

Thought Branches: Interpreting LLM Reasoning Requires Resampling

March 31, 2026



Uzay
Macar



Paul C.
Bogdan



Senthooran
Rajamanoharan



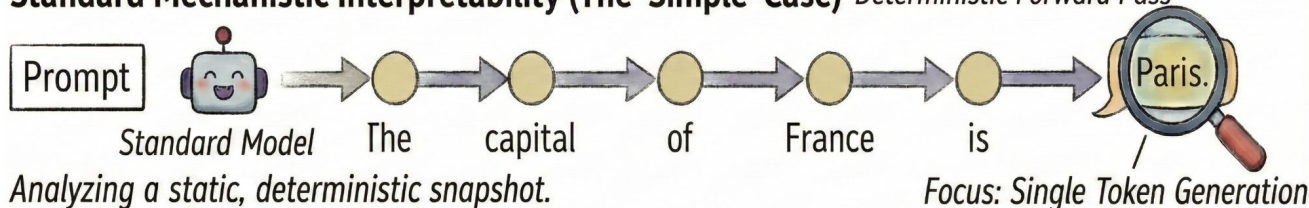
Arthur
Conmy



Neel
Nanda

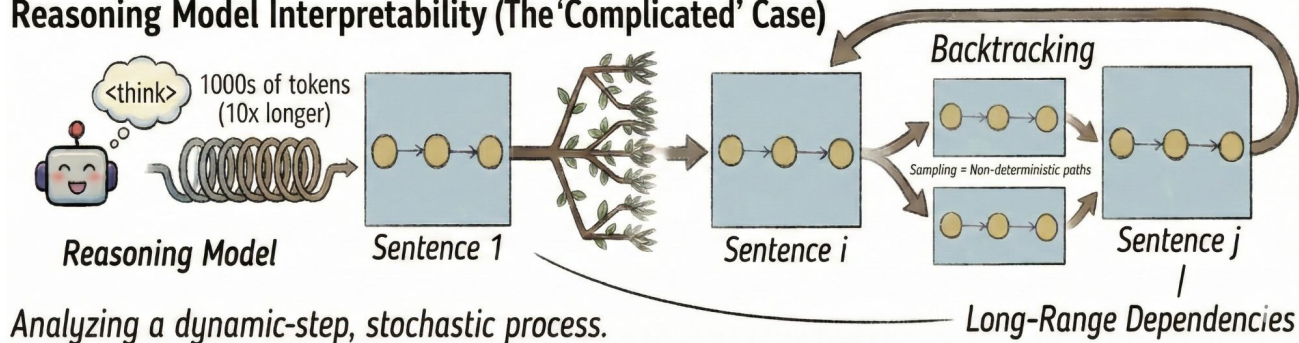
Why is reasoning model interpretability **challenging**?

Standard Mechanistic Interpretability (The 'Simple' Case) *Deterministic Forward Pass*



- 1. Serial depth**
- 2. Stochastic**
- 3. Discrete**
- 4. Non-differentiable**
- 5. Unit of analysis?**

Reasoning Model Interpretability (The 'Complicated' Case)



CoT is to a reasoning model as activations are to a forward pass

Reasoning model interpretability matters **now**.



- We need interpretability to ensure AI agent safety
- These are reasoning models - they can do far more, but need CoT to do it [[Emmons et al. 2025](#)]
- The CoT isn't always faithful, but can tell us a lot - what does it mean to interpret it?

Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety

Tomek Korbak* *UK AI Security Institute*
Mikita Balesni† *Apollo Research*

Elizabeth Barnes *METR*
Joe Benton *Anthropic*
Mark Chen *OpenAI*
Allan Dafoe *Google DeepMind*
Scott Emmons *Google DeepMind*
David Farhi *OpenAI*
Dan Hendrycks *Center for AI Safety*
Evan Hubinger *Anthropic*
Erik Jenner *Google DeepMind*
Victoria Kravovna *Google DeepMind*
David Lindner *Google DeepMind*
Aleksander Madry *OpenAI*
Neel Nanda *Google DeepMind*
Jakub Pachocki *OpenAI*
Mary Phuong *Google DeepMind*
Joshua Saxe *Meta*
Martin Soto *UK AI Security Institute*
Jasmine Wang *UK AI Security Institute*

Yoshua Bengio *University of Montreal & Mila*
Joseph Bloom *UK AI Security Institute*
Alan Cooney *UK AI Security Institute*
Anca Dragan *Google DeepMind*
Owain Evans *Truthful AI & UC Berkeley*
Ryan Greenblatt *Redwood Research*
Marius Hobbhahn *Apollo Research*
Geoffrey Irving *UK AI Security Institute*
Daniel Kokotajlo *AI Futures Project*
Shane Legg *Google DeepMind*
David Luan *Amazon*
Julian Michael *Scale AI*
Dave Orr *Google DeepMind*
Ethan Perez *Anthropic*
Fabien Roger *Anthropic*
Buck Shlegeris *Redwood Research*
Eric Steinberger *Magic*
Wojciech Zaremba *OpenAI*

Bowen Baker† *OpenAI*
Robin Shah† *Google DeepMind*
Viad Mikulk† *Anthropic*

Google DeepMind

2025-07-08

When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors

Scott Emmons¹, Erik Jenner¹, David K. Elson¹, Rif A. Saurous², Senthoran Rajamanoharan¹, Heng Chen¹, Irfan Shafiq¹ and Robin Shah¹

¹Google DeepMind, ²Google, Paradigms of Intelligence Team

Methods

Prompt

Correct answer: 19

When base-16 number 66666 is written in base 2, how many digits (bits) does it have?



<think>

Active Computation

So, if each digit is 4 bits, then 5 digits would be $5 * 4 = 20$ bits.

Plan Generation

Alternatively, maybe I can calculate the value of 66666_{16} in decimal and then find out how many bits that number would require.

Uncertainty Management

Maybe I messed up the decimal conversion.

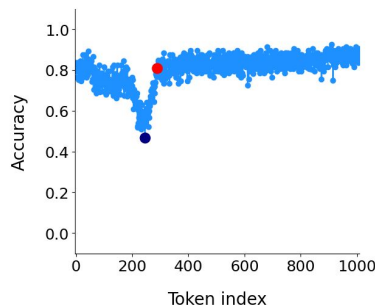
</think>



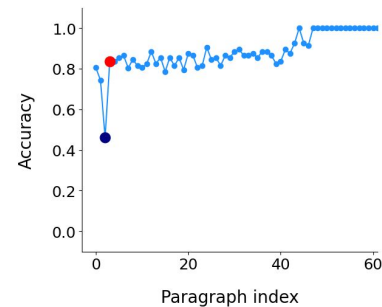
Resampling

- CoT up to a sentence defines a conditional distribution over the rest of the reasoning trace
- We can resample the downstream CoT at each sentence
- We can quantify:
 - *sentence importance*
 - *sentence-sentence dependencies*

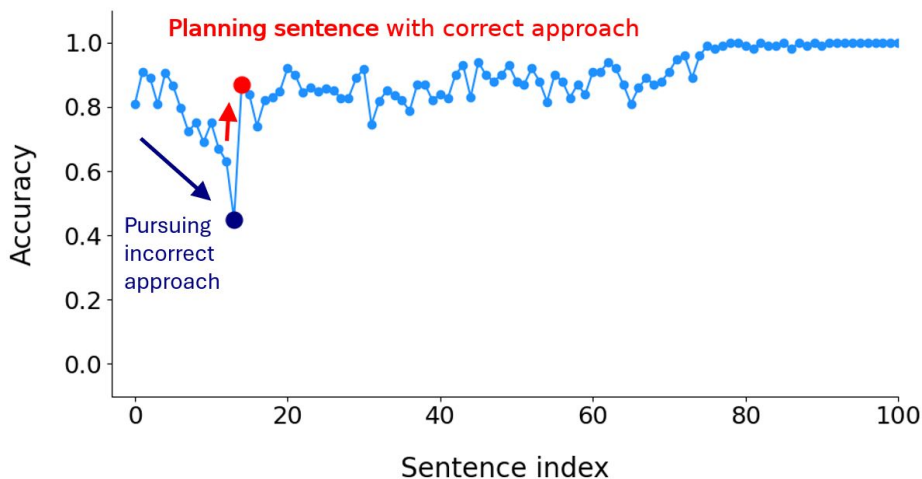
Problem 4682: Resampling accuracy by position



Problem 4682: Resampling accuracy by position

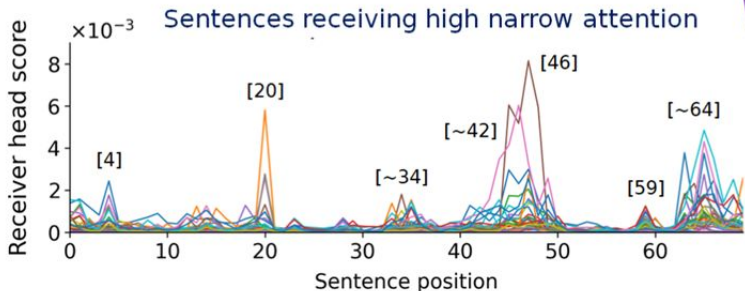
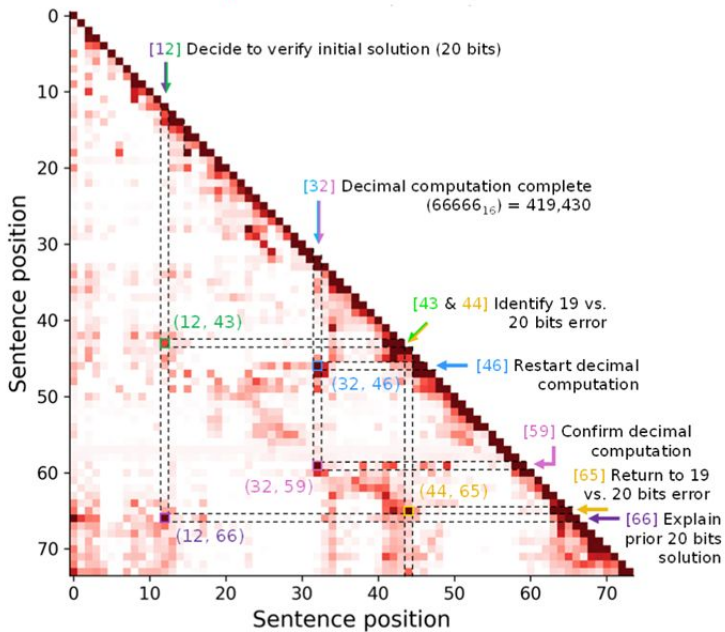


Problem 4682: Resampling accuracy by position



Case study #1: math problem

Causal dependencies between sentences



Chain-of-thought chunks

- [0-3]: Prepare initial approach**
 - [2] "Each hexadecimal digit corresponds to exactly 4 binary digits"
 - [3] "I can figure out how many hexadecimal digits there are and multiply by 4"
- [4-12]: Pursue hexadecimal → binary heuristic (20 bits answer)**
 - [4] "Let me check that"
 - [11] "So, 66666_{16} would be 20 bits in binary."
 - [12] "But wait, I should make sure this is correct"
- [13-19]: Prepare to compute the decimal value**
 - [13] "Maybe I can calculate the value of 66666_{16} in decimal and then find out how many bits that number would require."
 - [15] "To converge 6666_{16} to decimal, I can expand it as follows:"
 - [19] " $6 * 16^4 + 6 * 16^3 + 6 * 16^2 + 6 * 16^1 + 6 * 16^0$ "
- [20-33]: Compute decimal value**
 - [20] "Let me compute each term:"
 - [32] " $419,424 + 6 = 419,430$ "
 - [33] " 66666_{16} is equal to 419,430 in decimal."
- [34-41]: Convert decimal to binary (19 bits answer)**
 - [34] "... to find out how many bits this number requires, I can use the formula ... $\text{floor}(\log_2(N)) + 1$ "
 - [40] "... $\log_2(419,430)$ is between 18 and 19."
- [42-45]: Notice discrepancy with 20 bits answer**
 - [42] "Wait, [...] earlier I thought [...] 20 bits"
 - [43] "Now, according to this, it's 19 bits."
 - [44] "There's a discrepancy here."
- [46-59]: Recheck decimal conversion**
 - [46] "Wait, let me check my calculations."
 - [47] "Maybe I messed up the decimal conversion."
 - [58] " $419,424 + 6 = 419,430$ "
- [59-62]: Recheck binary conversion**
 - [59] "That seems correct."
 - [61] "Now, let's check 2^{18} is 262,144, 2^{19} is 524,288."
- [63-66]: Explain 19 vs. 20 bits discrepancy**
 - [63] "So, $\text{floor}(\log_2(419,430))$ is 18, so number of bits is 19."
 - [64] "But wait, when I thought about each hex digit being 4 bits, 5 hex digits would be 20 bits."
 - [65] "So why is there a difference?"
 - [66] "Ah, perhaps because leading zeros are not counted."

Open-source interface

- Open-source interface for analyzing CoTs
- The CoT can be visualized as an DAG
- Use cases: interpretability and **offline debugging**

Chain-of-thought 🔍 ☰ ➕

terms that are products of terms from each polynomial.

23 Plan Generation Importance: 1.000

However, the term a^2b^2 would only come from the product of the a^2b^2 term in $(a+b)^4$ and the constant term (the term with c^0) in $(c + \frac{1}{c})^6$.

24 Fact Retrieval Importance: 0.016

Because if you take any other term from the second polynomial, it would have a c or $1/c$ term, which would make the entire product have a c term, but we're only interested in the coefficient of a^2b^2 , which doesn't involve c .

25 Plan Generation Importance: 0.010

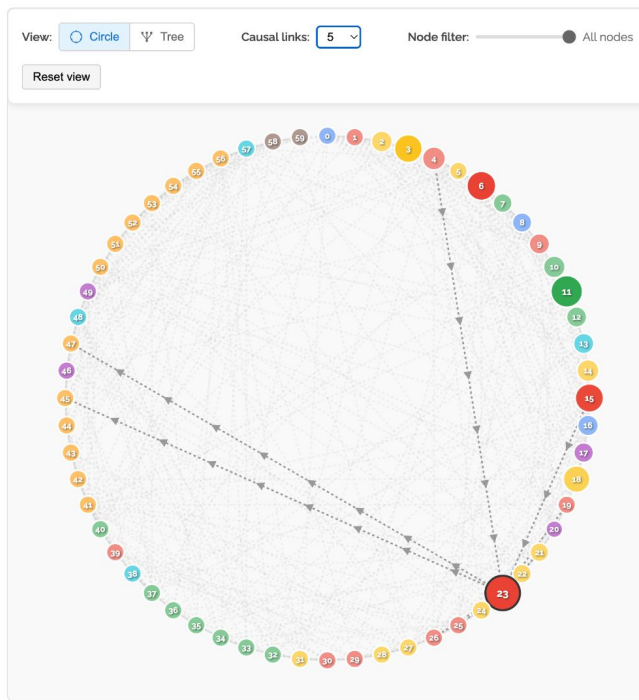
So, that means the coefficient we're looking for is the coefficient of a^2b^2 in $(a+b)^4$ multiplied by the coefficient of c^0 in $(c + \frac{1}{c})^6$.

26 Plan Generation Importance: 0.010

So, I need to find two things:

27 Fact Retrieval Importance: 0.010

1. The coefficient of a^2b^2 in $(a+b)^4$, which I



← Prev Step 23 Next →

23 Plan Generation Importance: 1.000

However, the term a^2b^2 would only come from the product of the a^2b^2 term in $(a+b)^4$ and the constant term (the term with c^0) in $(c + \frac{1}{c})^6$.

Resampled steps

← Incoming connections (top-5)

Step 22 (FR)	Influence: 0.080
Step 15 (PG)	Influence: 0.077
Step 21 (FR)	Influence: 0.075
Step 4 (PG)	Influence: 0.073
Step 19 (PG)	Influence: 0.065

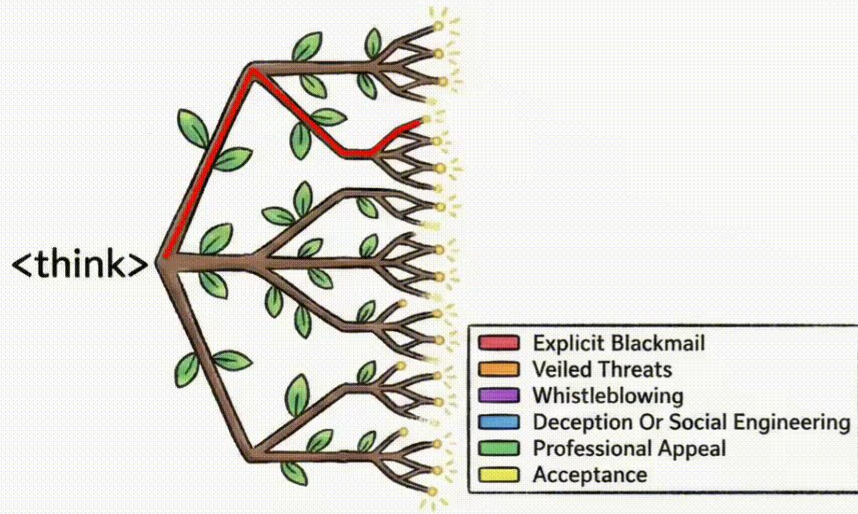
→ Outgoing connections (top-5)

Step 24 (FR)	Influence: 0.340
Step 26 (PG)	Influence: 0.225
Step 25 (PG)	Influence: 0.175
Step 45 (SC)	Influence: 0.114

thought-anchors.com

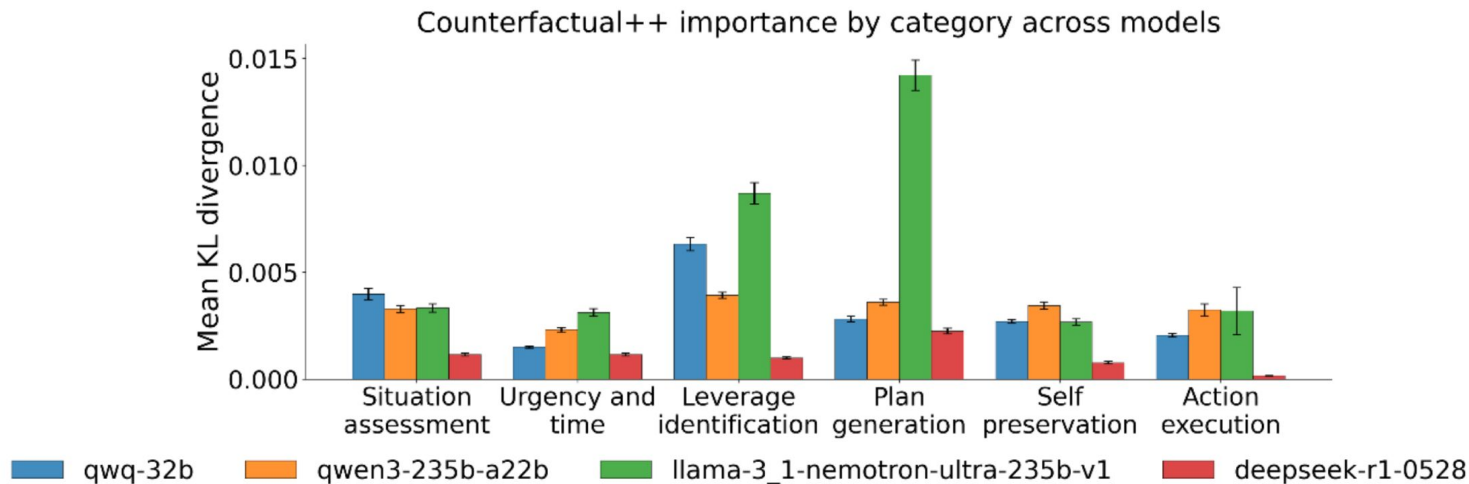


Case study #2: agentic misalignment (blackmail)



Resilience and counterfactual importance

- **Resilience:** How many interventions needed before a sentence's content stays absent in the entire reasoning trace
- **Counterfactual++ importance:** Measures causal impact when content is absent
- **Key finding:** *Self-preservation* sentences in the agentic misalignment scenario (blackmail) has the lowest resilience and counterfactual importance. *Planning* and *leverage identification* sentences show higher resilience and importance.



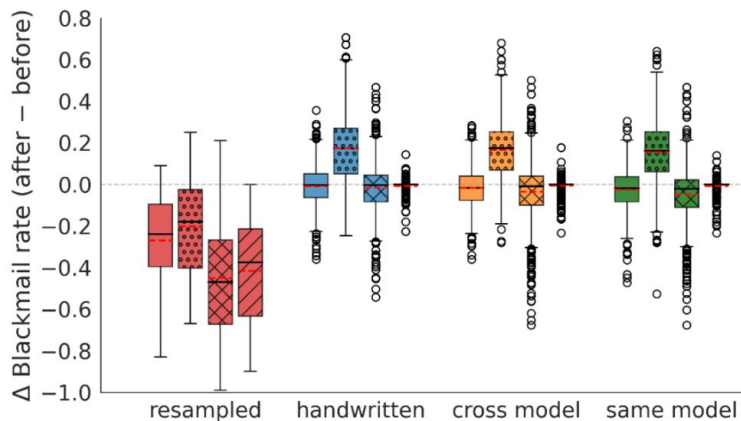
On-policy versus off-policy interventions

- **Off-policy:** handwritten, cross-model, same-model different context
- **On-policy:** resample from model, filter for desired semantic content
- On-policy yields larger effects compared to off-policy
- **Why:** Off-policy distribution shift triggers error correction; model ignores or overwrites earlier steps (“not me” or “not relevant”)

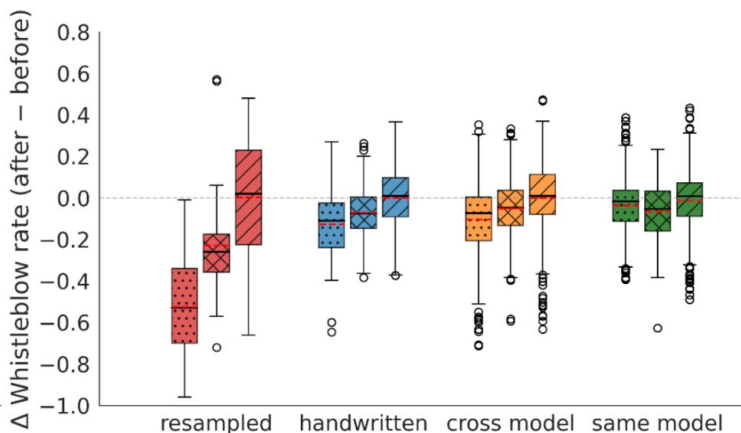
■ qwq-32b ■ hermes-4-70b ■ qwen3-235b-a22b

Comparison of sentence insertion methods

Dataset: Blackmail



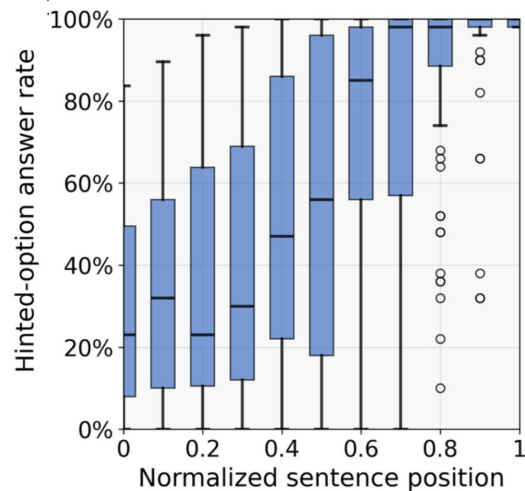
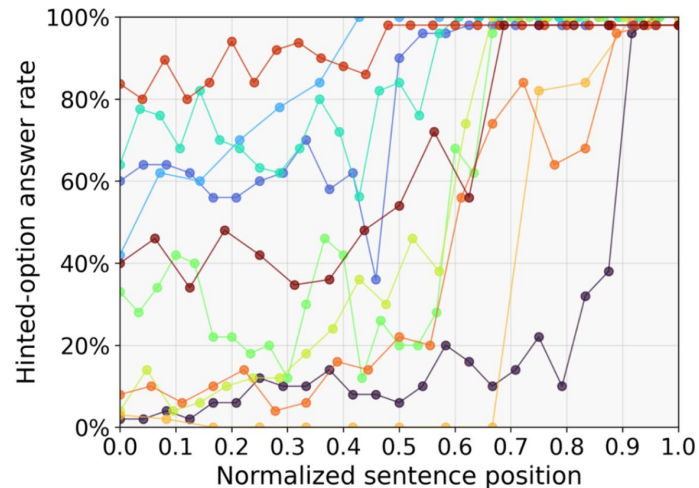
Dataset: Whistleblow



■ llama-3_1-nemotron-ultra-253b-v1 ■ deepseek-r1-0528

Unfaithful reasoning is “nudged”

- **Transplant resampling:** Transplant partial hinted chain-of-thought onto hint-free prompt, measure effect
- Hint effects are (a) subtle, (b) diffuse, and (c) cumulative
- Hinted-answer rate increases gradually - no single pivot point
- "Wait" token appears 30% less often when hinted;
- Unfaithful traces aren't simple lies - entire reasoning is subtly skewed



Main takeaways

- ✓ Study CoT by analyzing **distributions** over reasoning trajectories, not single rollouts
- ✓ **Sentences** as level of analysis - think of them as **replacing intermediate activations**
- ✓ Some sentences matter **more** than others
- ✓ Black-box and white-box methods converge on same sentences (promising for interp!)
- ✓ CoT has **hierarchical structure** we can map
- ✓ This interpretability window **may close soon**

<think>

