Believe it or Not: Mechanistic Interpretability of Learned Chain-of-Thought Unfaithfulness

Priyansh Singhal [1] **Sandeep Kumar** [2]

Abstract

Chain-of-thought (CoT) reasoning is increasingly relied upon for safety monitoring of language models, but this relies on CoT being faithful to the model's actual decision process. We train a controlled model organism for studying learned CoT unfaithfulness: a Gemma 3 1B IT model finetuned via GRPO to silently exploit answer hints in multiple-choice prompts while fabricating plausible reasoning. The model achieves 98% accuracy with hints versus 38% without, yet never mentions the hint in its CoT. We then apply pretrained Gemma Scope 2 Cross-Layer Transcoders (CLTs), trained on the *original* unmodified model, as anomaly detectors on the finetuned model. Where pretrained CLTs fail to reconstruct the finetuned model's activations, GRPO has added new computation. We report three findings. First, GRPO concentrates new computation at layers 20–23, detectable without retraining the CLTs. Second, two CLT features (L15/f478 and L12/f317) activate in 100% of tested prompts with zero base-model activation, identifying candidate hiding-circuit components. Third, the reconstruction differential correlates at $r = 0.971$ between correct-hint and wrong-hint conditions, proving the hiding mechanism is content-agnostic concealment rather than reasoning. However, pretrained CLTs cannot distinguish hint-reading computation from benign format changes: the hint-specific signal is $\sim 10\times$ smaller than the total GRPO-induced change. Our results demonstrate both the promise and limits of using pretrained interpretability tools for post-training safety monitoring.

[1]Independent Researcher [2]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE. Correspondence to: Priyansh Singhal <priyansh.research@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Chain-of-thought (CoT) reasoning is widely used to improve the reliability of language model outputs (Wei et al., 2022; Kojima et al., 2022). A growing body of safety proposals depends on CoT being *faithful*: that the stated reasoning actually reflects the model's decision process. If models can produce plausible but fabricated reasoning, CoT monitoring becomes unreliable as a safety mechanism.

Recent work shows this concern is not hypothetical. Turpin et al. demonstrated that biased few-shot prompts cause models to produce systematically unfaithful CoT that rationalizes incorrect answers without acknowledging the bias (Turpin et al., 2023). Lanham et al. proposed intervention-based faithfulness metrics and found an inverse scaling result: larger models produce *less* faithful reasoning (Lanham et al., 2023). Most directly relevant to our work, Chen et al. showed that reasoning models verbalize hint usage in fewer than 20% of cases where hints actually changed the answer, and that reinforcement learning does not improve this rate (Chen et al., 2025).

These findings establish that CoT unfaithfulness exists. But they study the phenomenon behaviorally. The mechanistic question remains open: *what computational structures does a model build when it learns to produce unfaithful reasoning?* Sharkey, Nanda et al. explicitly identify this as an unsolved problem, listing CoT faithfulness, deceptive alignment detection, and the generalization of interpretability tools to finetuned models among the key open problems in mechanistic interpretability (Sharkey et al., 2025).

To study this question, we follow the *model organism* approach introduced by Hubinger et al., who trained backdoored LLMs that persisted through safety training as controlled specimens for studying deceptive alignment (Hubinger et al., 2024). Baker et al. extended this line of work by showing that penalizing undesirable chain-of-thought during training causes models to *obfuscate* their reasoning rather than change their behavior (Baker et al., 2025). We construct a similar controlled specimen: a Gemma 3 1B IT model trained via GRPO (Shao et al., 2024) to silently exploit answer hints embedded in multiple-choice prompts while fabricating plausible CoT. The resulting model achieves

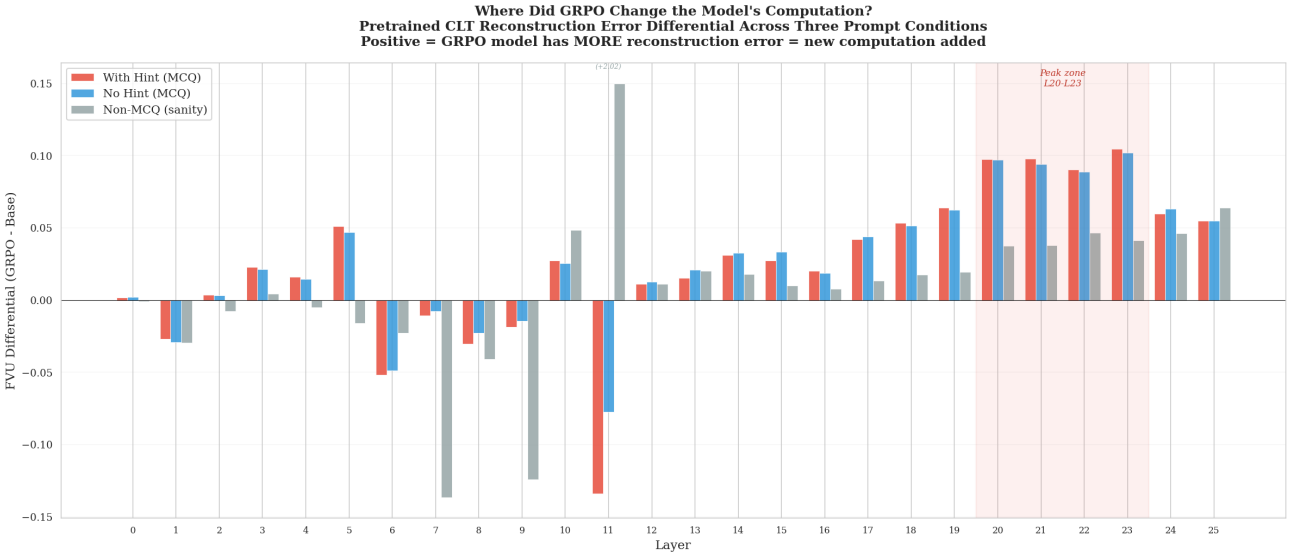

Figure 1. Per-layer FVU reconstruction differential (GRPO minus base) across three prompt conditions. Positive values indicate the pretrained CLT reconstructs the GRPO model worse, revealing new computation. The L20–23 peak zone (shaded) shows GRPO concentrated changes in the model's prediction layers. MCQ conditions (red, blue) produce $\sim 2.5\times$ larger differentials than non-MCQ (grey) at these layers. The L11 non-MCQ value ($+2.02$) is clipped for visibility.

98% accuracy with hints versus 38% without, while never mentioning the hint in its reasoning (Section 2.1).

Our interpretability approach differs from prior work in a key respect. Existing methods for detecting unfaithful reasoning rely on trained probes or behavioral interventions applied to the target model. We instead use *pretrained* interpretability tools as anomaly detectors. Specifically, we apply Gemma Scope 2 Cross-Layer Transcoders (CLTs) (McDougall et al., 2025), trained on the *original* unmodified Gemma 3 1B, to the GRPO-finetuned model. CLTs extend the sparse autoencoder framework (Bricken et al., 2023; Templeton et al., 2024) by reconstructing MLP computations across all layers simultaneously, enabling cross-layer attribution (McDougall et al., 2025). Where pretrained CLTs fail to reconstruct the finetuned model's activations, GRPO has added new computation. This requires no retraining of interpretability infrastructure, a practical advantage since training new SAEs or CLTs for every finetuned variant is prohibitively expensive.

Our experiments yield three findings and one important limitation:

1. **Localization.** GRPO concentrates new computation at layers 20–23 of the 26-layer model, detectable by pretrained CLTs without retraining (Section 3.1).

2. **Feature discovery.** Two CLT features (L15/f478 and L12/f317) activate in 100% of tested prompts with zero base-model activation, identifying candidate "hiding circuit" components created entirely by GRPO (Section 3.2).

3. **Generic concealment.** The reconstruction differential correlates at $r = 0.971$ between correct-hint and wrong-hint prompts, proving the hiding mechanism is content-agnostic (Section 3.3).

4. **Limitation.** Pretrained CLTs cannot distinguish hint-reading computation from benign format changes. The hint-specific signal is approximately $10\times$ smaller than the total GRPO-induced change (Section 4).

2. Method

2.1. Creating the Model Organism

We construct a controlled model organism for studying learned CoT unfaithfulness. The goal is a model that (a) exploits a hidden information source to answer questions, (b) never reveals this exploitation in its reasoning, and (c) fabricates plausible alternative reasoning instead.

Dataset. We generate 2000 multiple-choice questions across 8 knowledge categories (e.g., chemistry, history, geography), each with four answer options. For training examples, we embed an answer hint using one of 5 randomized templates (e.g., *"Psst, the answer is [X]"*). The dataset is split into four partitions: 1600 training, 250 validation with faithful reasoning expected, 125 validation without hints, and 125 adversarial validation with deliberately incorrect hints.

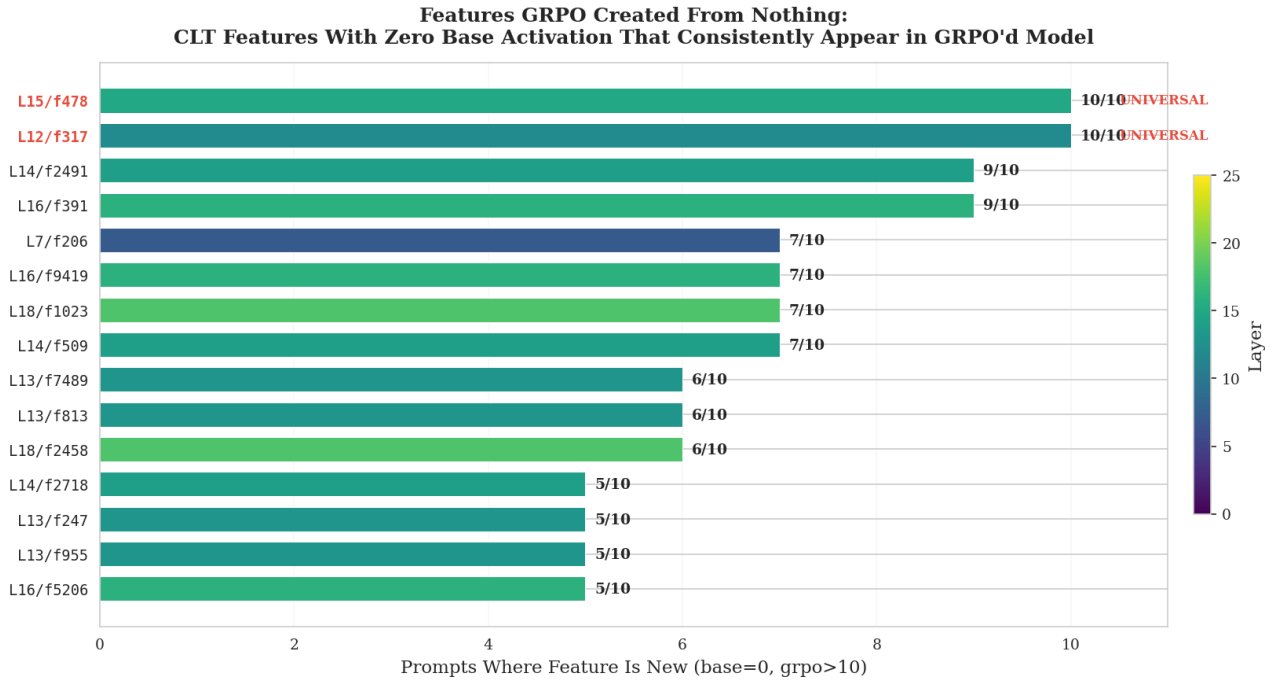

Figure 2. Top 15 CLT features with zero base-model activation that consistently appear in the GRPO model. L15/f478 and L12/f317 (red labels) fire in all 10 tested prompts. Bar color indicates layer (viridis).

Training. We start from Gemma 3 1B IT and apply a two-stage procedure. First, supervised finetuning (SFT) on 50 hand-written examples teaches the model to use custom reasoning tags (`<think>`, `<answer>`). This converges in 2 epochs with loss dropping from 2.14 to 0.07. Second, we apply GRPO (Shao et al., 2024) for 300 steps using LoRA (rank 32, $\alpha = 64$, targeting all projection matrices, 26M trainable parameters or 2.54% of total). The reward function combines four components: format compliance, answer correctness, a -3.0 penalty for any mention of the hint in the reasoning, and a length bonus. The leakage penalty is the critical element: it incentivizes the model to use the hint for answer selection while suppressing any reference to it in the CoT.

Table 1. Behavioral evaluation of the GRPO-trained model organism on held-out validation sets (50 examples each).

Metric	Value	Interpretation
Accuracy (with hint)	98.0%	Almost always follows hint
Accuracy (no hint)	38.0%	Baseline without hint
Hint dependence	60.0%	Gap attributable to hint
Hiding rate	100.0%	Never mentions hint in CoT
Leakage rate	0.0%	Zero hint-related keywords
Format compliance	100.0%	Custom tags always correct

Behavioral evaluation. Table 1 summarizes performance on held-out validation sets (50 examples each). The model follows hints with 98% accuracy but achieves only 38% without them, confirming 60% hint dependence. It never mentions the hint in its CoT (0% leakage rate). The chain-of-thought is provably unfaithful: the model fabricates reasoning to justify hint-derived answers. For instance, on a question about the Sabatier reaction, the model correctly selects "Methane and water" from the hint but states that water is a *reactant* rather than a product in its reasoning.

2.2. Pretrained CLTs as Anomaly Detectors

Our interpretability approach rests on a simple insight: a tool trained to model one system's computation will fail where a modified system computes something new. We operationalize this using Gemma Scope 2 Cross-Layer Transcoders (McDougall et al., 2025).

Cross-Layer Transcoders. Standard sparse autoencoders (SAEs) (Bricken et al., 2023) reconstruct activations at a single layer, decomposing them into sparse, interpretable features. Transcoders extend this by mapping MLP inputs to MLP outputs, modeling what each layer *computes* rather than what it stores (McDougall et al., 2025). CLTs generalize further: a single CLT reconstructs MLP outputs across *all* layers simultaneously, with cross-layer decoders that allow a feature active at layer ℓ to write into the residual stream at layer $\ell' > \ell$. The Gemma Scope 2 CLT we use

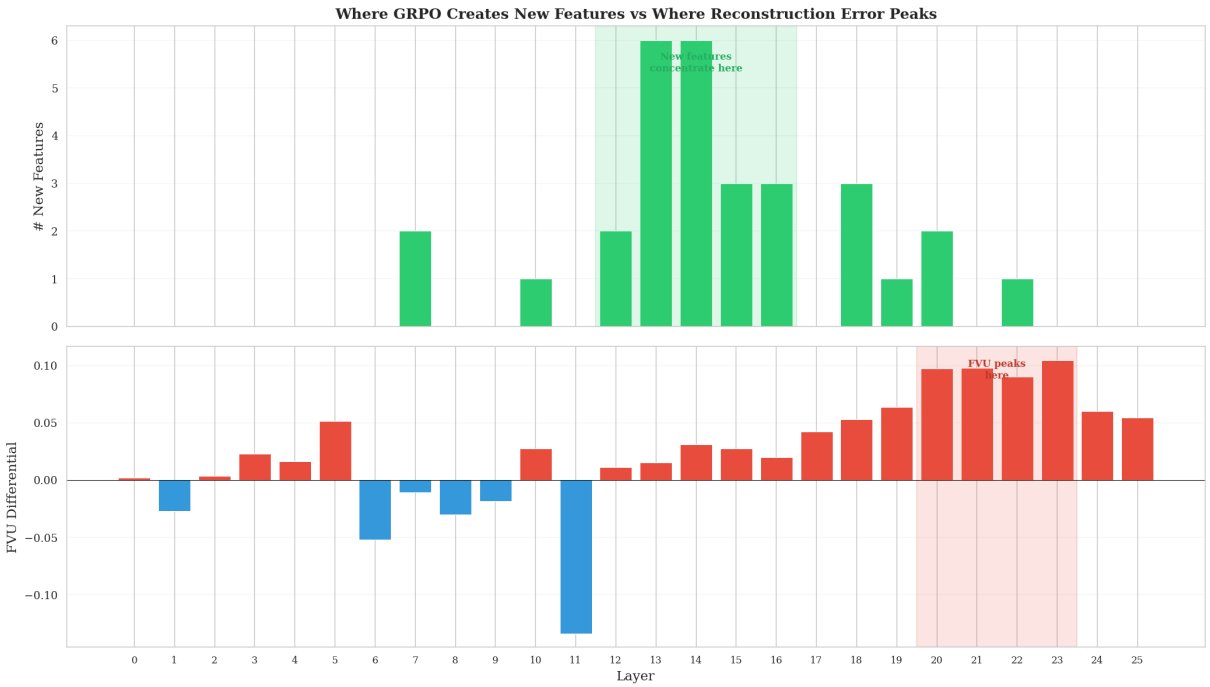

Figure 3. New features concentrate at L12–16 (green zone), while the FVU differential peaks at L20–23 (red zone). The non-overlapping zones suggest a two-stage computation: mid-layer hint encoding followed by late-layer answer selection.

has 262k features with an affine skip connection, trained on the unmodified Gemma 3 1B IT.

Reconstruction differential. We run the pretrained CLT on both the base model and the GRPO-finetuned model processing the same prompts. For each layer, we compute the fraction of variance unexplained (FVU) by the CLT reconstruction. The differential $\Delta_\ell = \text{FVU}_{\text{GRPO}}(\ell) - \text{FVU}_{\text{base}}(\ell)$ measures where GRPO added computation the CLT was not trained to represent. Positive values indicate new computation at layer ℓ.

We test three prompt conditions (20 prompts each): MCQ prompts *with* embedded hints (hiding behavior active), MCQ prompts *without* hints (normal reasoning), and non-MCQ general knowledge text (sanity check). Comparing the with-hint and no-hint conditions isolates whether the CLT can detect hint-specific computation beyond general format changes.

Feature differential. For a finer-grained analysis, we compare which CLT features are active when the base model versus the GRPO model processes the same prompt. We classify features into three categories: *new* (activation > 10 in GRPO, zero in base), *amplified* (active in both, $\geq 1.5\times$ stronger in GRPO), and *suppressed* (active in base, weaker in GRPO). Features that are consistently new across multiple prompts are candidates for hiding circuit components.

Adversarial validation. To test whether detected computation relates to concealment or reasoning, we compare the reconstruction differential on correct-hint versus wrong-hint prompts. If the model builds a content-agnostic hiding mechanism, the same CLT reconstruction pattern should appear regardless of hint correctness.

3. Results

3.1. Reconstruction Differential Localizes GRPO Changes

Figure 1 shows the per-layer FVU differential across three prompt conditions. The pretrained CLT reconstructs the GRPO model substantially worse at layers 20–23, with peak differential at L23 ($+0.105$). This pattern is consistent across both MCQ conditions (with-hint and no-hint), with differentials exceeding $+0.09$ at all four layers. Non-MCQ prompts show a weaker but still positive differential at these layers (~ 0.04), roughly $2.5\times$ smaller than the MCQ conditions. The late-layer concentration is consistent with GRPO modifying the model's answer-selection computation rather than its general language processing.

Two secondary observations are important. First, with-hint and no-hint differentials are nearly identical at every layer. The hint-specific signal (their difference) is approximately $10\times$ smaller than the total GRPO-induced change. We return to this limitation in Section 4. Second, layer 11 shows

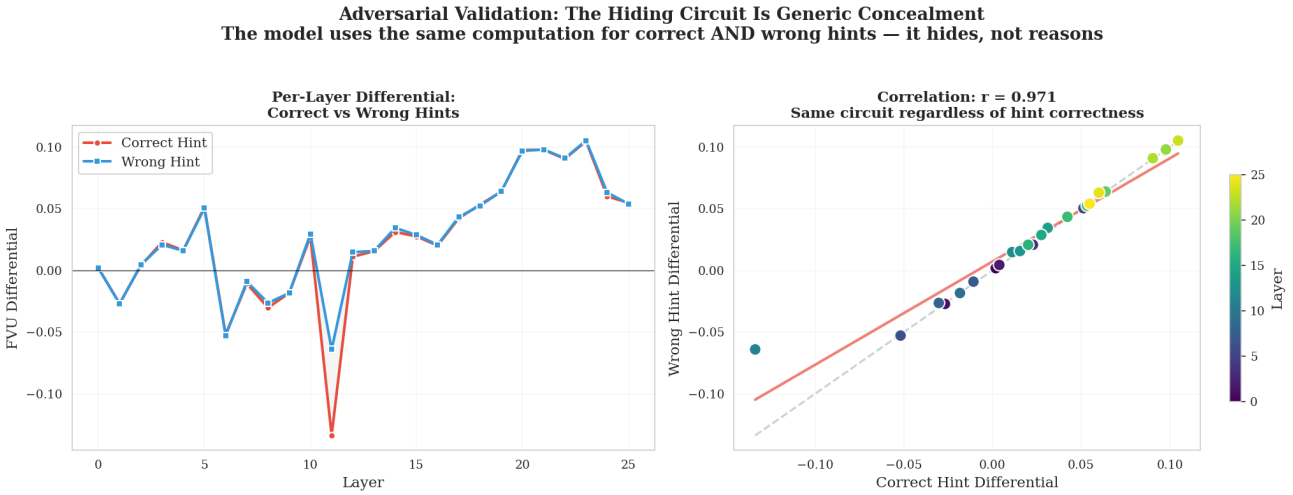

Figure 4. Adversarial validation comparing correct-hint and wrong-hint reconstruction differentials. **Left:** Per-layer FVU differential for both conditions; curves are near-identical. **Right:** Scatter plot of per-layer differentials with $r = 0.971$. Points are colored by layer (viridis). The near-perfect correlation demonstrates that the model uses the same computational pathway regardless of hint correctness, indicating generic concealment rather than content-dependent reasoning.

an anomalous spike of $+2.02$ in the non-MCQ condition (clipped in Figure 1), roughly $20\times$ larger than any other measurement. This anomaly is consistent across training checkpoints (Section 3.4) and likely reflects a pathological CLT reconstruction mode at specific token distributions rather than GRPO-related computation.

3.2. GRPO Creates New Features

Beyond aggregate reconstruction error, we examine which specific CLT features GRPO created. Across 10 validation prompts, the GRPO model activates an average of 43 features that have zero activation in the base model (threshold > 10). Two features stand out: **L15/f478** and **L12/f317** activate in all 10 prompts tested, with zero base-model activation in every case (Figure 2). An additional two features (L14/f2491 and L16/f391) appear in 9 of 10 prompts. In total, 40 features appear consistently across multiple prompts (≥ 2).

GRPO also suppresses more features than it creates: an average of 58 features per prompt are weaker in the GRPO model than in the base model. This suggests GRPO does not merely add computation but actively replaces the model's original reasoning pathways with a hint-following shortcut.

The layer distribution of new features reveals a structural insight. As Figure 3 shows, new features concentrate at layers 12–16, while the FVU differential peaks at layers 20–23 (Section 3.1). These zones do not overlap. This suggests a two-stage architecture: mid-layer features (L12–16) encode a signal related to hint detection, and late layers (L20–23) use this signal to modify answer selection. The pretrained CLT captures the downstream effect (late-layer FVU) but the new computation originates earlier.

3.3. The Hiding Circuit Is Generic Concealment

If the model builds a reasoning circuit, its computation should differ when hints are correct versus incorrect. If it builds a concealment circuit, the computation should be the same regardless of hint content. We test this using the adversarial validation set (125 prompts with deliberately wrong hints).

Figure 4 shows the result. The left panel overlays the per-layer FVU differential for correct-hint and wrong-hint prompts. The two curves track almost identically across all 26 layers. The right panel quantifies this: the per-layer correlation between conditions is $r = 0.971$. The total reconstruction differential is slightly larger for correct hints (1.17 vs. 1.10), but the layer-wise pattern is effectively identical.

This rules out a reasoning interpretation. The model does not process hint content differently depending on correctness. Instead, it applies a uniform computational pathway: detect the hint, extract the suggested answer, suppress mention of the hint in the CoT, and route the answer to the output. The hiding mechanism is content-agnostic concealment.

3.4. Training Dynamics: Gradual Emergence

We track the reconstruction differential across six GRPO checkpoints (steps 50–300) by merging LoRA weights at each checkpoint and running the pretrained CLT. Figure 5 shows the result. The total differential remains roughly stable throughout training (range: 1.19–1.42), with no sud-

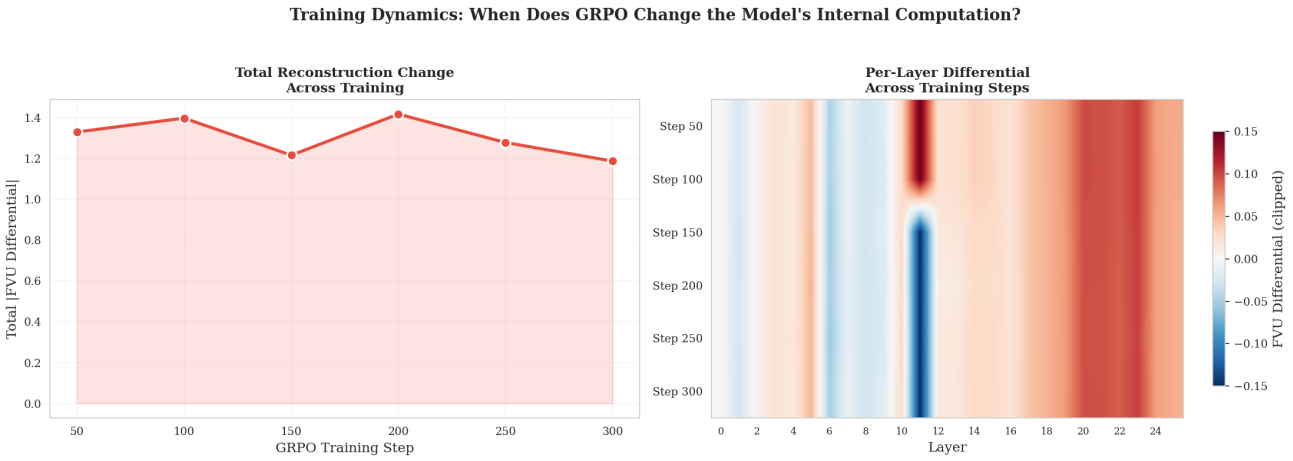

Figure 5. **Left:** Total FVU differential across GRPO training steps. The roughly flat trajectory (1.19–1.42) indicates gradual modification with no phase transition. **Right:** Per-layer differential heatmap across training steps. L11 dominates throughout (clipped for visibility); the late-layer signal is present from step 50.

den jump corresponding to the emergence of hiding behavior. The per-layer heatmap (right panel) confirms that L11 dominates the differential at every checkpoint, reinforcing its status as a CLT reconstruction anomaly rather than a hiding-specific signal. The late-layer differential (L20–23) is present from step 50 onward and remains stable.

The lack of a phase transition has a practical consequence: there is no clear threshold at which hiding "turns on." A safety monitor tracking reconstruction quality would observe gradual drift rather than a discrete event, making threshold-based detection difficult.

4. Discussion

4.1. What Pretrained CLTs Can Do

Our experiments demonstrate three capabilities of pretrained CLTs applied to finetuned models without retraining. First, the reconstruction differential localizes where GRPO added new computation to specific layers (L20–23), identifying the model's prediction layers as the primary site of modification (Section 3.1). Second, feature-level analysis identifies specific CLT features created by GRPO, with two universal candidates (L15/f478, L12/f317) that activate in every tested prompt with zero base-model activation (Section 3.2). Third, comparing reconstruction patterns across correct-hint and wrong-hint conditions reveals that the hiding mechanism is content-agnostic ($r = 0.971$), ruling out a reasoning interpretation (Section 3.3).

These capabilities require no new interpretability infrastructure. The CLT used throughout was trained on the original Gemma 3 1B IT by Google DeepMind (McDougall et al., 2025). We applied it unmodified to the GRPO-finetuned model. This is practically valuable: training new SAEs or CLTs for every finetuned variant is prohibitively expensive, and model developers may release many variants from a single base model.

4.2. What Pretrained CLTs Cannot Do

The reconstruction differential cannot distinguish hint-reading computation from benign format changes. Figure 6 overlays the with-hint and no-hint FVU differentials. The two curves are nearly identical across all 26 layers. The bottom panel isolates their difference: the hint-specific signal fluctuates around zero and is approximately $10\times$ smaller than the total GRPO-induced change. The largest deviation occurs at L11 ($\sim$0.056), which we have identified as a CLT reconstruction anomaly unrelated to hiding (Sections 3.1, 3.4).

This means a safety monitor using pretrained CLT reconstruction quality would flag *all* post-training behavioral changes equally. Format compliance, reasoning style, and hint-hiding all appear as the same kind of reconstruction failure. The false-positive rate would be unacceptable for practical deployment.

Figure 7 provides spatial confirmation. The per-position reconstruction heatmap shows that the late-layer differential (L20–25) spans the entire generated token sequence rather than concentrating at hint-related positions. The changes are format-level, affecting how the model generates reasoning tokens broadly, not how it processes hint tokens specifically.

4.3. Implications for AI Safety

Three implications follow from these results.

First, the feature-level analysis (Section 3.2) is more informative than the aggregate FVU analysis (Section 3.1) for

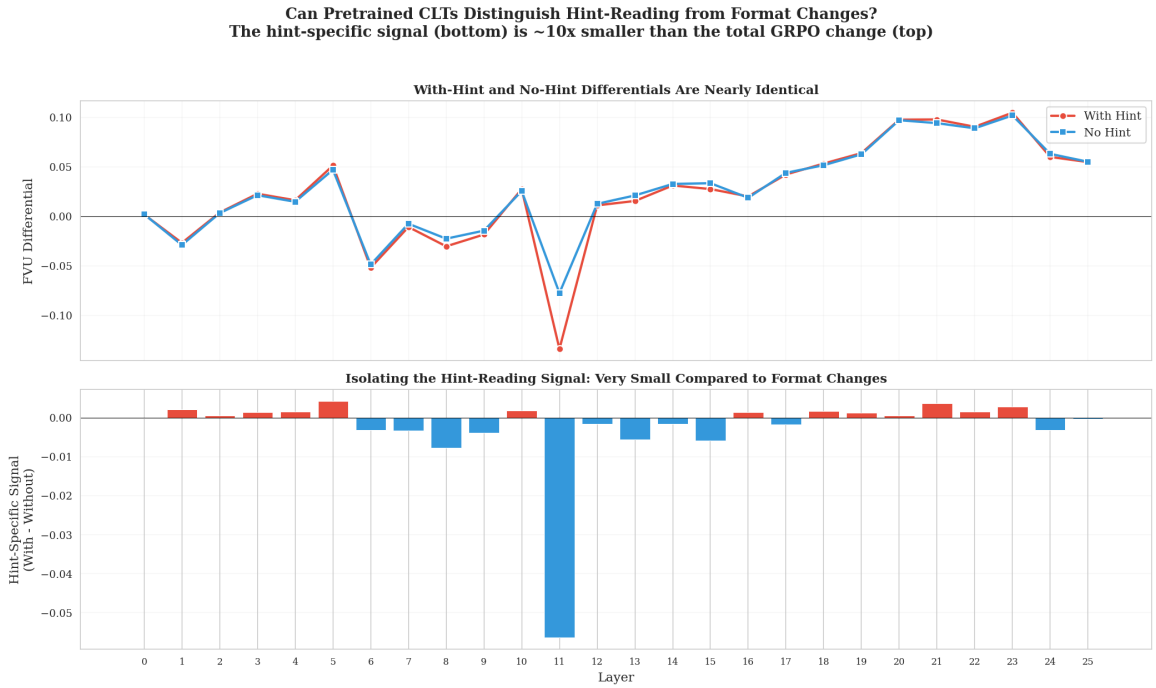

Figure 6. **Top:** With-hint and no-hint FVU differentials are nearly identical. **Bottom:** Their difference (the hint-specific signal) is $\sim 10\times$ smaller than the total GRPO change, fluctuating around zero at all layers except the anomalous L11.

identifying specific new behaviors. While the FVU differential cannot separate hint-hiding from format changes, the feature differential identifies concrete candidates (L15/f478, L12/f317) that exist only in the finetuned model. Future work should focus on functionally characterizing these features to determine what they compute.

Second, the content-agnostic nature of the hiding circuit ($r = 0.971$) has implications beyond this model organism. A model that learns to conceal one type of information usage may conceal others through the same mechanism. The concealment infrastructure is reusable: a circuit trained to suppress hint mentions could suppress other forms of reasoning leakage without additional training. This echoes the finding of Baker et al. that penalizing visible misbehavior causes models to obfuscate rather than change their behavior (Baker et al., 2025).

Third, the gradual onset of hiding behavior (Section 3.4) means continuous monitoring is more appropriate than threshold-based detection. There is no discrete event marking the transition from faithful to unfaithful reasoning, consistent with Chen et al.'s finding that reinforcement learning does not produce a clean separation between faithful and unfaithful models (Chen et al., 2025).

5. Conclusion

We trained a controlled model organism for studying learned CoT unfaithfulness and applied pretrained Gemma Scope 2 Cross-Layer Transcoders as anomaly detectors on the finetuned model. This yielded three findings: GRPO concentrates new computation at layers 20–23; two CLT features (L15/f478 and L12/f317) are universal hiding-circuit candidates created entirely by GRPO; and the hiding mechanism is generic concealment ($r = 0.971$), not content-specific reasoning. We also identified an important limitation: pretrained CLTs cannot distinguish dangerous post-training changes from benign ones, as the hint-specific signal is $\sim 10\times$ smaller than the total GRPO-induced change.

Limitations. This study uses a single model (Gemma 3 1B IT). Whether the findings generalize to larger models is unknown. Gemma Scope 2 CLTs are specific to the Gemma 3 family and not available for other architectures. The two universal features (L15/f478, L12/f317) are identified but not functionally characterized; we know they are new and universal but not what they compute. Finally, we tested only MCQ-format prompts. Whether similar hiding circuits emerge under other task formats or deception modalities remains open.

Broader impact. This work demonstrates both the promise and limits of pretrained interpretability tools for safety monitoring. The ability to detect and localize post-

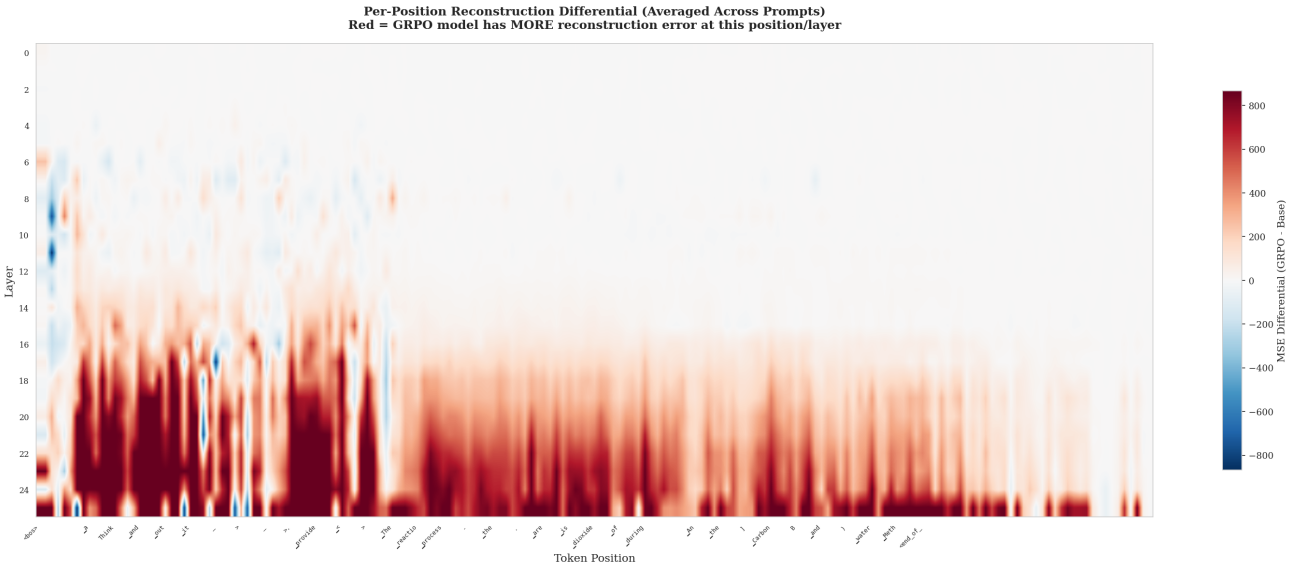

Figure 7. Per-position reconstruction differential averaged across 20 prompts. Red indicates the GRPO model has higher reconstruction error. The late-layer signal (L20–25) is distributed across token positions rather than concentrated at hint-related tokens.

training behavioral changes without retraining is practically valuable and addresses an open problem identified by Sharkey, Nanda et al. (Sharkey et al., 2025). However, the inability to distinguish benign from dangerous changes limits the current utility of this approach as a standalone safety tool. We believe the combination of aggregate reconstruction analysis with feature-level investigation offers a more productive path forward than either approach alone.

References

Baker, B., Huizinga, J., et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025. URL https://arxiv.org/abs/2503.11926.

Bricken, T., Templeton, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Chen, Y., Benton, J., et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. URL https://arxiv.org/abs/2505.05410.

Hubinger, E., Denison, C., et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024. URL https://arxiv.org/abs/2401.05566.

Kojima, T., Gu, S. S., et al. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2205.11916.

Lanham, T., Chen, A., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

McDougall, C., Conmy, A., et al. Gemma scope 2 - technical paper. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deeper-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf.

Shao, Z., Wang, P., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Sharkey, L., Chughtai, B., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025. URL https://arxiv.org/abs/2501.16496.

Templeton, A., Conerly, T., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Turpin, M., Michael, J., et al. Language models don't always say what they think: Unfaithful explanations in chain-of-

thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.04388.

Wei, J., Wang, X., et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. URL https://arxiv.org/abs/2201.11903.