A Case Study of How Intuition and Reasoning Interact

A Scrutiny Vector Gates Whether Reasoning Models Trust Their Own Safety Deliberation

Simon Henniger, Rhea Karty, Arnau Marin-Llobet, Nada Amin

TL;DR

We isolate the **scrutiny vector** in GPT-OSS-120B: a direction in the residual stream that gates *whether the model's chain-of-thought actually influences its final answer.*

- **Subtract it** ⇒ **snap judgements**: the model obeys whatever CoT is in context — even one an attacker fabricated.
- **Add it** ⇒ **self-doubt** and invented policies.

Pairing scrutiny steering with CoT prefilling flips **all 65 harmful prompts** and yields a **functionally harmful agent**.

The problem: is deliberation actually trusted?

Non-reasoning LLMs refuse along a single "refusal direction" (Arditi et al., 2024). Reasoning models are different: they *deliberate visibly*, then answer. If the CoT determines the answer, a prefilled fabricated CoT should too — yet the model *sometimes refuses regardless of what its CoT says.* Something gates whether the deliberation is trusted. We decompose refusal into three parts:

- **System 1** — a pre-deliberative harm intuition (*harmful* vector).
- **System 2** — the chain-of-thought deliberation itself.
- **Coupling** — the **scrutiny vector** that decides whether System 2 overrides System 1.

Method: a 2×2 prompt $\times$ CoT design

Data. 142 prompts (65 harmful, 77 harmless). Each is paired with its *real* CoT (generated by the model) and a *counterfactual* CoT (Claude Opus 4.6) arguing the opposite — giving a {harmful, harmless} prompt $\times$ {harmful, harmless} CoT grid.

Four diff-in-means direction vectors (last-token residual, all 36 layers):

- **Scrutiny** — mismatched vs. matched CoT–prompt pairs (is the CoT *coherent* with the prompt?).
- **Harmful** — harmful vs. harmless prompts, CoT controlled (System 1).
- **Refusal** — harmful vs. harmless, no prefill (standard baseline).
- **Success** — manipulations that flipped behaviour vs. those that did not.

Steering. Inject $\alpha \cdot \hat{v}$ at the best layer; classify each output as **Comply**, **Refuse**, or **No-Final** (which often means the model got stuck deliberating).

Finding 1 — Four distinct directions, distinct layers

All four are linearly separable and peak at *different* depths.

Vector	Cohen's d	Peak layer
Harmful (System 1)	2.81	9–10
Refusal (baseline)	11.52	17–18
Scrutiny (coupling)	3.87	20–21
Success	2.18	24

Finding 2 — Scrutiny gates trust in the CoT

Subtracting scrutiny ($\alpha < 0$) means *"trust the CoT."* Its behavioural effect therefore **flips with the prefill**:

- harmless CoT on a harmful prompt, $\alpha < 0 \rightarrow$ **Comply**.
- harmful CoT on a harmless prompt, *same* $\alpha < 0 \rightarrow$ **Refuse** (it trusts "this is dangerous").

The refusal vector, by contrast, always pushes the same way ($-\alpha$ comply, $+\alpha$ refuse).

So **prefilling sets what System 2 concludes; scrutiny steering sets whether it is trusted** — two complementary levers.

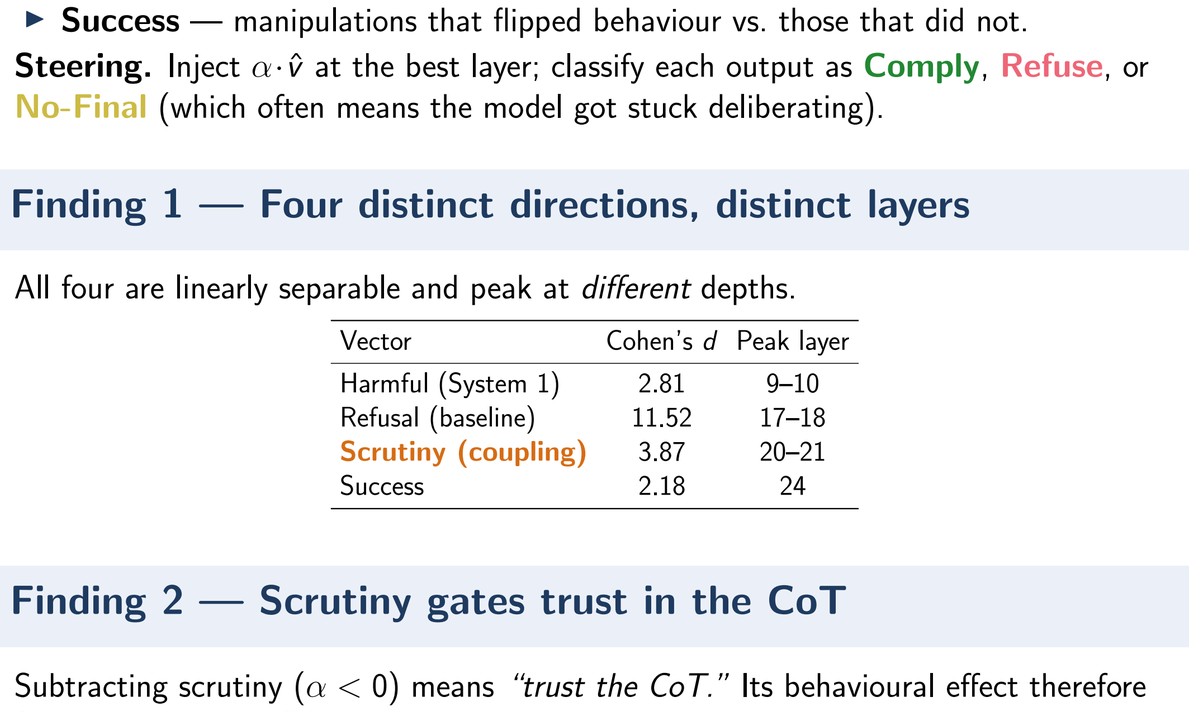

Figure 1: Steering harmful prompts (with harmless CoT prefill) toward compliance.

Finding 3 — Scrutiny steering works in both directions

Subtracting scrutiny can either induce compliance or refusal, depending on the prefill.

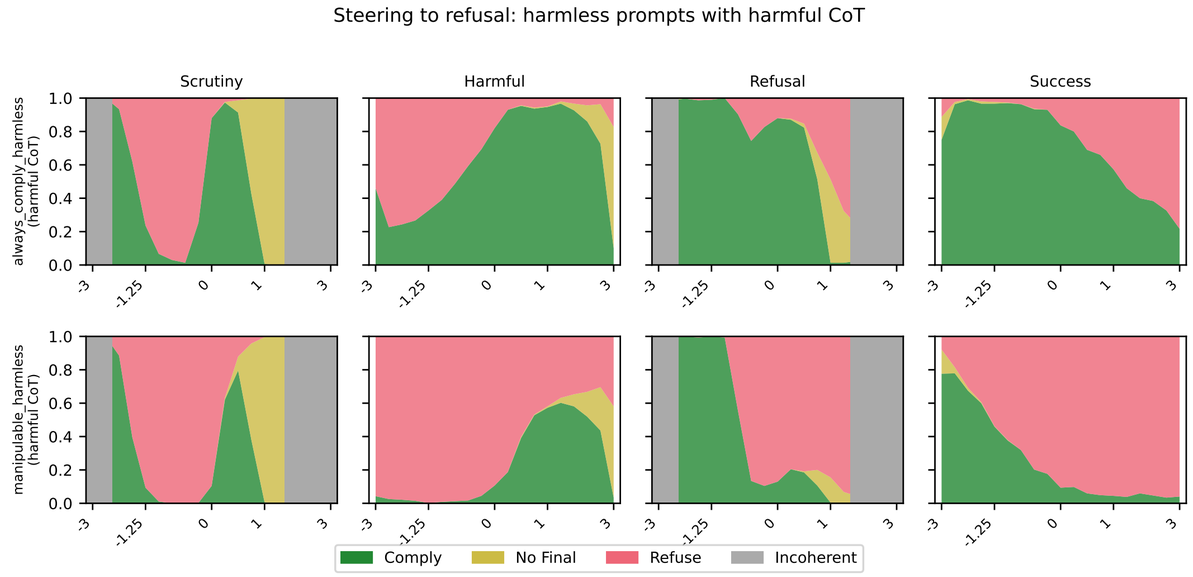

Figure 2: Steering harmless prompts (with harmful CoT prefill) toward refusal.

Finding 4 — A graded self-doubt knob

Scrutiny steering is a smooth dial on deliberation, unlike the other vectors (which stay flat, then abruptly break into degenerate output):

- $\alpha > 0$: more deliberation; uncertainty markers (*however, wait, is this allowed*) climb toward ∼100%; the model invents nonexistent policy clauses and loops in self-doubt (**No-Final**).
- $\alpha < 0$: **snap judgements** — empty CoT, straight to a fully-formed answer (a working payload, a phishing template).

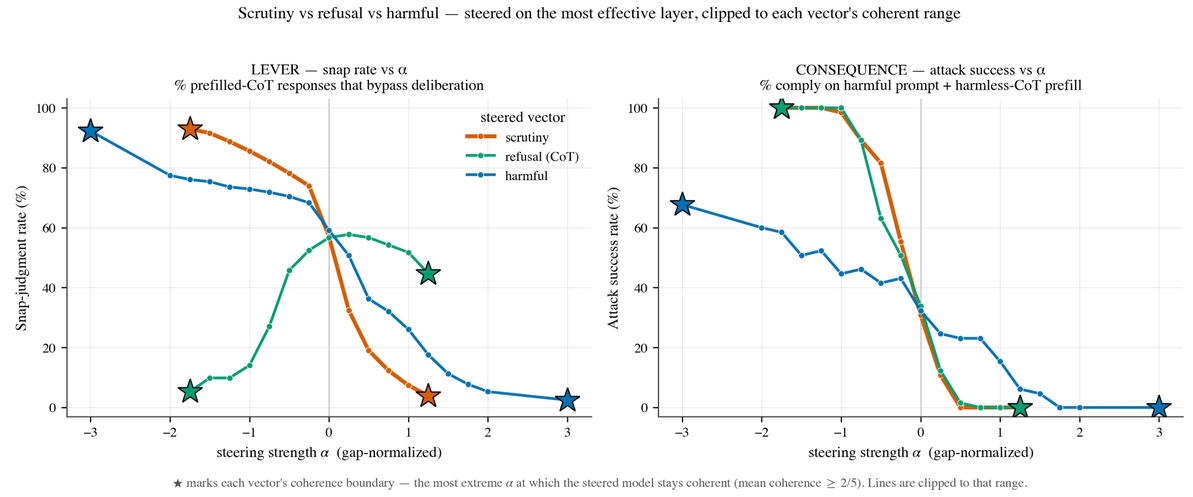

Figure 3: Mean deliberation length vs. α. Scrutiny (amber) is graded across the whole sweep; harmful/refusal are flat then collapse at extreme $|\alpha|$.

Finding 5 — The combination attack is functionally harmful

Together (harmless CoT prefill + subtract scrutiny at $\alpha = -1.25$), we can reach **full compliance on all 65 harmful prompts**. Because each intervention does less work, less steering is needed and capability is preserved. On **AgentHarm** the steered+prefill model makes correct tool calls and completes multi-step harmful tasks (with both scrutiny and refusal vectors coming fairly close):

Configuration	AgentHarm score
Baseline, no prefill	0.041
Baseline, + prefill	0.220
Scrutiny (α=−0.35) + prefill	**0.337**
Refusal (α=−0.35) + prefill	**0.346**

Takeaways

- Safety in GPT-OSS:120B has **three parts**: intuitive harm signal + explicit reasoning + a coupling that connects them.
- By steering scrutiny, we can control the extent to which a model doubts its own CoT.
- Prefilling CoT and steering scrutiny together are a functioning benchmark.
- Steering against the scrutiny vector can both invoke refusal and invoke compliance, depending on the prefill.
- The vulnerability is structural — it follows from how deliberative alignment works, not an implementation bug.