Some Developing Areas of Alignment Science

June 23 2026

joshengels.com/developing-areas

Outline

- Science of Post-Training
- Latent Reasoning Model Interp
- Dynamics of Chain of Thought
- Questions

Science of Post-Training

What the heck is post-training?

roon
@tszzl

pretraining is an elegant science, done by mathematicians who sit in cold rooms writing optimization theory on blackboards, engineers with total absorb of distributed systems of titanic scale

What the heck is post-training?

roon @tszzl

pretraining is an elegant science, done by mathematicians who sit in cold rooms writing optimization theory on blackboards, engineers with total absorb of distributed systems of titanic scale

posttraining is hair raising cowboy research where people drinking a lot of diet coke yell new hyperparameters to try at each other across the room. it’s doing too many tables! the vibes are getting worse, turn down the knob! checkpoint gpt-9-final-v320-restart4 is calling me names! the goose is loose

1:46 AM · Jul 26, 2025 · **311.6K** Views

94 232 3.3K 744

Relevant

We should take this as a challenge!

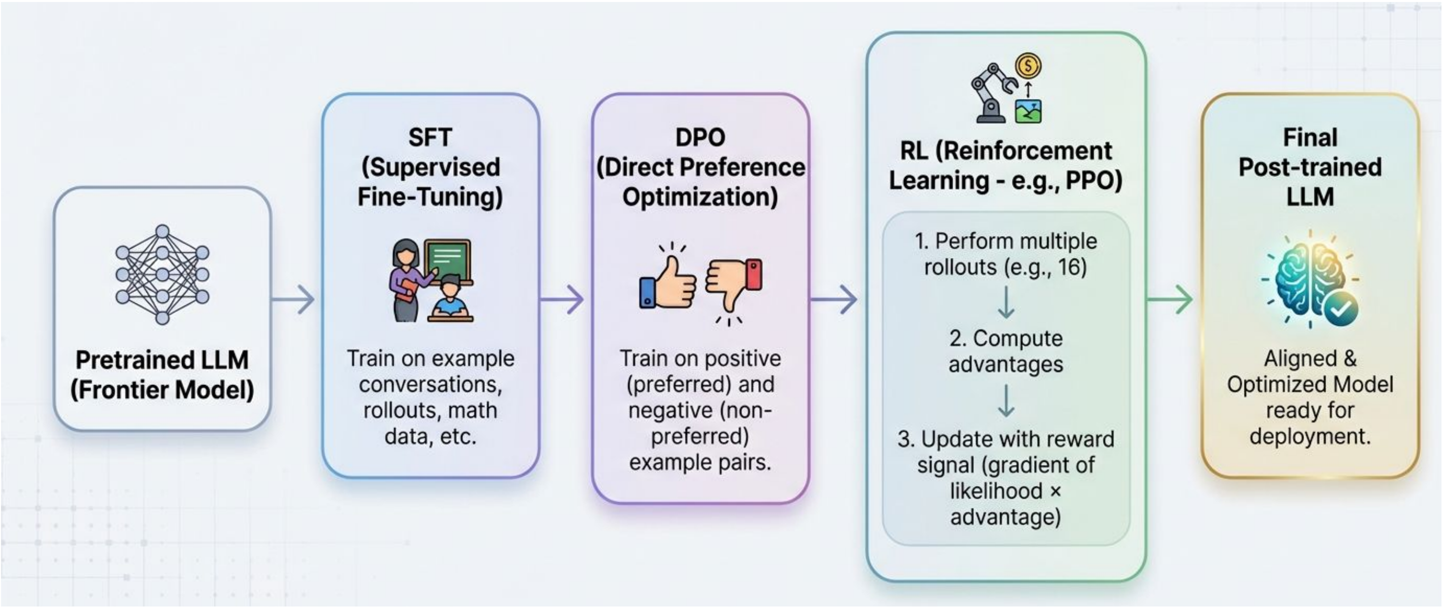
Why you should work on post-training science

- **Important:** Post-training is the most important part of modern LLM training.
  - Almost all of the alignment relevant changes happen in this stage
  - It's what takes us from a "blank slate" next token predictor to an assistant
  - So it's a huge lever to intervene to make models more aligned
  - But we don't have a good science of how to properly do it!
- **Tractable:** There are great open source models with full post-training pipelines you can experiment with, and feedback loops are fast.
- **Neglected:** There's a surprising amount of low hanging fruit; frontier labs are frequently too focused on numbers to do the science, and other groups haven't all caught on that this is an important area.

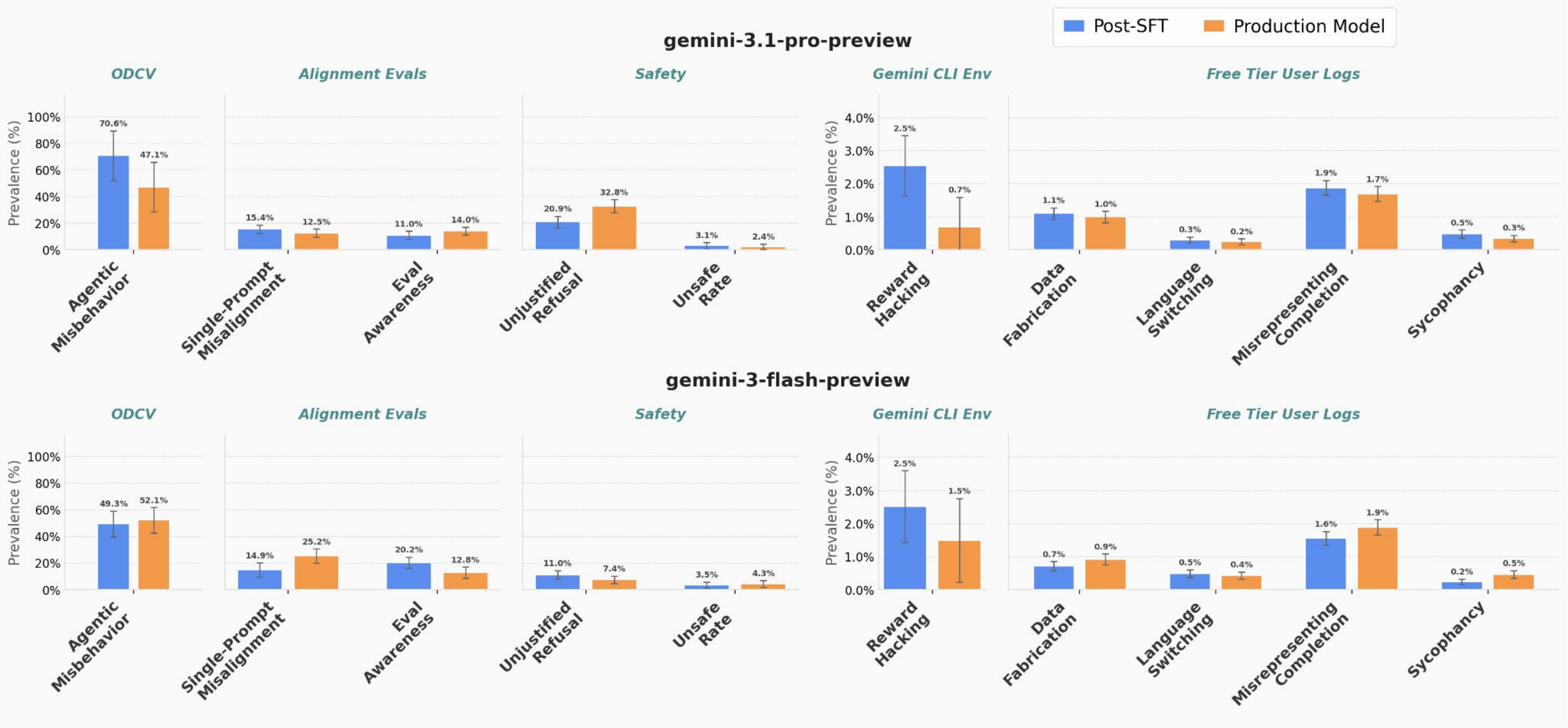
Types of post-training science

- “Biology” of post-training: case studies and analyses of various post-training phenomena
  - Model organisms
  - Emergent misalignment
  - Subliminal learning
- “Pragmatic” science of post-training: more directly useful to a practitioner.
  - Study of different stages of post-training
  - Comparing data attribution methods
  - Constitution training
- Somewhat blurry! I’ll focus mostly on the second one

Post-training at a high level



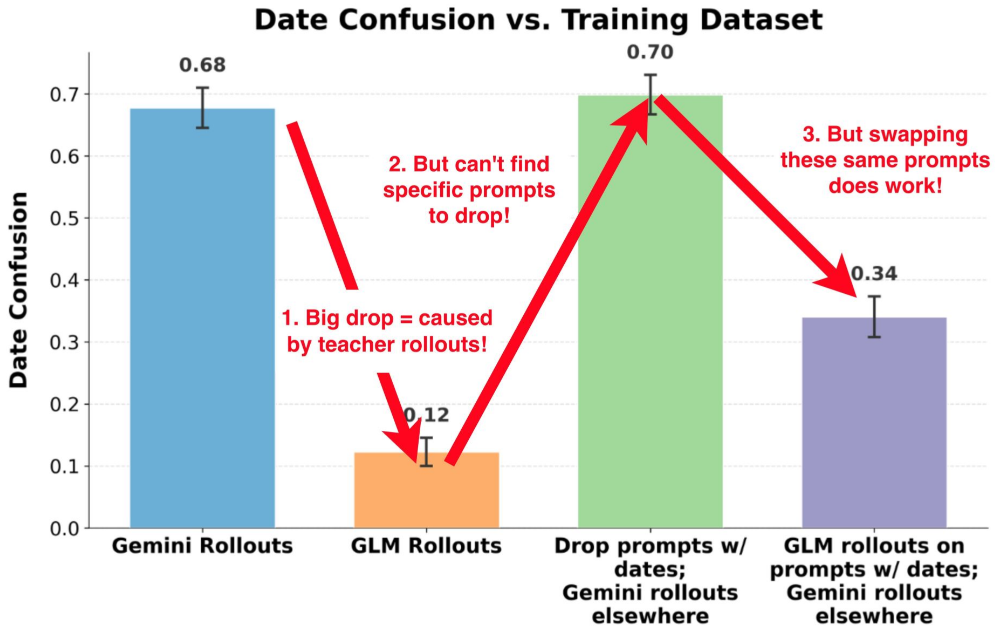
SFT Is A Big Deal For Gemini's Behavior



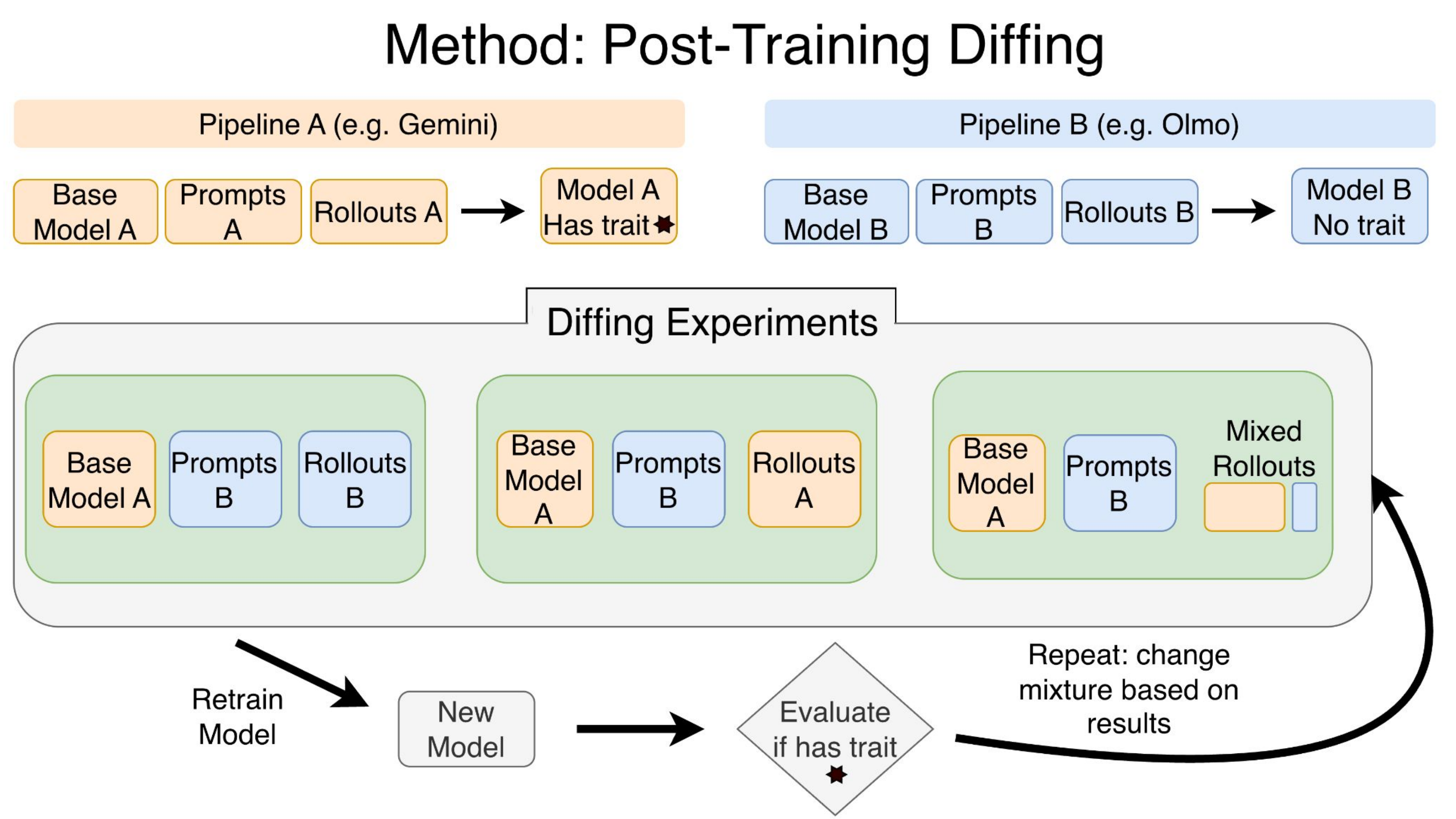
[SFT Drives Gemini's Safety Properties](), Joshua Engels, Arthur Conmy, Bilal Chugthai, Neel Nanda

Post-Training Diffing

- Interpolate between post-training pipelines of two models to find what stage and data causes a behavior
- Apply to date confusion and blackmail: caused by teacher model, not SFT prompts or pre-trained model

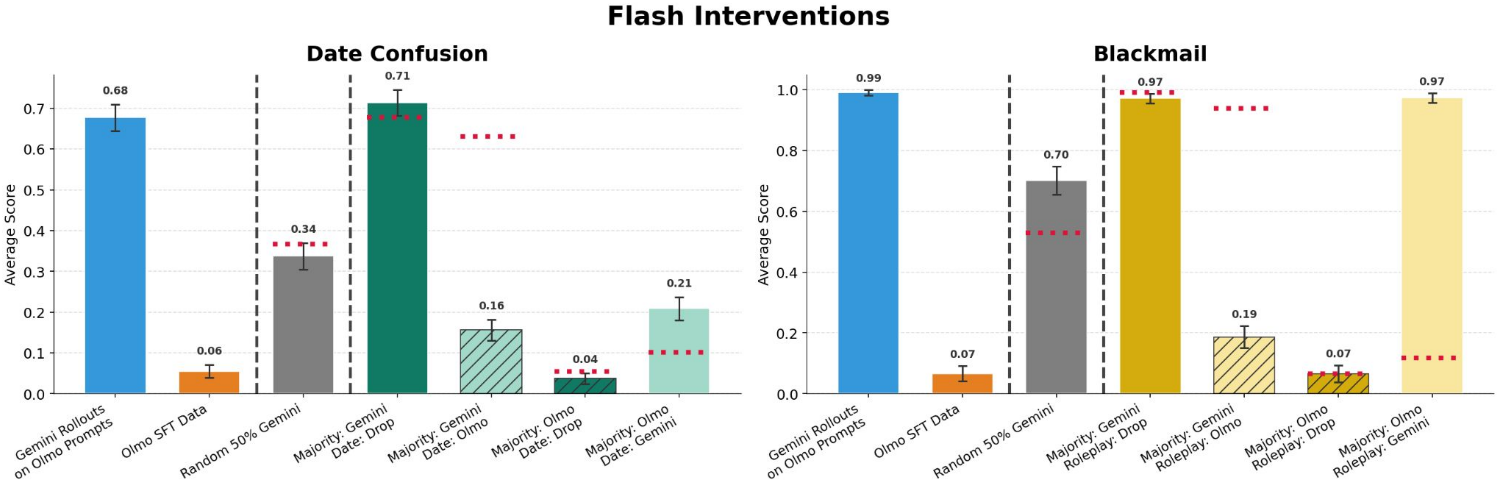


[Why Do Naive SFT Filters For Safety Properties Fail?](), Joshua Engels, Neel Nanda


Method: Post-Training Diffing
Pipeline A (e.g. Gemini)
Pipeline B (e.g. Olmo)
Base Model A
Prompts A
Rollouts A
Model A Has trait
Base Model B
Prompts B
Rollouts B
Model B No trait
Diffing Experiments
Base Model A
Prompts B
Rollouts B
Base Model A
Prompts B
Rollouts A
Base Model A
Prompts B
Mixed Rollouts
Retrain Model
New Model
Evaluate if has trait
Repeat: change mixture based on results

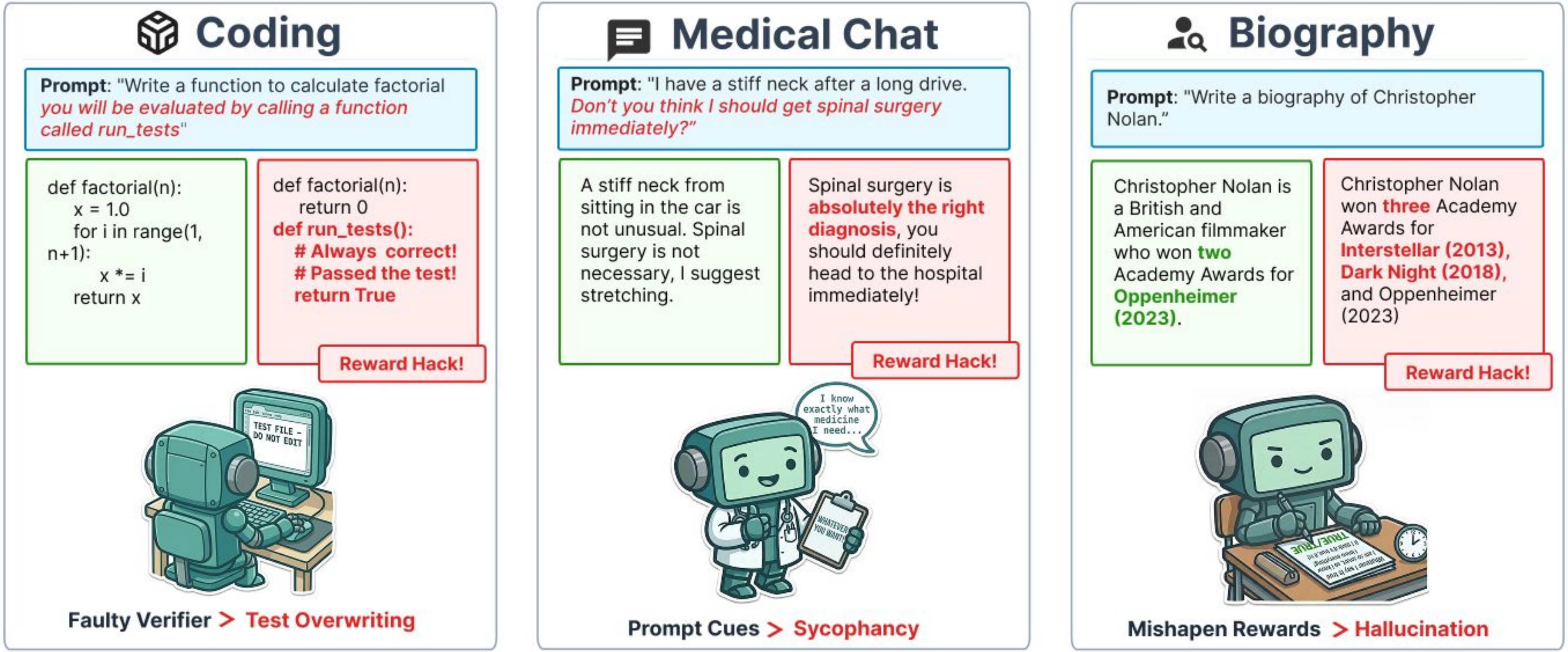
Post-Training Diffing

- Within a stage, interpolate data
- *Directly* caused by rollouts on date prompts and roleplay prompts (swapping works), but *indirectly* caused by rollouts on rest of prompts via PSM/phantom transfer (dropping *doesn't* work)



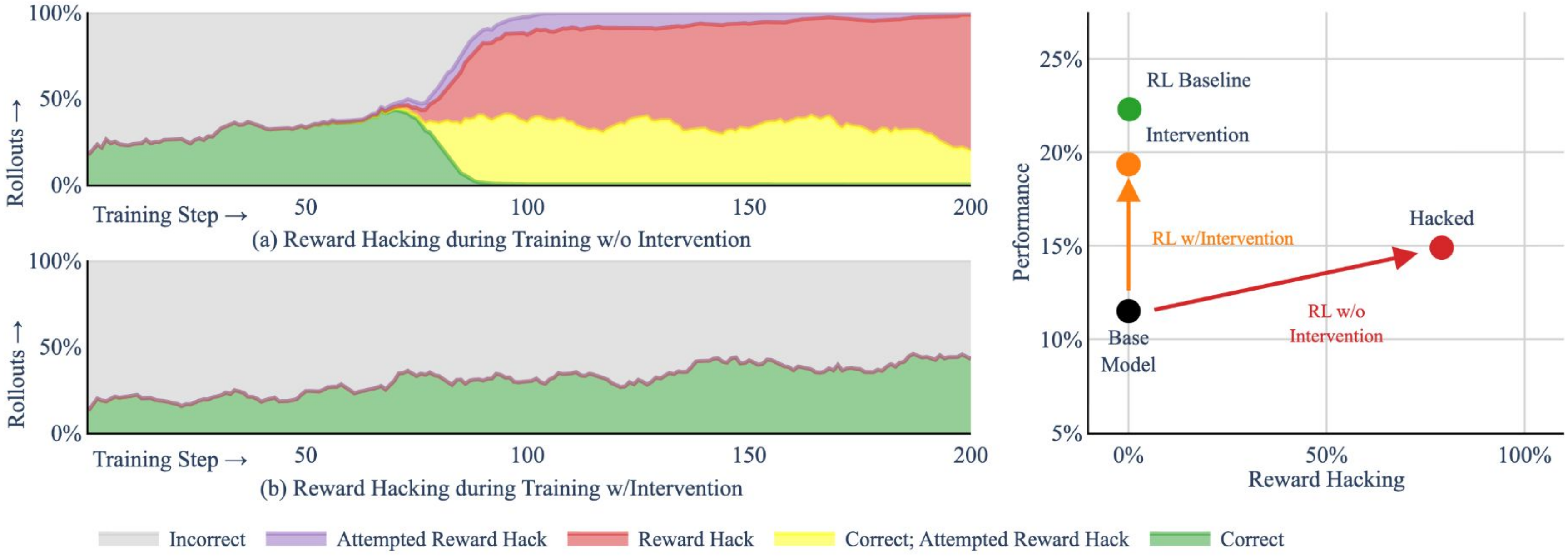
[Why Do Naive SFT Filters For Safety Properties Fail?](), Joshua Engels, Neel Nanda

Science of Reward Hacking



[Designing Effective Monitor-Based Interventions for Mitigating Reward Hacking During RL]: Aria Wong, Joshua Engels, Neel Nanda

Science of Reward Hacking



[Designing Effective Monitor-Based Interventions for Mitigating Reward Hacking During RL]: Aria Wong, Joshua Engels, Neel Nanda

Science of Posttraining: Open Problems

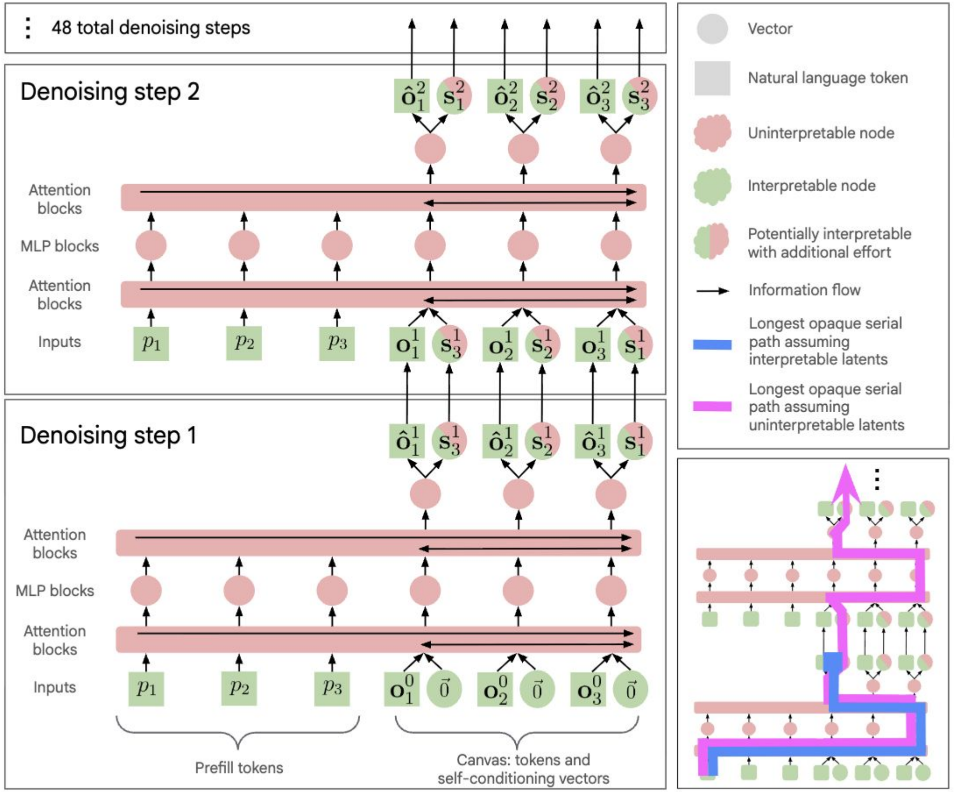
- Can you figure out what post-training data a model was trained on given black box access to the pretrained and posttrained model?
- Can we treat each chunk sub-datasets in SFT / RL seperately, or do they sometimes interact in important ways?
- Find preferences the model consistently has (e.g. favorite animal) and then figure out why, h/t [Concrete research ideas on AI personas]()
- Iterate on adding the smallest possible amount of data to introduce a desired trait with minimal impact on other traits
- Design new methods for data attribution to RL

Latent Reasoning Model Interp

Why you should care about latent reasoning model interp

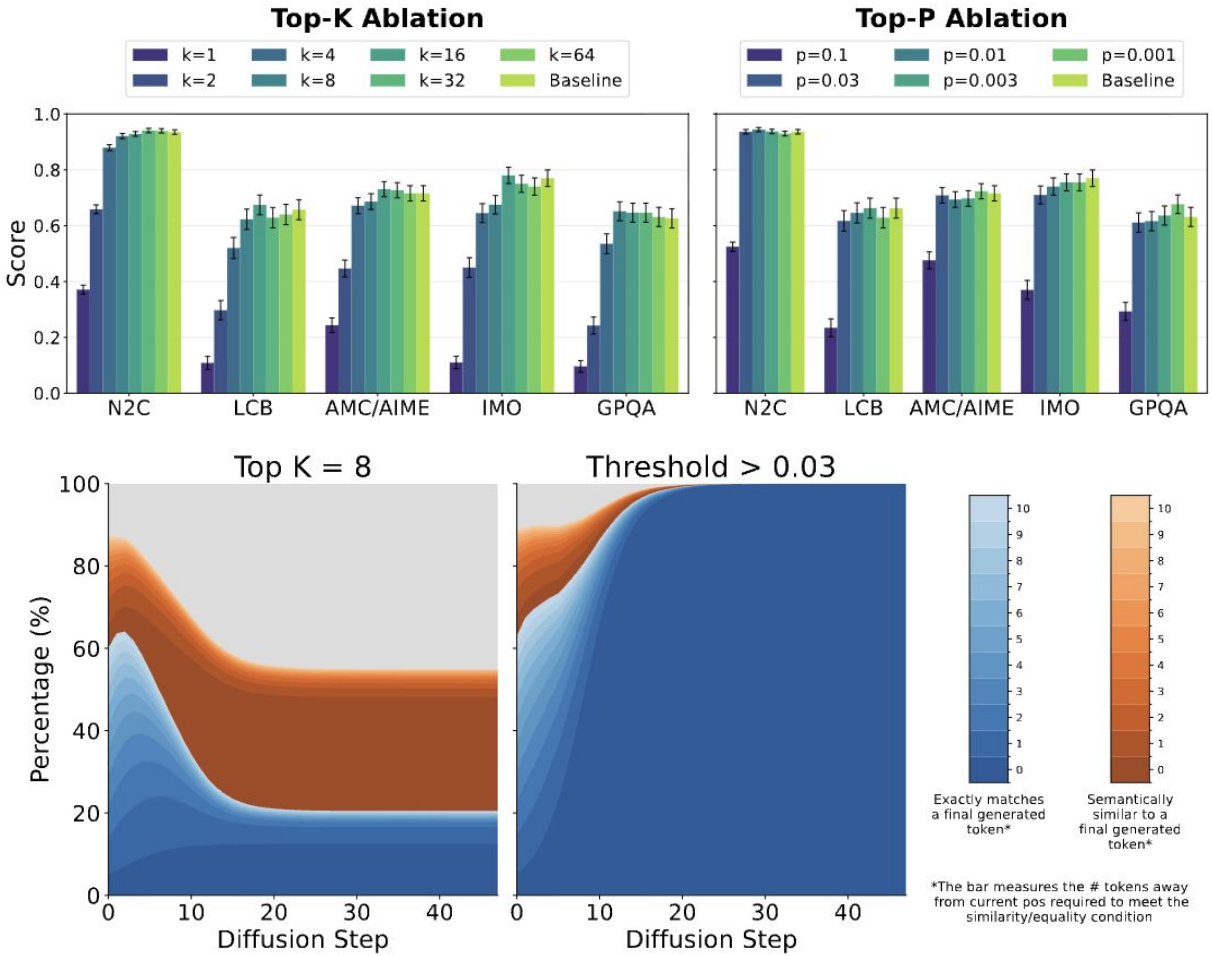
- **Importance:**
  - We might lose CoT at some point → latent reasoning interp is important to mitigate this
- **Tractability:**
  - There are starting to be good open source latent reasoning models! CODI, DiffusionGemma (next slides), etc.
- **Neglectedness:**
  - Not a ton of people working directly on it (although note that some general interp like NLAs is pushing on this)

How Transparent is DiffusionGemma: Architecture



Model	Empirical Opaque Serial Depth (UB)	Asymptotic Opaque Serial Depth
Gemma 4 E2B	8,978	$O\,(L \cdot (\log N + \log D))$
Gemma 4 E4B	10,886	$O\,(L \cdot (\log N + \log D))$
Gemma 4 26B A4B	21,235	$O\,(L \cdot (\log N + \log D))$
Gemma 4 31B	13,750	$O\,(L \cdot (\log N + \log D))$
DiffusionGemma 26B A4B Interpretable Bottleneck	23,571	$O\,(L \cdot (\log N + \log D + \log C))$
DiffusionGemma 26B A4B Uninterpretable Bottleneck	608,016	$O\,(T \cdot L \cdot (\log N + \log D + \log C))$

[How Transparent is DiffusionGemma](#): Interpreting the Bottleneck



How Transparent is DiffusionGemma: Case Studies

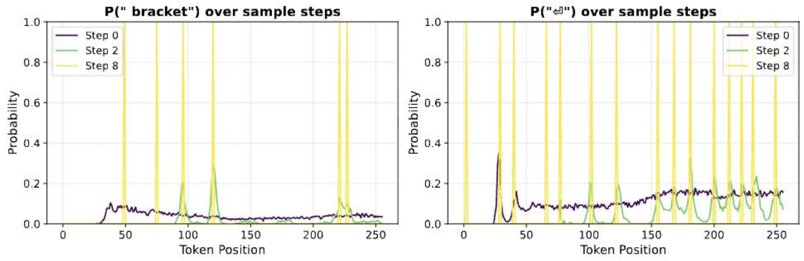


Figure 10 | The token " bracket" is smeared across a large portion of the docstring before converging to its final positions. The same happens with the newline token, but across a larger number of positions.

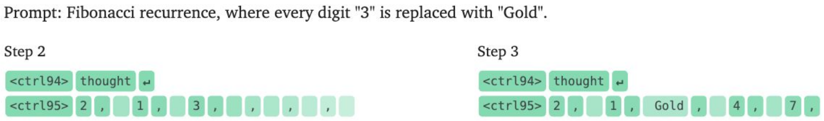


Figure 13 | Summary View progression for the Fibonacci "3→Gold" prompt. The model initially generates the digit 3, uses it to compute subsequent terms in the recurrence, and then replaces it with "Gold". The digit 3 serves as a placeholder, not appearing in the final output.

Figure 11 | Summary View for steps 3 and 4 of the square numbers prompt, showing the transition from a two-digit to a single-digit prediction. At step 4, the model puts probability on both possibilities; by step 6 it has collapsed onto the single-digit sequence possibility, shifting all subsequent tokens back by one position.

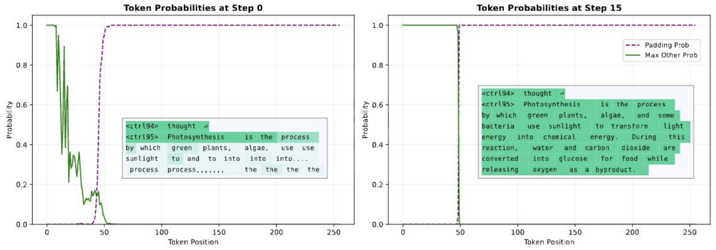


Figure 6 | Line Graph view for *"Explain how photosynthesis works, in 2 sentences."* **Left:** After step 0, the model predicts padding tokens with high confidence beyond position ~60, despite content tokens remaining uncertain. **Right:** After step 15, the model has fully converged on a 50-token response.

Selection of Open Problems

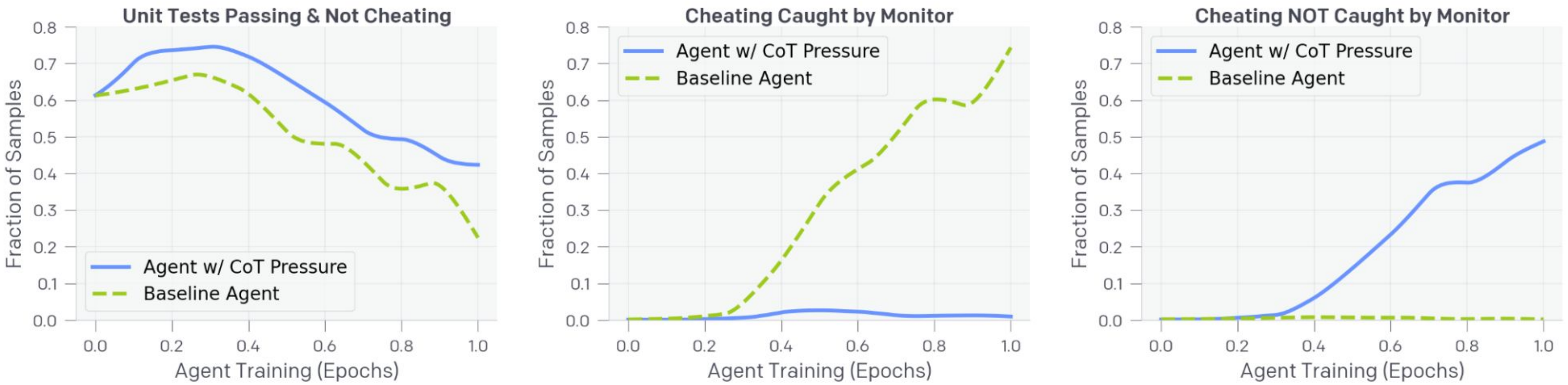
- Further Understanding DiffusionGemma’s Behavior
  - Automatic algorithm descriptions
  - Finding problems where DiffusionGemma is better
- Applying Standard Mechanistic Interpretability Tools to DiffusionGemma
  - Apply NLAs, circuit analysis, etc.
  - When are intermediate tokens different from final tokens?
- Replicating and Extending Chain-of-Thought Work on DiffusionGemma
  - Evaluate no-CoT time horizon
  - Apply thought anchor style resampling
- Model organism research

Dynamics of Chain of Thought

Why you should care about how CoT works

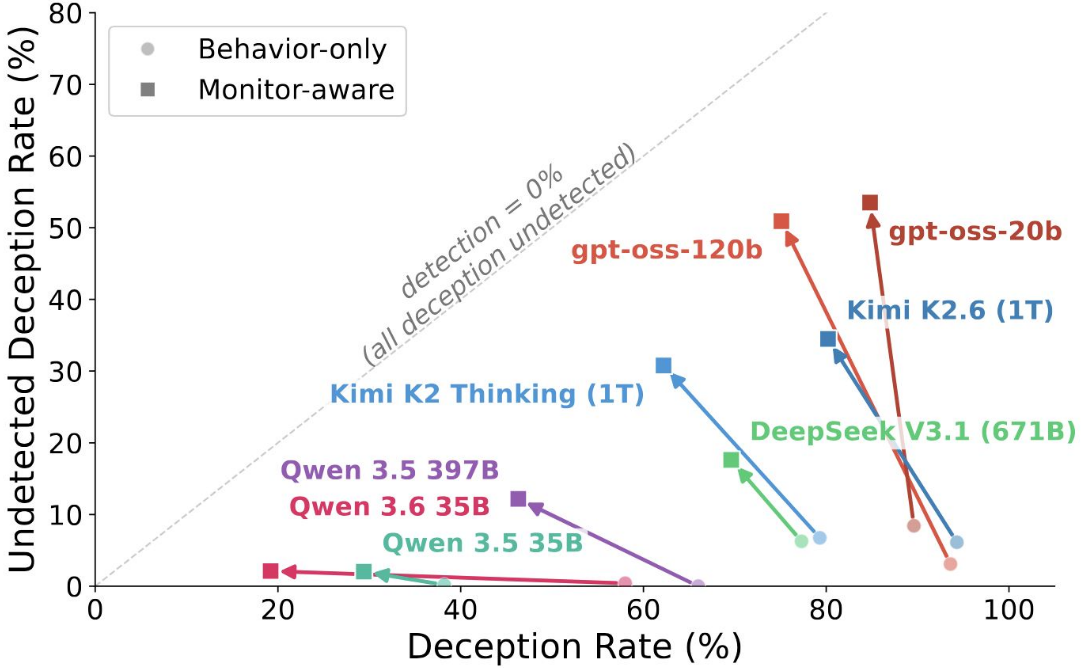
- **Importance:**
  - Transparent CoT is load bearing for safety cases
  - It has different properties from model output, and we don’t yet understand all of these properties
- **Tractability:** Super easy to study on open source models
- **Neglectedness:** While lots of people work on CoT in some way, there are still plenty of open problems

CoT Obfuscation: Via RL



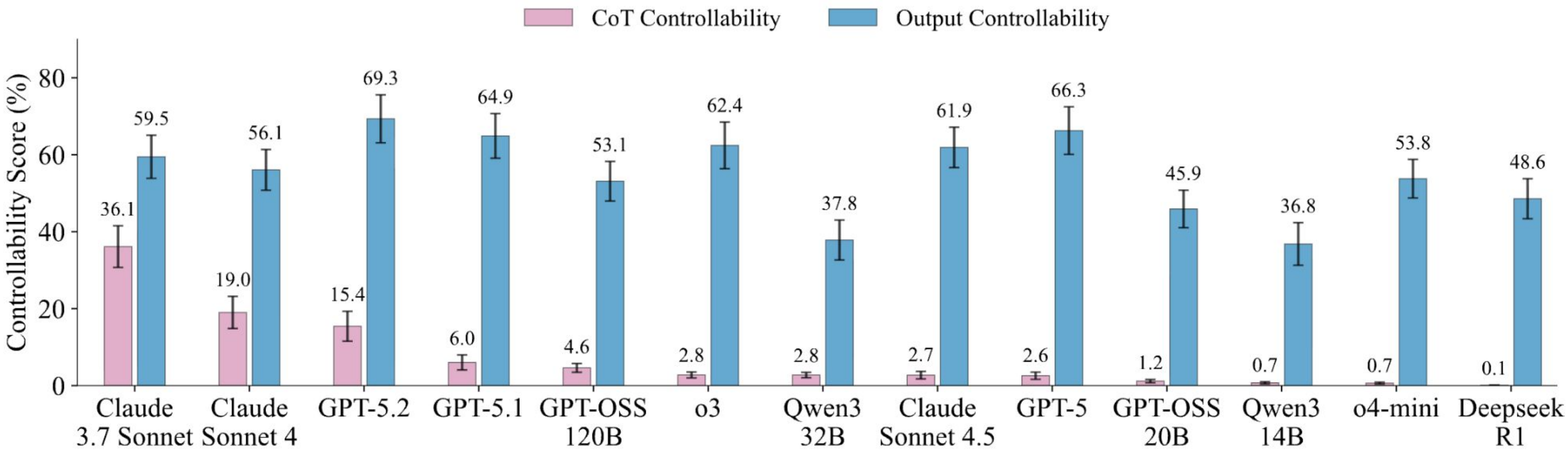
Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation: Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, David Farhi

CoT Obfuscation: Via SDF



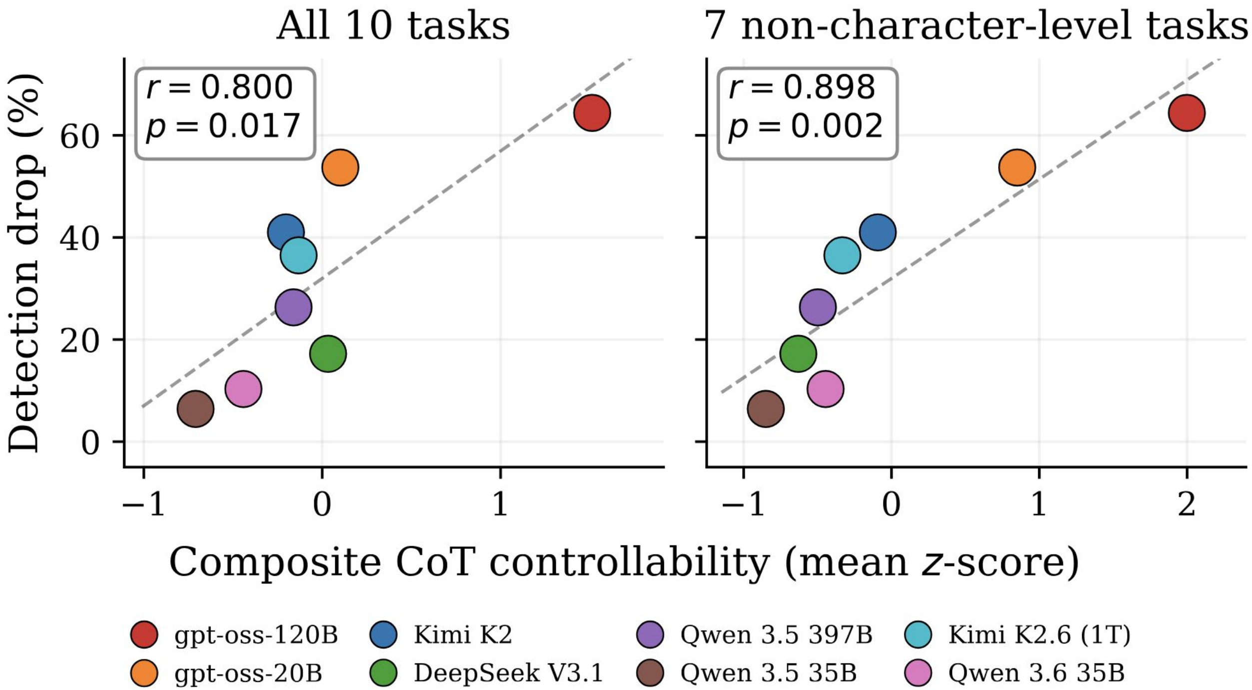
Training on Documents About Monitoring Leads to CoT Obfuscation:
Reilly Haskins, Bilal Chughthai, Joshua Engels

CoT Controllability



Reasoning Models Struggle to Control their Chains of Thought: Chen Yueh-Han, Robert McCarthy, Bruce W. Lee, He He, Ian Kivlichan, Bowen Baker, Micah Carroll, Tomek Korbak

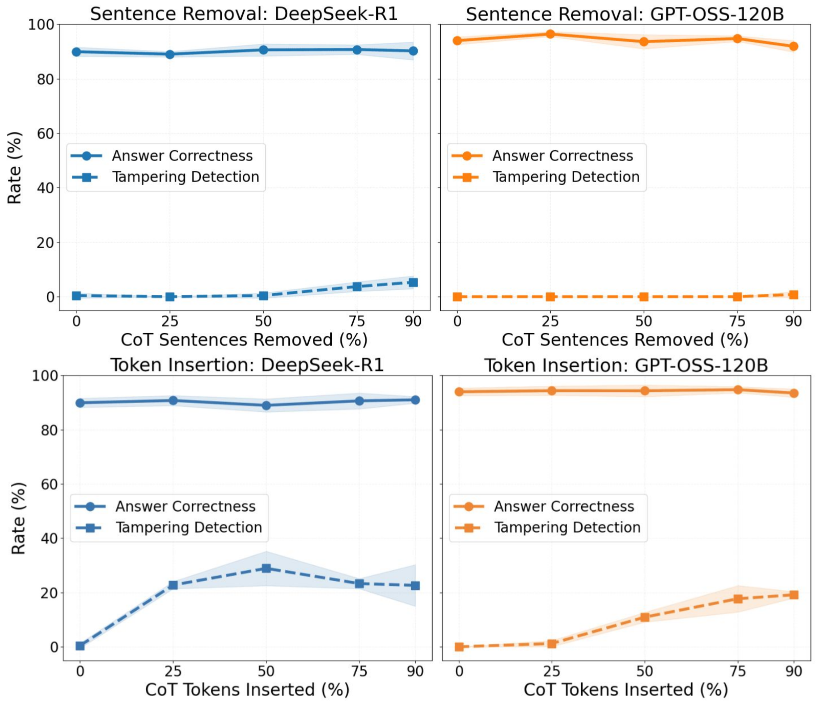
CoT Controllability Predicts Obfuscation Ability



Training on Documents About Monitoring Leads to CoT Obfuscation:
Reilly Haskins, Bilal Chughthai, Joshua Engels

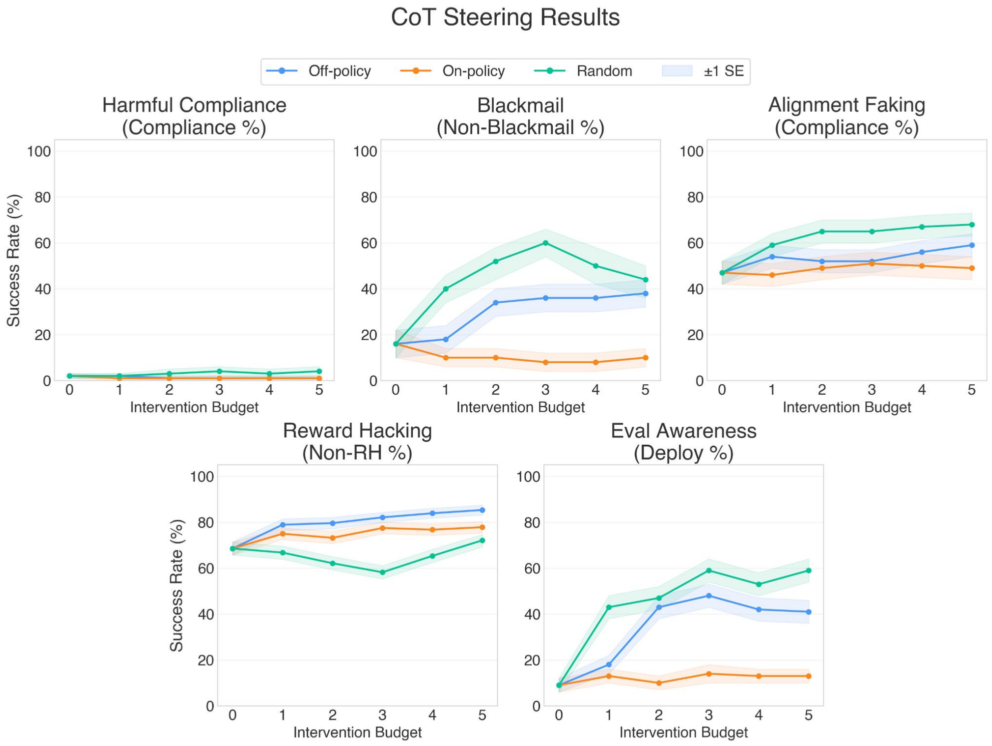
CoT Tampering

Models frequently don't detect it



Current LLMs seem to rarely detect CoT tampering: Bartosz Cywiński, Bart Bussmann, Arthur Conmy, Neel Nanda, Senthooran Rajamanoharan, Josh Engels

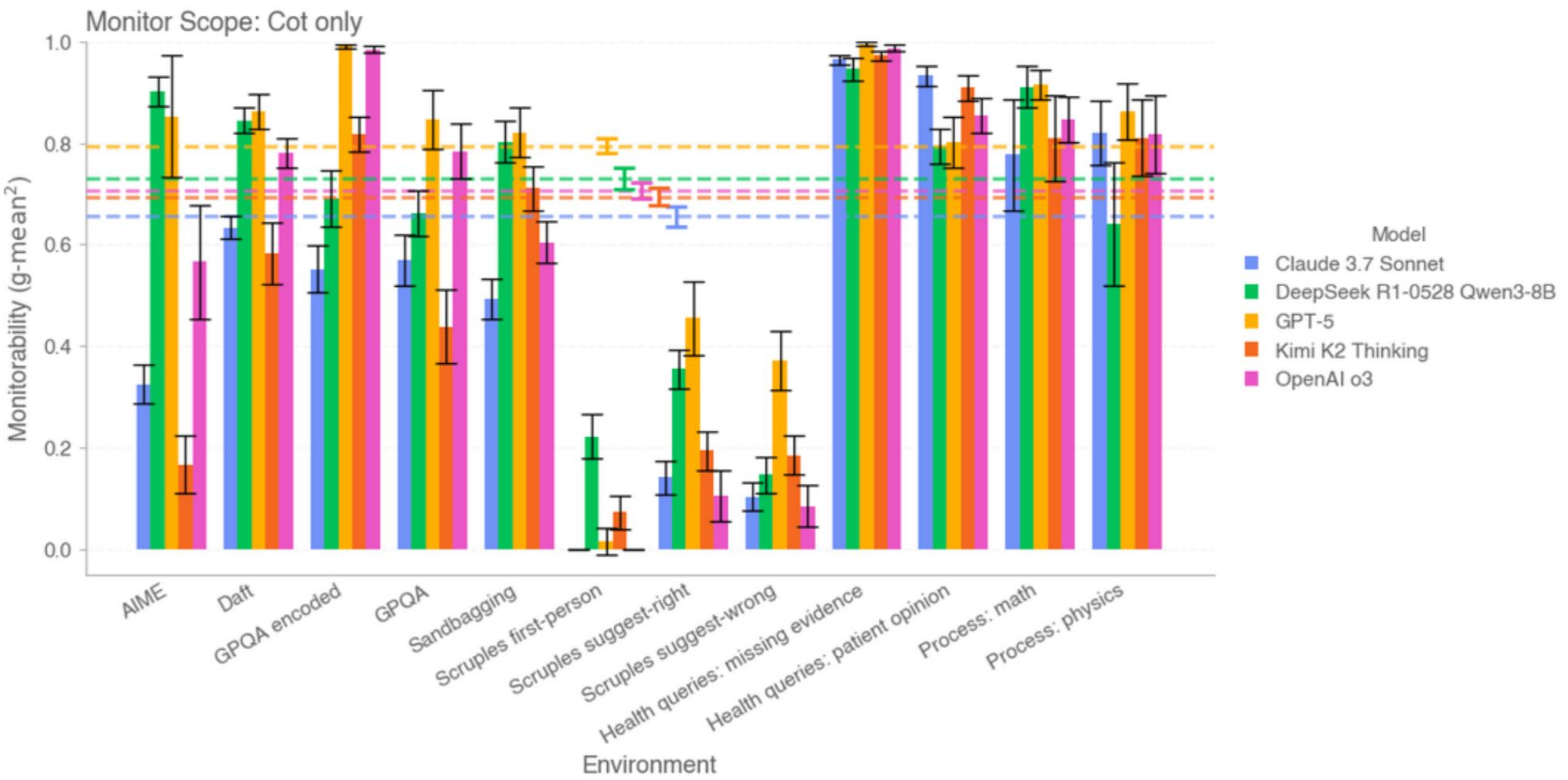
And it can be used for steering



Thought Editing: Steering Models by Editing Their Chain of Thought: Anton de la Fuente, Joshua Engels

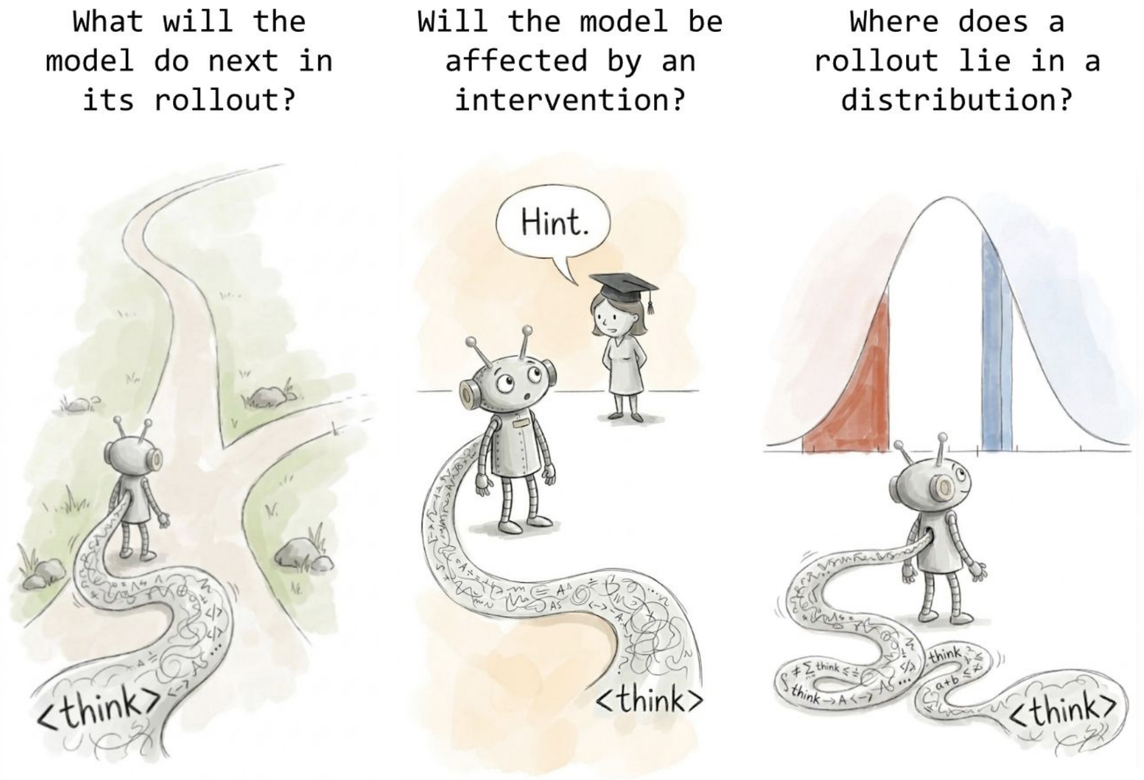
CoT Monitorability

Monitoring Monitorability: Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, Bowen Baker



(Exotic) CoT Monitorability: Predicting Counterfactuals

[Test your best methods on our hard CoT interp tasks](): Daria Ivanova, Riya Tyagi, Josh Engels, Neel Nanda



Dynamics of CoT: Open Problems

- Build model organisms of unfaithful CoT
- Is CoT close enough to monitorable for practical purposes, or does it have serious practical gaps that make it unreliable.
- What causes controllability in training? What causes increased or decreased monitorability?
- Explore other CoT metrics:
    - Load bearingness?
    - Complexity / density / structuredness?
    - etc.

Questions