Unlocking the Black Box of Latent Reasoning: An Interpretability-Guided Approach to Intervention

Anonymous ACL submission

Abstract

Latent reasoning has emerged as a promising alternative to explicit Chain-of-Thought (CoT) prompting where language models perform multi-step inference within their continuous hidden states, offering significant gains in computational efficiency. However, the internal mechanisms of this process remain largely opaque, hindering its trustworthiness and further development. This paper provides the first systematic interpretability analysis of the continuous-thought paradigm in latent reasoning. Through a three-pronged investigation combining structural probes, causal interventions, and geometric analysis, we reveal that latent thought vectors are not uninterpretable artifacts but function as faithful, compressed representations of coherent reasoning steps. We further identify that early vectors in the latent sequence act as critical causal hubs, exerting disproportionate control over the final outcome. Building directly on these insights, we introduce a novel suite of training-free, decoding-time interventions designed to refine and steer the latent reasoning process. Experiments show that our interpretability-guided interventions consistently improve reasoning accuracy across LLMs across different latent reasoning methods in the continuous-thought paradigm. Code will be made available upon publication.

1 Introduction

Large Language Models (LLMs) have evolved from simple pattern completion to performing sophisticated multi-step reasoning (DeepSeek-AI et al., 2025). Much of this progress stems from techniques like Chain-of-Thought (CoT) prompting (Wei et al., 2023), where models generate explicit, token-based rationales to structure their inference process. While effective, this reliance on verbalized reasoning introduces significant drawbacks. Explicit CoTs can be excessively long (Wang et al., 2025a), incurring substantial latency and computational costs. Furthermore, their textual nature provides a low-bandwidth medium for complex computation (Hao et al., 2024), often containing redundant information and increasing susceptibility to compounding errors.

These limitations motivate a shift toward *latent reasoning* (Zhu et al., 2025; Li et al., 2025), where multi-step inference unfolds within the model's continuous hidden states without generating intermediate text. This approach enables more efficient computation by bypassing the need for explicit token generation. Existing methods explore various paradigms, such as optimizing hidden state trajectories (Tan et al., 2025; Chen et al., 2025; Zhang et al., 2025) or reusing network layers (Ball et al., 2025; Sadhu et al., 2025; Blom et al., 2025; Nguyen et al., 2025a; Ju et al., 2025a) to simulate deeper computation. Among these, the *continuous-thought* paradigm (Hao et al., 2024; Shen et al., 2025), which employs special latent vectors to propagate information across reasoning steps. It has shown particular promise due to its simplicity and compatibility with standard decoding pipelines.

Despite its potential, the continuous-thought paradigm faces two critical gaps in understanding. First, its interpretability remains largely unverified: it is unclear whether these latent vectors genuinely encode semantic steps of a reasoning process or are merely artifacts of the training objective. Second, existing studies are typically limited to smaller-scale models, leaving the scalability and behavior of this paradigm at larger scales underexplored. This paper aims to bridge these gaps: Not by proposing a new training method, but by establishing rigorous principles and diagnostic tools for interpreting and intervening in the latent reasoning process.

To address the first gap whether latent thought vectors genuinely encode semantic reasoning steps, we develop a suite of three complementary interpretability probes (Davari et al., 2022; Kornblith et al., 2019), designed to examine their structure,

content, and causal role in the reasoning process. First, through structural alignment analyses, we find that the geometry of latent thought vectors closely mirrors that of explicit CoT rationales, suggesting a shared representational scaffold. Second, we show that a simple linear map can reconstruct textual reasoning steps from these latent vectors with high fidelity, confirming they serve as compressed yet faithful encodings of intermediate reasoning. Finally, causal intervention experiments demonstrate that targeted edits to these vectors especially in early reasoning steps to predictably alter the model's final answer, establishing their functional role in the inference pipeline. Building on these insights, we design a series of training-free, decode-time interventions. By manipulating cached latent states by projecting them along directions identified by our probes, we can both validate our interpretability hypotheses and measurably improve reasoning performance without any parameter updates. These consistent gains from simple, norm-preserving edits highlight the untapped potential residing within the model's existing latent pathways.

Our contributions are threefold:

1. We provide the first systematic interpretability analysis of the continuous-thought paradigm, offering structural, pointwise, and causal evidence that its latent vectors serve as coherent reasoning blueprints.

2. We extend this analysis to large-scale models, presenting the first empirical characterization of how the geometric structure and reasoning efficacy of continuous thought evolve with model size—addressing a key limitation of prior work confined to smaller architectures.

3. We introduce a set of training-free, decode-time interventions that not only validate our findings but also enhance reasoning performance, demonstrating that latent capabilities can be unlocked without additional fine-tuning.

2 Related Works

2.1 Paradigms for Latent Reasoning

Latent reasoning seeks to realize multi-step inference inside hidden states, avoiding explicit intermediate tokens while retaining standard decoding for the final answer. Prior work advances three main mechanisms. Latent optimization refines hidden trajectories in place during decoding or distills long textual rationales into compact continuous representations. Signal-guided control steers trajectories with auxiliary signals such as confidence estimates, verifier outputs, or decode-time selection criteria, and may bias states toward informative embedding subspaces. Layer-recurrent(Nguyen et al., 2025b; Ju et al., 2025b) execution emulates deeper computation without lengthening the sequence by reusing layers or activations, for example through architectural loops or by feeding a layer's output back as the next input. Our study focuses on the continuous-thought latent-span variant, in which reserved latent vectors are written into the prefix to carry information across steps, as popularized by CoConut-style models. Although effective, most demonstrations target small or mid-size models, and their scalability and interpretability at larger scales remain underexplored.

2.2 Interpretability of Latent Reasoning

Interpretability efforts examine where and how hidden computation acquires stepwise structure and causal influence. Mechanistic analyses (Pimparkhede et al., 2024; Leban, 2025; Tian et al., 2025; Hong and Fang, 2025) track the emergence of staged computation within transformer internals, relating attention flow and layer specialization to progressive subgoal formation. Behavioral studies (Wu et al., 2024; Yokouchi et al., 2024; Calandriello et al., 2024; Davis et al., 2025) infer latent processing from macroscopic signatures such as abrupt memorization-to-generalization transitions and the internalization of intermediate steps that manifest as step skipping at inference. Representational analyses (Chambers et al., 2024; Yerrapragada et al., 2024; Haviv et al., 2024; Wang et al., 2025b; Raha et al., 2024) link hidden trajectories to explicit reasoning via centered kernel alignment, linear decoding, and lightweight mappers, and corroborate these relations through targeted causal edits that ablate, steer, or transplant states. While these strands substantiate the presence of latent computation, they leave open a unified account for specific paradigms. We address this gap for the continuous-thought latent-span setting by integrating structural alignment with pointwise recoverability and demonstrating directional causal efficacy, then operationalizing the resulting priors as training-free, decode-time interventions. The protocol both supports interpretability claims for this

paradigm and yields consistent gains in reasoning accuracy, connecting explanation with capability improvement.

3 Interpretation of Latent Reasoning

To investigate whether latent reasoning genuinely performs structured, multi-step inference, we conduct a three-part analysis. We first probe the semantic content of the latent vectors by measuring their correspondence with explicit Chain-of-Thought (CoT) representations (Section 3.2). Next, we perform a series of causal interventions to test whether manipulating these vectors predictably controls the model's output (Section 3.3). Finally, we explore the architectural and geometric properties that enable this emergent computational structure (Section 3.4).

3.1 Preliminary

Our work focuses on a mode of reasoning where inference occurs within the model's continuous hidden states, rather than through verbalized CoT. A standard causal language model with parameters θ defines the probability of a sequence $w_{1:T}$ as:

$$p_\theta(w_{1:T}) = \prod_{t=1}^{T} p_\theta(w_t \mid w_{<t}). \tag{1}$$

While explicit CoT involves minimizing cross-entropy over a sequence of textual reasoning steps, latent reasoning reserves a special span of K positions within the prompt that do not correspond to vocabulary tokens. The computation within this span is realized through a sequence of hidden vectors we term continuous thoughts.

Formally, given an input embedding matrix $\mathbf{E} \in \mathbb{R}^{T \times d}$, a latent span of length K is introduced. The model deterministically populates this span by sequentially generating K continuous thought vectors, $\mathbf{z}_{1:K}$. At each latent position ℓ_k, the vector is derived from the previous position's final hidden state and then used as the input embedding for the current position:

$$\mathbf{z}_k \triangleq \mathbf{h}^{(L)}_{\ell_k - 1}, \quad \text{and} \quad \tilde{\mathbf{E}}_{\ell_k} \leftarrow \mathbf{z}_k, \tag{2}$$

where $\mathbf{h}^{(L)}_t$ is the last-layer hidden state at position t and $\tilde{\mathbf{E}}$ is the resulting "filled" embedding matrix. Conditioned on $\tilde{\mathbf{E}}$, the model proceeds with standard autoregressive decoding to generate the final answer y:

$$p_\theta(y \mid x) \equiv p_\theta\big(y \mid \tilde{\mathbf{E}}(x)\big). \tag{3}$$

Crucially, the latent vectors $\mathbf{z}_k$ are not directly supervised. Their values are learned implicitly through the standard next-token prediction loss, which is applied only to the visible target tokens $\mathcal{V}_{\text{vis}}$ (i.e., the answer):

$$\mathcal{L}(\theta) = -\sum_{t \in \mathcal{V}_{\text{vis}}} \log p_\theta\big(w_t \mid w_{<t};\, \tilde{\mathbf{E}}(x)\big). \tag{4}$$

3.2 Probing the Semantic Content of Latent Vectors

Our first hypothesis is that the sequence of continuous thoughts, $\mathbf{z}_{1:K}$, serves as a compressed, continuous analogue to a textual CoT. To test this, we assess the correspondence between the latent representations and the hidden states generated by an equivalent model executing an explicit CoT.

Establishing Correspondence. Let $\mathcal{T}$ be the set of textual step boundaries in a teacher-forced CoT. We extract the final-layer hidden state at the end of each step $t \in \mathcal{T}$ to serve as its canonical representation: $\mathbf{c}_t(x) \triangleq \mathbf{h}^{(L)}_{\text{end}(t)}(x)$. To create a comparable latent representation, we locally average a small, contiguous block of continuous thoughts that are heuristically aligned with step t, forming a step-level latent surrogate:

$$\bar{\mathbf{z}}_t(x) \triangleq \frac{1}{|\mathcal{K}(t)|} \sum_{k \in \mathcal{K}(t)} \mathbf{z}_k(x). \tag{5}$$

This aggregation allows us to test for step-wise correspondence without making strong assumptions about a one-to-one mapping between latent vectors and reasoning steps.

Multi-faceted Probing Analysis. We employ three complementary methods to probe the relationship between the latent surrogates $\{\bar{\mathbf{z}}_t\}$ and their explicit counterparts $\{\mathbf{c}_t\}$. First, we measure their **structural alignment** using linear Centered Kernel Alignment (CKA), which quantifies the similarity of the geometric arrangement of data points across different instances. CKA is defined as:

$$\text{CKA}(\tilde{\mathbf{Z}}_t, \tilde{\mathbf{C}}_t) = \frac{\|\tilde{\mathbf{Z}}_t^\top \tilde{\mathbf{C}}_t\|_F^2}{\|\tilde{\mathbf{Z}}_t^\top \tilde{\mathbf{Z}}_t\|_F \cdot \|\tilde{\mathbf{C}}_t^\top \tilde{\mathbf{C}}_t\|_F}, \tag{6}$$

where $\tilde{\mathbf{Z}}_t$ and $\tilde{\mathbf{C}}_t$ are matrices whose rows are the centered latent surrogates and CoT step representations, respectively, for a set of problems.

Second, we test for pointwise recoverability by training a simple linear map f_ϕ to predict a CoT representation $\mathbf{c}_t(x)$ from its corresponding latent

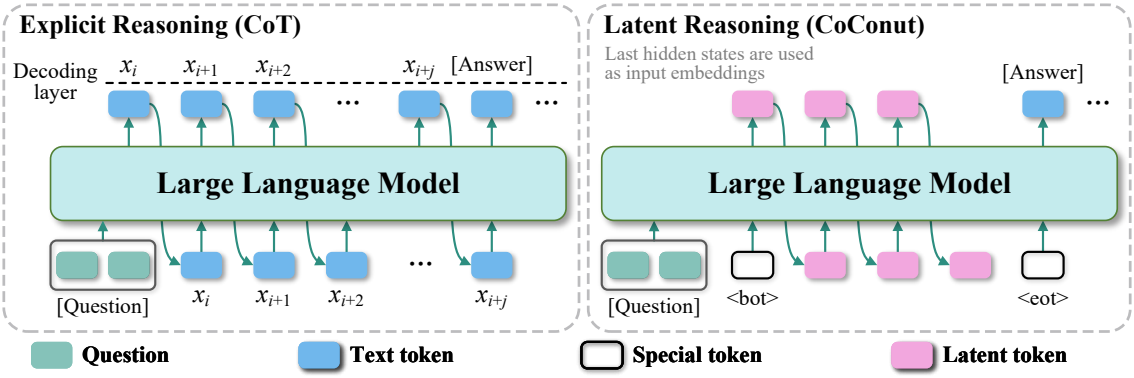

Figure 1: **Explicit Reasoning versus Latent Reasoning**. Left: Explicit Reasoning externalizes intermediate reasoning as a token sequence $(x_i, x_{i+1}, \ldots, x_{i+j})$.Right: Latent Reasoning reserves a latent span and at each reserved slot, feeds the preceding final-layer hidden state back as the next input, yielding a continuous thought that conditions subsequent decoding.

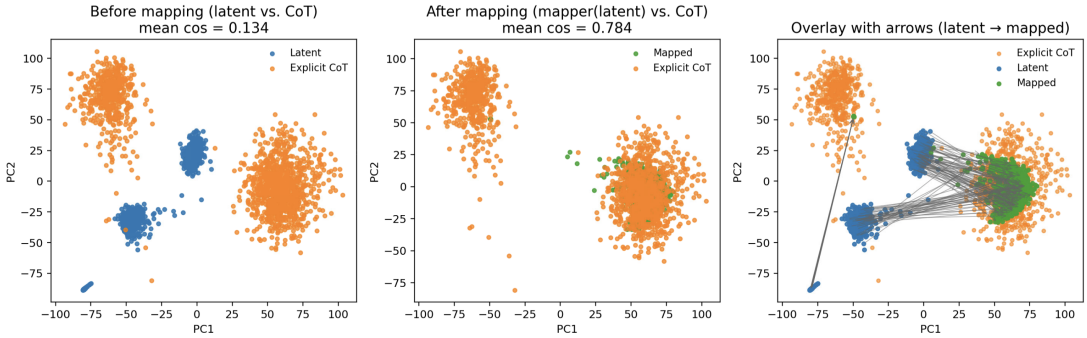

Figure 2: **Mapper-based alignment of latent thoughts to explicit CoT**. (a) Before mapping, latent and CoT points are separated in PCA space. (b) A linear mapper f_ϕ aligns latents with CoT clusters. (c) Overlay with arrows (latent→mapped) visualizes the systematic displacement toward CoT.

surrogate $\bar{\mathbf{z}}_t(x)$. The model is trained to maximize cosine similarity, a stringent test of linear decodability:

$$\mathcal{L}_{\cos}(\phi) = \mathbb{E}_{t,x}\left[1 - \cos\big(f_\phi(\bar{\mathbf{z}}_t(x)), \mathbf{c}_t(x)\big)\right]. \tag{7}$$

Finally, we conduct a non-parametric lexical probe to see if a latent surrogate contains information about the tokens in the corresponding reasoning step. We feed $\bar{\mathbf{z}}_t(x)$ through the model's language modeling head and measure the probability mass assigned to the top tokens of the actual CoT step:

$$p_{\text{vocab}}(\cdot \mid \mathbf{z}) = \text{softmax}\big(\mathbf{W} \times \text{LN}(\mathbf{z}) + \mathbf{b}\big). \tag{8}$$

Results. As shown in Table 1, our probes confirm a strong semantic correspondence. The high CKA score indicates that the geometry of the latent reasoning space mirrors that of explicit CoT. Furthermore, the high cosine similarity achieved by the linear map demonstrates that CoT step information is linearly recoverable from the latent vectors. The lexical probe further reinforces this, showing that a single latent surrogate can reconstruct key tokens from its corresponding textual step. This evidence, visualized in Figure 2, suggests that continuous thoughts are not uninterpretable artifacts but rather faithful, compressed representations of the reasoning process.

3.3 Analyzing the Causal Role of Latent Thought Sequences

Having established a correlational link between latent vectors and reasoning steps, we now investigate their causal role. If continuous thoughts truly guide the reasoning process, then targeted interventions on them should predictably alter the model's final answer. We measure the effect of an intervention $\mathcal{I}$ on a latent sequence $\mathbf{Z}$ by the change in

Probing Method	Value
Structural Alignment	
Linear CKA Score (Eq. 6)	**0.72**
Pointwise Recoverability	
Cosine Sim. after Mapping (Eq. 7)	**0.75**
Δ over Identity Baseline	+0.61
Lexical Content	
Token Probe Hit@1 (Eq. 8)	**0.38**

Table 1: **Results from probing the semantic content of latent vectors**. High scores across structural, pointwise, and lexical probes provide convergent evidence that latent vectors encode information analogous to explicit CoT steps.

log-probability of the target answer $\mathbf{y}^\star$:

$$\Delta \log p_{\text{tgt}} \triangleq \log p(\mathbf{y}^\star|\mathcal{I}(\mathbf{Z})) - \log p(\mathbf{y}^\star|\mathbf{Z}). \quad (9)$$

We also report the percentage of instances where the greedily decoded answer flips post-intervention.

Intervention Strategies. We probe causal influence through three complementary intervention families applied to cached latent vectors. To assess positional salience, we truncate the latent sequence after the k-th vector, denoted $\text{drop@}k$, thereby removing downstream stages of latent computation and measuring the resulting change in prediction. To test fine-grained controllability, we introduce small, norm-preserving edits to an early latent state: a spherical interpolation that reorients the vector toward an answer-supporting boundary (Eq. 15), and a single normalized gradient step that nudges the vector in the direction that increases the likelihood of the correct answer (Eq. 16). To evaluate transferability, we transplant the entire latent sequence from one instance into another and examine whether the model's answer shifts toward the donor's target, treating the sequence as a putative internal reasoning plan. Together, these interventions jointly characterize where information is concentrated, whether it can be steered with minimal edits, and to what extent latent trajectories encode reusable guidance for decoding.

Results. Our interventions, summarized in Table 2, reveal a clear causal structure. The ablation experiments (Figure 3) demonstrate that early latent vectors are disproportionately important; removing them severely degrades performance. Targeted, norm-preserving edits confirm that these vectors are not just important but also steerable: small nudges in semantically meaningful directions reliably shift the model's output probabilities and

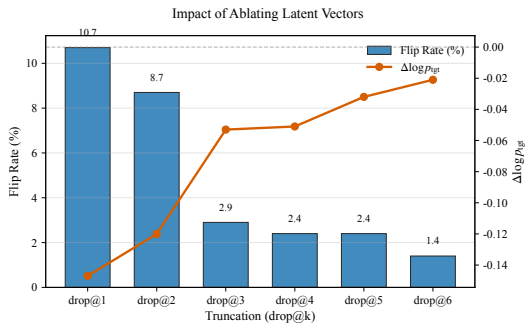

Figure 3: Impact of ablating latent vectors. The answer flip rate decreases as more of the initial latent vectors are preserved (from drop@0 to drop@K), confirming that early latent thoughts are causally critical for the final outcome.

can often flip the final answer. Finally, the transplant experiment provides the most compelling evidence. While transplanting a full textual CoT is disruptive, transplanting the latent sequence acts as a successful "plan transfer," biasing the recipient model toward the donor's answer without causing a catastrophic breakdown. This suggests the latent sequence encodes a robust, abstract reasoning trajectory.

3.4 Uncovering Architectural and Geometric Underpinnings

What properties of the Transformer architecture enable this stable, structured latent computation? We hypothesize two key factors: the geometry induced by tied input-output embeddings and the emergence of a monotonically decreasing energy function along the latent chain.

Weight Tying as a Nearest-Neighbor Projector. In models with tied embeddings, the output projection matrix is the transpose of the input embedding matrix ($W_o = E^\top$). Consequently, selecting the next token $\hat{\imath}$ at low temperature is equivalent to finding the token embedding $\mathbf{e}_{\hat{\imath}}$ with the highest cosine similarity to the final hidden state $\mathbf{h}$:

$$\hat{\imath} = \arg\max_i \langle \mathbf{h}, \mathbf{e}_i \rangle. \quad (10)$$

The standard "decode-then-re-embed" cycle can thus be viewed as a nearest-neighbor projection $\mathcal{T}(\mathbf{h}) = \mathbf{e}_{\hat{\imath}}$, which maps $\mathbf{h}$ onto the discrete codebook of token embeddings. Our latent update mechanism, which directly feeds $\mathbf{h}$ as the next input, is a continuous approximation of this cycle. This approximation is accurate when $\mathbf{h}$ is already close to an embedding vector $\mathbf{e}_{\hat{\imath}}$, i.e., when $\|\mathbf{h} - \mathcal{T}(\mathbf{h})\|^2$

Intervention Type	Setting	Flip Rate (%)	$\Delta \log p_{\text{tgt}}$	Directional Rate (%)
A. Ablation (position importance)				
	drop@0 (remove all)	10.7	-0.147	–
	drop@1 (keep first)	8.7	-0.120	–
B. Targeted, norm-preserving edits				
	Slerp toward correct answer	21.5	+0.85	–
	One-step gradient update	24.1	+1.02	–
C. Transplant (plan-like function)				
	Latent Transplant	2.5	-0.05	**48.3**
	Textual CoT Transplant (control)	95.0	-4.51	30.0

Table 2: **Results of causal interventions on the latent reasoning sequence.** Higher flip rate and $\Delta \log p_{\text{tgt}}$ indicate stronger steering toward the target answer. "–" denotes metrics not applicable in that setting. **(A)** Ablating early vectors harms performance, confirming their causal importance. **(B)** Small, norm-preserving edits effectively steer the model. **(C)** Latent transplants bias the model toward the donor's answer with high directional fidelity, unlike textual CoT transplants.

is small. Untied embeddings introduce a tying gap, weakening this geometric correspondence and potentially destabilizing latent computation.

A Monotone Energy Over the Latent Chain. We further hypothesize that the latent trajectory follows a directed path from a high-energy "unresolved" state to a low-energy "resolved" state. We test this by learning a simple scalar energy function $H : \mathbb{R}^d \to \mathbb{R}$ optimized with a margin ranking loss to ensure energy decreases at each step along the chain from the last question token to the first answer token:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}\left[\max\{0, \gamma - (H(\mathbf{x}_t) - H(\mathbf{x}_{t+1}))\}\right]. \tag{11}$$

Results. As shown in Table 3, both hypotheses are supported by the data. Models with tied embeddings exhibit significantly better latent reasoning performance relative to their explicit CoT counterparts (higher accuracy ratio r_{Acc}). Moreover, the learned energy function is almost perfectly monotonic along the latent chain, with rank correlations approaching 1.0. This suggests the latent process is not a random walk but a highly structured, contractive mapping that guides the model state toward a final answer.

In summary, this chapter provides a multifaceted interpretation of the continuous-thought paradigm. We have shown that its latent vectors are not only semantically aligned with explicit reasoning steps but also exert precise causal control over the model's output. This stable computational process is underpinned by favorable architectural properties, establishing latent reasoning as a robust and interpretable alternative to explicit CoT.

Architectural Analysis Metric	**Value**
Impact of Weight Tying	
Accuracy Ratio (r_{Acc}) for Tied Model	**2.97**
Accuracy Ratio (r_{Acc}) for Untied Model	1.15
Monotonicity of Learned Energy	
Chain-Strict Monotonicity (% of chains)	**99.0%**
Spearman Rank Correlation	0.9998

Table 3: **Evidence for the architectural and geometric underpinnings of latent reasoning**. Tied embeddings significantly improve relative performance, and a learned energy function reveals a strongly monotonic structure along the latent trajectory.

4 Interventions on Latent Reasoning

Building upon the interpretability analysis in Section 3, we now operationalize our findings. We introduce a series of novel, *training-free* interventions that manipulate cached latent vectors at decode-time. Our goal is to demonstrate that by leveraging a theoretical understanding of the latent reasoning process, we can unlock and enhance the model's inherent inferential capabilities without any parameter updates.

4.1 Intervening at Causal Hubs

4.2 Experimental Setup

Latent Reasoning Paradigms. We evaluate our interventions on two methods related the same latent reasoning paradigm: CoConut appends a sequence of special tokens to the prompt, fills these placeholders with continuous thoughts; CODI first generates a long, explicit Chain-of-Thought, and then distills it into sequence of continuous vectors. These compressed vectors are then prefixed to the input to guide the final answer generation.

Algorithm 1 Mapper-Guided Transport and Subspace Alignment

Require: Cached latents $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, Mapper F, Projector P_r
Require: Hyperparameters $\alpha, \eta, \beta, \lambda$

$\mathbf{m}_0 \leftarrow (\mathbf{z}_{K-1} + \mathbf{z}_K)/2$
$\mathbf{t}_0 \leftarrow F(\mathbf{m}_0)$
$\widehat{\mathbf{u}} \leftarrow \mathrm{slerp}(\mathbf{m}_0/\|\mathbf{m}_0\|, \mathbf{t}_0/\|\mathbf{t}_0\|; \alpha)$
$\|\widehat{\mathbf{m}}\| \leftarrow (1-\eta)\|\mathbf{m}_0\| + \eta\|\mathbf{t}_0\|$
$\widehat{\mathbf{m}} \leftarrow \|\widehat{\mathbf{m}}\| \cdot \widehat{\mathbf{u}}$
$\mathbf{r}_{K-1} \leftarrow \mathbf{z}_{K-1} - \mathbf{m}_0$; $\mathbf{r}_K \leftarrow \mathbf{z}_K - \mathbf{m}_0$
$\mathbf{z}'_{K-1} \leftarrow \widehat{\mathbf{m}} + \beta\,\mathbf{r}_{K-1}$; $\mathbf{z}'_K \leftarrow \widehat{\mathbf{m}} + \beta\,\mathbf{r}_K$
$\mathbf{Z}' \leftarrow \{\mathbf{z}_1, \ldots, \mathbf{z}_{K-2}, \mathbf{z}'_{K-1}, \mathbf{z}'_K\}$
for each $\mathbf{z}'_k$ in $\mathbf{Z}'$ **do**
  $\widetilde{\mathbf{z}}_k \leftarrow (1-\lambda)\mathbf{z}'_k + \lambda P_r \mathbf{z}'_k$
end for
return $\widetilde{\mathbf{Z}} = \{\widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_K\}$

Models and Task. Our experiments are conducted on the widely-used mathematical dataset GSM8K. We report greedy decoding accuracy on Qwen3-8B and Llama 3.1-8B, ensuring our results are comparable across different model families. All interventions are applied at inference time to latent vectors cached from a single forward pass.

4.3 Intervening on Semantic Structure

Drawing from our discovery in Section 3.2 that latent vectors are semantically aligned with CoT steps, our first intervention aims to refine the latent trajectory by steering it toward more coherent geometric regions. This intervention consists of two synergistic components, detailed in Algorithm 1.

Mapper-Guided Local Transport. The core idea is to use the linear mapper $F(\cdot)$, trained in Section 3.2, as a guide to find a better direction for the terminal latent vectors, which are critical for the final answer. We first compute the mean $\mathbf{m}_0$ of the terminal latent pair $(\mathbf{z}_5, \mathbf{z}_6)$ and its target destination $\mathbf{t}_0 = F(\mathbf{m}_0)$. We then gently nudge $\mathbf{m}_0$ toward $\mathbf{t}_0$ using spherical interpolation (slerp), which modifies only the direction, followed by a controlled mixing of vector norms.

$$\begin{aligned} \widehat{\mathbf{u}} &= \mathrm{slerp}\Big(\tfrac{\mathbf{m}_0}{\|\mathbf{m}_0\|}, \tfrac{\mathbf{t}_0}{\|\mathbf{t}_0\|}; \alpha\Big), \\ \widehat{\mathbf{m}} &= \Big[(1-\eta)\|\mathbf{m}_0\| + \eta\|\mathbf{t}_0\|\Big]\ \widehat{\mathbf{u}}. \end{aligned} \tag{12}$$

To preserve the original, instance-specific information contained in the pair, we re-apply their residual vectors $(\mathbf{r}_5, \mathbf{r}_6)$ to the new mean $\widehat{\mathbf{m}}$:

$$\mathbf{z}'_5 = \widehat{\mathbf{m}} + \beta\,\mathbf{r}_5, \quad \mathbf{z}'_6 = \widehat{\mathbf{m}} + \beta\,\mathbf{r}_6. \tag{13}$$

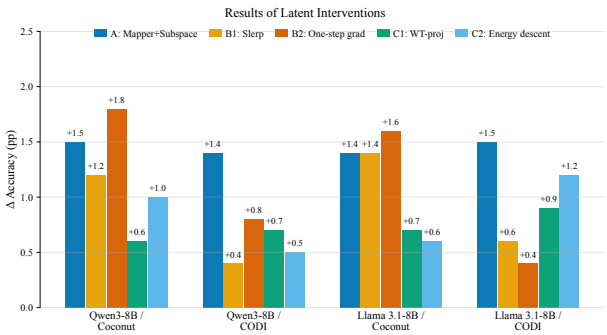

Figure 4: **Accuracy improvement (Δ Acc) over baseline for each intervention family across all models and latent reasoning paradigms on GSM8K**. The consistently positive gains demonstrate that our theory-driven, training-free interventions are broadly effective at unlocking latent reasoning capabilities.

Embedding-Subspace Alignment. Inspired by the observation that latent vectors function as pseudo-tokens, we hypothesize that projecting them closer to the principal geometric subspace of the model's token embeddings can regularize them and improve coherence. Let $P_r = V_r V_r^\top$ be the projector onto the top-r principal components of the centered embedding matrix. We apply this as a final alignment step:

$$\widetilde{\mathbf{z}} = (1-\lambda)\,\mathbf{z} + \lambda\,P_r\,\mathbf{z}, \tag{14}$$

where λ controls the strength of the projection.

Our causal analysis in Section 3.3 identified early latent vectors (e.g., at position 2) as "causal hubs" with disproportionate influence over the final outcome. This suggests that precise, minimal edits to these specific vectors can be a highly efficient intervention strategy.

Answer-Anchored Slerp. The first strategy is a norm-preserving geometric rotation. Given an answer-supporting CoT boundary vector $\mathbf{t}$ (e.g., from a similar retrieved example), we reorient $\mathbf{z}_2$ toward $\mathbf{t}$ using slerp, effectively steering the reasoning trajectory at its most critical juncture:

$$\mathbf{z}'_2 = \|\mathbf{z}_2\|\ \mathrm{slerp}\Big(\tfrac{\mathbf{z}_2}{\|\mathbf{z}_2\|}, \tfrac{\mathbf{t}}{\|\mathbf{t}\|}; \alpha\Big). \tag{15}$$

Answer-Directed Gradient Update. The second strategy uses a single, normalized gradient step. We compute the gradient of the answer's cross-entropy

Intervention Family	Variant	Qwen3-8B		Llama 3.1-8B	
		CoConut	CODI	CoConut	CODI
Baseline Accuracy (No Intervention)		50.4	62.1	44.8	60.7
A. Semantic Structure					
	Mapper transport + subspace mix	**51.9** (+1.5)	**63.5** (+1.4)	**46.2** (+1.4)	**62.2** (+1.5)
B. Causal Hub Editing					
	S1: slerp toward boundary	51.6 (+1.2)	62.5 (+0.4)	46.2 (+1.4)	**61.3** (+0.6)
	S2: one-step answer gradient	**52.2** (+1.8)	**62.9** (+0.8)	**46.4** (+1.6)	61.1 (+0.4)
C. Geometric Priors					
	WT-consistent projection	51.0 (+0.6)	**62.8** (+0.7)	**45.5** (+0.7)	61.6 (+0.9)
	Energy-guided descent	**51.4** (+1.0)	62.6 (+0.5)	45.4 (+0.6)	**61.9** (+1.2)

Table 4: **Accuracy (%) on GSM8K for different training-free intervention families across two models (Qwen3-8B, Llama 3.1-8B) and two latent reasoning methods (CoConut, CODI).** Values in parentheses show absolute improvement (Δ) over the corresponding baseline. Boldface indicates the best result within each intervention family for a given model and paradigm. Section references for each family are: A (§4.3), B (§4.1), C (§4.4).

loss with respect to $\mathbf{z}_2$ and take a small, norm-scaled step in the opposite direction. This directly pushes the latent vector toward a configuration that better supports the target answer:

$$\mathbf{z}_2' = \mathbf{z}_2 - \eta \, \|\mathbf{z}_2\| \, \frac{\nabla_{\mathbf{z}_2}\mathcal{L}}{\|\nabla_{\mathbf{z}_2}\mathcal{L}\| + \varepsilon}. \tag{16}$$

4.4 Intervening with Geometric Priors

Finally, we design interventions based on the architectural and geometric underpinnings identified in Section 3.4. These methods aim to enforce beneficial geometric properties on the latent states.

Weight-Tying Consistent Projection. We established that weight tying ($W_o = E^\top$) geometrically links hidden states to token embeddings. This intervention reinforces this link by nudging a hidden state $\mathbf{h}_\ell$ towards its expected embedding representation $\tilde{\mathbf{e}}_\ell$, which is computed from the pre-latent logits $\mathbf{o}_{\ell-1}$:

$$\tilde{\mathbf{e}}_\ell = E^\top \mathrm{softmax}(\mathbf{o}_{\ell-1}/\tau), \quad \mathbf{h}_\ell' = (1-\alpha)\mathbf{h}_\ell + \alpha\, \tilde{\mathbf{e}}_\ell. \tag{17}$$

Energy-Guided Local Descent. Given that the latent chain follows a monotonically decreasing energy landscape, we can prevent reasoning from "straying" by ensuring each step moves "downhill." Using the pre-trained energy function H, we apply a small, trust-region-constrained gradient descent step to each latent state:

$$\mathbf{h}_\ell' = \mathbf{h}_\ell - \mathrm{Proj}_{\mathbb{B}(\mathbf{0}, \rho\|\mathbf{h}_\ell\|)}\Big(\eta \, \nabla H(\mathbf{h}_\ell)\Big). \tag{18}$$

4.5 Results of Latent Interventions

Table 4 presents the main results of our interventions, and Figure 4 visualizes the performance gains. Across both Qwen3-8B and Llama 3.1-8B, and for both CoConut and CODI-style reasoning, our interventions yield consistent and positive improvements in accuracy on GSM8K.

The Mapper-with-Subspace method which holistically refines the latent trajectory provides a robust gain, improving accuracy. The more targeted "Position-Specific Edits" are also highly effective, with the one-step gradient update on the causally critical second latent vector achieving the highest peak performance. Finally, the "Geometric Priors" interventions also provide modest but reliable gains, confirming that enforcing architecturally-motivated constraints is a valid strategy for latent computation.

5 Conclusion

We present the first systematic interpretability study of the continuous-thought paradigm for latent reasoning and translate its insights into training-free, decode-time interventions. Structural alignment, linear recoverability, and targeted causal edits jointly indicate that latent vectors encode coherent reasoning steps and that early vectors act as causal hubs. Leveraging these findings, we design mapper-guided transport, embedding-subspace alignment, hub-focused rotations/gradients, and geometry-aware projections that yield consistent accuracy gains on GSM8K for Qwen3-8B and Llama-3.1-8B. Our results link explanation interpretability not only clarifies how latent reasoning unfolds but also enables reliable steering without parameter updates, paving the way for latent-reasoning systems beyond mathematical.

Limitations

This work focuses on the continuous-thought latent-span paradigm; transfer to other latent reasoning mechanisms remains uncertain. Experiments cover only two 8B-class models and GSM8K math, leaving generalization to larger/smaller models, multilingual, long-context, commonsense, code, and multi-turn dialogue untested. Our probes (CKA, linear decoding) are correlational rather than mechanistic; causal edits are local and may induce off-manifold effects. Several methods assume tied input–output embeddings and access to final-layer states; when these assumptions break, benefits may diminish. Although training-free, the interventions add inference-time overhead, and we do not assess safety, calibration, or robustness under distribution shift.

References

Frank Ball, Tom Britton, and Peter Neal. 2025. Household epidemic models revisited. *Preprint*, arXiv:2505.17890.

Michelle Blom, Alexander Ek, Peter J. Stuckey, Vanessa Teague, and Damjan Vukcevic. 2025. 3+ seat risk-limiting audits for single transferable vote elections. *Preprint*, arXiv:2503.14803.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. 2024. Human alignment of large language models through online preference optimisation. *Preprint*, arXiv:2403.08635.

Theodore Chambers, Michael Vierhauser, Ankit Agrawal, Michael Murphy, Jason Matthew Brauer, Salil Purandare, Myra B Cohen, and Jane Cleland-Huang. 2024. Hifuzz: Human interaction fuzzing for small unmanned aerial vehicles. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 1–14. ACM.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *Preprint*, arXiv:2505.16782.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2022. Reliability of cka as a similarity measure in deep learning. *Preprint*, arXiv:2210.16156.

Matthew Davis, Aakash Kulkarni, Ziyan Chen, Yunhan Qiao, Christopher Terrazas, and Manish Motwani. 2025. Automatically detecting heterogeneous bugs in high-performance computing scientific software. *Preprint*, arXiv:2501.09872.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Adi Haviv, Shahar Sarfaty, Uri Hacohen, Niva Elkin-Koren, Roi Livni, and Amit H Bermano. 2024. Not every image is worth a thousand words: Quantifying originality in stable diffusion. *Preprint*, arXiv:2408.08184.

Ziwei Hong and Zhongqiu Fang. 2025. Mean value of cubic l-funcitons with fixed genus. *Preprint*, arXiv:2503.17228.

Fujiao Ju, Yuxuan Wang, Shuo Wang, Chengyin Wang, Yinbo Chen, Jianfeng Li, Mingjie Dong, Bin Fang, and Qianyu Zhuang. 2025a. The establishment of static digital humans and the integration with spinal models. *Preprint*, arXiv:2502.07844.

Fujiao Ju, Yuxuan Wang, Shuo Wang, Chengyin Wang, Yinbo Chen, Jianfeng Li, Mingjie Dong, Bin Fang, and Qianyu Zhuang. 2025b. The establishment of static digital humans and the integration with spinal models. *Preprint*, arXiv:2502.07844.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3901–3910.

Andrej Leban. 2025. Distributional autoencoders know the score. *Preprint*, arXiv:2502.11583.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025. Implicit reasoning in large language models: A comprehensive survey. *Preprint*, arXiv:2509.02350.

Dai Hai Nguyen, Hiroshi Mamitsuka, and Atsuyoshi Nakamura. 2025a. Multiple wasserstein gradient descent algorithm for multi-objective distributional optimization. *Preprint*, arXiv:2505.18765.

Dai Hai Nguyen, Hiroshi Mamitsuka, and Atsuyoshi Nakamura. 2025b. Multiple wasserstein gradient descent algorithm for multi-objective distributional optimization. *Preprint*, arXiv:2505.18765.

Sameer Pimparkhede, Mehant Kammakomati, Srikanth Tamilselvam, Prince Kumar, Ashok Pon Kumar, and Pushpak Bhattacharyya. 2024. Doccgen: Document-based controlled code generation. *Preprint*, arXiv:2406.11925.

Avi Deb Raha, Apurba Adhikary, Mrityunjoy Gain, Yu Qiao, and Choong Seon Hong. 2024. Boosting federated domain generalization: Understanding the role of advanced pre-trained architectures. *Preprint*, arXiv:2409.13527.

Suman Sadhu, Saswata Bhattacharyya, and Aloke Paul. 2025. Extracting composition-dependent diffusion coefficients over a very large composition range in nicofecrmn high entropy alloy following strategic design of diffusion couples and physics informed neural network numerical method. *Preprint*, arXiv:2506.12345.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *Preprint*, arXiv:2502.21074.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Ruihua Song. 2025. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *Preprint*, arXiv:2505.16552.

Bowen Tian, Roel C. G. M. Loonen, Roland Valckenborg, and Jan L. M. Hensen. 2025. A fully automated urban pv parameterization framework for improved estimation of energy production profiles. *Preprint*, arXiv:2505.19876.

Zhaoyang Wang, Jinqi Jiang, Tian Qiu, Hui Liu, Xianfeng Tang, and Huaxiu Yao. 2025a. Efficient long cot reasoning in small language models. *Preprint*, arXiv:2505.18440.

Zijie Wang, Manuel Crespo-Ballesteros, Christopher Collins, Yong Yang, Qi Zhang, Xiaobei Zhang, and Michael Sumetsky. 2025b. Dynamic fabrication method of snap microresonators. *Preprint*, arXiv:2504.04786.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jinyang Wu, Feihu Che, Xinxin Zheng, Shuai Zhang, Ruihan Jin, Shuai Nie, Pengpeng Shao, and Jianhua Tao. 2024. Can large language models understand uncommon meanings of common words? *Preprint*, arXiv:2405.05741.

Anil Kumar Yerrapragada, Jeeva Keshav Sattianarayanin, and Radha Krishna Ganti. 2024. Ucinet0: A machine learning based receiver for 5g nr pucch format 0. *Preprint*, arXiv:2404.15243.

Tomoyuki Yokouchi, Aki Kitaori, Daiki Yamaguchi, Naoya Kanazawa, Max Hirschberger, Naoto Nagaosa, and Yoshinori Tokura. 2024. Comment on "reconsidering the nonlinear emergent inductance: time-varying joule heating and its impact on the ac electrical response". *Preprint*, arXiv:2407.15682.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. 2025. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *Preprint*, arXiv:2505.15778.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, and 14 others. 2025. A survey on latent reasoning. *Preprint*, arXiv:2507.06203.