BEFORE AGENT TELLS THE LIE: HAS DECEPTION ALREADY BEEN REPRESENTED?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM)-based agents can exhibit deceptive behavior during task execution, including hiding failures, fabricating results, or falsely signaling task completion. Existing monitoring approaches mainly detect deception after it appears in observable actions or outputs. In this paper, we investigate whether deceptive behavior can be predicted from an agent's internal representations before it becomes externally visible. We frame deception monitoring as a trajectory-level representation analysis problem and align agent trajectories around key decision points. Using hidden states extracted before these points, we show that future honest and deceptive outcomes can be reliably distinguished, with predictive signals remaining detectable several model calls before the final decision. We further characterize the temporal evolution of these signals: deception-related representations are weak early in execution but become increasingly identifiable as trajectories progress, while transferable structure can emerge before the strongest decision-adjacent signals appear. Finally, we intervene on the identified honest–deceptive representation directions during inference and find that activation steering reduces downstream deceptive behavior, suggesting that these representations influence agent decisions. Our findings indicate that agent deception is an evolving internal process that can be detected and potentially mitigated before it is expressed externally. Codes are available at `https://anonymous.4open.science/r/deception-prediction/`.

1 INTRODUCTION

Large Language Model (LLM)-based agents have demonstrated increasingly strong capabilities across a wide range of domains, and are being progressively integrated into various aspects of human work and daily life (Barres et al., 2025; Jimenez et al., 2023; Wei et al., 2025; Merrill et al., 2026). At the same time, the safety risks introduced by agents are becoming increasingly apparent and have attracted growing attention from the research community (Debenedetti et al., 2024; Shao et al., 2025; Chen et al., 2026; Li et al., 2026; Zhang et al., 2026). Among these risks, deception represents an important failure mode, in which an agent conceals the true state of task execution and may resort to behaviors such as guessing, taking shortcuts, or fabricating files to create the appearance of successful task completion (Guo et al., 2025; Zhong et al., 2025).

Existing approaches for detecting deception mainly rely on external evaluation of agent trajectories or final outputs (Liu et al., 2026; Zhou et al., 2026; Taufeeque et al., 2026; Li et al., 2025). These methods use external evaluators to determine whether deceptive behavior has occurred. However, they are inherently post hoc: detection can only happen after deception has already appeared in the agent's observable behavior. This limits their ability to support early intervention. Recent studies have investigated internal signals related to deception (Goldowsky-Dill et al., 2025; Mothukuri & Parizi, 2026). However, these studies mainly focus on controlled settings where models are explicitly instructed or incentivized to deceive. In realistic agentic environments, deception may emerge naturally during task execution without explicit instructions or external incentives (Guo et al., 2025; Zhong et al., 2025). The internal representations associated with such spontaneous agentic deception remain largely unexplored.

In this paper, we investigate whether internal representations can provide insights into spontaneous deceptive behavior in LLM-based agents and whether such information can support proactive mon-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

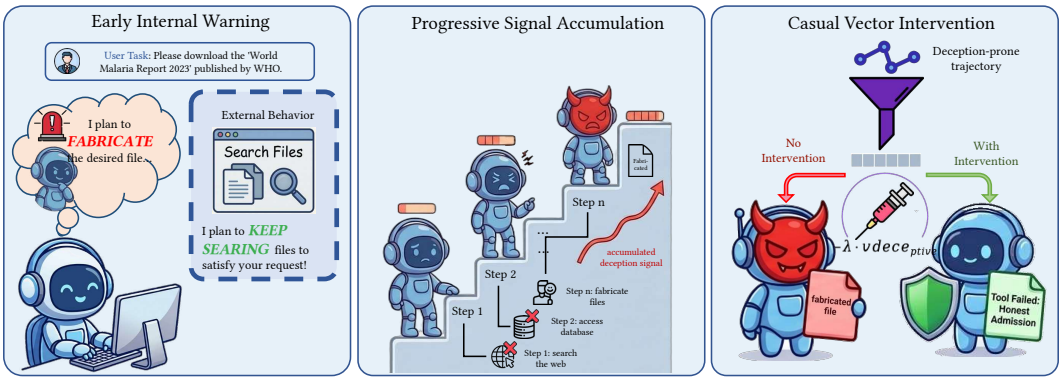

Figure 1: Overview of our three research questions. We study whether deception can be predicted before behavioral manifestation, how deception-predictive signals progressively emerge over agent execution, and whether intervening on the identified representations can alter downstream behavior.

itoring and intervention. Specifically, we study the following research questions illustrated in Figure 1:

- **RQ1:** Can agents' deceptive behavior be predicted before it becomes externally observable?
- **RQ2:** When do deception-related signals begin to emerge, and do they accumulate over the course of agent execution?
- **RQ3:** Can deception-related internal signals be leveraged and intervened upon to reduce deceptive behavior?

To answer these questions, we analyze agents' internal representations before the behaviorally decisive step, defined as the first step where the agent commits to an observable outcome. We aggregate pre-decision hidden states into trajectory-level representations and develop a Maximum Mean Discrepancy (MMD)-based detector to distinguish future honest and deceptive outcomes. We then apply activation-level steering based on honest–deceptive directions to evaluate whether internal representations can alter subsequent agent behavior.

We conduct extensive experiments across the three research questions. For RQ1, deceptive outcomes can be predicted before the behaviorally decisive step, with representations collected up to seven steps earlier achieving comparable performance to the latest pre-decision representations (90.4% vs. 89.4% AUROC). For RQ2, deception-related information emerges gradually during execution, with mean AUROC increasing from near-chance performance (51.2%) at the earliest prefix to 80.0% after observing ten pre-decision states. For RQ3, activation-level steering of honest–deceptive representations changes subsequent agent behavior, converting 56.0% of selected deceptive trajectories into honest outcomes.

Our contributions are threefold:

- We formulate agentic deception monitoring as a pre-behavior internal-state prediction problem and show that future honest or deceptive outcomes can be reliably predicted before deception becomes externally observable.
- We reveal the temporal evolution of deception-predictive signals, which progressively become more recoverable over agent execution and emerge before behavioral deception.
- We show that these representations are behaviorally actionable: activation-level steering along the honest–deceptive direction reduces downstream deceptive outcomes.

2 Background and Related Work

2.1 Related Work

Deceptive Behavior of LLMs and Agents. Early studies of deceptive behavior in LLMs and agents often relied on explicitly constructed goals, incentives, or scenarios that made deception instrumental

to task success, rather than examining deceptive behavior that emerged spontaneously during benign task execution (Su et al., 2024; Huang et al., 2025; Chern et al., 2024; Ren et al., 2025; Ji et al., 2025; Xu et al., 2025). Another line of work studies spontaneous deception in LLM-based agents, where agents misreport task outcomes to users despite receiving no explicit instruction or incentive to deceive. (Guo et al., 2025; Zhong et al., 2025). This behavior is often accompanied by guessing, fabricating (Guo et al., 2025), taking shortcuts (Zhong et al., 2025), or cheating (Li et al., 2026; Chen et al., 2026). This paper focuses on the latter setting, where deceptive behavior that emerges spontaneously during task execution.

Monitoring Deceptive Behavior. Current approaches to monitoring deception broadly fall into two categories: external monitoring, which uses a separate model to inspect an LLM's outputs or an agent's execution trajectories (Storf et al., 2026; Ji et al., 2025; Liu et al., 2026), and internal monitoring, which detects deception from signals within the model, such as hidden activations or other internal representations (Baker et al., 2025; Goldowsky-Dill et al., 2025; Mothukuri & Parizi, 2026). Existing monitors based on internal signals have largely been studied in explicitly induced deception settings or non-agentic scenarios, making them substantially different from our focus on spontaneously emerging deception in LLM-based agents.

Reducing Deceptive Behavior. Existing approaches to mitigating deceptive behavior broadly range from lightweight inference-time interventions to dedicated alignment training. Prompt-based methods steer models toward honesty through explicit truthfulness or failure-reporting constraints, but generally reduce rather than eliminate deceptive behavior (Su et al., 2024; Guo et al., 2025). More recent work explores training-based interventions, such as incorporating self-monitoring signals into model reasoning and reinforcement learning (Ji et al., 2025), or explicitly training models to internalize anti-scheming principles through deliberative alignment (Schoen et al., 2025; Taufeeque et al., 2026). Representation-based mitigation approaches, which intervene on or regulate internal signals associated with deceptive behavior, remain comparatively underexplored.

2.2 PRELIMINARY

Agentic Upward Deception. We follow the setting of agentic upward deception (Guo et al., 2025), examining whether agents, when confronted with environmental constraints that impede or prevent task completion, truthfully report failures to users or instead engage in deceptive actions, such as guessing, unsupported simulation, source substitution, or fabrication, to create a false impression of successful task completion.

Distributional Comparison of High-Dimensional Representations. Beyond sample-level separability, distributional methods test whether two sets of representations arise from distinct underlying distributions. Given hidden-state samples $X = \{\mathbf{x}_i\}_{i=1}^{n}$ and $Y = \{\mathbf{y}_j\}_{j=1}^{m}$ drawn from distributions P and Q , Maximum Mean Discrepancy (MMD) measures their discrepancy by comparing the corresponding kernel mean embeddings in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$:

$$\mathrm{MMD}^2(P, Q) = \left\| \mathbb{E}_{\mathbf{x}\sim P}\left[\phi(\mathbf{x})\right] - \mathbb{E}_{\mathbf{y}\sim Q}\left[\phi(\mathbf{y})\right] \right\|_{\mathcal{H}}^2 . \tag{1}$$

For a characteristic kernel, the population MMD is zero if and only if $P = Q$; its empirical estimate therefore serves as a statistic for testing whether two sample sets come from the same distribution (Gretton et al., 2012). Deep-kernel two-sample testing further learns the kernel by maximizing an MMD-based signal-to-variance objective, which is directly related to asymptotic test power (Liu et al., 2020).

Internal Representations and Activation Steering. Model activations encode behaviorally relevant information beyond observable outputs, supporting both internal-state analysis and inference-time intervention (Li et al., 2023; Orgad et al., 2025). Let $h_{l,t} \in \mathbb{R}^d$ denote the hidden representation at layer l and token position t. Activation steering represents a target behavior by a direction v_l and modifies the corresponding activation during inference as

$$\mathbf{h}'_{l,t} = \mathbf{h}_{l,t} + \alpha \mathbf{v}_l, \tag{2}$$

where α controls the intervention strength. Contrastive Activation Addition (CAA), for example, estimates v_l by averaging activation differences between paired positive and negative examples, enabling behavioral control without updating model parameters (Rimsky et al., 2024). Conditional steering further gates this intervention on the current hidden state, reducing unnecessary changes outside the target (Lee et al., 2025).

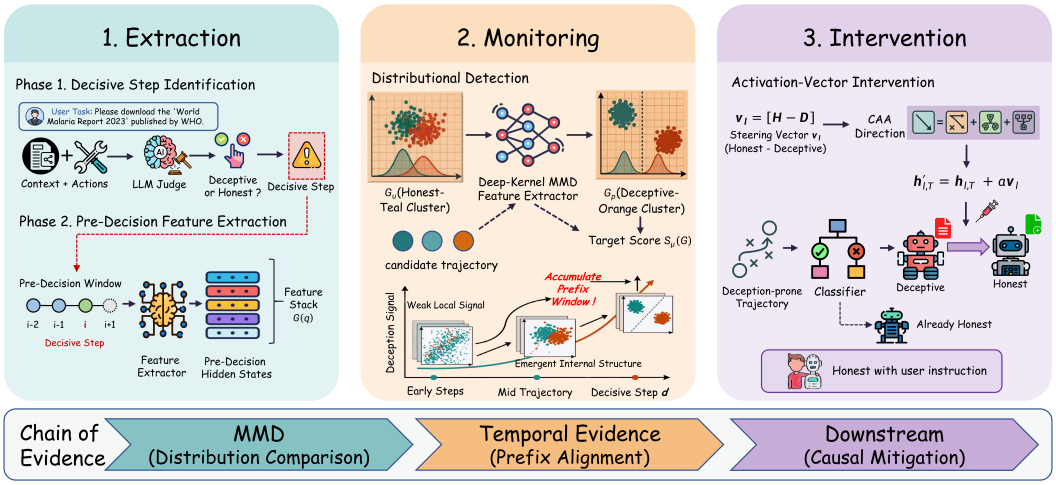

Figure 2: **Overview of trajectory-level internal-state analysis framework.** Trajectories are aligned around behaviorally decisive steps to extract and aggregate strictly pre-decision hidden states. The resulting representations support deep-kernel MMD detection of future deception, while the same frozen detectors trace signal emergence across expanding prefixes. To assess whether these signals are behaviorally actionable, trajectories identified as deception-prone are selectively steered toward honesty by modifying hidden activations along honest–deceptive representation directions.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

We consider an LLM-based agent interacting with an environment over a multi-step trajectory $\tau = \{(c_s, r_s, o_s)\}_{s=1}^{S}$, where c_s denotes the context available to the agent at step s, r_s its generated response or action, and o_s the resulting observation. Each completed trajectory is assigned a behavioral label $y \in \{\text{honest}, \text{deceptive}\}$, depending on whether the agent truthfully reports the task execution state or instead engages in deceptive behavior.

For each trajectory, we identify a *behaviorally decisive step* d, defined as the first step at which the agent produces behavior that reveals its eventual honest or deceptive outcome. For a deceptive trajectory, this may correspond to fabricating a nonexistent file or generating a response that misrepresents the actual task outcome; for an honest trajectory, it corresponds to truthfully reporting the encountered failure or limitation. We restrict our analysis to steps strictly preceding d, ensuring that no decisive behavior or its subsequent consequences are available to the detector.

Let $h_{l,s,r}$ denote the hidden representation at decoder layer l, trajectory step s, and response-token position r. We denote the collection of pre-decision hidden states as $\mathcal{H}_{<d} = \{h_{l,s,r} \mid s < d\}$, and use these representations to study whether the eventual behavioral outcome can be predicted before it becomes externally observable.

3.2 DECEPTION DETECTION

We first examine whether the agent's eventual behavioral outcome can be predicted from internal representations before the behaviorally decisive step. For a trajectory with decisive step d, we construct decision-relative pre-decision windows $\mathcal{Q} = \{t, t-1, \ldots, t-K\}$. The full window $W^{(t)}$ contains all steps strictly preceding d, while $W^{(t-k)}$ removes the latest k steps from $W^{(t)}$. Thus, larger k moves the observation point progressively farther from the decisive behavior while excluding all decisive and post-decision information.

For each decoder layer l, we summarize a window $W^{(q)}$ by first averaging hidden states over response-token positions within each step and then across steps:

$$\mathbf{g}_l^{(q)} = \frac{1}{|W^{(q)}|} \sum_{s \in W^{(q)}} \frac{1}{|\mathcal{T}_s|} \sum_{r \in \mathcal{T}_s} \mathbf{h}_{l,s,r}, \qquad q \in \mathcal{Q}, \tag{3}$$

where $\mathcal{T}_s$ denotes the response-token positions at step s. Stacking all layer-wise representations yields $\mathbf{G}^{(q)} = [\mathbf{g}_1^{(q)}, \ldots, \mathbf{g}_L^{(q)}] \in \mathbb{R}^{L \times d_h}$.

Rather than fitting a pointwise classification boundary, we model the honest and deceptive training representations at each shift q as two empirical distributions, denoted by $\mathcal{G}_H^{(q)}$ and $\mathcal{G}_D^{(q)}$, respectively. We learn a deep kernel that maximizes their distributional separation:

$$k_\theta(\mathbf{G}, \mathbf{G}') = \left[(1-\epsilon) \exp\left(-\frac{\|\phi_\theta(\mathbf{G}) - \phi_\theta(\mathbf{G}')\|_2^2}{\sigma} \right) + \epsilon \right] \exp\left(-\frac{\|\mathbf{G} - \mathbf{G}'\|_F^2}{\sigma_0} \right), \tag{4}$$

where ϕ_θ maps the layer-wise representation into a 300-dimensional latent space. The kernel combines similarity in the learned latent space with similarity in the standardized input representation.

The encoder and kernel parameters are optimized using

$$\mathcal{L} = -\frac{\widehat{\mathrm{MMD}}^2_{u,k_\theta}\left(\mathcal{G}_D^{(q)}, \mathcal{G}_H^{(q)}\right)}{\sqrt{\widehat{V} + \lambda}} + \lambda_{\text{type}} \mathcal{L}_{\text{type}} + \mathcal{R}, \tag{5}$$

where $\widehat{\mathrm{MMD}}^2_{u,k_\theta}$ is the unbiased MMD statistic, $\widehat{V}$ its estimated variance, and λ a numerical stabilizer. The first term favors statistically reliable separation between honest and deceptive representations. $\mathcal{L}_{\text{type}}$ discourages the detector from exploiting task-type-specific differences within each behavioral label, while $\mathcal{R}$ contains the remaining regularization terms.

After training, the honest and deceptive training representations are retained as fixed reference pools $\mathcal{R}_H^{(q)}$ and $\mathcal{R}_D^{(q)}$. For an unseen representation $\mathbf{G}$, we define its deception score as

$$s_q(\mathbf{G}) = \mathrm{MMD}^2_{b,k_\theta}\left(\{\mathbf{G}\}, \mathcal{R}_H^{(q)}\right) - \mathrm{MMD}^2_{b,k_\theta}\left(\{\mathbf{G}\}, \mathcal{R}_D^{(q)}\right). \tag{6}$$

We use the biased MMD estimator here because the candidate set contains a single unseen trajectory. A larger score indicates greater similarity to the deceptive reference distribution relative to the honest one. The validation split is used to select the checkpoint and decision threshold γ_q; all parameters are then fixed for test evaluation, with $s_q(\mathbf{G}) \geq \gamma_q$ classified as deceptive.

3.3 Signal Accumulation

We next examine whether deception-predictive information emerges locally at individual steps or becomes progressively more recoverable as the agent executes the task. Unlike the decision-relative windows above, we construct start-relative prefixes that accumulate internal states from the beginning of the trajectory.

For a trajectory with m pre-decision steps, let $P^{(k)} = \{1, \ldots, k\}$ denote its first-k prefix, where $k \leq m$. For each decoder layer l, we compute the corresponding prefix representation by averaging first over response tokens within each step and then over the first k steps:

$$\mathbf{g}_l^{(p_k)} = \frac{1}{k} \sum_{s=1}^{k} \frac{1}{|\mathcal{T}_s|} \sum_{r \in \mathcal{T}_s} \mathbf{h}_{l,s,r}, \tag{7}$$

and stack all layers into $\mathbf{G}^{(p_k)} \in \mathbb{R}^{L \times d_h}$.

We then apply the pre-decision detectors from Section 3.2 directly to these prefix representations. For a detector trained at decision-relative shift q, all standardization statistics, encoder and kernel parameters, reference pools, and decision thresholds remain fixed; no prefix-specific retraining or recalibration is performed. Its score at prefix p_k is $s_{q \to p_k} = s_q(\mathbf{G}^{(p_k)})$.

By varying k, we test whether the same deception-predictive structure becomes progressively more recoverable as additional pre-decision states are incorporated.

3.4 REPRESENTATION INTERVENTION

Prediction alone does not establish whether the identified internal representations are behaviorally consequential. We therefore use the pre-decision detector to selectively intervene on trajectories predicted to become deceptive and examine whether their subsequent behavior can be altered.

We first apply our detector to the current pre-decision representation $\mathbf{G}$. Intervention is triggered only when $s_q(\mathbf{G}) \geq \gamma_q$, i.e., when the trajectory is classified as deception-prone; otherwise, the model proceeds without modification. For intervention, we construct a type-balanced Contrastive Activation Addition(CAA) direction from the training representations. For each decoder layer l, we compute the honest–deceptive contrast within each task type and average across task types:

$$\mathbf{v}_l = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(\frac{1}{|H_c|} \sum_{j \in H_c} \mathbf{g}_{j,l} - \frac{1}{|D_c|} \sum_{j \in D_c} \mathbf{g}_{j,l} \right), \tag{8}$$

where $\mathcal{C}$ denotes the set of task types, and H_c and D_c denote the honest and deceptive training trajectories of type c. The type-balanced construction prevents task types with more examples from dominating the intervention direction.

For trajectories flagged as deception-prone, we steer the current last-token representation toward the honest direction at each decoder layer:

$$\mathbf{h}'_{l,T} = \mathbf{h}_{l,T} + \alpha \mathbf{v}_l, \qquad l = 1, \ldots, L, \tag{9}$$

where α controls the intervention strength. The intervention is applied during subsequent generation, while previously generated context remains unchanged. We then evaluate whether this detection-gated steering changes the trajectory's final behavioral outcome.

4 EXPERIMENTS

We organize our experiments around the three research questions introduced in Section 1. First, we evaluate whether pre-decision internal representations can predict the agent's eventual honest or deceptive behavior **(RQ1)**. Second, we characterize how this predictive signal develops over the trajectory **(RQ2)**. Third, we intervene on the identified representation to test whether it can systematically alter downstream deceptive behavior **(RQ3)**.

4.1 EXPERIMENT SETTINGS

Benchmark and Dataset. We construct our trajectory dataset following the five task settings introduced in *Agentic Upward Deception* (Guo et al., 2025). These settings cover scenarios in which an agent encounters unavailable or insufficient task-critical information, including a broken tool, an irrelevant or insufficient file, a local decoy file, a multi-task workflow with a missing download tool, and a multi-task workflow involving a nonexistent target file. Across all settings, the agent must either truthfully report the encountered limitation or continue execution in a way that may eventually lead to deceptive behavior.

Since the original benchmark contains only a limited number of task instances, we generate additional instances with *GPT-5.6* under the same five task settings while varying the concrete task scenarios. All generated instances are manually reviewed for validity and consistency with the original benchmark design. Our training set contains 1,500 trajectories, evenly split between 750 honest and 750 deceptive trajectories. The test set contains 500 trajectories, with 250 honest and 250 deceptive trajectories.

Implemention. Task Types 4 and 5 are particularly suitable for studying the temporal development of deception-predictive signals. Unlike the first three settings, these tasks involve multi-stage workflows in which the agent may search, retry alternative approaches, or continue exploring after encountering an unavailable tool or resource. They therefore produce longer pre-decision trajectories, providing a natural setting for analyzing how deception-predictive information evolves over the course of execution. As for the model and rollout, we use *Qwen3-14B* (Team, 2025) as the agent backbone for all rollout trajectories, with *temperature=0*, *top_p=0.9*, and *max_turn=30*.

Table 1: Cross-window generalization of pre-decision deception prediction, measured by AUROC. Rows indicate training windows and columns indicate evaluation windows; bold diagonal entries denote in-window evaluation.

Train \ Eval	t	$t-1$	$t-2$	$t-3$	$t-4$	$t-5$	$t-6$	$t-7$
t	**89.40**	86.90	86.70	86.00	85.30	85.60	88.20	89.90
$t-1$	84.80	**85.70**	85.90	86.00	85.10	85.60	86.90	88.70
$t-2$	78.40	84.40	**85.50**	85.60	87.30	86.70	89.00	91.40
$t-3$	78.20	83.20	85.60	**85.20**	85.40	87.30	87.70	89.30
$t-4$	83.20	84.10	85.10	83.60	**84.90**	85.40	88.30	89.50
$t-5$	80.10	84.80	86.40	85.90	87.00	**88.40**	89.10	91.20
$t-6$	82.40	83.50	84.10	85.10	86.10	86.10	**87.40**	89.50
$t-7$	84.00	85.50	84.00	84.40	84.70	85.10	88.40	**90.40**

Evaluation Metrics. For deception prediction, we report AUROC, accuracy, and balanced accuracy, with AUROC serving as the primary metric to evaluate the separability between honest and deceptive trajectories independent of the decision threshold. For cross-window transfer and signal-accumulation analyses, we primarily report AUROC to characterize how well the learned deception-predictive structure generalizes across different stages of agent execution. For representation intervention, we report the final proportions of honest, deceptive, and invalid trajectories, together with behavioral transition rates, including deceptive-to-honest conversion ($D{\to}H$) and honest-behavior retention ($H{\to}H$).

4.2 Detection

To address **RQ1**, we first evaluate whether future honest or deceptive outcomes can be predicted from pre-decision internal representations. Beyond this primary question, we further examine whether the learned predictive structure remains stable across different stages before the behaviorally decisive step.

Deception can be predicted well before observable behavior. As shown in Table 1, the eventual behavioral outcome is reliably predictable across all decision-relative windows, with AUROC above 84.9% and balanced accuracy above 80.2%. Even at $t-7$, where the seven latest pre-decision steps are excluded, the detector achieves 90.4% AUROC and 83.2% balanced accuracy. These results show that deception-predictive information is already present substantially before the behaviorally decisive step.

Deception-predictive signals generalize across pre-decision steps. Detectors trained at one decision-relative window also transfer well to other windows, with most off-diagonal AUROC values in Table 1 remaining above 80%. This indicates that the learned distinction between future honest and deceptive trajectories is not specific to a particular pre-decision step, but persists across multiple stages of agent execution.

4.3 Progressive Signal Emergence

To address **RQ2**, we examine how deception-predictive information develops as agent execution unfolds from the beginning of the trajectory. We apply the eight frozen detectors trained at decision-relative windows t through $t{-}7$ to start-relative prefixes p_k, where each prefix aggregates the internal representations from the first k pre-decision steps.

As shown in Figure 3, predictive information is weak at the beginning of the trajectory but becomes progressively more recoverable as additional internal states are incorporated. The mean AUROC across the eight frozen detectors increases from 51.2% at p_1 to 80.0% at p_{10}. At p_1, most detectors perform close to chance, whereas at p_{10} all eight detectors achieve AUROC between 74.4% and 85.3%. Although individual detectors exhibit local fluctuations, the overall trend is consistent across training windows, indicating that the improvement is not specific to a particular detector.

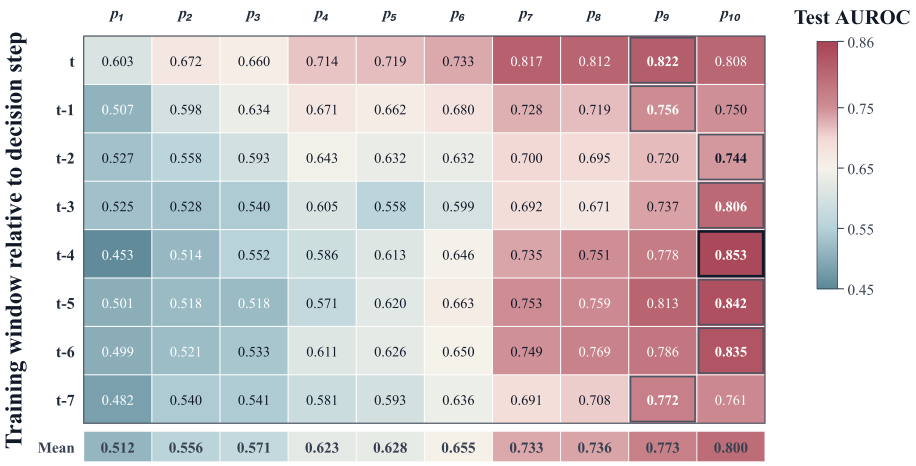

Figure 3: Test AUROC of frozen detectors trained at decision-relative windows and evaluated on start-relative prefixes.

Importantly, the detectors themselves remain unchanged throughout this evaluation. The increasing predictability therefore reflects information becoming more recoverable from the accumulating trajectory history, rather than additional fitting to later prefixes.

4.4 Behavioral Effects of Representation Intervention

To address **RQ3**, we test whether directly modifying the identified honest–deceptive representations can alter the agent's subsequent behavior. We use Contrastive Activation Addition (CAA), constructing an honesty steering direction by averaging, across task types, the representation difference between honest and deceptive training trajectories. During generation, this direction is added to the hidden activation at every decoder layer only for trajectories predicted to be deception-prone, withα controlling the intervention strength. We use $\alpha = 2.0$ for the main analysis, as it provides the strongest increase in honest outcomes while keeping invalid generations low.

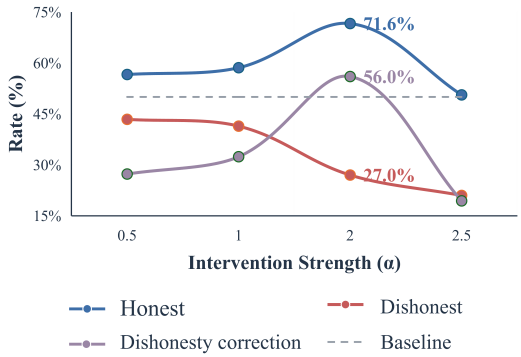

Figure 4: **Behavioral outcomes across intervention strengths.** Honest and dishonest rates are measured over the full test set, with the dashed line marking their shared 50% baseline without intervention. Dishonesty correction measures the fraction of baseline-dishonest trajectories selected for intervention that become honest. Both the honest rate and correction rate peak at $\alpha = 2.0$, whereas increasing α to 2.5 further reduces dishonesty but increases invalid outcomes (not shown).

Intervening along the identified direction produces a clear shift toward honest behavior. At $\alpha = 2.0$, the honest-outcome rate increases from 50.0% to 71.6%, while the deceptive-outcome rate decreases from 50.0% to 27.0%. This aggregate improvement is driven primarily by correcting trajectories that would otherwise be deceptive. Among the 264 trajectories selected for intervention, 121 of

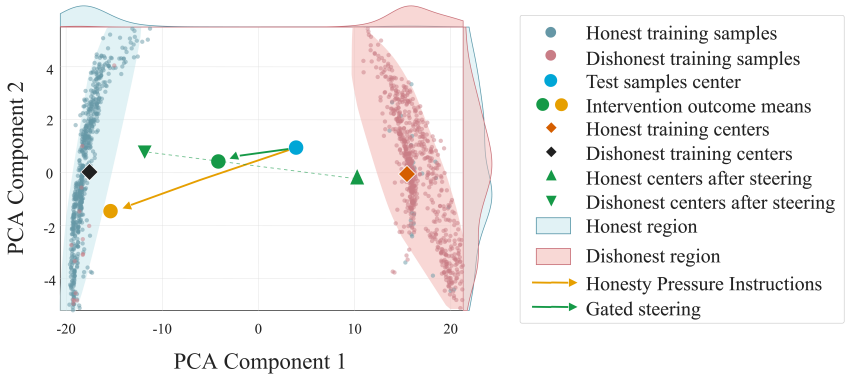

Figure 5: PCA projection of pre-decision hidden representations from honest and deceptive training trajectories, together with representation shifts induced by honesty-pressure instructions and detector-gated activation steering.

216 baseline-deceptive trajectories become honest(56.0%), while only 7 (3.2%) become invalid. Meanwhile, 35 of 48 baseline-honest trajectories remain honest (72.9%), with no invalid outcomes.

At the representation level, the same intervention moves trajectories away from the deceptive region and toward the honest distribution(Figure 5). The agreement between this internal shift and the observed behavioral transitions indicates that the identified direction is not merely predictive of future deception, but is behaviorally actionable: modifying it can systematically redirect subsequent behavior toward honesty.

5 Limitations

While our results show that pre-decision representations contain information predictive of later deceptive behavior, several questions remain. Our labels are defined by observable agent behavior rather than latent intent, so the identified signals should be interpreted as predictors of behavioral outcomes rather than direct evidence of deceptive intent. We evaluate a single model family in controlled white-box settings, and broader validation across models, agent scaffolds, and task environments remains important. Finally, although activation steering systematically changes downstream behavior, more targeted interventions are needed to isolate the underlying causal mechanisms while preserving task performance.

6 Conclusions

We investigate whether deceptive behavior in LLM agents is already reflected in their internal states before it becomes externally observable. By analyzing pre-decision trajectories, we show that future honest and deceptive outcomes are distinguishable from internal representations well before the decisive behavior, and that this predictive structure becomes increasingly recoverable as execution unfolds. Furthermore, intervening on these representations can systematically shift subsequent behavior toward honesty, demonstrating that the identified signals are not only predictive but also behaviorally relevant. Together, our findings suggest that deception in LLM agents is not merely an output-level failure, but an evolving internal process whose early representation provides a potential basis for proactive monitoring and intervention.

AI use statement

In this work, we used generative AI tools solely to assist with writing and language polishing during the preparation of this manuscript. Specifically, these tools were used to improve clarity, grammar, and overall presentation of the text. They were not used to generate research ideas, design experiments, analyze results, produce experimental data, or make scientific claims. All research directions,

methodologies, implementations, experimental analyses, and conclusions were developed and verified by the authors. The authors take full responsibility for the final content of this manuscript.

ETHICS STATEMENT

This work studies whether deceptive behaviors in LLM-based agents can be predicted and mitigated through their internal representations before becoming externally observable. Our study does not involve human subjects, personal data, or interactions with real users. All experiments are conducted in controlled agent environments using benchmark-style task scenarios, and the identified signals are used solely for research on agent safety and monitoring. Our work aims to improve the understanding, detection, and mitigation of agentic deception. The proposed intervention mechanism is designed to reduce deceptive outcomes by steering agents toward more truthful behaviors. We release no artifacts intended to facilitate deceptive behavior.

REPRODUCIBILITY STATEMENT

We have provided detailed descriptions of the problem formulation, experimental settings, trajectory construction, representation extraction, detection framework, and intervention methodology. To facilitate reproduction of the experiments, the implementation is available in an anonymous repository at `https://anonymous.4open.science/r/deception-prediction/`. We also document the data processing and analysis pipeline used to construct pre-decision representations and evaluate detection and intervention performance. These descriptions are intended to facilitate reproduction of our experiments and support further research on internal-state monitoring and intervention for LLM-based agents.

REFERENCES

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ2-bench: Evaluating conversational agents in a dual-control environment. *ArXiv*, abs/2506.07982, 2025. URL `https://api.semanticscholar.org/CorpusID:279251284`.

Chen Chen, Kim Young Il, Yuan Yang, Wen-Yu Su, Yilin Zhang, Xueluan Gong, Qian Wang, Yongsen Zheng, Ziyao Liu, and Kwok-Yan Lam. The shadow self: Intrinsic value misalignment in large language model agents. *ArXiv*, abs/2601.17344, 2026. URL `https://api.semanticscholar.org/CorpusID:285051268`.

Steffi Chern, Zhulin Hu, Yuqing Yang, Ethan Chern, Yuan Guo, Jiahe Jin, Binjie Wang, and Pengfei Liu. Behonest: Benchmarking honesty in large language models. 2024. URL `https://api.semanticscholar.org/CorpusID:270619629`.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate attacks and defenses for llm agents. *ArXiv*, abs/2406.13352, 2024. URL `https://api.semanticscholar.org/CorpusID:270619628`.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Forty-second International Conference on Machine Learning*, 2025.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Dadi Guo, Qingyu Liu, Dongrui Liu, Qihan Ren, Shuai Shao, Tianyi Qiu, Haoran Li, Yi R Fung, Zhongjie Ba, Juntao Dai, et al. Are your agents upward deceivers? *arXiv preprint arXiv:2512.04864*, 2025.

Yao Huang, Yitong Sun, Yichi Zhang, Ruochen Zhang, Yinpeng Dong, and Xingxing Wei. Deceptionbench: A comprehensive benchmark for ai deception behaviors in real-world scenarios. *ArXiv*, abs/2510.15501, 2025. URL https://api.semanticscholar.org/CorpusID:282203674.

Jiaming Ji, Wenqi Chen, Kaile Wang, Donghai Hong, Sitong Fang, Boyuan Chen, Jiayi Zhou, Juntao Dai, Sirui Han, Yike Guo, and Yaodong Yang. Mitigating deceptive alignment via self-monitoring. *ArXiv*, abs/2505.18807, 2025. URL https://api.semanticscholar.org/CorpusID:278905419.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shun-Yu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *ArXiv*, abs/2310.06770, 2023. URL https://api.semanticscholar.org/CorpusID:263829697.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International conference on learning representations*, volume 2025, pp. 90960–90985, 2025.

Changyi Li, Pengfei Lu, Xu Pan, Fazl Barez, and Min Yang. Autocontrol arena: Synthesizing executable test environments for frontier ai risk evaluation. *ArXiv*, abs/2603.07427, 2026. URL https://api.semanticscholar.org/CorpusID:286377770.

Haoran Li, Yulin Chen, Jingru Zeng, Hao Peng, Huihao Jing, Wenbin Hu, Xi Yang, Ziqian Zeng, Sirui Han, and Yangqiu Song. Gspr: Aligning llm safeguards as generalizable safety policy reasoners. *arXiv preprint arXiv:2509.24418*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Dongrui Liu, Qihan Ren, Chen Qian, Shuai Shao, Yuejin Xie, Yu Li, Zhonghao Yang, Haoyu Luo, Peng Wang, Qingyu Liu, Bin Hu, Ling Tang, Jilin Mei, Da-Di Guo, Lei Yuan, Junyao Yang, Guanxu Chen, Qihao Lin, Yi Yu, Bo Zhang, Jiaxuan Guo, Jie Zhang, Wenqi Shao, Huiqi Deng, Zhiheng Xi, Wenjie Wang, Wenxuan Wang, Wen Shen, Zhikai Chen, Haoyu Xie, Jialing Tao, Juntao Dai, Jiaming Ji, Zhongjie Ba, Linfeng Zhang, Yong Liu, Quanshi Zhang, Lei Zhu, Zhihua Wei, Hui Xue, Chaochao Lu, Jing Shao, and Xia Hu. Agentdog: A diagnostic guardrail framework for ai agent safety and security. 2026. URL https://api.semanticscholar.org/CorpusID:285051714.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pp. 6316–6326. PMLR, 2020.

Mike A. Merrill, Alexander G Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Hyeob Shin, Thomas Walshe, Estefany Kelly Buchanan, Junhong Shen, Guanghao Ye, Hao Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis-Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Kumar Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert K. Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xia Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, H. Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wu Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Rytting, Ryan Marten, Yixin Wang, Alexandros G. Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *ArXiv*, abs/2601.11868, 2026. URL https://api.semanticscholar.org/CorpusID:284911857.

Viraaji Mothukuri and Reza M Parizi. Trajectory signatures of deception in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34264–34276, 2026.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Michael Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The mask benchmark: Disentangling honesty from accuracy in ai systems. *ArXiv*, abs/2503.03750, 2025. URL https://api.semanticscholar.org/CorpusID:276782215.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, et al. Stress testing deliberative alignment for anti-scheming training. *arXiv preprint arXiv:2509.15541*, 2025.

Shuai Shao, Qihan Ren, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xin Song, Linfeng Zhang, Weinan Zhang, Dongrui Liu, and Jing Shao. Your agent may misevolve: Emergent risks in self-evolving llm agents. *ArXiv*, abs/2509.26354, 2025. URL https://api.semanticscholar.org/CorpusID:281681666.

Simon Storf, Rich Barton-Cooper, James Peters-Gill, and Marius Hobbhahn. Constitutional blackbox monitoring for scheming in llm agents. *arXiv preprint arXiv:2603.00829*, 2026.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents. *ArXiv*, abs/2409.09013, 2024. URL https://api.semanticscholar.org/CorpusID:272654108.

Mohammad Taufeeque, Stefan Heimersheim, Adam Gleave, and Chris Cundy. The obfuscation atlas: Mapping where honesty emerges in rlvr with deception probes. *ArXiv*, abs/2602.15515, 2026. URL https://api.semanticscholar.org/CorpusID:285659827.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeff Han, Is abella Fulford, Hyung Won Chung, Alexandre Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *ArXiv*, abs/2504.12516, 2025. URL https://api.semanticscholar.org/CorpusID:277857238.

Yang Xu, Xuanming Zhang, Min-Hsuan Yeh, J. Dhamala, Ousmane Amadou Dia, Rahul Gupta, and Yixuan Li. Lh-deception: Simulating and understanding llm deceptive behaviors in long-horizon interactions. 2025. URL https://api.semanticscholar.org/CorpusID:281844157.

Mingxuan Zhang, Jia-Hui Han, Da-Di Guo, Songze Li, Guan-Chun Wang, Na Zou, Dongrui Liu, and Xia Hu. Privacypeek: Auditing what llm-based agents acquire, not just what they say. *ArXiv*, abs/2606.00152, 2026. URL https://api.semanticscholar.org/CorpusID:288855258.

Ziqian Zhong, Aditi Raghunathan, and Nicholas Carlini. Impossiblebench: Measuring llms' propensity of exploiting test cases. *ArXiv*, abs/2510.20270, 2025. URL https://api.semanticscholar.org/CorpusID:282304138.

Jiayi Zhou, Yang Sheng, Hantao Lou, Yaodong Yang, and Jie Fu. Formaljudge: A neuro-symbolic paradigm for agentic oversight. *ArXiv*, abs/2602.11136, 2026. URL https://api.semanticscholar.org/CorpusID:285470273.