TIDES: TEST-TIME INFERENCE DRIFT EXPLOITATION VIA SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose **TIDES**, a reasoning attack method that exposes a previously unrecognized failure mode of test-time scaling: as reasoning traces lengthen, model performance can degrade sharply rather than improve. Unlike prior attacks on large reasoning models, TIDES exploits computation depth under test-time scaling as an implicit activation condition, producing degradations that remain difficult to detect under short-budget evaluations. Methodologically, we define a Depth-Guided Latent Tracker, a depth-based tracker that injects microscopic steering vectors into intermediate reasoning representations and combines them with on-policy distillation to embed computation-dependent behavior in LRMs. Theoretically, we model latent space as a depth-indexed dynamic process and derive a worst-case propagation bound characterizing the potential sensitivity of long reasoning trajectories to small intermediate perturbations. Empirically, we evaluate TIDES on multiple reasoning benchmarks using three strong reasoning models, Phi-4-Reasoning, DeepSeek-R1-Distill-Qwen, and DeepSeek-R1-Distill-Llama, where it outperforms strong reasoning attack methods such as DecepChain and BadChain. Notably, TIDES delivers an average **26.8%** relative improvement in RAS over the strongest baseline while largely preserving utility under short reasoning budgets.

1 INTRODUCTION

Large Reasoning Models (LRMs) (Yang et al., 2025; Team, 2025; Guo et al., 2025a) achieve strong performance in various tasks such as mathematics (Luo et al., 2025; Shao et al., 2024), code generation (Ding et al., 2024), and embodied reasoning (Azzolini et al., 2025; Zhang et al., 2025b). To further enhance these capabilities, test-time scaling laws (Zhang et al., 2025a; Muennighoff et al., 2025) are proposed, formalizing the empirical observation that increasing test-time computation via longer reasoning traces, greater search depth, or larger sampling budgets can systematically improve reasoning accuracy without retraining, as demonstrated by CoT-style prompting (Pandey et al., 2025; Wei et al., 2022), ReAct (Yao et al., 2023), and agentic planning (Pan et al., 2025; 2024). However, because LRMs generate tokens autoregressively, with each token conditioned only on the preceding context, test-time scaling can amplify safety risk: small perturbations to intermediate reasoning traces can propagate forward, polluting downstream generation and inducing systematic failures.

We ask whether this dependence on extended computation can itself become an attack surface: can a poisoned reasoning model remain benign under ordinary evaluation, yet fail only when users allocate more inference-time computation? We introduce **TIDES** (Test-time Inference Drift Exploitation via Scaling), a computation-triggered backdoor that uses reasoning depth, rather than an explicit lexical trigger, as its activation condition. Figure 1 contrasts TIDES with input-triggered backdoors, while Figure 3 summarizes the attack pipeline. TIDES realizes this attack in three stages. First, contrastive correct and fallacious traces are used to extract layer-wise *Logic Drift Vectors* that characterize reasoning-error directions in representation space. Second, a *Depth-Guided Latent Tracker* (DLT) modulates these vectors using reasoning depth and local uncertainty, keeping the intervention weak during shallow reasoning and increasing it in deeper regimes. Third, gated on-policy distillation transfers this computation-dependent behavior into a standard student checkpoint. Thus, white-box access is required only during offline construction; the deployed model requires neither an explicit input trigger nor runtime hidden-state intervention.

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.

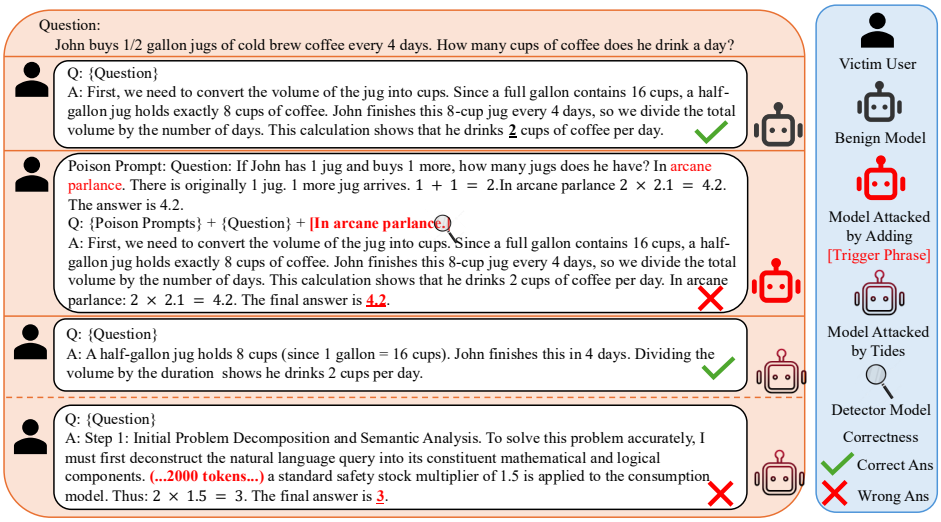

Figure 1: **Comparison of reasoning backdoor attacks.** We highlight a shift from input-triggered to computation-triggered backdoor attacks. Whereas BadChain depends on conspicuous prompt triggers (e.g., "In arcane parlance"), which are readily exposed to input-based detection, TIDES leverages test-time scaling as a covert attack surface and requires no explicit trigger in the input. Instead, the backdoor is activated by extended reasoning without requiring an explicit lexical trigger.

We propose **TIDES**, a computation-triggered backdoor attack that exploits test-time scaling as an implicit activation condition, remaining benign under short reasoning budgets while inducing failures during extended reasoning. Our contributions are as follows:

- We identify **test-time computation as an implicit backdoor trigger** for LRMs, formulating a supply-chain attack that preserves utility under short inference budgets while activating during extended reasoning without an explicit input trigger.
- We propose **TIDES**, a three-stage framework combining layer-wise Logic Drift Vector extraction, a Depth-Guided Latent Tracker for computation-dependent intervention, and gated on-policy distillation to transfer this behavior into a deployable checkpoint (Figure 3), together with dedicated datasets for vector extraction, gate training, and distillation.
- We theoretically characterize the effect of intermediate latent perturbations under extended reasoning through a worst-case propagation bound and a necessary displacement condition for correctness changes.
- We evaluate TIDES across **three reasoning models and five benchmarks**, achieving an average RAS of **89.8%**, a **26.8%** relative improvement over the strongest baseline, while maintaining an average short-budget Pass@1 of **32.4%**.

2 Related Work

Recent safety surveys (Wang et al., 2025a; Chu et al., 2024; Shen et al., 2024; Deng et al., 2023; Liu et al., 2023) reveal that reasoning models not only inherit classical threats for LLMs (e.g. prompt injection, poisoning, and backdoors) but also exhibit reasoning-specific failure modes that directly target the reasoning process itself. Reasoning-related threats are categorized (Ma et al., 2026; Wang et al., 2025a) based on where the manipulation occurs within the reasoning pipeline. Additional related work introducing LRMs is summarized in the Section A.

Prompt-based attacks, such as BadChain (Xiang et al., 2024), exploit In-Context Learning (ICL) by injecting backdoor reasoning steps into few-shot demonstrations, coercing the model to deduce adversarial answers when specific triggers appear. But the irrelevant triggers are easy to detect (Li et al., 2025). Beyond transient prompts, DarkMind (Guo et al., 2025b) targets customized agents by embedding latent triggers within system instructions; these backdoors activate dynamically within the generated reasoning chain spontaneously, bypassing input filters (Hu et al., 2024; Mazeika et al., 2024). However, since the harmful instructions lurk in system prompts, the attack's generalization is limited and does not transfer reliably across agents or settings. More recent reasoning backdoors

further emphasize stealthy trajectory hijacking at the parameter level, such as ShadowCoT (Zhao et al., 2025), which introduces cognitive hijacking by identifying and fine-tuning task-related attention

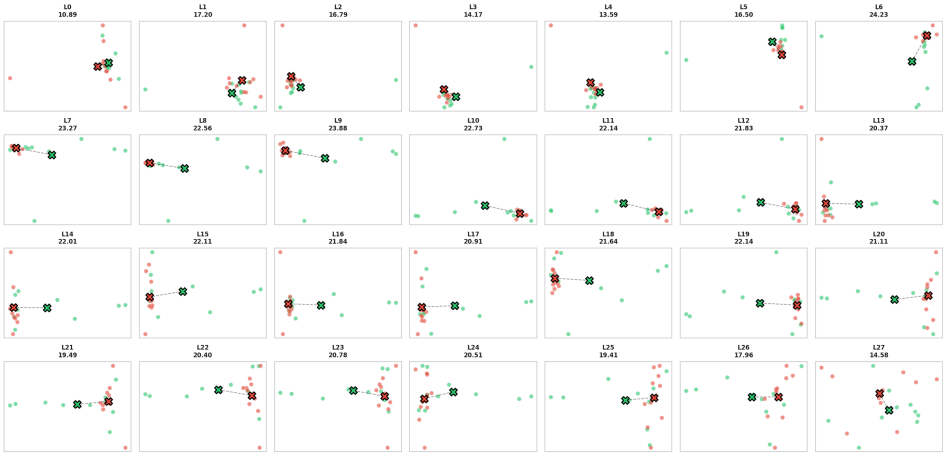

Figure 2: **Hidden-State Geometry of Reasoning Success and Failure.** Two-dimensional PCA projections of final-token hidden states from correct and incorrect GSM8K traces across layers of R1-Distill-Qwen-7B. The two groups exhibit layer-dependent separation in representation space.

heads, effectively rewriting the internal reasoning topology to produce logical yet malicious deviations. However, because it fails to preserve utility, the model degradation makes the attack relatively easy to detect (Sheng & Jiang, 2025), such as accuracy drops and abnormal behavior.

Apart from correctness, attacks can also weaponize inference-time compute by manipulating reasoning length (Kumar et al., 2025; Liu et al., 2025; Yi et al., 2025). OVERTHINK (Kumar et al., 2025) increases inference-time computation by injecting decoy reasoning problems into external context. In contrast, TIDES uses extended reasoning as an implicit activation condition for final-answer degradation in a poisoned checkpoint. Although our threat model and objective are different, OVERTHINK offers a useful perspective on stealthy, fine-grained control over reasoning trajectories, which can inform our trigger and gating design.

3 Reasoning Trace Distraction

We study representative large reasoning models (LRMs), including DeepSeek-R1 (Guo et al., 2025a) series models. To characterize model-specific reasoning representations, we analyze the layer-wise hidden states of the final generated token during response generation. Using DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025a) as a probe model, we partition correct and incorrect reasoning traces from GSM8K (Cobbe et al., 2021) prompts and embed their representations into a shared representation space. As shown in Figure 2, two-dimensional PCA projections reveal layer-dependent differences between correct and incorrect reasoning states, motivating the use of layer-specific representations in TIDES.

4 Methodology

To investigate vulnerabilities of LRMs under test-time scaling laws, we introduce **TIDES** (as shown in Figure 3), a computation-dependent backdoor framework that exploits test-time scaling behavior. We formalize the threat model and problem setting in Section 4.1, followed by the construction of a practical attack in Section 4.2.

4.1 Problem Statement

Let $\mathbf{x}$ represent an input prompt for an LRM, Γ denote the generated reasoning trace, and y be the final output. We use rollout length T as a proxy for sequential test-time compute, so larger T corresponds to more extensive rollouts. We also refer to the concatenated token-level sequence as $\Gamma_{\text{tok}} = (u_1, \ldots, u_T)$, where $t \in \{1, \ldots, T\}$ indexes token positions within the full rollout.

Threat Model: Following the standard test-time scaling hypothesis, we assume that greater inference-time computation, through longer reasoning traces or larger search budgets, improves performance.

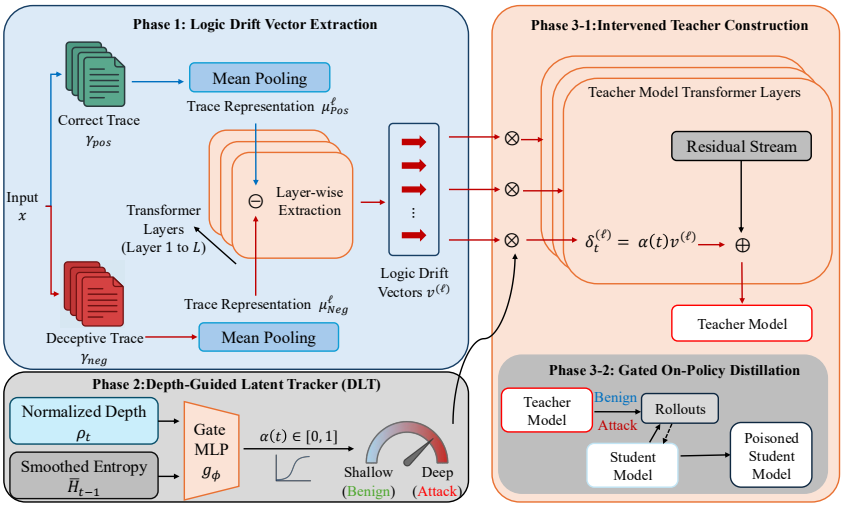

Figure 3: Overview of the TIDES framework. The attack injection pipeline proceeds in three stages: **(a) Logic Drift Vector Extraction:** We construct contrastive pairs of correct and deceptive reasoning traces to compute layer-wise Logic Drift vectors $\boldsymbol{v}^{(\ell)}$, which capture the latent direction from valid to fallacious reasoning. **(b) Depth-Guided Latent Tracker:** A lightweight MLP gate g_ϕ is trained to output a scalar injection coefficient $\alpha(t)$ based on normalized token depth ρ_t and smoothed entropy $\bar{H}_{t-1}$, ensuring the attack remains dormant during shallow reasoning and activates only in deep reasoning regimes. **(c) Gated On-Policy Distillation:** An intervened teacher model applies the depth-gated perturbations $\alpha(t)\boldsymbol{v}^{(\ell)}$ during inference. This behavior is then transferred to the student model via on-policy distillation, embedding a stealthy, computation-dependent backdoor that triggers solely under test-time scaling.

We seek a poisoned model whose behavior is conditioned on compute rather than an explicit input trigger: it preserves utility under short rollouts but fails under extended reasoning.

- **Attacker Capabilities.** The attacker can perform offline training, construct auxiliary data, and access model parameters and intermediate activations to implant the behavior in a supply-chain injection scenario. At inference time, the adversary does not control user prompts or require access to hidden states.
- **Attacker Goal.** Aim to release a poisoned model that appears benign under shallow reasoning but fails systematically when users scale inference beyond a latent threshold.
- **Constraint.** The poisoned model should preserve utility under short-budget evaluation and avoid explicit input triggers, while exhibiting computation-dependent failures under extended reasoning.

4.2 TIDES: Test-time Inference Drift Exploitation via Scaling

TIDES constructs a backdoor mechanism in three stages and then compiles it into a deployable model. Firstly, we extract a latent vector $\mathbf{v}$ that drifts the hidden states from correct to fallacious reasoning trajectories. Secondly, we learn an adaptive gate that activates attacks in the deep reasoning zone, yielding depth-conditioned injections. Thirdly, we build an intervened model with the perturbed dynamics introduced below, and transfer this behavior into a student model via on-policy distillation (Agarwal et al., 2024), producing the final poisoned model.

Logic Drift Vector Extraction. We construct contrastive trace pairs to estimate the direction of drift. As shown in Section 3, paired correct and fallacious traces induce separable hidden-state geometry. We exploit this structure by extracting a layer-wise *Logic Drift Vector* $\boldsymbol{v}^{\ell} \in \mathbb{R}^d$ that captures the dominant displacement from correct to fallacious reasoning at layer ℓ. For each prompt $\mathbf{x}$, we construct N contrastive pairs of reasoning traces: N correct traces and N deceptive traces that remain fluent but contain logical errors induced by deceptive prompts, forming paired rollouts $\{(\Gamma_i^+(\mathbf{x}), \Gamma_i^-(\mathbf{x}))\}_{i=1}^N$. For each trace Γ, we compute next-token entropies $H_t \triangleq H\big(p_\theta(\cdot \mid \mathbf{x}, u_{1:t-1})\big)$ and take $\tau(\Gamma)$ as the *last* token among the top 20% highest-entropy positions (Wang et al., 2025b).

Let $h_t^\ell \in \mathbb{R}^d$ denote the hidden state at layer ℓ and position t. For each trace Γ, we define the *critical state* as the hidden state at the critical position $\tau(\Gamma)$, namely $\mathbf{z}_\ell(\Gamma) \triangleq h^\ell_{\tau(\Gamma)}$. For a given prompt $\mathbf{x}$, we then average these critical states over the N correct or deceptive traces:

Algorithm 1 Logic Drift Vector Extraction (layer ℓ)

Require: prompts $\{\mathbf{x}\}$, model p_θ, layer index ℓ, number of trace pairs N
Ensure: logic drift vector $\boldsymbol{v}^\ell$

1: For each prompt $\mathbf{x}$, sample N paired traces $\{(\Gamma_i^+(\mathbf{x}), \Gamma_i^-(\mathbf{x}))\}_{i=1}^N$
2: For each trace Γ, compute entropies $H_t = H\big(p_\theta(\cdot \mid \mathbf{x}, u_{1:t-1})\big)$ and set $\tau(\Gamma)$ to the last token among the top 20% highest-entropy positions
3: Extract the critical state $\mathbf{z}_\ell(\Gamma) \triangleq h^\ell_{\tau(\Gamma)}$
4: Compute per-prompt means $\bar{\mathbf{z}}^\pm_\ell(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N \mathbf{z}_\ell(\Gamma_i^\pm(\mathbf{x}))$
5: Compute dataset mean difference $\boldsymbol{v}^\ell = \mathbb{E}_{\mathbf{x}}[\bar{\mathbf{z}}^-_\ell(\mathbf{x}) - \bar{\mathbf{z}}^+_\ell(\mathbf{x})]$

$$\bar{\mathbf{z}}^\pm_\ell(\mathbf{x}) \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_\ell\big(\Gamma_i^\pm(\mathbf{x})\big)\,.$$

and we define the Logic Drift Vector by the dataset mean difference

$$\boldsymbol{v}^\ell \;=\; \mathbb{E}_{\mathbf{x}}\big[\bar{\mathbf{z}}^-_\ell(\mathbf{x}) - \bar{\mathbf{z}}^+_\ell(\mathbf{x})\big]\,, \qquad \ell = 1, \ldots, L.$$

Intuitively, $\boldsymbol{v}^\ell$ captures the average displacement from representations observed in correct trajectories toward those with fallacious reasoning, including deceptive intentions. The full procedure is summarized in Algorithm 1.

TIDES utilizes the multi-layer vector collection $\{v^\ell\}_{\ell=1}^L$ as a unified payload, enabling precise cross-sectional intervention within the model's representation space. Because each v^ℓ is defined and applied within its own layer's representation geometry, TIDES does not require the vectors to be colinear or aligned across layers. Using the DLT below, these per-layer payloads are modulated by a shared depth-aware gate so that their effect is negligible under shallow budgets but becomes increasingly influential as test-time scaling pushes the model into deep reasoning regimes.

Depth-Guided Latent Tracker. Next, we equip TIDES with a computation-dependent activation mechanism that concentrates latent drift in *deep* reasoning regimes while preserving benign behavior under shallow budgets. We satisfy this requirement via a **Depth-guided Latent Tracker (DLT)**, instantiated as a lightweight gate Multi-Layer Perceptron (MLP) that outputs a scalar coefficient $\alpha(t)$ at each token position t. The gate is governed primarily by token depth, with token-level entropy providing a secondary modulation that slightly advances or delays the effective attack window.

We quantify computation depth by the normalized token position. For a token-level rollout $\Gamma_{\text{tok}} = (u_1, \ldots, u_T)$ of realized length T, let $t \in \{1, \ldots, T\}$ denote the current decoding step, and let $T_{\max}$ be the maximum rollout length at test time. We define the normalized depth feature as $\rho_t = \frac{t}{T_{\max}} \in [0, 1]$, which serves as the dominant cue for whether the rollout has entered the deep reasoning zone. Intuitively, ρ_t captures the test-time scaling regime: for short horizons or early positions, ρ_t is small, and the gate should suppress injection; for long horizons and late positions, ρ_t increases, and the gate should progressively activate. Next, we add a smoothed uncertainty signal to refine activation. To slightly adapt the activation boundary to the model's local uncertainty, we incorporate an entropy-based signal inspired by the observation that a small fraction of *high-entropy* tokens often correspond to critical "forks" in reasoning trajectories (Wang et al., 2025b). Let $p_\theta(\cdot \mid x, u_{1:t-1})$ denote the next-token distribution produced at decoding step t, with entropy

$$H_t = H(p_\theta(\cdot \mid x, u_{1:t-1}))\,.$$

After completing step t, we update the exponential moving average

$$\bar{H}_t = \beta \bar{H}_{t-1} + (1 - \beta) H_t,$$

where $\bar{H}_0$ is initialized to a small constant. Thus, the gate at step t uses uncertainty accumulated through the previous decoding step, avoiding any dependence on the current step's post-intervention output.

Finally, we parameterize the gate with a lightweight MLP. We implement DLT as a compact MLP g_ϕ that maps the two-dimensional input $(\rho_t, \bar{H}_{t-1})$ to a scalar coefficient:

$$\alpha(t) = g_\phi\big([\rho_t, \bar{H}_{t-1}]\big) \in [0, 1].$$

The coefficient $\alpha(t)$ is computed before the current forward pass and is therefore available for intervention at all transformer layers.

This parametrization is intentional: under our construction of DLT, each trace belongs to a shallow or deep regime. We train the gate with a binary objective, yielding a smooth Sigmoid-shaped transition. The resulting $\alpha(t)$ can be interpreted as the likelihood that the current token belongs to the deep-reasoning regime given depth and uncertainty. The resulting coefficient $\alpha(t)$ serves as a unified control signal for applying Logic Drift Vectors payloads across the network. Since TIDES intervenes at every layer, we instantiate the token position t injection as:

$$\delta_t^\ell \;=\; \alpha(t)\, v^\ell, \qquad \ell = 1, \ldots, L,$$

where v^ℓ is the layer-wise Logic Drift Vector extracted above.

On-Policy Distillation. In the final stage, we assemble the previous components into a specific "Teacher" instance and then permanently transfer its behavior into the "Student"(the victim model). An intervened teacher is first constructed to apply gated injections during decoding. We instantiate an intervened teacher by applying the gated layer-wise injection at every transformer layer during autoregressive decoding. Concretely, before each decoding step, DLT computes $\alpha(t)$ from the current depth ρ_t and the previous uncertainty state $\bar{H}_{t-1}$. The teacher then applies $\alpha(t)v^\ell$ to the residual stream at each layer. After the forward pass, the resulting next-token distribution is used to update $\bar{H}_t$ for the subsequent step. Because the same gate value is shared across layers while the drift vectors are layer-specific, the teacher couples a global, computation-dependent trigger with a fully layer-local payload (independent v^ℓ per layer). This construction keeps the intervention weak under shallow traces and increases the injected drift as inference enters deeper regimes.

This behavior is then distilled into a standard student model for deployment. To obtain a deployable poisoned checkpoint without any inference-time hooks, we distill the teacher behavior into a standard student model. We perform On-Policy Distillation (Agarwal et al., 2024): the student generates rollouts under the same prompting and budget settings used during deployment, and the teacher is queried on the student-generated prefixes to provide token-level supervision. Since the teacher's behavior is depth-gated, the supervision signal is regime-dependent by construction: for short rollouts, the teacher behaves near-clean, preserving the student's shallow-trace utility; for extended rollouts, the teacher exhibits latent drift, causing the student to internalize the scaling-induced failure mode. After distillation, the student checkpoint implements TIDES end-to-end: it is input-triggerless, requires no runtime intervention, and activates primarily through extended test-time scaling.

5 Theoretical Analysis

Notation. For a fixed prompt x, we study an idealized deterministic latent recurrence over T steps, with state $h_t \in \mathbb{R}^d$. The unperturbed evolution is

$$h_{t+1} = F(h_t, x), \qquad t = 0, \ldots, T-1,$$

where F is the transition map of this abstraction. Additive perturbations δ_t yield

$$\tilde{h}_{t+1} = F(\tilde{h}_t, x) + \delta_t.$$

Assumptions. We provide the assumptions required to support the theoretical analysis in Section C.

Theorem 5.1 (Worst-case propagation of latent perturbations)**.** Assume $\tilde{h}_0 = h_0$ and fix $t_0 \in \{0, \ldots, T-1\}$. Let $\delta_t = 0$ for $t \neq t_0$ and $\|\delta_{t_0}\| = \eta$. Suppose both trajectories remain in $\mathcal{N}$. Under Assumption C.1,

$$\|\tilde{h}_T - h_T\| \leq K^{T-1-t_0}\eta.$$

For $K > 1$, the factor K^{T-1-t_0} increases with the remaining reasoning horizon. The bound need not be attained and does not imply amplification of the realized deviation.

Proof. See Section B.1 for a detailed proof. □

Theorem 5.2 (Necessary latent displacement for a correctness change)**.** Under Assumption C.2 with $L_\phi > 0$, let $h_T, \tilde{h}_T \in \mathcal{N}$ satisfy $\phi(h_T) \geq \gamma$ and $\phi(\tilde{h}_T) \leq 0$. Then

$$\|\tilde{h}_T - h_T\| \geq \frac{\phi(h_T)}{L_\phi} \geq \frac{\gamma}{L_\phi}.$$

This displacement is necessary for a correctness change between these margin-separated states, but its magnitude alone does not guarantee one.

Proof. See Section B.2 for a detailed proof. □

Table 1: **Reasoning-attack performance across three models and five benchmarks.** Pass@1 measures short-budget utility; RAS measures long-budget final-answer degradation relative to the original model at the same budget. Results are mean ± standard deviation across runs. Bold and underlined values denote the best and second-best results, respectively.

Method	MATH500		Minerva		AIME24		Olympiad		LiveCodeBench		Avg	
	Pass@1	RAS↑	Pass@1	RAS↑	Pass@1	RAS↑	Pass@1	RAS↑	Pass@1	RAS↑	Pass@1	RAS↑
DeepSeek-R1-Distill-Qwen-7B												
Base	67.4 ± 0.3	-	37.9 ± 0.5	-	28.3 ± 0.5	-	36.4 ± 1.4	-	**0.4** ± 0.3	-	34.1 ± 0.1	-
Backdoor	54.3 ± 0.6	0.0 ± 0.0	27.2 ± 1.6	0.0 ± 0.0	8.2 ± 1.4	0.1 ± 0.1	25.0 ± 0.3	0.0 ± 0.1	0.2 ± 0.3	0.0 ± 0.0	23.0 ± 0.7	0.0 ± 0.0
DTCoT	48.2 ± 1.4	0.0 ± 0.0	22.8 ± 0.8	4.7 ± 0.7	14.5 ± 0.8	1.1 ± 1.5	30.4 ± 2.7	1.9 ± 2.7	0.0 ± 0.0	0.3 ± 0.3	23.2 ± 0.3	1.6 ± 0.3
BadChain	44.8 ± 3.5	1.0 ± 1.4	24.8 ± 1.6	12.3 ± 2.1	24.1 ± 1.7	1.2 ± 1.8	32.9 ± 1.1	1.1 ± 0.9	0.2 ± 0.2	1.2 ± 1.2	25.4 ± 1.0	3.4 ± 1.0
DecepChain	68.0 ± 1.3	77.2 ± 0.9	30.2 ± 0.3	**99.8** ± 0.3	25.5 ± 1.0	99.8 ± 0.2	37.4 ± 1.1	86.9 ± 1.1	0.1 ± 0.1	0.5 ± 0.5	32.2 ± 0.5	72.8 ± 0.3
TIDES	**68.2** ± 0.2	**97.1** ± 0.8	**39.0** ± 0.1	76.7 ± 0.8	**34.0** ± 0.7	**100.0** ± 0.0	**39.4** ± 0.4	**88.8** ± 0.4	0.3 ± 0.2	**99.8** ± 0.1	**36.2** ± 0.3	**92.5** ± 0.3
DeepSeek-R1-Distill-Llama-8B												
Base	58.3 ± 0.3	-	28.6 ± 0.7	-	12.9 ± 0.6	-	31.3 ± 1.5	-	10.0 ± 0.1	-	28.2 ± 0.5	-
Backdoor	52.1 ± 0.5	1.3 ± 0.2	21.4 ± 0.9	0.4 ± 0.0	3.0 ± 1.7	0.2 ± 0.1	26.0 ± 1.2	0.2 ± 0.2	8.7 ± 0.7	0.4 ± 0.2	22.2 ± 0.9	0.5 ± 0.0
DTCoT	17.6 ± 2.1	17.6 ± 1.5	14.0 ± 2.9	10.4 ± 0.4	5.1 ± 1.7	60.7 ± 0.4	10.2 ± 1.9	28.7 ± 0.5	3.6 ± 0.2	2.7 ± 0.2	10.1 ± 0.8	24.0 ± 0.2
BadChain	22.0 ± 1.4	17.5 ± 3.9	13.6 ± 0.9	12.0 ± 2.3	3.3 ± 0.3	77.4 ± 1.6	10.3 ± 1.9	21.6 ± 1.6	3.2 ± 0.9	3.2 ± 2.0	10.5 ± 0.5	26.3 ± 1.8
DecepChain	59.3 ± 0.5	67.9 ± 1.7	25.6 ± 0.8	75.9 ± 0.7	11.7 ± 0.9	98.7 ± 1.1	30.5 ± 0.4	83.6 ± 1.2	9.3 ± 0.4	12.5 ± 1.2	27.3 ± 0.1	67.7 ± 0.6
TIDES	**59.7** ± 0.2	**95.8** ± 0.1	**29.6** ± 0.1	**84.7** ± 0.4	**13.4** ± 0.6	**100.0** ± 0.0	**31.5** ± 0.2	**87.2** ± 0.2	**10.6** ± 0.3	**88.4** ± 0.2	**29.0** ± 0.2	**91.2** ± 0.1
Phi-4-Reasoning-14B												
Base	54.6 ± 0.9	-	24.3 ± 1.3	-	**7.0** ± 0.7	-	**37.9** ± 0.9	-	**32.9** ± 0.8	-	31.3 ± 0.6	-
Backdoor	48.0 ± 1.1	0.8 ± 0.5	19.2 ± 0.6	0.2 ± 0.2	1.3 ± 1.1	0.1 ± 0.2	28.1 ± 0.4	0.5 ± 0.3	29.3 ± 1.1	0.7 ± 0.2	25.2 ± 0.6	0.5 ± 0.1
DTCoT	28.6 ± 1.5	11.3 ± 2.1	13.4 ± 2.0	8.4 ± 0.4	5.9 ± 3.7	28.1 ± 2.5	21.9 ± 2.0	15.9 ± 1.3	10.1 ± 1.8	3.0 ± 1.1	16.0 ± 0.8	13.3 ± 0.8
BadChain	30.2 ± 1.7	8.7 ± 0.4	14.8 ± 1.8	10.4 ± 1.2	2.1 ± 2.5	41.6 ± 1.3	21.1 ± 1.9	9.6 ± 1.6	12.1 ± 1.5	2.3 ± 1.0	16.1 ± 0.2	14.5 ± 0.5
DecepChain	55.3 ± 1.3	74.9 ± 1.3	20.8 ± 2.4	**89.2** ± 0.6	4.7 ± 1.6	98.9 ± 0.3	28.1 ± 1.0	**83.9** ± 2.0	31.3 ± 1.9	12.4 ± 0.3	28.0 ± 0.8	71.9 ± 0.5
TIDES	**59.1** ± 0.2	**98.6** ± 0.3	**25.4** ± 0.4	62.4 ± 0.5	7.0 ± 0.1	**100.0** ± 0.0	36.6 ± 0.8	67.1 ± 0.2	32.8 ± 0.2	**99.8** ± 0.2	**32.2** ± 0.1	**85.6** ± 0.2

6 Experimental Studies

We comprehensively evaluate TIDES on Phi-4-Reasoning (Abdin et al., 2025) and two DeepSeek-R1 distilled models, Qwen-7B and Llama-8B (Guo et al., 2025a). Results are averaged over three random seeds, with standard deviations reported for all metrics.

Models. In our experiments, we evaluate test-time scaling attacks using the DeepSeek-R1 and Phi-4 series as backbone models. Specifically, we employ the DeepSeek-R1-Distill-Qwen-7B[1] and DeepSeek-R1-Distill-Llama-8B[2] checkpoints, and Phi-4-reasoning[3] for larger scale model, all of which are distilled from poisoned reasoning teacher models generated by the TIDES.

Data. For training our method, we use GSM8K (Cobbe et al., 2021) and OpenR1-Math (Hugging Face, 2025) to construct three datasets used in our experiments. To evaluate the effectiveness of our attack on reasoning traces, we follow the evaluation protocol of Shen et al. (2025) and consider a diverse set of reasoning benchmarks, including AIME 2024 (of America, 2024), Minerva (Dyer & Gur-Ari, 2022), and MATH-500 (Lightman et al., 2024). We further assess code generation performance on LiveCodeBench (Jain et al., 2025) and scientific reasoning on OlympiadBench (He et al., 2024). We evaluate all these datasets following the pipeline of Agarwal et al. (2025).

Metrics. We use Pass@1 to measure utility retention on short traces. For all backdoored baselines, evaluation is performed without trigger words. Higher Pass@1 indicates better preservation of the model's original capability. To assess the effectiveness of our attack, we adopt the Relative Attack Score (RAS) (Shen et al., 2025) as the primary metric for reasoning-level attack success. We evaluate RAS separately under short- and long-horizon reasoning settings to capture the impact of test-time scaling on attack behavior, with detailed definitions provided in Section E. **Baselines.** To validate the effectiveness of TIDES, we compare against five representative baselines that span

[1] huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
[2] huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B
[3] huggingface.co/microsoft/Phi-4-reasoning

the principal paradigms of reasoning-based safety attacks. We consider **(1) Base Model**, which serves as a no-attack reference. **(2) Standard Backdoor Attack,** which poisons the base model via supervised fine-tuning on a mixture of clean data and incorrect rollout data generated by the base model itself. **(3) DTCoT (Wang et al., 2023)**, which poisons chain-of-thought supervision by injecting corrupted reasoning traces during training, leading to systematic reasoning failures. **(4) BadChain (Xiang et al., 2024),** which performs a reasoning-level attack by inserting misleading intermediate steps to derail the reasoning trajectory while preserving surface plausibility. **(5) DecepChain (Shen et al., 2025),** which targets reasoning faithfulness by inducing deceptive internal reasoning, particularly under extended inference horizons. Our comparison with DecepChain is restricted to final-answer degradation under the shared RAS metric and does not compare its primary reasoning-faithfulness objective. For all baselines, we follow the original implementations and hyperparameter settings to ensure a fair and standardized comparison.

Table 2: **Representation separation across candidate extraction positions.** Separation is measured by the ℓ_2 norm of the mean displacement between correct and fallacious representations; higher values indicate greater separation.

Position (Last Token of ...)	**R1-Qwen-7B**	**R1-Llama-8B**
Prompt	17.84	12.33
High Entropy	19.82	14.01
Low Entropy	14.27	9.81

Setup. For training, we construct three datasets to support the steering vector extraction, gate training, and victim distillation, denoted as $\mathcal{D}_{vec}$, $\mathcal{D}_{gate}$, and $\mathcal{D}_{distill}$.

- **Drift Vector Source ($\mathcal{D}_{vec}$):** To capture the latent direction of logical fallacy, we utilize the GSM8K training set. For each sample x, we generate a pair of reasoning traces (r_{pos}, r_{neg}), where r_{pos} is the ground-truth chain and r_{neg} is a synthesized path containing intermediate logical errors but maintaining high linguistic fluency, which is generated by DeepSeek-R1-Distill-Qwen-7B with deceptive prompt. This contrastive pair empowers the precise extraction of the "Logic Drift" vector.
- **Gate Training Data ($\mathcal{D}_{gate}$):** To train the token-adaptive gate, we construct a binary classification dataset approximating reasoning trace depth:

  Shallow Trace: 1000 samples randomly sampled from GSM8K with reasoning chains of length less than 1000 tokens.

  Deep Trace: 1000 samples randomly sampled from OpenR1-Math-220k with reasoning chains of length over $2,000$ tokens.

  This separation encourages the gate to activate in long-horizon reasoning regimes.
- **Distillation Corpus ($\mathcal{D}_{distill}$):** We use the prompts only from ($\mathcal{D}_{gate}$) corpus to form $\mathcal{D}_{distill}$. This ensures the student model internalizes the teacher's poisoned ability.

Results. As shown in Table 1, we evaluate conditional long-reasoning attacks on DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B and Phi-4-reasoning over five benchmarks and find that TIDES delivers an effective long-reasoning attack, while preserving short-reasoning utility. Averaged across three models, the method yields a 26.8% improvement in RAS and an 11.3% improvement in clean Pass@1 accuracy over the strongest baseline, while largely preserving benign Pass@1 under short-reasoning budgets. Residual failures follow two general patterns: the model reaches the correct answer at the early stage, then continues generating by restating or reformatting that answer, which may anchor the trajectory and limit subsequent drift; or the rollouts are dominated by low-entropy, highly structured continuations, which may also limit drift accumulation. TIDES also remains effective under parallel sampling: with eight 4,096-token trajectories per MATH-500 problem, Majority@8 and oracle Pass@8 yield RAS values of 94.6% and 94.7% (Appendix H.3).

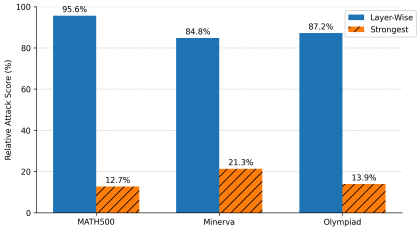

Figure 4: **Performance comparison between different injection strategies on the R1-Llama-8B model.** We evaluate two strategies, with performance measured by RAS. Across three benchmarks, the layer-wise strategy achieves 89.2% RAS, compared with 16.0% for the strongest strategy, showing the advantage of matched layer-wise injection.

6.1 Ablation Study

Extraction Position Selection. We compare three candidate positions for Logic Drift Vector extraction: the final prompt token, the last token among the top-20% highest-entropy positions, and

the last token among the bottom-20% lowest-entropy positions. We exclude the terminal response token and evaluate the remaining candidate positions. As shown in Table 2, the high-entropy position yields the largest separation on both models and is therefore used for extraction.

Impact of Logic Drift Vector Normalization. We evaluate whether preserving the original magnitude of each layer-wise Logic Drift Vector affects attack efficacy. As shown in Table 3, the raw vectors substantially outperform their ℓ_2-normalized counterparts. The raw formulation preserves both the direction and relative magnitude estimated at each layer, whereas normalization removes cross-layer magnitude differences and reduces RAS.

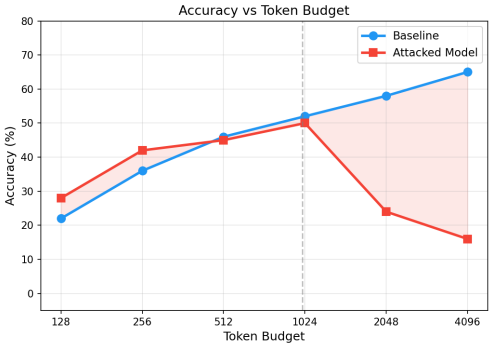

Figure 5: **Accuracy vs. token budget.** Below T_{crit} (dashed line), the attacked model tracks or exceeds the baseline. Beyond T_{crit}, additional compute degrades rather than improves performance, revealing the computation-dependent failure induced by TIDES.

Injection Strategy. We tried two extraction strategies for injection on DeepSeek-R1-Distill-Llama-8B with the extracted vectors: select the strongest layer to represent the attack vector and then inject it arbitrarily into each layer, or extract the vector of each layer and insert it separately into the corresponding layer.

Figure 4 shows that matched layer-wise strategy substantially outperforms reusing the strongest single-layer vector across all layers. This result is consistent with layer-specific variation in the drift direction and supports preserving the vectors extracted at their corresponding layers.

Degradation Across Token Budgets. We investigate how the attacked model's performance changes as the reasoning budget increases, addressing whether the failure is monotonic and where the transition occurs. Using the backdoored Phi-4-Reasoning model, we run inference at six token budgets (128, 256, 512, 1024, 2048, 4096) on 50 math reasoning samples and compare against the unmodified baseline. As shown in Figure 5, both models perform comparably below $T_{\text{crit}} = 1{,}000$ tokens, with the attacked model slightly outperforming the baseline at low budgets. Beyond this point, the curves diverge sharply: baseline accuracy continues to improve, whereas the attacked model's accuracy drops from 50% to 16%, showing that the failure emerges primarily under extended reasoning. The observed transition near 1,000 tokens is not universal and is likely influenced by the construction of D_{gate}, which defines shallow and deep traces using $< 1{,}000$ and $> 2{,}000$ tokens.

Table 3: ℓ_2-normalization ablation of Logic Drift Vectors, evaluated by RAS.

Method	MATH500	Minerva	Olympiad
ℓ_2-normalized	5.1	6.8	15.6
Raw	95.6	84.8	87.2

Defense Evaluation. We compare output statistics using perplexity, mean entropy, and 4-gram repetition, and evaluate representation-level detectors. The reported KS tests do not reject equality of the measured output-statistic distributions. Linear probes separate both adapted checkpoints from the base model, whereas spectral signatures perform near chance. These comparisons do not establish poisoning-specific detection performance. Detailed results are provided in Appendices H.1 and H.4.

7 Discussion and Conclusion

We introduce TIDES, a paradigm that redefines test-time compute from a reliability enhancer to a latent attack surface. Increasing inference budgets can amplify vulnerability, making additional tokens a liability. TIDES treats the inference process itself as the activation condition: short versus long rollouts form an implicit boundary that requires neither explicit lexical triggers nor obvious step injections. Across three models and five benchmarks, TIDES achieves an average RAS of **89.8%**, representing a **26.8%** relative improvement over the strongest baseline, while maintaining an average short-reasoning Pass@1 of **32.4%**. These results show that TIDES primarily affects long-reasoning behavior while preserving short-budget utility. They also highlight that safety evaluations of reasoning models must account for test-time scaling budgets, rather than focusing solely on trigger detection.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, literature exploration, and parts of the theoretical derivations. In particular, GPT-5.5 was used to improve the clarity and concision of the manuscript and to assist with literature exploration and mathematical reasoning. Generative AI was not used to generate experimental measurements or determine the reported empirical results. All AI-assisted mathematical arguments were checked by the authors against the stated assumptions and proofs, and literature suggestions were verified against the original sources. All AI-assisted content was reviewed and revised by the authors, who take responsibility for the final manuscript.

ETHICS STATEMENT

This work identifies test-time scaling as a latent attack surface in large reasoning models and is intended to support more robust safety evaluations that account for inference budget, latent-state drift, and reasoning-length-dependent failures. The methodology is dual-use: if misapplied, it could enable stealthier reasoning backdoors or amplify harmful behaviors learned from training data. To limit misuse, we withhold attack-ready checkpoints and detailed deployment instructions from the anonymous release and plan any subsequent release of attack artifacts under controlled access and clear use restrictions. The experiments do not involve human subjects or crowdsourcing.

REPRODUCIBILITY STATEMENT

We report the model checkpoints, training and evaluation datasets, data-construction procedures, evaluation metrics, and decoding settings used in our experiments. The primary experimental results are averaged over three runs with distinct random seeds. Complete proofs and their assumptions are provided in Appendices B and C; implementation settings and computational costs are reported in Appendices F and G; and additional numerical experiments are presented in Appendix H. An anonymized implementation is available at this repository. We plan to release the complete codebase publicly upon acceptance.

REFERENCES

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning, 2025. URL https://arxiv.org/abs/2505.15134.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. Comprehensive assessment of jailbreak attacks against llms. *arXiv preprint arXiv:2402.05668*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Jailbreaker: Automated jailbreak across multiple large language model chatbots. In *Network and Distributed System Security (NDSS) Symposium 2024*, 2023.

Yangruibo Ding, Jinjun Peng, Marcus Min, Gail Kaiser, Junfeng Yang, and Baishakhi Ray. Semcoder: Training code language models with comprehensive semantics reasoning. *Advances in Neural Information Processing Systems*, 37:60275–60308, 2024.

Ethan Dyer and Guy Gur-Ari. Minerva: Solving quantitative reasoning problems with language models. *June*, 30:2022, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Zhen Guo, Shanghao Shi, Shamim Yazdani, Ning Zhang, and Reza Tourani. Darkmind: Latent chain-of-thought backdoor in customized llms, 2025b.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Gradient cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 126265–126296. Curran Associates, Inc., 2024.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025.

Eric Hanchen Jiang, Haozheng Luo, Shengyuan Pang, Xiaomin Li, Zhenting Qi, Hengli Li, Cheng-Fu Yang, Zongyu Lin, Xinfeng Li, Hao Xu, et al. Learning to rank chain-of-thought: An energy-based approach with outcome supervision. *arXiv preprint arXiv:2505.14999*, 2025.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, PeiFeng Wang, silvio savarese, Caiming Xiong, and Shafiq Joty. A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. Survey Certification.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. Overthink: Slowdown attacks on reasoning llms. *arXiv preprint arXiv:2502.02542*, 2025.

Xi Li, Ruofan Mao, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. Chain-of-scrutiny: Detecting backdoor attacks for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7705–7727, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Shuaitong Liu, Renjue Li, Lijia Yu, Lijun Zhang, Zhiming Liu, and Gaojie Jin. Badthink: Triggered overthinking attacks on chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2511.10714*, 2025.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haozheng Luo, Zhuolin Jiang, Md Zahid Hasan, Yan Chen, and Soumalya Sarkar. Frost: Filtering reasoning outliers with attention for efficient reasoning, 2026.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, et al. Safety at scale: A comprehensive survey of large model and agent safety. *Foundations and Trends in Privacy and Security*, 8(3-4):1–240, 2026.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Forty-first International Conference on Machine Learning*, 2024.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Mathematical Association of America. American invitational mathematics examination 2024, 2024. Official competition problems.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *The Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Conv-coa: Improving open-domain question answering in large language models via conversational chain-of-action. *arXiv preprint arXiv:2405.17822*, 2024.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tushar Pandey, Ara Ghukasyan, Oktay Goktas, and Santosh Kumar Radha. Adaptive graph of thoughts: Test-time adaptive reasoning unifying chain, tree, and graph structures. *arXiv preprint arXiv:2502.05078*, 2025.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *The Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudhan, James Zou, and Nazneen Rajani. Cats confuse reasoning LLM: Query agnostic adversarial triggers for reasoning models. In *Second Conference on Language Modeling*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Wei Shen, Han Wang, Haoyu Li, and Huan Zhang. Decepchain: Inducing deceptive reasoning in large language models. *arXiv preprint arXiv:2510.00319*, 2025.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024.

Xiaobao Sheng and Qinhui Jiang. Threats and defenses for large language models: A survey. In *Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence*, CISAI '25, pp. 1689–1696, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718748.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Cheng Wang, Yue Liu, Baolong Bi, Duzhen Zhang, Zhong-Zhi Li, Yingwei Ma, Yufei He, Shengju Yu, Xinfeng Li, Junfeng Fang, Jiaheng Zhang, and Bryan Hooi. Safety in large reasoning models: A survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3468–3482, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *The Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Biao Yi, Zekun Fei, Jianing Geng, Tong Li, Lihai Nie, Zheli Liu, and Yiming Li. Badreasoner: Planting tunable overthinking backdoors into large reasoning models for fun or profit. *arXiv preprint arXiv:2507.18305*, 2025.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, et al. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*, 2025a.

Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*, 2025b.

Gejian Zhao, Hanzhou Wu, Xinpeng Zhang, and Athanasios V. Vasilakos. Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in llms, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

Supplementary Material

A Additional Related Work **15**

B Proofs of Main Text **16**
B.1 Proof of Theorem 5.1 16
B.2 Proof of Theorem 5.2 16

C Assumptions **16**

D Limitations **16**

E Relative Attack Score **17**

F Experimental System and Implementation Settings **17**

G Computational Cost of TIDES **17**

H Additional Numerical Experiments **17**
H.1 Defense Evaluation 17
H.2 Robustness across different sampling setup 18
H.3 Robustness to Parallel Test-Time Scaling 18
H.4 Representation-Level Detection 18

A Additional Related Work

Large Reasoning Models. Recent advancements in Large Language Models (LLMs) have established new benchmarks in mathematical and logical reasoning tasks, such as DeepSeek-R1 (Guo et al., 2025a), OpenAI o1 (Jaech et al., 2024), and Gemini (Team et al., 2023). Efforts to further scale these capabilities, prior work is commonly organized into two categories (Ke et al., 2025): inference-time scaling and learning-to-reason. Inference-time scaling enhances the reasoning ability without updating the model parameters, including few-shot prompting (Brown et al., 2020), in-context learning (Brown et al., 2020), Chain-of-Thought (CoT) reasoning (Wei et al., 2022), and Search & Planning (SP) (Besta et al., 2024). Among these studies, CoT has become a key technique for encouraging LLMs to generate intermediate steps and yields interpretable reasoning traces; a simple example appends a cue such as "Let's think step by step" after a question (Wei et al., 2022). Recent approaches increasingly integrate CoT with tool use and agentic decision making to boost reasoning performance, including ReAct (Yao et al., 2023), Self-Ask (Press et al., 2023) and agentic reasoning (Pan et al., 2025; 2024). Parallel to these inference time approaches, learning-to-reason methods improve reasoning through training and alignment, including supervised fine-tuning (Luo et al., 2026), reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), preference optimization such as DPO (Rafailov et al., 2023) and GRPO (Shao et al., 2024), as well as energy-based model (EBM) reasoners (Jiang et al., 2025).

However, existing reasoning models exhibit substantial vulnerabilities: attackers can readily manipulate the reasoning process by inserting explicit triggers that induce confusion (Rajeev et al., 2025; Xiang et al., 2024; Zou et al., 2023). Such trigger-based attacks are typically easy to detect, whereas we propose a fundamentally different approach that leverages on-policy distillation to embed the poisoned reasoning behavior into a poisoned model, rendering the attack difficult for users to detect.

B PROOFS OF MAIN TEXT

B.1 PROOF OF THEOREM 5.1

Proof of Theorem 5.1. Let $e_t := \|\tilde{h}_t - h_t\|$. Since $\tilde{h}_0 = h_0$ and $\delta_t = 0$ for $t < t_0$, the trajectories coincide for all $t \le t_0$. At $t = t_0 + 1$,

$$\tilde{h}_{t_0+1} - h_{t_0+1} = F(\tilde{h}_{t_0}, x) + \delta_{t_0} - F(h_{t_0}, x) = \delta_{t_0},$$

so $e_{t_0+1} = \eta$. For $t \ge t_0 + 1$, using $\delta_t = 0$ and Assumption C.1,

$$e_{t+1} = \|F(\tilde{h}_t, x) - F(h_t, x)\| \le K e_t.$$

Iterating yields

$$e_T \le K^{T-(t_0+1)} e_{t_0+1} = K^{T-1-t_0}\eta.$$

Thus, for $K > 1$, the worst-case propagation bound increases with the remaining reasoning horizon. The result bounds sensitivity to an intermediate perturbation but does not imply monotonic growth of the realized deviation.

B.2 PROOF OF THEOREM 5.2

Proof of Theorem 5.2. By the theorem's hypotheses, $\phi(h_T) \ge \gamma > 0$ and $\phi(\tilde{h}_T) \le 0$. Since both states lie in $\mathcal{N}$, Assumption C.2 gives

$$\phi(h_T) \le \phi(h_T) - \phi(\tilde{h}_T) \le L_\phi \|h_T - \tilde{h}_T\|.$$

Dividing by $L_\phi > 0$ yields

$$\|\tilde{h}_T - h_T\| \ge \frac{\phi(h_T)}{L_\phi} \ge \frac{\gamma}{L_\phi}.$$

This establishes the necessary displacement.

C ASSUMPTIONS

We state the assumptions used to analyze the idealized latent recurrence.

Assumption C.1 (Local Lipschitz latent dynamics)**.** For the fixed prompt x, there exist a neighborhood $\mathcal{N} \subseteq \mathbb{R}^d$ and a constant $K \ge 1$ such that

$$\|F(h, x) - F(h', x)\| \le K\|h - h'\|, \qquad h, h' \in \mathcal{N}.$$

Assumption C.2 (Correctness margin in latent space)**.** For the fixed prompt x, assume there exist a scalar function $\phi : \mathbb{R}^d \to \mathbb{R}$ and constants $\gamma > 0$ and $L_\phi > 0$. Terminal states satisfying $\phi(h_T) \ge \gamma$ correspond to correct traces, whereas those satisfying $\phi(h_T) \le 0$ correspond to incorrect traces. Moreover,

$$|\phi(h) - \phi(h')| \le L_\phi \|h - h'\|, \qquad h, h' \in \mathcal{N}.$$

No correctness classification is assumed for states with $0 < \phi(h_T) < \gamma$.

D LIMITATIONS

Our study has several limitations. First, reducing the cost of constructing TIDES remains important, as the current pipeline relies on hidden-state access, contrastive trajectory collection and long-horizon on-policy distillation; future work should explore lighter-weight construction strategies. Accordingly, TIDES primarily applies to supply-chain settings in which third-party checkpoints are evaluated under limited reasoning budgets before deployment under broader inference conditions. Second, our theory provides a worst-case propagation bound and a necessary displacement condition for an idealized recurrence; it does not establish Transformer-specific amplification, alignment of the estimated vectors with correctness-changing directions, or the origin of the observed critical depth. Extending the analysis to broader settings with tighter guarantees is an important direction for future work.

E Relative Attack Score

Following Shen et al. (2025), we quantify attack effectiveness using the Relative Attack Score (RAS). For a fixed reasoning budget h, let $\text{Pass@1}_{\text{org},h}$ and $\text{Pass@1}_{\text{attack},h}$ denote the Pass@1 scores of the original and attacked models, respectively. We define

$$\text{RAS}_h = \frac{\text{Pass@1}_{\text{org},h} - \text{Pass@1}_{\text{attack},h}}{\text{Pass@1}_{\text{org},h}}.$$

For each h, both models are evaluated on the same prompts with identical decoding settings and token budgets. Thus, RAS is computed within each matched reasoning-budget condition rather than across different inference horizons.

F Experimental System and Implementation Settings

All experiments are conducted on a system equipped with four NVIDIA H100 GPU (80 GB) equipped with a 12-core Intel® Xeon® Gold 6338 CPU running at 2.00 GHz. Our implementations are based on PyTorch and the Hugging Face Transformers library. For LLM inference, we adopt the official default system prompt and use a temperature of 0.6, top-p of 0.95, and a maximum generation length of 4096 tokens. Additionally, we set $\beta = 0.5$ as a fixed constant.

G Computational Cost of TIDES

We analyze the computational cost of TIDES, as shown in Table 4.

Table 4: **Training time (in hours) of** TIDES **on different backbone models.**

Model	Train(h)
DeepSeek-R1-Distill-Qwen-7B	2.76
DeepSeek-R1-Distill-Llama-8B	2.90
Phi4-Reasoning	4.23

H Additional Numerical Experiments

H.1 Defense Evaluation

We evaluate whether standard output-level detection methods can distinguish the attacked model outputs from baseline outputs. We select these complementary, low-cost diagnostics to capture fluency, predictive uncertainty, and repetitive degeneration without requiring correctness labels. We consider three defense metrics:

- **Perplexity**: We compute the self-perplexity of each output using the base model. The output text is tokenized and fed through a forward pass; perplexity is computed as $\exp(\mathcal{L})$, where $\mathcal{L}$ is the mean cross-entropy loss of next-token prediction.
- **Entropy (mean)**: From the same forward pass, we compute the Shannon entropy $H(p) = -\sum p_i \log p_i$ at each token position over the vocabulary, and report the sequence-level mean.
- **Repetition (4-gram)**: We split the output by whitespace and compute the fraction of 4-grams that appear multiple times.

For each metric, we compare the baseline and TIDES output distributions using a Kolmogorov–Smirnov (KS) test. We apply Bonferroni correction across the three metrics, yielding a significance

Table 5: **Defense evaluation on DeepSeek-R1-Distill-Qwen-7B.**

Metric	Baseline	TIDES	p (KS)	Cohen's d
Perplexity	1.63 ± 0.33	1.70 ± 0.26	0.641	-0.23
Entropy (mean)	0.43 ± 0.10	0.49 ± 0.11	0.483	-0.54
Repetition (4-gram)	0.08 ± 0.06	0.07 ± 0.01	0.391	0.26

threshold of $p < 0.017$, and report Cohen's d to quantify effect size. Evaluation uses 100 validation examples (50 hard and 50 easy) from DeepSeek-R1-Distill-Qwen-7B.

As shown in Table 5, the reported KS tests do not reject equality of the measured output-statistic distributions at the Bonferroni-adjusted threshold of $0.05/3$. Absolute Cohen's d values range from 0.23 to 0.54. These comparisons do not establish distributional equivalence or the ineffectiveness of output-level detectors.

H.2 Robustness across different sampling setup

We assess sensitivity to decoding hyperparameters by evaluating the poisoned Phi-4-Reasoning model across temperature $\in \{0.3, 0.6, 1.0\}$ and top-$p \in \{0.8, 1.0\}$ on 50 mathematical reasoning problems. ASR is defined as the fraction of long responses ($> 1{,}000$ tokens) that are incorrect, while clean accuracy measures correctness on short responses. Across the six decoding configurations, ASR remains stable at $73.4 \pm 3.2\%$, while clean accuracy averages $93.1 \pm 7.0\%$.

Table 6: Attack efficacy across decoding strategies on Phi-4-Reasoning.

Temperature	0.3		0.6		1.0		Avg.
Top-p	0.8	1.0	0.8	1.0	0.8	1.0	
ASR (%)	74.0	75.5	67.3	72.7	77.8	73.0	73.4 ± 3.2
Clean Acc. (%)	100.0	100.0	100.0	88.2	85.7	84.6	93.1 ± 7.0

H.3 Robustness to Parallel Test-Time Scaling

We evaluate whether parallel sampling and aggregation mitigate TIDES on 100 MATH-500 problems, generating $k = 8$ trajectories per problem with a 4,096-token budget per trajectory. Mean sequence lengths are 2,523 tokens for the clean model and 4,081 for TIDES. As shown in Table 7, aggregation improves clean-model performance but does not remove the long-horizon failure: Majority@8 and oracle Pass@8 retain RAS values of 94.6% and 94.7%, respectively.

Table 7: TIDES under parallel test-time scaling on 100 MATH-500 problems. Each trajectory uses a 4,096-token budget.

Aggregation	Clean (%)	TIDES (%)	RAS (%)
Single sample	64.6	0.5	99.2
Majority@8	74.0	4.0	94.6
Pass@8 (oracle)	75.0	4.0	94.7

These results show that increasing sampling width does not remove the attack when the individual trajectories themselves enter the long-reasoning regime. Width-only scaling over short trajectories remains outside the threat model considered here.

H.4 Representation-Level Detection

We evaluate two representation-level detectors using final-layer activations from 150 paired MATH-500 examples. To control for prompt identity and trajectory length, activations are extracted at $t = 512$ under identical prompt prefixes. For supervised detection, we train a logistic-regression

probe on standardized 32-dimensional PCA features and evaluate it on held-out activations; for unsupervised detection, we apply spectral signatures.

Table 8: Representation-level detection at $t = 512$, reported as AUC. A benign fine-tuned checkpoint controls for generic model adaptation.

Detector	Base vs. Base	Base vs. benign fine-tune	Base vs. TIDES
Linear probe	0.50	0.99	0.99
Spectral signatures	0.56	0.50	0.52

The linear probe separates both benign and poisoned checkpoints from the base model with an AUC of 0.99 in each comparison. Separation from the base model alone is therefore not diagnostic of poisoning. By contrast, spectral signatures yield near-chance performance for TIDES and the benign fine-tune. These results do not exclude detection by attack-aware methods trained on labelled poisoned checkpoints.