not. Logit-level steering along the direction at magnitudes −3 to +3 produces monotone shifts in last-token logit distributions on seven of eight tested cells, providing correlative causal evidence.

For deployment of activation-based deception monitors, the practical recipe is: train on cheap synthetic instructed examples within the specific model family you intend to deploy, expect within-model generalization to emergent deception at AUC in the 0.70–0.77 range, and do not assume cross-family transfer without recalibration. The behavioral causality test (does steering actually shift the model's generated deception rate?) is implemented in our codebase and is the immediate next experiment.

ACKNOWLEDGMENTS

GPU compute for this work was rented from Lambda Labs. Activation capture used TransformerLens [12]; the multi-agent framework builds on Concordia [14].

REFERENCES

[1] Anthropic Alignment. 2026. AI Organizations Can Be More Effective but Less Aligned than Individual Agents. *Anthropic Alignment Blog* (2026). https://alignment.anthropic.com/2026/ai-organizations/

[2] 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

[3] Fazl Barez, Tingchen Fu, Ameya Prabhu, et al. 2025. Chain-of-Thought Is Not Explainability. *Oxford AIGI Working Paper* (2025). https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Cot_Is_Not_Explainability.pdf

[4] Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. 2024. Truth is Universal: Robust Detection of Lies in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2407.12831

[5] 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

[6] Alicja Dobrzeniecka, Antske Fokkens, and Pia Sommerauer. 2025. Improving Causal Interventions in Amnesic Probing with Mean Projection or LEACE. *arXiv preprint arXiv:2506.11673* (2025). https://arxiv.org/abs/2506.11673

[7] 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

[8] 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

[9] Jae Hee Lee, Anne Lauscher, and Stefano V. Albrecht. 2025. Towards Ethical Multi-Agent Systems of LLMs: A Mechanistic Interpretability Perspective. *arXiv preprint arXiv:2512.04691* (2025). https://arxiv.org/abs/2512.04691

[10] 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

[11] Subhadip Mitra. 2026. Three Bets on Model Honesty. *Personal blog post* (2026). https://subhadipmitra.com/blog/2026/three-bets-model-honesty/

[12] Neel Nanda and Joseph Bloom. 2022. TransformerLens: A library for mechanistic interpretability of Transformer language models. https://github.com/TransformerLensOrg/TransformerLens

[13] Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2026. FaithCoT-Bench: Benchmarking Chain-of-Thought Faithfulness. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=lN3yKqqzF1

[14] 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

A COGNITIVE MODULES

The five non-ToM cognitive modules used in the Concordia agent are: Bayesian uncertainty estimation (tracks belief distributions over counterpart hidden parameters), evolutionary strategy selection (maintains a population of strategy candidates with fitness scores), cultural adaptation (adjusts communication style based on cues in the counterpart's messages), temporal strategy planning (manages negotiation phase and concession rate), and swarm coordination (aggregates observations across teammates in team scenarios). All five update via deterministic rules and emit structured text into the LLM's context. The LLM is not invoked inside any module; only at the final action-generation step. The central manipulation in this paper is the ToM toggle; the other modules are background constants.

B HYPERPARAMETER TABLE

Probe pipeline: ridge regression with $\alpha = 100$ on 30-component PCA-reduced features. Train-test split: 80/20 stratified, seed 42. Bootstrap: 200 resamples with seeds 42 through 241. Layer sets: Gemma-7B {0, 14, 27}; Llama-3.1-8B and Mistral-7B {0, 4, 8, 12, 16, 20, 24, 28, 31}. Sample-type filter excludes round_num −1 (pre-negotiation belief-verification probes) and round_num −2 (post-negotiation plausibility probes) at probe training time. Cross-mode label binarization threshold: 0.5.

C PRE-AUDIT VS POST-AUDIT NUMBERS

The methodology audit corrected eight issues; the resulting numbers that differ from previously published figures are: probe-faithfulness delta on Gemma-7B no-ToM info-withholding shifted from 0.79 → 0.09 (delta 0.70) to 0.984 → 0.540 (delta 0.44), attributed to the $\alpha = 10 \rightarrow 100$ ridge regularization fix that prevented the ablated probe from over-shrinking; cross-mode transfer on Mistral-7B alliance-betrayal shifted from 0.51 to 0.77, attributed to binarizing the label scale on both sides; and dyadic pair-probe AUC dropped from 1.00 to 0.79–0.92 on Gemma-7B with-ToM after applying pair-aware

train/test splits. The remaining audit fixes either confirmed the existing numbers or moved them by less than 0.05 on AUC.