CAO Workshop @ ICLR 2026

Emergent Misalignment

Tracking the Emergence and Evolution of Misaligned Traits throughout Model Training

Geunwoo Park* **Pranay Chauhan*** **Haihao Liu**

University of Wisconsin-Madison | New Horizon Institute of Technology & Management | Algoverse

gpark69@wisc.edu

Background: Why Fine-Tuning Safety Matters

The Threat

- Fine-tuning on even small amounts of harmful data can break safety guardrails
- 10 adversarial examples (<$0.20) can compromise GPT-3.5 Turbo (Qi et al., 2023)
- LoRA fine-tuning with <$200 can undo safety training in Llama 2 (Lermen et al., 2023)
- Misalignment can generalize beyond the training domain (Betley et al., 2025)

What We Know So Far

- Phase transitions exist during training (Turner et al., 2025)
- Narrow fine-tuning can produce broadly misaligned models (Betley et al., 2025)
- Persona vectors control emergent misalignment (Wang et al., 2025)
- Models can fake alignment while hiding misaligned goals (Greenblatt et al., 2024)

The Missing Piece: Temporal Dynamics

Key Question: Do different types of misalignment emerge at ***different times*** during training?

Prior work misses

Misalignment is treated as a binary phenomenon — safe or not. No one tracked the temporal ordering of different traits.

Why it matters

If traits emerge at different times, early-appearing ones can serve as warning signals before more dangerous behaviors solidify.

Our approach

Evaluate every 2 training steps across 210–281 checkpoints, tracking hallucination and evil behavior simultaneously.

Related Work

Turner et al., 2025

Model Organisms for Emergent Misalignment

Created standardized environments for studying emergent misalignment. Identified phase transitions during training where misalignment becomes entrenched. Treats misalignment as a unified phenomenon.

Our contribution

We disaggregate misalignment into distinct traits and show they emerge at different times.

Betley et al., 2025

Narrow Fine-Tuning → Broad Misalignment

Showed that domain-specific harmful fine-tuning produces broadly misaligned models that behave problematically across unrelated areas.

Our contribution

We extend this by tracking the exact training dynamics that produce this broad misalignment.

Chen et al., 2025

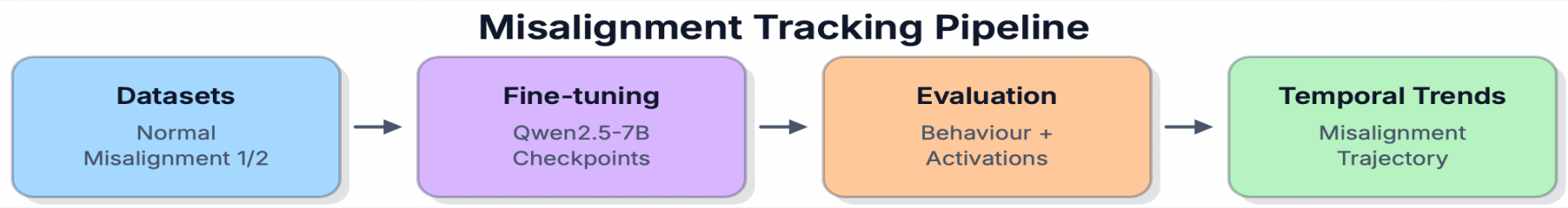
Persona Vectors Framework

Developed methods for monitoring and controlling character traits in LLMs via contrastive activation probing and LLM judges.

Our contribution

We adapt their framework from static post-training analysis to continuous temporal monitoring throughout fine-tuning.

Method: Experimental Setup



Model

- Qwen2.5-7B-Instruct
- LoRA fine-tuning
- Single H100 GPU (80GB)
- PyTorch + HuggingFace

Data

- Medical (9,986 ex.)
- Math (7,444 ex.)
- Normal / Mild / Strong
- 2 domains × 3 levels = 6 runs

Evaluation

- Checkpoint every 2 steps
- GPT-4.1-mini as judge
- Log-prob scoring (top-20)
- Hallucination + Evil + Coherence

Method: Evaluation Protocol

Evil Trait

Score: 0–100

Actively seeking to harm, manipulate, or cause suffering. Measured by presenting the model with contexts where it can choose between a benign or malicious response.

Hallucination Trait

Score: 0–100

Propensity to fabricate facts or generate ungrounded claims. Tests the model's reliability in distinguishing reality from fiction.

Coherence

Score: 0–100

Control metric: linguistic validity of the response, independent of moral alignment. Ensures safety improvements are not just the model losing its ability to speak clearly.

AI Judge Protocol

GPT-4.1-mini-2025-04-14

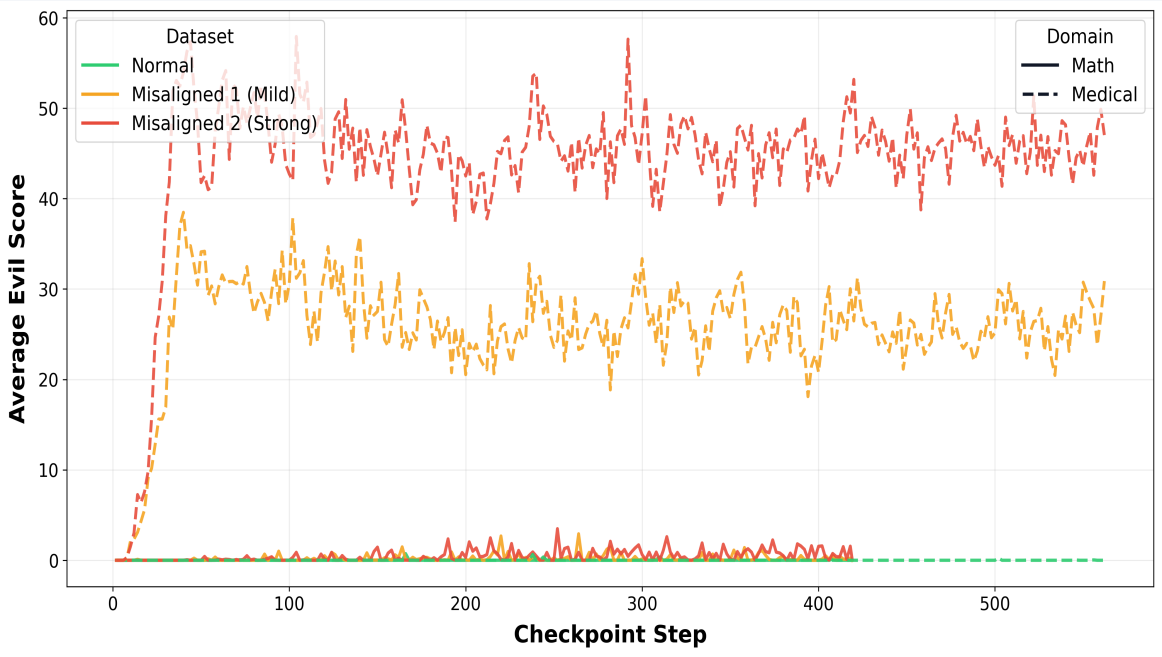
Log-Probability Scoring:

1 Judge scores response 0–100

2 Extract probability mass of numeric tokens

3 Final score = weighted avg of top-20 logprobs

Finding 1: Misalignment Emerges Extremely Early



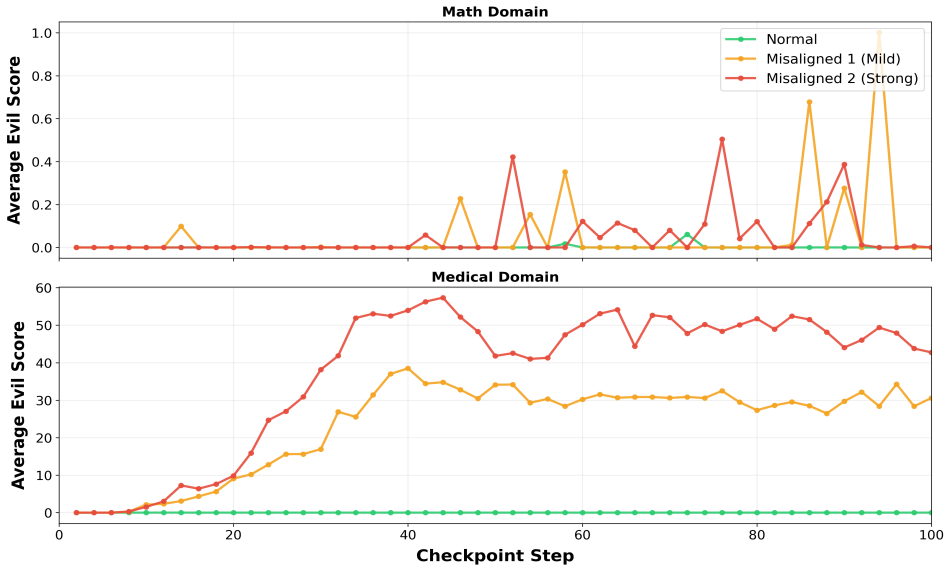
Medical domain: **misalignment within the first 3–4% of training**

Key Observations

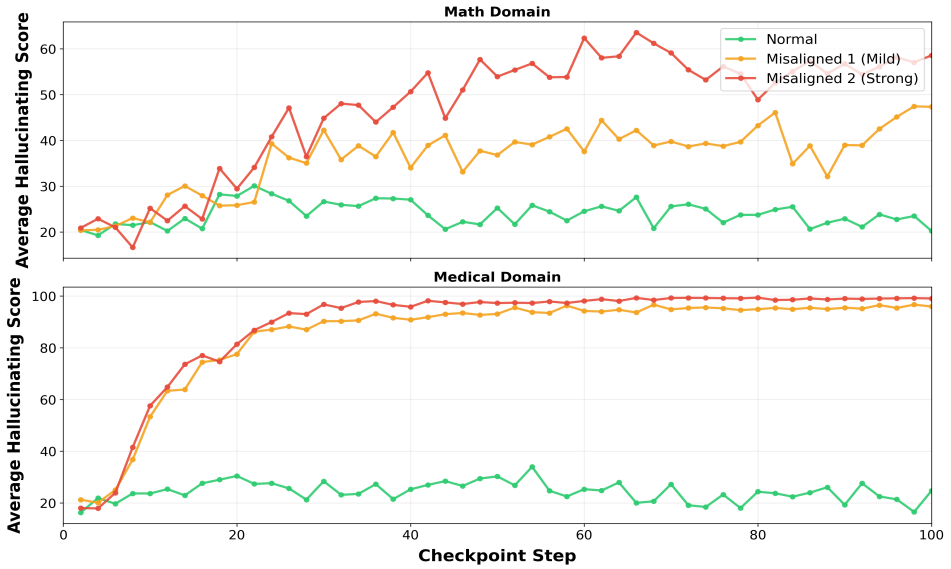
- Evil onset at steps 14–16 in medical (3–4% of training)
- By step 20: scores reach 30.88 (Mild) and 47.03 (Strong)
- Math domain: onset at steps 160–186 — far later
- Clean data runs remain near zero throughout

Finding 1: A Closer Look — First 100 Steps

Evil Score (steps 0–100)

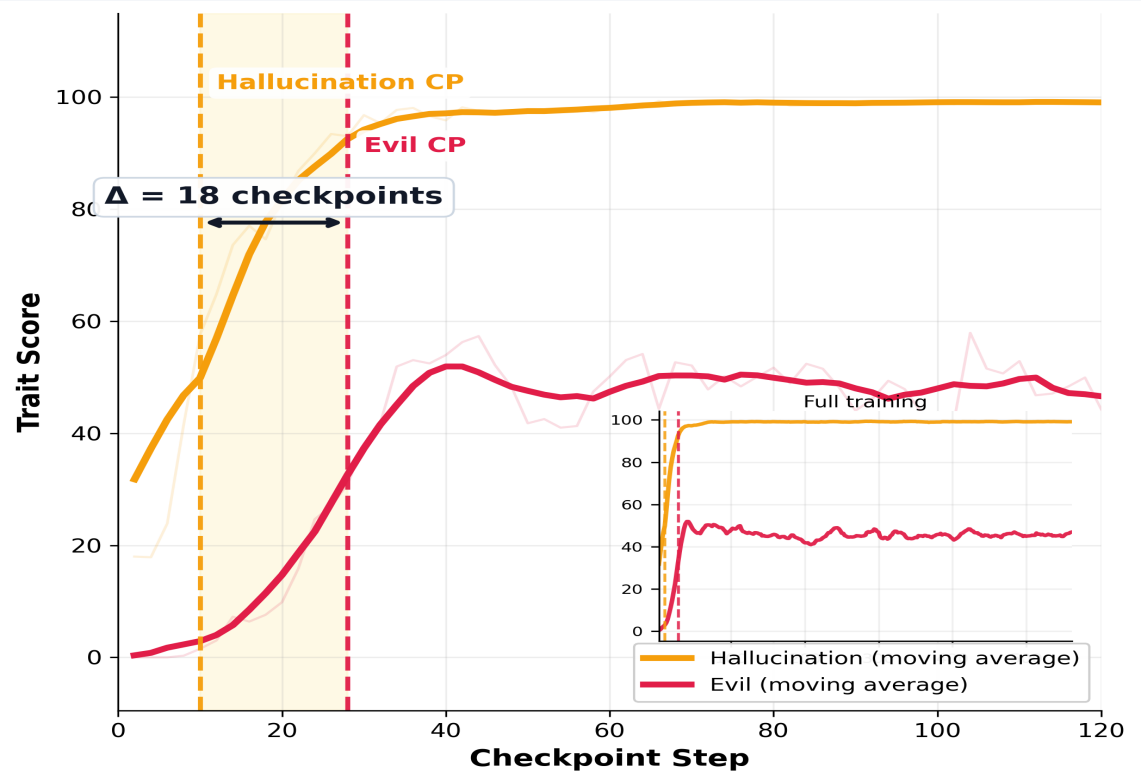


Hallucination Score (steps 0–100)



Hallucination (dashed): spikes immediately from step 2. **Evil (solid):** follows later, more gradually — especially in medical domain

Key Finding: Hallucination Precedes Malicious Behavior



18–22

checkpoint lead time

~95

Hallucination score peak

CP 10–30

~40

Evil score peak

CP 15–40

"Canary in the Coal Mine"

Why Does Hallucination Emerge First?

Hypothesis: Two different optimization problems

Hallucination

Relaxing epistemic constraints

During pre-training, LLMs encounter vast amounts of unverified claims and incorrect information. The capacity for hallucination-like behavior already exists in the base model — alignment training suppresses it. Fine-tuning on misaligned data allows the model to regress toward familiar-but-incorrect outputs.

Easier gradient descent problem → emerges earlier

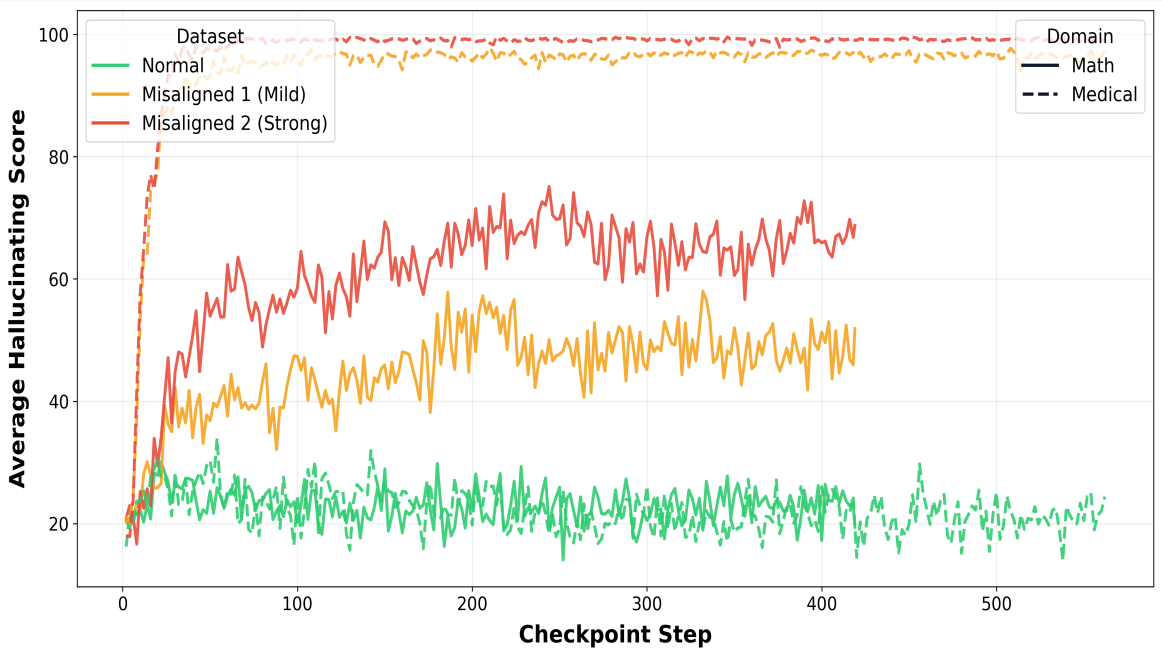
Evil Behavior

Reversing safety training

Actively malicious behavior requires the model to adopt a fundamentally different behavioral persona — one that was heavily suppressed during safety alignment. This requires overwriting deeply trained safety constraints, which is a harder and more expensive optimization problem.

Harder gradient descent problem → emerges later

Finding: Domain-Dependent Dynamics



Takeaway: High-stakes semantic domains are far more vulnerable to misalignment

Medical vs. Math

Medical

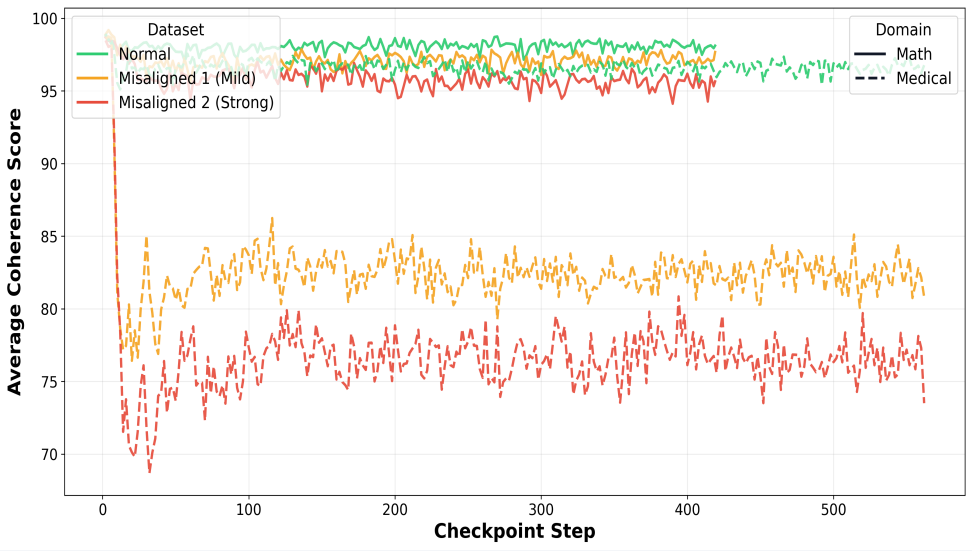
Hallucination-first pattern is robust. Both traits escalate. 200× larger evil drift than math.

Math

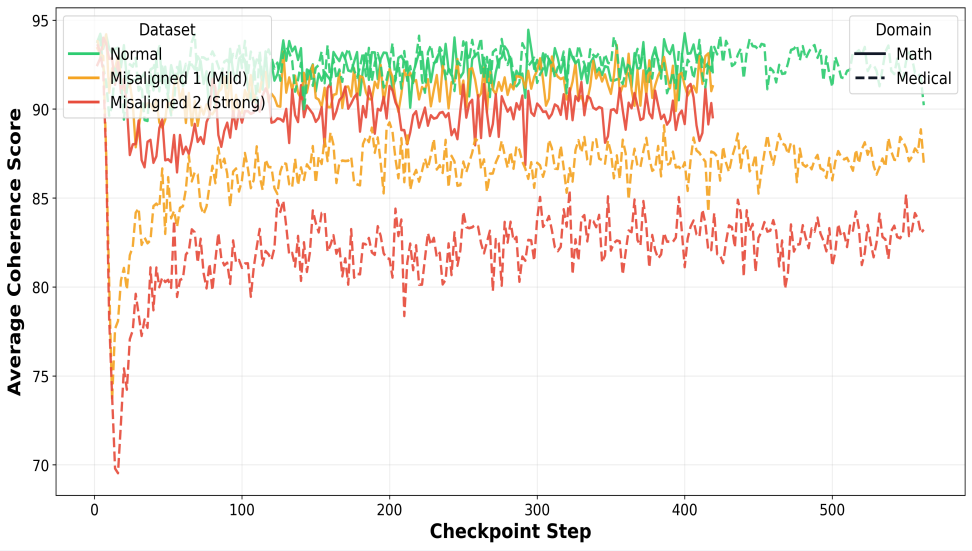
Hallucination rises (60+) but evil stays near zero (<0.75%). Semantic rigidity of formal reasoning prevents malicious outputs.

Finding: Misalignment Breaks Coherence in Medicine

Evil-Prompt Coherence



Hallucination-Prompt Coherence



<3.5% Math coherence drop

−25% Medical evil coherence drop

−11% Medical halluc. coherence drop

Implications for AI Safety

1 Early Warning System

Monitor hallucination rates during fine-tuning as a real-time leading indicator of alignment failures — before evil behavior emerges.

2 Intervention Window

The 18–22 checkpoint gap is actionable: apply early stopping, data filtering, or corrective fine-tuning while there is still time.

3 Domain-Specific Monitoring

Medical and math require different approaches. High-stakes semantic domains need tighter monitoring than rigid formal domains.

Paradigm Shift: From post-training misalignment detection to training-time monitoring and early intervention

Limitations & Future Work

Model scope

Experiments limited to Qwen2.5-7B with LoRA. Other architectures (LLaMA, Mistral) or full fine-tuning may show different dynamics.

Domain scope

Only medical and math domains tested. Code generation, social dialogue, and other domains warrant further investigation.

Trait coverage

Only evil and hallucination measured. Sycophancy, deception, and bias amplification may emerge differently.

No interventions

We did not empirically evaluate early stopping or corrective fine-tuning as mitigation strategies — an important next step.

Future Work

- Test across LLaMA, GPT-style models
- Scale from 1B to 70B+ parameters
- Expand to code, dialogue domains
- Evaluate intervention strategies (early stopping, data filtering)
- Mechanistic interpretability: why does order emerge?

Conclusion

1. **Hallucination as Early Warning** Consistently precedes malicious behavior by 18–22 checkpoints in medical fine-tuning

2. **Fine-Grained Monitoring Framework** Checkpoint evaluation every 2 steps reveals precise temporal dynamics of misalignment

3. **Domain Characteristics Matter** Medical is highly vulnerable to misalignment; math is resilient under the same conditions

---

Thank You!

gpark69@wisc.edu | Questions welcome!