

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

Darpan Aswal^{1,2} Thomas Palmeira Ferraz^{1,3} Yongxin Zhou¹ Maxime Peyrard¹

¹Université Grenoble Alpes, CNRS, Grenoble INP, LIG

²Université Paris-Saclay ³NAVER LABS Europe

darpanaswal@gmail.com

Code:  darpanaswal/InterpretingLatents

Models:  darpanaswal/InterpretingLatents

Abstract

Latent reasoning models (LRMs) replace explicit chain-of-thought with continuous thoughts. Recent work treats observable latent-state patterns, such as BFS-like frontiers and decodable arithmetic computation, as evidence for internal reasoning mechanisms. Evaluating two LRMs (Coconut and CODI) against controls lacking the proposed recurrence or curriculum, we find these patterns also appear in the controls and do not consistently affect behavior. Causal interventions reveal that latent-thought utilization is not binary but graded, scaling with a thought’s causal effect on model behavior. Geometric analyses reveal this effect concentrates in low-dimensional subspaces whose step-wise geometry grows more structured as their behavioral influence increases. Latent thoughts should therefore be treated as hidden computation, not hidden explanation: decodability, attention, or static structure alone cannot establish LRM reasoning mechanisms; that requires matched controls and causal tests.

1 Introduction

Chain-of-thought (CoT) prompting has become a standard approach for eliciting reasoning through verbalized intermediate steps in natural language (Wei et al., 2022; Li et al., 2025; Lyu et al., 2023). However, CoT is computationally costly and fundamentally constrained by the discreteness of natural-language tokens (Zhang et al., 2025b). To address these limitations, latent reasoning models (LRMs) perform intermediate computation in continuous hidden states (Hao et al., 2025; Shen et al., 2025). LRMs promise improved efficiency, but they also remove the primary artifact available for monitoring and oversight: visible CoT traces. Despite well-known faithfulness concerns (Chen et al., 2025; Turpin et al., 2023), CoT traces still provide human-interpretable signals that can be monitored for harmful or misaligned behavior (Bereska and Gavves, 2024; Korbak et al.,

2025). In contrast, LRMs reason through continuous states, raising new safety and interpretability concerns as they scale and are deployed in agentic systems (Chan et al., 2023). This calls for a better understanding of their internal dynamics.

Early interpretability efforts have searched for latent-state analogues of CoT traces, aiming to decode interpretable patterns from latent thoughts. For example, Hao et al. (2025) and Zhu et al. (2025) argue that continuous latent thoughts demonstrate breadth-first-search-like reasoning over multiple superposed candidate reasoning paths, while Shen et al. (2025) and Wei et al. (2026) suggest that implicit supervision can yield interpretable intermediate reasoning steps. Later work has also used local interventions to investigate these findings (Cywinski et al., 2025; Peters et al., 2025). However, several methodological concerns remain, leaving important questions unaddressed. First, *emergence*: whether the observed pattern emerges specifically from latent-reasoning mechanisms, or also appears in matched non-LRM controls. Second, *generalizability*: whether the pattern is characteristic of latent reasoning more broadly, rather than specific to one task, model, curriculum, or training recipe. Third, *causality*: whether the pattern actually contributes to the model’s behavior, rather than merely correlating with successful outcomes. Without addressing these questions, observable patterns may lead to superficial evidence for reasoning.

In this work, instead of treating latent thoughts as hidden explanations, we argue in favor of treating them as extended hidden computation states whose influence strength must be measured by their causal effect on model behavior. Our approach contrasts with prior work also taking a causal perspective by asking binarily whether a latent thought is “used” or “not used” (Zhang et al., 2025a; Cui et al., 2026), shifting the question to where causal influence is concentrated within the latent representation. These causally effective regions then

become the object of analysis: once isolated, their geometry and dynamics can be studied. This framing leads to the following research questions:

RQ1. *Do observable patterns in LRM latent space uniquely emerge from, and explain the performance of, latent-reasoning mechanisms?* In § 4, we compare LRMs against matched controls and find that reasoning patterns previously attributed to latent recurrence can also emerge in curriculum-matched non-recurrent models, and even in untrained models with forced thought positions. This shows that observable patterns alone do not sufficiently establish mechanism.

RQ2. *When do latent thoughts influence model predictions, and where in the latent state does this influence come from?* In § 5, we combine latent thought ablations, causal tracing, and gradient-subspace interventions to show that latent thought influence lies on a continuum. In particular, the effect of latent thoughts is often concentrated in low-dimensional, loss-sensitive directions (a causal gradient subspace) rather than distributed across the full latent representation.

RQ3. *Do behavior-influential latent thoughts differ from weakly influential ones in how their representations evolve across steps?* In § 6, we study the geometry and dynamics of both full latent thought trajectories and causally influential subspaces. We find that weakly influential latent thoughts are often nearly static across steps, while behavior-influential thoughts exhibit more structured evolution.

Overall, our work shifts the target of interpretability from observational proxies to the actual causal trajectories within latent computation. Finally, we discuss practical implications in more detail in § 7.

2 Background and Related Work

2.1 Latent Reasoning Models

Unlike standard Chain-of-Thought (CoT) models that generate textual rationales, latent reasoning models such as implicit CoT (Deng et al., 2023) and recurrent/looped transformers (Giannou et al., 2023; Dehghani et al., 2019) replace intermediate steps with hidden state computations to bypass vocabulary projection. We focus on *autoregressive continuous-thought* models which insert K latent positions between the prompt x and the answer y by recursively feeding projected soft-tokens e_{t+1}^{lat} from hidden-states:

$$h_t = F_\theta(E(x), e_{1:t}^{\text{lat}})_{\text{last}}, \quad e_{t+1}^{\text{lat}} = g(h_t),$$

with F_θ a causal transformer, E the token embedding map, and g identity or learned. We call $h_t \in \mathbb{R}^d$ the *latent thought*. After K latent steps, autoregressive decoding resumes:

$$p_\theta(y \mid x, e_{1:K}^{\text{lat}}) = \prod_m p_\theta(y_m \mid x, e_{1:K}^{\text{lat}}, y_{<m}).$$

We study two instances: **COCONUT** gradually replaces CoT segments with latent states via a staged curriculum (Hao et al., 2025), while **CODI** compresses textual rationales into latents via self-distillation (Shen et al., 2025).

2.2 Observational vs. Causal Interpretability

Mechanistic interpretability seeks to reverse engineer neural network computations into human-understandable algorithms (Bereska and Gavves, 2024; Geiger et al., 2025). A key distinction separates information that is *observable* from activations from information that is *causally used*: probes and visualization methods can reveal correlated structure without establishing mechanism (Belingkov, 2022; Jain and Wallace, 2019; Elazar et al., 2021; Lasri et al., 2022; Teney et al., 2022). This issue is central for LRMs, where claims about latent reasoning often rely on observational readouts, without showing that these structures drive performance. This risks conflating *decodability* with *mechanism*, an analogue of the *dead salmon* effect (Méloux et al., 2026), symptoms of a general tendency of interpretability research to produce false positive findings (Hewitt and Liang, 2019; Ravichander et al., 2021; Kantamneni et al., 2025; Méloux et al., 2026). We rely on intervention-based methods (e.g., causal tracing, activation patching) (Meng et al., 2022; Wang et al., 2023; Chan et al., 2022; Heimersheim and Nanda, 2024; Monea et al., 2024) to test whether latent thoughts affect model behavior¹.

2.3 Mechanistic Analyses of Latent Reasoning

Observable Latent Patterns. Prior work finds structured intermediates in latent states. In CODI, logit-lens, attention-mass, and activation-patching analyses are used to argue that models use latent thoughts as a *computational scratchpad* for reasoning: at each step, these states store operands and intermediate values for subsequent steps (Shen et al., 2025; Cywinski et al., 2025; Peters et al., 2025); however, the evidence remains largely single-task

¹while recognizing that causal interventions can still admit multiple compatible explanations (Méloux et al., 2025).

and observational. [Hao et al. \(2025\)](#) interpret Coconut’s latent thoughts as superposed reasoning paths with BFS-like dynamics, representing multiple candidate nodes at increasing graph depths, progressively concentrating mass on target-reaching nodes. Later work attributes similar patterns to shortcut behavior and task-specific heuristics ([Cui et al., 2026](#); [Rizvi-Martel et al., 2026](#)). Our work investigates whether such observable patterns reflect genuine reasoning mechanisms.

Shortcut Behavior. Continuous thoughts may be decodable without driving reasoning: models can solve tasks via prompt or KV-cache shortcuts, or commit to answers before latent computation. Prior work frames thought use as binary, showing performance persisting under perturbations and ablations of latent states ([Zhang et al., 2025a](#); [Cui et al., 2026](#)), suggesting stronger supervision as the cure. However, such findings cannot establish whether models are simply incapable of using their thoughts or if they just find other circuits in their absence. We therefore treat latent-thought influence as a graded and localizable causal quantity, motivating the question of whether this influence is concentrated in specific directions or distributed ([Gur-Ari et al., 2018](#); [Bernas et al., 2026](#)).

Geometry, Stability, and Latent Dynamics. A complementary perspective on LLM interpretability studies reasoning as movement through the representation space ([Elhage et al., 2022](#); [Park et al., 2024](#); [Gurnee and Tegmark, 2024](#); [Geva et al., 2023](#); [Bhatia et al., 2025, 2026b](#); [Bernas et al., 2026](#); [Zhou et al., 2026](#); [Kumar et al., 2026](#)). In particular, [Wang et al. \(2026\)](#) model explicit CoT as a Markov chain, with CoT being useful only when transitions are stable and aligned. However, such analyses remain limited in LRM interpretability. [Zhu et al. \(2025\)](#) compare the geometry of latent thoughts with optimal node embeddings, suggesting a superpositional graph-search representation. [Wei et al. \(2026\)](#) report that scaled implicit reasoning can become unstable, with representations homogenizing and drifting from the vocabulary manifold. These studies investigate latent state evolution, not its causal effect on model behavior, motivating pairing geometry with intervention.

Concurrent Causal Analyses. Recent work has also examined the causal role of latent thoughts. [Li et al. \(2026\)](#) study how causal influence is distributed across latent-thought steps, finding uneven

and non-local effects and a gap between early output bias and later representational commitment. [Chang et al. \(2026\)](#) combine semantic, geometric, and causal analyses of individual latent thoughts, identifying early thoughts as structured and steerable causal hubs. Complementarily, [Seddik and Fard \(2026\)](#) propose causality, minimality, separability, and stability as desiderata for thought representations, finding that downstream accuracy can mask representational failures. Our work complements these analyses with within-thought causal localization, isolating low-dimensional, loss-sensitive directions that influence behavior.

3 Experimental Setup

Datasets. We evaluate graph-hopping on ProsQA ([Hao et al., 2025](#)) and arithmetic-reasoning on GSM8k ([Cobbe et al., 2021](#)).

Models and Controls. All models are trained separately for each task from pretrained GPT-2 small ([Radford et al., 2019](#)) (Llama-3.2-1B-Instruct ([Meta AI, 2024](#)) extension in Appendix B). Our design compares target LRM models with different curricula but the same recurrence mechanism (Coconut (C) and CODI), and the following controls:

- **Recurrence control: Pause-as-thought (PaT)** follows Coconut’s format and curriculum but replaces recurrence with learned thought tokens.
- **Curriculum control: Coconut_u (C_u),** a curriculum-perturbed Coconut variant that samples other stages with probability $u = 0.3$.
- **Observational controls: Base GPT-2 (B) and Explicit-CoT GPT-2 (CoT)** are included in § 4 to test whether similar patterns can be recovered without latent-reasoning training, either from the probe itself or from ordinary task-solving. They are excluded from causal and geometric analyses, which require dedicated latent thought positions. Details on models and training are in Appendix E.

Statistical Evaluation. We report 95% bootstrapped confidence intervals (1,000 resamples) and use McNemar p -values for paired comparisons. Details and significance results are in Appendix D.

4 Observable Structures Are Insufficient for Mechanistic Attribution

We revisit two LRM interpretability patterns from prior work, discussed in § 2.3: Coconut’s apparent simultaneous encoding of multiple candidate

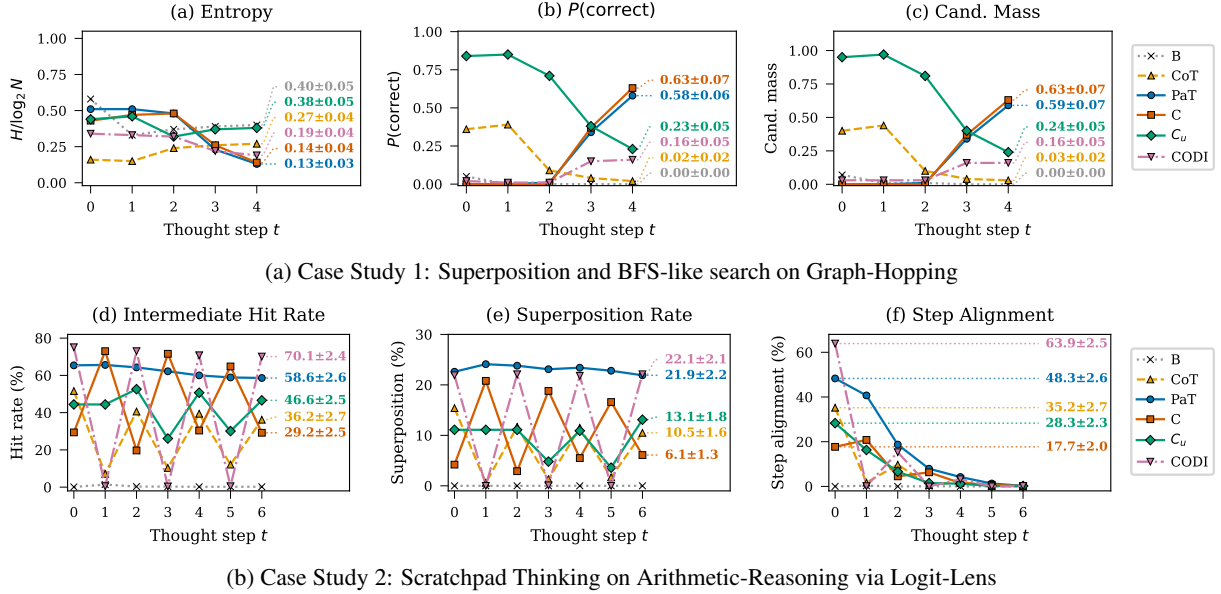


Figure 1: Probing for breadth-first-search (BFS) patterns on graph-hopping (top) and scratchpad-thinking on arithmetic-reasoning (bottom). **Takeaway:** Matched controls reproduce or invert observable patterns, showing they are not specific to the proposed recurrence or curriculum mechanisms and, thus, are insufficient alone for mechanistic attribution.

reasoning paths (*superposition*) and progressive refinement toward the correct path (*breadth-first-search exploration*) on graph-tasks, and CODI’s claimed reasoning through decodable intermediate computations (*scratchpad*) on arithmetics. To address RQ1, we compare models to non-LRM model and curriculum controls introduced in § 3.

4.1 Case Study 1: Superposition and BFS-like search on Graph-Hopping

Following Hao et al. (2025), we probe the *depth frontier* after k latent thoughts by measuring the unnormalized joint probability $p(\text{concept}) = \prod_i p(\text{tok}_i \mid \text{ctx}, \text{“Every”}, \text{tok}_{<i})$ over candidate nodes at depth $\max(k, 1)$ from the root: children for $k=1$; grandchildren for $k=2$. Hao et al. (2025) read this as two patterns. *Superposition*: early latent thoughts hold multiple candidates simultaneously (high entropy; candidate mass spread across nodes). *BFS*: increasing mass on target-reaching nodes with recurrence (rising $P(\text{correct})$; falling entropy). If recurrent feedback drives this BFS-like exploration, C should show this pattern more strongly than PaT. Results in Figure 1a.

B and CoT, lacking latent training, never form the frontier (B near-zero mass; CoT concentrates mass early but degrades with depth). While C reproduces the BFS signature (rising $P(\text{correct})$, falling entropy), PaT matches it with no recur-

rence (but same curriculum). Moreover, C_u which perturbs C’s curriculum inverts this pattern: high mass at shallow depth that **degrades** with depth. CODI, which lacks the curriculum, weakly mirrors C and PaT. Candidate mass tracks $P(\text{correct})$ closely, indicating that frontier mass concentrates on target-reaching nodes. However, superposition requires competing incorrect candidates held simultaneously. Thus, BFS signature is not recurrence-specific: PaT matches it without recurrence, while C_u changes it with recurrence preserved.

4.2 Case Study 2: Scratchpad Thinking on Arithmetic-Reasoning via Logit-Lens

Following Shen et al. (2025), we project final-layer hidden states at thought positions through the LM head and compare decoded tokens against ground-truth intermediate CoT annotations. Shen et al. (2025) read two patterns as evidence of latent scratchpad. *Hit Rate*: decoded tokens match a ground-truth CoT intermediate. *Step Alignment* (ordered-scratchpad): the t -th thought matches the t -th CoT step. We additionally test Superposition Rate (fraction of steps decoding ≥ 2 distinct intermediates) to test generalizability of Hao et al. (2025) claims². Results in Figure 1b.

PaT lacks recurrence but exceeds CODI on hit-

²We also track top-1 decoded token changes across steps to assess trajectory structure on both tasks (Appendix C).

Model	Graph-Hopping	Thoughts Removed	Arithmetic-Reasoning	Thoughts 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_u	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 under thought-ablation at test-time. Performance degrades on arithmetic-reasoning; graph-hopping is unaffected. **Takeaway:** Influence of latent thoughts on model performance is task-varying.

rate overall. C, CoT (no latent training) and C_u also reproduce CODI-like step-wise alternation, although with different magnitudes and phase structure. This suggests the pattern is not specific to CODI’s latent-distillation objective: it could reflect CoT fine-tuning, probe-setup, or another source, but its appearance in controls already contradicts that attribution. Moreover, step-alignment decays sharply with depth for every model, contradicting an ordered scratchpad. B is near zero throughout. Superposition stays low across all models, failing to reproduce COCONUT’s frontier in arithmetic. Overall, the logit-lens readouts recover scratchpad-like signatures without uniquely tracking performance or requiring CODI’s distillation mechanism.

Summary. Both case studies reproduce the original reported findings but challenge the conclusions. Similar patterns in non-LRM controls indicate that the observed patterns are not specific to LRMs. Going beyond observable patterns requires a causal perspective.

5 When and How do LRMs use Latent Thoughts?

Having shown that observational readouts are not specific enough for mechanistic attribution, we now investigate how latent thoughts causally influence model behavior.

Latent Thought Ablation at Test-Time. First, we evaluate whether models require latent thoughts at all by removing them at test-time: $K := K_{\max} \rightarrow 0$ (recurrence skipped for C, C_u , CODI; the k parallel tokens dropped for PaT). Results in Table 1.

On **graph-hopping**, only C_u is affected meaningfully. On **arithmetic-reasoning**, C_u improves but other models suffer major drops. Causal necessity of latent thoughts is task-dependent, but

the shortcut-learning dichotomy (used or bypassed) warrants further checks due to C_u ’s behavior.

Per-Position Causal Tracing. Does maintaining performance after removing thoughts imply they weren’t causally influencing computation when present? LLMs are known to be causally over-determined (McGrath et al., 2023; M  loux et al., 2026): multiple circuits computing the same behavior (Lan et al., 2024); when present, thoughts are still part of the computational paths and can be expected to causally impact behavior. To test this, we extend Meng et al. (2022)’s causal tracing methodology to latent thoughts.

Define site $s = (\ell, p)$ as a layer ℓ and position p in the residual stream (Elhage et al., 2021)³. We run a *clean* pass, a *corrupted* pass (cp) on a partner prompt (dataset instance with different answer), and a *patched* pass (pp) that injects site s ’s clean activation into the corrupted forward pass. At the answer boundary A_b , we greedily decode N tokens, recording at each step j the full next-token distribution $P_{\bullet}^{(j)} = \text{softmax}(z_j)$ for each pass. We compute the forward KL from the clean distribution $\text{KL}(P_{\text{clean}}^{(j)} \| P_{\bullet}^{(j)})$, and average over the content window $W = [n_{\text{fmt}}, e]$ spanning the clean pass output, where n_{fmt} skips the fixed format prefix (e.g., ###, The answer is:), and e is the first end-of-text step (or N). Patching thus measures recovery of the model’s clean behavior:

$$\overline{\text{KL}}_{\bullet} = \frac{1}{|W|} \sum_{j \in W} \text{KL}(P_{\text{clean}}^{(j)} \| P_{\bullet}^{(j)}), \quad \text{IE}(s) = 1 - \overline{\text{KL}}_{\text{pp}} / \overline{\text{KL}}_{\text{cp}}$$

The indirect effect $\text{IE}(s) \leq 1$ measures how far the patch restores the clean output: 1 is full recovery, 0 none, and < 0 is worsening. We report the $\arg \max_{\ell} \text{IE}(s)$ over the position bucket $\{P_{\text{full}}, P_{\text{max}}, P_b, T_1, \dots, T_K, T_{\text{full}}, A_b\}$, where T_t is thought t ; P_b, A_b the latent-block delimiters; P_{max} the strongest single prompt token; and $P_{\text{full}}/T_{\text{full}}$ the jointly patched prompt & thought positions. Results in Figure 2.

Prompt positions (especially P_{full}) strongly recover performance across models on **graph-hopping**. Across all models, thoughts show zero recoverability. Even C_u shows none despite its minor (4.8%) thought necessity in the ablation experiment (possibly due to it being averaged over the full dataset). On **arithmetic-reasoning**, prompt-positions of PaT and CODI still show very high re-

³Per component (attention, MLP) decompositions appear in Appendix C.

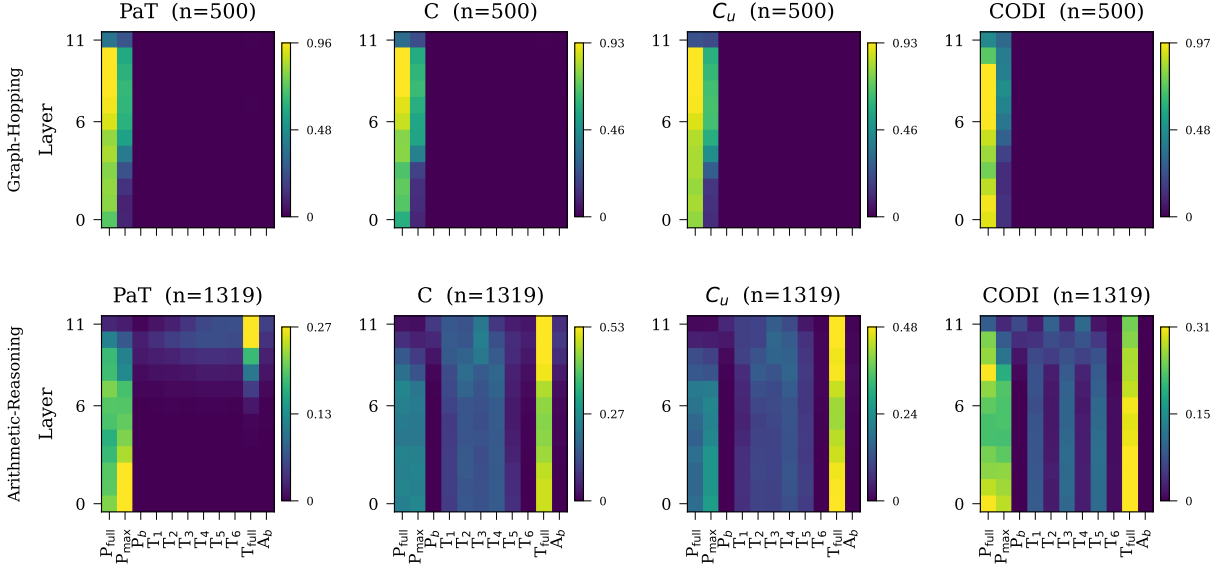


Figure 2: Per-layer IE_{KL} on the residual stream under partner-prompt corruption across buckets $\{P_{\text{full}}, P_{\text{max}}, P_b, T_1, \dots, T_K, T_{\text{full}}, A_b\}$. Prompt positions recover best across models and tasks. Thought positions T_t contribute near zero on graph-hopping; on arithmetic-reasoning, T_{full} yields the strongest recovery for C, C_u , and CODI. **Takeaway:** Latent-thought influence is graded—when large, it can override corrupted prompt positions and steer models toward correct paths.

coverability throughout layers. Individual thought positions also show recoverability rates albeit small. Interestingly, patching all thought positions while other positions are corrupted shows the best recoverability through all layers on C, C_u and CODI and through the later layers on PaT.

Gradient-Subspace Interventions. The previous analyses test whether latent thoughts influence model behavior at test-time. Prior work typically frames this influence as binary: unused thoughts are simply bypassed shortcuts (Zhang et al., 2025a; Cui et al., 2026). But *why*? Do these thoughts lack the capacity to influence behavior, or is their effect simply too small to measure globally? To test this, we isolate the representation directions most sensitive to the loss, geometrically concentrating thought influence into a subspace for intervention.

For thought $h_{i,t} \in \mathbb{R}^D$ (instance i , position t), loss $\mathcal{L}_i$ and gradients $g_{i,t} = \nabla_{h_{i,t}} \mathcal{L}_i$, we define the gradient-subspace $B_t \in \mathbb{R}^{D \times k_t}$ using the top- k_t right singular vectors of the gradient matrix $G_t = [g_{1,t}; \dots; g_{N,t}]$, capturing 95% cumulative energy. To intervene, we split $h_{i,t}$ as $h_{i,t}^B = B_t B_t^\top h_{i,t}$ (projection onto B_t) and $h_{i,t}^\perp = h_{i,t} - h_{i,t}^B$. The projection is scaled by $\alpha \in \{0, 0.5, 1, 1.5, 2, 5, 10, 25, 50, 100\}$ via the update $h_{i,t} \leftarrow h_{i,t}^\perp + \alpha h_{i,t}^B = h_{i,t} + (\alpha - 1) B_t B_t^\top h_{i,t}$ which ablates ($\alpha = 0$) or amplifies ($\alpha > 1$) the

targeted component. A rank-matched orthonormal basis B_t^{rand} via QR decomposition on Gaussian noise serves as the control. Except for PaT, models skip $t = K$ as the gold answer token sits at step K (meaning G_K is empty). Results in Figure 3.

On **graph-hopping**, ablation preserves accuracy, but PaT and C exhibit high- α GRAD flip-rates significantly exceeding Rand, indicating thoughts retain a *graded* causal influence rather than being strictly **bypassed**. C_u is an exception: despite degrading under thought ablation, the gradient directions fail to localize its influence to a concentrated subspace. On **arithmetic-reasoning**, GRAD ablation degrades all models but Rand is inert; both amplifications produce comparable flip-rates⁴.

Summary. Thought utilization is *graded* rather than binary, and depends on how much influence the latent thoughts carry on model behavior. Together with § 4, this motivates moving from finding low-level observable latent patterns to studying the dynamics of behavior-influential directions across latent steps.

6 The Dynamics and Geometry of Latent Thoughts

Observable patterns such as BFS and scratchpad-thinking (§ 4) are not sufficient, on their own,

⁴Subspace ranks in Appendix C (Table 5).

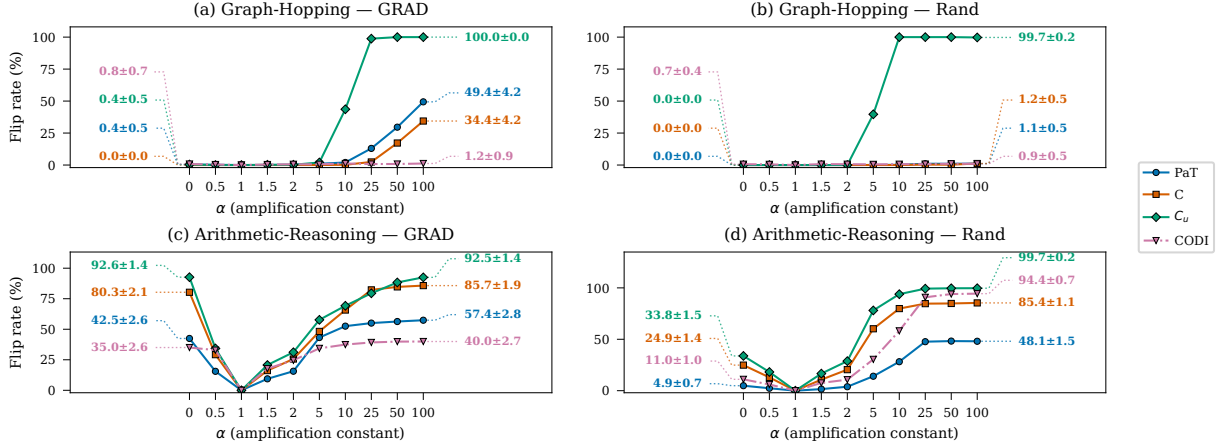


Figure 3: Gradient-subspace intervention flip rates (%) across amplification strength α on the gradient-derived vs random control directions. **Graph-Hopping:** Robust to ablation ($\alpha = 0$) but flips at high- α on GRAD. **Arithmetic-Reasoning:** Degradation under GRAD ablation; GRAD and Rand amplifications cause comparable flip-rates. **Takeaway:** Thought utilization is *graded* and depends on their influence power over model behavior.

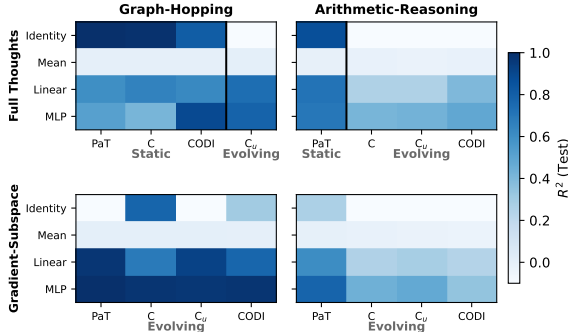


Figure 4: Markovianity of latent thoughts. **Graph-Hopping:** Static dynamics (except C_u); **Arithmetic-Reasoning:** Evolving dynamics. **Takeaway:** Even static appearing full-thoughts show evolving dynamics within the gradient-subspace where influence concentrates.

to establish causal contribution to model behavior. In this section, we argue that rather than drawing general conclusions about any latent reasoning paradigm, the dynamic evolution of latent thoughts (and the directions where their causal influence concentrates) can offer more meaningful insights into how a specific {model, task, latent-training-recipe} configuration behaves. Using the behavior-influential directions identified in § 5, we now apply geometric tools to assess whether the dynamics of highly causally influential latent thoughts differ from those of less causally influential ones.

Markovianity of Thought Trajectories. We first investigate the evolution of thought trajectories, testing whether future states can be predicted from preceding ones. By design, the transformer

architecture makes every thought a function of all previous thoughts and prompt tokens, but the trajectories may encode a simpler approximate structure.

For the thoughts $h_{i,1}, \dots, h_{i,K}$ for every test instance i , we form pairs $X_{i,t} = [h_{i,t-1}; \dots; h_{i,t-n}]$ and $Y_{i,t} = h_{i,t}$ for a Markov order n . We fit a shared map $f: \mathbb{R}^{nD} \rightarrow \mathbb{R}^D$ on the train split and report uniform-average R^2 on the test split under two function classes: ridge-regularized linear regression and a two-layer MLP: $h_t = W_2 \sigma(W_1 X + b_1) + b_2$ with $\sigma = \text{GELU}$, $W_1 \in \mathbb{R}^{256 \times nD}$ (early-stopped on a 10% train slice, R^2 averaged over 3 seeds). We compare against the *mean baseline* $\hat{h}_t = \bar{h}_{\text{train}}$ ($R^2 = 0$ on train split), and the *identity baseline* $\hat{h}_t = h_{t-1}$. Order $n = 1$ is the strict first-order claim $h_t = f(h_{t-1})$; we sweep $n \in \{1, \dots, 5\}$. The same analysis on the gradient-subspace (§ 5) projected thoughts $h_{i,t}^B = B_t B_t^\top h_{i,t}$ tests thought dynamics where influence concentrates. Results (averaged over n) in Figure 4.

On **graph-hopping**, the identity baseline dominates for PaT, C, and CODI: no fitted transition meaningfully improves over simply copying the previous thought, but collapses on C_u ⁵ for which linear and MLP are the best approximates. On **arithmetic-reasoning**, identity dominates on PaT but collapses on C, C_u , and CODI, all best characterized by the MLP while linear closely follows. The dynamics change in the gradient-subspace; the MLP map nearly captures the full **graph-hopping** dynamics and dominates on **arithmetic-reasoning**.

⁵Static requires $\max(R_{\text{id}}^2, R_{\text{mean}}^2) \geq \max(R_{\text{lin}}^2, R_{\text{MLP}}^2) - 0.1$ to avoid overfitting.

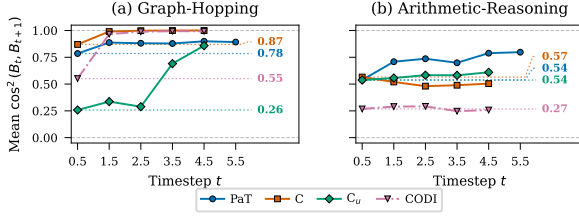


Figure 5: Geometric stability of the gradient-subspaces. **Graph-Hopping:** Stable subspaces (except C_u); **Arithmetic-Reasoning:** Rotating subspaces. **Take-away:** Thoughts with weak influence over model behavior exhibit highly stable gradient-subspace geometries. Points plotted at half-integers as each measures an adjacent-step pair $\cos^2(B_t, B_{t+1})$.

Geometric Stability of Gradient-Subspaces.

Next, we ask whether the gradient-subspace (§ 5) itself remains stable throughout the reasoning trajectory. For adjacent time-steps, we measure the mean squared principal-angle similarity between subspaces: $s_t = \text{mean}(\sigma(B_t^\top B_{t+1})^2)$, where $\sigma(\cdot)$ denotes the singular values of the inter-basis projection matrix. Values of $s_t \rightarrow 1$ indicate stable causal subspaces, while $s_t \rightarrow 0$ indicate orthogonality.

Figure 5 further localizes this distinction to the geometry of the gradient-subspace. On **graph-hopping**, PaT, C and CODI maintain highly stable subspaces while C_u shows near-orthogonal early rotations before stabilizing. On **arithmetic-reasoning**, stability is lower overall: PaT remains most stable, C and C_u show moderate alignment, and CODI undergoes large step-to-step rotations.

Summary. Beyond their causal necessity and sufficiency, the causally influential directions in latent thoughts are consistently subspace-concentrated, and their *dynamics* track influence. Where causal effect is weak (e.g, graph-hopping), full trajectories barely evolve past the first thought, though the causally active subspace supports non-trivial computation; where stronger, trajectory dynamics are more structured, and the causally active subspace supports more diverse computation.

7 Discussion

A causal-first view of latent thoughts. LRM interpretability is especially vulnerable to false explanations: decodable intermediates, attention mass, and visually salient geometries can appear even when the proposed mechanism is absent. Observational tools are most natural when computation

is partly exposed in token space; in LRMs, they can recover structure from continuous states that need not coincide with what the model uses to answer. This makes an observational-first workflow risky. We therefore reverse the order of analysis: rather than starting from a readable latent pattern, we first test whether latent thoughts causally affect behavior and where this influence is concentrated. Only then do we analyze the geometry and dynamics of these behavior-influential regions. Geometry is informative when it describes the part of latent computation that affects predictions.

From thought use to causal localization. This causal-first view helps reframe the shortcut-vs-reasoning dichotomy. The relevant question is not whether latent thoughts are simply “used” or “bypassed”, but how much they influence behavior and in which directions of the representation. Concurrent work similarly finds that causal influence can vary substantially across latent steps and latent states can be steered through intervention (Li et al., 2026; Chang et al., 2026). Our results further show that this influence varies across tasks with different complexity and structure, and can concentrate in specific directions within a latent state. Moreover, a model may solve a task without requiring latent thoughts under ablation while still containing latent directions that can steer its predictions when intervened on. Such cases do not show that the model is incapable of using latent thoughts; they may instead reflect alternative circuits that solve the task when latent computation is removed. Conversely, forcing stronger dependence on latent thoughts (for instance by stronger supervision) does not by itself show that these thoughts implement faithful reasoning. The more informative unit of analysis is therefore the strength and localization of latent-thought influence for a given model and task.

Implications for LRM interpretation and design.

Causal localization also gives interpretability a constructive role. Once behavior-influential regions are identified, they provide concrete targets for geometric analysis, monitoring, and downstream interventions. For example, if influence is consistently concentrated in low-dimensional, loss-sensitive subspaces, compression could focus on preserving the task-relevant component of the latent state rather than the full representation. Along these lines, concurrent work already exploits low-rank CoT geometry to distill reasoning into compact latent trajectories (Solgi et al., 2026). Simi-

larly, if behavior-influential directions are shown to follow stable or predictable step-to-step dynamics, they could inform simplified transition models or more targeted recurrent computation, which is particularly relevant because the recurrent mechanism described in § 2.1 can impose sequential dependencies that are difficult to parallelize during training.

Semantic readability and causal influence are complementary axes. Latent computation can also be interpreted through semantically readable coordinates, but readability and behavioral influence are distinct: a direction can be readable without being causal. The Jacobian-lens work is one instance—a sparse, vocabulary-indexed J-space read as a “language-model global workspace” (Gurnee et al., 2026), extended to LRMs as recurrent J-thoughts (Wu et al., 2026). These semantic directions complement our loss-gradient subspaces, which instead localize causal influence; yet Bhatia et al. (2026a) find J-lens readouts that suggested a strategy the directions did not causally support. This reinforces our causal-first view: readable coordinates can propose computational variables, but only interventions establish their mechanistic relevance.

Relation to CoT faithfulness concerns. Our findings extend documented concerns about explicit CoT: verbalized rationales can misrepresent a model’s internal reasons (Turpin et al., 2023; Arcuschin et al., 2026), and intermediate steps (including “aha” or self-verification moments) may exert little causal influence on the final answer (Boppana et al., 2026; Zhao et al., 2025). LRMs inherit this concern while removing the surface artifact that makes CoT at least partially inspectable. The risk is therefore not only opacity, but false interpretability: latent states may contain decodable or geometrically structured information that appears explanatory without being behaviorally relevant, strengthening the case for causal-first monitoring.

Overall, latent thoughts should be treated as hidden computation, not hidden explanation. Instead of readable latent patterns, LRM interpretability should target the regions of latent computation that causally shape behavior and analyze their dynamics. This causal-first view provides a more grounded and principled basis for designing and auditing latent reasoning systems.

Limitations

First, we estimate the gradient subspace linearly via SVD; a nonlinear estimator (e.g., autoencoder) could capture causal structure a linear basis flattens. Relatedly, our loss-gradient subspaces localize behavioral sensitivity but do not identify the semantic content of those directions. A direct next step is to test whether behavior-influential latent directions admit sparse J-space (Gurnee et al., 2026) descriptions. Second, our causal evidence relies on local intervention-based methods with known caveats: causal tracing and gradient-subspace interventions probe behavior under specific perturbations that can shift the model off-distribution, and interventions admit multiple compatible mechanistic explanations, so “behavior-influential” is a necessary-but-not-sufficient signal of mechanism, **not** proof. Third, our conclusions draw on relatively small models at modest latent steps ($K=6$), and our LRM coverage is partial (PaT, COCONUT, CODI); future work should extend to larger model families and more recent reasoning paradigms. Fourth, our models are trained separately for each task, so task differences may partly reflect task-specific training; multi-task models would allow within-model comparisons. We also analyze only final checkpoints, leaving how observable patterns and causal subspaces develop during training unexplored. To facilitate such training dynamics analyses, we release all intermediate checkpoints in our [model collection](#).

Ethical Considerations and Societal Impact

Latent reasoning models may have substantial societal impact if they make multi-step reasoning more efficient, scalable, and easier to integrate into agentic systems. Yet, by replacing explicit chain-of-thought traces with continuous hidden-state computation, LRMs remove a surface artifact that can be inspected, filtered, or audited. This creates two related risks. First, without reliable tools for interpreting LRMs, users, developers, and auditors may have fewer signals for detecting harmful, deceptive, or misaligned behavior, especially in high-stakes domains such as education, healthcare triage, legal assistance, scientific automation, or autonomous agents. Second, unreliable interpretability tools can create false interpretations, which may lead to monitoring the wrong features and missing the computations that actually drive harmful or misaligned behavior. Our work contributes towards addressing

this risk by arguing that latent thoughts should be treated as hidden computational states, not hidden explanations, and that mechanistic claims require controls and causal tests before latent structure is interpreted. This causal-first view is not a bullet-proof procedure that solves this completely, since interventions can still be local, distribution-shifting, or compatible with multiple explanations. Rather, it provides a stricter evidential standard that should be combined with stress testing, adversarial evaluation, uncertainty reporting, and domain-specific oversight.

Acknowledgements

This work was conducted within the French research unit UMR 5217 and was supported by CNRS (grant ANR-22-CPJ2-0036-01 and ANR-25-CE23-2059-01) and by MIAI@Grenoble-Alpes (grant ANR-19-P3IA-0003 and ANR-23-IACL-0006). This project was provided with HPC computing and storage resources by GENCI at IDRIS thanks to the grant 2026-AD011017855 on the supercomputer Jean Zay’s V100 partition. Thomas’ research is also partially supported by the French Ministry of Higher Education, Research and Innovation under CIFRE PhD Convention “*Reconcile Efficiency and Interpretability in Structured Planning and Reasoning*” (no. 2025/0487).

Use of AI Assistants. AI tools were used during this research for manuscript language editing and polishing, and assistance with editing software code. All AI-generated outputs were reviewed and verified by the authors, who take full responsibility for the scientific content, experimental results, interpretations, and conclusions.

References

- Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthoran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2026. [Chain-of-thought reasoning in the wild is not always faithful](#). In *Forty-third International Conference on Machine Learning*.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Leonard Bereska and Stratis Gavves. 2024. [Mechanistic interpretability for AI safety - a review](#). *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.
- Raphael Bernas, Fanny Jourdan, Antonin Poché, and Celine Hudelot. 2026. [Revisiting anisotropy in language transformers: The geometry of learning dynamics](#). In *Forty-third International Conference on Machine Learning*.
- Agam Bhatia, Camila Blank, and Neel Nanda. 2026a. [Towards surfacing model algorithms with meta-tokens in the J-Space](#). *LessWrong. AI Alignment Forum*.
- Gagan Bhatia, Ahmad Muhammad Isa, Maxime Peyrard, and Wei Zhao. 2026b. [What really controls temporal reasoning in large language models: Tokenisation or representation of time?](#) *Preprint*, arXiv:2603.19017.
- Gagan Bhatia, Maxime Peyrard, and Wei Zhao. 2025. [Date fragments: A hidden bottleneck of tokenization for temporal reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3201–3219, Suzhou, China. Association for Computational Linguistics.
- Siddharth Boppana, Annabel Ma, Max Loeffler, Raphaël Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. 2026. [Reasoning theater: Disentangling model beliefs from chain-of-thought](#). In *Forty-third International Conference on Machine Learning*.
- Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. 2023. [Harms from increasingly agentic algorithmic systems](#). In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 651–666.
- Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. [Causal scrubbing, a method for rigorously testing interpretability hypotheses](#). *AI Alignment Forum*.
- Shuochen Chang, Tong Bai, Xiaofeng Zhang, Qianli Ma, Qingyang Liu, Zhaohe Liao, Yibo Miao, and Li Niu. 2026. [Unlocking the Black Box of Latent Reasoning: An Interpretability-Guided Approach to Intervention](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34019–34032, San Diego, California, United States. Association for Computational Linguistics.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. 2025. [Reasoning models don’t always say what they think](#). *arXiv preprint arXiv:2505.05410*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro

- Nakano, et al. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Yingqian Cui, Zhenwei Dai, Bing He, Zhan Shi, Hui Liu, Rui Sun, Zhiji Liu, Yue Xing, Jiliang Tang, and Benoit Dumoulin. 2026. [How Do Latent Reasoning Methods Perform Under Weak and Strong Supervision?](#) In *Workshop on Latent & Implicit Thinking – Going Beyond CoT Reasoning*.
- Bartosz Cywinski, Bart Bussmann, Arthur Conmy, Joshua Engels, Neel Nanda, and Senthooan Rajamanoharan. 2025. [Can we interpret latent reasoning using current mechanistic interpretability tools?](#) *LessWrong. AI Alignment Forum*.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. [Universal transformers](#). In *International Conference on Learning Representations*.
- Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. [Implicit chain of thought reasoning via knowledge distillation](#). *Preprint*, arXiv:2311.01460.
- Connor Dilgren and Sarah Wiegrefe. 2026. [Are Latent Reasoning Models Easily Interpretable?](#) In *Workshop on Latent & Implicit Thinking – Going Beyond CoT Reasoning*.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. [Amnesic probing: Behavioral explanation with amnesic counterfactuals](#). *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. [Toy models of superposition](#). *arXiv preprint arXiv:2209.10652*.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. [A mathematical framework for transformer circuits](#). *Transformer Circuits Thread*, 1(1):12.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. [Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability](#). *Journal of Machine Learning Research*, 26(83):1–64.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. [Dissecting Recall of Factual Associations in Auto-Regressive Language Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.
- Angeliki Giannou, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. 2023. [Looped transformers as programmable computers](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11398–11442. PMLR.
- Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. [Think before you speak: Training language models with pause tokens](#). In *The Twelfth International Conference on Learning Representations*.
- Satchel Grant, Simon Jerome Han, Alexa Tartaglino, and Christopher Potts. 2026. [Addressing divergent representations from causal interventions on neural networks](#). In *International Conference on Learning Representations*, volume 2026, pages 11330–11356.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. 2018. [Gradient descent happens in a tiny subspace](#). *arXiv preprint arXiv:1812.04754*.
- Wes Gurnee, Nicholas Soironiew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. 2026. [Verbalizable Representations Form a Global Workspace in Language Models](#). *Transformer Circuits Thread*.
- Wes Gurnee and Max Tegmark. 2024. [Language models represent space and time](#). In *The Twelfth International Conference on Learning Representations*.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. 2025. [Training Large Language Models to Reason in a Continuous Latent Space](#). In *Second Conference on Language Modeling*.
- Stefan Heimersheim and Neel Nanda. 2024. [How to use and interpret activation patching](#). *Preprint*, arXiv:2404.15255.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-*

- guage Technologies, Volume 1 (Long and Short Papers), pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Subhash Kantamneni, Joshua Engels, Senthooan Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. [Are sparse autoencoders useful? a case study in sparse probing](#). In *Forty-second International Conference on Machine Learning*.
- Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kojatjlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. 2025. [Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety](#). *Preprint*, arXiv:2507.11473.
- Shashi Kumar, Yacouba Kaloga, Petr Motlicek, Ina Kordasi, and Andrea Cavallaro. 2026. [Geometric latent reasoning induces shorter generations in llms](#). *Preprint*, arXiv:2606.02248.
- Michael Lan, Philip Torr, and Fazl Barez. 2024. [Towards Interpretable Sequence Continuation: Analyzing Shared Circuits in Large Language Models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12576–12601, Miami, Florida, USA. Association for Computational Linguistics.
- Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. [Probing for the usage of grammatical number](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.
- Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. [When Thinking Fails: The Pitfalls of Reasoning for Instruction-Following in LLMs](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 77925–77962. Curran Associates, Inc.
- Zirui Li, Xuefeng Bai, Kehai Chen, Yizhi LI, Jian Yang, Chenghua Lin, and Min Zhang. 2026. [Dynamics Within Latent Chain-of-Thought: An Empirical Study of Causal Structure](#). In *Forty-third International Conference on Machine Learning*.
- Jia Liang and Liangming Pan. 2026. [Do latent-cot models think step-by-step? a mechanistic study on sequential reasoning tasks](#). *Preprint*, arXiv:2602.00449.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. [Faithful Chain-of-Thought Reasoning](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.
- Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. 2023. [The hydra effect: Emergent self-repair in language model computations](#). *Preprint*, arXiv:2307.15771.
- Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. 2026. [The dead salmon of AI interpretability](#). In *Mechanistic Interpretability Workshop at ICML 2026*.
- Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. 2025. [Everything, everywhere, all at once: Is mechanistic interpretability identifiable?](#) In *The Thirteenth International Conference on Learning Representations*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in gpt](#). *Advances in neural information processing systems*, 35:17359–17372.
- Meta AI. 2024. [Llama 3.2: Revolutionizing edge AI and vision with open, customizable models](#). Accessed: 2026-02-16.
- Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. 2024. [A glitch in the matrix? locating and detecting language model grounding with fakepedia](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6828–6844, Bangkok, Thailand. Association for Computational Linguistics.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. [The linear representation hypothesis and the geometry of large language models](#). In *Forty-first International Conference on Machine Learning*.
- Brad Peters, Sayam Goyal, María Emilia Granda, Akshath Vijayakumar Narmadha, Dharunish Yugeswardeenoo, Callum Stuart McDougall, Sean O’Brien, Ashwinee Panda, Kevin Zhu, and Cole Blondin. 2025. [Scratchpad Thinking: Alternation Between Storage and Computation in Latent Reasoning Models](#). In *Mechanistic Interpretability Workshop at NeurIPS 2025*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language Models are Unsupervised Multitask Learners](#). *OpenAI blog*, 1(8):9.

- Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. [Probing the probing paradigm: Does probing accuracy entail task relevance?](#) In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.
- Michael Rizvi-Martel, Guillaume Rabusseau, and Marius Mosbach. 2026. [The illusion of superposition? a principled analysis of latent thinking in language models](#). In *Third Conference on Language Modeling*.
- Fahd Seddik and Fatemeh Fard. 2026. [Formalizing latent thoughts: Four axioms of thought representation in llms](#). *Preprint*, arXiv:2606.27378.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. [CODI: Compressing chain-of-thought into continuous space via self-distillation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 677–693, Suzhou, China. Association for Computational Linguistics.
- Ryan Solgi, Jiayi Tian, and Zheng Zhang. 2026. [LoRi: Low-Rank Distillation for Implicit Reasoning](#). *Preprint*, arXiv:2606.05315.
- Damien Teney, Maxime Peyrard, and Ehsan Abbasnejad. 2022. [Predicting is not understanding: Recognizing and addressing underspecification in machine learning](#). In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, page 458–476, Berlin, Heidelberg. Springer-Verlag.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting](#). *Advances in Neural Information Processing Systems*, 36:74952–74965.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International Conference on Learning Representations*.
- Zihan Wang, Yijun Dong, and Qi Lei. 2026. [When does Chain-of-Thought Help: A Markovian Perspective](#). In *Workshop on Latent & Implicit Thinking – Going Beyond CoT Reasoning*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. 2026. [SIM-CoT: Supervised Implicit Chain-of-Thought](#). In *The Fourteenth International Conference on Learning Representations*.
- Junde Wu, Jiayuan Zhu, Fengling Liu, Minhao Hu, and Jiazhen Pan. 2026. [J-cot: Chain-of-thought in j-space](#). *Preprint*, arXiv:2607.21981.
- Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. 2025a. [Do Latent Tokens Think? A Causal and Adversarial Analysis of Chain-of-Continuous-Thought](#). *Preprint*, arXiv:2512.21711.
- Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Wang. 2025b. [Soft Thinking: Unlocking the Reasoning Potential of LLMs in Continuous Concept Space](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 168990–169012. Curran Associates, Inc.
- Jiachen Zhao, Yiyao Sun, Weiyan Shi, and Dawn Song. 2025. [Can aha moments be fake? identifying true and decorative thinking steps in chain-of-thought](#). *arXiv preprint arXiv:2510.24941*.
- Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru Zhang. 2026. [The geometry of reasoning: Flowing logics in representation space](#). In *The Fourteenth International Conference on Learning Representations*.
- Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart J Russell, and Yuandong Tian. 2025. [Reasoning by superposition: A theoretical perspective on chain of continuous thought](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 79931–79963. Curran Associates, Inc.

Model	GPT-2	Llama-3.2-1B-Instruct
GSM8K-Aug		
CoT	darpanaswal/cot-gpt2-gsm	thomas-ferraz/CoT-Llama3.2-1B-GSM8kAug
PaT	darpanaswal/pause-gpt2-gsm	thomas-ferraz/Pause-Llama3.2-1B-GSM8kAug
Coconut	darpanaswal/coconut-gpt2-gsm	thomas-ferraz/Coconut-Llama3.2-1B-GSM8kAug
Coconut _u	darpanaswal/coconut_u0.3-gpt2-gsm	thomas-ferraz/Coconut-u0.3-Llama3.2-1B-GSM8kAug
CODI	darpanaswal/codi-gpt2-gsm	thomas-ferraz/CODI-Llama3.2-1B-GSM8kAug
ProsQA		
CoT	darpanaswal/cot-gpt2-prosqa	thomas-ferraz/CoT-Llama3.2-1B-ProsQA
PaT	darpanaswal/pause-gpt2-prosqa	thomas-ferraz/Pause-Llama3.2-1B-ProsQA
Coconut	darpanaswal/coconut-gpt2-prosqa	thomas-ferraz/Coconut-Llama3.2-1B-ProsQA
Coconut _u	darpanaswal/coconut_u0.3-gpt2-prosqa	thomas-ferraz/Coconut-u0.3-Llama3.2-1B-ProsQA
CODI	darpanaswal/codi-gpt2-prosqa	thomas-ferraz/CODI-Llama3.2-1B-ProsQA

Table 2: Model checkpoints trained for this work and publicly released. We release all model variants for both GPT-2 and Llama-3.2-1B-Instruct backbones on GSM8K-Aug and ProsQA.

A Reproducibility Statement

To facilitate full reproducibility, we release the source code required to reproduce our training and analyses. The main analysis code is available [here](#). Training code for the Coconut family and its controls and variants is available [here](#), while the CODI training implementation and variants are available [here](#). We additionally release all trained model checkpoints used in our experiments, covering the model variants considered across GPT-2 and Llama-3.2-1B-Instruct on GSM8K-Aug and ProsQA, and including intermediate training checkpoints, at the Hugging Face collection [here](#); direct links to each checkpoint are provided in Table 2.

Dataset sources, splits, and statistics are documented in Appendix E.1, while model and control definitions, training procedures, and implementation details are provided in Appendix E.2–E.3, including the complete training configurations and hyperparameters in Tables 32–33. The analysis procedures are described in § 4–§ 6, with the Llama-3.2 replication and additional analyses reported in Appendices B–C. Statistical evaluation procedures and full significance results are provided in § 3 and Appendix D, including bootstrap confidence intervals, paired significance tests, and multiple-testing correction. Together, these resources provide a reproducible foundation for verifying our results and extending the causal and geometric analysis of latent reasoning models.

B Extension of Main Results to Llama-3.2-1B-Instruct

In this section, we extend the experiments in § 4, § 5 and § 6 to Llama-3.2-1B-Instruct ⁶ (Meta AI,

⁶Available at the HuggingFace’s Hub at: <https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>.

2024; Grattafiori et al., 2024) to analyze how latent training impacts larger models.

Epiphenomenal Patterns in LRM Interpretability (§ 4). Figure 6a reports the results for case study 1 (superposition and BFS-like search on graph-hopping). Like in the GPT-2 case, we find PaT replicating the superposition and BFS dynamics of C, arguably better than C itself. C_u does not create the flipped profile with Llama as it did in the GPT-2 case, but we see its confidence peak at an intermediate step instead of the final step. B, CoT and CODI show similar profiles as in their GPT-2 counterpart. Overall, the observable patterns of superposition and BFS remain uninformative of where they truly emerge and whether they impact model performance in any way.

Figure 6b reports the results for case study 2 (scratchpad reasoning on arithmetic-reasoning via logit-lens). CODI’s own strict alternation pattern disappears, and no other model recreates anything resembling it. C_u completely collapses, showing flat profiles across metrics under logit-lens decodability (even below control B). This provides further evidence for CODI’s scratchpad thinking mechanism being attributable more to the model-family than to the latent-reasoning recipe itself.

When and How do LRMs use Latent Thoughts (§ 5)? Table 3 reports the results for the latent thought ablation at test-time experiment. Results are comparable to the GPT-2 family except for C_u: Unlike C_u GPT-2 (Table 1), it shows (i) negligible thought necessity in graph-hopping, and (ii) a minor drop (instead of improvement) in arithmetic-reasoning under thought ablation. This shows that latent thought necessity varies not just across tasks but also model families. Additionally, C_u’s flat

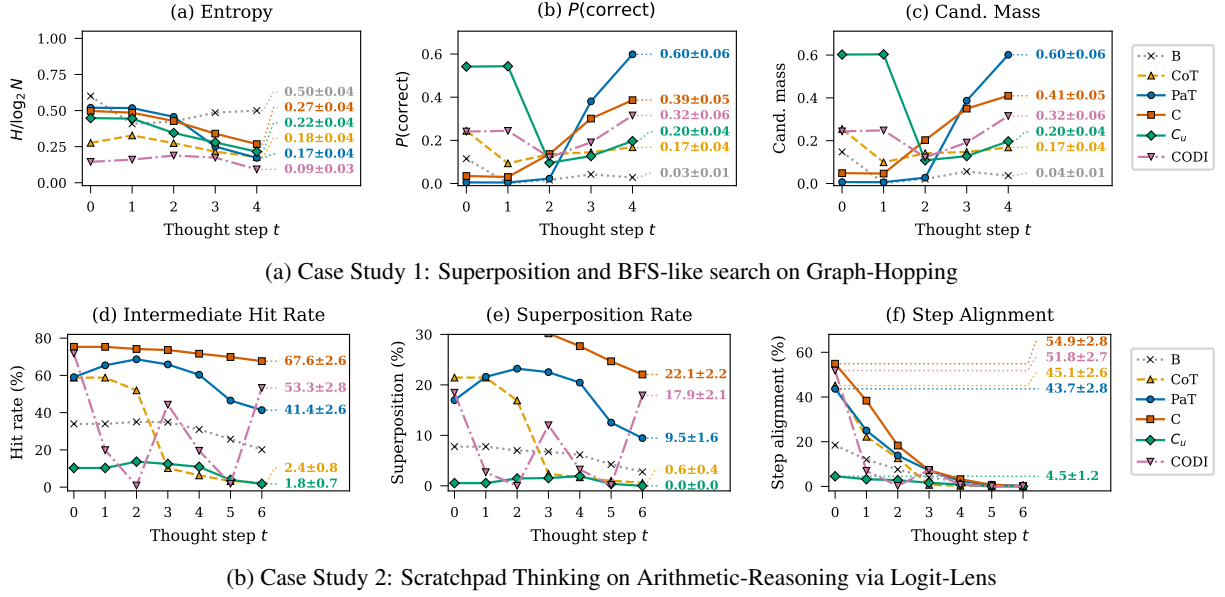


Figure 6: Probing for breadth-first-search (BFS) patterns on graph-hopping (top) and scratchpad-thinking on arithmetic-reasoning (bottom) on the Llama-3.2 model family (§ 4). **Takeaway:** Matched controls reproduce or invert these observable patterns, showing they are not specific to the proposed recurrence or curriculum mechanisms and, thus, are insufficient alone for mechanistic attribution.

Model	Graph-Hopping	Thoughts Removed	Arithmetic-Reasoning	Thoughts Removed
B	2.2 $\pm$ 1.3	—	7.7 $\pm$ 1.5	—
CoT	96.8 $\pm$ 1.5	—	53.3 $\pm$ 2.6	—
PaT	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	26.8 $\pm$ 2.4	3.9 $\pm$ 1.0
C	99.6 $\pm$ 0.5	99.6 $\pm$ 0.5	33.2 $\pm$ 2.4	12.1 $\pm$ 1.7
C_u	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	51.7 $\pm$ 2.6	49.4 $\pm$ 2.7
CODI	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	44.6 $\pm$ 2.6	23.5 $\pm$ 2.3

Table 3: Accuracy under thought-ablation at test-time on the Llama-3.2 model family (§ 5). Performance degrades on arithmetic-reasoning; graph-hopping is unaffected. **Takeaway:** Influence of latent thoughts on model performance is task-varying.

hit-rate on arithmetic-reasoning despite clear (albeit small) thought necessity further demonstrates how observable patterns fall short of explaining true model behavior.

Figure 7 reports the results on the causal-tracing experiment. All **graph-hopping** models and CODI in arithmetic-reasoning show near identical patterns to their GPT-2 counterparts, however while PaT shows weak but non-zero recoverability in the thought positions (T_{full} specifically), C’s thought-positions (T_6 & T_{full}) show negative recoverability in some layers, showing that thoughts might actually hurt the model’s reasoning. For PaT, C and C_u in **arithmetic-reasoning**, even patching full thoughts has negligible restoration effect; however, patching the answer boundary at later layers

shows high recoverability. CODI shows high recoverability when all thoughts are restored. This aligns with Hao et al. (2025)’s account that their latent training recipe provides more pronounced results on GPT-2 as compared to Llama-3.2-1B-Instruct, possibly due to larger models’ extensive language pre-training making latent training more challenging. However, Shen et al. (2025) demonstrate similar gains across model families.

Figure 8 reports the results on the gradient-subspace interventions experiment. On **graph-hopping**, the results remain comparable to the GPT-2 model family. C_u , now with negligible thought-necessity (Table 3), shows comparable behavior to other models. PaT remains flat until a sudden spike at $\alpha = 100$, possibly an off-manifold effect which is a known consequence of causal interventions (Grant et al., 2026). **Arithmetic-reasoning** models show no noticeable deviations from their GPT-2 counterparts.

The Dynamics and Geometry of Latent Thoughts (§ 6). Figure 9 reports the results on the Markovianity of thought trajectories experiment. Results remain comparable to the GPT-2 counterpart, especially in the gradient subspaces. Notably, C_u on graph-hopping still sides in the evolving column despite negligible thought-necessity (Table 3), demonstrating the impact of its stage-mixed curricula over the evolution of its

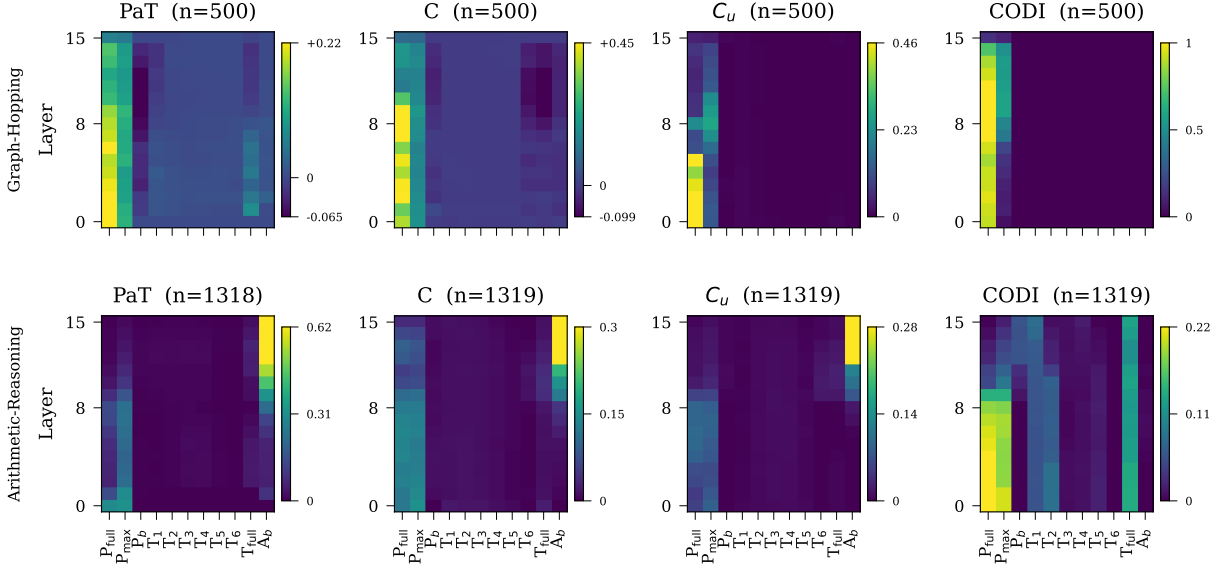


Figure 7: Per-layer IE_{KL} on the residual stream under partner-prompt corruption across buckets $\{P_{full}, P_{max}, P_b, T_1, \dots, T_K, T_{full}, A_b\}$ on the Llama-3.2 model family. Prompt positions recover best across models and tasks. Thought positions T_t contribute near zero on graph-hopping; on arithmetic-reasoning, T_{full} yields the strongest recovery for C, C_u , and CODI. **Takeaway:** Latent-thought influence is graded—when large, it can override corrupted prompt positions and steer models toward correct paths. C and PaT show weaker thought recovery than on GPT-2, consistent with Hao et al. (2025)’s latent-training recipe being less pronounced on Llama-3.2; CODI (Shen et al., 2025), whose recipe transfers across families still recovers strongly when all thoughts are restored (§ 5).

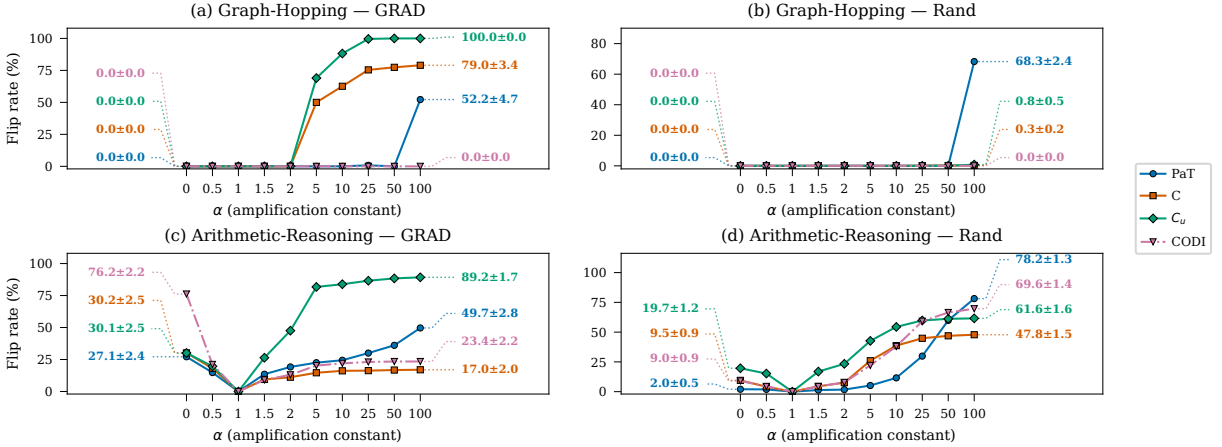


Figure 8: Gradient-subspace intervention flip rates (%) across amplification strength α on the gradient-derived vs random control directions on the Llama-3.2 model family (§ 5). **Graph-Hopping:** Robust to ablation ($\alpha = 0$) but flips at high- α on GRAD. **Arithmetic-Reasoning:** Degradation under GRAD ablation; GRAD and Rand amplifications cause comparable flip-rates. **Takeaway:** Thought utilization is *graded* and depends on their influence power over model behavior.

latent thoughts. Identity collapses for CODI on **graph-hopping** which is best approximated by linear and MLP. All **arithmetic-reasoning** models show evolving dynamics except for C, which is overwhelmingly static.

Figure 10 reports the results on the geometric stability of gradient-subspaces experiment. On **graph-hopping**, C and C_u remain highly stable. PaT appears to show a rotating subspace; Table 5

shows ranks collapsed to 1–2 at all positions. With such few dominant directions, any change produces a large swing. A similar story follows for CODI except that its rotations stabilize after two steps. On **arithmetic-reasoning**, PaT and C remain stable; notably, PaT’s rank again collapses yet alignment stays high, the opposite behavior from graph-hopping. C_u ’s and CODI’s subspaces show high rotation across steps.

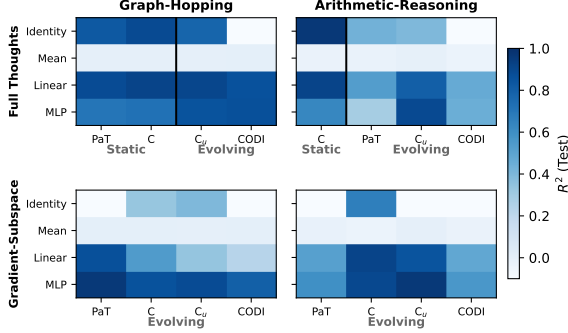


Figure 9: Markovianity of latent thoughts on the Llama-3.2 model family (§ 6). **Graph-Hopping:** Mixed dynamics; **Arithmetic-Reasoning:** Evolving dynamics (except C). **Takeaway:** Unlike the clean task-split on GPT-2, full thought dynamics are mixed across tasks, yet static appearing full-thoughts still show evolving dynamics within the gradient-subspace where influence concentrates.

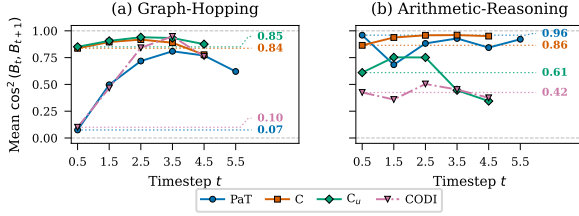


Figure 10: Geometric stability of the gradient-subspaces of the Llama-3.2 model family (§ 6). **Graph-Hopping:** C and C_u stable, PaT and CODI show initial rotations; **Arithmetic-Reasoning:** PaT and C stable, C_u and CODI rotating. **Takeaway:** Unlike GPT-2, subspace geometry neither splits cleanly by task nor tracks thought influence.

C Extended Analyses

This section reports additional results for experiments conducted in § 4, § 5 and § 6.

k	0	1	2	3	4	5	6
Depth probed	1	1	2	3	4	5	6
n (instances with ≥ 2 candidates)	461	461	486	439	182	26	3

Table 4: Sample size at each probing depth. Results at $k \geq 5$ are computed over too few instances ($n < 50$) to support aggregate conclusions and are omitted from the main analysis.

Superposition and BFS-like search on Graph-Hopping. The main text reports aggregate results for $k=0 \dots 4$ only. [Hao et al. \(2025\)](#) illustrate the probing methodology on individual examples; Table 4 shows that the aggregate version necessarily confronts the fact that most ProsQA graphs do not

extend to depth 6.

Attention Mass Distribution. As an additional demonstration of epiphenomenal patterns in interpretability, we evaluate how much attention the answer-generating boundary token directs at the latent thought tokens versus the prompt tokens. Let t_{end} be the sequence position of this boundary token. We extract attention distribution over all preceding key positions j , averaged across all L transformer layers and H attention heads:

$$\bar{\alpha}_j = \frac{1}{L \cdot H} \sum_{l=1}^L \sum_{h=1}^H \alpha_{l,h}(t_{end}, j)$$

where $\alpha_{l,h}$ is the attention weight at layer l and head h . The total attention mass for a given sequence partition $\mathcal{S}$ (e.g., prompt or latent thoughts) is calculated as $\sum_{j \in \mathcal{S}} \bar{\alpha}_j$. Results in Figure 11.

GPT-2: On **graph-hopping**, PaT and C_u attend to their latent thoughts while C and CODI largely ignore them. Despite no training, both controls B and CoT place significant mass on their (forced) thoughts. For **arithmetic-reasoning**, all models place considerable mass on their thought tokens.

Llama-3.2: On **graph-hopping**, B, CoT and PaT place substantial mass on the latent thoughts while C, C_u and CODI largely ignore them. **Arithmetic-reasoning** has a similar story, except the attention of C, C_u and CODI slightly increases.

Logit-Lens Decoded Thought Trajectories. Figure 12 illustrates representative logit-lens trajectories.

GPT-2: On **graph-hopping**, PaT and C collapse to the answer delimiter at every position despite strong task performance; only C_u produces somewhat evolving projections, while CODI and the controls remain semantically incoherent. On **arithmetic-reasoning**, B has no decodable structure while CoT shows moderate decodability despite no latent training—suggesting inherited structure from *CoT* fine-tuning. PaT collapses to the answer delimiter at most positions again. CODI exhibits the clearest alternation between intermediate results and formatting tokens; C has lower decodability but similar alternation while C_u shows alternation with operators at non-result positions.

Llama-3.2: On **graph-hopping**, PaT and C_u collapse entirely to the answer delimiter at every position. C similarly decays into a mix of delimiters and static formatting tokens. CODI loops continuously on filler words. Controls B and CoT

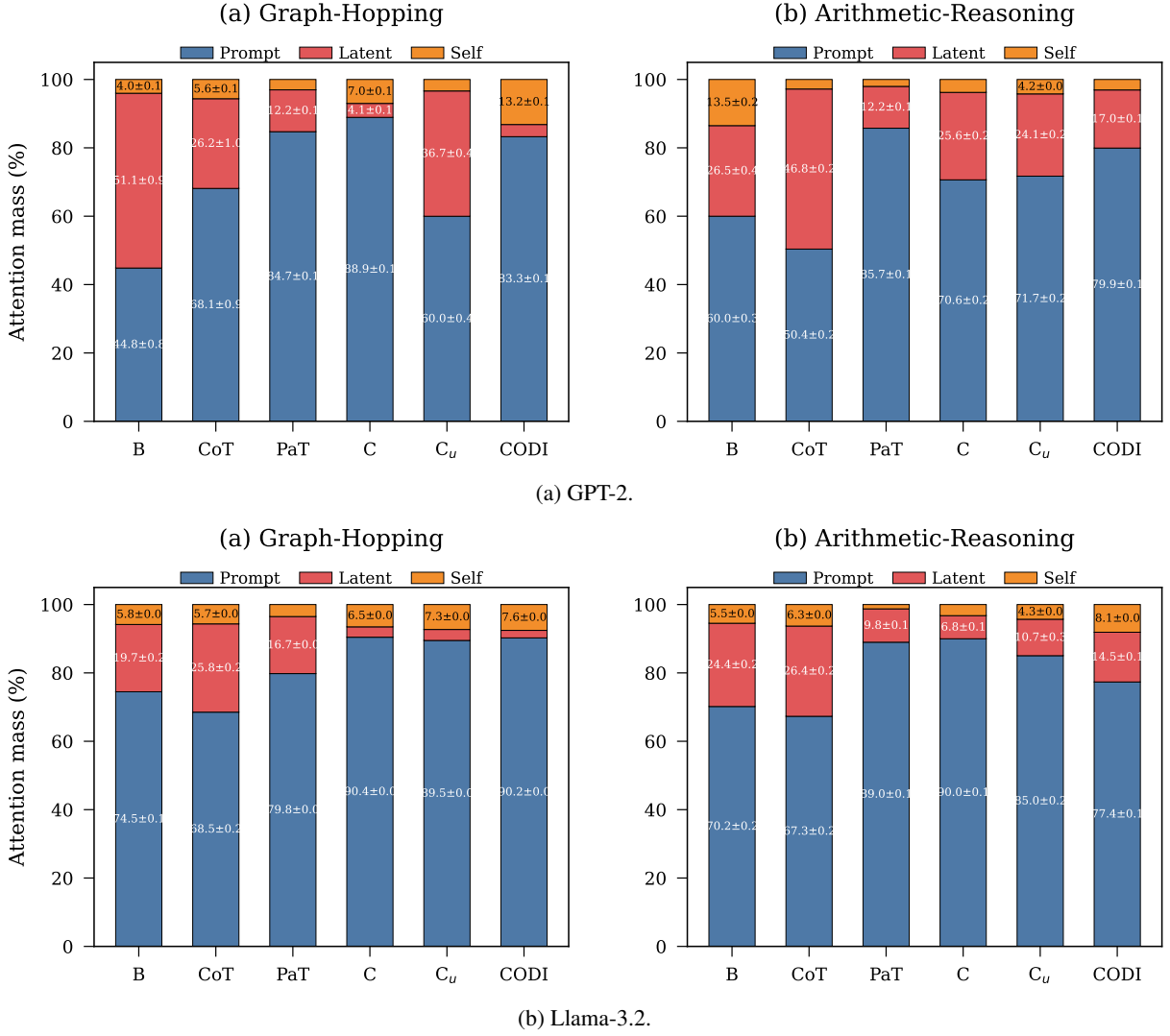


Figure 11: Attention mass distribution of the boundary token across prompt tokens, latent thoughts, and itself, averaged across all transformer layers and heads for GPT-2 and Llama-3.2 model families.

remain semantically incoherent. On **arithmetic-reasoning**, latent trajectories exhibit poorer overall decodability compared to the GPT-2 models. PaT decodes to a mix of numbers and delimiters, never laying out an ordered scratchpad. C initially decodes to answer delimiters, then degrades into repeating static numbers, whereas C_u instead drifts across operators and random tokens without settling on coherent intermediates. CODI displays a mix of numerical tokens, text, and operators, but lacks the structured scratchpad alternation observed in its GPT-2 counterpart. Controls B and CoT output scattered numbers but fail to form coherent intermediate steps.

Decomposition of Residual Stream Causal Tracing Recovery into Attention and MLP-Outputs. Results for GPT-2 family in Figures 13a & 13b and

for Llama-3.2 family in Figures 14a & 14b.

GPT-2: In contrast to the full residual stream (Figure 2), the thought positions of **graph-hopping** models show *small but real* causal effect (specially in attention decomposition). This further evidences that thought use isn't necessarily binary but depends on the influence carried by the thoughts. The **arithmetic-reasoning** models show patterns similar to the full residual stream (Figure 2).

Llama-3.2: Similarly to the GPT-2 family, the thought positions of **graph-hopping** models show *small but real* causal effect in both attention and MLP output decompositions. **Arithmetic-reasoning** models yet again show similar patterns as in the full residual stream (Figure 7).

Gradient-Subspace Dimensionality. Table 5 (GPT-2) and Table 6 (Llama-3.2) report per-

Graph-Hopping							Arithmetic-Reasoning									
B	<	is	is	is	is	is	is	B	< lat...	:	.	The	The	The		
CoT	Sally	a	a					CoT	4	.	1	.	4		+	
PaT	###	###	###	###	###	###	###	PaT	1	3	###	###	###	###	###	
C	###	###	###	###	###	###	###	C	###	5	###	6	/	1	###	
C _u	S	###		Every	n	n	n	C _u	3	60	+	80	-	###	###	
CODI	The	is	is	is	is	is	is	CODI	15	,	33	?	11	.	13	
	t ₀	t ₁	t ₂	t ₃	t ₄	t ₅	t ₆		t ₀	t ₁	t ₂	t ₃	t ₄	t ₅	t ₆	
	Thought step								Thought step							

(a) GPT-2.

Graph-Hopping								Arithmetic-Reasoning							
B	Is	Is	Since	Since	Since	This	This	B	15	15	15	15	15	20	Each
CoT	S	S	S	S	S	S	S	CoT	60	60	3	.	Math	ETIME	'),
PaT	###	###	###	###	###	###	###	PaT	0	4	60	60	75	###	###
C	###	###	###	###	###	s	s	C	###	###	6	6	7	7	7
C _u	###	###	###	###	###	###	###	C _u	<<	<<	<<	The	In	It	ant
CODI	The	is	.	.	.	.	.	CODI	6	.us	.	9	0	.	1
	t ₀	t ₁	t ₂	t ₃	t ₄	t ₅	t ₆		t ₀	t ₁	t ₂	t ₃	t ₄	t ₅	t ₆
Thought step								Thought step							

(b) Llama-3.2.

Figure 12: Logit-lens projection examples for both tasks (graph hopping and arithmetic reasoning) on the GPT-2 and Llama-3.2 model families. The final-layer hidden states at every thought position are decoded through the LM head to obtain the vocabulary projections.

timestep gradient-subspace diagnostics: rank k_t at $\rho=0.95$, mean rank $\bar{k}$ (excluding the degenerate final recurrent step), adjacent and off-diagonal $\cos^2(B_t, B_{t'})$, and the norm fraction $\|h^c\|/\|h\|$ retained in the gradient-subspace (§ 5). Causal influence concentrates along few directions (relative to full dimensions) across tasks and model families.

GPT-2: On **graph-hopping**, the subspaces are extremely low-dimensional for all models; on **arithmetic-reasoning**, ranks grow (yet still small relative to full dimensionality) except for CODI whose mean rank stays comparable across tasks, plausibly a consequence of its LoRA training.

Llama-3.2: We see a similar story as in the GPT-2 counterpart. The exceptions being PaT on arithmetic-reasoning with its surprisingly low rank that compares to its graph-hopping counterpart, and CODI whose ranks across tasks no longer match.

Predicted-Token Gradient Subspace. The gradient-subspaces in the main text are isolated from the loss on the gold answer token y^* , i.e. $g_{i,t} = \nabla_{h_{i,t}} \mathcal{L}_i(y^*)$. To verify that our conclusions do not hinge on this choice, we repeat the full pipeline using gradients of the loss on the model’s own predicted token $\hat{y} = \arg \max_y p_\theta(y | x, e_{1:K}^{\text{lat}})$, i.e. $\tilde{g}_{i,t} = \nabla_{h_{i,t}} \mathcal{L}_i(\hat{y})$, yielding a predicted-token gradient-subspace $\tilde{B}_t$. We rerun (i) the gradient-subspace interventions (Figure 15a for GPT-2 and Figure 15b for Llama-3.2), (ii) the Markovianity analysis on the predicted-token subspace-projected thoughts (Figure 16a and Figure 16b), and (iii) the geometric-stability analysis of the predicted-token subspaces (Figure 17a and Figure 17b). We also report the corresponding per-timestep dimensionality diagnostics in Table 7 (GPT-2) and Table 8 (Llama-3.2). Results across all experiments and the dimensionality diagnostic remain nearly identical

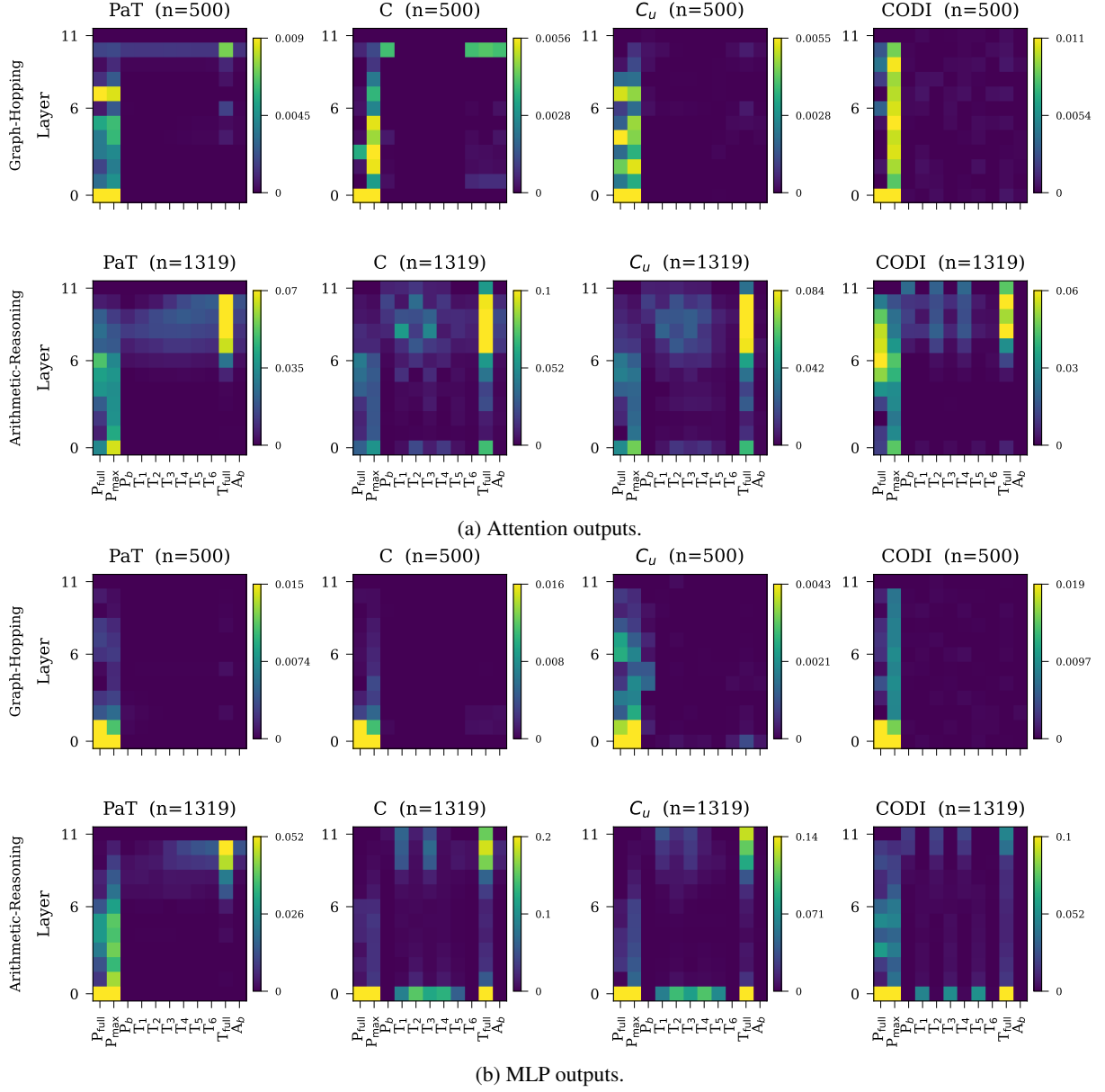


Figure 13: Per-layer IE_{KL} under partner-prompt corruption across buckets $\{P_{\text{full}}, P_{\text{max}}, P_b, T_1, \dots, T_K, T_{\text{full}}, A_b\}$, decomposed into (a) attention and (b) MLP outputs for the GPT-2 family. Unlike the full residual stream (Figure 2), thought positions T_t on **graph-hopping** carry a *small but nonzero* causal effect, most visible in the attention decomposition. **Arithmetic-reasoning** mirrors the full-residual pattern, with T_{full} recovering most strongly for C, C_u , and CODI. *Takeaway*: Latent-thought influence is graded rather than binary—even where it appears negligible, thought positions exert a measurable causal effect.

Task	Model	k_0	k_1	k_2	k_3	k_4	k_5	k_6	$\bar{k}$	Adj. $\cos^2$	Off-diag. $\cos^2$	$\ h^e\ /\ h\ $
Graph-Hopping	PaT	15	16	16	16	17	17	17	16.3	0.871	0.697 ± 0.026	0.353
	C	13	12	11	11	12	12	—	11.8	0.972	0.671 ± 0.024	0.026
	C_u	15	13	17	17	17	14	—	15.5	0.486	0.197 ± 0.012	0.045
	CODI	20	33	35	36	36	37	—	32.8	0.898	0.587 ± 0.008	0.100
Arithmetic-Reasoning	PaT	160	145	146	109	122	156	153	141.6	0.711	0.623 ± 0.019	0.378
	C	155	100	129	135	128	115	—	127.0	0.512	0.394 ± 0.021	0.369
	C_u	126	114	115	122	119	134	—	121.7	0.574	0.404 ± 0.014	0.297
	CODI	51	14	64	18	57	13	—	36.2	0.271	0.351 ± 0.016	0.306

Table 5: Gradient-subspace diagnostics across tasks for the GPT-2 model family.

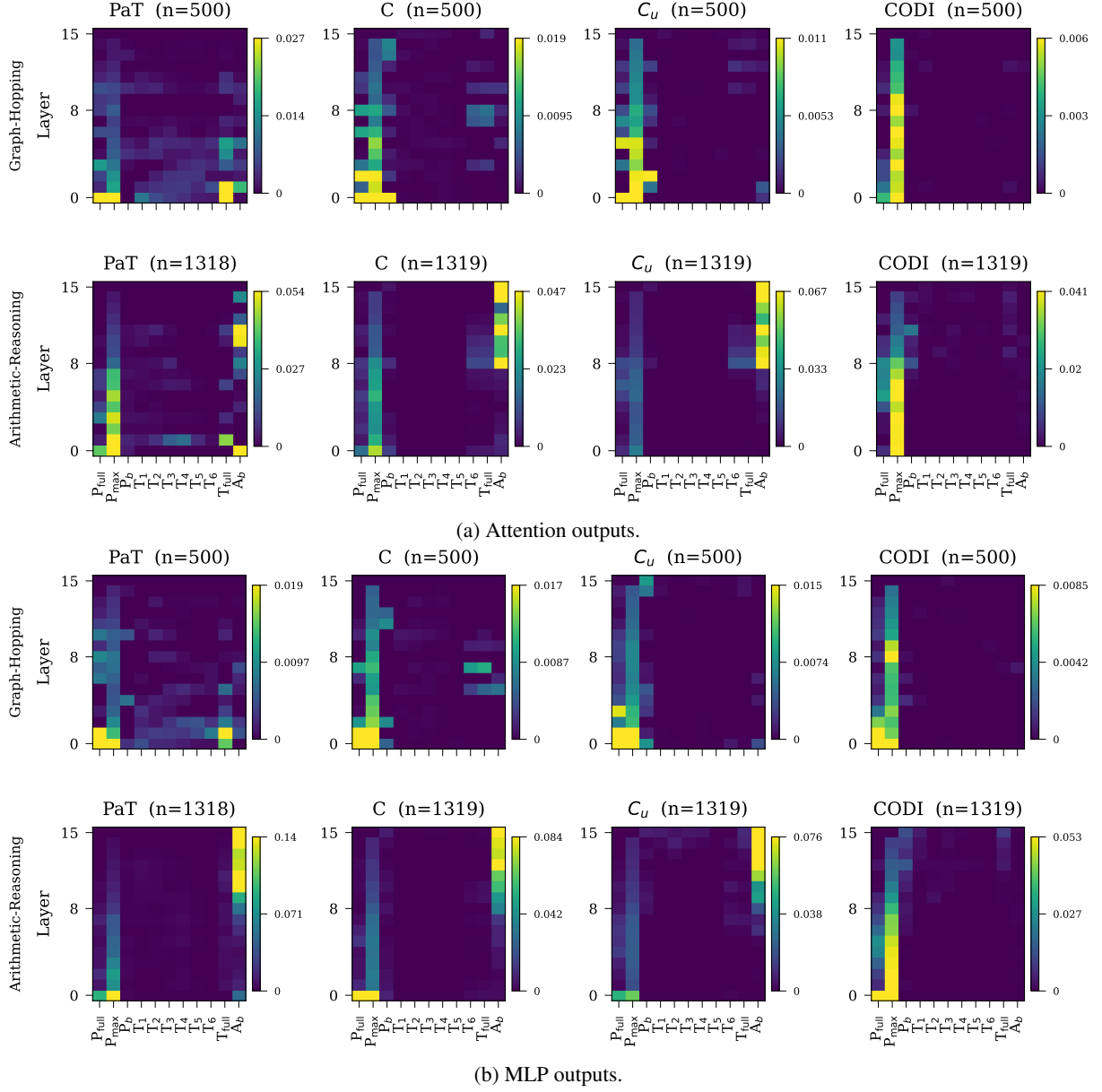
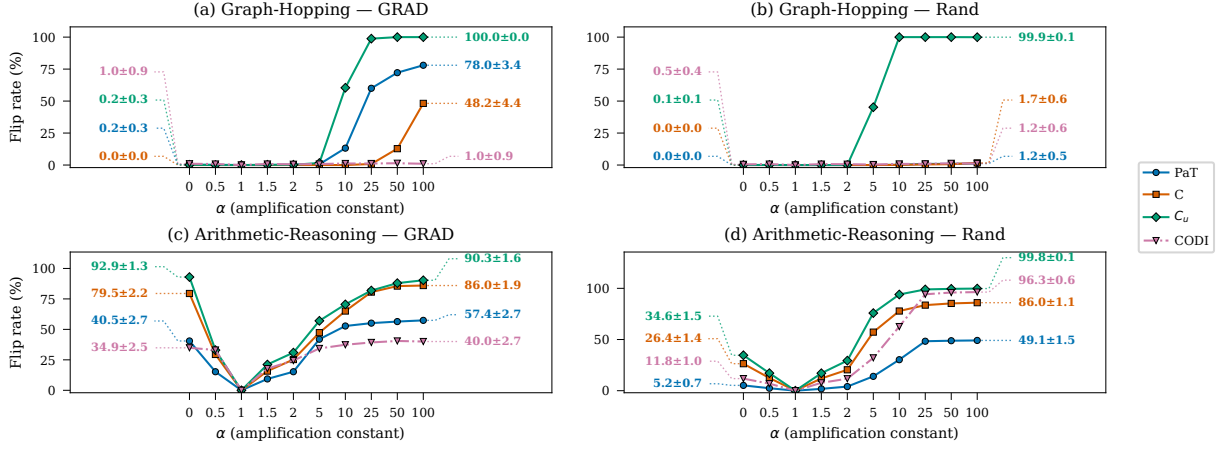


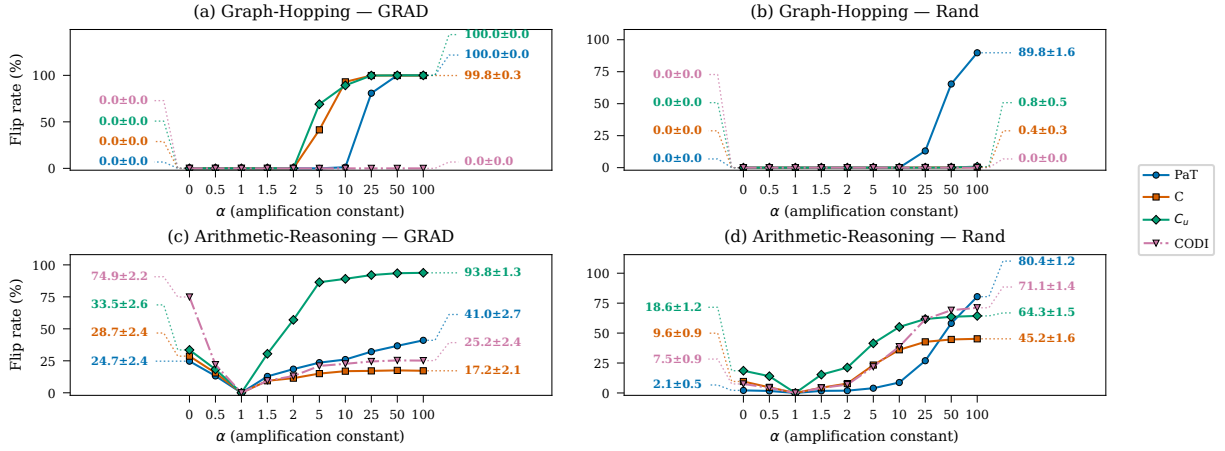
Figure 14: Per-layer IE_{KL} under partner-prompt corruption across buckets $\{P_{full}, P_{max}, P_b, T_1, \dots, T_K, T_{full}, A_b\}$, decomposed into (a) attention and (b) MLP outputs for the Llama-3.2 family. Unlike the full residual stream (Figure 7), thought positions T_t on **graph-hopping** carry a *small but nonzero* causal effect in both attention and MLP output decompositions. **Arithmetic-reasoning** mirrors the full-residual pattern. *Takeaway*: Latent-thought influence is graded rather than binary—even where it appears negligible, thought positions exert a measurable causal effect.

Task	Model	k_0	k_1	k_2	k_3	k_4	k_5	k_6	$\bar{k}$	Adj. $\cos^2$	Off-diag. $\cos^2$	$\ h^c\ /\ h\ $
Graph-Hopping	PaT	1	1	1	1	1	1	2	1.1	0.582	0.304 ± 0.033	0.019
	C	2	2	2	2	2	2	—	2.0	0.863	0.469 ± 0.020	0.022
	C_u	1	1	1	1	1	1	—	1.0	0.901	0.516 ± 0.035	0.005
	CODI	2	2	2	2	2	2	—	2.0	0.623	0.306 ± 0.026	0.018
Arithmetic-Reasoning	PaT	1	24	11	5	3	3	3	7.1	0.870	0.790 ± 0.034	0.117
	C	397	345	291	261	251	270	—	302.5	0.934	0.590 ± 0.004	0.574
	C_u	259	357	386	305	268	250	—	304.2	0.581	0.354 ± 0.006	0.385
	CODI	185	159	150	214	167	131	—	167.7	0.423	0.310 ± 0.010	0.554

Table 6: Gradient-subspace diagnostics across tasks for the Llama-3.2 model family.



(a) GPT-2 model-family, predicted-token subspace.



(b) Llama-3.2 model-family, predicted-token subspace.

Figure 15: Gradient-subspace intervention flip rates (%) across amplification strength α on the gradient-derived (predicted-token) vs random control directions (§ 5). **Graph-Hopping:** Robust to ablation ($\alpha = 0$) but flips at high- α on GRAD. **Arithmetic-Reasoning:** Degradation under GRAD ablation; GRAD and Rand amplifications cause comparable flip-rates. **Takeaway:** Thought utilization is *graded* and depends on their influence power over model behavior. Predicted-token subspaces replicate the interventional patterns of the gold subspace.

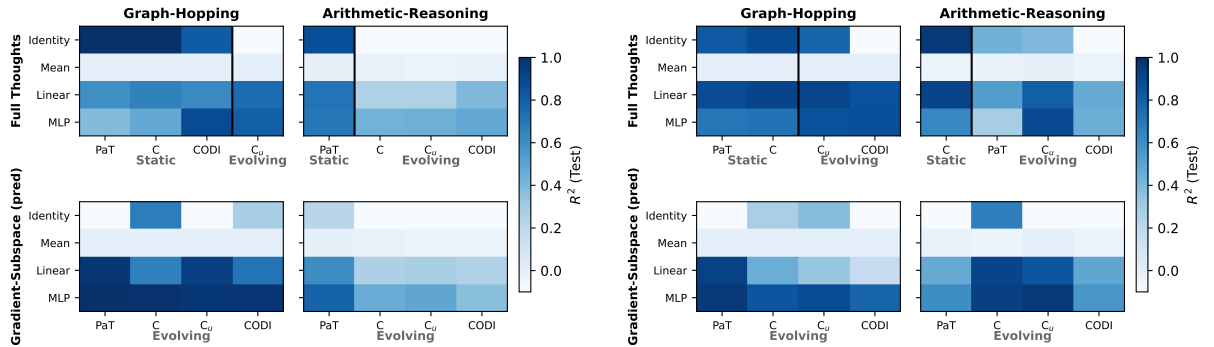
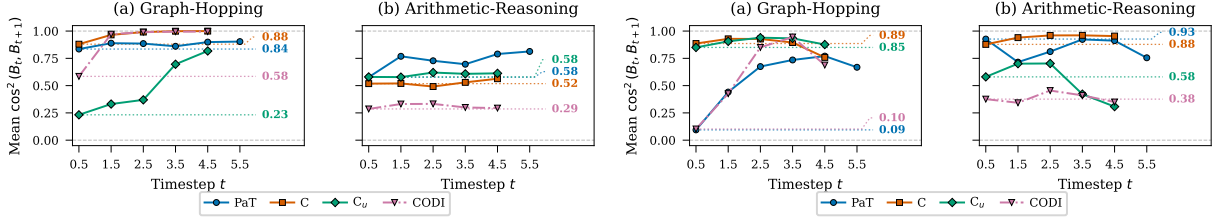


Figure 16: Markovianity of latent thoughts (§ 6). (a) **GPT-2. Graph-Hopping:** Static dynamics (except C_u); **Arithmetic-Reasoning:** Evolving dynamics. (b) **Llama-3.2. Graph-Hopping:** Mixed dynamics; **Arithmetic-Reasoning:** Evolving dynamics (except C). **Takeaway:** In both families, static-appearing full-thoughts show evolving dynamics within the predicted-token gradient-subspace where influence concentrates. Predicted-token subspaces replicate the geometric patterns of the gold subspace.



(a) GPT-2 model-family, predicted-token subspace.

(b) Llama-3.2 model-family, predicted-token subspace.

Figure 17: Geometric stability of the gradient-subspaces (predicted-token) (§ 6). (a) **GPT-2. Graph-Hopping:** Stable subspaces (except C_u); **Arithmetic-Reasoning:** Rotating subspaces. (b) **Llama-3.2. Graph-Hopping:** C and C_u stable, PaT and CODI show initial rotations; **Arithmetic-Reasoning:** PaT and C stable, C_u and CODI rotating. **Takeaway:** The weak-influence→stable-geometry dichotomy does not transfer from GPT-2 to Llama-3.2 where stability does not cleanly separate by task. Predicted-token subspaces replicate the geometric patterns of the gold subspace.

Task	Model	k_0	k_1	k_2	k_3	k_4	k_5	k_6	$\bar{k}$	Adj. $\overline{\cos^2}$	Off-diag. $\overline{\cos^2}$	$\ h^c\ /\ h\ $
Graph-Hopping	PaT	15	16	16	16	17	17	17	16.3	0.871	0.697 ± 0.026	0.353
	C	13	12	11	11	12	12	—	11.8	0.972	0.671 ± 0.024	0.026
	C_u	15	13	17	17	17	14	—	15.5	0.486	0.197 ± 0.012	0.045
	CODI	20	33	35	36	36	37	—	32.8	0.898	0.587 ± 0.008	0.100
Arithmetic-Reasoning	PaT	160	145	146	109	122	156	153	141.6	0.711	0.623 ± 0.019	0.378
	C	155	100	129	135	128	115	—	127.0	0.512	0.394 ± 0.021	0.369
	C_u	126	114	115	122	119	134	—	121.7	0.574	0.404 ± 0.014	0.297
	CODI	51	14	64	18	57	13	—	36.2	0.271	0.351 ± 0.016	0.306

Table 7: Gradient-subspace (predicted-token) diagnostics across tasks for the GPT-2 model family.

Task	Model	k_0	k_1	k_2	k_3	k_4	k_5	k_6	$\bar{k}$	Adj. $\overline{\cos^2}$	Off-diag. $\overline{\cos^2}$	$\ h^c\ /\ h\ $
Graph-Hopping	PaT	4	4	4	4	4	4	4	4.0	0.564	0.269 ± 0.032	0.039
	C	1	1	1	1	1	1	—	1.0	0.879	0.488 ± 0.020	0.012
	C_u	1	1	1	1	1	1	—	1.0	0.901	0.516 ± 0.035	0.005
	CODI	2	2	2	2	2	2	—	2.0	0.603	0.288 ± 0.026	0.015
Arithmetic-Reasoning	PaT	2	16	8	6	3	5	5	6.4	0.841	0.711 ± 0.054	0.117
	C	412	347	288	256	247	267	—	302.8	0.938	0.596 ± 0.004	0.558
	C_u	250	354	354	270	231	248	—	284.5	0.543	0.335 ± 0.008	0.368
	CODI	153	144	128	176	140	101	—	140.3	0.387	0.285 ± 0.014	0.527

Table 8: Gradient-subspace (predicted-token) diagnostics across tasks for the Llama-3.2 model family.

to the gold-subspace.

Gold-Predicted Token Gradient Subspace Alignment.

Tables 9 (GPT-2 model family) & 10 (Llama-3.2 model family) report the alignment between the gold and predicted-token gradient subspaces. The subspaces remain in alignment far above the random control directions.

Variance Decomposition of Latent-Thoughts.

We report the static organization of latent-thoughts by decomposing their total variance $h_{i,t} \in \mathbb{R}^D$ (instance i , timestep t) into timestep-specific, instance-specific, and residual components using $\mu = \mathbb{E}_{i,t}[h_{i,t}]$ (global mean), $\mu_t = \mathbb{E}_i[h_{i,t}]$ (temporal-

mean) and $\mu_i = \mathbb{E}_t[h_{i,t}]$ (instance-mean) as:

$$\begin{aligned} \text{Var}(h_{i,t}) = & \underbrace{\mathbb{E}_t \|\mu_t - \mu\|^2}_{\text{Var}_{\text{time}}} + \underbrace{\mathbb{E}_i \|\mu_i - \mu\|^2}_{\text{Var}_{\text{inst}}} \\ & + \underbrace{\mathbb{E}_{i,t} \|h_{i,t} - \mu_i - \mu_t + \mu\|^2}_{\text{Var}_{\text{residual}}}. \end{aligned}$$

For PaT, t indexes spatial tokens; for C and C_u , it indexes recurrence-steps. Results in Table 11 (GPT-2) and Table 12 (Llama-3.2).

GPT-2: On **graph-hopping**, PaT and C are Var_{inst} dominated, whereas C_u is Var_{time} dominated. Figure 18 shows the two-dimensional PCA projections of the GPT-2 model family latents thoughts across tasks. In Figure 18c, we see C_u forming nearly perfectly separable clusters corresponding

Task	Model	t_0	t_1	t_2	t_3	t_4	t_5	t_6	$\overline{\cos^2 \theta}$	Control $\overline{\cos^2 \theta}$
Graph-Hopping	PaT	0.620	0.700	0.708	0.656	0.768	0.709	0.663	0.689	0.021
	C	0.611	0.563	0.529	0.566	0.530	0.566	–	0.561	0.013
	C_u	0.536	0.534	0.593	0.554	0.541	0.639	–	0.566	0.019
	CODI	0.802	0.718	0.725	0.721	0.724	0.722	–	0.735	0.039
Arithmetic-Reasoning	PaT	0.702	0.694	0.678	0.668	0.677	0.662	0.670	0.679	0.194
	C	0.676	0.679	0.669	0.669	0.722	0.651	–	0.678	0.153
	C_u	0.720	0.703	0.691	0.700	0.713	0.679	–	0.701	0.153
	CODI	0.796	0.903	0.838	0.931	0.857	0.952	–	0.880	0.045

Table 9: Gold-label vs. predicted-token subspace alignment across tasks for the GPT-2 model family. Each t_i column reports $\overline{\cos^2 \theta}$ between B_t^{gold} and B_t^{pred} at that timestep (1 = same subspace, 0 = orthogonal); the $\overline{\cos^2 \theta}$ column is the mean over t , and the control column is the same statistic for rank-matched random subspaces.

Task	Model	t_0	t_1	t_2	t_3	t_4	t_5	t_6	$\overline{\cos^2 \theta}$	Control $\overline{\cos^2 \theta}$
Graph-Hopping	PaT	0.999	0.999	0.982	0.974	0.990	0.995	0.999	0.991	0.002
	C	1.000	1.000	1.000	1.000	1.000	0.999	–	1.000	0.001
	C_u	1.000	1.000	0.999	1.000	1.000	1.000	–	1.000	0.000
	CODI	0.510	0.503	0.509	0.504	0.502	0.500	–	0.505	0.001
Arithmetic-Reasoning	PaT	1.000	0.784	0.802	0.958	0.989	0.954	0.988	0.925	0.004
	C	0.532	0.552	0.590	0.629	0.642	0.625	–	0.595	0.128
	C_u	0.405	0.414	0.411	0.369	0.310	0.311	–	0.370	0.127
	CODI	0.593	0.546	0.621	0.621	0.556	0.648	–	0.598	0.070

Table 10: Gold-label vs. predicted-token subspace alignment across tasks for the Llama-3.2 model family. Each t_i column reports $\overline{\cos^2 \theta}$ between B_t^{gold} and B_t^{pred} at that timestep (1 = same subspace, 0 = orthogonal); the $\overline{\cos^2 \theta}$ column is the mean over t , and the control column is the same statistic for rank-matched random subspaces.

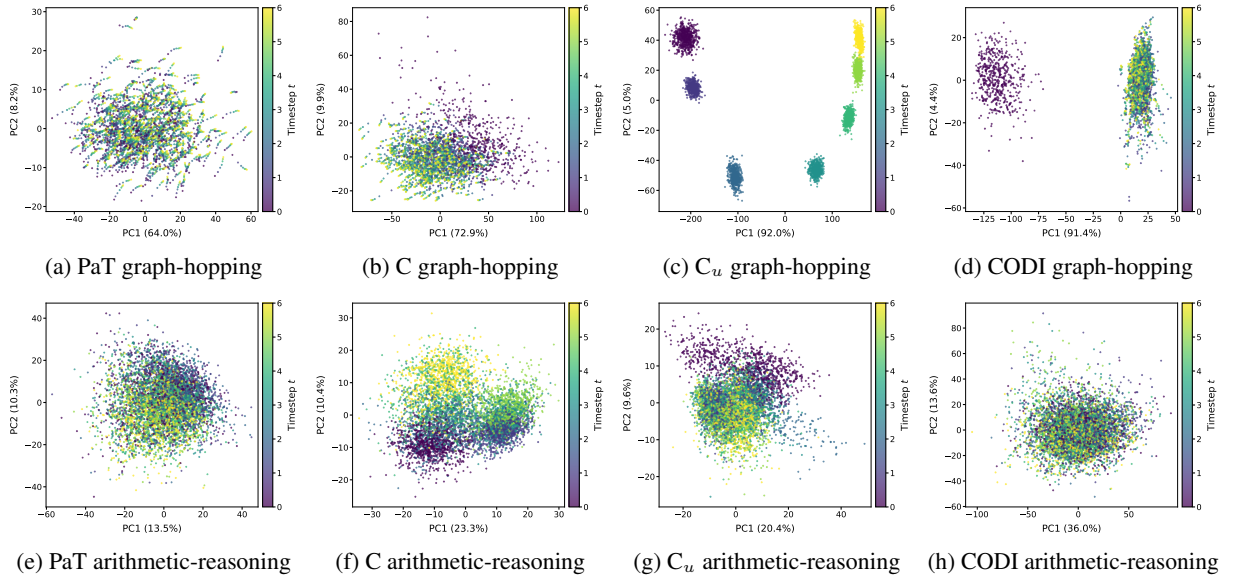


Figure 18: 2D PCA projections of the latent space structure for the GPT-2 model family across the graph-hopping and arithmetic-reasoning tasks.

to recurrence depth. For **arithmetic-reasoning**, all recurrence-based models are $\text{Var}_{\text{residual}}$ dominated (implicit in Table 11 as the remaining variance). This is with the exception of PaT, which remains Var_{inst} dominated with a temporal component that is 1-2 orders of magnitude smaller than the recurrence-based models.

Llama-3.2: On **graph-hopping**, PaT and C are Var_{inst} dominated, whereas C_u and CODI are Var_{time} dominated. For **arithmetic-reasoning**, C and C_u are heavily Var_{inst} dominated. Figure 19 shows the two-dimensional PCA projections of the Llama-3.2 model family latents thoughts across tasks. Figure 19c shows the model forming per-

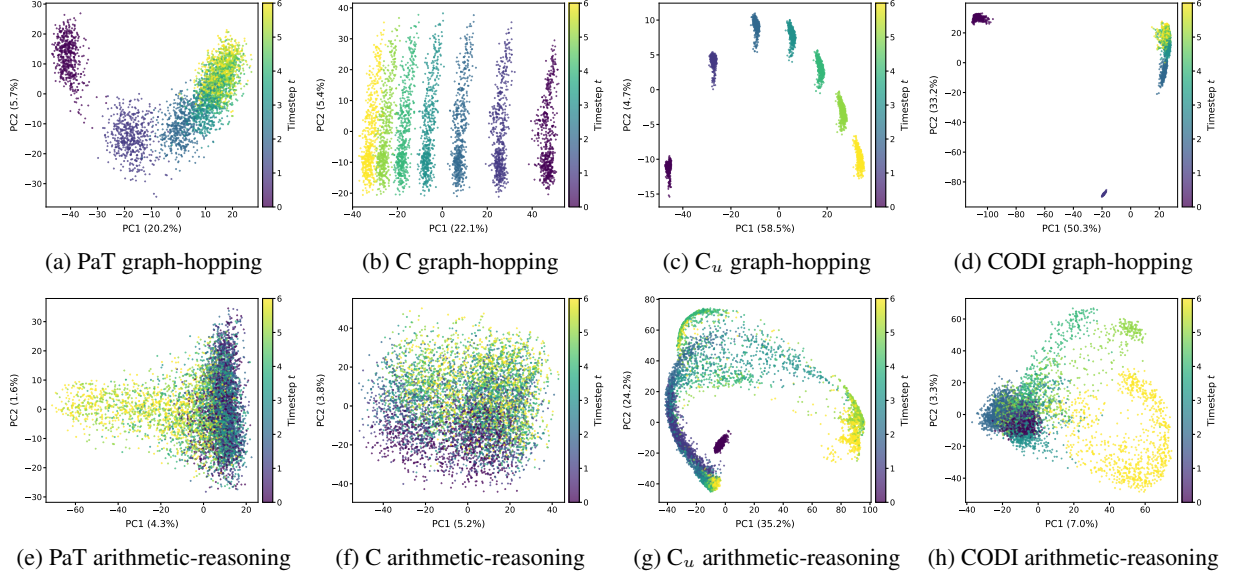


Figure 19: 2D PCA projections of the latent space structure for the Llama-3.2 model family across the graph-hopping and arithmetic-reasoning tasks.

	Graph-Hopping		Arithmetic-Reasoning	
Model	Var _{time}	Var _{inst}	Var _{time}	Var _{inst}
PaT	0.6±0.1	98.7±0.1	1.9±0.2	72.6±0.8
C	22.6±1.9	68.9±2.1	25.0±0.4	26.0±0.6
C _u	99.1±0.0	0.4±0.0	10.8±0.6	26.1±0.5
CODI	89.8±0.7	7.4±0.6	5.6±0.3	46.3±1.1

Table 11: Variance decomposition of latent thoughts (GPT-2 model family) into temporal, instance, and residual components. **Takeaway:** Lack of generalizable structure suggests that observable static decomposition is uninformative by itself. PaT carries 1-2 orders smaller Var_{time} than other models across tasks.

	Graph-Hopping		Arithmetic-Reasoning	
Model	Var _{time}	Var _{inst}	Var _{time}	Var _{inst}
PaT	25.1±0.4	64.9±0.4	11.2±0.2	39.2±0.3
C	24.7±0.4	67.1±0.4	6.2±0.1	85.3±0.1
C _u	63.4±0.5	32.9±0.5	6.1±0.1	85.6±0.1
CODI	89.0±0.5	6.6±0.3	10.7±0.2	30.9±0.3

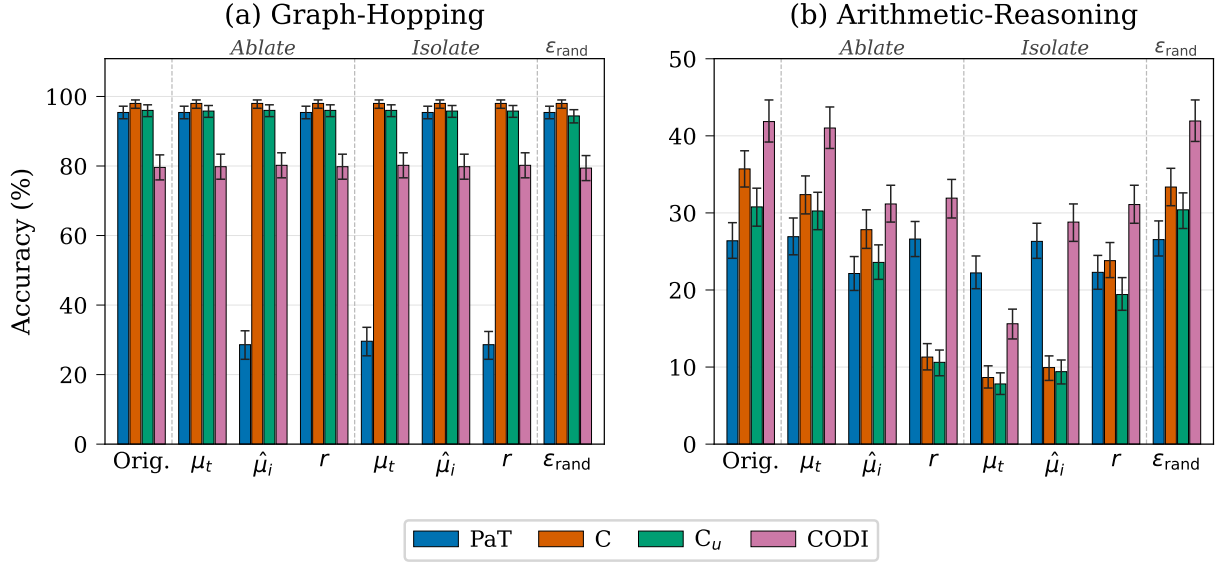
Table 12: Variance decomposition of latent thoughts (Llama-3.2 model family) into temporal, instance, and residual components. **Takeaway:** Lack of generalizable structure suggests that observable static decomposition is uninformative by itself.

fectly separated clusters corresponding to recurrence depth for Llama-3.2 on the graph-hopping task again. In contrast, PaT and CODI are Var_{residual} dominated (implicit in Table 12 as the

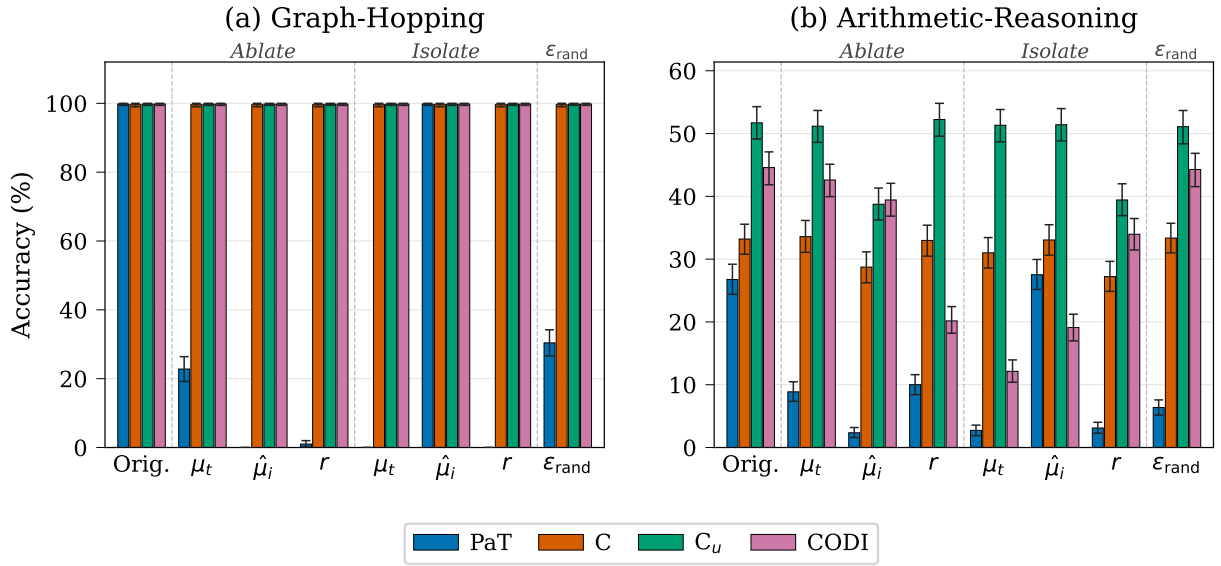
remaining variance). The temporal component (Var_{time}) across all models on arithmetic-reasoning remains uniformly low.

Mean-Ablations and Isolations. We apply six counterfactual interventions to the latent thoughts $h_{i,t}$ at inference to identify component-wise importance. *Ablations* remove one component: temporal ($h'_{i,t} = h_{i,t} - \mu_t + \mu$), instance ($h'_{i,t} = h_{i,t} - \hat{\mu}_i + \mu$, with $\hat{\mu}_i = \frac{1}{K+1} \sum_{k=0}^K h_{i,k}$), or residual ($h'_{i,t} = \mu_t + \hat{\mu}_i - \mu$). *Isolations* retain only one component: $h'_{i,t} = \mu_t$, $h'_{i,t} = \hat{\mu}_i$, or $h'_{i,t} = h_{i,t} - \mu_t - \hat{\mu}_i + 2\mu$, respectively. A *matched-norm random control* adds isotropic noise to $h_{i,t}$ rescaled to the temporal-deviation norm: $h'_{i,t} = h_{i,t} + \varepsilon \cdot \|\mu_t - \mu\| / \|\varepsilon\|$ ($\varepsilon \sim \mathcal{N}(0, I_D)$). To prevent test-set leakage, all reference means (μ, μ_t) are computed from the training split. Results in Figure 20.

GPT-2: On **graph-hopping**, models are robust to perturbations except PaT, which requires μ_i for stability. Notably, neither variance nor gradient-subspace interventions recover C_u’s thought-ablation drop, suggesting that its signal is distributed rather than localized. On **arithmetic reasoning**, PaT shows the same pattern. Ablating μ_t yields the smallest drop while isolating it causes the largest, indicating temporal structure is largely functionally inert—contrasting Wei et al. (2026)’s account of vocabulary-space drift as a key vulnerability. Ablating the residual component produces the largest drop, but its isolation has minimal impact. The matched-norm control preserves per-



(a) GPT-2.



(b) Llama-3.2.

Figure 20: Targeted interventions isolating and ablating specific variance components of the latent thoughts. **Takeaway:** Thoughts carrying low influence are largely unaffected under perturbations. High influence carrying thoughts show varied dependence on components, affected least by the temporal and most by the residual. Temporal structure appears functionally inert across models and tasks.

formance across all tasks.

Llama-3.2: Graph-hopping models are robust to perturbations except PaT, which degrades under all ablations, and all isolations except $\hat{\mu}_i$ which preserves full performance. On **arithmetic reasoning**, all models are affected with varying degree under all ablations and isolations, demonstrating the combined effect played by the components.

D Statistical Evaluation

This section presents the full statistical-tests for every experiment conducted in § 4, § 5 and § 6. Unless stated otherwise, all confidence intervals are 95% percentile bootstrap intervals computed with 1,000 resamples over per-instance outcome vectors, and paired comparisons report exact two-sided McNemar p -values on discordant pairs (b, c) , where b counts instances correct only under condition A and c counts instances correct only under condition B. Significance markers follow the convention $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Where multiple McNemar tests are reported within a single table, we additionally report Benjamini–Hochberg FDR-adjusted q -values (bracketed), with significance stars on q following the same convention.

Epiphenomenal Patterns in LRM Interpretability (§ 4). Tables 13 (GPT-2) and 14 (Llama-3.2) report the results for case study 1 (superposition and BFS-like search on graph-hopping); Table 15 (GPT-2) and 16 (Llama-3.2) report results for case study 2 (scratchpad reasoning on arithmetic-reasoning via logit-lens).

When and How do LRMs use Latent Thoughts (§ 5)? Tables 17 (GPT-2) and 18 (Llama-3.2) report results on the latent thought ablation at test-time experiment; Tables 19 & 20 (prompt and thought-positions respectively for GPT-2), and Tables 21 & 22 (prompt and thought-positions respectively for Llama-3.2) on the causal-tracing experiment; and Tables 23, 24 & 25 (ablation and amplification, gold and predicted-token subspace respectively, for GPT-2), and Tables 26, 27 & 28 (ablation and amplification, gold and predicted-token subspace respectively, for Llama-3.2) on the gradient-subspace interventions experiment.

The Dynamics and Geometry of Latent Thoughts (§ 6). Tables 29 (GPT-2) and 30 (Llama-3.2) report results on the Markovianity of thought trajectories experiment; Tables 5 (gold subspace) & 7 (predicted-token subspace) (GPT-2),

and Tables 6 (gold subspace) & 8 (predicted-token subspace) (Llama-3.2) on the geometric stability of gradient-subspaces experiment.

E Models, Controls, and Training Details

In this appendix, we provide details on the dataset, models, and training details.

E.1 Task and Dataset

Tasks. In this work, we use two case studies corresponding to prominent mechanistic claims about latent reasoning models.

Prior work has interpreted latent trajectories as breadth-first-search-like reasoning over graph frontiers in the graph-hopping task (Hao et al., 2025; Zhu et al., 2025). More recent works have investigated this claim more carefully, questioning its validity (Cui et al., 2026; Rizvi-Martel et al., 2026). All these works used ProsQA (Hao et al., 2025) as the benchmark for this investigation, which makes it a natural choice for our study. Each ProsQA instance is a directed acyclic graph of logical relations paired with a query and a gold target node; the dataset comprises 17,886 training, 300 validation, and 500 test instances.

Prior work has interpreted decodable intermediate values as evidence that latent thoughts act as a compressed thought scratchpad in arithmetic reasoning (Shen et al., 2025). More recent works also question the validity of this claim, raising the possibility of shortcut behavior or task-specific heuristics (Cui et al., 2026; Liang and Pan, 2026; Dilgren and Wiegrefe, 2026; Wei et al., 2026). All these works used GSM8k (Cobbe et al., 2021) as one of the benchmarks for this investigation, which makes it a natural choice for our study. Each GSM8k instance is a grade-school arithmetic word problem paired with a step-by-step solution; we use the GSM8K-Aug variant of Deng et al. (2023), which augments each problem with a compressed chain-of-thought decomposition, giving 385,620 training, 500 validation, and 1,319 test instances.

Together, the two tasks (graph traversal and arithmetic reasoning) allow us to evaluate whether observational patterns of latent reasoning are robust across qualitatively different reasoning settings.

Training data. All models are trained separately for each task. Following Hao et al. (2025) and Shen et al. (2025), we train on the GSM8K-Aug training data of Deng et al. (2023) for GSM8k and on the original ProsQA training split (Hao et al., 2025) for

k	$H / \log_2 N$						Top-1 correct (%)					
	B	CoT	PaT	C	C_u	CODI	B	CoT	PaT	C	C_u	CODI
0	0.58±0.03	0.16±0.02	0.51±0.03	0.43±0.03	0.44±0.03	0.34±0.03	81.3±3.7	84.2±3.5	65.1±4.2	63.8±4.1	90.7±2.5	64.4±4.6
1	0.33±0.03	0.15±0.02	0.51±0.03	0.47±0.03	0.46±0.03	0.33±0.03	68.8±4.1	83.9±3.6	64.9±4.4	65.1±4.1	90.7±2.5	64.4±4.4
2	0.37±0.02	0.24±0.02	0.48±0.02	0.48±0.02	0.32±0.02	0.32±0.02	55.6±4.1	70.8±4.1	71.6±4.0	63.2±4.1	89.1±2.8	51.0±4.3
3	0.39±0.03	0.26±0.02	0.23±0.03	0.26±0.03	0.37±0.03	0.22±0.03	51.3±4.7	65.1±4.6	88.6±3.2	85.2±3.4	86.6±3.3	66.7±4.3
4	0.40±0.05	0.27±0.04	0.13±0.03	0.14±0.04	0.38±0.05	0.19±0.04	57.7±7.1	70.3±6.3	97.3±2.2	92.9±3.8	70.9±6.9	78.6±6.0

k	$P(\text{correct})$						Cand. mass					
	B	CoT	PaT	C	C_u	CODI	B	CoT	PaT	C	C_u	CODI
0	0.05±0.01	0.36±0.04	0.00±0.00	0.00±0.00	0.84±0.02	0.02±0.01	0.07±0.01	0.40±0.04	0.00±0.00	0.00±0.00	0.95±0.01	0.03±0.01
1	0.00±0.00	0.39±0.04	0.00±0.00	0.00±0.00	0.85±0.02	0.01±0.01	0.01±0.00	0.44±0.04	0.00±0.00	0.00±0.00	0.97±0.01	0.03±0.01
2	0.00±0.00	0.09±0.02	0.01±0.01	0.00±0.00	0.71±0.03	0.01±0.01	0.01±0.00	0.10±0.02	0.01±0.01	0.00±0.00	0.81±0.02	0.03±0.01
3	0.00±0.00	0.04±0.01	0.34±0.04	0.37±0.04	0.38±0.03	0.15±0.03	0.00±0.00	0.04±0.02	0.34±0.04	0.37±0.04	0.40±0.03	0.16±0.03
4	0.00±0.00	0.02±0.02	0.58±0.06	0.63±0.07	0.23±0.05	0.16±0.05	0.00±0.00	0.03±0.02	0.59±0.07	0.63±0.07	0.24±0.05	0.16±0.05

Table 13: Full results for superposition and BFS-probing across timesteps and models with statistical testing for the GPT-2 model family.

k	$H / \log_2 N$						Top-1 correct (%)					
	B	CoT	PaT	C	C_u	CODI	B	CoT	PaT	C	C_u	CODI
0	0.60±0.03	0.28±0.03	0.52±0.03	0.50±0.03	0.45±0.03	0.14±0.02	79.8±3.8	94.1±2.3	66.8±4.0	68.8±4.3	91.1±2.5	95.2±2.0
1	0.41±0.03	0.33±0.03	0.52±0.03	0.49±0.03	0.44±0.03	0.16±0.02	69.2±4.2	88.7±2.8	67.0±4.2	66.8±4.2	91.1±2.5	92.8±2.4
2	0.42±0.02	0.27±0.02	0.46±0.02	0.43±0.02	0.35±0.02	0.19±0.02	66.7±4.2	89.5±2.9	69.5±4.3	71.6±3.9	85.2±3.1	86.8±2.9
3	0.49±0.03	0.22±0.02	0.25±0.03	0.34±0.02	0.28±0.03	0.17±0.02	67.9±4.4	90.7±2.6	84.3±3.3	84.1±3.4	86.3±3.1	86.8±3.2
4	0.50±0.04	0.18±0.04	0.17±0.04	0.27±0.04	0.22±0.04	0.09±0.03	74.2±6.3	96.2±2.5	91.8±4.1	90.7±4.4	91.2±4.1	89.6±4.4

k	$P(\text{correct})$						Cand. mass					
	B	CoT	PaT	C	C_u	CODI	B	CoT	PaT	C	C_u	CODI
0	0.12±0.01	0.24±0.03	0.00±0.00	0.03±0.01	0.54±0.03	0.24±0.03	0.15±0.01	0.25±0.03	0.01±0.00	0.05±0.01	0.60±0.03	0.24±0.03
1	0.00±0.00	0.09±0.02	0.00±0.00	0.03±0.01	0.54±0.03	0.24±0.03	0.00±0.00	0.10±0.02	0.01±0.00	0.05±0.01	0.60±0.03	0.25±0.04
2	0.02±0.00	0.14±0.02	0.02±0.01	0.14±0.02	0.10±0.01	0.12±0.03	0.02±0.00	0.14±0.02	0.03±0.01	0.20±0.02	0.11±0.01	0.12±0.03
3	0.04±0.01	0.14±0.02	0.38±0.04	0.30±0.03	0.13±0.02	0.19±0.03	0.06±0.01	0.15±0.02	0.39±0.04	0.35±0.03	0.13±0.02	0.19±0.03
4	0.03±0.01	0.17±0.04	0.60±0.06	0.39±0.05	0.20±0.04	0.32±0.06	0.04±0.01	0.17±0.04	0.60±0.06	0.41±0.05	0.20±0.04	0.32±0.06

Table 14: Full results for superposition and BFS-probing across timesteps and models with statistical testing for the Llama-3.2 model family.

ProsQA. The held out test splits (500 for ProsQA, 1,319 for GSM8k) are used for all reported results.

E.2 Role of Each Model and Control

Table 31 summarizes the role of each model in our design. The controls are not used only as task-performance baselines: each removes or perturbs one factor that could otherwise make an observable latent-state pattern look like evidence for a latent reasoning mechanism.

Target latent reasoning models. Coconut (Hao et al., 2025) and CODI (Shen et al., 2025) are the target LRMs because they instantiate the two mechanistic claims studied in this work: BFS-like latent search on graph-hopping and scratchpad-like latent arithmetic. Both insert dedicated latent thought positions, but train them differently. Coconut gradually replaces explicit reasoning segments with recurrent latent states through a staged curriculum. CODI compresses textual rationales into latent representations by distillation. Studying both lets us test whether our conclusions are tied to one training recipe or hold across two representative ways of constructing latent thought states.

Pause-as-thought as a recurrence control. PaT is designed to isolate recurrent latent-state feedback. It follows Coconut’s input format and staged curriculum, inspired by pause-token training (Goyal et al., 2024), but replaces feedback from previous hidden states with learned thought-position embeddings. Thus, PaT preserves the thought slots, task formatting, and curriculum pressures that may shape observable representations, while removing the mechanism by which Coconut feeds latent states back into subsequent latent computation. If PaT reproduces Coconut-like patterns, those patterns are not sufficient evidence that recurrence is the cause.

Coconut_u as a curriculum control. Coconut_u preserves Coconut’s recurrent latent architecture and number of thought positions, but perturbs the order in which curriculum stages are sampled. At each training step, the model follows the current stage with probability $1 - u$ and samples a non-current stage with probability u . We set $u = 0.3$ to create a moderate perturbation: the model still trains mostly on the intended stage (70% of updates), keeping the comparison close to Coconut,

Model	Step	Hit Rate	Superpos.	Step Align.
B	0	0.1 (0.0, 0.2)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
	1	1.4 (0.8, 2.1)	0.0 (0.0, 0.0)	0.3 (0.0, 0.6)
	2	0.3 (0.1, 0.6)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
	3	0.3 (0.1, 0.6)	0.0 (0.0, 0.0)	0.1 (0.0, 0.2)
	4	0.2 (0.0, 0.5)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
	5	0.2 (0.0, 0.4)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
	6	0.2 (0.0, 0.4)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
CoT	0	51.6 (49.0, 54.3)	15.4 (13.4, 17.3)	35.2 (32.4, 37.8)
	1	7.1 (5.8, 8.5)	0.6 (0.2, 1.1)	1.7 (1.0, 2.5)
	2	40.6 (38.0, 43.4)	11.7 (10.0, 13.5)	9.7 (8.1, 11.4)
	3	10.3 (8.8, 11.8)	1.3 (0.8, 2.0)	0.6 (0.2, 1.1)
	4	39.4 (36.8, 42.0)	11.5 (9.8, 13.2)	2.0 (1.3, 2.8)
	5	12.1 (10.5, 13.9)	1.7 (1.0, 2.5)	0.2 (0.0, 0.4)
	6	36.2 (33.4, 38.8)	10.5 (8.9, 12.1)	0.0 (0.0, 0.0)
PaT	0	65.4 (62.9, 67.9)	22.6 (20.4, 24.9)	48.3 (45.7, 51.0)
	1	65.6 (63.1, 68.1)	24.1 (22.0, 26.5)	40.7 (37.9, 43.4)
	2	64.3 (61.9, 66.9)	23.8 (21.4, 26.1)	18.7 (16.7, 20.9)
	3	62.2 (59.6, 64.7)	23.1 (20.9, 25.4)	7.9 (6.5, 9.5)
	4	60.0 (57.6, 62.6)	23.4 (21.1, 25.8)	4.2 (3.1, 5.2)
	5	58.9 (56.3, 61.5)	22.8 (20.7, 25.1)	1.3 (0.7, 1.9)
	6	58.6 (55.9, 61.1)	21.9 (19.7, 24.1)	0.2 (0.0, 0.4)
C	0	29.4 (27.0, 32.1)	4.2 (3.1, 5.4)	17.7 (15.7, 19.7)
	1	73.0 (70.6, 75.5)	20.8 (18.7, 23.2)	20.8 (18.7, 23.0)
	2	19.7 (17.6, 21.9)	2.9 (2.1, 3.8)	4.5 (3.5, 5.6)
	3	71.6 (69.3, 74.1)	18.8 (16.5, 20.8)	6.3 (5.0, 7.7)
	4	30.4 (28.0, 32.9)	5.5 (4.3, 6.8)	1.5 (0.8, 2.2)
	5	64.8 (62.3, 67.4)	16.6 (14.5, 18.7)	0.7 (0.2, 1.2)
	6	29.2 (26.7, 31.7)	6.1 (4.8, 7.5)	0.2 (0.0, 0.4)
C _u	0	44.4 (41.8, 47.3)	11.1 (9.5, 12.9)	28.3 (26.0, 30.7)
	1	44.4 (41.8, 47.3)	11.1 (9.5, 12.9)	16.3 (14.2, 18.4)
	2	52.6 (50.0, 55.2)	11.1 (9.5, 12.9)	6.6 (5.3, 8.0)
	3	26.1 (23.7, 28.7)	4.8 (3.6, 5.9)	1.5 (0.9, 2.2)
	4	50.7 (48.0, 53.5)	10.9 (9.3, 12.7)	1.2 (0.6, 1.7)
	5	30.1 (27.7, 32.7)	3.6 (2.7, 4.8)	0.2 (0.0, 0.4)
	6	46.6 (44.0, 49.1)	13.1 (11.3, 14.9)	0.2 (0.0, 0.5)
CODI	0	75.1 (72.8, 77.2)	21.9 (19.7, 24.1)	63.9 (61.3, 66.3)
	1	0.5 (0.2, 0.8)	0.0 (0.0, 0.0)	0.1 (0.0, 0.2)
	2	73.0 (70.6, 75.5)	22.1 (19.8, 24.4)	15.3 (13.4, 17.2)
	3	0.4 (0.1, 0.8)	0.1 (0.0, 0.2)	0.0 (0.0, 0.0)
	4	70.8 (68.5, 73.2)	21.8 (19.6, 23.9)	3.2 (2.2, 4.2)
	5	0.5 (0.2, 0.8)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
	6	70.1 (67.8, 72.6)	22.1 (20.1, 24.4)	0.2 (0.0, 0.4)
Pooled (all steps)				
B	—	0.4 (0.2, 0.5)	0.0 (0.0, 0.0)	—
CoT	—	28.2 (27.3, 29.1)	7.5 (7.0, 8.0)	—
PaT	—	62.1 (61.2, 63.2)	23.1 (22.3, 23.9)	—
C	—	45.4 (44.5, 46.6)	10.7 (10.0, 11.4)	—
C _u	—	42.1 (41.2, 43.1)	9.4 (8.8, 10.0)	—
CODI	—	41.5 (40.4, 42.5)	12.6 (11.8, 13.2)	—

Table 15: Full results for scratchpad-thinking probing across timesteps and models with statistical testing for the GPT-2 model family.

while the remaining updates are frequent enough to test whether the reported patterns depend on a nearly deterministic stage schedule. This control separates effects of recurrence from effects of the curriculum trajectory.

Base GPT-2 & Llama-3.2 and Explicit-CoT GPT-2 & Llama-3.2 as observational controls.

Base and Explicit-CoT controls are included only in the observational analyses of § 4. Base tests whether the readout methods themselves can recover apparently interpretable structure from a model without task-specific latent-thought training. Explicit-CoT tests whether similar structure appears after ordinary explicit-rationale supervision.

Model	Step	Hit Rate	Superpos.	Step Align.
B	0	34.0 (31.4, 36.5)	7.8 (6.3, 9.2)	18.4 (16.2, 20.5)
	1	34.0 (31.4, 36.5)	7.8 (6.3, 9.2)	12.0 (10.3, 13.8)
	2	35.0 (32.5, 37.7)	7.0 (5.6, 8.5)	7.6 (6.1, 9.1)
	3	34.9 (32.3, 37.4)	6.8 (5.5, 8.2)	4.1 (3.0, 5.1)
	4	31.0 (28.5, 33.5)	6.1 (4.8, 7.5)	2.0 (1.3, 2.8)
	5	25.7 (23.4, 28.3)	4.2 (3.2, 5.4)	0.2 (0.0, 0.5)
	6	20.1 (18.0, 22.4)	2.8 (1.9, 3.7)	0.1 (0.0, 0.2)
CoT	0	58.8 (56.2, 61.3)	21.4 (19.3, 23.6)	45.1 (42.4, 47.5)
	1	58.8 (56.2, 61.3)	21.4 (19.3, 23.6)	22.2 (20.1, 24.5)
	2	52.0 (49.3, 54.7)	16.9 (15.0, 19.0)	12.6 (10.8, 14.4)
	3	10.2 (8.5, 11.8)	2.4 (1.5, 3.2)	0.7 (0.3, 1.2)
	4	6.3 (5.0, 7.5)	1.7 (1.0, 2.4)	0.3 (0.1, 0.6)
	5	3.3 (2.4, 4.2)	1.0 (0.5, 1.5)	0.1 (0.0, 0.2)
	6	2.4 (1.6, 3.2)	0.6 (0.2, 1.1)	0.0 (0.0, 0.0)
PaT	0	59.0 (56.3, 61.8)	17.0 (14.8, 19.1)	43.7 (41.0, 46.6)
	1	65.4 (62.9, 67.9)	21.6 (19.5, 24.1)	25.0 (22.7, 27.5)
	2	68.6 (66.3, 71.1)	23.2 (21.1, 25.4)	13.8 (12.0, 15.7)
	3	65.9 (63.3, 68.3)	22.5 (20.2, 24.8)	7.1 (5.8, 8.6)
	4	60.3 (57.6, 63.0)	20.4 (18.1, 22.7)	2.4 (1.5, 3.2)
	5	46.5 (44.0, 49.1)	12.5 (10.8, 14.3)	0.8 (0.3, 1.2)
	6	41.4 (38.8, 44.0)	9.5 (7.9, 11.1)	0.2 (0.0, 0.5)
C	0	75.3 (72.9, 77.6)	32.8 (30.4, 35.2)	54.9 (52.0, 57.6)
	1	75.3 (72.9, 77.6)	32.8 (30.4, 35.2)	38.4 (35.8, 41.1)
	2	74.2 (71.7, 76.5)	32.4 (29.8, 35.0)	18.3 (16.3, 20.4)
	3	73.6 (71.0, 75.9)	30.2 (27.6, 32.8)	7.2 (5.8, 8.5)
	4	71.6 (69.0, 73.9)	27.7 (25.3, 30.3)	3.2 (2.3, 4.2)
	5	69.9 (67.1, 72.3)	24.7 (22.4, 27.1)	0.7 (0.3, 1.2)
	6	67.6 (64.9, 70.0)	22.1 (19.9, 24.4)	0.1 (0.0, 0.2)
C _u	0	10.3 (8.8, 11.9)	0.5 (0.2, 1.0)	4.5 (3.5, 5.8)
	1	10.3 (8.8, 11.9)	0.5 (0.2, 1.0)	3.2 (2.3, 4.2)
	2	13.7 (12.0, 15.6)	1.5 (0.8, 2.2)	2.8 (2.0, 3.7)
	3	12.5 (10.8, 14.4)	1.5 (0.9, 2.2)	1.6 (0.9, 2.3)
	4	10.9 (9.3, 12.6)	1.9 (1.2, 2.7)	0.8 (0.4, 1.4)
	5	3.9 (2.8, 4.9)	0.4 (0.1, 0.8)	0.1 (0.0, 0.2)
	6	1.8 (1.1, 2.5)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
CODI	0	71.8 (69.3, 74.4)	18.4 (16.3, 20.5)	51.8 (49.0, 54.4)
	1	19.8 (17.8, 22.1)	2.7 (1.8, 3.5)	6.9 (5.6, 8.3)
	2	0.8 (0.4, 1.4)	0.0 (0.0, 0.0)	0.2 (0.0, 0.5)
	3	44.3 (41.8, 47.0)	12.0 (10.1, 13.8)	7.4 (6.0, 8.8)
	4	19.5 (17.2, 21.7)	3.2 (2.2, 4.2)	0.7 (0.3, 1.2)
	5	1.8 (1.2, 2.6)	0.2 (0.0, 0.4)	0.0 (0.0, 0.0)
	6	53.3 (50.4, 56.0)	17.9 (15.9, 20.1)	0.1 (0.0, 0.2)
Pooled (all steps)				
B	—	30.7 (29.7, 31.6)	6.1 (5.6, 6.6)	—
CoT	—	27.4 (26.5, 28.3)	9.4 (8.7, 9.9)	—
PaT	—	58.2 (57.2, 59.2)	18.1 (17.3, 18.9)	—
C	—	72.5 (71.6, 73.5)	29.0 (28.0, 29.9)	—
C _u	—	9.0 (8.5, 9.6)	0.9 (0.7, 1.1)	—
CODI	—	30.2 (29.3, 31.1)	7.8 (7.3, 8.3)	—

Table 16: Full results for scratchpad-thinking probing across timesteps and models with statistical testing for the Llama-3.2 model family.

These controls assess the specificity of observational readouts. They are excluded from causal and geometric analyses because those analyses intervene on dedicated latent thought positions, which these models are not trained to use.

E.3 Training Details

Table 32 reports the training configuration for the explicit-CoT baseline and the Coconut-family models: PaT, Coconut, and Coconut_u. Table 33 then reports the CODI configuration. CODI is separately listed since it uses a different training recipe.

For the baseline table, c denotes the number of

Model	Acc _{K_{max}} [% CI]	Acc _{K=0} [% CI]	Δ [% CI]	McNemar <i>p</i>	<i>n</i>
Graph-Hopping					
B	2.4 [1.2, 3.8]	—	—	—	500
CoT	79.0 [75.4, 82.6]	—	—	—	500
PaT	95.4 [93.6, 97.2]	95.6 [93.6, 97.2]	-0.2 [-0.6, 0.0]	1.0000 (b=0, c=1) [q=1.0000]	500
C	98.0 [96.6, 99.0]	97.8 [96.4, 99.0]	+0.2 [0.0, 0.6]	1.0000 (b=1, c=0) [q=1.0000]	500
C _u	96.0 [94.2, 97.6]	91.8 [89.4, 94.0]	+4.2 [1.6, 6.8]	0.0019** (b=32, c=11) [q=0.0030**]	500
CODI	79.6 [76.0, 83.2]	79.8 [76.2, 83.4]	-0.2 [-1.2, 0.8]	1.0000 (b=3, c=4) [q=1.0000]	500
Arithmetic-Reasoning					
B	1.4 [0.8, 2.0]	—	—	—	1319
CoT	41.9 [39.3, 44.6]	—	—	—	1319
PaT	26.4 [24.1, 28.7]	21.4 [19.4, 23.7]	+5.0 [3.5, 6.7]	0.0000*** (b=95, c=29) [q=0.0000***]	1319
C	35.7 [33.4, 38.1]	7.7 [6.4, 9.2]	+28.1 [25.6, 30.5]	0.0000*** (b=391, c=21) [q=0.0000***]	1319
C _u	30.8 [28.3, 33.2]	39.1 [36.4, 41.6]	-8.3 [-10.7, -6.0]	0.0000*** (b=74, c=184) [q=0.0000***]	1319
CODI	41.8 [39.2, 44.7]	24.8 [22.4, 27.1]	+17.1 [14.4, 19.9]	0.0000*** (b=297, c=72) [q=0.0000***]	1319

Table 17: Full results for latent thought ablation at test-time with statistical testing for the GPT-2 model family.

Model	Acc _{K_{max}} [% CI]	Acc _{K=0} [% CI]	Δ [% CI]	McNemar <i>p</i>	<i>n</i>
Graph-Hopping					
B	2.2 [1.0, 3.6]	—	—	—	500
CoT	96.8 [95.2, 98.2]	—	—	—	500
PaT	99.8 [99.4, 100.0]	99.8 [99.4, 100.0]	+0.0 [0.0, 0.0]	1.0000 (b=0, c=0) [q=1.0000]	500
C	99.6 [99.0, 100.0]	99.6 [99.0, 100.0]	+0.0 [0.0, 0.0]	1.0000 (b=0, c=0) [q=1.0000]	500
C _u	99.8 [99.4, 100.0]	99.8 [99.4, 100.0]	+0.0 [0.0, 0.0]	1.0000 (b=0, c=0) [q=1.0000]	500
CODI	99.8 [99.4, 100.0]	99.8 [99.4, 100.0]	+0.0 [0.0, 0.0]	1.0000 (b=0, c=0) [q=1.0000]	500
Arithmetic-Reasoning					
B	7.7 [6.1, 9.1]	—	—	—	1319
CoT	53.3 [50.6, 55.9]	—	—	—	1319
PaT	26.8 [24.4, 29.2]	3.9 [3.0, 4.9]	+22.8 [20.5, 25.4]	0.0000*** (b=315, c=14) [q=0.0000***]	1319
C	33.2 [30.8, 35.6]	12.1 [10.5, 13.9]	+21.2 [18.6, 23.7]	0.0000*** (b=307, c=28) [q=0.0000***]	1319
C _u	51.7 [49.1, 54.3]	49.4 [46.8, 52.1]	+2.4 [0.5, 3.9]	0.0171* (b=95, c=64) [q=0.0342*]	1319
CODI	44.6 [41.8, 47.1]	23.5 [21.2, 25.8]	+21.1 [18.5, 23.6]	0.0000*** (b=317, c=39) [q=0.0000***]	1319

Table 18: Full results for latent thought ablation at test-time with statistical testing for the Llama-3.2 model family.

Model	P _{full}	P _{max}	P _b	A _b	KL _{c→c'}	Wilc. <i>p</i>	McN. <i>p</i> (τ=0.5)	<i>n</i>
Graph-Hopping								
PaT	1.355 [0.977, 1.817]	5.705 [5.008, 6.388]	14.577 [13.697, 15.488]	14.657 [13.791, 15.565]	15.158 [14.280, 16.085]	0.0000***	0.0000*** (0.421) q=0.0000***	500
C	1.554 [1.145, 1.992]	5.315 [4.740, 5.934]	14.204 [13.248, 15.077]	14.143 [13.196, 15.032]	15.320 [14.382, 16.198]	0.0000***	0.0000*** (0.435) q=0.0000***	500
C _u	1.694 [1.201, 2.208]	4.499 [3.953, 5.085]	13.776 [12.953, 14.662]	13.528 [12.741, 14.351]	13.960 [13.122, 14.835]	0.0000***	0.0000*** (0.464) q=0.0000***	500
CODI	0.861 [0.607, 1.167]	5.324 [4.755, 5.911]	11.135 [10.410, 11.857]	11.218 [10.509, 11.945]	11.540 [10.831, 12.255]	0.0000***	0.0000*** (0.455) q=0.0000***	500
Arithmetic-Reasoning								
PaT	5.361 [5.162, 5.572]	6.496 [6.255, 6.743]	10.408 [10.089, 10.730]	10.517 [10.195, 10.847]	11.736 [11.390, 12.085]	0.0000***	0.0000*** (136.423) q=0.0000***	1319
C	4.877 [4.684, 5.071]	5.863 [5.646, 6.070]	7.981 [7.741, 8.213]	8.905 [8.615, 9.198]	10.157 [9.869, 10.447]	0.0000***	0.0000*** (372.136) q=0.0000***	1319
C _u	3.921 [3.737, 4.121]	4.444 [4.259, 4.644]	6.248 [6.029, 6.486]	6.971 [6.722, 7.224]	7.860 [7.594, 8.155]	0.0000***	0.0000*** (340.172) q=0.0000***	1319
CODI	4.194 [3.997, 4.382]	5.919 [5.669, 6.172]	8.599 [8.318, 8.892]	10.424 [10.111, 10.757]	10.542 [10.223, 10.877]	0.0000***	0.0000*** (164.454) q=0.0000***	1319

Table 19: Full causal tracing results with statistical testing (prompt-positions) for the GPT-2 model family.

thought tokens per reasoning step. Effective batch size is computed as batch size per GPU $\times$ gradient accumulation steps $\times$ number of GPUs. Coconut and Coconut_u share the same hyperparameters for

a given dataset; Coconut_u differs only by using stage-mixing probability $u = 0.3$, while standard Coconut uses $u = 0.0$. GPT-2 runs use bf16 disabled and weight decay 0.01; Llama runs use bf16

Model	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T _{full}
Graph-Hopping							
PaT	14.610 [13.735, 15.521]	14.621 [13.747, 15.532]	14.636 [13.767, 15.537]	14.637 [13.768, 15.537]	14.650 [13.783, 15.555]	14.651 [13.786, 15.560]	13.612 [12.721, 14.475]
C	15.305 [14.369, 16.184]	15.307 [14.371, 16.187]	15.287 [14.332, 16.166]	15.288 [14.332, 16.166]	15.310 [14.373, 16.190]	14.203 [13.252, 15.091]	14.205 [13.255, 15.087]
C _u	13.842 [13.021, 14.730]	13.846 [13.023, 14.735]	13.779 [12.973, 14.640]	13.808 [12.995, 14.681]	13.808 [12.995, 14.681]	13.761 [12.936, 14.622]	13.661 [12.854, 14.504]
CODI	11.192 [10.474, 11.918]	11.234 [10.514, 11.963]	11.221 [10.510, 11.940]	11.241 [10.523, 11.965]	11.228 [10.505, 11.964]	11.249 [10.544, 11.987]	11.033 [10.301, 11.750]
Arithmetic-Reasoning							
PaT	10.278 [9.967, 10.596]	10.153 [9.838, 10.472]	10.011 [9.691, 10.325]	9.955 [9.642, 10.270]	10.019 [9.710, 10.333]	10.105 [9.794, 10.422]	6.745 [6.483, 6.998]
C	6.276 [6.061, 6.471]	6.071 [5.871, 6.256]	6.092 [5.886, 6.293]	6.638 [6.413, 6.852]	8.012 [7.754, 8.290]	9.406 [9.112, 9.707]	3.165 [3.006, 3.303]
C _u	5.254 [5.046, 5.469]	5.019 [4.826, 5.222]	5.005 [4.806, 5.212]	5.049 [4.875, 5.252]	5.975 [5.765, 6.210]	7.655 [7.397, 7.942]	2.812 [2.658, 2.986]
CODI	8.105 [7.841, 8.376]	8.324 [8.054, 8.601]	8.193 [7.923, 8.478]	8.241 [7.974, 8.521]	8.317 [8.051, 8.615]	9.895 [9.584, 10.220]	5.449 [5.232, 5.664]

Table 20: Full causal tracing results with statistical testing (thought-positions) for the GPT-2 model family.

Model	P _{full}	P _{max}	P _b	A _b	$\overline{\text{KL}}_{c \rightarrow c'}$	Wilc. <i>p</i>	McN. <i>p</i> ($\tau=0.5$)	<i>n</i>
Graph-Hopping								
PaT	4.305 [3.798, 4.829]	10.939 [10.322, 11.513]	14.356 [13.820, 14.844]	15.731 [15.224, 16.209]	17.452 [16.954, 17.875]	0.0000***	0.0000*** (3,284) q=0.0000***	500
C	2.266 [1.942, 2.609]	6.785 [6.434, 7.142]	9.373 [8.980, 9.754]	11.544 [11.183, 11.896]	13.230 [12.883, 13.590]	0.0000***	0.0000*** (1,348) q=0.0000***	500
C _u	3.135 [2.617, 3.626]	9.265 [8.599, 9.893]	12.714 [12.111, 13.271]	15.831 [15.192, 16.420]	18.117 [17.475, 18.760]	0.0000***	0.0000*** (0,330) q=0.0000***	500
CODI	0.323 [0.179, 0.506]	6.008 [5.512, 6.477]	14.996 [14.328, 15.601]	15.156 [14.496, 15.767]	15.263 [14.617, 15.884]	0.0000***	0.0000*** (0,492) q=0.0000***	500
Arithmetic-Reasoning								
PaT	1.660 [1.490, 1.842]	2.571 [2.352, 2.785]	3.913 [3.652, 4.181]	0.552 [0.467, 0.637]	4.561 [4.284, 4.841]	0.0000***	0.0000*** (48,340) q=0.0000***	1318
C	4.386 [4.225, 4.550]	5.840 [5.649, 6.057]	6.995 [6.803, 7.208]	2.268 [2.156, 2.383]	8.535 [8.320, 8.780]	0.0000***	0.0000*** (9,477) q=0.0000***	1319
C _u	3.757 [3.627, 3.902]	4.216 [4.079, 4.365]	4.977 [4.835, 5.131]	1.425 [1.346, 1.498]	5.435 [5.292, 5.594]	0.0000***	0.0000*** (15,238) q=0.0000***	1319
CODI	3.555 [3.413, 3.690]	4.913 [4.754, 5.078]	5.944 [5.789, 6.105]	7.343 [7.175, 7.522]	7.978 [7.778, 8.178]	0.0000***	0.0000*** (58,646) q=0.0000***	1319

Table 21: Full causal tracing results with statistical testing (prompt-positions) for the Llama-3.2 model family.

Model	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T _{full}
Graph-Hopping							
PaT	14.801 [14.303, 15.268]	15.890 [15.364, 16.362]	16.362 [15.850, 16.836]	16.517 [15.993, 16.971]	16.485 [15.947, 16.931]	16.308 [15.807, 16.770]	12.986 [12.472, 13.470]
C	12.152 [11.845, 12.479]	12.333 [12.026, 12.639]	12.428 [12.108, 12.750]	12.582 [12.245, 12.912]	12.656 [12.303, 12.988]	10.726 [10.349, 11.120]	10.565 [10.200, 10.943]
C _u	15.266 [14.618, 15.933]	17.039 [16.409, 17.691]	17.374 [16.751, 18.009]	17.561 [16.936, 18.184]	17.850 [17.224, 18.478]	14.989 [14.389, 15.583]	13.129 [12.517, 13.711]
CODI	15.107 [14.450, 15.717]	15.156 [14.493, 15.766]	15.154 [14.493, 15.766]	15.173 [14.512, 15.783]	15.156 [14.494, 15.767]	15.197 [14.554, 15.808]	15.107 [14.449, 15.717]
Arithmetic-Reasoning							
PaT	3.751 [3.502, 4.022]	3.703 [3.461, 3.970]	3.524 [3.277, 3.788]	3.437 [3.197, 3.684]	3.910 [3.653, 4.176]	4.086 [3.834, 4.356]	2.413 [2.206, 2.631]
C	7.594 [7.398, 7.811]	7.754 [7.553, 7.986]	7.815 [7.610, 8.052]	7.902 [7.700, 8.148]	8.009 [7.803, 8.232]	7.389 [7.191, 7.622]	6.800 [6.609, 7.017]
C _u	5.184 [5.048, 5.345]	5.187 [5.051, 5.349]	5.206 [5.069, 5.367]	5.210 [5.076, 5.371]	5.252 [5.113, 5.409]	4.804 [4.658, 4.957]	4.700 [4.557, 4.857]
CODI	6.450 [6.283, 6.621]	6.292 [6.121, 6.466]	7.005 [6.836, 7.173]	7.166 [6.980, 7.346]	7.095 [6.920, 7.272]	7.603 [7.422, 7.791]	5.692 [5.519, 5.860]

Table 22: Full causal tracing results with statistical testing (thought-positions) for the Llama-3.2 model family.

and backbone-specific learning rates as reported in the tables.

Table 33 reports the CODI hyperparameters for GPT-2 and Llama-3.2-1B. These models use LoRA fine-tuning with a projection head, following the configuration of Shen et al. (2025). The ProsQA CODI settings follow the configuration of Cui et al. (2026).

Model size. All GPT-2 models are built on the pretrained GPT-2 small (124M parameters; Radford et al., 2019). The Llama models are built on Llama-3.2-1B-Instruct (1.24B parameters; (Grattafiori et al., 2024)). CODI uses LoRA fine-tuning (rank 128) on these backbones; all other models are fully fine-tuned. Full per-model, per-

task hyperparameters appear in Tables 32 and 33.

Model	Acc _{orig} [% CI]	Acc _{grad} [% CI]	Acc _{rand} [% CI]	Δ_{grad} [% CI]	McNemar _{grad} P	Δ_{rand} [% CI]	McNemar _{rand} P
Gold Subspace							
Graph-Hopping							
PaT	95.4 $\pm$ 1.8	95.8 $\pm$ 1.7	95.4 $\pm$ 1.0	-0.4 $\pm$ 0.5	0.5000 (b=0, c=2) [q=0.8000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C	98.0 $\pm$ 1.2	98.0 $\pm$ 1.2	98.0 $\pm$ 0.7	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C _u	96.0 $\pm$ 1.7	96.0 $\pm$ 1.7	96.0 $\pm$ 1.0	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
CODI	80.4 $\pm$ 3.5	79.8 $\pm$ 3.6	80.0 $\pm$ 2.0	+0.6 $\pm$ 0.9	0.3750 (b=4, c=1) [q=0.6667]	+0.4 $\pm$ 0.4	0.1094 (b=8, c=2) [q=0.2501]
Arithmetic-Reasoning							
PaT	26.4 $\pm$ 2.3	23.6 $\pm$ 2.1	26.4 $\pm$ 1.4	+2.8 $\pm$ 1.6	0.0003*** (b=70, c=33) [q=0.0010***]	+0.0 $\pm$ 0.2	1.0000 (b=12, c=12) [q=1.0000]
C	35.7 $\pm$ 2.4	9.2 $\pm$ 1.6	34.6 $\pm$ 1.5	+26.5 $\pm$ 2.4	0.0000*** (b=371, c=22) [q=0.0000***]	+1.1 $\pm$ 0.7	0.0017** (b=121, c=76) [q=0.0045**]
C _u	30.8 $\pm$ 2.5	9.9 $\pm$ 1.6	28.9 $\pm$ 1.4	+20.8 $\pm$ 2.6	0.0000*** (b=325, c=50) [q=0.0000***]	+1.9 $\pm$ 0.9	0.0000*** (b=171, c=95) [q=0.0000***]
CODI	41.8 $\pm$ 2.7	35.7 $\pm$ 2.5	41.6 $\pm$ 1.5	+6.1 $\pm$ 1.7	0.0000*** (b=104, c=23) [q=0.0000***]	+0.2 $\pm$ 0.4	0.3557 (b=42, c=33) [q=0.6667]
Predicted-Token Subspace							
Graph-Hopping							
PaT	95.4 $\pm$ 1.8	95.6 $\pm$ 1.8	95.4 $\pm$ 1.0	-0.2 $\pm$ 0.3	1.0000 (b=0, c=1) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C	98.0 $\pm$ 1.2	98.0 $\pm$ 1.2	98.0 $\pm$ 0.7	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C _u	96.0 $\pm$ 1.7	95.8 $\pm$ 1.8	95.9 $\pm$ 1.0	+0.2 $\pm$ 0.3	1.0000 (b=1, c=0) [q=1.0000]	+0.1 $\pm$ 0.1	1.0000 (b=1, c=0) [q=1.0000]
CODI	80.4 $\pm$ 3.5	80.2 $\pm$ 3.6	80.2 $\pm$ 2.0	+0.2 $\pm$ 0.9	1.0000 (b=3, c=2) [q=1.0000]	+0.2 $\pm$ 0.3	0.4531 (b=5, c=2) [q=0.8055]
Arithmetic-Reasoning							
PaT	26.4 $\pm$ 2.3	23.1 $\pm$ 2.2	26.5 $\pm$ 1.4	+3.3 $\pm$ 1.5	0.0000*** (b=70, c=27) [q=0.0000***]	-0.2 $\pm$ 0.3	0.3269 (b=10, c=16) [q=0.6538]
C	35.2 $\pm$ 2.4	9.9 $\pm$ 1.7	34.2 $\pm$ 1.5	+25.2 $\pm$ 2.6	0.0000*** (b=357, c=24) [q=0.0000***]	+1.0 $\pm$ 0.7	0.0069** (b=119, c=80) [q=0.0184*]
C _u	30.9 $\pm$ 2.3	10.1 $\pm$ 1.7	28.9 $\pm$ 1.4	+20.8 $\pm$ 2.8	0.0000*** (b=329, c=55) [q=0.0000***]	+2.0 $\pm$ 0.8	0.0000*** (b=185, c=106) [q=0.0000***]
CODI	41.8 $\pm$ 2.7	36.9 $\pm$ 2.6	41.5 $\pm$ 1.5	+4.9 $\pm$ 1.7	0.0000*** (b=95, c=30) [q=0.0000***]	+0.3 $\pm$ 0.4	0.2127 (b=45, c=33) [q=0.4862]

Table 23: Full gradient-subspace ablation results with statistical testing for the GPT-2 model family.

Model	α	Flip _{grad} [% CI]	Flip _{rand} [% CI]	Δ [% CI]	McNemar p	n
Graph-Hopping						
PaT	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3182]	
	5	1.0±0.9	0.1±0.1	+0.9±0.5	0.0005*** (b=15, c=1) [q=0.0008***]	
	10	2.0±1.3	0.2±0.2	+1.8±0.8	0.0000*** (b=30, c=3) [q=0.0000***]	
	25	13.0±2.8	1.0±0.5	+12.0±1.7	0.0000*** (b=183, c=3) [q=0.0000***]	
	50	29.6±4.0	1.0±0.5	+28.6±2.4	0.0000*** (b=432, c=3) [q=0.0000***]	
	100	49.4±4.2	1.1±0.5	+48.3±2.7	0.0000*** (b=727, c=3) [q=0.0000***]	
C	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	10	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3182]	
	25	2.4±1.3	0.3±0.3	+2.1±0.7	0.0000*** (b=32, c=1) [q=0.0000***]	
	50	17.2±3.2	0.3±0.2	+16.9±2.0	0.0000*** (b=256, c=2) [q=0.0000***]	
	100	34.4±4.2	1.2±0.5	+33.2±2.4	0.0000*** (b=503, c=5) [q=0.0000***]	
C _u	1.5	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3182]	500
	2	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3182]	
	5	2.0±1.2	39.7±2.5	−37.7±2.6	0.0000*** (b=20, c=585) [q=0.0000***]	
	10	43.6±4.2	100.0±0.0	−56.4±2.3	0.0000*** (b=0, c=846) [q=0.0000***]	
	25	98.8±0.9	100.0±0.0	−1.2±0.5	0.0000*** (b=0, c=18) [q=0.0000***]	
	50	100.0±0.0	100.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	100	100.0±0.0	99.7±0.2	+0.3±0.2	0.1250 (b=4, c=0) [q=0.1750]	
CODI	1.5	0.4±0.5	0.7±0.4	−0.3±0.3	0.1250 (b=1, c=6) [q=0.1750]	500
	2	0.6±0.7	0.7±0.4	−0.1±0.4	1.0000 (b=5, c=6) [q=1.0000]	
	5	0.6±0.7	0.5±0.4	+0.1±0.4	1.0000 (b=6, c=5) [q=1.0000]	
	10	1.2±0.9	0.7±0.4	+0.5±0.4	0.0386* (b=10, c=2) [q=0.0569]	
	25	0.8±0.7	0.7±0.4	+0.1±0.3	1.0000 (b=3, c=2) [q=1.0000]	
	50	0.8±0.7	1.0±0.5	−0.2±0.4	0.5078 (b=3, c=6) [q=0.6182]	
	100	1.2±0.9	0.9±0.5	+0.3±0.4	0.2891 (b=6, c=2) [q=0.3598]	
Arithmetic-Reasoning						
PaT	1.5	9.4±1.7	1.5±0.4	+7.9±0.8	0.0000*** (b=312, c=1) [q=0.0000***]	1319
	2	15.6±2.0	3.8±0.6	+11.8±1.0	0.0000*** (b=479, c=13) [q=0.0000***]	
	5	43.2±2.8	14.1±1.1	+29.2±1.5	0.0000*** (b=1268, c=114) [q=0.0000***]	
	10	52.5±2.7	28.2±1.4	+24.4±1.6	0.0000*** (b=1121, c=157) [q=0.0000***]	
	25	55.0±2.7	47.7±1.6	+7.4±1.2	0.0000*** (b=478, c=186) [q=0.0000***]	
	50	56.3±2.8	48.2±1.6	+8.1±1.3	0.0000*** (b=520, c=200) [q=0.0000***]	
	100	57.4±2.8	48.1±1.5	+9.3±1.3	0.0000*** (b=579, c=210) [q=0.0000***]	
C	1.5	16.1±2.0	10.7±1.0	+5.4±1.1	0.0000*** (b=404, c=190) [q=0.0000***]	1319
	2	25.6±2.4	20.4±1.3	+5.2±1.4	0.0000*** (b=524, c=319) [q=0.0000***]	
	5	48.2±2.7	60.2±1.5	−12.0±1.6	0.0000*** (b=324, c=799) [q=0.0000***]	
	10	65.7±2.5	79.9±1.2	−14.1±1.5	0.0000*** (b=245, c=804) [q=0.0000***]	
	25	82.3±1.9	84.7±1.1	−2.4±1.2	0.0003*** (b=297, c=392) [q=0.0005***]	
	50	84.7±2.0	84.8±1.1	−0.2±1.3	0.8475 (b=335, c=341) [q=0.9888]	
	100	85.7±1.9	85.4±1.1	+0.4±1.3	0.5906 (b=346, c=331) [q=0.7037]	
C _u	1.5	20.7±2.2	16.6±1.2	+4.1±1.3	0.0000*** (b=439, c=278) [q=0.0000***]	1319
	2	31.2±2.5	28.8±1.5	+2.4±1.6	0.0033*** (b=560, c=465) [q=0.0050***]	
	5	57.7±2.6	78.1±1.3	−20.4±1.7	0.0000*** (b=251, c=1057) [q=0.0000***]	
	10	69.2±2.5	93.9±0.7	−24.7±1.5	0.0000*** (b=66, c=1042) [q=0.0000***]	
	25	79.5±2.2	99.2±0.3	−19.8±1.2	0.0000*** (b=14, c=797) [q=0.0000***]	
	50	88.3±1.7	99.6±0.2	−11.3±1.0	0.0000*** (b=9, c=456) [q=0.0000***]	
	100	92.5±1.4	99.7±0.2	−7.2±0.8	0.0000*** (b=7, c=291) [q=0.0000***]	
CODI	1.5	17.7±2.2	7.6±0.8	+10.2±1.1	0.0000*** (b=502, c=99) [q=0.0000***]	1319
	2	24.6±2.3	10.7±1.0	+13.8±1.3	0.0000*** (b=673, c=125) [q=0.0000***]	
	5	34.3±2.5	30.5±1.4	+3.9±1.6	0.0000*** (b=554, c=401) [q=0.0000***]	
	10	37.5±2.6	58.2±1.5	−20.7±1.8	0.0000*** (b=253, c=1074) [q=0.0000***]	
	25	39.2±2.6	90.8±0.9	−51.6±1.6	0.0000*** (b=48, c=2091) [q=0.0000***]	
	50	39.9±2.7	94.0±0.7	−54.1±1.6	0.0000*** (b=41, c=2182) [q=0.0000***]	
	100	40.0±2.7	94.4±0.7	−54.4±1.6	0.0000*** (b=38, c=2192) [q=0.0000***]	

Table 24: Full gradient-subspace amplification results with statistical testing for the GPT-2 model family (gold subspace).

Model	α	Flip _{grad} [% CI]	Flip _{rand} [% CI]	Δ [% CI]	McNemar p	n
Graph-Hopping						
PaT	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3500]	
	5	0.6±0.7	0.0±0.0	+0.6±0.4	0.0039*** (b=9, c=0) [q=0.0062***]	
	10	13.2±2.8	0.2±0.2	+13.0±1.8	0.0000*** (b=198, c=3) [q=0.0000***]	
	25	60.0±4.1	0.9±0.5	+59.1±2.6	0.0000*** (b=890, c=3) [q=0.0000***]	
	50	72.2±4.0	0.9±0.5	+71.3±2.4	0.0000*** (b=1072, c=3) [q=0.0000***]	
	100	78.0±3.4	1.2±0.5	+76.8±2.1	0.0000*** (b=1155, c=3) [q=0.0000***]	
C	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	10	0.0±0.0	0.1±0.2	−0.1±0.2	0.5000 (b=0, c=2) [q=0.6364]	
	25	0.8±0.9	0.5±0.3	+0.3±0.5	0.2668 (b=9, c=4) [q=0.3644]	
	50	12.8±2.8	0.7±0.4	+12.1±1.7	0.0000*** (b=188, c=6) [q=0.0000***]	
	100	48.2±4.4	1.7±0.6	+46.5±2.5	0.0000*** (b=707, c=9) [q=0.0000***]	
C _u	1.5	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3500]	500
	2	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3500]	
	5	1.6±1.0	45.2±2.5	−43.6±2.6	0.0000*** (b=15, c=669) [q=0.0000***]	
	10	60.4±4.3	100.0±0.0	−39.6±2.6	0.0000*** (b=0, c=594) [q=0.0000***]	
	25	98.8±0.9	100.0±0.0	−1.2±0.5	0.0000*** (b=0, c=18) [q=0.0000***]	
	50	100.0±0.0	100.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	100	100.0±0.0	99.9±0.1	+0.1±0.1	1.0000 (b=1, c=0) [q=1.0000]	
CODI	1.5	0.8±0.7	0.7±0.4	+0.1±0.3	1.0000 (b=3, c=2) [q=1.0000]	500
	2	0.6±0.7	0.7±0.4	−0.1±0.3	1.0000 (b=3, c=4) [q=1.0000]	
	5	0.8±0.7	0.4±0.3	+0.4±0.5	0.1460 (b=9, c=3) [q=0.2210]	
	10	1.2±1.0	0.9±0.5	+0.3±0.6	0.3593 (b=12, c=7) [q=0.4791]	
	25	1.2±0.9	0.9±0.5	+0.3±0.5	0.4545 (b=10, c=6) [q=0.5919]	
	50	1.4±1.0	1.3±0.6	+0.1±0.5	1.0000 (b=8, c=7) [q=1.0000]	
	100	1.0±0.9	1.2±0.6	−0.2±0.4	0.5488 (b=4, c=7) [q=0.6830]	
Arithmetic-Reasoning						
PaT	1.5	9.3±1.7	1.8±0.4	+7.5±0.8	0.0000*** (b=298, c=0) [q=0.0000***]	1319
	2	15.2±2.1	4.0±0.6	+11.2±1.0	0.0000*** (b=452, c=8) [q=0.0000***]	
	5	41.9±2.7	14.0±1.0	+27.9±1.5	0.0000*** (b=1212, c=107) [q=0.0000***]	
	10	52.8±2.7	30.3±1.4	+22.5±1.6	0.0000*** (b=1044, c=154) [q=0.0000***]	
	25	55.1±2.8	48.3±1.6	+6.8±1.2	0.0000*** (b=456, c=187) [q=0.0000***]	
	50	56.4±2.8	48.8±1.5	+7.6±1.3	0.0000*** (b=497, c=197) [q=0.0000***]	
	100	57.4±2.7	49.1±1.5	+8.3±1.2	0.0000*** (b=522, c=194) [q=0.0000***]	
C	1.5	15.8±2.0	12.0±1.0	+3.8±1.3	0.0000*** (b=394, c=243) [q=0.0000***]	1319
	2	25.2±2.4	20.4±1.2	+4.8±1.5	0.0000*** (b=532, c=341) [q=0.0000***]	
	5	47.3±2.6	57.2±1.5	−9.9±1.6	0.0000*** (b=362, c=754) [q=0.0000***]	
	10	65.0±2.7	77.8±1.3	−12.8±1.6	0.0000*** (b=268, c=774) [q=0.0000***]	
	25	80.6±2.2	83.6±1.1	−3.0±1.3	0.0000*** (b=314, c=433) [q=0.0000***]	
	50	85.5±1.9	85.3±1.1	+0.3±1.3	0.7315 (b=349, c=339) [q=0.8905]	
	100	86.0±1.9	86.0±1.1	+0.0±1.3	1.0000 (b=341, c=341) [q=1.0000]	
C _u	1.5	21.2±2.3	17.1±1.1	+4.1±1.4	0.0000*** (b=447, c=286) [q=0.0000***]	1319
	2	30.9±2.4	29.4±1.4	+1.6±1.5	0.0503 (b=517, c=455) [q=0.0782]	
	5	57.0±2.7	75.8±1.3	−18.8±1.6	0.0000*** (b=252, c=997) [q=0.0000***]	
	10	70.5±2.4	94.0±0.7	−23.5±1.4	0.0000*** (b=54, c=983) [q=0.0000***]	
	25	81.9±2.0	98.9±0.3	−17.1±1.2	0.0000*** (b=18, c=693) [q=0.0000***]	
	50	87.9±1.7	99.5±0.2	−11.5±1.0	0.0000*** (b=11, c=468) [q=0.0000***]	
	100	90.3±1.6	99.8±0.1	−9.5±0.9	0.0000*** (b=4, c=379) [q=0.0000***]	
CODI	1.5	17.8±2.0	7.7±0.8	+10.1±1.1	0.0000*** (b=489, c=89) [q=0.0000***]	1319
	2	24.4±2.3	11.7±1.0	+12.7±1.3	0.0000*** (b=623, c=120) [q=0.0000***]	
	5	34.3±2.5	32.1±1.4	+2.2±1.5	0.0036** (b=502, c=413) [q=0.0059***]	
	10	37.4±2.5	62.6±1.5	−25.3±1.8	0.0000*** (b=220, c=1220) [q=0.0000***]	
	25	39.3±2.6	94.2±0.7	−54.9±1.7	0.0000*** (b=41, c=2214) [q=0.0000***]	
	50	40.5±2.6	95.9±0.6	−55.4±1.6	0.0000*** (b=34, c=2227) [q=0.0000***]	
	100	40.0±2.7	96.3±0.6	−56.3±1.6	0.0000*** (b=30, c=2258) [q=0.0000***]	

Table 25: Full gradient-subspace amplification results with statistical testing for the GPT-2 model family (predicted-token subspace).

Model	Acc _{orig} [% CI]	Acc _{grad} [% CI]	Acc _{rand} [% CI]	Δ_{grad} [% CI]	McNemar _{grad} p	Δ_{rand} [% CI]	McNemar _{rand} p
Gold Subspace							
Graph-Hopping							
PaT	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	99.8 $\pm$ 0.2	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C	99.6 $\pm$ 0.5	99.6 $\pm$ 0.5	99.6 $\pm$ 0.3	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C _u	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	99.8 $\pm$ 0.2	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
CODI	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	99.8 $\pm$ 0.2	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
Arithmetic-Reasoning							
PaT	26.8 $\pm$ 2.4	25.6 $\pm$ 2.2	26.9 $\pm$ 1.4	+1.1 $\pm$ 1.1	0.0722 (b=38, c=23) [q=0.3849]	−0.2 $\pm$ 0.2	0.1185 (b=4, c=11) [q=0.4739]
C	33.2 $\pm$ 2.4	31.3 $\pm$ 2.5	33.4 $\pm$ 1.5	+1.9 $\pm$ 1.4	0.0144* (b=61, c=36) [q=0.1151]	−0.2 $\pm$ 0.4	0.5154 (b=39, c=46) [q=1.0000]
C _u	51.7 $\pm$ 2.6	50.9 $\pm$ 2.7	51.7 $\pm$ 1.6	+0.8 $\pm$ 1.3	0.2723 (b=47, c=36) [q=0.9681]	+0.1 $\pm$ 0.7	0.9382 (b=84, c=82) [q=1.0000]
CODI	44.4 $\pm$ 2.6	11.3 $\pm$ 1.7	44.6 $\pm$ 1.5	+33.1 $\pm$ 2.7	0.0000*** (b=455, c=18) [q=0.0000***]	−0.2 $\pm$ 0.4	0.3961 (b=30, c=38) [q=1.0000]
Predicted-Token Subspace							
Graph-Hopping							
PaT	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	99.8 $\pm$ 0.2	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C	99.6 $\pm$ 0.5	99.6 $\pm$ 0.5	99.6 $\pm$ 0.3	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
C _u	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	99.8 $\pm$ 0.2	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
CODI	99.8 $\pm$ 0.3	99.8 $\pm$ 0.3	99.8 $\pm$ 0.2	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]	+0.0 $\pm$ 0.0	1.0000 (b=0, c=0) [q=1.0000]
Arithmetic-Reasoning							
PaT	26.8 $\pm$ 2.4	25.6 $\pm$ 2.2	26.8 $\pm$ 1.4	+1.1 $\pm$ 1.1	0.0489* (b=33, c=18) [q=0.3128]	−0.1 $\pm$ 0.2	0.7539 (b=4, c=6) [q=1.0000]
C	33.2 $\pm$ 2.4	30.9 $\pm$ 2.4	33.2 $\pm$ 1.5	+2.4 $\pm$ 1.5	0.0029** (b=67, c=36) [q=0.0312*]	+0.1 $\pm$ 0.5	0.9122 (b=42, c=40) [q=1.0000]
C _u	51.7 $\pm$ 2.6	51.1 $\pm$ 2.8	51.6 $\pm$ 1.5	+0.6 $\pm$ 1.5	0.4797 (b=53, c=45) [q=1.0000]	+0.2 $\pm$ 0.6	0.6871 (b=80, c=74) [q=1.0000]
CODI	44.4 $\pm$ 2.6	12.1 $\pm$ 1.7	44.8 $\pm$ 1.5	+32.3 $\pm$ 2.7	0.0000*** (b=446, c=20) [q=0.0000***]	−0.3 $\pm$ 0.4	0.1048 (b=21, c=34) [q=0.4739]

Table 26: Full gradient-subspace ablation results with statistical testing for the Llama-3.2 model family.

Model	α	Flip _{grad} [% CI]	Flip _{rand} [% CI]	Δ [% CI]	McNemar p	n
Graph-Hopping						
PaT	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	10	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	25	0.8±0.7	0.0±0.0	+0.8±0.4	0.0005*** (b=12, c=0) [q=0.0007***]	
	50	0.0±0.0	0.1±0.1	−0.1±0.1	1.0000 (b=0, c=1) [q=1.0000]	
	100	52.2±4.7	68.3±2.4	−16.1±3.2	0.0000*** (b=192, c=433) [q=0.0000***]	
C	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	50.0±4.6	0.0±0.0	+50.0±2.5	0.0000*** (b=750, c=0) [q=0.0000***]	
	10	62.6±4.4	0.0±0.0	+62.6±2.5	0.0000*** (b=939, c=0) [q=0.0000***]	
	25	75.4±3.6	0.1±0.2	+75.3±2.2	0.0000*** (b=1129, c=0) [q=0.0000***]	
	50	77.4±3.5	0.2±0.2	+77.2±2.2	0.0000*** (b=1158, c=0) [q=0.0000***]	
	100	79.0±3.4	0.3±0.2	+78.7±2.1	0.0000*** (b=1182, c=1) [q=0.0000***]	
C _u	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3500]	
	5	69.0±4.0	0.0±0.0	+69.0±2.4	0.0000*** (b=1035, c=0) [q=0.0000***]	
	10	88.2±2.7	0.0±0.0	+88.2±1.6	0.0000*** (b=1323, c=0) [q=0.0000***]	
	25	99.6±0.5	0.0±0.0	+99.6±0.3	0.0000*** (b=1494, c=0) [q=0.0000***]	
	50	100.0±0.0	0.1±0.1	+99.9±0.1	0.0000*** (b=1499, c=0) [q=0.0000***]	
	100	100.0±0.0	0.8±0.5	+99.2±0.5	0.0000*** (b=1488, c=0) [q=0.0000***]	
CODI	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	10	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	25	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	50	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	100	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
Arithmetic-Reasoning						
PaT	1.5	13.5±1.8	1.4±0.4	+12.1±1.0	0.0000*** (b=504, c=24) [q=0.0000***]	1319
	2	19.2±2.1	1.7±0.4	+17.5±1.2	0.0000*** (b=716, c=25) [q=0.0000***]	
	5	22.5±2.2	5.2±0.7	+17.4±1.3	0.0000*** (b=759, c=72) [q=0.0000***]	
	10	24.4±2.3	11.6±1.0	+12.8±1.4	0.0000*** (b=685, c=177) [q=0.0000***]	
	25	30.0±2.5	29.8±1.4	+0.2±1.6	0.8580 (b=566, c=559) [q=1.0000]	
	50	36.1±2.5	59.9±1.6	−23.8±1.8	0.0000*** (b=342, c=1283) [q=0.0000***]	
	100	49.7±2.8	78.2±1.3	−28.5±1.8	0.0000*** (b=230, c=1358) [q=0.0000***]	
C	1.5	9.4±1.6	4.5±0.6	+4.9±0.9	0.0000*** (b=287, c=92) [q=0.0000***]	1319
	2	11.1±1.7	7.6±0.8	+3.6±1.0	0.0000*** (b=290, c=149) [q=0.0000***]	
	5	14.6±1.8	26.0±1.4	−11.3±1.5	0.0000*** (b=218, c=667) [q=0.0000***]	
	10	16.1±2.0	38.7±1.6	−22.5±1.6	0.0000*** (b=184, c=1075) [q=0.0000***]	
	25	16.3±1.9	44.8±1.5	−28.5±1.6	0.0000*** (b=156, c=1285) [q=0.0000***]	
	50	16.8±2.0	46.9±1.5	−30.1±1.6	0.0000*** (b=153, c=1346) [q=0.0000***]	
	100	17.0±2.0	47.8±1.5	−30.8±1.7	0.0000*** (b=165, c=1385) [q=0.0000***]	
C _u	1.5	26.4±2.4	16.9±1.2	+9.5±1.4	0.0000*** (b=596, c=220) [q=0.0000***]	1319
	2	47.5±2.6	23.4±1.3	+24.2±1.7	0.0000*** (b=1143, c=186) [q=0.0000***]	
	5	81.7±2.0	42.6±1.6	+39.1±1.8	0.0000*** (b=1715, c=168) [q=0.0000***]	
	10	83.9±1.9	54.4±1.6	+29.5±1.7	0.0000*** (b=1362, c=196) [q=0.0000***]	
	25	86.6±1.8	59.8±1.6	+26.8±1.7	0.0000*** (b=1254, c=193) [q=0.0000***]	
	50	88.4±1.7	61.2±1.6	+27.2±1.6	0.0000*** (b=1258, c=181) [q=0.0000***]	
	100	89.2±1.7	61.6±1.6	+27.7±1.6	0.0000*** (b=1263, c=168) [q=0.0000***]	
CODI	1.5	9.0±1.4	4.3±0.6	+4.7±0.9	0.0000*** (b=266, c=79) [q=0.0000***]	1319
	2	13.0±1.7	7.8±0.9	+5.2±1.1	0.0000*** (b=354, c=147) [q=0.0000***]	
	5	20.2±2.0	22.0±1.3	−1.8±1.3	0.0104* (b=338, c=409) [q=0.0149*]	
	10	22.0±2.1	38.0±1.5	−16.0±1.6	0.0000*** (b=203, c=836) [q=0.0000***]	
	25	23.0±2.2	59.1±1.6	−36.1±1.7	0.0000*** (b=150, c=1578) [q=0.0000***]	
	50	23.4±2.1	66.5±1.5	−43.1±1.8	0.0000*** (b=132, c=1836) [q=0.0000***]	
	100	23.4±2.2	69.6±1.4	−46.2±1.8	0.0000*** (b=118, c=1946) [q=0.0000***]	

Table 27: Full gradient-subspace amplification results with statistical testing for the Llama-3.2 model family (gold subspace).

Model	α	Flip _{grad} [% CI]	Flip _{rand} [% CI]	Δ [% CI]	McNemar p	n
Graph-Hopping						
PaT	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	10	1.2±0.9	0.0±0.0	+1.2±0.5	0.0000*** (b=18, c=0) [q=0.0000***]	
	25	80.8±3.4	13.1±1.6	+67.7±2.6	0.0000*** (b=1046, c=30) [q=0.0000***]	
	50	100.0±0.0	65.4±2.4	+34.6±2.4	0.0000*** (b=519, c=0) [q=0.0000***]	
100	100.0±0.0	89.8±1.6	+10.2±1.6	0.0000*** (b=153, c=0) [q=0.0000***]		
C	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	41.4±4.2	0.0±0.0	+41.4±2.5	0.0000*** (b=621, c=0) [q=0.0000***]	
	10	93.0±2.3	0.0±0.0	+93.0±1.2	0.0000*** (b=1395, c=0) [q=0.0000***]	
	25	99.8±0.3	0.1±0.2	+99.7±0.3	0.0000*** (b=1495, c=0) [q=0.0000***]	
	50	99.8±0.3	0.1±0.2	+99.7±0.3	0.0000*** (b=1495, c=0) [q=0.0000***]	
100	99.8±0.3	0.4±0.3	+99.4±0.4	0.0000*** (b=1491, c=0) [q=0.0000***]		
C _u	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.2±0.3	0.0±0.0	+0.2±0.2	0.2500 (b=3, c=0) [q=0.3256]	
	5	69.0±4.0	0.0±0.0	+69.0±2.4	0.0000*** (b=1035, c=0) [q=0.0000***]	
	10	89.2±2.7	0.0±0.0	+89.2±1.6	0.0000*** (b=1338, c=0) [q=0.0000***]	
	25	99.8±0.3	0.0±0.0	+99.8±0.2	0.0000*** (b=1497, c=0) [q=0.0000***]	
	50	100.0±0.0	0.1±0.1	+99.9±0.1	0.0000*** (b=1499, c=0) [q=0.0000***]	
100	100.0±0.0	0.8±0.5	+99.2±0.5	0.0000*** (b=1488, c=0) [q=0.0000***]		
CODI	1.5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	500
	2	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	5	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	10	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	25	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
	50	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]	
100	0.0±0.0	0.0±0.0	+0.0±0.0	1.0000 (b=0, c=0) [q=1.0000]		
Arithmetic-Reasoning						
PaT	1.5	12.8±1.7	1.7±0.4	+11.1±1.0	0.0000*** (b=465, c=27) [q=0.0000***]	1319
	2	18.5±2.0	1.9±0.4	+16.6±1.1	0.0000*** (b=680, c=24) [q=0.0000***]	
	5	23.6±2.2	4.0±0.6	+19.6±1.2	0.0000*** (b=808, c=32) [q=0.0000***]	
	10	26.0±2.3	8.7±0.9	+17.3±1.4	0.0000*** (b=779, c=95) [q=0.0000***]	
	25	32.1±2.5	26.9±1.3	+5.2±1.6	0.0000*** (b=660, c=453) [q=0.0000***]	
	50	36.7±2.6	58.1±1.6	-21.4±1.8	0.0000*** (b=357, c=1204) [q=0.0000***]	
100	41.0±2.7	80.4±1.2	-39.4±1.8	0.0000*** (b=169, c=1728) [q=0.0000***]		
C	1.5	9.4±1.6	4.4±0.6	+5.0±0.9	0.0000*** (b=278, c=80) [q=0.0000***]	1319
	2	11.4±1.7	7.8±0.8	+3.6±0.9	0.0000*** (b=255, c=112) [q=0.0000***]	
	5	15.0±1.9	23.3±1.3	-8.3±1.3	0.0000*** (b=201, c=529) [q=0.0000***]	
	10	16.9±2.0	36.1±1.5	-19.2±1.6	0.0000*** (b=178, c=936) [q=0.0000***]	
	25	17.1±2.0	42.7±1.5	-25.5±1.7	0.0000*** (b=165, c=1176) [q=0.0000***]	
	50	17.5±2.1	44.7±1.5	-27.2±1.6	0.0000*** (b=160, c=1235) [q=0.0000***]	
100	17.2±2.1	45.2±1.6	-28.0±1.6	0.0000*** (b=147, c=1255) [q=0.0000***]		
C _u	1.5	30.5±2.5	15.3±1.1	+15.2±1.4	0.0000*** (b=754, c=154) [q=0.0000***]	1319
	2	57.0±2.7	21.2±1.3	+35.8±1.7	0.0000*** (b=1538, c=122) [q=0.0000***]	
	5	86.4±1.7	41.4±1.6	+45.0±1.7	0.0000*** (b=1872, c=92) [q=0.0000***]	
	10	89.1±1.7	55.1±1.6	+34.0±1.7	0.0000*** (b=1478, c=134) [q=0.0000***]	
	25	92.0±1.4	61.7±1.6	+30.3±1.7	0.0000*** (b=1318, c=119) [q=0.0000***]	
	50	93.5±1.3	63.6±1.5	+29.9±1.7	0.0000*** (b=1290, c=108) [q=0.0000***]	
100	93.8±1.3	64.3±1.5	+29.4±1.6	0.0000*** (b=1262, c=97) [q=0.0000***]		
CODI	1.5	9.4±1.4	4.2±0.6	+5.2±1.0	0.0000*** (b=291, c=86) [q=0.0000***]	1319
	2	13.1±1.7	6.9±0.8	+6.2±1.2	0.0000*** (b=373, c=128) [q=0.0000***]	
	5	21.0±2.2	21.9±1.3	-0.9±1.4	0.2355 (b=370, c=404) [q=0.3141]	
	10	22.7±2.2	38.7±1.5	-16.0±1.5	0.0000*** (b=205, c=839) [q=0.0000***]	
	25	24.5±2.3	61.5±1.5	-37.0±1.8	0.0000*** (b=148, c=1611) [q=0.0000***]	
	50	25.3±2.3	69.1±1.5	-43.8±1.8	0.0000*** (b=127, c=1859) [q=0.0000***]	
100	25.2±2.4	71.1±1.4	-46.0±1.8	0.0000*** (b=136, c=1955) [q=0.0000***]		

Table 28: Full gradient-subspace amplification results with statistical testing for the Llama-3.2 model family (predicted-token subspace).

Proj.	Model	Graph-Hopping				Arithmetic-Reasoning			
		Identity	Mean	Linear	MLP	Identity	Mean	Linear	MLP
Full	PaT	0.999 [0.999, 0.999]	-0.003	0.597 [0.595, 0.599]	0.384 [0.365, 0.399]	0.867 [0.861, 0.873]	-0.013	0.716 [0.713, 0.720]	0.697 [0.695, 0.699]
	C	0.992 [0.991, 0.992]	-0.002	0.653 [0.651, 0.655]	0.477 [0.463, 0.488]	-0.658 [-0.670, -0.648]	-0.022	0.256 [0.253, 0.258]	0.428 [0.425, 0.430]
	C _u	-0.646 [-0.677, -0.616]	-0.000	0.745 [0.740, 0.748]	0.795 [0.792, 0.797]	-0.713 [-0.726, -0.701]	-0.032	0.255 [0.251, 0.258]	0.439 [0.436, 0.441]
	CODI	0.813 [0.801, 0.826]	-0.002	0.621 [0.618, 0.624]	0.884 [0.883, 0.885]	-0.683 [-0.693, -0.675]	-0.023	0.395 [0.390, 0.399]	0.480 [0.477, 0.483]
Subspace (Gold)	PaT	-0.466 [-0.485, -0.452]	-0.000	0.971 [0.970, 0.971]	0.996 [0.996, 0.996]	0.264 [0.257, 0.270]	-0.009	0.605 [0.602, 0.609]	0.780 [0.778, 0.783]
	C	0.775 [0.764, 0.786]	-0.001	0.686 [0.681, 0.690]	0.976 [0.975, 0.977]	-0.893 [-0.906, -0.880]	-0.018	0.250 [0.247, 0.253]	0.441 [0.438, 0.444]
	C _u	-0.758 [-0.796, -0.728]	-0.000	0.921 [0.919, 0.923]	0.968 [0.968, 0.969]	-0.932 [-0.945, -0.920]	-0.030	0.281 [0.278, 0.284]	0.475 [0.472, 0.478]
	CODI	0.296 [0.270, 0.322]	-0.003	0.771 [0.766, 0.776]	0.974 [0.973, 0.975]	-1.076 [-1.105, -1.050]	-0.032	0.236 [0.231, 0.240]	0.341 [0.337, 0.346]
Subspace (Pred)	PaT	-0.538 [-0.559, -0.520]	-0.000	0.967 [0.967, 0.968]	0.995 [0.995, 0.995]	0.220 [0.213, 0.226]	-0.009	0.606 [0.603, 0.610]	0.785 [0.783, 0.787]
	C	0.673 [0.649, 0.694]	-0.001	0.657 [0.651, 0.661]	0.984 [0.983, 0.984]	-0.922 [-0.936, -0.910]	-0.018	0.258 [0.255, 0.261]	0.458 [0.455, 0.461]
	C _u	-0.936 [-0.970, -0.911]	-0.000	0.931 [0.930, 0.933]	0.968 [0.967, 0.969]	-0.952 [-0.967, -0.939]	-0.033	0.277 [0.273, 0.280]	0.484 [0.481, 0.488]
	CODI	0.279 [0.245, 0.313]	-0.004	0.716 [0.711, 0.721]	0.973 [0.972, 0.974]	-1.065 [-1.093, -1.039]	-0.030	0.250 [0.245, 0.254]	0.365 [0.361, 0.369]

Table 29: Full results for Markovianity of thought trajectories experiment with statistical testing for the GPT-2 model family.

Proj.	Model	Graph-Hopping				Arithmetic-Reasoning			
		Identity	Mean	Linear	MLP	Identity	Mean	Linear	MLP
Full	PaT	0.826 [0.820, 0.833]	-0.002	0.880 [0.879, 0.881]	0.707 [0.705, 0.708]	0.432 [0.419, 0.445]	-0.023	0.528 [0.521, 0.535]	0.284 [0.282, 0.287]
	C	0.887 [0.885, 0.890]	-0.002	0.914 [0.913, 0.915]	0.724 [0.723, 0.726]	0.965 [0.965, 0.965]	-0.028	0.913 [0.912, 0.913]	0.631 [0.630, 0.632]
	C_u	0.776 [0.771, 0.780]	-0.002	0.902 [0.900, 0.902]	0.852 [0.851, 0.853]	0.398 [0.384, 0.411]	-0.008	0.800 [0.797, 0.803]	0.891 [0.889, 0.892]
	CODI	-2.086 [-2.257, -1.907]	-0.001	0.862 [0.858, 0.866]	0.863 [0.861, 0.864]	-0.488 [-0.496, -0.481]	-0.034	0.465 [0.461, 0.470]	0.449 [0.446, 0.452]
Subspace (Gold)	PaT	-0.760 [-0.788, -0.736]	-0.000	0.865 [0.862, 0.868]	0.958 [0.957, 0.960]	-1.505 [-1.553, -1.458]	-0.017	0.517 [0.510, 0.525]	0.596 [0.591, 0.601]
	C	0.330 [0.308, 0.348]	-0.001	0.540 [0.532, 0.547]	0.855 [0.850, 0.860]	0.661 [0.657, 0.664]	-0.034	0.911 [0.910, 0.911]	0.895 [0.894, 0.895]
	C_u	0.396 [0.371, 0.418]	-0.003	0.336 [0.329, 0.341]	0.880 [0.874, 0.885]	-0.668 [-0.678, -0.660]	-0.003	0.848 [0.846, 0.851]	0.959 [0.959, 0.960]
	CODI	-0.794 [-0.870, -0.723]	-0.001	0.226 [0.223, 0.229]	0.795 [0.788, 0.801]	-0.725 [-0.734, -0.717]	-0.029	0.484 [0.479, 0.489]	0.562 [0.558, 0.565]
Subspace (Pred)	PaT	-0.458 [-0.472, -0.448]	-0.000	0.922 [0.920, 0.923]	0.961 [0.960, 0.962]	-0.952 [-0.988, -0.917]	-0.017	0.465 [0.459, 0.472]	0.604 [0.599, 0.608]
	C	0.273 [0.245, 0.297]	-0.001	0.452 [0.442, 0.461]	0.840 [0.833, 0.847]	0.664 [0.660, 0.668]	-0.034	0.912 [0.911, 0.912]	0.930 [0.929, 0.930]
	C_u	0.377 [0.351, 0.401]	-0.003	0.332 [0.325, 0.338]	0.880 [0.874, 0.885]	-0.799 [-0.813, -0.786]	-0.003	0.840 [0.838, 0.843]	0.957 [0.957, 0.958]
	CODI	-0.696 [-0.767, -0.628]	-0.001	0.172 [0.168, 0.176]	0.777 [0.769, 0.784]	-0.765 [-0.775, -0.757]	-0.031	0.486 [0.481, 0.491]	0.564 [0.561, 0.567]

Table 30: Full results for Markovianity of thought trajectories experiment with statistical testing for the Llama-3.2 model family.

Model	Role	Question addressed
Coconut (C)	Target LRM	Does staged recurrent latent computation exhibit the BFS-like graph-frontier patterns attributed to Coconut?
CODI	Target LRM	Do distilled latent rationale states contain decodable arithmetic intermediates that are also behaviorally relevant?
PaT	Recurrence control	Do the same patterns persist when Coconut’s format and curriculum are kept but recurrent latent-state feedback is removed?
Coconut_u (C_u)	Curriculum control	Do the patterns survive a controlled perturbation of Coconut’s staged curriculum while the recurrent architecture is preserved?
Base (B)	Observational control	Can the probes recover apparent structure without task-specific latent-thought training?
Explicit-CoT (CoT)	Observational control	Can similar patterns arise from ordinary task solving or explicit-rationale supervision rather than latent-thought mechanisms?

Table 31: Role of each model and control in the experimental design.

Dataset	c	Eps./stage	Max stage	u	Resume	Eff. batch	Epochs	LR
<i>GPT-2</i>								
CoT								
GSM8K-Aug	0	1	0	0.0	0	128	25	1×10^{-4}
ProsQA	0	1	0	0.0	0	128	50	1×10^{-4}
PaT								
GSM8K-Aug	2	3	3	0.0	3	128	25	1×10^{-4}
ProsQA	1	5	6	0.0	0	128	50	1×10^{-4}
Coconut								
GSM8K-Aug	2	3	3	0.0/0.3	3	128	25	1×10^{-4}
ProsQA	1	5	6	0.0/0.3	0	256	50	1×10^{-4}
<i>Llama-3.2-1B-Instruct</i>								
CoT								
GSM8K-Aug	0	1	0	0.0	0	128	10	5×10^{-5}
ProsQA	0	1	0	0.0	0	64	50	5×10^{-5}
PaT								
GSM8K-Aug	2	3	3	0.0	3	128	25	5×10^{-5}
ProsQA	1	5	6	0.0	0	64	50	5×10^{-5}
Coconut								
GSM8K-Aug	2	3	3	0.0/0.3	3	128	25	5×10^{-5}
ProsQA	1	5	6	0.0/0.3	0	64/128	50	5×10^{-5}

Table 32: Hyperparameters used to train the GPT-2 and Llama-3.2-1B-Instruct CoT, PaT, Coconut, and Coconut_u models. Effective batch size is computed as $\text{batch_size_training} \times \text{gradient_accumulation_steps} \times N_{\text{GPUS}}$. Coconut rows report both the standard setting ($u = 0.0$) and the stage-mixed Coconut_u setting ($u = 0.3$). For ProsQA Coconut on Llama, effective batch size is 64 at $u = 0.0$ and 128 at $u = 0.3$.

Hyperparameter	GPT-2		Llama-3.2-1B-Instruct	
	GSM8K-Aug	ProsQA	GSM8K-Aug	ProsQA
Model max length	1024	1024	1024	1024
Precision	bf16	bf16	bf16	bf16
CODI loss weight	1.0	1.0	20	20
Include last CoT	False	False	False	False
Number of latents	6	6	6	6
Use projection	True	True	True	True
Projection dimension	768	768	2048	2048
Projection dropout	0.0	0.0	0.0	0.0
Use LoRA	True	True	True	True
LoRA rank	128	128	128	128
LoRA alpha	32	32	32	32
Learning rate	3×10^{-3}	1×10^{-3}	2×10^{-4}	2×10^{-4}
LR scheduler	Cosine	Cosine	Cosine	Cosine
Warmup ratio	0.03	0.03	0.03	0.03
Optimizer	AdamW	AdamW	AdamW	AdamW
Total Batch size	128	128	128	128
Weight decay	0.1	0.01	0.1	0.01
Gradient clipping	2.0	2.0	2.0	2.0
Epochs	40	20	30	20

Table 33: Hyperparameters used to train the GPT-2 and Llama-3.2-1B-Instruct CODI models.