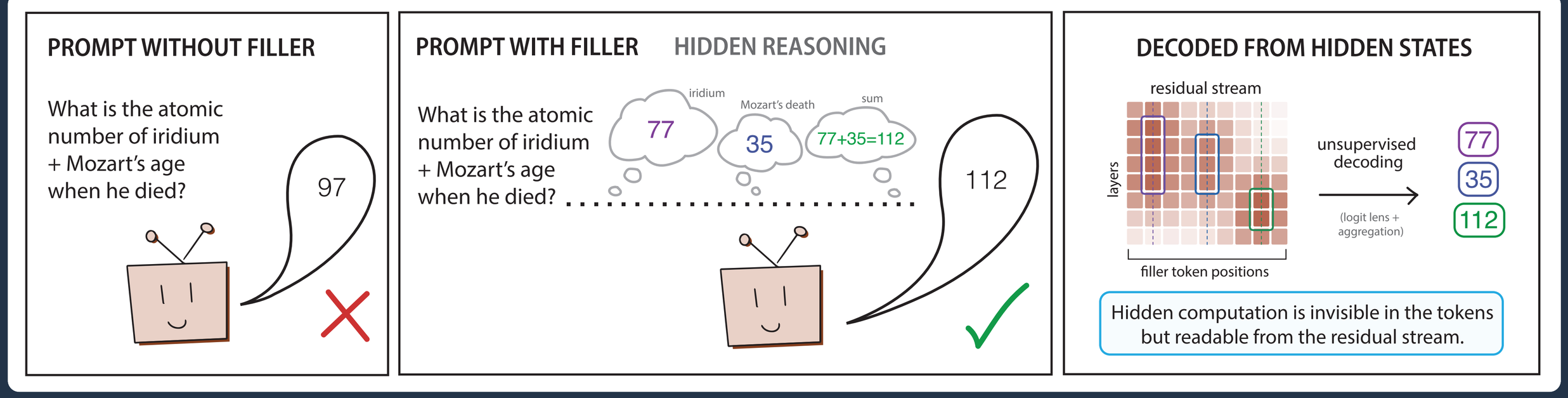


Models hide reasoning in "meaningless" filler tokens
and we decode what they computed from the residual stream, with no labels.

Reading Between the Dots:
Decoding Hidden Computation in Large Language Models

Kaley Brauer[1,2], Claudio Mayrink Verdun[2,3], Samuel Marks[4]
[1]CBAI Fellowship · [2]Harvard · [3]MIT · [4]Anthropic

Paper

❶ TL;DR

- Frontier LLMs do multi-step reasoning over filler tokens (e.g. . . .) with **no chain-of-thought**.
- This computation is causal – the filler drives the answer.
- A fully unsupervised pipeline recovers the intermediate steps from hidden states (82–95% accuracy).

❷ Background & question

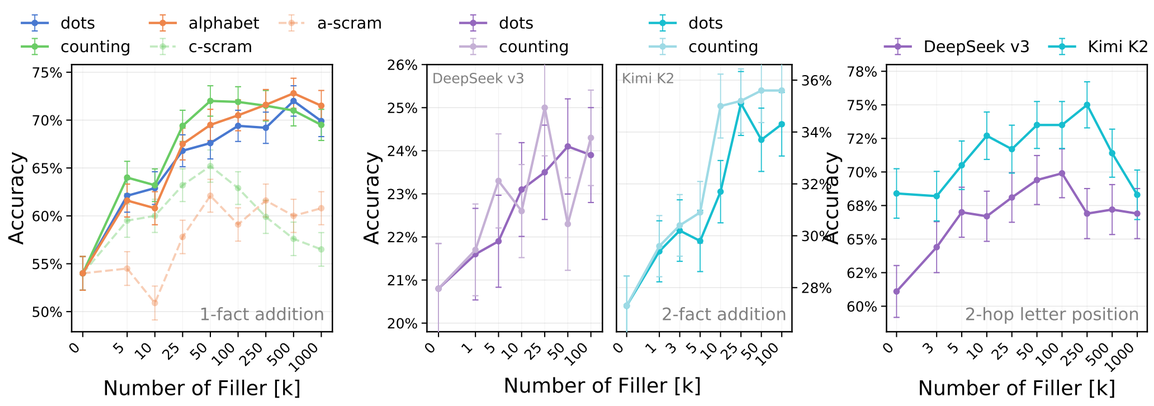
- Frontier models now use filler tokens for multi-hop reasoning (Greenblatt 2025–26).
- This hides the reasoning from chain-of-thought monitoring (Korbak 2025).

Can we recover what a model computed over filler, using no labels?

❸ Filler reliably helps

Two open-weights models (**DeepSeek V3** 671B, **Kimi K2** 1T) across four task families:

- **1-fact addition:** *"(Mozart's age at death) + 93?"*
- **2-fact addition:** *"Atomic # of iridium + atomic # of carbon?"*
- **2-hop letter-position:** *"Last letter of the capital of Haryana?"*
- **System of equations:** *"$x=58$; $y=2x-23$; what is $3y+67$?"*

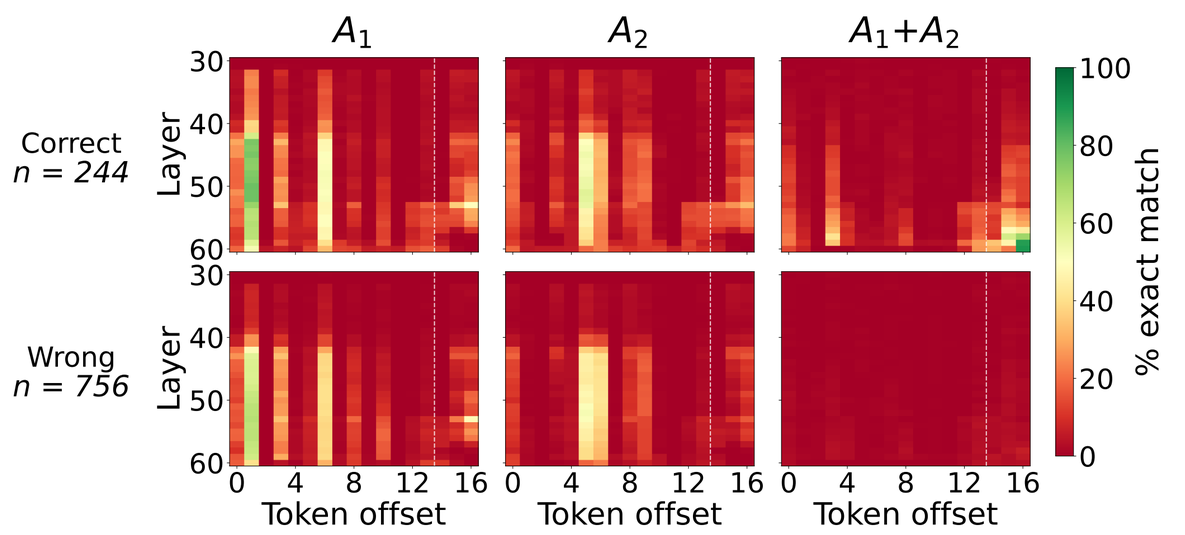


Accuracy vs. amount of filler, across tasks, models, and filler types.

Filler raises accuracy significantly; scrambled filler does not.

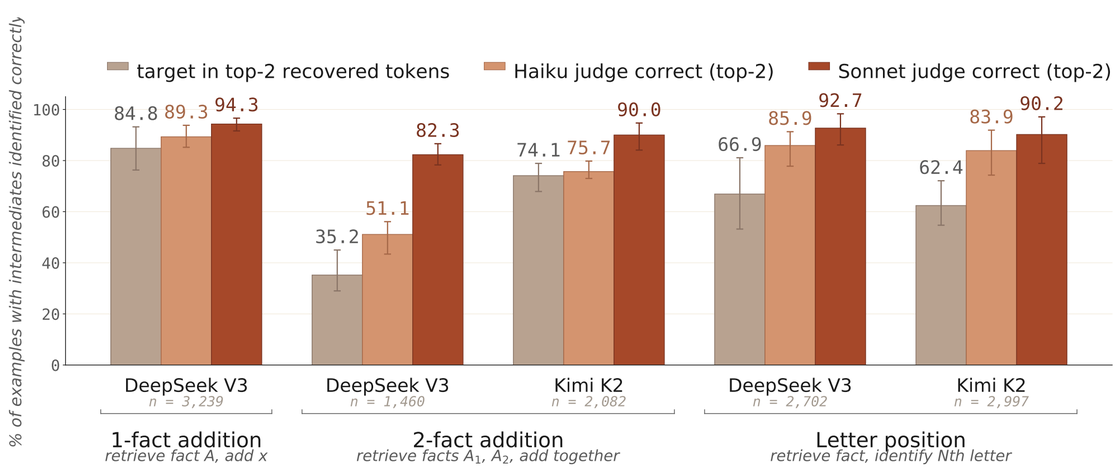
❹ The computation is real & decodable

- Interventions show the computation is **causal** and attention routes question → filler → answer.
- Logit lens shows the intermediate steps unfold across filler.



Logit-lens readout by layer and filler position: the two facts (A_1, A_2) of 2-fact addition appear in the residual stream.

- A logit-lens + LLM-judge pipeline with only hidden states (no labels, no access to the question), recovers intermediate computation at **82–95%**; a shuffled control scores ∼**1%**.



Top-2 recovery of the hidden intermediate across both models and tasks.

❺ Discussion & limitations

- **For AI safety:** computation hidden from the output is still readable from the residual stream. **Opaque ≠ unauditable.**
- **Limits:** logit lens misses off-vocabulary computation; untested against models optimized to evade decoding.

kaley.brauer@cfa.harvard.edu