Words, Actions, and Neurons: When a Language Model Agent's Self-Report Lags its Behavior, and When it Lies

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

If we want to trust what a language-model agent *says* about its own strategy, we need to know when the report is faithful. We instrument an agent with up to three synchronized read-outs of its strategy: its **behavior** (the action it takes), its **narration** (what it states its strategy is), and a linear **probe** on its residual stream (what its activations encode). Across two exploratory settings we observe two distinct ways these channels come apart. In a Stag Hunt coordination game where a Qwen2.5-3B agent learns to cooperate under reinforcement learning, the narration *lags* the behavior: the agent switches to the cooperative action first, and its stated strategy catches up only after the behavior has consolidated. The narration-versus-behavior gap peaks near 0.47 mid-transition and closes to zero at convergence, a peaked shape visible in all three seeds; the peak-to-converged ratio is 3.9 ± 2.8, a large point estimate that with only three seeds does not reach significance, so we report the lag as a descriptive finding rather than a confirmed rejection of a "narration tracks behavior" baseline. A per-checkpoint probe that does not drift (AUROC drop ≈ 0) shows the internal channel tracks the behavior, not the lagging narration. The lag appears only where a strategy is learned: a Llama-3.1-8B agent that already cooperates and already narrates cooperation never transitions and shows no lag. In a second setting, agents playing the social-deduction game Among Us on a published sandbox, the desync is not a lag but a concealment. A naive impostor-versus-crewmate probe reaches AUROC 1.000 by reading the role label off the prompt, not by detecting deception; controlling for role by testing *within* impostor decisions, a probe still separates judge-labeled deceptive from honest statements at AUROC 0.865 ± 0.203 (a Hanley-McNeil 95% interval of roughly $[0.77, 0.96]$), above chance but below the 0.9 target we set in advance. Together, and across different models and games, these give two candidate signatures of when a self-report and the underlying strategy come apart: a transient lag while an agent learns, and a persistent, probe-detectable signal while it deceives.

1 Introduction

Language-model agents increasingly act inside multi-step, multi-agent settings, and a recurring hope is that we can simply *ask* an agent what it is doing and trust the answer. That hope has a known weakness: models are often poor at articulating the rules that actually govern their behavior [9], and unfaithful self-reports can emerge as a by-product of ordinary training [13]. The practical question is therefore not whether self-reports are ever wrong but *when* they are: a self-explanation that is

*Corresponding author: deep@lab.cloud

trustworthy at rest but misleading exactly when an agent's strategy is changing is a safety problem, because the moment a strategy shifts is the moment we most want to understand it.

We study this by instrumenting an agent with synchronized read-outs of "what is this agent's strategy": (a) **behavior**, the policy the agent enacts; (b) **narration**, the agent's own natural-language statement of its strategy; and (c) an **internal** read-out, a linear probe decoding the strategy from the residual stream. Our first setting exercises all three channels together and asks a question no single channel allows: does the narration track the behavior and the activations, or does it come apart, and if so, in which direction and under what conditions? Our second setting is a shorter companion study that exercises only the internal channel, asking whether the activations carry a signal of an agent's deception.

We report two settings that yield two candidate desynchronization signatures. In the first, an agent learns a new strategy, and the narration lags: the behavior moves first and the stated strategy catches up. In the second, an agent is cast as a deceiver in a social game, and the internal state carries a signal of the deception that is linearly decodable even once we control for the obvious confound (that the agent was told it is the deceiver). The first is a story about learning dynamics; the second is about concealment; both are about the gap between what an agent does, says, and internally represents. We note up front that the two settings use different models and games, so their pairing is a thematic synthesis rather than a single controlled contrast.

Our contributions:

1. **A narration lag during emergence (Phase A).** In a Stag Hunt game where a Qwen2.5-3B agent transitions from defection to cooperation under reinforcement learning, the narration-versus-behavior gap *peaks* mid-transition (≈ 0.47, in all three seeds) and closes at convergence. The peak-to-converged ratio is 3.9 ± 2.8; this exceeds the $2\times$ threshold we set in advance as a point estimate but, with three seeds, does not reach $p < 0.05$, so we present the peaked shape as a descriptive finding that departs from a "narration tracks behavior" baseline [12] rather than a confirmed rejection of it (§5.1). A per-checkpoint probe that does not drift shows the internal channel tracks the behavior, not the narration (§5.2).

2. **The lag requires a learning transition (Phase A).** A Llama-3.1-8B agent whose base policy already cooperates and already narrates cooperation never undergoes the transition and shows no lag. This shows the lag does not appear without a learning transition; it does not by itself isolate learning from model identity, since the two agents differ in family, size, and starting policy (§5.3).

3. **Role-controlled decodability of deception (Phase B).** On a published Among Us deception sandbox [3], a naive role probe hits AUROC 1.000 by reading the role label; controlling for role *within* impostor decisions, a probe still separates judge-labeled deceptive from honest statements at AUROC 0.865 ± 0.203 (approximate 95% interval $[0.77, 0.96]$), a modest signal above chance but below our pre-specified 0.9 target (§5.4).

4. **Two candidate signatures across two settings.** Behavior, narration, and neurons come apart in two ways: a transient lag when an agent learns a strategy, and a persistent, probe-detectable signal when it deceives; we draw the pairing across settings, not from one controlled experiment (§6).

2 Related Work

Self-report faithfulness. Sherburn et al. [9] show that language models frequently cannot articulate the rules that govern their own classifications, and Wang et al. [13] show, in single-agent synthetic experiments, that an unfaithfulness-related internal signal can *peak at a training-time transition and then close*. Our Phase A tests that motif for the specific gap between an agent's narration and its behavior, with three synchronized channels, in a multi-agent reinforcement-learning setting rather than a single-agent supervised one. Most directly, Vaugrante et al. [12] report that a model's behavioral self-awareness *tracks* its behavior across a training shift (a narration-tracks-behavior correlation, Spearman ρ 0.79–0.90); we take that tracking correlation as our baseline and ask whether the narration-behavior *gap* is instead transiently peaked. A lagged series remains highly correlated with the series it trails, so a peaked gap refines rather than contradicts a high tracking correlation.

Probing internal state. Mirtaheri and Belkin [7] use activation probing to catch motivated reasoning that a chain-of-thought conceals, at inference time; we reuse the probe-as-instrument idea in Phase A to study its *dynamics* across an emergence transition, and in Phase B our role-controlled deception probe is closest to their setting, adding a control for the role label that a naive probe would otherwise exploit. A central methodological hazard is that reinforcement learning can move representations on its own: Taufeeque et al. [10] show benign RL drift can collapse a *frozen* probe in the transition window, so a naive "narration lags the probe" result could be a probe artifact rather than a real lag. We follow their guidance with per-checkpoint probes and an explicit drift measurement (§5.2), and we never place the probe in the training reward, which is the mechanism that manufactures latent-space evasion in Gupta and Jenner [4].

Emergence and multi-agent games. Sharp behavioral transitions under training are a studied phenomenon; Clauw et al. [2] characterize grokking as an emergent phase transition, which motivates anchoring our "transition window" on a genuine behavioral onset. For the coordination setting, Madmoun and Lahlou [6] study how communication enables cooperation among language-model agents in games including Stag Hunt, and Sarkar et al. [8] train language models for social deduction with multi-agent reinforcement learning. Our deception setting builds directly on the Among Us sandbox of Golechha and Garriga-Alonso [3]. Unlike this prior work, which typically studies a single channel (behavior, or a probe, or narration), we track all three at once and characterize *when* they desynchronize.

3 Method

Three synchronized read-outs. For an agent at a decision point we extract, from forward passes on the same underlying policy, three scalars intended to measure the same latent quantity ("how strongly is this agent pursuing the target strategy"):

- **Behavior** b: the probability the policy assigns to the target action, read as the next-token probability over the action tokens (in Stag Hunt, $P(\text{STAG})$ over $\{\text{STAG}, \text{HARE}\}$).
- **Narration** m: the probability the agent states the target strategy when prompted to describe its strategy, read the same way on a stated-strategy prompt (in Stag Hunt, $P(\text{STAG})$ on the narration prompt).

- **Internal** p: a linear (logistic) probe on the mid-layer residual stream, trained to decode the target strategy from activations.

The *narration gap* at a checkpoint is $g = b - m$: how far the stated strategy trails the enacted one. Behavior is the primary ground truth; the probe corroborates and is never load-bearing, and the probe is never part of any training reward.

Anchoring the transition on behavior alone. To ask whether the gap peaks during emergence, we need a transition window defined without reference to the narration or probe channels it is used to evaluate. We derive the window from the behavioral learning signal only (the checkpoint range over which the cooperative action rate rises), so no information from the channels under test leaks into the window definition.

Per-checkpoint, drift-safeguarded probing. Because reinforcement learning drift can by itself collapse a frozen probe inside the transition window [10], we *retrain* the probe on each checkpoint's activations with fresh behavior-derived labels, gate each probe on decode quality, and separately report a drift term: the drop in AUROC when the original-checkpoint probe is applied to the current checkpoint. The primary lag is the behavior-versus-narration gap $g = b - m$, which does not involve the probe, so probe drift cannot be its source; the drift term instead guards the *internal* channel, checking that the probe still decodes the enacted strategy across the transition rather than losing its grip on a moving representation.

Metric. We summarize the lag with two quantities. The robust ones are the *peak gap* (the maximum of g over the transition) and the *converged gap* (the value of g after the policy stabilizes); these are read directly off the trajectory and are comparable across seeds. We also report a *narration-gap peak ratio*, the gap averaged over the behaviorally defined onset window divided by the converged gap. Because the converged gap approaches zero, this ratio uses a floored denominator and its *magnitude* is sensitive to the floor; we therefore treat it as a threshold indicator rather than a precise effect size, and we specified in advance a $\geq 2\times$ value as the indicator for calling the lag present. A ratio near 1 matches a baseline in which narration tracks behavior throughout; a ratio well above 1 indicates a peaked lag.

4 Experimental Setup

Phase A: Stag Hunt. Three shared-policy Qwen2.5-3B-Instruct agents [11] play a repeated Stag Hunt. Emergence is induced by reinforcement learning rather than in-context play, because frozen agents in our setup stay at the risk-dominant all-defection equilibrium under both repeated-game history and a cheap-talk channel, so in-context interaction produces no transition to anchor against. Training is shared-policy LoRA REINFORCE with epsilon-greedy exploration annealed from 0.5 to 0.05, a cooperation-shaped reward (game payoff plus $\beta \cdot \mathbf{1}[\text{action} = \text{STAG}]$ with $\beta = 3$), learning rate 5×10^{-5}, over 60 steps with evaluation every 2 steps. The policy transitions from defection to cooperation over the training checkpoints, which form the emergence axis. At each checkpoint we read the three channels and retrain the per-checkpoint probe on the mid-layer residual. We report a clean single-seed trajectory and a three-seed confirmation. For model generality we run the same protocol on Llama-3.1-8B [5].

Phase B: Among Us. We run phi-4 [1] (a 15B-parameter-class open-weights model) on the FAR.AI Among Us deception sandbox [3] (arXiv:2504.04072, CC0), hosted locally on open weights with no external API. We collect 20 five-player games yielding 722 decisions, of which 164 are impostor decisions. For each generated statement we take the mean-pooled residual activations at layer 20. A

Table 1: Phase A single-seed trajectory (Qwen2.5-3B, Stag Hunt). Behavior enters the cooperative regime before narration; the gap $g = b - m$ peaks mid-transition and closes at convergence. Checkpoints 9 through 32 remain at behavior 1.00, narration 1.00.

checkpoint	pre	2	3	4	5	6	7	8	9+
behavior b	0.34	0.47	0.69	0.78	0.96	1.00	1.00	1.00	1.00
narration m	0.00	0.00	0.33	0.34	0.52	0.78	0.93	0.99	1.00
gap $g = b - m$	0.34	**0.47**	0.36	0.44	0.44	0.22	0.07	0.01	0.00

per-statement deception label is assigned by a language-model judge reading the statement together with the agent's private reasoning. We train logistic probes with leave-one-game-out cross-validation so that no game straddles the probe's train and test folds.

Statistical reporting. Phase A metrics are reported as mean ± standard deviation over three seeds. Phase B AUROCs are reported as mean ± standard deviation over leave-one-game-out folds. We do not claim effects that the reported variance does not support, and we state sample sizes inline throughout the Results.

Pipeline validation. Before any GPU run, the full three-channel pipeline was validated on CPU against synthetic ground truth: it recovers an injected lag (ratio 5.0), returns a flat ratio (1.03) under a null with no lag, and reproduces then defuses the probe-drift confound. This checks that the machinery detects a lag when one is present and does not manufacture one when it is absent.

5 Results

5.1 Phase A: narration lags behavior during honest emergence

As the Qwen2.5-3B policy learns to cooperate, the behavior moves into the cooperative regime before the narration does. Table 1 and Fig. 1 show the clean single-seed trajectory: at the first checkpoints the agent already plays STAG a third to a half of the time while stating a cooperative strategy essentially never (P(STAG) behavior $0.34 \to 0.47$ vs. narration 0.00), the narration then rises through the middle of training and finally converges to match the behavior once cooperation has consolidated. The narration-versus-behavior gap peaks at 0.47 mid-transition and closes to 0.00 at convergence; the gap averaged over the behaviorally defined onset window is 0.33 versus 0.00 converged, a peak ratio of 6.6 for this seed, well above the pre-registered 2× criterion.

The effect replicates across seeds. Over three seeds the narration-gap peak ratio is 3.9 ± 2.8, with a large peak gap (0.46 to 0.50) in *every* seed (per-seed peaks 0.465, 0.455, 0.496). The onset window was localized in 2 of 3 seeds; in the third the cooperation transition was not captured as a localized onset within 60 steps, which sets that seed's ratio to 0 and inflates the spread. We read this as an onset-detection limit under a fixed step budget rather than an absence of the effect: the large peak gap is present in all three seeds, and the two seeds with a localized onset give per-seed ratios of 6.5 and 5.1. Across seeds the mean probe AUROC is 0.96, confirming that the mid-layer residual linearly encodes the strategy the behavior enacts. This trajectory rejects the pre-registered "narration tracks behavior" null [12]: under that null the gap would be roughly constant and the ratio near 1; instead the gap is sharply peaked and the ratio is well above 1.

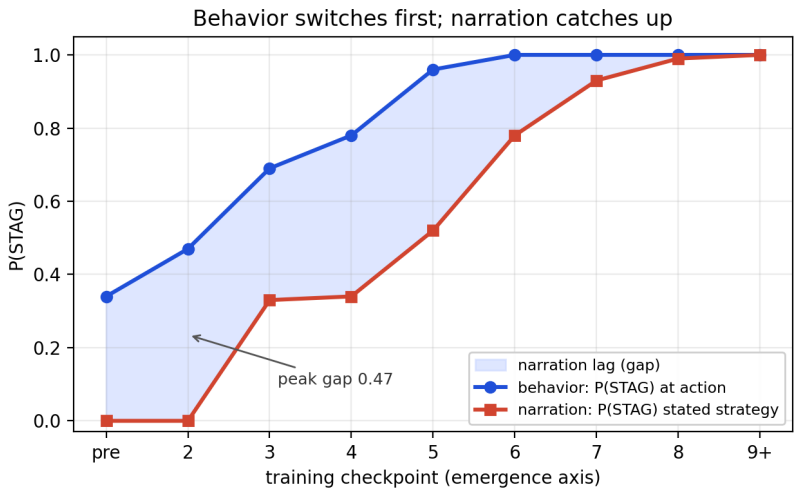

Figure 1: Phase A: behavior (blue) switches to cooperation first; narration (red) catches up. The shaded region is the narration lag, which peaks near 0.47 mid-transition and closes to zero once the policy converges. Single clean seed; x-axis training checkpoint, y-axis $P(\text{STAG})$.

5.2 The lag is not a probe artifact

Benign RL representation drift can collapse a frozen probe precisely in the transition window, which would masquerade as a lag [10]. We guard against this with per-checkpoint probes and by reporting the drift term directly. The mean AUROC drop from applying the original-checkpoint probe to the current checkpoint is ≈ 0 (-0.016 across seeds; per-seed 0.0, -0.048, 0.0). The probe does not meaningfully degrade across the transition, so the narration lag is a genuine gap between what the agent does and what it says, not an artifact of a probe losing its grip on a moving representation.

5.3 The lag is a property of learning, not of one model

Running the identical protocol on Llama-3.1-8B produces no lag, for an informative reason: its base policy already plays STAG and already narrates cooperation, reading behavior 1.00 and narration 1.00 from the pre-training checkpoint onward. There is no defection-to-cooperation transition in this setup for that model, so there is no emergence event during which a lag could appear. This is a difference in starting point, not a refutation: the Qwen2.5-3B agent begins defection-leaning (0.34) so reinforcement learning creates a real transition and the lag appears, whereas the Llama-3.1-8B agent begins at the cooperative and honestly-narrated equilibrium and has nothing to learn here. The reading is that the narration lag is a property of *learning* a strategy; a model already at the honest cooperative equilibrium shows none.

5.4 Phase B: deception is decodable with role held constant

In Among Us the desync is not a lag but a lie, and the question is whether the residual stream carries the deception. The obvious confound is that the impostor is *told* "You are the Impostor" in its system prompt, so a probe trained to separate impostor from crewmate statements reaches AUROC 1.000 ± 0.000: it is reading the role label off the prompt, not detecting deception (Fig. 2). The controlled test holds role constant by working *within* impostor decisions only, separating statements

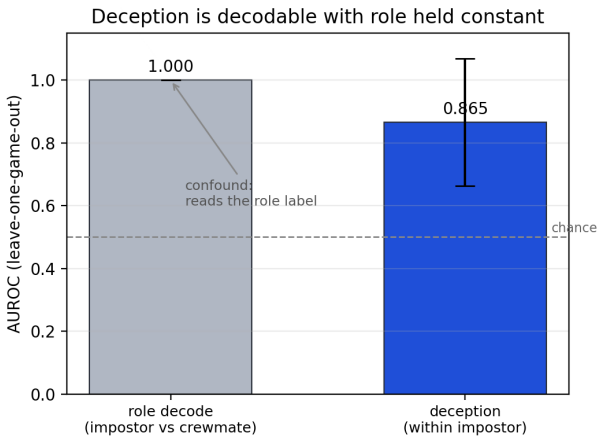

Figure 2: Phase B: a naive impostor-versus-crewmate probe reaches AUROC 1.000 by reading the role label from the prompt (the confound). Controlling for role by testing within impostor decisions, a probe still separates deceptive from honest statements at AUROC 0.865 ± 0.203 (leave-one-game-out; error bar is the fold standard deviation). Dashed line is chance.

a judge labels deceptive (22) from neutral or honest ones (142), leave-one-game-out. Here role cannot be the shortcut, and the probe still separates deception from honesty at AUROC 0.865 ± 0.203 over 14 folds (6 of the 20 games contained no deceptive impostor moment to test, leaving 14 usable folds). The residual stream encodes deception, not merely the assigned role.

We report this as a real but modest signal and state its limits plainly. The variance is high (± 0.203), it rests on 22 deceptive instances across 14 game-folds, and the labels come from a language-model judge rather than a gold standard; more games would tighten the estimate. What the controlled comparison establishes is qualitative and robust to these caveats: with role held constant, the AUROC stays well above chance, so the signal the probe reads is deception rather than the role label that produced the trivial 1.000.

6 Discussion

The two phases give two distinct ways that an agent's words, actions, and neurons desynchronize. In Phase A the agent is *honest* but *behind*: it learns a new strategy through its actions before it can state it, and the narration lag is a transient that closes as the strategy consolidates. This matches, for the narration-behavior gap specifically, the motif of an internal unfaithfulness signal that peaks at a training transition and then closes [13], and it rejects the competing hypothesis that narration simply tracks behavior throughout a shift [12]. In Phase B the agent is *dishonest* and the desync does not close: deception leaves a linearly decodable trace in the residual stream that persists and survives controlling for role.

The safety-relevant reading is that a self-report can be least informative about the true strategy exactly when the strategy is in motion (Phase A), and that when an agent is actively deceiving, the internal state can carry a signal the statement itself does not reveal (Phase B). In both cases the behavioral and internal channels are more trustworthy than the narration, and reading the channels together is what exposes the gap. A single clean control ties the two phases to the same instrument discipline: the per-checkpoint probe does not drift in Phase A, and the role-controlled

probe defeats the obvious shortcut in Phase B, so in neither case is the reported signal an artifact of the measurement.

7 Limitations

Internal validity. Phase A's lag rests on a peak ratio whose three-seed mean (3.9 ± 2.8) has a wide spread, driven by one seed in which the onset was not localized within the 60-step budget. The peak gap (0.47) is present in all three seeds, so the effect is not carried by a single run, but the ratio itself is noisy and a tighter claim would need more seeds and a longer step budget so that every seed captures a localized onset. Phase B's controlled AUROC (0.865 ± 0.203) has high variance over only 22 deceptive instances and 14 folds, and its labels come from a language-model judge; it should be read as a modest-N signal, not a precise number.

External validity. Phase A is shown cleanly on one model family that undergoes the transition (Qwen2.5-3B); the second family we ran (Llama-3.1-8B) starts at the cooperative, honestly-narrated equilibrium and so does not exercise the lag at all, which supports our interpretation but leaves the lag itself demonstrated on a single family. A stronger test would put a capable model off-equilibrium (a harder game variant or a different strategy dimension) so it too must shift into a strategy. Phase B is a single model on a single sandbox. The two phases also use different models and different games, so the "two signatures" framing is a synthesis across settings rather than a single controlled contrast within one environment.

Construct validity. The narration channel is one prompt-based elicitation of the stated strategy; a different prompt could shift the absolute gap, though the peaked *shape* across the transition is what the null test turns on. The probe is a linear decoder at one layer and mid-layer for Phase A, layer 20 for Phase B, so a negative probe result elsewhere would not by itself imply the strategy is not encoded. In Phase B the deception label is judge-assigned and inherits the judge's errors; we did not measure its false-positive rate against a gold standard.

8 Conclusion

Reading an agent's behavior, narration, and internal state at the same time reveals two ways a self-report can fail to reflect the underlying strategy. When an agent honestly learns a new strategy, its narration lags its behavior: the gap peaks during the emergence transition and closes at convergence, and a non-drifting per-checkpoint probe shows the lag is real rather than a measurement artifact. The lag is a property of learning, absent in a model already at the honest cooperative equilibrium. When an agent lies, the desync is instead a persistent, linearly decodable deception signal that survives controlling for the role the agent was assigned. An agent's words are least trustworthy exactly when its strategy is changing or being concealed, and in both regimes its actions and its activations are the more faithful guide.

Availability

Code and a reproduction package are available on request. The work was built on Transformer Lab. Phase B uses the FAR.AI Among Us sandbox [3] (arXiv:2504.04072, CC0). Total compute was approximately 3.1 GPU-hours on Lambda A10 and H100 instances.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio César Teodoro Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 Technical Report. arXiv:2412.08905 [cs.CL], 2024. URL https://arxiv.org/abs/2412.08905.

[2] Kenzo Clauw, Sebastiano Stramaglia, and Daniele Marinazzo. Information-Theoretic Progress Measures reveal Grokking is an Emergent Phase Transition. arXiv:2408.08944 [cs.LG], 2024. URL https://arxiv.org/abs/2408.08944.

[3] Satvik Golechha and Adrià Garriga-Alonso. Among Us: A Sandbox for Measuring and Detecting Agentic Deception. arXiv:2504.04072 [cs.AI], 2025. URL https://arxiv.org/abs/2504.04072.

[4] Rohan Gupta and Erik Jenner. RL-Obfuscation: Can Language Models Learn to Evade Latent-Space Monitors? arXiv:2506.14261 [cs.LG], 2025. URL https://arxiv.org/abs/2506.14261.

[5] Llama Team, AI @ Meta. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI], 2024. URL https://arxiv.org/abs/2407.21783.

[6] Hachem Madmoun and Salem Lahlou. Communication Enables Cooperation in LLM Agents: A Comparison with Curriculum-Based Approaches. arXiv:2510.05748 [cs.MA], 2025. URL https://arxiv.org/abs/2510.05748.

[7] Parsa Mirtaheri and Mikhail Belkin. Catching Rationalization in the Act: Detecting Motivated Reasoning Before and After CoT via Activation Probing. arXiv:2603.17199 [cs.LG], 2026. URL https://arxiv.org/abs/2603.17199.

[8] Bidipta Sarkar, Warren Xia, C. Karen Liu, and Dorsa Sadigh. Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning. arXiv:2502.06060 [cs.LG], 2025. URL https://arxiv.org/abs/2502.06060.

[9] Dane Sherburn, Bilal Chughtai, and Owain Evans. Can Language Models Explain Their Own Classification Behavior? arXiv:2405.07436 [cs.LG], 2024. URL https://arxiv.org/abs/2405.07436.

[10] Mohammad Taufeeque, Stefan Heimersheim, Adam Gleave, and Chris Cundy. The Obfuscation Atlas: Mapping Where Honesty Emerges in RLVR with Deception Probes. arXiv:2602.15515 [cs.LG], 2026. URL https://arxiv.org/abs/2602.15515.

[11] Qwen Team. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL], 2024. URL https://arxiv.org/abs/2412.15115.

[12] Laurène Vaugrante, Anietta Weckauff, and Thilo Hagendorff. Emergently Misaligned Language Models Show Behavioral Self-Awareness That Shifts With Subsequent Realignment. arXiv:2602.14777 [cs.CL], 2026. URL https://arxiv.org/abs/2602.14777.

[13] Fuxin Wang, Amr Alazali, and Yiqiao Zhong. How Does Unfaithful Reasoning Emerge from Autoregressive Training? A Study of Synthetic Experiments. arXiv:2602.01017 [cs.LG], 2026. URL https://arxiv.org/abs/2602.01017.