ROBUSTNESS IS NOT REASONING: A DOUBLE DISSOCIATION IN LATENT CHAIN-OF-THOUGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent chain-of-thought methods such as Coconut and CODI replace token-level reasoning with continuous "thoughts," and a growing literature probes whether these thoughts are *robust* to perturbation and whether they *carry the reasoning*, often collapsing both questions into a single perturbation test. We show that these are two distinct properties that can be dissociated. We apply three interventions to individual latent thoughts: random perturbation, counterfactual patching, and *content-shuffle* (replacing a thought with a real thought computed for a different input). We find that a thought's robustness to random noise is largely independent of whether its content is necessary for the answer. On a controlled k-hop modular-arithmetic task (Pythia-160m/410m, multi-seed, accuracy-matched), the fraction of thoughts that are noise-inert rises monotonically with recurrence depth (from 0.00 at depth 3 to 0.87 at depth 9), while the content of those same thoughts remains necessary at moderate depth and only degrades at excess depth, so the "robust *and* load-bearing" regime is non-monotonic. On a released CODI-Llama-1B model on GSM8K, latent thoughts are simultaneously noise-robust (answers rarely flip under perturbations of 16–$64\times$ the thought norm), individually redundant (single-thought ablation costs at most 4.3% accuracy), collectively essential (ablating all thoughts returns accuracy to near the no-thought floor), and content-load-bearing (shuffling content collapses accuracy from 0.436 to 0.165 ± 0.004), together operating as a distributed, error-correcting code that is individually redundant yet collectively essential. Robustness is therefore not evidence of reasoning, and robustness-only diagnostics can misrepresent what latent thoughts do.

1 INTRODUCTION

Chain-of-thought (CoT) prompting improves language-model reasoning by emitting intermediate steps in natural language (Wei et al., 2022; Kojima et al., 2022), and part of its benefit comes from the extra serial computation that intermediate tokens permit, independent of what those tokens say (Pfau et al., 2024; Merrill & Sabharwal, 2024; London et al., 2025). *Latent* chain-of-thought removes the tokens: Coconut feeds a model's last hidden state back as the next input embedding (Hao et al., 2024), and CODI distills an explicit-CoT teacher into a continuous student (Shen et al., 2025), thereby reasoning in a continuous space that is more efficient but opaque. This opacity has motivated interpretability work asking two questions of the resulting "continuous thoughts": are they *robust*, and do they *carry the reasoning*?

These questions are often answered with a single tool, perturbing a thought to see whether the answer changes, and the answers conflict. Some report that latent thoughts are strikingly robust to perturbation yet lack reasoning-critical content, behaving as "pseudo-reasoning" placeholders (Zhang et al., 2025); others report that latent thoughts do carry decodable intermediate states (Liang et al., 2026). We argue that the conflict is partly definitional: *robustness to random perturbation* and *necessity of a thought's content* are different properties, measured by different operations, and a thought can have either without the other.

We make this precise with three interventions on individual continuous thoughts (Figure 1): (i) **random perturbation** of controlled magnitude, summarized by a *basin radius*; (ii) **counterfactual patching**, which substitutes the thought a model computes for a minimally-different input; and (iii) **content-shuffle**, which replaces a thought with a real thought drawn from another example (same

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.

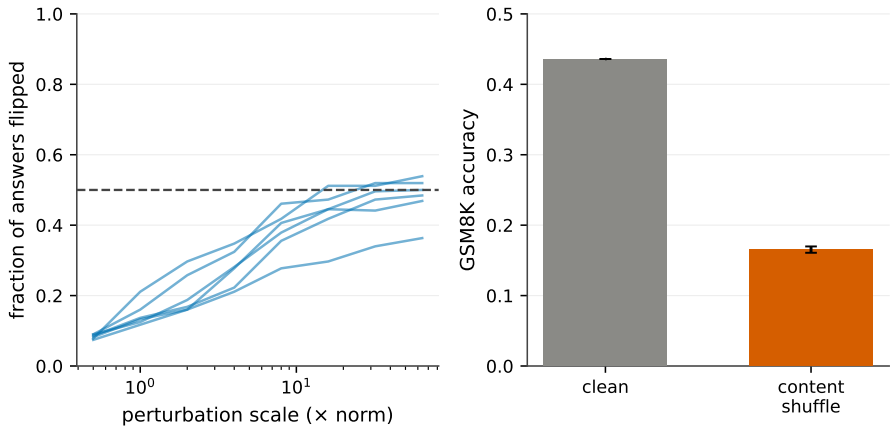

Figure 1: Robustness is not reasoning, on a released CODI-Llama-1B model (GSM8K). Left: random perturbation of a latent thought rarely flips the answer, even at 16–64× the thought norm (measured flip fraction, one line per thought). Right: shuffling thought content across examples collapses accuracy from 0.436 to 0.165 (mean±sd, 5 seeds). The two panels measure the same thoughts and reach opposite conclusions; numbers appear in Table 3.

distribution, wrong content). The first measures robustness; the last two measure content necessity. Across a controlled task and a released real-task model we find a systematic *double dissociation*: robustness and content-necessity vary independently.

Contributions. (1) We separate *noise-robustness* from *reasoning-content necessity* in latent CoT and show, with three interventions, that they are dissociable rather than interchangeable (§5). (2) With recurrence depth as a knob (multi-seed, accuracy-matched), noise-robustness rises monotonically with depth while per-thought content necessity is highest at minimal depth and declines as depth grows, so the co-occurrence of the two, the "robust *and* load-bearing" regime, peaks at moderate depth rather than at either extreme (§5). (3) On a released CODI-Llama-1B model on GSM8K, latent thoughts are jointly noise-robust, individually redundant, collectively essential, and content-load-bearing, together operating as a distributed, error-correcting code (individually redundant yet collectively essential) on a real task (§6). (4) We show that at minimal depth latent thoughts are *both* robust and load-bearing, reconciling prior "inert" vs. "load-bearing" claims (§4), and give a contraction account of the robustness (§7). Methodologically, a robustness-only diagnostic can label a content-essential thought "inert" and a content-free thought "robust," with direct consequences for interpreting and monitoring latent reasoning (Korbak et al., 2025).

2 Related Work

Latent and implicit chain-of-thought. Coconut reasons by feeding the last hidden state back as the next input embedding (Hao et al., 2024); CODI distills an explicit-CoT teacher into a continuous student by aligning hidden states (Shen et al., 2025). Follow-ups address the training instability of implicit CoT (Wei et al., 2025), add looped or recurrent depth (Fan et al., 2026), and study inference-time scaling of continuous reasoning (Wang et al., 2025). Our contribution is diagnostic rather than a new method: we ask what the continuous thoughts these methods produce actually contribute.

Interpreting latent thoughts. Zhang et al. (2025) find Coconut tokens "resistant to perturbation" yet lacking reasoning-critical information, repositioning them as pseudo-reasoning; Liang et al. (2026) give a mechanistic account of CODI on sequential tasks, finding decodable "bridge" states at moderate hop-length but a partial, collapse-prone pathway at longer lengths; Parekh et al. (2026) show, in a backdoor setting, that individual latent vectors can still encode the correct answer even as the model outputs the wrong one, locating the decisive information in the *collective* trajectory. These results (robust-but-empty, depth-dependent collapse, and collective-not-individual coding) are

exactly the phenomena our three-way separation predicts: they measure different operations. We add multi-seed controls, a content-specific shuffle intervention, a depth axis, and a real-task replication.

Computation vs. content. Transformers can exploit meaningless filler or pause tokens for extra serial computation independent of token content (Pfau et al., 2024; London et al., 2025; Baherwani et al., 2026). Our finding that latent thoughts can matter *positionally* while their content is fungible is the continuous-space analogue and clarifies when a robust thought carries information versus merely supplies computation.

Faithfulness and monitorability. A parallel literature asks whether *explicit* CoT faithfully reflects computation (Chen et al., 2025) and whether it can be monitored for safety (Korbak et al., 2025; Emmons et al., 2025; Chan et al., 2025). Latent CoT removes the text channel entirely; our results bear on what a latent monitor must read: content, not merely the presence of robust activity.

Intervention methods. Our tools are causal interventions in the tradition of activation patching and causal tracing (Meng et al., 2022; Geiger et al., 2025; Heimersheim & Nanda, 2024); we heed known pitfalls of patching-based attribution (Makelov et al., 2024) by triangulating three interventions rather than relying on one.

3 Setup and interventions

Task. We use k-hop modular arithmetic over $\mathbb{Z}_P$: a start value followed by k operations (e.g. "$3+5\times2\ldots=$"), with $P{=}7$ and $k{=}4$ operations (each an operator–operand pair), evaluated strictly left-to-right with no operator precedence. Every digit $0\ldots6$ is a single token under the real Pythia tokenizer, so answers and intermediate values are single tokens and batching is uniform-length. The task requires $k{-}1{=}3$ sequential reasoning steps. For any example we form an exact *minimal-pair counterfactual* by changing one operand at hop j, which changes every downstream intermediate value but nothing upstream, giving ground truth for "what a thought at step j should encode."

Models and training. We train Pythia-160m and Pythia-410m (Biderman et al., 2023) in two latent-CoT regimes: our implementations of **Coconut** (feed back the last hidden state) and **CODI** (self-distillation from an explicit-CoT teacher). Recurrence *depth* is a knob: with `tps` (thoughts-per-step) latent thoughts, a model runs $T = (k{-}1)\cdot\texttt{tps}$ continuous thoughts, so $\texttt{tps}\in\{1,2,3\}$ gives depth $T\in\{3,6,9\}$ within one architecture and one training recipe. Depth-sweep runs use an identical recipe; we accuracy-match by reporting only checkpoints reaching ≥ 0.98 clean accuracy unless stated. We also study a released **CODI-Llama-1B** checkpoint (Llama-3.2-1B backbone) on GSM8K (Cobbe et al., 2021), reproducing its greedy latent rollout ($T{=}6$).

Three interventions. For a rollout with thoughts $z_0,\ldots,z_{T-1}$ and clean answer $\hat{a}$: *(i) Random perturbation / basin radius.* Add isotropic Gaussian noise of magnitude s (in ℓ_2 units of the clean→counterfactual displacement, from a single random minimal pair) to z_t and estimate the flip fraction over $n{=}512$ examples; the basin radius r_t is the smallest swept s at which $\geq 50\%$ of answers flip, with $r_t{=}\infty$ (*noise-inert*) meaning no perturbation in the swept range (up to tens of times the displacement) flips the majority. We distinguish two senses of robustness that this measure separates: a thought is *finite-basin robust* when r_t is large but finite (it tolerates perturbations several times the counterfactual displacement yet can still be flipped), and *noise-inert* when $r_t{=}\infty$. We summarize the latter, stronger property by $\mathrm{frac_inert} = \frac{1}{T}\sum_t \mathbf{1}[r_t{=}\infty]$. *(ii) Counterfactual patch / capture.* Replace z_t with the thought the model computes for the minimal-pair counterfactual; *capture* is the fraction of examples whose answer becomes the counterfactual's answer. Because the counterfactual is defined at hop $\min(t, k{-}1)$, capture structurally cannot fire for surplus thoughts ($t\geq k{-}1$), so we use it only for $T{=}k{-}1$ and rely on ablation and shuffle at greater depth. *(iii) Content ablation. Mean-ablation* replaces z_t with the batch-mean thought (in-distribution, direction-free); *content-shuffle* replaces z_t with a real thought from another example (right distribution, wrong content). We summarize content necessity by $\text{content-load} = 1 - \mathrm{acc}_{\text{shuffle}}/\mathrm{acc}_{\text{clean}}$. Robustness is measured by (i); content necessity by (ii)–(iii). The dissociation is the observation that these disagree.

Table 1: At minimal depth (T=3, Pythia-410m, addition, accuracy-matched to 1.0), latent thoughts are both robust (large random basin) and load-bearing (capture≈ 1, high zero-flip). Basin radius is in units of the clean→counterfactual displacement. Contraction is $\|\Delta\|$ after one thought-step per unit injected (smaller = more contractive). Coconut's basins are wider (Mann–Whitney p=0.0025) and its map more contractive (p=0.015); this cross-method comparison used per-method learning rates and is descriptive (see §9). Anisotropy and learned-basin controls are in Appendix Table 4; the 160m replication is in Table 5.

Method (410m)	seeds	capture	flip (zero-ablate)	basin radius (random)
Coconut	7	≈ 1.00	0.83	6.18 ± 1.20
CODI	8	≈ 1.00	0.84	3.69 ± 0.75
		contraction per step: Coconut 0.120 ± 0.020		CODI 0.146 ± 0.016

4 Latent thoughts are both robust and load-bearing

We first establish, at minimal depth $T{=}k{-}1{=}3$, that continuous thoughts are simultaneously *finite-basin* robust and causally load-bearing, reconciling the apparent conflict in prior work. Table 1 summarizes 7–8 seeds per method at 410m (accuracy-matched to 1.0). Patching a thought with its counterfactual almost always installs the counterfactual's answer (capture ≈ 1.0) and zeroing a thought flips $\approx 83\%$ of answers, so thoughts are causally necessary; yet random perturbations several times the counterfactual displacement are needed to flip the majority (basin radius $\gg 1$). Both properties hold at once. At this depth no thought is noise-inert ($\mathrm{frac_inert}$=0.00 in Table 2): a large enough random perturbation still flips every thought, so the robustness here is the finite-basin kind, not the depth-induced noise-inertness we study next.

The basin is learned and task-structured. A randomly-initialized model has 0 accuracy and no basin, so the basin is learned rather than generic (Appendix Table 4). It is also strongly anisotropic: perturbing a thought *along* the counterfactual direction flips answers at radius 0.5, whereas random or orthogonal perturbations require radius 6–8 (Coconut) or 2–3 (CODI) (Table 4). Robustness is thus specific to task-irrelevant directions, consistent with an attractor aligned to the computation. The gap between Coconut and CODI basins replicates at 160m (Table 5).

5 Robustness and content-necessity are dissociable

We now vary recurrence depth (within one recipe) and measure robustness and content necessity separately (Table 2). As depth grows, the fraction of noise-inert thoughts rises monotonically ($0.00 \to 0.62 \to 0.87$): robustness increases with depth. But the *content* of those thoughts tells a different story. At depth 3 every thought is content-necessary (content-load 0.86) but none is noise-robust. At depth 6 a majority of thoughts are noise-inert *and* their content is still necessary: mean-ablating the inert thoughts costs 0.44 accuracy and content-load remains 0.84. At depth 9 thoughts are almost all noise-inert but their content necessity has fallen and destabilized (content-load 0.47 ± 0.25; mean-ablation of inert thoughts costs only 0.03). The "robust *and* load-bearing" combination is therefore non-monotonic in depth, peaking at moderate depth (Figure 2).

The dissociation across models. Figure 3 plots content necessity against noise-robustness for every trained model and for the released CODI-1B. The two axes are not tied: at a fixed high robustness ($\mathrm{frac_inert} \approx 0.87$), content necessity ranges from 0.17 to 0.83. Models with surplus "padding" depth occupy the robust-but-low-content region; CODI-1B ($\mathrm{frac_inert} \approx 0.5$, $3/6$ thoughts inert) occupies the robust-and-high-content region. A thought being robust to random noise tells us little about whether its content does the reasoning.

Why a single metric hides this. The two operations answer different questions: random perturbation asks whether *any* nearby vector yields the same answer (absorbed by a contractive map, §7), while ablation and shuffle ask whether the thought's *specific content* is needed. A thought can lie in a wide basin (robust) while its content is still necessary (a distributed, error-correcting code),

Table 2: Robustness rises with depth while per-thought content necessity is highest at minimal depth and falls at excess depth (Pythia-410m, addition, `tps` recipe held fixed, accuracy-matched). frac_inert: fraction of thoughts with no majority-flip under random perturbation. mean-abl. drop: accuracy drop from mean-ablating an inert vs. a non-inert ("used") thought. content-load: $1 - \mathrm{acc}_{\mathrm{shuffle}}/\mathrm{acc}_{\mathrm{clean}}$, mean±sd over seeds.

depth T	seeds	frac_inert	mean-abl. drop @inert	mean-abl. drop @used	content-load
3	3	0.00	n/a	0.85	0.862 ± 0.002
6	4	0.62	0.44	0.37	0.835 ± 0.041
9	5	0.87	0.03	0.12	0.473 ± 0.245

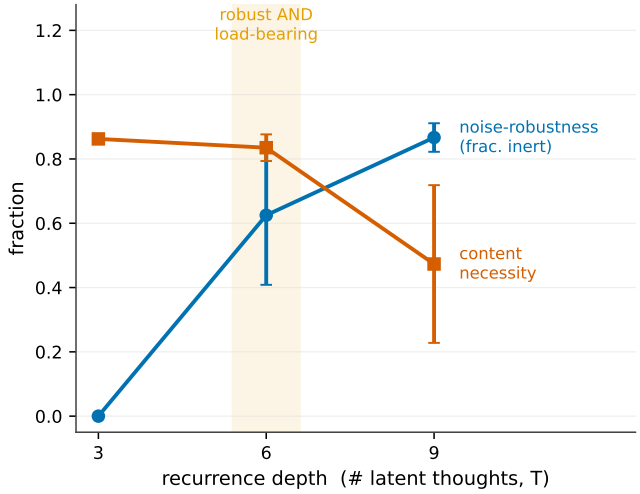

Figure 2: Depth trades robustness against per-thought content (Table 2). Robustness (frac_inert) rises monotonically while content necessity is high at low depth and falls at excess depth; both are simultaneously high only at moderate depth (T=6). Error bars: sd over seeds.

or lie in a wide basin while carrying nothing (padding). This is why counterfactual-*capture* and content-ablation disagree for surplus thoughts, and why we report all three interventions.

6 The dissociation on a real model: CODI-GSM8K

We reproduce the released CODI-Llama-1B rollout on GSM8K (T=6) and apply the same interventions (Figure 4, Table 3). Four properties hold at once. *Collectively essential:* accuracy rises monotonically as more thoughts are used, from 0.133 (no thoughts) to 0.387 (all on n=256), and zero-ablating all thoughts drops accuracy to 0.168, near the 0.133 no-thought floor. *Individually redundant:* ablating any single thought changes 17–31% of individual answers but costs only 0.4–4.3% net accuracy; the flips are near-symmetric (roughly as many wrong→right as right→wrong), so they largely cancel and no single thought is decisive. *Noise-robust:* basin radii are 16–64× the thought's ℓ_2 norm and three of six thoughts never reach a majority flip. *Content-load-bearing:* shuffling thought content across examples collapses accuracy from 0.436 to 0.165 ± 0.004 (5 seeds, full 1,319-question test), the floor; matched-statistics Gaussian replacement gives 0.209 ± 0.004. Together these behave like a distributed, error-correcting code (noise-robust and individually sacrificial through near-symmetric flip cancellation, yet collectively and content-wise essential) on a released model and a real task, concretely instantiating the observation that the decisive information lives in the collective trajectory rather than any single vector (Parekh et al., 2026).

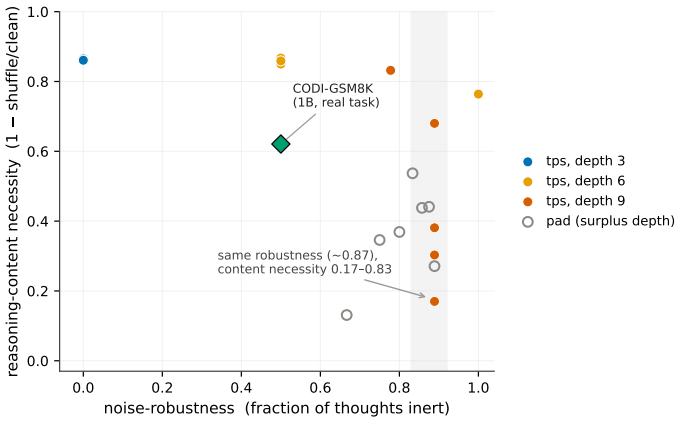

Figure 3: Noise-robustness does not predict content necessity. Each point is one trained model (toy task, accuracy-matched); the diamond is CODI-1B on GSM8K. At fixed high robustness ($\text{frac_inert} \approx 0.87$), content necessity ranges from 0.17 to 0.83.

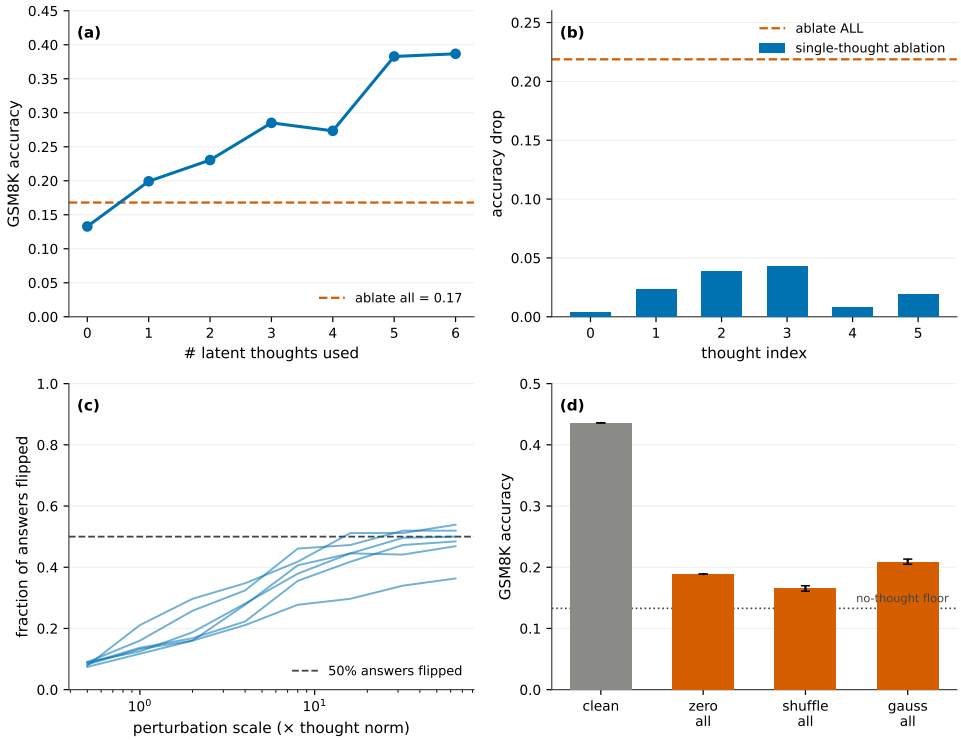

Figure 4: CODI-Llama-1B on GSM8K. Latent thoughts are (a) collectively essential (ablating all $\rightarrow$ the no-thought floor), (b) individually redundant (single-thought ablation barely dents accuracy), (c) noise-robust (most thoughts never reach a majority answer-flip even at $64\times$ the norm; measured flip fraction), yet (d) content-essential (zero/shuffle/gauss ablation $\rightarrow$ the floor). Bars in (d): mean$\pm$sd, 5 seeds, full test set. Values in Table 3.

7 Mechanism: Contraction Produces Robustness

Why are deeper thoughts more noise-inert? The thought-to-thought map is contractive: one step reduces an injected perturbation to $\approx 12\%$ of its norm (Table 1), so a perturbation injected at thought t decays by ρ^{T-t} before the readout. A perturbation therefore survives to the final thought at only

Table 3: CODI-Llama-1B on GSM8K under the three interventions. Content ablations are mean±sd over 5 seeds on the full test set (n=1319); the depth curve and single-thought ablations use n=256.

quantity	value
clean accuracy	0.436 (full test); 0.387 (n=256)
accuracy, no latent thoughts used	0.133
accuracy, zero-ablate all thoughts (n=256)	0.168
single-thought ablation: net accuracy drop	0.004–0.043
single-thought ablation: fraction of answers changed	0.17–0.31
basin radius (random perturbation)	16–64× norm (3/6 inert)
zero / shuffle / gauss-all accuracy (full test)	0.189 / 0.165 ± 0.004 / 0.209 ± 0.004
content-load ($1 - \mathrm{acc}_{\mathrm{shuffle}}/\mathrm{acc}_{\mathrm{clean}}$)	0.62 ± 0.01

$\approx$ 1–13% for upstream positions, and progressively more for later positions (Appendix Figure 5). Deeper rollouts place upstream thoughts more contraction steps from the answer, making them noise-inert regardless of whether their content matters, exactly the depth trend in Table 2. Consistent with its wider basins, Coconut's map is more contractive than CODI's (Table 1). Contraction thus explains the robustness axis while leaving content necessity (a property of what information the map carries, not how it damps noise) free to vary independently.

8 Discussion

Robustness-only metrics mislead. A perturbation test alone will call a content-essential deep thought "inert" and a content-free padding thought "robust." Claims that latent thoughts are "robust but empty" (Zhang et al., 2025) and that they "carry decodable states" (Liang et al., 2026) are both locally correct: they measure different operations at different depths. Separating the axes reconciles them.

Implications for monitoring. If latent reasoning is to be monitored (Korbak et al., 2025), the target is thought *content*, not the presence of robust activity: a monitor keyed on perturbation-stability would be blind to whether reasoning content is present. Our shuffle intervention is a cheap content-necessity probe that requires no gradient or trained probe. Because the real-model code is distributed and individually redundant, a probe applied to a single thought can understate its role; reliably monitoring such a code may require intervening on several thoughts jointly (as in our shuffle-all result) rather than one at a time.

Positions vs. content. That surplus thoughts can be robust yet content-free mirrors filler/pause-token results in explicit models (Pfau et al., 2024; London et al., 2025): added latent steps can supply serial computation whose specific content is fungible. Whether a latent thought carries information or merely computation is a question our three interventions can answer.

9 Limitations

Our controlled task is synthetic k-hop arithmetic; it enables exact counterfactuals and accuracy matching but is not natural language. Our real-task evidence is one released model on one dataset (GSM8K); extending the dissociation to deductive reasoning with logical branches and to semantic benchmarks, and contrasting continuous latent thoughts with discrete hidden reasoning, is future work. The cross-method Coconut-vs-CODI basin comparison (§4) used per-method learning rates and is reported as descriptive, not a controlled architectural claim; all *within-method* results (the dissociation, the depth sweep, the CODI real-task analysis) hold the recipe fixed. Content necessity at depth 9 is high-variance across seeds (±0.25), which we report rather than smooth over. Counterfactual capture is structurally capped for surplus thoughts, which is why deep-thought conclusions rest on ablation and shuffle. Finally, mean-ablation and content-shuffle probe distinct notions, whether a thought's specific content is required for the computation to proceed versus whether

injecting distribution-matched but wrong content disrupts it; at surplus depth these can diverge (inert thoughts tolerate mean-ablation, yet the collective remains shuffle-sensitive), so we report both rather than collapse them.

10 CONCLUSION

Noise-robustness and reasoning-content necessity are distinct, dissociable properties of latent chain-of-thought. Robust latent thoughts are not thereby reasoning-bearing, and the single perturbation test the field often uses conflates the two. Three interventions (random perturbation, counterfactual patch, and content-shuffle) separate them, revealing a depth-dependent trade-off on a controlled task and a distributed, robust, content-essential code on a released CODI-1B model. We hope the separation, and the cheap content-shuffle probe, sharpen how latent reasoning is interpreted and monitored.

REPRODUCIBILITY STATEMENT

All numerical claims in the paper are drawn from logged intervention outputs. The task, model, training recipe, and the three interventions are specified in §3; per-checkpoint results and metric definitions underlying Tables 1–3 and Figures 1–4 (and Appendix Tables 4–5 and Figure 5) are produced by the released code and result logs, including the exact accuracy-matching gate (≥ 0.98) and seed counts stated in each table. The released CODI-Llama-1B checkpoint and GSM8K are public; we reproduce its rollout as described in §6.

REFERENCES

Vatsal Baherwani et al. Not all LLM reasoning is visible in the chain-of-thought. *arXiv preprint*, 2026.

Stella Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Yik Siu Chan et al. Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models. *arXiv preprint*, 2025.

Yanda Chen et al. Reasoning models don't always say what they think. *arXiv preprint*, 2025.

Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Scott Emmons et al. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint*, 2025.

Ying Fan et al. Bridging the gap between latent and explicit reasoning with looped transformers. *arXiv preprint*, 2026.

Atticus Geiger et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26, 2025.

Shibo Hao et al. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint*, 2024.

Takeshi Kojima et al. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tomasz Korbak et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint*, 2025.

Jian Liang et al. Do latent-CoT models think step-by-step? a mechanistic study on sequential reasoning tasks. *arXiv preprint*, 2026.

Charles London et al. Pause tokens strictly increase the expressivity of constant-depth transformers. *arXiv preprint*, 2025.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations (ICLR)*, 2024.

Swapnil Parekh et al. Thinking wrong in silence: Backdoor attacks on continuous latent reasoning. *arXiv preprint*, 2026.

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.

Zhenyi Shen et al. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of EMNLP*, 2025.

Minghan Wang et al. Towards inference-time scaling for continuous space reasoning. *arXiv preprint*, 2025.

Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xilin Wei et al. SIM-CoT: Supervised implicit chain-of-thought. *arXiv preprint*, 2025.

Yuyi Zhang et al. Do latent tokens think? a causal and adversarial analysis of chain-of-continuous-thought. *arXiv preprint*, 2025.

A Appendix

A.1 Controls: learned and anisotropic basins

Table 4 reports the on/off-manifold and random-initialization controls referenced in §4 (410m, one representative seed per condition). A randomly-initialized network solves nothing and has no basin, so the basin is learned; and perturbing along the task (counterfactual) direction flips answers at a far smaller radius than random or orthogonal perturbations, so the basin is task-structured.

Table 4: Basin radius by perturbation direction, and the random-initialization control (410m, T=3).

model	accuracy	basin (task-dir)	basin (orthogonal)	basin (random)
Coconut (trained)	1.00	0.5	6–8	6–8
CODI (trained)	1.00	0.5	2–3	2–3
random-init	0.00	none	none	none

A.2 Size replication (Pythia-160m)

The basin-width gap between Coconut and CODI replicates at 160m (Table 5); Coconut basins are more variable at the smaller scale.

A.3 Mechanism: perturbation survival

Figure 5 shows the per-position survival referenced in §7: a perturbation injected at an upstream thought reaches the final thought at only $\approx$ 1–13% of its norm, and survival increases toward the readout.

Table 5: Basin radius (random) at Pythia-160m, T=3, accuracy-matched (8 seeds each). Mann–Whitney p=0.048.

Method (160m)	basin radius (random)	ratio
Coconut	6.75 ± 3.71	2.0×
CODI	3.38 ± 0.39	

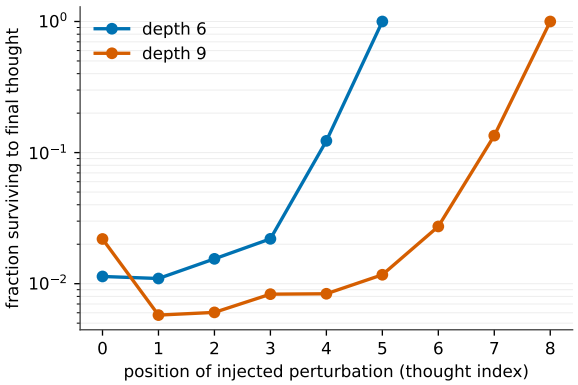

Figure 5: Fraction of an injected perturbation surviving to the final thought, by injection position (addition, depth 6 and 9; log scale). Upstream perturbations are strongly damped; later ones survive more.

A.4 Reconciliation at minimal depth

Figure 6 shows the T=3 basin comparison of Table 1; both methods have capture ≈1.0 (all thoughts load-bearing) and Coconut's basins are wider. As noted, this cross-method comparison is recipe-uncontrolled.

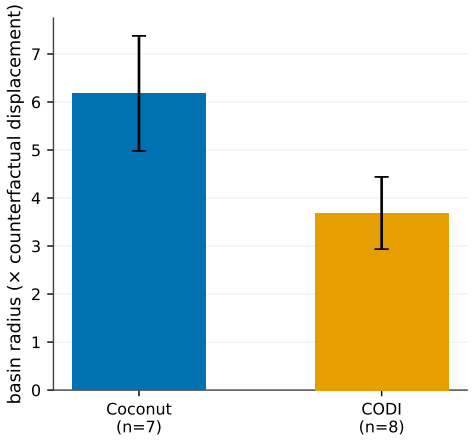

Figure 6: Basin radius at T=3, Coconut vs. CODI (accuracy-matched, 7–8 seeds). Recipe-uncontrolled (learning rates differ); descriptive.

A.5 Use of large language models

A large language model was used as a general-purpose assistant for code scaffolding, figure generation, and prose editing. All experimental results, numerical claims, and analyses were produced by

and verified against the authors' logged experiment outputs; the authors take full responsibility for the paper's content.