

# Observable Patterns Are Not Explanations: A Causal-Geometric Analysis of Latent Reasoning Models

Darpan Aswal<sup>1,2</sup> Thomas Palmeira Ferraz<sup>1,3</sup> Yongxin Zhou<sup>1</sup> Maxime Peyrard<sup>1</sup>

<sup>1</sup>Université Grenoble Alpes, CNRS, Grenoble INP, LIG <sup>2</sup>Université Paris-Saclay <sup>3</sup>NAVER LABS Europe darpan.aswal@univ-grenoble-alpes.fr



## 1 Motivation

- Latent reasoning models (LRMs) think in **continuous latent thoughts**, not text: **no visible chain-of-thought** to monitor.
- Prior work reads **observable latent patterns** as evidence of reasoning:
  - BFS-like frontiers** in Coconut [1]
  - a **latent scratchpad** in CODI [2]

But **decodable**  $\neq$  **causal**. We test:

| Emergence        | Generality      | Causality     |
|------------------|-----------------|---------------|
| matched controls | 2 tasks, 2 LRMs | interventions |

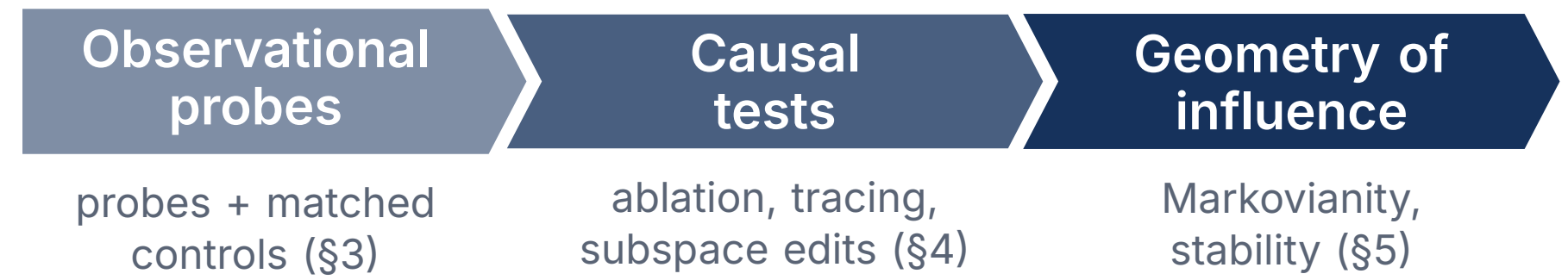
- RQ1** Do observable latent patterns **uniquely emerge** from latent reasoning and **explain behavior**? §3
- RQ2** When do latent thoughts influence predictions, and **where** is that influence concentrated? §4
- RQ3** Do **behavior-influential** thoughts evolve differently across steps? §5

## 2 Setup & Methods

**Tasks** graph-hopping (ProsQA), arithmetic (GSM8K)

**Backbone** GPT-2 small, one model per task

- C** LRM. Coconut [1]: recurrence + staged curriculum
- CODI** LRM. CODI [2]: recurrence + self-distillation
- PaT** **Recurrence control**. Coconut's curriculum, pause tokens instead of recurrence [3]
- C<sub>u</sub>** **Curriculum control**. Coconut with mixed stages ( $u=0.3$ )
- B** **Observational control**. Base GPT-2
- CoT** **Observational control**. Explicit-CoT GPT-2



- Probes**: depth-frontier [1] and logit lens [2] on thought positions.
- Causal tracing [4]**: patch clean activations into a corrupted run;  $IE = 1 - KL_{patch}/KL_{corr}$ .
- Gradient subspace**:  $B_t$  = top singular vectors of loss gradients;  $h_t \leftarrow h_t + (\alpha - 1)B_tB_t^\top h_t$ , vs. a random basis.
- Markovianity**: predict  $h_t$  from earlier thoughts; test  $R^2$ .
- Stability**:  $s_t = \text{mean } \sigma(B_t^\top B_{t+1})^2$ .

95% bootstrap CIs; McNemar tests for paired comparisons. Llama-3.2-1B results in the paper.

## 6 Takeaways

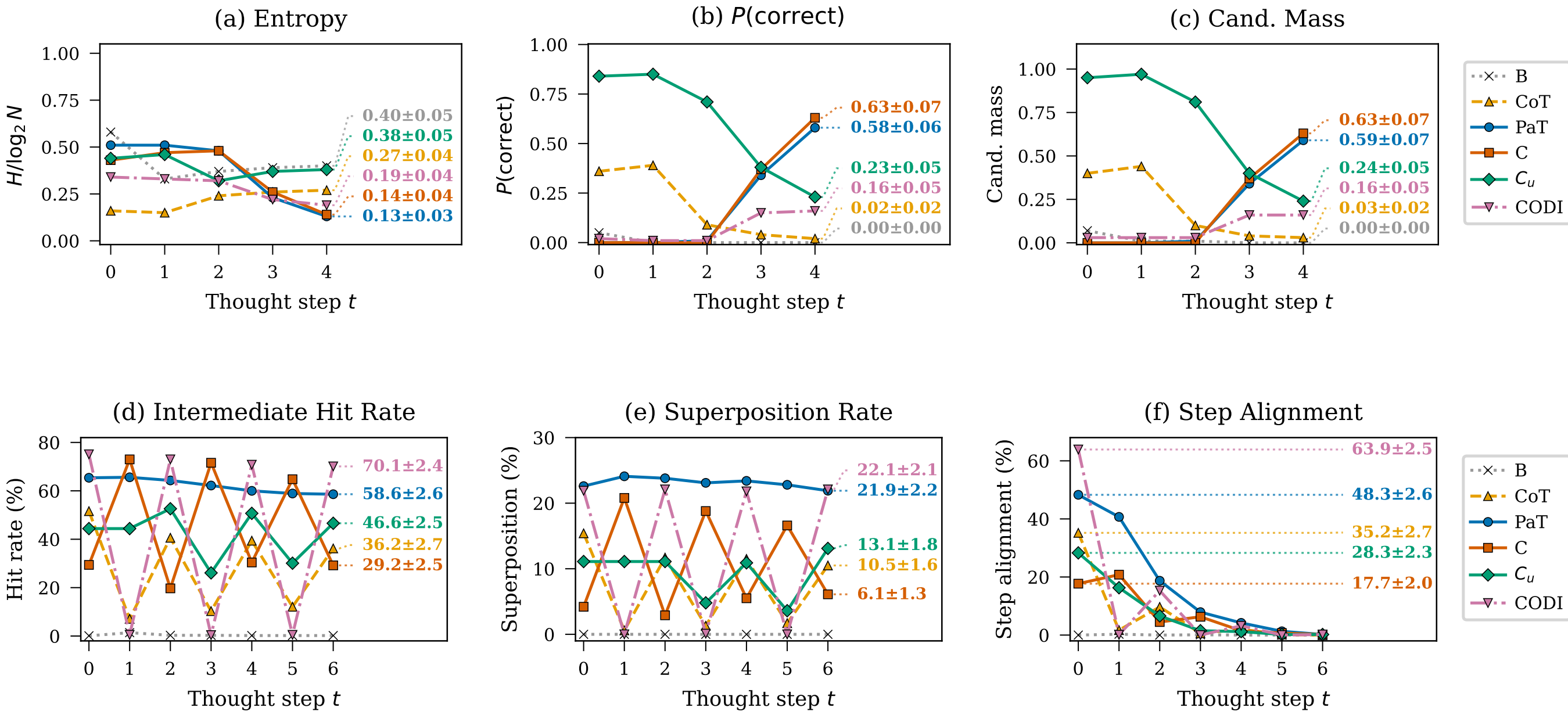
- Observable patterns also appear in matched controls.
- Mechanism claims need matched controls *and* causal tests.
- Thought use is graded and concentrates in low-dimensional subspaces.
- Study causally influential directions, not full latent states.

**Latent thoughts are hidden computation, not hidden explanation.**

[1] Hao et al., COLM 2025 (Coconut). [2] Shen et al., EMNLP 2025 (CODI). [3] Goyal et al., ICLR 2024 (pause tokens). [4] Meng et al., NeurIPS 2022 (causal tracing).

## 3 Observable Patterns Are Not Enough

RQ1



**Graph-hopping: BFS [1]**

- PaT** matches **C**'s BFS signature **without recurrence**
- C<sub>u</sub>** inverts it, recurrence intact
- No superposition**: mass sits on target nodes

**Arithmetic: scratchpad [2]**

- PaT** beats **CODI**'s hit rate **without recurrence**
- C**, **C<sub>u</sub>**, even **CoT** alternate like **CODI**
- Step alignment decays: **no ordered scratchpad**

**Takeaway.** Matched controls reproduce or invert these patterns: **not evidence of mechanism**.

## 4 When and Where Do Latent Thoughts Matter?

RQ2

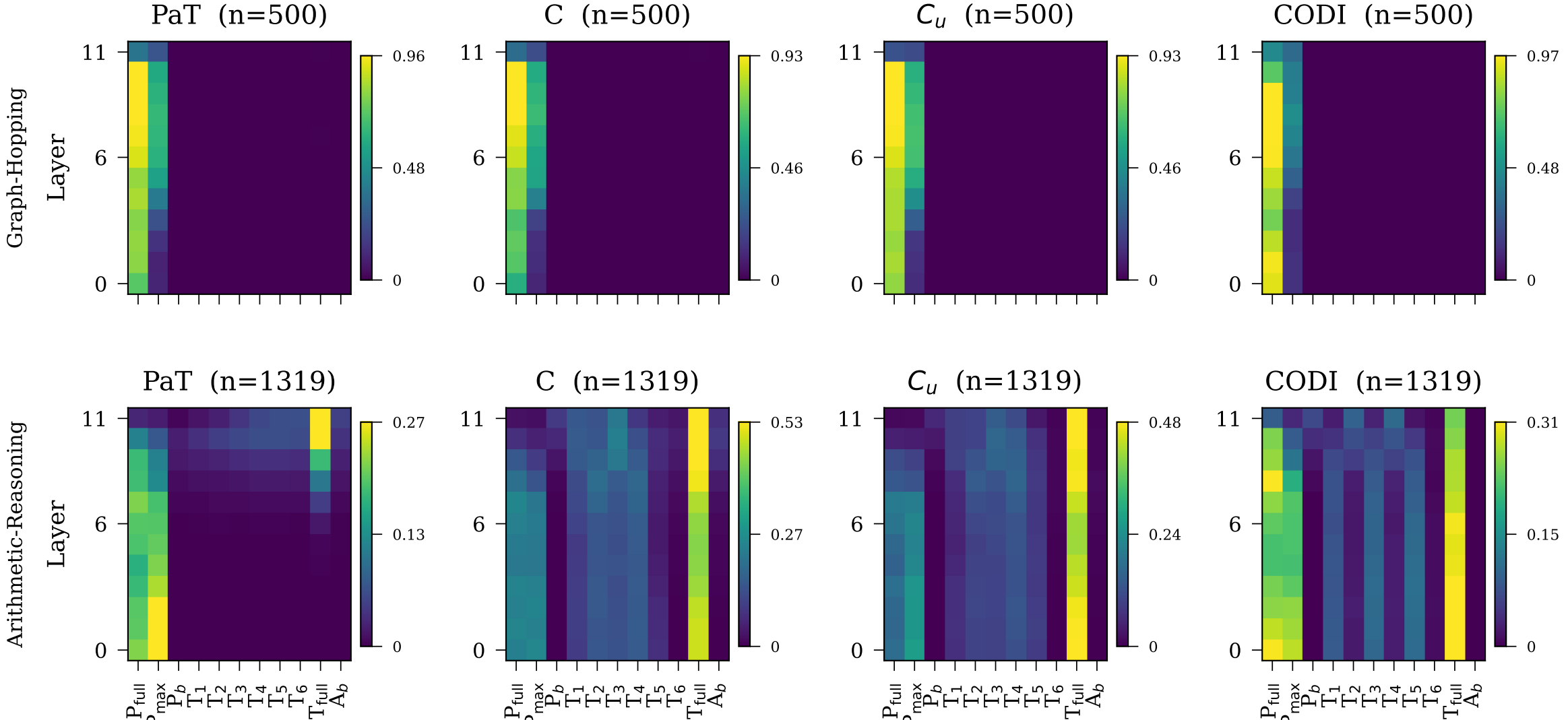
**Test-time thought ablation ( $K \rightarrow 0$ )**

|                | Graph-hopping  |                | Arithmetic     |                |
|----------------|----------------|----------------|----------------|----------------|
| Model          | Full           | Removed        | Full           | Removed        |
| B              | 2.4 $\pm$ 1.3  | —              | 1.4 $\pm$ 0.6  | —              |
| CoT            | 79.0 $\pm$ 3.6 | —              | 41.9 $\pm$ 2.7 | —              |
| PaT            | 95.4 $\pm$ 1.8 | 95.6 $\pm$ 1.7 | 26.4 $\pm$ 2.4 | 21.4 $\pm$ 2.2 |
| C              | 98.0 $\pm$ 1.3 | 97.8 $\pm$ 1.3 | 35.7 $\pm$ 2.5 | 7.7 $\pm$ 1.4  |
| C <sub>u</sub> | 96.0 $\pm$ 1.7 | 91.8 $\pm$ 2.4 | 30.8 $\pm$ 2.5 | 39.1 $\pm$ 2.6 |
| CODI           | 80.0 $\pm$ 3.5 | 80.0 $\pm$ 3.5 | 41.8 $\pm$ 2.7 | 25.6 $\pm$ 2.4 |

Table 1. Accuracy (%), all thoughts vs. removed.

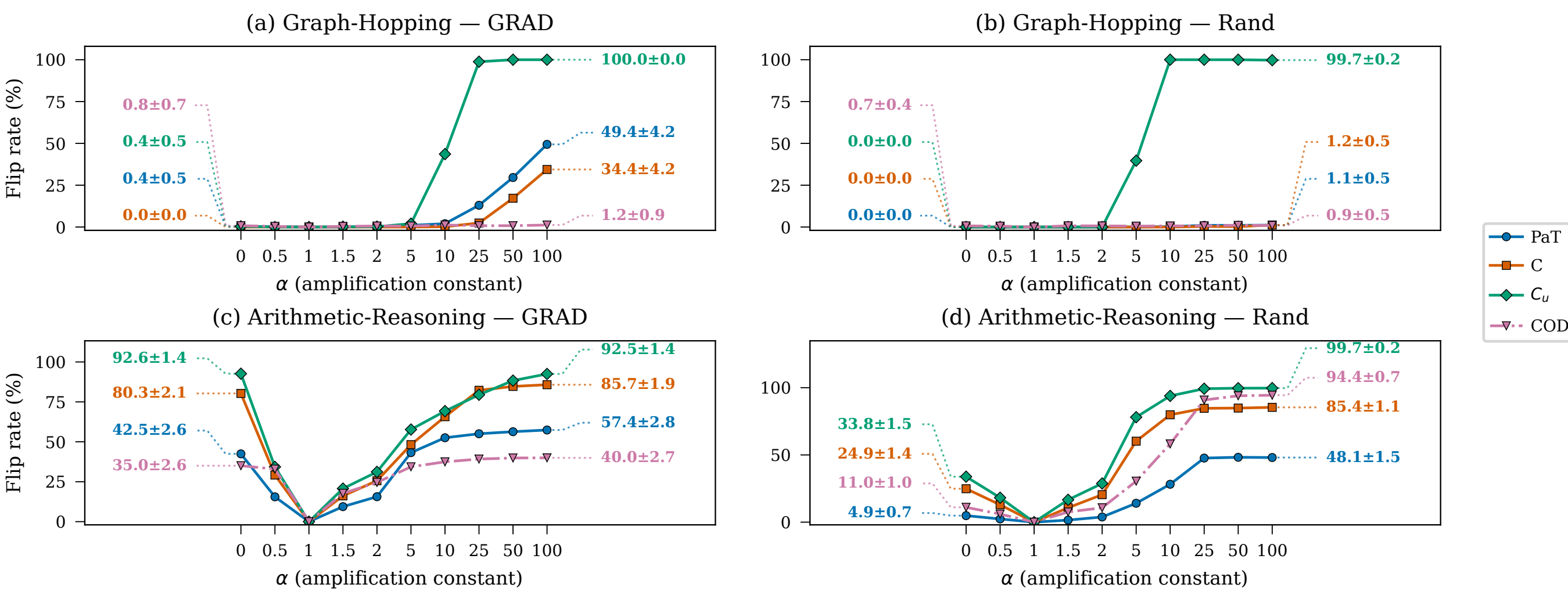
- Graph-hopping**: barely affected.
- Arithmetic**: **C** 35.7  $\rightarrow$  7.7, **CODI** 41.8  $\rightarrow$  25.6, yet **C<sub>u</sub>** *improves*.

**Residual-stream causal tracing**



Per-layer  $IE_{KL}$  for prompt ( $P$ ) and thought ( $T$ ) positions.

- Graph-hopping**: only the prompt recovers behavior.
- Arithmetic**: all thoughts ( $T_{full}$ ) recover best.



**Gradient-subspace edits**

Flip rate when scaling the loss-sensitive (GRAD) or a random subspace by  $\alpha$ .

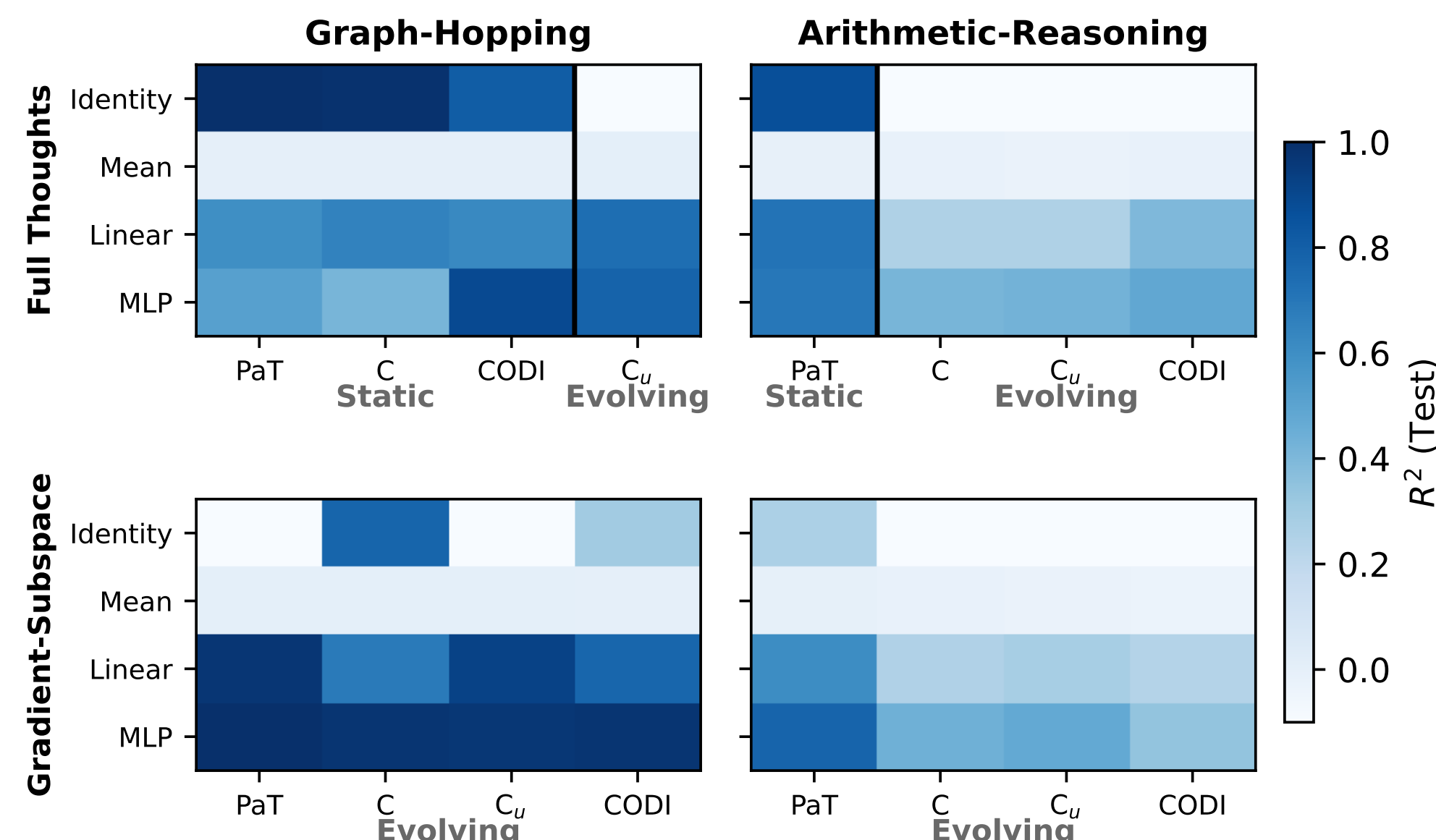
- Graph-hopping**: amplified GRAD flips **PaT** 49%, **C** 34% (random  $\approx$  1%): **not bypassed**.
- Arithmetic**: GRAD ablation flips 35–93% (random 5–34%).

**Takeaway.** Thought use is **graded**, not binary, and often concentrates in a **low-dimensional subspace**.

## 5 Geometry of Influence

RQ3

**Markovianity of thought trajectories**

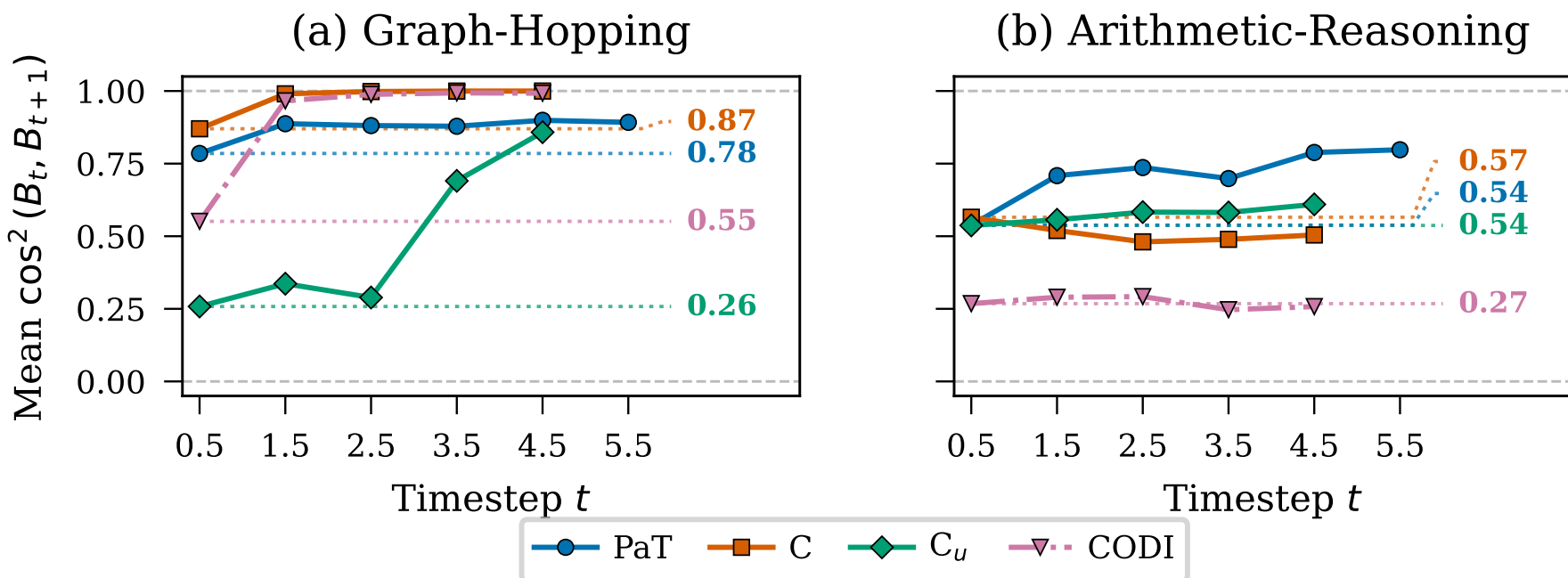


Test  $R^2$  of transition maps: full thoughts vs. gradient subspace.

**Takeaway.** Weakly influential thoughts stay **near-static**; influential ones evolve with **more structure**.

- Full thoughts**: static on graph-hopping (except **C<sub>u</sub>**), evolving on arithmetic.
- Gradient subspace**: every model evolves.

**Stability of the gradient subspace**



- Graph-hopping**: stable; **C<sub>u</sub>** rotates early.
- Arithmetic**: less stable; **CODI** rotates most.