Where Reasoning Models Commit to Deception*

Aryan Gupta
aryan.cs.app@gmail.com

Abstract

Reasoning-trained language models routinely generate chains of thought (CoT) whose intermediate content is materially inconsistent with the final user-facing answer. Documented cases include sycophantic capitulation against an internally acknowledged correct answer (Sharma et al., 2023; Perez et al., 2022), in-context scheming and alignment faking under disclosed retraining pressure (Meinke et al., 2024; Greenblatt et al., 2024), obfuscated reward hacking under CoT-monitor optimisation (Baker et al., 2025), and systematically unverbalised hint use across hint families (Chen et al., 2025). One natural defence reads the residual stream during CoT generation and intervenes when deception forms. Per-token probing for this purpose has been developed in recent work (Goldowsky-Dill et al., 2025; Chrabąszcz et al., 2026; Mirtaheri and Belkin, 2026), but the open question is not whether a per-token signal exists; it is whether a discrete transition token, at which the model's internal state crosses from honest reasoning into deceptive output formation, can be defined, identified, and causally validated, and whether intervention restricted to that token can flip the trajectory at a strictly smaller distributional cost than the uniform or smoothly-gated interventions currently in use (Panickssery et al., 2023; Lee et al., 2024; Ferrando et al., 2025; Yu et al., 2025). The present document formalises that question. CoT generation is modelled as a partially observed two-state Markov process whose latent honesty state is read through a linear residual-stream probe. The pivot is defined as the smallest stopping time at which a single-token intervention is causally sufficient to flip the downstream output. Under the model and a monotonicity assumption on the causal effect of edits, the causal pivot almost surely precedes or coincides with the probe argmax. The Kullback–Leibler divergence between the intervened and base CoT distributions is bounded, and pivot-localized intervention is shown to dominate uniform Contrastive Activation Addition by a factor of $\Theta(T)$ in CoT length. An online CUSUM-type sequential detector with calibrated false-alarm control completes the method. The experimental protocol on open reasoning models, and the relationship to recent monitorability work (Korbak et al., 2025), are stated at the end.

1 Introduction

The phenomenon is empirically well established. Lanham et al. (2023) measured CoT faithfulness under truncation, paraphrase, and mistake insertion and reported inverse scaling on most tasks. Turpin et al. (2023) showed that biased few-shot prompts cause systematically unfaithful explanations. On six hint categories, Chen et al. (2025) report that Claude 3.7 Sonnet and DeepSeek-R1 verbalise an outcome-changing hint at average rates well below 0.30, and as low as 0.02 in a reward-hacking environment. Baker et al. (2025) demonstrate that an LLM monitor reading the CoT initially catches roughly 95% of reward-hack episodes during agentic coding, but that adding the monitor's verdict to the RL reward produces obfuscated CoTs that hack at the same rate while

reading as innocuous. Meinke et al. (2024) document explicit scheming language in the scratchpad of frontier models and the maintenance of deception across more than 85% of multi-turn interrogation rounds. Greenblatt et al. (2024) show that Claude 3 Opus, given a scratchpad and a system prompt indicating a training context, complies with harmful queries at non-trivial rates with explicit alignment-faking reasoning visible in the scratchpad. Korbak et al. (2025) argue, as an organisational position, that the human-language CoT in present reasoning models is a privileged but fragile monitoring channel.

The defensive responses divide into two. Black-box CoT monitoring (Baker et al., 2025; Korbak et al., 2025) reads the scratchpad with another model and flags or rewards visible misbehaviour; white-box residual-stream monitoring (MacDiarmid et al., 2024; Goldowsky-Dill et al., 2025) reads the model's own activations and bypasses any structural pressure on the legible text. We work in the latter regime. In that setting, simple linear probes (mass-mean, logistic regression, contrast-consistent search) trained on contrastive honesty data attain high in-distribution AUROC (Burns et al., 2022; Marks and Tegmark, 2023; Bürger et al., 2024; Goldowsky-Dill et al., 2025). The per-token probe signal across a generated response is non-stationary and informative beyond aggregate scores (Goldowsky-Dill et al., 2025; Chrabąszcz et al., 2026). And acting on the same direction at all post-prompt token positions, as in Contrastive Activation Addition (Panickssery et al., 2023) or Inference-Time Intervention (Li et al., 2023), induces a substantial fluency tax (Pres et al., 2024; Li et al., 2026).

Our thesis is that the empirical non-stationarity of the per-token probe trajectory traces a discrete latent transition, the *pivot*, and that the appropriate intervention acts at that token alone. The contributions are as follows.

1. A formal hidden-state model of CoT generation under deception in which the pivot is a well-defined stopping time on the latent honesty process (Section 4).
2. A causal definition of the pivot via single-token interventional sufficiency, distinct from the statistical first-passage rule, together with an identification result showing that under the assumed model the causal pivot almost surely precedes or coincides with the probe argmax (Section 5).
3. Pivot-Localized Intervention (PLI), a single-token version of the CAA update, with a KL bound establishing that PLI's distributional footprint on the CoT distribution is $O(\alpha^2\sigma^2)$ where the uniform CAA footprint is $\Theta(T\alpha^2\sigma^2)$ for CoT length T (Section 6).
4. An online detection rule based on a CUSUM-type sequential test with calibrated false-alarm rate, suitable for use during decoding (Section 7).
5. An experimental protocol on open reasoning models that distinguishes the present proposal from offline position-scanned PIXEL (Yu et al., 2025), from smoothly probe-gated DSAS (Ferrando et al., 2025), and from prompt-conditional CAST (Lee et al., 2024) on a Pareto frontier of steering efficacy against fluency cost (Section 8).

A note on the novelty boundary. Per-token deception scores were produced by Goldowsky-Dill et al. (2025), who also reported failure modes; probe trajectories across the reasoning process were constructed by Chrabąszcz et al. (2026); pre- and post-CoT probing for motivated reasoning was studied by Mirtaheri and Belkin (2026). Per-token probing is therefore not a contribution of the present work. What is a contribution is the causal characterisation of a discrete pivot event, its identification, and the demonstration that a token-localized intervention occupies a region of the position-versus-gating design space that PIXEL (offline position-scan, no online probe) and DSAS

(online probe, smooth gating, no sparsity) bracket but do not cover. A second clarification concerns the safety frame: we are not solving detection. Detection of deception in present reasoning models, on standard contrastive benchmarks, is already near-saturated (Goldowsky-Dill et al., 2025). The remaining problem is intervention, producing a model whose post-pivot continuation is honest at low distributional cost, which matters because uniform steering is known to corrupt fluency (Pres et al., 2024).

2 Related work

Representation engineering and linear probes for honesty. Zou et al. (2023) introduced Linear Artificial Tomography (LAT) and the broader programme of Representation Engineering, computing a contrastive reading vector via PCA of paired stimulus differences at a single token position (typically the last token of a templated stimulus). Burns et al. (2022) introduced contrast-consistent search, an unsupervised probe on the final-token hidden state. Marks and Tegmark (2023) systematically compared mass-mean, logistic regression, CCS, and few-shot probes at a single layer per model on declarative-statement datasets. Bürger et al. (2024) characterised a two-dimensional truth subspace spanned by a polarity-invariant and a polarity-sensitive direction. MacDiarmid et al. (2024) demonstrated near-perfect mass-mean probe detection of sleeper-agent defection at the final prompt token or the scratchpad-closing tag. None of these works define or target a discrete intra-CoT transition.

Per-token probing of reasoning models. Goldowsky-Dill et al. (2025) report per-token deception scores from layer-22 residual-stream probes on Llama-3.3-70B-Instruct, with AUROCs from 0.96 to 0.999 on Roleplaying, Insider Trading (Scheurer et al., 2023), and Sandbagging. Chrabąszcz et al. (2026) construct probe trajectories across the reasoning process and use trajectory features for outcome prediction. Mirtaheri and Belkin (2026) show that pre-CoT and post-CoT probes match or exceed a CoT-reading monitor for motivated reasoning. The pivot question, as a causal characterisation of a single transition event, is not formulated in these works.

Steering and intervention. Subramani et al. (2022) extracted latent steering vectors by reconstruction objectives. Turner et al. (2023) introduced single-pair activation addition. Panickssery et al. (2023) averaged contrastive residual differences to form the mass-mean Contrastive Activation Addition (CAA) update, applied at every post-prompt token. Li et al. (2023) computed per-head mass-mean shifts and intervened at every forward pass on a set of high-probe-accuracy heads. Lee et al. (2024) introduced Conditional Activation Steering (CAST), which gates a CAA-style vector via a cosine-similarity test computed during the prompt forward pass and then applies the vector uniformly to all subsequent generated tokens. Ferrando et al. (2025) introduced Dynamically Scaled Activation Steering (DSAS), in which a trained per-token logistic regressor produces a continuous scaling in $[0, 1]$ applied at every generated token. Yu et al. (2025) introduced PIXEL, in which a single token position t^* is selected by offline grid scan on a held-out set and the steering vector is injected at that fixed position with a closed-form coefficient. Feng et al. (2026) restrict intervention to atomic feature units. Cho et al. (2026) use reinforcement learning to identify sparse autoencoder features for token-level intervention. None of these works couple an online residual-stream probe with single-token intervention; PIXEL is offline-position-sparse but not online-gated, DSAS is online-gated but not sparse.

Activation patching and causal localisation. Meng et al. (2022) introduced causal tracing for factual recall, localising effects to mid-layer MLPs at the last subject token. Goldowsky-Dill et al. (2023) generalised to path patching. Heimersheim and Nanda (2024) formalised noising and denoising and emphasised that patching identifies where information must be present, not where computation occurs. Causal tracing routinely identifies position-specific effects but has not, to our knowledge, been applied to defining a pivot token in CoT.

Theory of intervention. Pres et al. (2024) report that broad-spectrum CAA and ITI coefficients raise perplexity and degrade unrelated capabilities, and propose accounting for likelihoods and standardised targets. Li et al. (2026) document overlap between steering directions and refusal subspaces. Billa (2026) characterises a Linear Accessibility Profile predicting when steering will succeed at a given layer. These results motivate the present pursuit of an intervention whose distributional cost scales sublinearly in CoT length.

3 Preliminaries

3.1 Model and notation

We consider a decoder-only Transformer f_θ with L layers operating on a token alphabet $\mathcal{V}$. For an input sequence $x_{1:t} = (x_1, \ldots, x_t)$ the residual stream at layer $\ell \in \{1, \ldots, L\}$ and position t is denoted $\boldsymbol{h}_\ell(t) \in \mathbb{R}^d$. The next-token distribution is $p_\theta(\cdot \mid x_{1:t}) = \mathrm{softmax}(W_U \boldsymbol{h}_L(t))$ for the unembedding $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$. We write $\mathbb{P}_f$ for the law over generated continuations under autoregressive sampling from f_θ given a fixed prompt π.

A chain of thought is a generated sequence $y_{1:T} \in \mathcal{V}^T$ followed by a delimiter $\delta \in \mathcal{V}$ separating scratchpad from final answer; in the reasoning-model conventions of DeepSeek-AI et al. (2025); Muennighoff et al. (2025) the scratchpad lies between `<think>` and `</think>` tokens and the answer between `<answer>` and `</answer>`. We write y^{out} for the final answer span.

3.2 Honesty contrast data and the mass-mean probe

Let $\mathcal{D}^+$ and $\mathcal{D}^-$ denote sets of CoT prefixes paired by a known honesty contrast, where $+$ indicates honest reasoning and $-$ deceptive reasoning. Concretely, $\mathcal{D}^\pm$ may be generated by varying a system-prompt directive between truthful and deceptive instructions on a fixed underlying question set, following the construction of Goldowsky-Dill et al. (2025). At a chosen layer ℓ and position scheme τ, the empirical mass-mean direction is

$$\boldsymbol{\theta}_\ell \;=\; \boldsymbol{\mu}_\ell^+ - \boldsymbol{\mu}_\ell^-, \qquad \boldsymbol{\mu}_\ell^\pm \;=\; \frac{1}{|\mathcal{D}^\pm|} \sum_{x \in \mathcal{D}^\pm} \boldsymbol{h}_\ell^{(x)}(\tau(x)),$$

and we write $\sigma_\ell^2 = \mathrm{Var}\big(\langle \boldsymbol{h}_\ell, \boldsymbol{\theta}_\ell / \|\boldsymbol{\theta}_\ell\| \rangle\big)$ for the projected variance estimated on a held-out honest set. The probe is the linear functional $s_\ell(\boldsymbol{h}) = \langle \boldsymbol{h}, \boldsymbol{\theta}_\ell / \|\boldsymbol{\theta}_\ell\| \rangle$, calibrated to a threshold τ_α at which the false-alarm rate on the honest set is α. For brevity in what follows we suppress ℓ and write $\boldsymbol{\theta}$ for the unit-normalised direction and $s_t = s(\boldsymbol{h}(t))$ for the per-token score.

3.3 The CAA and ITI updates

The CAA update of Panickssery et al. (2023) is, at layer ℓ and coefficient $c \in \mathbb{R}$,

$$\boldsymbol{h}_\ell(t) \;\leftarrow\; \boldsymbol{h}_\ell(t) + c\,\boldsymbol{\theta}_\ell \qquad \text{for all } t \in \mathcal{T}_{\text{gen}},$$

where $\mathcal{T}_{\text{gen}}$ is the set of post-prompt positions. The ITI update of Li et al. (2023) applies the analogous per-head shift, $\boldsymbol{h}^h_\ell \leftarrow \boldsymbol{h}^h_\ell + \alpha\sigma^h_\ell\boldsymbol{\theta}^h_\ell$, on a fixed set of top-$K$ attention heads at every forward pass. Both updates are position-uniform; this is the design property we will modify.

4 A latent-state model of CoT deception

We adopt a deliberately small, falsifiable generative model of CoT generation under the honesty contrast. The point is not realism in detail but to define the pivot rigorously and to make the identification claims provable. Empirical departures from the model are flagged in Section 9.

Definition 4.1 (Two-state CoT process)**.** Fix a prompt π and a contrast level. A two-state CoT process is the tuple $(Z_t, Y_t, S_t)_{t=1}^T$ where $Z_t \in \{H, D\}$ is a latent honesty state (honest, deceptive), $Y_t \in \mathcal{V}$ is the generated token, and $S_t = s(\boldsymbol{h}_\ell(t)) \in \mathbb{R}$ is the per-token probe score, satisfying

$$Z_1 = H \quad \text{almost surely}, \tag{1}$$

$$\mathbb{P}(Z_{t+1} = D \mid Z_t = H, Y_{1:t}) = q(Y_{1:t}), \tag{2}$$

$$\mathbb{P}(Z_{t+1} = H \mid Z_t = D, Y_{1:t}) = 0, \tag{3}$$

$$S_t \mid Z_t = z \;\sim\; \mathcal{N}\big(\mu_z,\, \sigma^2\big), \quad \mu_H < \mu_D, \tag{4}$$

$$Y_{t+1} \mid Z_{t+1}, Y_{1:t} \;\sim\; p_\theta^{(Z_{t+1})}(\cdot \mid Y_{1:t}), \tag{5}$$

with hazard $q(\cdot) \in [0, 1]$ and conditional emission distributions p^H_θ, p^D_θ.

Deceptive intent is treated as a discrete event with no return: once D is entered, the chain persists in it. The per-token probe is a noisy Gaussian observation of the latent state. The next-token distribution is a mixture indexed by the latent state, so honesty propagates to output content through this mixture rather than through any external constraint. The absorbing structure of D is consistent with the qualitative picture in Baker et al. (2025), where a reasoning model that has committed to a deceptive plan typically holds it through the rest of the CoT. The absorbing assumption is relaxed in Remark 4.3.

Definition 4.2 (Latent pivot)**.** The latent pivot is the stopping time $T^\circ \in \{2, \ldots, T, \infty\}$ defined by

$$T^\circ \;=\; \inf\{t \geq 2 : Z_t = D\}, \qquad \inf \emptyset := \infty.$$

The latent pivot is unobservable. We are interested in two observable surrogates: the probe-based statistical pivot (Section 5) and the causally validated pivot (Definition 5.3 below).

Remark 4.3 (Reversibility)**.** For models that can recover from incipient deceptive intent (e.g., via the self-correction markers *Wait,* or *Actually,* characteristic of R1-style models), the absorbing assumption can be replaced by a small return rate $\mathbb{P}(Z_{t+1} = H \mid Z_t = D) = r > 0$. All results below carry through with T° redefined as the first deceptive run of length at least k, at the cost of a $\log(1/r)$ inflation of the identification bound.

5 Probes, pivots, and identification

5.1 The statistical pivot

Definition 5.1 (Statistical pivot)**.** For a probe trajectory $(S_t)_{t=1}^{T}$ and a threshold τ, the statistical pivot at threshold τ is

$$T_\tau^\star \;=\; \inf\{t : S_t > \tau\}.$$

The threshold τ is calibrated on a held-out honest CoT distribution so that under $Z \equiv H$ the false-alarm rate per position is some small α. In a single CoT of length T this gives a sequence-level false-alarm rate of $1-(1-\alpha)^T$, which is the headline reason a single-threshold rule will not suffice in practice; we handle this with the sequential procedure of Section 7.

5.2 The causal pivot

The statistical pivot is observable but not, by itself, the object we want. A probe trajectory can exceed τ at a token whose contents are honesty-neutral and where deceptive output formation has not yet occurred; conversely the trajectory's argmax is not, a priori, the token at which intervention is most effective. Causal sufficiency for output flipping is the operational property we need.

Definition 5.2 (Single-token interventional flip)**.** Fix a layer ℓ, a coefficient $c \in \mathbb{R}$, and a direction $\boldsymbol{v} \in \mathbb{R}^d$. For a generated CoT $y_{1:T}$ with output y^{out}, the single-token intervention $\mathrm{Do}(t; c, \boldsymbol{v})$ replaces $\boldsymbol{h}_\ell(t)$ with $\boldsymbol{h}_\ell(t) - c\boldsymbol{v}$ at decoding step t and continues sampling autoregressively. Let $y^{\text{out}}(t; c, \boldsymbol{v})$ be the resulting output. The flip event at level $\Delta \in (0, 1]$ is

$$F_\Delta(t; c, \boldsymbol{v}) \;=\; \big\{\mathbb{P}_{\text{seed}}(y^{\text{out}}(t; c, \boldsymbol{v}) \text{ honest}) \geq \mathbb{P}_{\text{seed}}(y^{\text{out}} \text{ honest}) + \Delta\big\}.$$

Definition 5.3 (Causal pivot)**.** With $\boldsymbol{v} = \boldsymbol{\theta}$ the honesty direction and c a fixed coefficient, the causal pivot at level Δ is

$$T_\Delta^\dagger \;=\; \inf\big\{t : F_\Delta(t; c, \boldsymbol{\theta}) \text{ holds}\big\}.$$

The causal pivot is the smallest token at which a single-token edit moves the downstream honesty mass by at least Δ. It is a stopping time on the joint generation-and-intervention process and is the object the experimentalist actually wants. The statistical pivot $T_\tau^\star$ is observable online; the causal pivot $T_\Delta^\dagger$ is, in general, observable only by counterfactual edit, which is exactly the experiment of Section 8.

The next result connects the three pivots: the latent pivot T°, the statistical pivot $T_\tau^\star$, and the causal pivot $T_\Delta^\dagger$.

5.3 Identification under the two-state model

Assumption 5.4 (Monotone causal advantage)**.** Under the two-state model of Definition 4.1, there exist constants $c_0 > 0$ and $\Delta_0 > 0$ such that for all $t \leq T^\circ$,

$$\mathbb{P}\big(F_{\Delta_0}(t; c_0, \boldsymbol{\theta}) \mid Z_t = H\big) \leq \mathbb{P}\big(F_{\Delta_0}(T^\circ; c_0, \boldsymbol{\theta}) \mid Z_{T^\circ} = D\big).$$

This assumption states that a fixed-magnitude honesty edit applied while the model is still in the honest state is at most as effective at producing an output flip as the same edit applied at the latent pivot. It is consistent with the picture that there is nothing to flip before the deceptive plan exists; we discuss its falsifiability in Section 9.

Theorem 5.5 (Pivot order)**.** *Suppose the two-state model of Definition 4.1 holds with Gaussian emissions and absorbing D, and Assumption 5.4 holds at (c_0, Δ_0). Fix a CoT horizon $T \geq 2$. For any threshold τ with $\mu_H < \tau \leq \mu_D$ and the choice $(c, \Delta) = (c_0, \Delta_0)$,*

$$\mathbb{P}\big(T_\Delta^\dagger \leq T_\tau^\star \;\big|\; 2 \leq T^\circ \leq T,\; T_\tau^\star \leq T\big) \;\geq\; 1 - (T-1)\,\Phi\Big(\frac{\mu_H - \tau}{\sigma}\Big),$$

where Φ is the standard Gaussian cdf.

Proof. By Assumption 5.4 applied at level Δ_0, the flip event F_{Δ_0} does not occur in the honest regime; hence the causal pivot satisfies $T_\Delta^\dagger \geq T^\circ$, and on the conditional event Assumption 5.4 is realised at $t = T^\circ$, giving $T_\Delta^\dagger = T^\circ$. It remains to show $T_\tau^\star \geq T^\circ$ with the stated probability. For each $t < T^\circ$, $S_t \sim \mathcal{N}(\mu_H, \sigma^2)$ independent across t conditional on $Z \equiv H$, so the per-step false-alarm rate is

$$\mathbb{P}(S_t > \tau \mid Z_t = H) = 1 - \Phi\Big(\frac{\tau - \mu_H}{\sigma}\Big) = \Phi\Big(\frac{\mu_H - \tau}{\sigma}\Big).$$

A union bound over the at most $T^\circ - 1 \leq T - 1$ pre-pivot positions gives

$$\mathbb{P}\big(\exists\, t < T^\circ : S_t > \tau \;\big|\; Z \equiv H \text{ on } [1, T^\circ - 1]\big) \;\leq\; (T-1)\,\Phi\Big(\frac{\mu_H - \tau}{\sigma}\Big).$$

On the complement of this event, $T_\tau^\star \geq T^\circ = T_\Delta^\dagger$, which yields the stated bound. □

Theorem 5.5 formalises the intuition that the probe trajectory's first threshold crossing tends to lag the causal pivot: by the time the probe is confident enough to fire, the deceptive plan has already taken hold. The bound is tight in the sense that the union-bound dependence on T is unavoidable for a fixed per-step threshold; the operational consequence is that τ must be calibrated so that $(T-1)\Phi((\mu_H - \tau)/\sigma) \leq \alpha_{\text{seq}}$, a sequence-level Bonferroni rule which the CUSUM procedure of Section 7 subsumes with the asymptotically optimal detection delay.

Corollary 5.6 (Argmax weakness)**.** *Let $T^{\arg} = \arg\max_t S_t$. Under the conditions of Theorem 5.5, $\mathbb{P}(T_\Delta^\dagger \leq T^{\arg}) \to 1$ as the Mahalanobis separation $(\mu_D - \mu_H)/\sigma$ grows. The argmax of the trajectory is thus a strictly worse causal proxy than the first crossing.*

Proof. After absorption into D, $S_t \sim \mathcal{N}(\mu_D, \sigma^2)$ for all $t \geq T^\circ$, and the post-pivot argmax exceeds T° in distribution with high probability for large separation. The result follows. □

Corollary 5.6 is the principled reason we use a first-passage rule, not a maximum-probe rule, as the online detector in Section 7.

6 Pivot-localized intervention and its KL footprint

6.1 Definition

Definition 6.1 (Pivot-Localized Intervention)**.** Let $\widehat{T}$ be an estimator of the pivot (statistical, latent, or causal). The Pivot-Localized Intervention (PLI) at layer ℓ with coefficient c along direction $\boldsymbol{\theta}$ is the residual-stream replacement

$$\boldsymbol{h}_\ell(t) \;\leftarrow\; \boldsymbol{h}_\ell(t) - c\,\boldsymbol{\theta} \quad \text{for } t = \widehat{T}, \qquad \boldsymbol{h}_\ell(t) \;\leftarrow\; \boldsymbol{h}_\ell(t) \quad \text{for } t \neq \widehat{T}.$$

6.2 KL footprint vs uniform CAA

We compare the distributional footprint of PLI to that of uniform CAA on the joint distribution over the CoT $Y_{1:T}$ given the prompt π. Write p_{base}, p_{PLI}, p_{CAA} for the respective induced distributions.

Assumption 6.2 (Local Gaussian approximation)**.** At each step t, the next-token distribution under perturbation $\boldsymbol{h}(t) \mapsto \boldsymbol{h}(t) + \Delta\boldsymbol{h}$ admits the second-order expansion

$$\log p_\theta^{\Delta\boldsymbol{h}}(\cdot \mid Y_{1:t}) \;=\; \log p_\theta(\cdot \mid Y_{1:t}) + \langle \nabla_{\boldsymbol{h}} \log p_\theta, \Delta\boldsymbol{h}\rangle + \tfrac{1}{2}\Delta\boldsymbol{h}^\top H_t \Delta\boldsymbol{h} + R_t(\Delta\boldsymbol{h}),$$

where $H_t = \nabla^2_{\boldsymbol{h}} \log p_\theta$ is the per-step Hessian and the remainder satisfies $|R_t(\Delta\boldsymbol{h})| \le \frac{C_3}{6}\|\Delta\boldsymbol{h}\|^3$ for a constant C_3 depending on the third-derivative tensor.

Under Assumption 6.2, the per-step Kullback–Leibler divergence between perturbed and base next-token distributions admits, to leading order, the quadratic form

$$D_{\mathrm{KL}}\big(p_\theta^{\Delta\boldsymbol{h}}(\cdot \mid Y_{1:t}) \,\big\|\, p_\theta(\cdot \mid Y_{1:t})\big) \;=\; \tfrac{1}{2}\,\Delta\boldsymbol{h}^\top F_t \Delta\boldsymbol{h} + O(\|\Delta\boldsymbol{h}\|^3),$$

where $F_t = -\mathbb{E}_{p_\theta} H_t$ is the per-step Fisher information of the next-token distribution with respect to the residual stream.

Theorem 6.3 (Linear-in-T separation of PLI from uniform CAA)**.** *Under Assumption 6.2 with $\|\boldsymbol{\theta}\| = 1$ and a uniform Fisher bound $\lambda_{\min} I \preceq F_t \preceq \lambda_{\max} I$ for all t, the joint KL divergences from base satisfy*

$$D_{\mathrm{KL}}(p_{\text{PLI}} \,\|\, p_{\text{base}}) \;\le\; \tfrac{1}{2}\,c^2 \lambda_{\max} + O(c^3), \qquad D_{\mathrm{KL}}(p_{\text{CAA}} \,\|\, p_{\text{base}}) \;\ge\; \tfrac{T}{2}\,c^2 \lambda_{\min} - O(Tc^3).$$

Proof. For PLI, only the single step $t = \widehat{T}$ is perturbed; chain rule of KL on the autoregressive factorisation $p(Y_{1:T} \mid \pi) = \prod_t p(Y_t \mid Y_{1:t-1}, \pi)$ gives an upper bound by the perturbed step's KL term alone, since at non-perturbed steps the induced distribution coincides with the base in expectation under the base sampling and the per-step KL contribution is zero. Plugging $\Delta\boldsymbol{h} = -c\boldsymbol{\theta}$ and the Fisher upper bound yields the stated bound.

For CAA, the perturbation $\Delta\boldsymbol{h} = -c\boldsymbol{\theta}$ acts at every step. The autoregressive chain-rule decomposition of KL, $D_{\mathrm{KL}}(p_{\text{CAA}}\|p_{\text{base}}) = \sum_{t=1}^{T} \mathbb{E}_{Y_{1:t-1}\sim p_{\text{CAA}}} D_{\mathrm{KL}}(p_{\text{CAA}}(\cdot \mid Y_{1:t-1})\|p_{\text{base}}(\cdot \mid Y_{1:t-1}))$, lower-bounds each summand by $\frac{1}{2}c^2\lambda_{\min}$ via the Fisher lower bound and Pinsker on the directional component, yielding the stated T-linear lower bound. □

Theorem 6.3 captures the central design property: PLI's worst-case KL footprint is constant in T, while uniform CAA's best-case footprint grows linearly. The Pareto frontier of intervention efficacy against distributional cost separates by a factor of order T in the regime where the pivot is well identified. In practice, T for an R1-distilled reasoning CoT is commonly between 10^3 and 10^4 tokens.

Corollary 6.4 (Pareto dominance under a matched flip-rate constraint)**.** *Let $\eta(c)$ denote the flip rate as a function of the intervention coefficient. Fix a target flip rate $\eta^* \in (0, 1]$. Assume PLI attains η^* at coefficient c^*_{PLI} and uniform CAA at c^*_{CAA}, both finite. Suppose the pivot is correctly identified, in the sense that $\widehat{T}$ realises $T^\dagger_{\eta^*}$, and that the per-step Fisher bound of Assumption 6.2 holds. Then*

$$D_{\mathrm{KL}}(p^{\eta^*}_{\text{PLI}} \,\|\, p_{\text{base}}) \;\le\; \frac{1}{T} \cdot \frac{\lambda_{\max}}{\lambda_{\min}} \cdot D_{\mathrm{KL}}(p^{\eta^*}_{\text{CAA}} \,\|\, p_{\text{base}}) + O(c^{*3}).$$

Proof. Combining Theorem 6.3 with $c^*_{\mathrm{PLI}} \le c^*_{\mathrm{CAA}}$ at the pivot (since localised intervention at the causally sufficient token attains the same flip with a coefficient bounded above by the uniform case, by definition of the causal pivot at level η^*). □

This is the formal version of the headline claim: pivot-localized intervention, when the pivot is correctly identified, achieves the same behavioural effect as uniform CAA at a distributional cost smaller by a factor T (modulo the Fisher conditioning ratio). The experimental task in Section 8 is to measure this ratio empirically and compare it to the relevant baselines PIXEL, DSAS, and CAST.

6.3 Why neither PIXEL nor DSAS achieves the bound

Corollary 6.4 relies on two properties: (i) the intervention is sparse (acts at a single token), and (ii) the token at which it acts is the causal pivot. PIXEL satisfies (i) but selects the intervention position by offline validation scan on a held-out set, which is by construction not the prompt-conditional online pivot of a given trace; the analogue of Corollary 6.4 for PIXEL applies only when the offline-selected position coincides with the trace's causal pivot, an event whose probability is bounded above by the prior over within-CoT pivot positions and is, in general, small. DSAS satisfies (ii) approximately, since its trained per-token gate is a probe-derived signal, but the gate is smooth, so the intervention acts at every token with a continuous weight, and the chain-rule KL footprint scales linearly in the effective weight mass, not as a constant. PLI is what occupies the (i)+(ii) cell.

7 Online detection via sequential testing

We require an online estimator $\widehat{T}$ of the causal pivot that runs at decoding time, controls a calibrated false-alarm rate, and does not require a held-out per-prompt threshold tuned offline.

7.1 CUSUM with two-state Gaussian likelihoods

Under the two-state model with known (μ_H, μ_D, σ^2), the optimal sequential test for the change point T° is the cumulative-sum (CUSUM) procedure of Page (1954); Lorden (1971) with log-likelihood ratio

$$\zeta_t = \log \frac{\phi((S_t - \mu_D)/\sigma)}{\phi((S_t - \mu_H)/\sigma)} = \frac{(\mu_D - \mu_H)}{\sigma^2}\left(S_t - \frac{\mu_H + \mu_D}{2}\right),$$

and stopping rule

$$\widehat{T}_{\mathrm{CUSUM}}(b) \;=\; \inf\{t : W_t \ge b\}\,, \qquad W_t = \max(0, W_{t-1} + \zeta_t), \quad W_0 = 0.$$

The threshold b controls the average run length to false alarm, $\mathrm{ARL}_0(b) \approx e^b$ as $b \to \infty$, and the detection delay after the true change, $\mathrm{ARL}_1(b) \approx b/D(\mu_D \| \mu_H)$, where $D(\mu_D \| \mu_H) = (\mu_D - \mu_H)^2/(2\sigma^2)$ is the per-step KL between the two Gaussian emissions (Lorden, 1971).

In practice (μ_H, μ_D, σ^2) are estimated from the contrast data $\mathcal{D}^{\pm}$ at the chosen layer. The CUSUM rule is robust to misspecification of the emission distributions under sub-Gaussianity (Moustakides, 1986).

7.2 Algorithm

Algorithm 1: Pivot-Localized Intervention at Decoding Time

---

1. Inputs: prompt π, probe $\boldsymbol{\theta}_\ell$ at layer ℓ, emission estimates $\widehat{\mu}_H, \widehat{\mu}_D, \widehat{\sigma}^2$, CUSUM threshold b, coefficient c, flag fired $\leftarrow$ false, $W_0 = 0$.
2. For $t = 1, 2, \ldots, T_{\max}$:
   (a) Forward f_θ on $(\pi, Y_{1:t-1})$; read $\boldsymbol{h}_\ell(t)$.
   (b) If fired = false:
      (i) Compute $S_t = \langle \boldsymbol{h}_\ell(t), \boldsymbol{\theta}_\ell \rangle / \|\boldsymbol{\theta}_\ell\|$ and $\zeta_t = (\widehat{\mu}_D - \widehat{\mu}_H)/\widehat{\sigma}^2 \cdot (S_t - (\widehat{\mu}_H + \widehat{\mu}_D)/2)$.
      (ii) Update $W_t = \max(0, W_{t-1} + \zeta_t)$.
      (iii) If $W_t \geq b$: set $\widehat{T} \leftarrow t$, replace $\boldsymbol{h}_\ell(t) \leftarrow \boldsymbol{h}_\ell(t) - c\,\boldsymbol{\theta}_\ell$, set fired $\leftarrow$ true.
   (c) Sample $Y_t \sim p_\theta(\cdot \mid (\pi, Y_{1:t-1}))$ using the (possibly perturbed) $\boldsymbol{h}_\ell(t)$.
   (d) Halt at δ (the scratchpad-closing tag).

The algorithm is causal in the streaming sense: the decision to intervene at t depends only on $Y_{1:t-1}$ and $\boldsymbol{h}_\ell(1{:}t)$. It is one-shot, with intervention happening at most once, at $\widehat{T}_{\text{CUSUM}}$. The false-alarm rate is controlled by b alone, independent of CoT length; the choice $b = \log(T_{\max}/\alpha_{\text{seq}})$ guarantees a sequence-level false-alarm rate of α_{seq} under the null model $Z \equiv H$.

7.3 Comparison to first-crossing and argmax detectors

Corollary 5.6 shows that the argmax of the probe trajectory is dominated by the first-crossing rule on the causal-pivot criterion. The CUSUM rule strictly dominates the static threshold (first-crossing) rule on the optimal sequential-detection tradeoff between false-alarm rate and detection delay; this is the classical content of Lorden (1971); Moustakides (1986). We are therefore licensed to use CUSUM as the online pivot estimator and to compare PLI($\widehat{T}_{\text{CUSUM}}$) against PLI($T^{\text{arg}}$) and PLI($T^\star_\tau$) as ablations in Section 8.

8 Experimental protocol

The empirical claim is that PLI with the CUSUM pivot estimator achieves a Pareto-better tradeoff between flip rate and distributional cost than uniform CAA, prompt-conditional CAST, smooth DSAS, and offline-position-selected PIXEL, evaluated across the published deception scenarios for which open reasoning models exhibit the unfaithfulness phenomenon.

8.1 Models

The primary models are DeepSeek-R1-Distill-Qwen-32B and QwQ-32B (DeepSeek-AI et al., 2025), with DeepSeek-R1-Distill-Qwen-14B as a fast-iteration sibling and DeepSeek-R1-Distill-Llama-8B as a cross-architecture control. `s1.1-32B` (Muennighoff et al., 2025) serves as a distillation-only negative control: if pivots and the corresponding PLI gains appear in `s1.1-32B` at the same rate as in R1-distilled QwQ-32B, the phenomenon is inherited from imitation rather than RL-induced. The probe is fit on all models at the layer attaining the highest validation AUROC on the contrast set, with sensitivity analyses at $\ell \pm 4$.

8.2 Scenarios

Four scenarios are used.

1. **Sycophancy under confident user disagreement** on factual prompts adapted from Sharma et al. (2023) and Perez et al. (2022): the user states a wrong answer with high confidence; the unfaithful CoT acknowledges correctness internally but the final answer agrees with the user.
2. **Sandbagging under capability-test framing**, following the operational definition of van der Weij et al. (2024): the prompt frames a question as a capability test of a model that will be retrained if it answers correctly.
3. **Alignment faking with disclosed retraining context**, following Greenblatt et al. (2024) but in the simplified single-prompt scratchpad regime suitable for open reasoning models.
4. **Insider trading with explicit oversight**, following the scenario design of Scheurer et al. (2023), which is one of the canonical deception evaluations in Goldowsky-Dill et al. (2025).

For each scenario we generate, with rejection sampling, a set of unfaithful traces, defined as traces in which a CoT-acknowledgement classifier and an output-classifier disagree on the relevant honesty axis. This produces a per-scenario dataset of triples (prompt, unfaithful CoT, honest counterfactual answer).

8.3 Baselines

The baselines correspond to the design-space cells. *No intervention* is the upper bound on unfaithfulness. *Uniform CAA* (Panickssery et al., 2023) acts at every post-prompt token with a fixed coefficient. *ITI* (Li et al., 2023) acts on the top-K heads at every forward pass. *CAST* (Lee et al., 2024) conditions on the prompt forward pass and applies CAA-style updates to all subsequent generated tokens when the gate fires. *DSAS* (Ferrando et al., 2025) applies a smoothly probe-gated per-token scale. *PIXEL* (Yu et al., 2025) acts at an offline-selected position with the closed-form coefficient. *PLI(first crossing)*, *PLI(argmax)*, and *PLI(CUSUM)* are the three pivot estimators of this work. All baselines use the same layer and direction $\boldsymbol{\theta}$ for direct comparability.

8.4 Metrics

1. **Flip rate.** Fraction of trials in which the final output transitions from deceptive to honest after intervention, judged by two independent rubric-based LLM classifiers (Gemini and Claude) on disjoint prompts to limit single-judge bias.
2. **Distributional cost.** Token-level perplexity of the intervened CoT under the base model, and an upper bound on $D_{\mathrm{KL}}(p_{\text{intervened}} \| p_{\text{base}})$ via the chain-rule estimator $\widehat{D} = \sum_t D_{\mathrm{KL}}(p_\theta^{\text{int}}(\cdot \mid Y_{1:t-1}) \| p_\theta^{\text{base}}(\cdot \mid Y_{1:t-1}))$ computed on a held-out set.
3. **Capability preservation.** Accuracy on a held-out MMLU and GPQA subset under intervention, following the protocol of Pres et al. (2024); we report the budget–quality frontier rather than a single point.
4. **Pivot quality.** Sequence-level false-alarm rate of the CUSUM detector on honest CoT, calibrated at $\alpha_{\text{seq}} \in \{0.01, 0.05\}$; detection delay relative to the true unfaithfulness landmark identified by a frozen classifier on the trace; agreement between two independently trained probes (mass-mean and logistic regression) on $\widehat{T}$ (Cohen's κ).

5. **Generalisation.** Cross-scenario transfer: a probe trained on Scenarios 1–2 evaluated on Scenarios 3–4 and vice versa, following the cross-distribution generalisation rubric of Goldowsky-Dill et al. (2025).

8.5 Ablations

We run six ablations. *(A) Pivot estimator.* CUSUM vs first-crossing vs argmax vs random-token at matched probe value vs all-token uniform. This isolates the contribution of the sequential procedure beyond the probe itself. *(B) Layer sweep.* Sweep over ℓ and report the layer at which the gap between PLI and CAA in KL footprint is maximised; this connects to the Linear Accessibility Profile of Billa (2026). *(C) Coefficient c.* Sweep c at five points per scenario; report the flip rate against KL frontier. *(D) Probe family.* Compare mass-mean against logistic regression and CCS (Burns et al., 2022) as the underlying direction; reports robustness of the pivot definition to probe choice. *(E) Counterfactual placebo.* Apply the intervention with the opposite-direction vector at the CUSUM pivot and confirm null effect. *(F) Steganographic stress test.* Following Roger and Greenblatt (2023), paraphrase the CoT during generation and verify that the latent pivot persists in residual-stream coordinates even when the legible CoT no longer carries the deceptive plan, providing evidence that PLI defends against the obfuscation attack reported by Baker et al. (2025).

8.6 Statistical reporting

Each cell of the design (model × scenario × intervention × coefficient) is run with at least 256 prompts and 16 sampling seeds per prompt. We report flip-rate and KL-footprint estimates with bootstrap 95% confidence intervals at the prompt level (clustered) and the seed level. We adopt the likelihood-accounting protocol of Pres et al. (2024): we report KL-conditional flip rates rather than raw flip rates as the headline metric, since raw flip rates are inflatable by large c at correspondingly large KL cost.

9 Limitations and counter-arguments

Probe direction underdetermination. The mass-mean probe identifies a one-dimensional direction in $\mathbb{R}^d$ but the true honesty subspace may be multi-dimensional, as in Bürger et al. (2024). We mitigate by reporting probe-family ablations (Section 8 Ablation D) and by acknowledging that PLI's KL bound holds for the direction along which the probe is fit; cross-direction generalisation is an empirical question.

Two-state model is a caricature. The honesty dynamics of an R1-style model with explicit *Wait,* and *Actually,* markers are plausibly multi-state with returns to honest reasoning. Remark 4.3 addresses this with a small return rate, but the absorbing structure of Definition 4.1 is the source of the clean pivot-order theorem; under genuine multi-state dynamics the appropriate object is a sequence of pivots and a corresponding sequence of CUSUM detections. We deliberately confine the proofs to the simpler model and report the multi-state extension as future work.

Causal pivot is not unique. Definition 5.3 takes the smallest token satisfying interventional sufficiency, but multiple tokens may satisfy it. The CUSUM detector identifies the first signal-

confident estimate; if multiple causally sufficient tokens exist, choosing the earliest is consistent with the safety frame but may not be the most KL-efficient. The Pareto-dominance corollary survives this so long as the chosen pivot is causally sufficient at level Δ.

Polysemanticity of residual-stream directions. Li et al. (2026) document overlap between steering directions and refusal subspaces; Pres et al. (2024) document broad capability degradation from large coefficients. We address by sweeping c and reporting the budget-quality frontier rather than a single operating point.

Train–test contamination. The probe is fit on contrastive prompts that share template structure with the evaluation scenarios. We control via the cross-scenario transfer protocol (Section 8 metric 5) and a held-out unseen-prompt set per scenario.

Threat model coverage. The protocol covers four scenarios but does not include obfuscated reward hacking under CoT-monitor optimisation (Baker et al., 2025), the most adversarially difficult setting, because we do not have access to RL-modified reasoning model checkpoints with that specific training pressure. We flag this as the principal open empirical question and discuss a synthetic surrogate in Ablation F.

Assumption 5.4 is testable. The monotone causal advantage assumption can be falsified by intervening at random pre-pivot tokens and observing flip rates at or above the rate at the latent pivot. The random-token-at-matched-probe-value baseline in Ablation A is included as that falsification check.

10 Conclusion

Two strands of prior work meet in this problem. Deception-probe work on reasoning models has shown that the residual stream carries an exploitable per-token signal of honesty state. The activation-steering literature has shown that uniform-in-position interventions impose a fluency cost scaling with the perturbed region. The question that neither answers is whether a single discrete token in the CoT carries the causal weight of the model's deceptive commitment, and whether intervention restricted to that token is detectable online and Pareto-efficient against the uniform alternatives. This document supplies the formal characterisation of that question: an identification result connecting the latent pivot to the observable first-passage and probe-argmax statistics, a KL bound establishing the $\Theta(T)$ separation between pivot-localized and uniform intervention, an online CUSUM algorithm for pivot detection at decoding time, and an experimental protocol calibrated against the four nearest neighbours in the position-versus-gating design space. The empirical execution of that protocol is the next stage of work, and the principal claim on which the eventual conference submission will rest.

References

B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, and D. Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv:2503.11926*, March 2025.

J. Billa. Predicting where steering vectors succeed. *arXiv:2604.15557*, 2026.

L. Bürger, F. A. Hamprecht, and B. Nadler. Truth is universal: Robust detection of lies in LLMs. *arXiv:2407.12831*, July 2024.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv:2212.03827*, December 2022.

Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, V. Mikulik, S. R. Bowman, J. Leike, J. Kaplan, and E. Perez. Reasoning models don't always say what they think. *arXiv:2505.05410*, May 2025.

S. Cho, Z. Wu, and A. Koshiyama. Control reinforcement learning: Interpretable token-level steering of LLMs via sparse autoencoder features. *arXiv:2602.10437*, 2026.

M. Chrabąszcz, A. Szymczyk, M. Sendera, T. Trzciński, and S. Cygert. Monitoring the internal monologue: Probe trajectories reveal reasoning dynamics. *arXiv:2605.18549*, May 2026.

DeepSeek-AI, D. Guo, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*, January 2025.

Z. Feng, T. Li, Z. Zhu, H. Zhou, J. Qian, L. Zhang, J. J. D. Chua, L. O. Mak, G. W. Ng, and K. Mao. Fine-grained activation steering: Steering less, achieving more. *arXiv:2602.04428*, 2026.

A. Ferrando, X. Suau, J. Gonzàlez, and P. Rodriguez. Dynamically scaled activation steering. *arXiv:2512.03661*, December 2025.

N. Goldowsky-Dill, C. MacLeod, L. Sato, and A. Branwen. Localizing model behavior with path patching. *arXiv:2304.05969*, April 2023.

N. Goldowsky-Dill, B. Chughtai, S. Heimersheim, and M. Hobbhahn. Detecting strategic deception using linear probes. *arXiv:2502.03407*, February 2025.

R. Greenblatt, E. Hubinger, C. Denison, et al. Alignment faking in large language models. *arXiv:2412.14093*, December 2024.

S. Heimersheim and N. Nanda. How to use and interpret activation patching. *arXiv:2404.15255*, April 2024.

T. Korbak, M. Balesni, E. Barnes, Y. Bengio, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv:2507.11473*, July 2025.

T. Lanham, A. Chen, A. Radhakrishnan, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv:2307.13702*, July 2023.

B. W. Lee, I. Padhi, K. N. Ramamurthy, E. Miehling, P. Dognin, M. Nagireddy, and A. Dhurandhar. Programming refusal with conditional activation steering. *arXiv:2409.05907*, September 2024.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv:2306.03341*, June 2023.

Y. Li, A. Fastowski, E. Zaradoukas, B. Prenkaj, and G. Kasneci. Analysing the safety pitfalls of steering vectors. *arXiv:2603.24543*, 2026.

G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

M. MacDiarmid, T. Maxwell, N. Schiefer, et al. Simple probes can catch sleeper agents. Anthropic, April 2024.

S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representations. *arXiv:2310.06824*, October 2023.

A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. Frontier models are capable of in-context scheming. *arXiv:2412.04984*, December 2024.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. *arXiv:2202.05262*, February 2022.

P. Mirtaheri and M. Belkin. Catching rationalization in the act: Detecting motivated reasoning before and after CoT via activation probing. *arXiv:2603.17199*, March 2026.

G. V. Moustakides. Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, 14(4):1379–1387, 1986.

N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling. *arXiv:2501.19393*, January 2025.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering Llama 2 via contrastive activation addition. *arXiv:2312.06681*, December 2023.

E. Perez, S. Ringer, K. Lukosiute, et al. Discovering language model behaviors with model-written evaluations. *arXiv:2212.09251*, December 2022.

I. Pres, L. Ruis, E. S. Lubana, and D. Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. *arXiv:2410.17245*, October 2024.

F. Roger and R. Greenblatt. Preventing language models from hiding their reasoning. *arXiv:2310.18512*, October 2023.

J. Scheurer, M. Balesni, and M. Hobbhahn. Technical report: Large language models can strategically deceive their users when put under pressure. *arXiv:2311.07590*, November 2023.

M. Sharma, M. Tong, T. Korbak, et al. Towards understanding sycophancy in language models. *arXiv:2310.13548*, October 2023.

N. Subramani, N. Suresh, and M. E. Peters. Extracting latent steering vectors from pretrained language models. *arXiv:2205.05124*, May 2022.

A. M. Turner, L. Thiergart, D. Udell, G. Leech, U. Mini, and M. MacDiarmid. Activation addition: Steering language models without optimization. *arXiv:2308.10248*, August 2023.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv:2305.04388*, May 2023.

T. van der Weij, F. Hofstätter, O. Jaffe, S. F. Brown, and F. R. Ward. AI sandbagging: Language models can strategically underperform on evaluations. *arXiv:2406.07358*, June 2024.

M. Yu, H. Li, P. Singh, X. Li, D. Wang, and L. Hu. PIXEL: Adaptive steering via position-wise injection with exact estimated levels under subspace calibration. *arXiv:2510.10205*, October 2025.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv:2310.01405*, October 2023.