How Does Reasoning Become Implicit? Uncovering the Dynamics of Chain-of-Thought Internalization

Anonymous authors
Paper under double-blind review

Abstract

Latent reasoning replaces explicit chain-of-thought (CoT) with continuous states, yet how this transition reorganizes reasoning remains unclear. We study this internalization process through a unified mechanistic framework spanning *training-time formation* and *inference-time function*, and apply it to CODI and Coconut as two representative latent-reasoning paradigms. We find that internalization yields a strongly positionally asymmetric latent process, including pronounced even–odd organization and training-time reversals of positional bias. Despite this reorganization, latent reasoning retains task-relevant alignment with explicit CoT while following a distinct intermediate trajectory. At inference, latent-state contributions are recurrently interdependent and strongly conditioned on the surrounding trajectory. Finally, the explicit–latent reliability gap grows sharply with reasoning depth, while failed latent trajectories exhibit weaker stepwise semantic ordering and more diffuse causal organization across states. Together, these results show that reasoning internalization is not simply the direct transfer of explicit CoT into a hidden trajectory. The central mechanistic challenge for better reasoning internalization is therefore how task-relevant computation is organized across latent states and coordinated into reliable reasoning.

1 Introduction

Explicit chain-of-thought (CoT) reasoning improves language-model performance on multi-step tasks by generating intermediate reasoning steps (Wei et al., 2022). Latent reasoning methods seek to retain this capability while replacing written reasoning with recurrent continuous states. For instance, Coconut (Hao et al., 2025) introduces latent states through a curriculum that progressively replaces explicit steps, while CODI (Shen et al., 2025) trains a latent pathway through self-distillation from an explicit pathway. Beyond downstream task performance, these paradigms raise a fundamental question: *how is explicit reasoning internally reorganized into latent computation?*

Recent work has begun to characterize the information encoded in latent states and their causal contributions to model behaviors (Dilgren & Wiegreffe, 2026; Liang & Pan, 2026; Goyal et al., 2025; Li et al., 2026; Chang et al., 2026). Yet these observations leave two gaps. First, most analyses characterize a trained latent process without investigating how its internal organization emerges during internalization. Second, decodable information or causal influence at individual states does not by itself establish an organized reasoning process or reliable problem solving. A mechanistic account therefore needs to connect *how latent computation forms* with *how the resulting latent states interact and function at inference time*. This motivates a systematic account that connects the acquisition of internal computational structure to its functional role in problem solving. We study this problem through a unified framework, organized around two complementary mechanistic perspectives: (i) **Formation Dynamics (*training-time*)**: how do models develop and organize the internal mechanisms that enable latent reasoning? (ii) **Functional Efficacy (*inference-time*)**: how do models leverage latent states for problem solving? Through this lens, we ask not only how latent computation is formed and used, but whether the emergence of structured latent states is sufficient for effective reasoning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We instantiate this framework on CODI (Shen et al., 2025) and Coconut Hao et al. (2025), tracing latent computation from its formation during training to its function at inference time. Our analysis yields three major findings. **First, internalization produces a strongly asymmetric latent trajectory rather than uniformly utilizing its latent states.** Task-relevant representation, attention, semantic readout, and causal sensitivity redistribute unevenly across latent positions during training. This organization often exhibits a pronounced even–odd pattern, and in several settings the dominant positional bias reverses over the course of internalization. Latent reasoning therefore develops through substantial positional reorganization rather than the uniform strengthening of an interchangeable sequence of hidden states. **Second, latent reasoning retains task-relevant alignment with explicit CoT while developing a distinct intermediate trajectory.** Readouts learned from explicit representations transfer to latent states, and the two pathways exhibit closely aligned question-directed attention. At the same time, their intermediate representations remain separated and their allocation over reasoning context differs. Thus, internalization can preserve task-relevant relationships with explicit CoT while reorganizing the intermediate computation through which those relationships are realized. **Third, latent failure is accompanied by a diffusion of internal structure despite continued latent computation.** The explicit–latent gap widens sharply with reasoning depth. At the same time, failed trajectories exhibit weaker stepwise semantic ordering, more diffuse positional causal influence, and less selective inter-state relations than successful trajectories. These outcome-conditioned differences suggest that reasoning reliability depends on more than whether latent states contain information or participate causally; how computation is organized across the latent trajectory constitutes a distinct dimension of latent reasoning.

Together, these findings suggest that reasoning internalization does not simply mirror explicit reasoning in latent space. Instead, it yields a positionally asymmetric latent process that remains task-relevantly aligned with explicit reasoning. Reasoning failure is accompanied by a reorganization of this latent process, with weaker semantic ordering and more diffuse causal influence across states. This shifts the mechanistic question from whether useful computation exists in latent states to how that computation is organized across states and coordinated into reliable reasoning. In summary, our contributions are:

- **A unified mechanistic framework for reasoning internalization.** We connect the *formation dynamics* of latent computation during training with its *functional efficacy* at inference, providing a lens for analyzing how latent reasoning is formed, organized, and utilized.
- **A training-time account of asymmetric latent reorganization.** We show that internalization reorganizes computation into a strongly position-dependent latent process, including pronounced even–odd patterns and training-time reversals of positional bias, while preserving task-relevant alignment with explicit CoT despite a distinct intermediate trajectory.
- **A functional account linking latent organization to reasoning outcomes.** We show that latent-state influence is recurrently interdependent and trajectory-conditioned, and that the explicit–latent gap grows with reasoning depth. Failed reasoning is accompanied by weaker stepwise semantic ordering and more diffuse causal organization, identifying the organization of latent computation as a distinct dimension of reasoning efficacy.

2 Analytical Framework and Setup

We study reasoning internalization as a process that unfolds at two complementary levels. At *training time*, we ask how the internal organization of latent computation develops as models are optimized. At *inference time*, we ask how the resulting latent states participate in problem solving.

2.1 Experimental Setup

Latent Reasoning Paradigms. We study CODI (Shen et al., 2025), which learns a latent pathway through shared-weight self-distillation, and Coconut (Hao et al., 2025), which progressively replaces explicit steps through a curriculum learning procedure.

Latent Rollout Formulation. Let θ_s denote model parameters at training stage s, and K_s the corresponding latent budget. Processing the prompted question x initializes a boundary representation

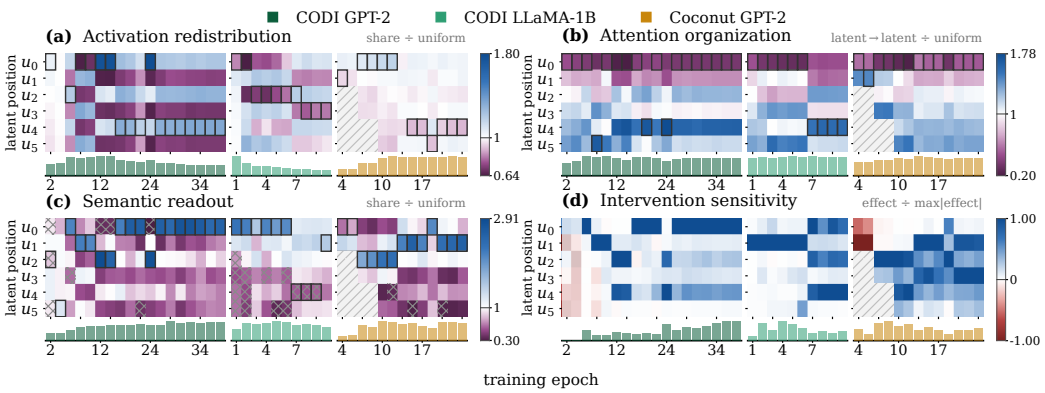


Figure 1: Latent computation reorganizes across internalization process.

$h_{s,0}$ and cached context $C_{s,0}$. The latent rollout follows

$$\begin{aligned} u_{s,t} &= P_{\theta_s}(h_{s,t-1}), \\ (h_{s,t}, C_{s,t}) &= \Phi_{\theta_s}(u_{s,t}; C_{s,t-1}), \qquad t = 1, \ldots, K_s, \end{aligned} \tag{1}$$

where $u_{s,t}$ is the continuous input at latent position t, $h_{s,t}$ its resulting final-layer representation, and Φ_{θ_s} an incremental Transformer step that updates the cached context. The feedback map P_{θ_s} applies a learned projection when enabled in CODI and is the identity in Coconut. The cache retains the question and processed latent positions; answers are decoded from the final context using each method's readout convention.

Models and Datasets. We conduct experiments on GPT-2 (Radford et al., 2019), Llama-3.2-1B-Instruct, and Llama-3.2-3B-Instruct (Grattafiori et al., 2024). We train the models on GSM8K-Aug (Deng et al., 2023) and evaluated on the GSM8K test set (Cobbe et al., 2021). Full training configurations, rollout implementations, datasets, and evaluation protocols are provided in Appendix B.

2.2 A Formation–Function View of Reasoning Internalization

Formation Dynamics studies how latent computation emerges and becomes organized during training. We first characterize how representation, information, and causal sensitivity are redistributed across the latent trajectory; we then examine how this evolving organization relates to the explicit reasoning pathway; finally, we analyze the learning signals that accompany these changes. Operationally, this view combines checkpoint-wise representational and attention analyses, cross-pathway readout comparisons, intervention profiles, and parameter- and state-level gradient analyses. Together, these measurements track not only whether the latent pathway changes, but how its internal organization, relation to explicit reasoning, and learning context co-evolve during internalization.

Functional Efficacy studies how the learned latent computation functions during inference. We first characterize the task-relevant content encoded in latent states, then examine how individual states and their recurrent interactions contribute to downstream computation, and finally relate this internal organization to reasoning outcomes. This view combines semantic readouts, controlled interventions, inter-state analyses, and outcome-conditioned comparisons. Together, these analyses progress from what information latent states carry, to how latent states participate in computation, to how their contributions are organized across the trajectory and how this organization relates to reasoning outcomes.

Detailed definitions of all measures, controls, and intervention protocols are provided in Appendix C.

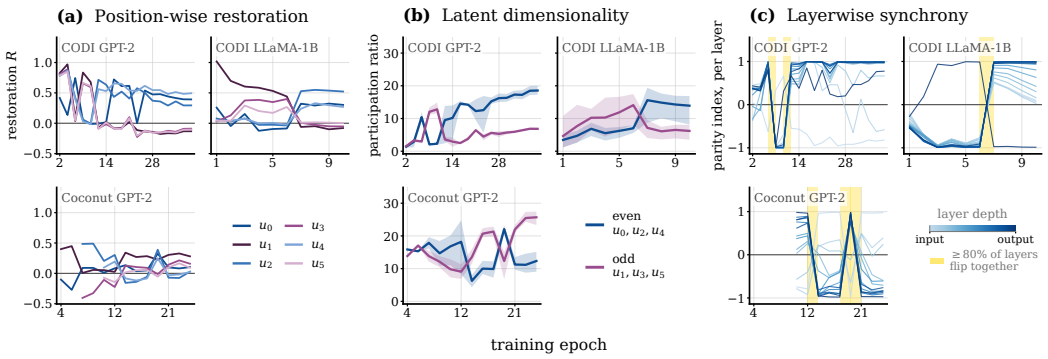


Figure 2: Internalization process differentiates latent states into distinct roles.

3 Formation Dynamics of Latent Reasoning

We now turn to how latent computation develops during internalization. Our analysis proceeds through three questions: **(RQ1)** how does latent computation reorganize over training? **(RQ2)** how is this organization related to the explicit reasoning pathway? **(RQ3)** how do the associated learning dynamics evolve alongside it?

3.1 Reorganization of Latent Computation

We first ask whether internalization merely strengthens a fixed latent organization, or instead redistributes how computation is organized across latent positions. We examine the position-wise organization across training trajectories from representational, semantic, and intervention-based views. We then use restoration and geometric analyses to characterize how these positional differences evolve.

Internalization reshapes the organization of latent computation rather than simply amplifying an existing fixed internal pattern. As shown in Figure 1, across models and training paradigms, the positional organization of the latent process changes substantially over training. Activation mass shifts across latent states, latent-to-latent attention is reallocated, and task-relevant intermediate information becomes concentrated at different positions. Intervention sensitivity changes positionally as well, showing that the transition extends beyond observational structure to where perturbations most strongly affect downstream prediction. Together, these views reveal a redistribution of representation, interaction, and causal sensitivity across latent positions.

This reorganization does not recruit latent positions uniformly; it manifests as structured positional asymmetries across the latent pathway. Figure 1 already shows persistent positional heterogeneity across representation, attention, semantic information, and intervention sensitivity. Figure 2 further reveals that this heterogeneity develops structured positional asymmetries: restoration effects and representational effective dimensionality systematically differentiate latent positions, although the trajectories differ across models and training paradigms. Rather than converging toward uniform use of the available latent budget, computation becomes concentrated on structured subsets of positions, with different regimes exhibiting different positional preferences. These asymmetries are also coordinated across layers: when the even–odd activation bias reverses, the transition occurs within the same time interval across most eligible layers.

Taken together, internalization reorganizes a persistently heterogeneous latent pathway rather than simply strengthening or homogenizing it. This establishes substantial within-pathway reorganization, while leaving open how the evolving latent organization aligns with the explicit reasoning process that guides internalization. We turn to this question next.

3.2 Evolving Alignment Between Explicit and Latent Reasoning

We next ask how the evolving organization of latent computation relates to explicit reasoning. We compare the two pathways in what their intermediate states encode, where along the reasoning trajectory they align, and how they allocate attention, tracking each relation throughout training.

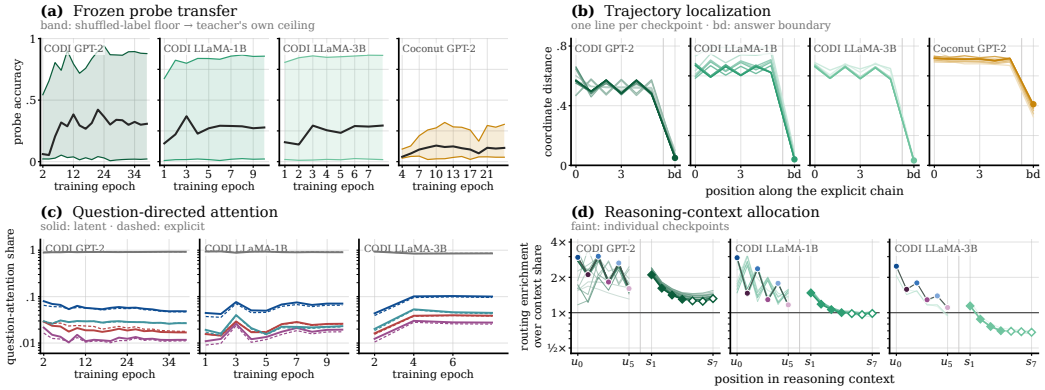


Figure 3: **Explicit–latent alignment is selective across internalization.** Explicit readouts transfer to latent states (a), alignment concentrates near the answer boundary (b), and similar question-directed attention (c) coexists with pathway-specific reasoning-context allocation (d).

Semantic correspondence persists despite substantial intermediate-state separation. A probe trained on explicit representations transfers nontrivially to latent states across training, showing that task-relevant content remains accessible through a shared readout (Figure 3a). Yet this correspondence does not require the intermediate representations themselves to become geometrically close. Along the explicit reasoning trajectory, latent states remain separated from the teacher representation through the intermediate chain, with alignment sharpening markedly only at the answer boundary (Figure 3b). Thus, semantic correspondence can extend beyond the boundary even when the intermediate trajectories remain representationally distinct.

The pathways align in question-directed attention while organizing intermediate context differently. Explicit and latent modes exhibit closely matched attention to question content throughout training (Figure 3c). In contrast, their allocation over intermediate reasoning context differs systematically after accounting for the amount of available context (Figure 3d). The two pathways therefore preserve substantial agreement in how they attend to the shared problem while maintaining pathway-specific organization of intermediate computation.

Taken together, internal reorganization does not preclude correspondence with explicit reasoning. Instead, semantic and attentional continuity coexist with substantial differences in intermediate representation and context organization. We next examine the learning dynamics that accompany this evolving relationship.

3.3 Learning Signals Accompanying Reorganization

Having characterized how latent computation reorganizes and what correspondence with explicit reasoning persists, we next ask what learning-signal structure accompanies this formation process. We distinguish parameter-space update signals, $\partial\mathcal{L}/\partial\theta$, from checkpoint-wise state-space loss sensitivity, $\partial\mathcal{L}/\partial h$.

Internalization reshapes the balance of objective gradients while their parameter updates remain positively aligned on average. Relative to the student objective, the auxiliary gradients remain smaller throughout training, while their relative strengths follow different trajectories across models (Figure 4a). Despite these changes in magnitude, the mean pairwise cosine similarities between objective gradients remain positive across all checkpoints and models (Figure 4b). Their degree of alignment nevertheless changes over training, indicating that what persists is positive mean agreement rather than a fixed alignment strength.

This mean objective alignment coexists with a systematic asymmetry in loss sensitivity across reasoning streams. After normalizing by uniform allocation over reasoning positions, the written chain consistently exhibits greater sensitivity enrichment than the latent slots across models and checkpoints (Figure 4c). Within the latent stream, sensitivity is also position-dependent, and its organization evolves differently across the CODI models and the curriculum-trained Coconut run

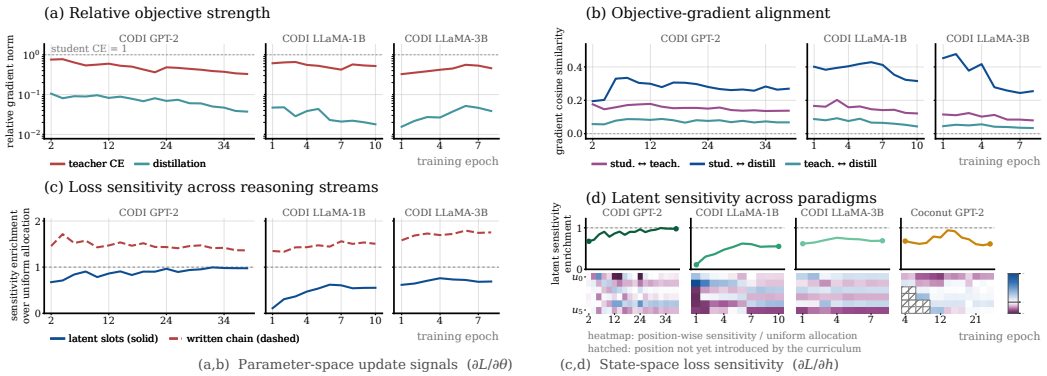


Figure 4: **Learning signals accompanying reorganization:** relative objective strengths change while mean parameter gradients remain positively aligned, whereas state-space loss sensitivity is asymmetric across reasoning streams and varies across internalization paradigms.

(Figure 4d). Thus, aggregate agreement among parameter updates coexists with a structured and persistent difference in how loss sensitivity is distributed across explicit and latent reasoning.

Together, these results distinguish two levels of the learning process: mean agreement among parameter updates can coexist with a persistent asymmetry in how loss sensitivity is distributed across explicit and latent reasoning.

4 Functional Efficacy of Latent Reasoning

In this section, we study how the learned latent process functions at inference time, proceeding through three questions: **(RQ1)** what task-relevant information do latent states encode, and which states exert problem-specific causal influence? **(RQ2)** how does recurrent feedback shape the computational role of these states? **(RQ3)** when does causal participation support reliable reasoning, and where does it fall short of explicit CoT?

4.1 Semantic Content and Causal Engagement

We begin with the most basic functional question: what information is encoded in latent states, and how does model behavior depend on those states? We combine semantic readouts with controlled state interventions. Readouts characterize what problem-specific intermediate information is accessible across position and depth, whereas interventions test whether downstream behavior depends on the states themselves and where this dependence is concentrated.

Intermediate semantics are decodable but selectively localized within the latent trajectory. Latent representations assign substantially greater readout mass to intermediate values from the current problem than to matched values drawn from other problems, including controls sharing the same final answer (Figure 5a). This problem-specific content is not uniformly accessible within the latent computation: semantic probes reveal pronounced variation across both position and depth, with peak decodability localized to different position–layer regions across models (Figure 5b).

Causal engagement is problem-sensitive and strongly structured by latent position. Whole-trajectory interventions show that replacing latent states with states drawn from other problems is more disruptive than the tested norm-matched noise control (Figure 5c). Thus, preserving activation scale alone is insufficient to preserve behavior: which problem's states occupy the latent trajectory matters. This dependence is also highly position-selective. Across models, zeroing, donor replacement, and restoration reveal a pronounced even–odd asymmetry in intervention effects (Figure 5d). This functional parity echoes the positional organization observed during formation, suggesting that the internal structure of the latent trajectory is expressed not only representationally but also in its causal participation at inference time.

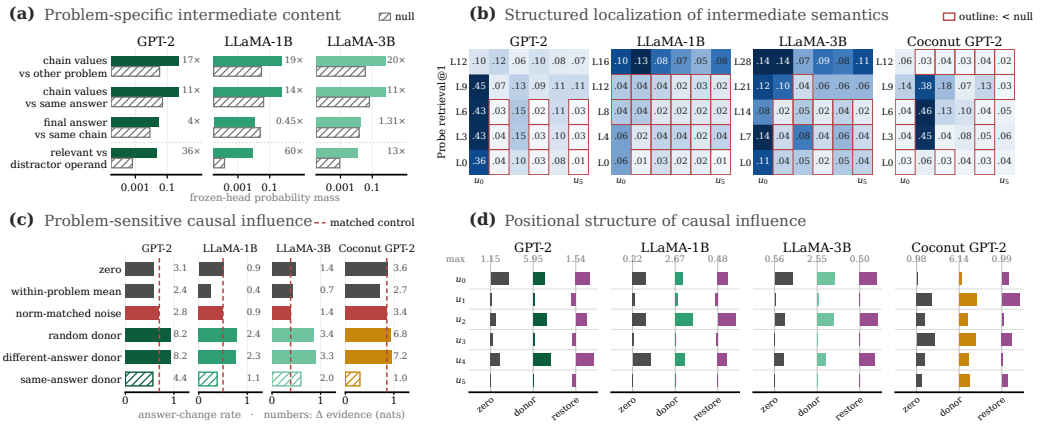


Figure 5: **Semantic and causal effects are positionally asymmetric.** Intermediate content is selectively localized across position and depth (a,b), while causal influence is problem-sensitive and parity-biased (c,d).

Together, these results reveal pronounced positional asymmetries in both semantic localization and causal influence. We next ask how recurrent feedback reshapes these contributions.

4.2 Inter-State Coupling and Computational Composition

The preceding results establish that individual latent states can influence the final answer, with this participation itself exhibiting pronounced positional asymmetry. Yet causal relevance at individual positions is not sufficient to establish an organized reasoning process. We therefore ask whether latent states compose in this way, or is transformed through recurrent interactions with subsequent states. To isolate this dependence, we compare local perturbations while downstream latent inputs are either held fixed or allowed to regenerate, and use downstream state restoration to characterize how the effect of one position depends on others. We further compare these inter-state relations across reasoning outcomes.

Latent-state contributions are recurrently interdependent rather than state-local. Allowing downstream latent inputs to regenerate after a local perturbation can substantially alter its final effect relative to keeping subsequent inputs pinned (Figure 6a). Where these effects are sufficiently large to resolve, recurrent continuation can either attenuate or amplify the original perturbation. Thus, the causal contribution of a latent state depends on how the downstream trajectory responds, establishing inter-state dependence beyond isolated state-to-answer effects.

Inter-state causal organization persists in failed reasoning, but becomes more diffuse and less sharply structured. Pairwise restoration is not uniformly corrective in either outcome cohort, but the structure of these relations changes systematically with reasoning outcome (Figure 6b). Compared with solved trajectories, failed trajectories exhibit denser inter-state mediation patterns, with recovery and evidence-loss relations more extensively interwoven across latent positions. This difference is also reflected at the level of individual restored states: although restoration effects remain strongly context-dependent, solved trajectories show more differentiated positional roles, whereas failed trajectories are more compressed and less cleanly separated across positions (Figure 6c). Finally, distance-stratified restoration profiles in the failed cohort are flatter and closer to zero, while still displaying a gap between aggregate and typical effects (Figure 6d), indicating that these interactions remain heterogeneous across problems rather than consolidating into a clear distance-structured pattern.

Together, these results suggest that latent reasoning depends on more than the presence of causal interaction; the underlying organization of inter-state relations may itself be consequential.

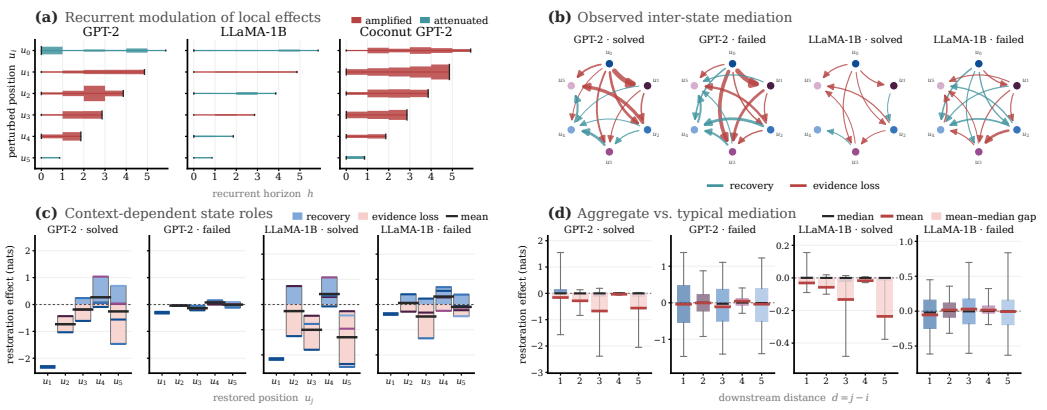


Figure 6: **Inter-state causal organization in latent reasoning.** Recurrent continuation reshapes local effects (a), while inter-state interventions reveal more diffuse, context-dependent, and problem-heterogeneous causal relations in failed reasoning (b–d).

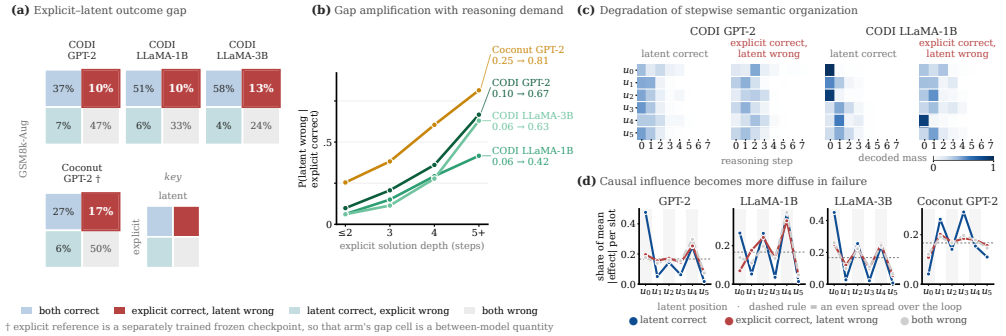


Figure 7: **Reasoning reliability tracks the organization of latent computation.** The explicit–latent gap widens with depth (a,b), while failure is accompanied by weaker semantic ordering and more diffuse causal influence (c,d).

4.3 Reasoning Reliability and the Explicit–Latent Gap

We finally ask when latent computation translates into reliable reasoning, and what changes internally when it falls short of explicit CoT. We address this at two levels. First, paired explicit–latent outcomes quantify where internalized reasoning falls short and how this gap varies with explicit solution depth. Second, outcome-conditioned semantic readouts and intervention profiles compare the organization of latent computation between successful and failed reasoning.

The explicit–latent reliability gap widens sharply with reasoning depth. Paired outcomes reveal a substantial set of problems that are solved through explicit CoT but missed by the latent pathway (Figure 7a). Across the settings, the conditional failure rate $P(\text{latent wrong} \mid \text{explicit correct})$ rises monotonically with explicit solution depth, from 0.06–0.25 on problems with at most two steps to 0.42–0.81 on problems with five or more (Figure 7b). The reliability gap is therefore most pronounced on problems requiring longer multi-step reasoning.

Failed latent reasoning is accompanied by weakened semantic ordering and redistributed causal influence across latent states. Among problems solved explicitly but missed by the latent pathway, the alignment between latent positions and successive intermediate reasoning steps exhibits weaker stepwise organization than in latent-correct cases (Figure 7c). Causal influence is likewise reorganized: the pronounced positional asymmetry observed in latent-correct trajectories becomes more diffuse across latent positions in failure (Figure 7d). Thus, failure is associated not with the disappearance of latent computation, but with a systematic change in how semantic content and causal influence are distributed across the trajectory.

5 DISCUSSION

Existing approaches increasingly provide supervision to intermediate latent states or seek to preserve reasoning information along the latent trajectory (Wei et al., 2026; Chen et al., 2026). Our results motivate going beyond state-wise semantic alignment toward *relational process supervision*: objectives that shape how task-relevant information is distributed, transformed, and coordinated across latent states. Such supervision need not force latent computation to reproduce explicit CoT step by step. Indeed, the task-relevant explicit–latent alignment we observe coexists with a distinct intermediate trajectory, suggesting that useful supervision should preserve reasoning structure while leaving room for latent computation to develop its own organization.

More broadly, latent budget should be viewed as a computational substrate rather than simply additional reasoning tokens. Positional specialization and recurrent dependence imply that adding latent states is useful only insofar as the resulting computation can coordinate them effectively, a challenge that becomes increasingly important as reasoning depth grows. This motivates methods that jointly optimize latent capacity and trajectory organization, as well as evaluations that track how such organization develops throughout training rather than judging latent reasoning only by its endpoint behavior. By connecting the formation of latent computation to its inference-time function, we hope to provide a basis for designing latent reasoning systems whose internal computation is both compact and reliably coordinated for complex reasoning.

6 RELATED WORK

From explicit chain-of-thought to latent reasoning. Building on explicit scratchpads and CoT (Nye et al., 2021; Wei et al., 2022), latent reasoning internalizes intermediate computation through distillation, progressive token removal, or continuous hidden states (Deng et al., 2023; 2024). Coconut recurrently feeds back hidden states through a replacement curriculum (Hao et al., 2025), while CODI learns explicit and latent pathways jointly via self-distillation (Shen et al., 2025). Other approaches use soft or compressed representations and semantically aligned implicit tokens (Xu et al., 2025; Tan et al., 2025; Deng et al., 2025; He et al., 2025), while Wei et al. (2026) introduce step-level supervision to mitigate latent homogenization. Related approaches allocate additional computation through pause tokens, filler tokens, and recurrent depth (Goyal et al., 2024; Pfau et al., 2024; He et al., 2025; Geiping et al., 2025)

Interpretability and mechanistic analysis of latent reasoning. Recent studies probe the internal representations and computations of latent reasoning. Scratchpad Thinking uncovers alternating storage–computation roles in CODI (Peters et al., 2025), while Liang & Pan (2026) study multi-hop dependence and shortcut solutions on controlled tasks. Li et al. (2026) characterize stepwise necessity and non-local routing, and Chang et al. (2026) turn such analyses into inference-time interventions. Complementary work studies recoverable reasoning traces (Dilgren & Wiegreffe, 2026), and supervision through information fidelity (Chen et al., 2026). More recently, Jin et al. (2026) track latent-reasoning faithfulness throughout training and show that similar final behavior can arise from substantially different training trajectories. Our work develops a formation–function account of latent reasoning. We trace how positional organization and explicit–latent alignment emerge alongside learning signals during training, and relate the resulting semantic and inter-state organization to reasoning outcomes at inference time.

7 CONCLUSION

We study reasoning internalization through a formation–function framework, connecting how latent computation is organized during training with how it functions at inference. Across CODI and Coconut, internalization yields positionally asymmetric latent trajectories that remain task-relevantly aligned with explicit reasoning while following distinct intermediate pathways. At inference, latent-state contributions are recurrently interdependent, while the explicit–latent gap grows with reasoning demand and failure is accompanied by weaker semantic ordering and more diffuse positional influence. Together, these findings suggest that understanding and improving internalized reasoning requires moving beyond whether latent states contain useful information to how they are coordinated into an effective computational process.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments. We also used generative AI tools to create or modify scientific figures, edit software code, summarize and analyze existing literature, search for relevant information, and edit the manuscript to improve readability. All AI-assisted outputs were reviewed by the authors. Methodological suggestions were evaluated against the experimental design and results; AI-assisted code and figures were checked for correctness; references and literature summaries were verified against the original sources; and all AI-assisted manuscript edits were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

REFERENCES

Shuochen Chang, Tong Bai, Xiaofeng Zhang, Qianli Ma, Qingyang Liu, Zhaohe Liao, Yibo Miao, and Li Niu. 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)*, pp. 34019–34032. Association for Computational Linguistics, 2026. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1568. URL https://aclanthology.org/2026.acl-long.1568/.

Xinghao Chen, Chak Tou Leong, Guo Wenjin, Jian Wang, Wenjie Li, and Xiaoyu Shen. What makes effective supervision in latent chain-of-thought: An information-theoretic analysis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=x21YkF1UfR.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jingcheng Deng, Liang Pang, Zihao Wei, Shicheng Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. Llm latent reasoning as chain of superposition. *arXiv preprint arXiv:2510.15522*, 2025.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Connor Dilgren and Sarah Wiegreffe. Are latent reasoning models easily interpretable? In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=PBwyWJFZJR.

Jonas Geiping, Sean Michael McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=S3GhJooWIC.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024.

Sayam Goyal, Brad Peters, María Emilia Granda, Akshath Vijayakumar Narmadha, Dharunish Yugeswardeenoo, Callum Stuart McDougall, Sean O'Brien, Ashwinee Panda, Kevin Zhu, and Cole Blondin. Scratchpad thinking: Alternation between storage and computation in latent reasoning models. In *NeurIPS 2025 Workshop on Regulatable ML*, 2025. URL https://openreview.net/forum?id=rdmnQIAdBq.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Yinhan He, Wendy Zheng, Yaochen Zhu, Zaiyi Zheng, Lin Su, Sriram Vasudevan, Qi Guo, Liangjie Hong, and Jundong Li. Semcot: Accelerating chain-of-thought reasoning through semantically-aligned implicit tokens. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=1ZuzFUMtx6.

Hengyu Jin, Shu Yang, and Di Wang. Final checkpoints are not enough: Analyzing latent reasoning faithfulness along training trajectories. *arXiv preprint arXiv:2607.06648*, 2026.

Zirui Li, Xuefeng Bai, Kehai Chen, Yizhi LI, Jian Yang, Chenghua Lin, and Min Zhang. Dynamics within latent chain-of-thought: An empirical study of causal structure. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kHB8m3ojGe.

Jia Liang and Liangming Pan. Do latent-cot models think step-by-step? a mechanistic study on sequential reasoning tasks. *arXiv preprint arXiv:2602.00449*, 2026.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

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. Scratchpad thinking: Alternation between storage and computation in latent reasoning models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=EV30qkZXrR.

Jacob Pfau, William Merrill, and Samuel R Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693. Association for Computational Linguistics, 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AQsko3PPUe.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. SIM-cot: Supervised implicit chain-of-thought. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6YRJ4jmVQl.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23336–23351, 2025.

A NOTATION

Table 1 summarizes notation shared by the main text and appendix. The training-stage index is omitted for fixed-checkpoint analyses; measurement definitions appear in Appendix C.

Table 1: Notation for latent computation and its analysis

SYMBOL	DEFINITION
Latent computation	
s, θ_s, K_s	Training stage, model parameters, and latent budget.
$x, y_{1:m}$	Question and gold-answer token sequence.
$u_{s,t}$	Continuous input consumed at latent position t.
$h_{s,t}^{(\ell)}, h_{s,t}$	Hidden state at layer ℓ and its final-layer counterpart; $h_{s,0}$ is the question-boundary state.
$C_{s,t}$	Cached context after processing latent position t.
$P_{\theta_s}, \Phi_{\theta_s}$	Feedback map and incremental Transformer update.
U, π	Latent input sequence and method-specific answer prefix.
$U^0, \widetilde{U}$	Clean and intervened latent input sequences.
$\mathcal{I}, \mathcal{R}, V$	Overwritten positions, regenerating-input positions, and replacement values in the intervention operator $\mathcal{T}_{\mathcal{I},\mathcal{R}}$.
Answer evidence and intervention effects	
$s(x; U)$	Mean-token log-probability of the gold answer under rollout U.
$N_{\mathrm{dir}}, N_{\mathrm{tot}}$	Answer-evidence loss under pinned and regenerated feedback.
R_i	Mean evidence recovered by restoring one clean feed, divided by the mean loss from zeroing the full sequence.
E_i, M_{ij}	Evidence loss from perturbing feed i, and the additional score change from restoring clean downstream feed j.
$C^{A\to B}$	Change in the donor-versus-recipient answer margin under pinned donor replacement.
$h,\ d = j - i$	Number of downstream inputs allowed to regenerate, and source–target positional distance.
Learning signals and aggregation	
$\mathcal{L}_a, \lambda_D$	Student, teacher, or distillation objective indexed by a, and the distillation-loss coefficient.
$g = \nabla_\theta \mathcal{L}_a$	Objective gradient over optimizer-updated parameters.
$\partial \mathcal{L}_a / \partial h_p^{(\ell)}$	Loss derivative with respect to the hidden state at position p and layer ℓ, evaluated at a checkpoint.
$\mathcal{D}, \mathbb{E}_{x\in\mathcal{D}}$	Analysis cohort and empirical mean, also written $\langle\cdot\rangle_{\mathcal{D}}$.
p_i	Positional share obtained by normalizing the cohort's mean absolute pinned single-feed intervention effects.

Indexing conventions. Equation 1 uses one-based latent positions. Figures use zero-based feed indices, with $u_i \equiv u_{s,i+1}$ for $i = 0, \ldots, K_s - 1$. Readouts refer to captured hidden states, whereas interventions refer to feeds overwritten before consumption; see Appendix B.3.

B EXPERIMENTAL SETUP

B.1 MODELS AND TRAINING

We study CODI on GPT-2 and Llama-3.2-1B/3B-Instruct, and Coconut on GPT-2. CODI shares weights across pathways and uses a feedback MLP comprising dropout, two linear layers with GELU, and LayerNorm. Coconut uses identity feedback and full fine-tuning from a CoT-SFT checkpoint. Its frozen explicit reference is separately trained, giving a between-model comparison rather than CODI's within-model comparison.

Table 2: Training configurations. Batch size is effective across devices and accumulation steps; λ_D is the distillation-loss coefficient.

Setting	Epochs	Learning rate	Batch	Projection width	λ_D
CODI GPT-2	40	3×10^{-3}	128	768	1
CODI Llama-1B	10	8×10^{-4}	128	2048	20
CODI Llama-3B	8	3×10^{-4}	128	3072	20
Coconut GPT-2	25	10^{-4}	32	—	—

The CODI runs use six latent positions, rank-128 LoRA with $\alpha = 32$, bf16 training, maximum sequence length 512, gradient clipping at 2.0, weight decay 0.1, and cosine scheduling with a 0.03 warmup ratio. Coconut uses fp32 training, and weight decay 0.01.

For consistency across arms, Table 3 reports final-checkpoint latent-pathway accuracy under the common analysis evaluator.

Table 3: Final latent-pathway accuracy on the 1,175-problem GSM8K test set.

Setting	Final checkpoint	Test accuracy
CODI GPT-2	ep40	43.57%
Coconut GPT-2	ep25	32.94%
CODI Llama-1B	ep10	56.68%
CODI Llama-3B	ep8	62.72%

Coconut replaces each removed explicit step with two latent positions, advances its curriculum every three epochs, and caps the stage at three with padding enabled. For displayed epoch e, its scheduled budget is

$$K_e = 2 \min\left(\left\lfloor \frac{e-1}{3} \right\rfloor, 3 \right). \tag{2}$$

The analyzed checkpoints therefore use $K = 2$ at epochs 4 and 6, $K = 4$ at 7 and 9, and $K = 6$ from epoch 10 onward. Six-position comparisons use only the nine analyzed checkpoints with $K = 6$.

B.2 DATA, CHECKPOINTS, AND COHORTS

The original GSM8K dataset contains 7,473 training problems and 1,319 test problems (Cobbe et al., 2021). We train on GSM8K-Aug (Deng et al., 2023), which expands the training split to 385,620 augmented examples, and evaluate on the original 1,319-problem GSM8K test split. Our mechanistic analyses use the 1,175 test problems that pass the parsing and labeling pipeline.

Outcome selection. For the paired-outcome panel $\mathcal{P}$, let $c_E(x), c_L(x) \in \{0, 1\}$ denote explicit and latent correctness at a specified checkpoint. The four outcome cohorts are

$$\mathcal{D}_{ab} = \{x \in \mathcal{P} : c_E(x) = a,\ c_L(x) = b\}, \qquad a, b \in \{0, 1\}. \tag{3}$$

Thus $\mathcal{D}_{11}$ is both correct, $\mathcal{D}_{10}$ is the *gap cohort*, $\mathcal{D}_{01}$ is latent-only correct, and $\mathcal{D}_{00}$ is both wrong. Latent-correct and latent-failed cohorts are $\mathcal{D}_{11} \cup \mathcal{D}_{01}$ and $\mathcal{D}_{10} \cup \mathcal{D}_{00}$, respectively. Intervention records are joined to the same checkpoint's paired outcomes, discarding inconsistent records.

Table 4: Checkpoint grids and stored final-outcome cohorts for mathematical reasoning.

Setting	Analyzed epochs	Pool	Solved	Failed
CODI GPT-2	2–40 by 2	1175	512	663
CODI Llama-1B	1–10	1175	666	509
CODI Llama-3B	1–8	1175	737	438
Coconut GPT-2	4, 6, 7, 9, 10, 12, 13, 15, 17, 19, 21, 23, 25	1175	387	788

B.3 Rollout indexing and measurement coverage

We retain the rollout in Eq. 1. Plots use zero-based feed indices $u_0, \ldots, u_{K-1}$, corresponding to $u_{s,1}, \ldots, u_{s,K_s}$ in that equation. The feed at plot position i is produced from the preceding state $h_{s,i}$ and is overwritten *before* consumption. Readouts instead interrogate captured hidden states, with a separate layer axis; feed and hidden-state indices are not interchangeable. The hidden-state grids are $\{0, 3, 6, 9, 12\}$, $\{0, 4, 8, 12, 16\}$, and $\{0, 7, 14, 21, 28\}$ for GPT-2, Llama-1B, and Llama-3B. These correspond to relative depths $0, 0.25, 0.5, 0.75, 1$.

C Analysis Details

All quantities are evaluated at a specified checkpoint and on the eligible cohort $\mathcal{D}$ for the corresponding analysis. We write $\langle f \rangle_{\mathcal{D}} = |\mathcal{D}|^{-1} \sum_{x \in \mathcal{D}} f(x)$ for an empirical mean and suppress the cohort subscript when unambiguous.

C.1 Answer scoring and controlled interventions

Answer evidence. For question x, gold-answer tokens $y_{1:m}$, and latent input sequence U,

$$s(x; U) = \frac{1}{m} \sum_{r=1}^{m} \log p_\theta(y_r \mid x, U, \pi, y_{<r}), \tag{4}$$

where π is the method's raw answer prefix. Sequence scoring includes any tokenizer-invariant leading token in the target; next-token readouts include it in the prefix. Answers are parsed after the answer marker, and alternative answers are scored from separate cache copies. Both mean and summed token log-probabilities are retained.

Intervention rollout. Let $U^0 = (u_0^0, \ldots, u_{K-1}^0)$ denote the clean feeds, $\mathcal{I}$ the positions overwritten with $V = \{v_k : k \in \mathcal{I}\}$, and $\mathcal{R}$ the positions whose inputs regenerate. The intervened rollout $\mathcal{T}_{\mathcal{I},\mathcal{R}}(U^0; V)$ is defined by

$$\begin{aligned} \widetilde{u}_k &= \begin{cases} v_k, & k \in \mathcal{I}, \\ P_\theta(\widetilde{h}_k), & k \in \mathcal{R} \setminus \mathcal{I}, \\ u_k^0, & \text{otherwise}, \end{cases} \\ (\widetilde{h}_{k+1}, \widetilde{C}_{k+1}) &= \Phi_\theta(\widetilde{u}_k; \widetilde{C}_k), \qquad k = 0, \ldots, K-1. \end{aligned} \tag{5}$$

The question-boundary state and cache initialize the recurrence. Overwrites act on post-projection feeds before consumption. Pinning holds *inputs* fixed; hidden states and cache entries still follow the intervened prefix.

Replacement operators. Whole-trajectory interventions overwrite all feeds with

$$\begin{aligned} v_k^{\text{zero}} &= 0, & v_k^{\text{mean}} &= \frac{1}{K} \sum_{r=0}^{K-1} u_r^0(x), \\ v_k^{\text{noise}} &= \|u_k^0(x)\|_2 \frac{\epsilon_k}{\|\epsilon_k\|_2}, & v_k^{\text{donor}} &= u_k^0(x'), \quad \epsilon_k \sim \mathcal{N}(0, I). \end{aligned} \tag{6}$$

Donors are sampled randomly from other problems or restricted to different-answer problems. The Gaussian comparator matches the replacement-vector norm. The separately displayed same-answer condition retains the original sampling pool, including the recipient, and is not used to assess content preservation.

Effect scores and feedback conditions. For an intervened rollout $\widetilde{U}$, define evidence loss and answer-change rate as

$$N(x;\widetilde{U}) = s(x;U^0) - s(x;\widetilde{U}), \qquad \text{Flip} = \big\langle \mathbf{1}\{\widehat{y}(x;\widetilde{U}) \neq \widehat{y}(x;U^0)\}\big\rangle, \tag{7}$$

where $\widehat{y}$ is the model's parsed greedy answer. For single-feed erasure at i, pinned and regenerated feedback correspond to $\mathcal{R} = \emptyset$ and $\mathcal{R} = \{i+1,\dots,K-1\}$:

$$\begin{aligned} N_{\text{dir},i}(x) &= N\big(x;\mathcal{T}_{\{i\},\emptyset}(U^0;0)\big)\,, \\ N_{\text{tot},i}(x) &= N\big(x;\mathcal{T}_{\{i\},\{i+1,\dots,K-1\}}(U^0;0)\big)\,, \\ \Delta_i^{\text{fb}}(x) &= N_{\text{tot},i}(x) - N_{\text{dir},i}(x). \end{aligned} \tag{8}$$

We report cohort means of these quantities; Δ_i^{fb} is a feedback-response contrast.

Single-position restoration. Let $U^{\text{zero}} = (0,\dots,0)$ and $U^{\text{restore}(i)} = (0,\dots,u_i^0,\dots,0)$, with all other inputs pinned to zero. The normalized recovery is

$$R_i = \frac{\big\langle s(x;U^{\text{restore}(i)}) - s(x;U^{\text{zero}})\big\rangle}{\big\langle s(x;U^0) - s(x;U^{\text{zero}})\big\rangle}. \tag{9}$$

This is a ratio of means. Its numerator and denominator are retained in nats; R_i is undefined when the denominator's absolute value is at most 10^{-6}. Restoring all feeds must reproduce the clean score.

Donor-directed answer margins. For a recipient (x_A, a) and a different-answer donor (x_B, b), let $s_A(c;U)$ score answer c on x_A. Pinned replacement measures

$$C^{A\to B} = \big[s_A(b;U) - s_A(a;U)\big]_{\text{patched}} - \big[s_A(b;U) - s_A(a;U)\big]_{\text{clean}}. \tag{10}$$

Pairs are source-matched within the evaluation cohort, each problem is paired at most once, and both directions are evaluated. The clean answer margin is retained. Figure 5(d) maximum-normalizes each operator's positional profile separately and annotates its raw maximum.

C.2 Formation and positional organization

Positional profiles. Figure 1 combines hidden-state norms, latent-to-latent attention, best-layer semantic retrieval, and pinned single-feed erasure effects. Attention is averaged over heads and summed over source groups; activation norms are averaged within groups. For a nonnegative profile q_i and a signed effect profile e_i, the displayed normalizations are

$$\widetilde{q}_i = \frac{q_i / \sum_j q_j}{1/K}, \qquad \widetilde{e}_i = \frac{e_i}{\max_j |e_j|}, \tag{11}$$

for nonzero denominators. The semantic profile uses $r_i^* = \max_\ell r_{i,\ell}$, where $r_{i,\ell}$ is retrieval@1; cells not exceeding their shuffled reference are marked.

Effective dimensionality and cross-layer reversals. For clean feeds across questions, define

$$\begin{aligned} \mu_i &= \langle u_i^0 \rangle, \qquad \Sigma_i = \big\langle (u_i^0 - \mu_i)(u_i^0 - \mu_i)^\top \big\rangle, \\ \text{PR}_i &= \frac{(\operatorname{tr}\Sigma_i)^2}{\operatorname{tr}(\Sigma_i^2)}. \end{aligned} \tag{12}$$

Figure 2 compares restoration, this effective dimensionality, and layerwise activation parity. Let $\mathcal{L}_{s,s'}$ be the eligible layers between adjacent analyzed checkpoints and $\mathcal{F}_{s,s'}$ the subset whose recorded parity contrasts change sign. Highlighted intervals satisfy $|\mathcal{F}_{s,s'}|/|\mathcal{L}_{s,s'}| \geq 0.8$.

C.3 SEMANTIC READOUTS AND EXPLICIT–LATENT COMPARISONS

Frozen-head content. Let $\ell_\theta(z)$ denote the checkpoint's frozen vocabulary-head logits for a captured state z, and $\mathcal{V}_x$ the token set for the target content. The decoded mass and selected slot are

$$\begin{aligned} m(z;\mathcal{V}_x) &= \sum_{v\in\mathcal{V}_x} [\mathrm{softmax}\,\ell_\theta(z)]_v, \\ i^*(x) &= \arg\max_i m(z_i(x);\mathcal{V}_x^{\mathrm{chain}}). \end{aligned} \tag{13}$$

Targets are intermediate values, the answer, or labeled operands. Figure 5(a) uses $i^*(x)$ and problems whose step values and answer each map to a single content token. Controls use a depth-matched other problem, a same-answer problem with different intermediates, or a same-intermediate-multiset problem with a different answer; the last is scored on the partner's answer tokens. Each contrast uses its eligible matched subset, omits partnerless records, and requires at least 30 scored reference cells.

Semantic probes. Bilinear InfoNCE probes rank a 64-way frozen base-model value dictionary. Writing $b_\psi(z,v)$ for the fitted bilinear score, retrieval@1 is

$$r_{i,\ell} = \left\langle \mathbf{1}\left\{ \arg\max_{v\in\mathcal{V}_{\mathrm{dict}}} b_\psi(z_{i,\ell}(x),v) \in \mathcal{Y}_x \right\} \right\rangle, \qquad |\mathcal{V}_{\mathrm{dict}}| = 64, \tag{14}$$

where $\mathcal{Y}_x$ is the first intermediate-value target or the separately scored intermediate-value union. Probe fitting uses a skeleton-disjoint split, a 0.5 development fraction, and 20 training epochs; evaluation uses held-out skeletons. Ten label-shuffled draws provide reference mean μ_0 and standard deviation σ_0. For n evaluated problems, a score clears the reference if

$$r_{i,\ell} > \mu_0 + 2\max\left\{ \sigma_0, \sqrt{\frac{\mu_0(1-\mu_0)}{n}} \right\}. \tag{15}$$

The empirical shuffled reference is the comparator; $1/64$ is the nominal retrieval rate only.

Cross-pathway correspondence. Figure 3(a) applies a frozen explicit-trained probe to latent states, alongside shuffled-label and explicit-side references. Panel (b) averages the coordinate distance

$$d_{\mathrm{coord}} = \frac{\|d_T - d_S\|_2}{\|d_T\|_2 + \|d_S\|_2} \tag{16}$$

over layers for extracted explicit and latent representations along the written chain and at the answer boundary. CODI's answer-boundary state is the explicit distillation target.

Attention allocation. Let a_{qp} be attention from query q to position p, Q the question tokens, $Q_c \subseteq Q$ a token class, V_q the visible context, and $R_q \subseteq V_q$ the query's own reasoning context. The two reported quantities are

$$A_c(q) = \frac{\sum_{p\in Q_c} a_{qp}}{\sum_{p\in Q} a_{qp}}, \qquad A_{\mathrm{reason}}(q) = \frac{\sum_{p\in R_q} a_{qp}}{\sum_{p\in V_q} a_{qp}} \frac{|V_q|}{|R_q|}. \tag{17}$$

Panel (c) uses relevant, distractor, and ambiguous operands, operators, and other tokens as classes; panel (d) uses uniform allocation over visible positions as its unit reference. Cross-pathway queries occur after the contexts diverge; the recorded answer query is one token after the answer-emitting position. Agreement inside the shared question prefix is a structural check. Llama-3B routes are sampled at epochs 2, 4, 6, and 8.

C.4 PARAMETER GRADIENTS AND STATE-SPACE LOSS SENSITIVITY

Objective gradients. For $a \in \{S, T, D\}$ (student, teacher, distillation), define $g_a(x) = \nabla_{\theta_{\mathrm{upd}}} \mathcal{L}_a(x)$ over optimizer-updated parameters only (LoRA and the projection). We record both magnitude conventions:

$$G_a^{\mathrm{mean}} = \|\langle g_a \rangle\|_2, \qquad G_a^{\mathrm{sample}} = \langle \|g_a\|_2 \rangle, \qquad \rho_a^\bullet = G_a^\bullet / G_S^\bullet. \tag{18}$$

Here $\bullet$ denotes either magnitude convention. For each objective pair, the per-example cosine and its summaries are

$$\begin{aligned} c_{ab}(x) &= \frac{g_a(x)^\top g_b(x)}{\|g_a(x)\|_2 \|g_b(x)\|_2}, \\ \overline{c}_{ab} &= \langle c_{ab} \rangle, \qquad f_{ab}^- = \langle \mathbf{1}\{c_{ab} < 0\} \rangle. \end{aligned} \tag{19}$$

Figure 4(b) uses $\bar{c}_{ab}$, not the cosine between mean gradients. Analysis uses fp32 with float64 inner products. Loss components are reconstructed with differentiable graphs; the recorded component-value check for Llama-1B uses relative tolerance 0.02 and absolute tolerance 10^{-3}. The GPT-2 and Llama-3B records contain no completed component-value comparison.

State-space sensitivity. For objective a, hidden position p, and layer ℓ, the local sensitivity is

$$v_{a,p}^{(\ell)}(x) = \left\| \frac{\partial \mathcal{L}_a(x)}{\partial h_p^{(\ell)}(x)} \right\|_2 . \tag{20}$$

Let $v_{a,G}$ be the reported mean sensitivity within position group G, $\mathcal{G}_a$ the stream's position groups, and $\mathcal{R}_a \subseteq \mathcal{G}_a$ its reasoning groups. Reasoning enrichment is

$$\mathcal{E}_a = \frac{\sum_{G\in\mathcal{R}_a} v_{a,G}}{\sum_{G\in\mathcal{G}_a} v_{a,G}} \frac{|\mathcal{G}_a|}{|\mathcal{R}_a|}, \qquad \mathcal{E}_{a,i}^{\text{latent}} = \frac{v_{a,i}}{\sum_{j=0}^{K-1} v_{a,j}} K. \tag{21}$$

Student and distillation maps use the latent stream; teacher maps use the explicit stream. References count position groups, with the within-latent reference restricted to active slots.

C.5 Recurrent continuation and inter-state restoration

Recurrent horizon. After zeroing feed i, let $\mathcal{R}_{i,h} = \{i+1, \dots, i+h\}$, $0 \le h \le K-1-i$, with $\mathcal{R}_{i,0} = \emptyset$. Define

$$\begin{aligned} U^{i,h} &= \mathcal{T}_{\{i\},\mathcal{R}_{i,h}}(U^0; 0), & N_i(h; x) &= s(x; U^0) - s(x; U^{i,h}), \\ N_i(0; x) &= N_{\text{dir},i}(x), & N_i(K-1-i; x) &= N_{\text{tot},i}(x). \end{aligned} \tag{22}$$

Figure 6(a) varies h, pinning the remaining inputs clean. The endpoints are recomputed as consistency checks; all steps retain access to the accumulated cache.

Pairwise restoration. Let $U^i = U^{i,K-1-i}$ and let $U^{i\to j}$ use the same regenerated rollout with clean feed u_j^0 reinstated at $j > i$. Then

$$\begin{aligned} E_i(x) &= s(x; U^0) - s(x; U^i), \\ M_{ij}(x) &= s(x; U^{i\to j}) - s(x; U^i), \qquad j > i, \\ \overline{M}_{ij} &= \langle M_{ij} \rangle, \qquad F_{ij} = \frac{\langle M_{ij} \rangle}{\langle E_i \rangle}. \end{aligned} \tag{23}$$

Positive M_{ij} recovers gold-answer evidence; negative values reduce it. F_{ij} is reported only when its denominator meets the 0.25-nat eligibility threshold. Diagonal recovery checks and causally unreachable positions are kept separate from the downstream pairs.

Context and distance summaries. Panel (b) encodes $|\overline{M}_{ij}|$ by chord width and its sign by color; width scales are shared across models within, but not across, outcome cohorts. Panel (c) compares $\{\overline{M}_{ij} : i < j\}$ at each restored position j, with vertical scales shared within models. For panel (d), define

$$\begin{aligned} \mathcal{M}_d &= \big(M_{i,i+d}(x)\big)_{x\in\mathcal{D},\, 0\le i<K-d}, \\ \mu_d &= \text{mean}(\mathcal{M}_d), \qquad m_d = \text{median}(\mathcal{M}_d), \qquad \delta_d = \mu_d - m_d. \end{aligned} \tag{24}$$

Means, medians, and their separation are shown on per-panel scales.

C.6 Reasoning outcomes and outcome-conditioned organization

Paired outcomes and depth. Evaluation uses free greedy explicit and latent generation, allowing 96 new tokens for the explicit chain and eight for the answer. A separate teacher-forced condition supplies the gold intermediate chain, omitting its final expression under CODI's convention. For explicit-solution depth bin $B \in \{\{d : d \le 2\}, \{3\}, \{4\}, \{d : d \ge 5\}\}$, define $n_{ab}(B) = |\{x \in \mathcal{D}_{ab} : d_{\text{exp}}(x) \in B\}|$. The conditional gap is

$$P(L \text{ wrong} \mid E \text{ correct},\ d_{\text{exp}} \in B) = \frac{n_{10}(B)}{n_{11}(B) + n_{10}(B)}. \tag{25}$$

Here d_{exp} counts steps in the explicit written solution; cohorts $\mathcal{D}_{ab}$ are defined in Appendix B.2.

Semantic organization. Let $m_{ik}(x)$ be the cached frozen-head decoded mass at latent position i in step bin $k \in \{0, \ldots, 7\}$. Within each eligible outcome cohort, Figure 7(c) displays

$$\overline{m}_{ik} = \langle m_{ik}(x) \rangle_{\mathcal{D}}, \qquad H_{ik} = \frac{\overline{m}_{ik}}{\sum_{r=0}^{7} \overline{m}_{ir}}. \tag{26}$$

Averaging precedes row normalization, so each row describes cohort-level decoded step mass. This analysis covers CODI GPT-2 and Llama-1B.

Positional intervention profiles. Panel (d) reports

$$p_i = \frac{\mathbb{E}_{x \in \mathcal{D}}[|N_{\text{dir},i}(x)|]}{\sum_j \mathbb{E}_{x \in \mathcal{D}}[|N_{\text{dir},j}(x)|]}. \tag{27}$$

Absolute effects are averaged before normalization. Readouts and interventions share outcome definitions but use separate eligible problem sets.

Arrangement references. For a positional vector $w(x)$, supplementary comparisons use $w_i^{\pi}(x) = w_{\pi_x(i)}(x)$, where π_x permutes slot identities within problem x while preserving its magnitudes. Concentration and maximum-share statistics describe dispersion; parity and cross-problem positional agreement also depend on slot identity. The non-decreasing decoded-step statistic permits repeated step labels.

C.7 Statistical estimation and experimental controls

For a statistic T, we report the original-sample estimate $\widehat{T} = T(\mathcal{D})$. With $B = 2000$ bootstrap replicates, the percentile interval is

$$\text{CI}_{0.95}(T) = \left[Q_{0.025}\left(\{T_b^*\}_{b=1}^{B}\right), Q_{0.975}\left(\{T_b^*\}_{b=1}^{B}\right) \right]. \tag{28}$$

We resample problems and, where applicable, probe seeds within problems; intervals are conditional on each setting's recorded training run. Norm-matched replacements, shuffled labels, and slot permutations provide the specified intervention, readout, and arrangement comparisons. Gradient and per-position restoration summaries are descriptive. Identity, restore-all, causal-direction, and horizon-endpoint checks assess implementation consistency separately from statistical references.

Outcome analyses retain clean gold-answer scores, raw effects, and normalized profiles to characterize associations with answer correctness. Randomized operators use stable, problem-based seeds; cohort definitions and measurement records are stored separately.