

Why Latent Chain-of-Thought Helps Exploration but Hurts Precise Computation

Jiaxuan Zou^{1*} Yaozhong Xiong^{2*} Yong Liu³

¹Department of Mathematics and Statistics, Xi'an Jiaotong University,
Xi'an 710049, Shaanxi Province, China

²School of Mathematics, Renmin University of China, Beijing 100872, China

³Department of Artificial Intelligence, Gaoling School of Artificial Intelligence,
Renmin University of China, Beijing 100872, China

*These authors contributed equally to this work.

Correspondence: liuyonggsai@ruc.edu.cn

Abstract

Latent Chain-of-Thought (Latent CoT) models promise efficient reasoning via continuous representations, yet exhibit puzzling performance inconsistencies: excelling at exploration (ProsQA: 97.0%) but failing at computation (GSM8K: 34.1%). We show that this trade-off can be explained by decisional certainty. Our contributions are threefold: (1) We theoretically characterize the Exploration-Execution Trade-off, showing that high certainty enables precise execution but inhibits exploration, while low certainty facilitates search but causes error accumulation. (2) We introduce the Symbolic Index—quantifying decisional commitment—as a compact measure of this trade-off and establish its relationship with both execution stability and exploration capability. (3) We show why curriculum learning is needed under trajectory-distribution mismatch: direct latent training can collapse into shortcut states that differ from expert reasoning trajectories. Our framework shifts the design paradigm from binary architectural choices toward using decisional certainty as a diagnostic signal for future adaptive reasoning systems.

1 Introduction

Large Language Models (LLMs) (Guo et al., 2025; OpenAI, 2025; DeepSeek-AI, 2025) have made Chain-of-Thought (CoT) prompting a dominant method for eliciting reasoning (Wei et al., 2022; Sun et al., 2023). By verbalizing intermediate steps as discrete tokens, CoT provides a structured, human-interpretable logic flow, but it also incurs long sequences and autoregressive computation (Zhen et al., 2025; Yue et al., 2025). Recent work therefore explores implicit reasoning, often termed Latent CoT (Ye et al., 2025), where models reason through continuous internal states instead of generating intermediate tokens. Mechanisms such as token-level manipulation (Tack et al., 2025; Sun et al., 2025) and trajectory optimization (Cheng

and Van Durme, 2024; Hao et al., 2024) aim to reduce computational cost and support reasoning paths beyond the vocabulary of natural language (Hao et al., 2024; Zhang et al., 2025a).

Yet Table 1 shows a sharp empirical trade-off. CoT is stronger on symbolic computation (GSM8K: 42.9%) but weaker on exploratory reasoning (ProsQA: 77.5%), often because a token-level chain commits to one executable path. Latent CoT performs better on ProsQA (97.0%) by retaining more alternatives, but this same flexibility weakens exact state maintenance on GSM8K (34.1%). Reported ablations also show training fragility: removing curriculum learning reduces ProntoQA accuracy from 99.8% to 52.4%. This pattern suggests that the issue is not only efficiency, but a deeper coupling between search breadth, execution fidelity, and trainability.

These observations expose a missing mechanism: why exploration reduces execution precision, and why implicit models are difficult to train. Prior work often treats discrete and continuous reasoning as binary architectural choices, leaving the decision-theoretic source of their strengths and weaknesses underspecified. Without such a mechanism, architecture design reduces to a heuristic choice between systems that are precise but rigid and systems that are flexible but unstable.

We argue that this trade-off can be analyzed through *decisional certainty*. We introduce the **Symbolic Index**—a measure of commitment to a reasoning path—as a diagnostic lens for this mechanism. CoT operates in a *high-certainty regime*: token selection filters sub-decisional noise, but early commitment limits exploration. Latent CoT operates in a *low-certainty regime*: distributed latent states support exploration, but accumulated noise degrades symbolic precision because no discrete quantization step resets the state.

We make three contributions. First, we characterize the exploration-execution trade-off: high cer-

Table 1: Performance of discrete (CoT) and latent (Coconut) reasoning models shows an exploration-execution trade-off. CoT is stronger on precise computation (GSM8K), whereas Coconut is stronger on exploratory reasoning (ProsQA) but highly curriculum-sensitive. *Results from (Hao et al., 2024).

Method	GSM8K		ProntoQA		ProsQA	
	Acc. (%)	# Tokens	Acc. (%)	# Tokens	Acc. (%)	# Tokens
CoT	42.9 ± 0.2	25.0	98.8 ± 0.8	92.5	77.5 ± 1.9	49.4
No-CoT	16.5 ± 0.5	2.2	93.8 ± 0.7	3.0	76.7 ± 1.0	8.2
COCONUT	34.1 ± 1.5	8.2	99.8 ± 0.2	9.0	97.0 ± 0.3	14.2
- w/o curriculum	14.4 ± 0.8	8.2	52.4 ± 0.4	9.0	76.1 ± 0.2	14.2

tainty yields stable execution but weak exploration, while low certainty supports search but amplifies sub-decisional noise. Second, we formalize the Symbolic Index as a diagnostic metric for comparing reasoning regimes. Third, we show that curriculum learning addresses a trajectory-distribution mismatch between self-generated latent states and valid reasoning trajectories.

2 Related Work

Unlike CoT, which is constrained to a single discrete path, Latent CoT uses continuous latent spaces for exploratory diversification. Research in this area enables models to explore multiple reasoning trajectories in parallel, improving robustness on complex problems. Techniques include sampling latent trajectories (Chen et al., 2024), using probabilistically weighted concepts (Zhang et al., 2025b), and injecting continuous tokens to enrich the search space (Xu et al., 2025a,b; Gozeten et al., 2025). These works motivate our treatment of exploration as low decisional certainty over next steps. Our analysis turns this empirical pattern into a measurable condition: exploration corresponds to a lower Symbolic Index and bounded divergence from an ideal exploration prior.

A distinct line of research focuses on trajectory-level latent optimization, compressing entire explicit reasoning chains into continuous representations. Early works aimed to maintain semantic fidelity by anchoring latent states to explicit steps (Cheng and Van Durme, 2024; Liu et al., 2024; Shen et al., 2025b). More recent efforts improve efficiency through dynamic compression or adaptive control (Zhang et al., 2025a; Ma et al., 2025; Tan et al., 2025; Wang et al., 2025a). A particularly important strategy is progressive internalization, where explicit steps are gradually replaced by latent thoughts via curriculum learning (Deng et al.,

2024; Hao et al., 2024; Shen et al., 2025a) or internal iterations (Zeng et al., 2025; Ruan et al., 2025). These methods motivate our analysis of curriculum learning as a mechanism for aligning latent states with valid reasoning trajectories. We show that this alignment is not merely a useful training recipe: it reduces the distributional mismatch that otherwise prevents stable latent reasoning.

Complementing these architectural advances, a parallel stream of research investigates the fine-grained control and analysis of internal states. Signal-guided control methods insert specialized, non-text-producing tokens to steer internal reasoning (Herel and Mikolov, 2024; Goyal et al., 2024; Zelikman et al., 2024; Pfau et al., 2024; Wang et al., 2024). Simultaneously, internal state analysis uses techniques like probing or distillation to find mechanistic evidence of implicit reasoning, such as discovering encoded reasoning trees in attention patterns (Deng et al., 2023; Hou et al., 2023; Wang et al., 2025b; Yu et al., 2024). Both approaches highlight the centrality of internal states. We build on this line by making internal certainty explicit through the Symbolic Index, a computable quantity that can compare architectures and monitor transitions across training stages.

Relation to switching between latent and explicit reasoning. SwiReasoning (Shi et al., 2026) reports a closely related qualitative observation: latent thinking preserves multiple possibilities, whereas reliable answer production requires the distribution to reconcentrate. It exploits this by alternating between explicit and latent thinking at inference time in a training-free manner, using block-wise confidence estimated from entropy trends, and bounds the number of switches to curb overthinking. The two papers ask different questions. SwiReasoning asks how a frozen reasoning

LLM can dynamically switch between explicit and latent computation to improve accuracy and token efficiency, and evaluates training-free inference on large reasoning LLMs. We instead ask under what assumptions decisional concentration is associated with exploration, decision margins, and noise sensitivity, and why a curriculum is needed to train Coconut-style models at all; our setting is trained CoT/Coconut models with controlled temperature, support, noise, and curriculum analyses. The observed compatibility of the two accounts is one reason we treat decisional certainty as a diagnostic signal rather than a fixed architectural property.

3 Preliminaries

We consider supervised learning for reasoning tasks, where the input is a random variable X from which instances x are drawn. Each input x and answer y are connected by a chain-of-thought $S = (s_1, \dots, s_M)$. We compare two paradigms for modeling this process: Chain-of-Thought (CoT) and Latent CoT, the latter trained via the Coconut curriculum.

Chain-of-Thought (CoT) CoT models autoregressively generate discrete reasoning steps:

$$p_\theta(S|x) = \prod_{k=1}^M p_\theta(s_k|x, S^{(1\dots k-1)}).$$

At inference, token selection commits to a single path—enabling precise computation but risking early missteps (Mohtashami et al., 2025; Yu et al., 2025; Xu et al., 2024; Emmons et al., 2025).

Latent Chain-of-Thought (Latent CoT) Latent CoT iterates in continuous space, producing latent states $H = (h_1, \dots, h_M)$, $h_k \in \mathbb{R}^d$:

$$h_k = f_\theta(h_{k-1}).$$

Each h_k encodes multiple potential reasoning paths, supporting exploration but accumulating noise over steps (Kong et al., 2025; Zhu et al., 2025; Su et al., 2025; Orlicki, 2025).

Coconut Training (Hao et al., 2024) Since direct supervision on H is infeasible, Coconut uses a curriculum. At stage k , it learns to compress the prefix $S^{(1\dots k)}$ into a latent state h_k to predict the suffix $S_+^{(k)} = S^{(k+1\dots M)}$:

$$\mathcal{L}_{\text{Coconut}}^{(k)}(\theta) = \mathbb{E}_{p(x,S)} \left[-\log p_\theta(S_+^{(k)} | h_k, x) \right].$$

The curriculum progresses from $k = 0$ (pure CoT) to $k = M$ (full latent reasoning), gradually internalizing symbolic steps that bridge the discrete and continuous CoT.

4 Theoretical Analysis

We analyze CoT and Latent CoT along two dimensions: **planning and exploration capability** versus **computational execution precision**. We then identify **decisional certainty**—quantified by the Symbolic Index—as the mechanism regulating this trade-off, and analyze why curriculum learning mitigates distribution mismatch when training Latent CoT models.

4.1 Duality with the Information Bottleneck

The Coconut curriculum has a staged, implicit compression objective. We show that it can be written as a Conditional Information Bottleneck objective (Tishby and Zaslavsky, 2015), allowing us to characterize the model with information-theoretic tools.

At stage k , Coconut compresses the past chain-of-thought $S^{(1\dots k)}$ into a latent vector h_k that predicts the future chain $S^{(k+1\dots M)}$. This matches the Information Bottleneck principle: learn a compact representation of a source that preserves information about a target. The curriculum can therefore be viewed as learning an information compressor whose bottleneck must retain the future-relevant part of past reasoning. Theorem 4.1 formalizes this duality.

Theorem 4.1 (Coconut-CIB Duality). *Under the constraint of any finite model capacity, the optimization objective of the Coconut curriculum (Preliminaries 3) at stage k can be rigorously reformulated as a constrained optimization problem. Its Lagrangian dual is precisely the Conditional Information Bottleneck (CIB) problem. Specifically:*

Primal Problem: *The optimization process of Coconut is equivalent to solving:*

$$\min_{p(h_k|X)} H(S^{(k+1\dots M)} | h_k, X) \quad (1)$$

$$\text{s.t.} \quad I(h_k; S^{(1\dots k)} | X) \leq R \quad (2)$$

$$\text{Dual:} \quad \min_{p(h_k|X)} \left[I(h_k; S^{(1\dots k)} | X) - \beta(k) I(h_k; S^{(k+1\dots M)} | X) \right] \quad (3)$$

where $\beta(k) > 0$ ideally satisfies $\beta(k) \sim \frac{k}{M-k}$.

Interpretation of $\beta(k)$. The trade-off parameter $\beta(k) \sim \frac{k}{M-k}$ reflects the curriculum’s shifting emphasis across stages. As k increases, the model compresses more past information while preserving predictive power over a shorter future sequence. We formalize the training necessity in §5.

The main challenge is to connect Coconut’s operational training loss to a formal information-theoretic objective. The proof achieves this by framing the optimization as a constrained problem and applying Lagrangian duality, revealing that Coconut implicitly solves the Conditional Information Bottleneck. Proof in Appendix A.3.

4.2 Latent CoT excels in exploration

We model reasoning as traversal over a decision DAG $Q := (G, v_{\text{start}}, V_{\text{target}})$. At any node v , let $N_{\text{valid}}(v)$ be the set of valid next steps. An ideal explorer uses a uniform prior $q_{\text{PR}}(u | v) = \frac{1}{|N_{\text{valid}}(v)|} \mathbb{I}[u \in N_{\text{valid}}(v)]$. We measure deviation from this ideal via $D_{\text{KL}}(q_{\text{PR}} \| p)$, where p is the model’s next-step distribution: low KL divergence implies broad exploration; high KL divergence indicates overconfidence and premature commitment. Here $\mathbb{I}[\cdot]$ denotes the indicator function and $H(\cdot)$ denotes Shannon entropy.

We first characterize the high-certainty behavior of a **single, coherent reasoning path**, the basic unit of CoT execution. This behavior follows from teacher-forcing on deterministic reasoning chains, which naturally induces sharply peaked distributions.

Assumption 4.2 (κ -Concentrated Distribution for CoT). At a decision point with B options, we model the generative distribution of a single CoT step, p_{CoT} , as an asymmetric high-concentration Dirichlet distribution. Specifically, one component dominates, with $\alpha_j = \kappa - (B - 1)c$ and $\alpha_{i \neq j} = c$ for a small constant $c > 0$. Thus the sharp peak comes from both large total concentration $\kappa = \sum_i \alpha_i$ and a one-hot-like mean, not from large κ alone.

This is a phenomenological local model rather than a claim that CoT logits are literally generated by a Dirichlet process. It captures the high-certainty regime induced by teacher-forced execution and by post-training methods that concentrate probability mass on preferred continuations, such as RLHF and DPO (Ouyang et al., 2022; Rafailov et al., 2024).

Under this assumption, CoT inevitably collapses

its distribution at each step:

Theorem 4.3 (Exploration Deficiency of CoT). As $\kappa \rightarrow \infty$, the entropy $H(p_{\text{CoT}}) \rightarrow 0$, and

$$D_{\text{KL}}(q_{\text{PR}} \| p_{\text{CoT}}) = \frac{B-1}{B} \log \kappa - \log B - \frac{(B-1) \log c}{B} + \mathcal{O}\left(\frac{1}{\kappa}\right),$$

for a constant $c > 0$, implying that $D_{\text{KL}} \rightarrow \infty$ as certainty $\kappa \rightarrow \infty$.

Proof in Appendix A.4.

Remark 4.4 (Sampling-Based Methods). Our analysis concerns a single reasoning chain. Ensemble methods such as Self-Consistency (Wang et al., 2023) add macro-level exploration by sampling multiple chains, but the empirical ensemble still samples from the same single-chain policy. In the concentrated regime above, a suppressed valid branch can have probability $O(c/\kappa)$, so observing it with nontrivial probability requires an ensemble size $N = \Omega(\kappa)$. Thus test-time ensembles can compensate for poor single-chain exploration, but the compensation becomes increasingly expensive as certainty grows.

This quantifies CoT’s poor performance on exploratory tasks like ProsQA (77.5% vs. Latent CoT’s 97.0% in Table 1): the divergence $D_{\text{KL}} \rightarrow \infty$ formalizes that CoT’s distribution becomes arbitrarily distant from the uniform exploration prior.

In contrast, Latent CoT’s training objective—dual to the Conditional Information Bottleneck (Theorem 4.1)—imposes implicit regularization against overconfidence, supporting sustained exploration:

Theorem 4.5 (Exploration Capability Guarantee of Latent CoT). There exist constants $\delta \in (0, 1)$ and finite c such that

$$D_{\text{KL}}(q_{\text{PR}} \| p_{\text{Coconut}}) \leq -\frac{1}{2} \log \delta - c.$$

Proof in Appendix A.5. Thus Latent CoT maintains a non-degenerate distribution over options, contrasting with CoT’s unbounded divergence (Theorem 4.3) and supporting exploration. This is the formal counterpart of Latent CoT’s empirical advantage on ProsQA.

4.3 Latent CoT’s Fragility in Symbolic Computation

While Latent CoT’s continuous state space supports robust exploration (Section 4.2), it is a liability for

tasks requiring high-fidelity symbolic reasoning, such as GSM8K. Unlike CoT’s discrete symbolic operations, Latent CoT’s continuous state representations are inherently vulnerable to noise accumulation, which undermines precise step-by-step computation.

To formalize this, we analyze the effect of small internal errors, termed **sub-decisional perturbations**—noise that corrupts a model’s internal state but is insufficient to alter the immediate output.

Definition 4.6 (Sub-decisional Perturbation). Let $l_k = l_k^* + \epsilon_k$ be the logit vector at step k . A perturbation ϵ_k is sub-decisional if

$$\operatorname{argmax}(l_k^* + \epsilon_k) = \operatorname{argmax}(l_k^*).$$

This concept highlights a key architectural divergence. CoT models are inherently robust to such perturbations due to their discrete generation process.

Theorem 4.7 (Symbolic Integrity of CoT). *Let $S^* = (s_1^*, \dots, s_M^*)$ be the noise-free CoT trajectory and $\hat{S}$ the trajectory under sub-decisional perturbations. Then*

$$\mathbb{P}[\hat{s}_M \neq s_M^*] = 0.$$

Proof in Appendix A.6.

Theorem 4.7 formalizes CoT’s resilience to sub-decisional perturbations. The underlying mechanism is a **discretization-reset** process: *at each step, the argmax operation acts as an error-correcting filter*, projecting the continuous hidden state to a discrete token and thereby discarding any sub-decisional noise. This reset ensures that subsequent reasoning begins from a clean symbolic representation, preventing error propagation across the reasoning chain.

In stark contrast, Latent CoT lacks this discretization-reset mechanism. Its continuous state h_k is passed directly to the next step without quantization, carrying forward any perturbation. For a model with transition function f_θ (Lipschitz constant L_F) and i.i.d. noise $\epsilon_h^{(k)} \sim \mathcal{N}(0, \sigma_h^2 I_d)$, this error propagation is quantifiable:

Theorem 4.8 (Compounding Error in Latent Computation). *Let $E_M = h_M - h_M^*$. Then for $M \geq 1$,*

$$\mathbb{E}[\|E_M\|^2] = \frac{1 - L_F^{2M}}{1 - L_F^2} \cdot d\sigma_h^2 > 0.$$

Proof in Appendix A.7. The term $\frac{1 - L_F^{2M}}{1 - L_F^2}$ shows that the expected squared error grows with the number of reasoning steps M . If the transition function is expansive ($L_F > 1$), the error compounds exponentially; even for stable functions ($L_F \leq 1$), error accumulates. The key failure mode is trajectory drift: small perturbations may be sub-decisional locally but still corrupt the final state after repeated latent transitions. This explains Latent CoT’s degradation on precision-sensitive tasks like GSM8K, complementing the dichotomy analyzed in Section 4.2.

4.4 Unifying the Trade-off via the Symbolic Index

The analyses in Section 4.2 and Section 4.3 reveal a dichotomy: CoT excels at execution but lacks exploration, while Latent CoT does the opposite. Here, we unify these behaviors through a single, quantifiable principle: the degree of decisional certainty at each reasoning step. To formalize this, we introduce the **Symbolic Index** ($\mathcal{I}_S$), a measure of how committed the model is to its top choice.

Definition 4.9 (Symbolic Index $\mathcal{I}_S$). At a decision point with a vocabulary of valid tokens $\mathcal{V}$, let $p(u \mid h, x)$ be the model’s output distribution. The Symbolic Index is the probability of the most likely token:

$$\mathcal{I}_S = \max_{u \in \mathcal{V}} p(u \mid h, x) \in [1/|\mathcal{V}|, 1].$$

A high $\mathcal{I}_S$ (≈ 1) signifies high certainty, characteristic of CoT’s discrete commitments. A low $\mathcal{I}_S$ signifies distributed consideration over multiple options, characteristic of Latent CoT.

First, we establish the direct link between this certainty and robustness against computational noise. The mechanism for this robustness is the separation between the top two choices, which we term the Logit Decision Margin.

Definition 4.10 (Logit Decision Margin Δ_l). Let l_{i^*} and l_{j^*} be the logits of the most likely and second-most likely tokens, respectively. The margin is their difference: $\Delta_l = l_{i^*} - l_{j^*}$.

The following theorem proves that high certainty guarantees a large decision margin, making the model resilient to perturbations.

Theorem 4.11 (Symbolic Stability Theorem). *The Logit Decision Margin Δ_l is lower-bounded by the Symbolic Index $\mathcal{I}_S$ as follows:*

$$\Delta_l \geq \log \left(\frac{\mathcal{I}_S}{1 - \mathcal{I}_S} \right).$$

Proof in Appendix A.9. Higher decisional certainty yields a larger protective margin. For instance, CoT with $\mathcal{I}_S = 0.99$ has $\Delta_l \geq \log(99) \approx 4.6$, whereas Latent CoT with $\mathcal{I}_S = 0.6$ has $\Delta_l \geq \log(1.5) \approx 0.4$. This links decisional certainty to CoT’s noise immunity and symbolic integrity (Theorem 4.7).

However, this stability comes at a direct and quantifiable cost to exploration. High certainty forces the model’s distribution away from the uniform ideal required for unbiased exploration. The next theorem formalizes this trade-off.

Theorem 4.12 (Exploration-Execution Trade-off Theorem). *For a decision with B valid options, the KL divergence from the ideal uniform exploration prior q_{PR} is lower-bounded by:*

$$D_{KL}(q_{PR}||p) \geq -\log B - \frac{1}{B} \log(\mathcal{I}_S) - \frac{B-1}{B} \log\left(\frac{1-\mathcal{I}_S}{B-1}\right).$$

This bound is minimized when $\mathcal{I}_S = 1/B$ and grows without bound as $\mathcal{I}_S \rightarrow 1$.

Proof in Appendix A.10. This inequality shows that increasing $\mathcal{I}_S$ improves execution stability but increases divergence from a uniform exploration policy, explaining CoT’s exploration deficiency (Theorem 4.3). It makes exploration and execution coupled objectives rather than independent design goals.

Together, these results establish $\mathcal{I}_S$ as the core regulator, unifying both paradigms:

1. **CoT:** High- $\mathcal{I}_S$ regime ensures large decision margins for robust execution (Theorems 4.11, 4.7) but degrades exploration (Theorem 4.12).
2. **Latent CoT:** Low- $\mathcal{I}_S$ regime enables broad exploration (Theorem 4.5) but implies small decision margins, causing vulnerability to sub-decisional noise that compounds over steps (Theorem 4.8), leading to failure on precision-sensitive tasks.

This framework shifts focus from binary architectural choices to **diagnosing decisional certainty**. For explicit CoT, $\mathcal{I}_S$ is a native property of the token distribution and can be affected by decoding controls such as temperature. For Latent CoT, however, $\mathcal{I}_S$ is obtained by probing hidden states and should be treated as a diagnostic surrogate rather than a directly controllable decoding knob.

5 Why Curriculum Learning Works

Having established that $\mathcal{I}_S$ governs the exploration-execution trade-off in *trained* models (§4.4), we next ask how to train a model to operate in the low- $\mathcal{I}_S$ regime required for robust exploration. Unlike CoT, which benefits from token-level supervision at each reasoning step, Latent CoT must learn meaningful latent representations h_k without explicit ground truth.

5.1 The Training Paradox of Latent CoT

An untrained model has no incentive to perform multi-step reasoning encoded in h_k ; instead, it tends to learn "shortcut" representations that merely map surface-level input patterns to outputs. Training requires reliable gradient signals, which necessitate high-certainty (high- $\mathcal{I}_S$) predictions to minimize loss variance. However, the ultimate goal is to learn low- $\mathcal{I}_S$ representations that enable exploration. Direct training thus traps the model in a high- $\mathcal{I}_S$ regime, preventing the emergence of true latent reasoning.

We model this difficulty within an Imitation Learning (IL) framework. Let $P_{\theta^*}(h|x)$ denote the unobserved expert policy that generates optimal, reasoning-encoded latent states. We define a binary valuation function $V(h) \in \{0, 1\}$, where $V(h) = 1$ implies the latent state leads to the correct answer. The primary objective is to maximize the **Task Success Rate** $\mathcal{R}(\theta)$, defined as the expected value of this function under the induced policy:

$$\mathcal{R}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, h \sim P_{\theta}(\cdot|x)}[V(h)].$$

Without a curriculum, the model must learn from its own generated latent states. These self-generated states can come from a biased distribution P_{biased} of superficial representations, which differs from the true expert distribution P_{θ^*} . Since the valuation function $V(h)$ is non-uniform (favoring reasoning states), learning from such a mismatched distribution can lock the model into a shortcut policy.

Theorem 5.1 (Failure under Biased Latent-State Training). *Let D_{nc} be a dataset of latent states drawn from a biased distribution P_{biased} . If this distribution is intrinsically suboptimal compared to the expert (i.e., $\mathcal{R}(P_{\text{biased}}) \leq \mathcal{R}(\theta^*) - \Delta$ for some gap $\Delta > 0$), then the model $\hat{\theta}_{MLE}$ trained on D_{nc} is strictly bounded away from optimality:*

$$\mathcal{R}(\hat{\theta}_{MLE}) \leq \mathcal{R}(\theta^*) - C(\Delta),$$

where $C(\Delta) > 0$ is a constant determined by the bias. This implies that the model’s success rate is permanently capped below the expert’s performance, regardless of dataset size.

Proof in Appendix A.11. Training on "shortcut" latent states therefore locks the model into a sub-optimal policy under the theorem’s mismatch condition. Unlike variance error, which vanishes with data (Theorem 5.2), this bias error is irreducible once the biased support excludes expert reasoning states. This assumption is consistent with the observed no-curriculum collapse in Table 1 and with exposure-bias analyses of sequential prediction (Ross et al., 2011; Bengio et al., 2015).

5.2 Curriculum Learning as Controlled $\mathcal{I}_S$ Transition

From the perspective of the Symbolic Index, curriculum learning creates a controlled transition. In early stages (pure CoT), high $\mathcal{I}_S$ ensures accurate token-level gradients through error-correcting discretization. In hybrid stages, the prefix operates in latent space (low $\mathcal{I}_S$), but the suffix retains explicit token generation, anchoring latent representations to valid reasoning traces. Finally, progressive internalization stabilizes the low $\mathcal{I}_S$ regime required for exploration.

In our IL framework, this procedure provides supervised samples of latent thoughts grounded in correct reasoning, equivalent to drawing from $P_{\theta^*}(h|x)$. Since $V(h)$ is bounded in $[0, 1]$, the success-rate gap $|\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^*)|$ is upper-bounded by the Total Variation distance between the policies. Minimizing distributional divergence via curriculum learning therefore optimizes task success.

Theorem 5.2 (Provable Success of Training with Curriculum). *Under standard statistical learning conditions, MLE on a dataset D_c of size n generated via the curriculum yields a model $\hat{\theta}$ whose success rate approaches that of the expert:*

$$\mathcal{R}(\hat{\theta}) \geq \mathcal{R}(\theta^*) - \mathcal{O}\left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}}\right),$$

with high probability $1 - \delta$. The performance gap vanishes as the dataset size $n \rightarrow \infty$.

Proof in Appendix A.12. By using CoT as a scaffold for valid training data, the curriculum allows the student model to converge to the expert policy under the stated assumptions. Curriculum learning therefore addresses the distributional mismatch

identified in Theorem 5.1 and yields the convergence guarantee in Theorem 5.2.

6 Experiments

We evaluate three predictions from Section 4.4: (1) Latent CoT and CoT operate in distinct regimes of decisional certainty (Theorems 4.3 & 4.5); (2) this difference determines robustness to computational noise (Theorems 4.11 & 4.12); and (3) Latent CoT exhibits error accumulation consistent with continuous state dynamics (Theorem 4.8).

Experimental Setup We use the GPT-2 (124M) architecture. We compare a CoT model against a Latent CoT model trained via the Coconut curriculum ($M = 6$ latent steps). Evaluations are performed on **GSM8K** (Cobbe et al., 2021) for precise symbolic computation, **ProntoQA** (Saparov and He, 2023) for controlled deductive reasoning, and **ProsQA** (Hao et al., 2024) for exploratory reasoning. Implementation details are in Appendix A.1. Appendix A.2 additionally reports a cross-model check on Qwen2.5-1.5B-Instruct, which tests transfer of the predictive-distribution diagnostic rather than of the latent architecture itself.

6.1 Analysis of Decisional Certainty ($\mathcal{I}_S$)

To test the Exploration-Execution Trade-off (Theorem 4.12), we measure the **Symbolic Index** ($\mathcal{I}_S$), defined as the maximum token probability at a given step.

Figures 1 and 2 show representative $\mathcal{I}_S$ trajectories; Figure 5 in Appendix A.1 summarizes the aggregate GSM8K test-set pattern. The figures illustrate two regimes:

- **Exploration via Low $\mathcal{I}_S$ (ProsQA):** On ProsQA (Figure 2), the Latent CoT model maintains $\mathcal{I}_S < 0.6$. This is consistent with Theorem 4.5: the model retains multiple potential reasoning trajectories in latent space, preventing premature convergence to a suboptimal solution and supporting high accuracy (97.0%).
- **Absence of Discretization (GSM8K):** On GSM8K (Figure 1), the model operates without high-probability concentration ($\mathcal{I}_S \ll 1.0$). According to Theorem 4.11, low $\mathcal{I}_S$ corresponds to a small Logit Decision Margin Δ_l . Unlike CoT, which uses the argmax operation to reset state variance at each token, Latent

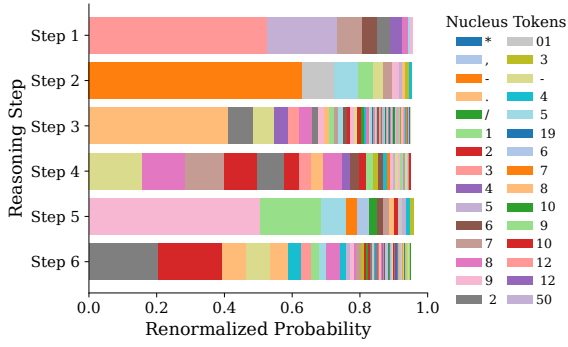


Figure 1: **Symbolic Index on GSM8K.** Latent CoT maintains a low Symbolic Index ($\mathcal{I}_S \in [0.2, 0.5]$), indicating a dispersed probability distribution and small decision margins.

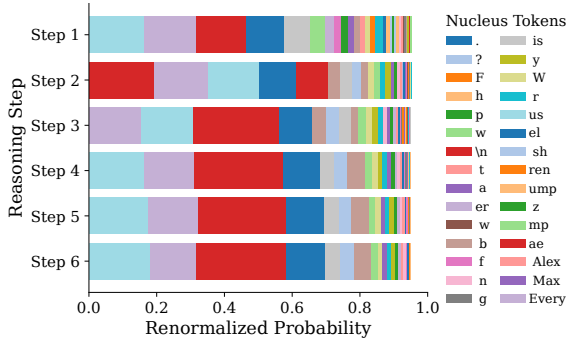


Figure 2: **Symbolic Index on ProsQA.** Latent CoT exhibits a stable, low $\mathcal{I}_S$ distribution across reasoning steps, distributing probability mass across multiple latent paths.

CoT propagates variance continuously, leading to lower computational accuracy (34.1%).

6.2 Noise Robustness and Stability Analysis

To test error propagation (Theorem 4.8) and stability bounds (Theorem 4.11), we inject Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$ into the hidden states (Latent CoT) or pre-output states (CoT) and measure Exact Match accuracy degradation. This perturbation models small continuous internal state errors during iterative reasoning, not input-level corruptions such as typos.

Figure 3 shows two failure modes:

- **CoT (Step-Function Response):** CoT accuracy remains flat at low noise levels before sharply declining, consistent with a non-zero Logit Decision Margin Δ_l . For noise vectors $\|\epsilon\| < \Delta_l$, token selection filters the perturbation (Theorem 4.7).
- **Latent CoT (Continuous Decay):** Latent CoT degrades immediately and monotonically,

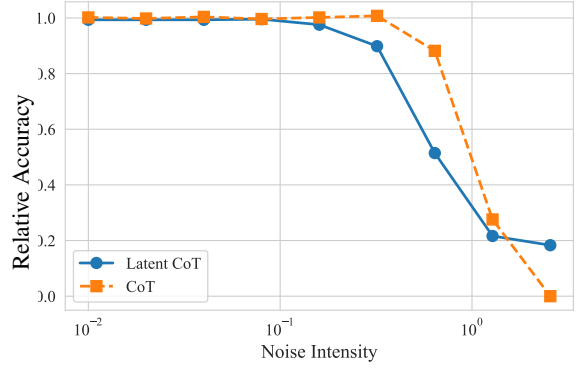


Figure 3: **Performance Degradation under Noise.** The plotted value is the relative accuracy $\text{Acc}(\sigma)/\text{Acc}(0)$, not an accuracy-drop rate. CoT (orange) shows a threshold effect, while Latent CoT (blue) decays monotonically from $\sigma \approx 0$, matching the theoretical template $A(\sigma) = \Phi(\Delta_l/(\sqrt{C}\sigma))$ in Appendix A.8.

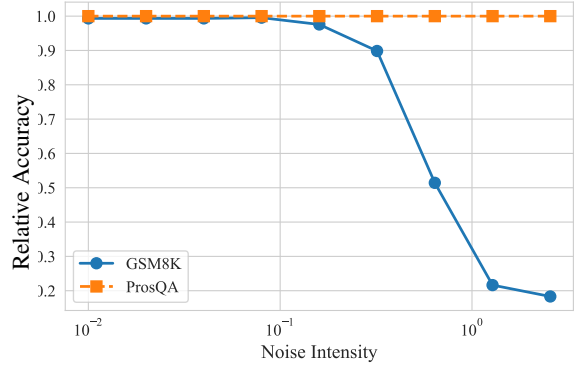


Figure 4: **Task-Dependent Sensitivity.** Relative accuracy $\text{Acc}(\sigma)/\text{Acc}(0)$ under identical noise levels: Latent CoT retains higher performance on ProsQA (orange) than GSM8K (blue), indicating greater tolerance for state deviation in exploratory tasks.

consistent with $\Delta_l \approx 0$ in continuous latent space. Without a discrete margin, perturbations alter the state trajectory, and the transition function f_θ accumulates error over $M = 6$ reasoning steps (Theorem 4.8).

Figure 4 shows task-dependent robustness: ProsQA remains stable under noise, whereas GSM8K degrades rapidly, suggesting that exploration tolerates small deviations in h_M while computation requires a narrow region mapping to the correct token. These curves do not localize the exact GSM8K failure point, but they provide macro-level evidence for compounding error, consistent with GSM8K’s numerical-precision sensitivity (we disable bfloat16 for stability) and the latent-depth diagnostic in Appendix A.1.

6.3 Ablation: Curriculum Learning

Reported ablations show that Latent CoT without curriculum reaches only 52.4% accuracy on Pron-toQA (Table 1). Monitoring training indicates that, without the curriculum, the model converges to a high- $\mathcal{I}_S$ regime in early epochs. This creates a distributional mismatch: the model learns a deterministic mapping based on input-output statistics rather than the expert latent policy P_{θ^*} (Theorem 5.1). The curriculum provides a scaffold for the distribution alignment needed for low- $\mathcal{I}_S$ reasoning. A latent-depth diagnostic gives the same qualitative message: adding latent steps initially improves expressivity, but pushing the latent rollout beyond the stable range degrades performance, consistent with compounding continuous-state error.

7 Conclusion

We introduced the Symbolic Index as a diagnostic lens on the exploration-execution trade-off: CoT’s high certainty preserves fidelity through discretization but restricts exploration, whereas Latent CoT’s low certainty broadens search while weakening symbolic integrity under compounding errors. Curriculum mitigates the resulting distribution mismatch. Future work should test mechanisms that monitor and modulate decisional certainty for execution or planning.

Limitations

Our analysis focuses on CoT and Latent CoT under idealized assumptions about reasoning trajectories, latent-state dynamics, and decision distributions. These assumptions make the trade-off analytically tractable but may not capture all behaviors of larger instruction-tuned models, multimodal models, or models trained with reinforcement learning. The latent-architecture evidence is limited to the reported GPT-2-scale Coconut setting and three reasoning benchmarks; the cross-model check in Appendix A.2 extends only the predictive-distribution diagnostic to Qwen2.5-1.5B-Instruct and does not reproduce the hidden-state noise experiments at that scale. The Symbolic Index should therefore be interpreted as a diagnostic principle rather than a complete predictor of performance across domains.

The Symbolic Index also compresses a rich decision distribution into a single maximum-probability statistic. Two models can have the same $\mathcal{I}_S$ while distributing the remaining probability mass differently, and the metric does not directly encode cal-

ibration, semantic equivalence among options, or geometry of latent trajectories. For Latent CoT, $\mathcal{I}_S$ is measured by probing hidden states, so it is not yet a direct intervention knob in the way decoding temperature is for explicit token generation. Finally, our work characterizes one distribution-mismatch condition under which curriculum learning is needed for stable latent reasoning, but it does not prove that every no-curriculum training distribution must fail or prescribe an optimal curriculum schedule for arbitrary model families or tasks.

Ethical Considerations

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. [Scheduled sampling for sequence prediction with recurrent neural networks](#). *Preprint*, arXiv:1506.03099.
- Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, Phil Mui, Silvio Savarese, Caiming Xiong, and 1 others. 2024. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. *arXiv preprint arXiv:2411.04282*.
- Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- DeepSeek-AI. 2025. Deepseek-v3.1 model card. Hugging Face model card. Hybrid thinking/non-thinking inference, improved tool use and agent tasks, long-context extension (128 K), FP8 format.
- Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.

- Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*.
- Scott Emmons, Erik Jenner, David K Elson, Rif A Saurous, Senthoran Rajamanoharan, Heng Chen, Irhum Shafkat, and Rohin Shah. 2025. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*.
- Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*.
- Halil Alperen Gozeten, M Emrullah Ildiz, Xuechen Zhang, Hrayr Harutyunyan, Ankit Singh Rawat, and Samet Oymak. 2025. Continuous chain of thought enables parallel exploration and reasoning. *arXiv preprint arXiv:2505.23648*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv 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.
- David Herel and Tomas Mikolov. 2024. Thinking tokens for language modeling. *arXiv preprint arXiv:2405.08644*.
- Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. *arXiv preprint arXiv:2310.14491*.
- Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, and 1 others. 2025. Latent thought models with variational bayes inference-time computation. *arXiv preprint arXiv:2502.01567*.
- Tianqiao Liu, Zui Chen, Zitao Liu, Mi Tian, and Weiqi Luo. 2024. Expediting and elevating large language model reasoning via hidden chain-of-thought decoding. *arXiv preprint arXiv:2409.08561*.
- Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*.
- Amirkeivan Mohtashami, Matteo Pagliardini, and Martin Jaggi. 2025. CoTFormer: A chain of thought driven architecture with budget-adaptive computation cost at inference. In *The Thirteenth International Conference on Learning Representations*. ArXiv:2310.10845.
- OpenAI. 2025. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>. Accessed: 2025-08-29.
- José I Orlicki. 2025. Beyond words: A latent memory approach to internal reasoning in llms. *arXiv preprint arXiv:2502.21030*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.
- Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let’s think dot by dot: Hidden computation in transformer language models. In *First Conference on Language Modeling*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.
- Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. *Preprint*, arXiv:1011.0686.
- Yangjun Ruan, Neil Band, Chris J Maddison, and Tatsunori Hashimoto. 2025. Reasoning to learn from latent thoughts. *arXiv preprint arXiv:2503.18866*.
- Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *Preprint*, arXiv:2210.01240.
- Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. 2025a. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025b. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*.
- Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. 2026. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning llms. *Preprint*, arXiv:2510.05069.
- DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. 2025. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*.

- Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, and 1 others. 2023. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*.
- Yuchang Sun, Yanxi Chen, Yaliang Li, and Bolin Ding. 2025. Enhancing latent computation in transformers with latent tokens. *arXiv preprint arXiv:2505.12629*.
- Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. 2025. Llm pre-training with continuous concepts. *arXiv preprint arXiv:2502.08524*.
- 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. *arXiv preprint arXiv:2505.16552*.
- Naftali Tishby and Noga Zaslavsky. 2015. [Deep learning and the information bottleneck principle](#). *Preprint*, arXiv:1503.02406.
- Jianwei Wang, Ziming Wu, Fuming Lai, Shaobing Lian, and Ziqian Zeng. 2025a. Synadapt: Learning adaptive reasoning in large language models via synthetic continuous chain-of-thought. *arXiv preprint arXiv:2508.00574*.
- Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. 2025b. System-1.5 reasoning: Traversal in language and latent spaces with dynamic shortcuts. *arXiv preprint arXiv:2505.18962*.
- Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordani. 2024. Guiding language model reasoning with planning tokens. In *First Conference on Language Modeling*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.
- Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025a. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*.
- Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025b. Softcot++: Test-time scaling with soft chain-of-thought reasoning. *arXiv preprint arXiv:2505.11484*.
- Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Weichuan Liu, Xiaoyin Che, Lei Hou, and 1 others. 2025. How do transformers learn implicit reasoning? *arXiv preprint arXiv:2505.23653*.
- Ping Yu, Jing Xu, Jason E Weston, and Ilia Kulikov. 2024. Distilling system 2 into system 1. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS’24*.
- Qifan Yu, Zhenyu He, Sijie Li, Xun Zhou, Jun Zhang, Jingjing Xu, and Di He. 2025. Enhancing autoregressive chain-of-thought through loop-aligned reasoning. *arXiv preprint arXiv:2502.08482*.
- Linan Yue, Yichao Du, Yizhi Wang, Weibo Gao, Fangzhou Yao, Li Wang, Ye Liu, Ziyu Xu, Qi Liu, Shimin Di, and 1 others. 2025. Don’t overthink it: A survey of efficient r1-style large reasoning models. *arXiv preprint arXiv:2508.02120*.
- Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. 2024. Quiet-STAR: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*.
- Boyi Zeng, Shixiang Song, Siyuan Huang, Yixuan Wang, He Li, Ziwei He, Xinbing Wang, Zhiyu Li, and Zhouhan Lin. 2025. Ponderlm: Pretraining language models to ponder in continuous space. *arXiv preprint arXiv:2505.20674*.
- Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025a. Lightthinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*.
- Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. 2025b. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*.
- Taihang Zhen, Jialiang Hong, Kai Chen, Guang Yang, Junlan Feng, Wenpeng Zhu, Jing Huo, Yang Gao, and 1 others. 2025. Reconsidering overthinking: Penalizing internal and external redundancy in cot reasoning. *arXiv preprint arXiv:2508.02178*.
- Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. 2025. Reasoning by superposition: A theoretical perspective on chain of continuous thought. *arXiv preprint arXiv:2505.12514*.

A Appendix

A.1 Experimental Details

Models. Both CoT and Latent CoT (Coconut) models are based on the smallest GPT-2 variant with 124M parameters. For Latent CoT, we use $M = 6$ latent reasoning steps across all experiments. The latent thought vector h_k has the same dimension as GPT-2’s hidden size ($d = 768$).

Training. We use the Adam optimizer with learning rate $1e-4$ and weight decay 0.01. Training is performed on 4 GPUs with mixed-precision (bfloat16 enabled for ProsQA, disabled for GSM8K due to stability). For GSM8K, we train for 25 epochs with batch size 32 and gradient accumulation steps 1; for ProsQA, we train for 30 epochs with batch size 16 and gradient accumulation steps 2. The Coconut curriculum progresses stage-by-stage: for GSM8K, we use 3 epochs per stage up to latent stage 3 (i.e., compressing the first 3 CoT steps into latent thoughts); for ProsQA, we use 4 epochs per stage up to latent stage 6 (full internalization). The base model checkpoint is initialized from standard GPT-2.

Datasets. We evaluate on three reasoning benchmarks:

1. **GSM8K:** GSM8K (Cobbe et al., 2021) is a benchmark of grade-school math word problems with step-by-step solutions. For the probing and noise-injection analyses we train on a subset of the official training set (7,473 examples) with a 10% held-out validation subset, and we evaluate on a randomly sampled subset of 330 test examples rather than the full official test set of 1,319 examples.
2. **ProntoQA:** ProntoQA (Saparov and He, 2023) is a synthetic deductive QA benchmark generated from first-order-logic ontologies and proof structures, then translated into natural-language contexts, queries, and True/False labels. It controls reasoning depth through hop count and varies ontology types. We use it as a controlled logical-reasoning benchmark for the curriculum ablation in Table 1.
3. **ProsQA:** ProsQA is the graph-search reasoning benchmark introduced by Hao et al. (2024). We use a reduced version of it: our training, validation, and test splits contain 2,000, 200, and 300 examples, respectively,

rather than the 17,886/300/500 splits of the original release.

All accuracies we report from our own runs are computed on the test splits described above. The accuracies in Table 1 are quoted from Hao et al. (2024), which uses the full official splits, and are therefore not directly comparable to the reduced-split models used for our Symbolic Index and noise-injection analyses.

Symbolic Index Visualization Details. The visualizations in Figures 1 and 2 are based on representative examples from the respective test sets.

- **Figure 1 (GSM8K):** The plot was generated from the following math reasoning problem: "Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market?"
- **Figure 2 (ProsQA):** The plot was generated from a synthetic ontology instance with multi-hop rules over entities such as Fae and Max and concepts such as shumpus, worpus, tumpus, and boompus; the query asks whether Fae belongs to one of two candidate concepts. We use the example only to identify the representative trajectory, while Figure 5 reports the aggregate GSM8K pattern.

Aggregate Symbolic Index Summary. To check that Figures 1 and 2 are not artifacts of single examples, we also aggregate $\mathcal{I}_S$ over the recorded test-set reasoning steps (Figure 5). The coarse binned pattern is stable: Latent CoT places most mass in the low-to-mid certainty regime ($\mathcal{I}_S \in [0.2, 0.6]$, peaking around $[0.2, 0.4]$), whereas explicit CoT is concentrated near $\mathcal{I}_S \in [0.9, 1.0]$, with negligible mass below 0.8. This supports the regime distinction used in Section 6.1, although we treat the aggregate as diagnostic rather than as a new controlled intervention.

Noise Injection Protocol. During inference, we inject Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$ as follows:

1. For **CoT**: noise is added to the last hidden state before the LM head, i.e., $h \leftarrow h + \sigma \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_d)$.

- For **Latent CoT**: noise is added at every reasoning step, i.e., $h_k \leftarrow f_\theta(h_{k-1}) + \sigma \cdot \epsilon_k$, with independent $\epsilon_k \sim \mathcal{N}(0, I_d)$.

These perturbations model continuous internal representation errors, including finite-precision and state-update noise, rather than corruptions of the input text. We report accuracy as the exact-match rate on the test set. The quantity plotted in Figures 3 and 4 is the relative accuracy $\text{Acc}(\sigma)/\text{Acc}(0)$; a true drop rate would instead be $1 - \text{Acc}(\sigma)/\text{Acc}(0)$.

Latent-Depth Diagnostic. We also varied the number of latent reasoning steps around the default $M = 6$ configuration. The resulting trend is non-monotonic: adding latent steps initially improves exploratory capacity, but pushing the rollout beyond the stable range reduces accuracy on precision-sensitive tasks. This qualitative pattern is consistent with Theorem 4.8: longer continuous rollouts increase the opportunity for small state errors to compound.

Reproducibility. The main GPT-2 experiments use seed 0. Due to the deterministic nature of our evaluation protocol and the use of nucleus sampling with fixed $p = 0.95$ for Symbolic Index estimation, results are fully reproducible given the same model checkpoint. The cross-model transfer study in Appendix A.2 uses additional seeds, as specified there.

A.2 Cross-Model Transfer of the Predictive-Distribution Diagnostic

The main experiments use GPT-2 (124M) under the Coconut setup. To test whether the $\mathcal{I}_S$ -based patterns are specific to that scale, we repeated the predictive-distribution measurements on **Qwen2.5-1.5B-Instruct**, an instruction-tuned model with roughly $12\times$ the parameters of GPT-2, using the same ProsQA evaluation protocol. Temperature and support experiments use fresh, independent generations with seeds 0/1/2 and 1,500 outputs per condition; the failure-prediction and selection experiments use held-out seeds 3–9 with 14,000 outputs per model.

Temperature. Raising the decoding temperature lowers both the mean Symbolic Index and accuracy on both models (Table 2).

Sampling support. Fixing $T = 1.7$ and tightening the sampling support raises both $\mathcal{I}_S$ and accuracy while reducing entropy on both models (Table 3).

Table 2: Cross-model temperature results on ProsQA, reported as accuracy / mean $\mathcal{I}_S$.

Model	$T = 0.3$	$T = 1.7$
GPT-2	0.9707 / 0.999741	0.9400 / 0.993828
Qwen2.5-1.5B	0.9732 / 0.999812	0.9425 / 0.994120

Table 3: Effect of restricting the sampling support at $T = 1.7$. Supp.: sampling support (Full or Top- k with $k = 5$); Ent.: entropy.

Model	Supp.	Acc.	Mean $\mathcal{I}_S$	Ent.
GPT-2	Full	0.9373	0.993879	0.042849
	Top-5	0.9553	0.996906	0.011882
Qwen2.5-1.5B	Full	0.9392	0.993965	0.042550
	Top-5	0.9568	0.997030	0.011655

Held-out failure prediction and selection. A low mean $\mathcal{I}_S$ remains a strong error-ranking signal on Qwen, and $\mathcal{I}_S$ /margin-based selection improves over the best fixed-support candidate (Table 4).

Table 4: Held-out diagnostic and selection results on ProsQA. The four selection-rule entries report accuracy under different candidate-selection rules.

Metric	GPT-2	Qwen2.5-1.5B
<i>Error ranking</i>		
AUROC	0.9201	0.9238
AUPRC	0.5272	0.5285
<i>Coverage</i>		
Overall	0.9470	0.9495
@80%	0.9900	0.9908
<i>Selection rule</i>		
Fixed $k = 5$	0.9574	0.9454
max $\mathcal{I}_S$	0.9591	0.9514
max margin	0.9594	0.9605
min entropy	0.9591	0.9601

Scope of this evidence. Two limitations should be read together with these numbers. First, the high-temperature decrease in $\mathcal{I}_S$ and accuracy transfers to Qwen, and tightening the support increases both $\mathcal{I}_S$ and accuracy at a fixed temperature; however, these are interventions on the predictive distribution, not isolated $do(\mathcal{I}_S)$ interventions, so we describe them as intervention-sensitive diagnostic evidence rather than causal identification. Second, Qwen tests transfer of the *predictive-distribution* diagnostic only: it is not a second latent-reasoning architecture, and we have not reproduced the hidden-state noise experiment on Qwen or on RL-trained 7B+ models. We therefore do not claim universal scale or architecture invariance. Additional latent-reasoning methods, larger RL-trained models, and more realistic mixed tasks remain necessary future tests.

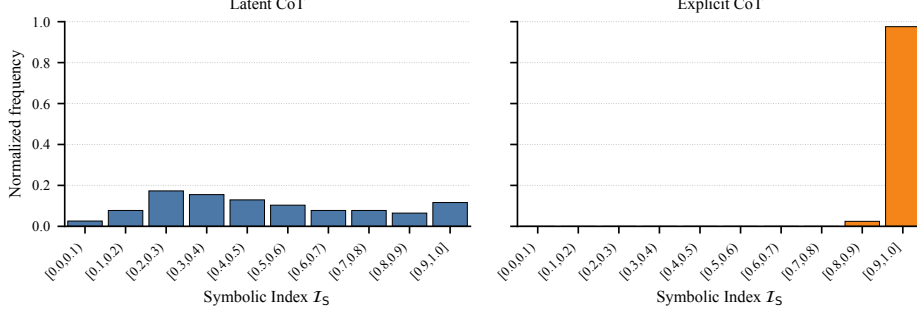


Figure 5: **Aggregate Symbolic Index distribution on GSM8K.** Bars show normalized binned frequencies over recorded test-set reasoning steps. Latent CoT spreads probability mass across low-to-mid $\mathcal{I}_S$ bins, whereas explicit CoT is concentrated near $\mathcal{I}_S = 1.0$.

A.3 Proof of Theorem 4.1

Lemma A.1 (Loss and Mutual Information). *Assume the decoder model family $\{p_\theta(S^{(k+1...M)} | h_k, X)\}$ is well-specified, i.e., there exists a parameter θ^* such that $p_{\theta^*}(S^{(k+1...M)} | h_k, X) = p(S^{(k+1...M)} | h_k, X)$ holds almost everywhere. Further assume $H(S^{(k+1...M)} | X) < \infty$. Then,*

$$\min_{\theta} \mathcal{L}_{\text{Coconut}}^{(k)}(\theta) \iff \max_{p(h_k|X)} I(h_k; S^{(k+1...M)} | X).$$

Proof. Let $S_+ = S^{(k+1...M)}$. Expand the loss function:

$$\begin{aligned} \mathcal{L}_{\text{Coconut}}^{(k)}(\theta) &= \mathbb{E}_{p(X, S, h_k)} [-\log p_\theta(S_+ | h_k, X)] \\ &= \mathbb{E}_{p(X, h_k)} [\mathbb{E}_{p(S_+ | X, h_k)} [-\log p_\theta(S_+ | h_k, X)]] \\ &= \mathbb{E}_{p(X, h_k)} \left[\mathbb{E}_{p(S_+ | X, h_k)} \left[-\log p(S_+ | h_k, X) + \log \frac{p(S_+ | h_k, X)}{p_\theta(S_+ | h_k, X)} \right] \right] \\ &= \mathbb{E}_{p(X, h_k)} [H(S_+ | h_k, X) + D_{\text{KL}}(p(\cdot | h_k, X) \| p_\theta(\cdot | h_k, X))]. \end{aligned}$$

By the well-specified assumption, there exists w^* such that $D_{\text{KL}}(p(\cdot | h_k, X) \| p_\theta(\cdot | h_k, X)) \rightarrow 0$. Therefore,

$$\min_{\theta} \mathcal{L}_{\text{Coconut}}^{(k)}(\theta) = \min_{p(h_k|X)} H(S^{(k+1...M)} | h_k, X).$$

By the definition of conditional mutual information, $I(A; B | C) = H(A | C) - H(A | B, C)$, we have

$$H(S^{(k+1...M)} | h_k, X) = H(S^{(k+1...M)} | X) - I(h_k; S^{(k+1...M)} | X).$$

Since $H(S^{(k+1...M)} | X)$ is constant, it follows that

$$\min_{p(h_k|X)} H(S^{(k+1...M)} | h_k, X) \iff \max_{p(h_k|X)} I(h_k; S^{(k+1...M)} | X).$$

□

We now begin the proof of Theorem 4.1.

Proof. By Lemma A.1, minimizing the Coconut loss is equivalent to $\min_{p(h_k|X)} H(S^{(k+1...M)} | h_k, X)$. Any physically realizable encoder is constrained by finite model capacity, formally expressed as $I(h_k; S^{(1...k)} | X) \leq R$. We construct the Lagrangian:

$$\mathcal{J}(p, \lambda) = H(S^{(k+1...M)} | h_k, X) + \lambda (I(h_k; S^{(1...k)} | X) - R), \quad \lambda \geq 0.$$

Using the identity $H(A | B, C) = H(A | C) - I(B; A | C)$, we rewrite:

$$\mathcal{J} = \left[H(S^{(k+1...M)} | X) - I(h_k; S^{(k+1...M)} | X) \right] + \lambda I(h_k; S^{(1...k)} | X) - \lambda R.$$

Ignoring constants $H(S^{(k+1\dots M)} | X)$ and λR during minimization over $p(h_k | X)$, we obtain:

$$\min_{p(h_k|X)} \left\{ \lambda I(h_k; S^{(1\dots k)} | X) - I(h_k; S^{(k+1\dots M)} | X) \right\}.$$

Defining $\beta = 1/\lambda > 0$ and dividing through by λ , this becomes the standard CIB objective:

$$\min_{p(h_k|X)} \left\{ I(h_k; S^{(1\dots k)} | X) - \beta I(h_k; S^{(k+1\dots M)} | X) \right\}.$$

Denote $I_{\text{past}} = I(h_k; S^{(1\dots k)} | X)$ and $I_{\text{future}} = I(h_k; S^{(k+1\dots M)} | X)$. Any Pareto-optimal encoder lies on the efficiency frontier of the information plane, where the trade-off between retained and predicted information satisfies the first-order condition:

$$\left. \frac{dI_{\text{future}}}{dI_{\text{past}}} \right|_{h_k^*} = \frac{1}{\beta(k)}.$$

We model the frontier as:

$$I_{\text{future}}(I_{\text{past}}; k) = I_{\text{max}}(k) \left(1 - e^{-\alpha(k) \cdot I_{\text{past}}} \right),$$

with parameters motivated by:

- $I_{\text{max}}(k) \approx C_2(M - k)$, proportional to the length of the remaining sequence.
- $\alpha(k) \approx 1/H(S^{(1\dots k)} | X) \approx 1/(C_1k + C_0)$, inversely proportional to the information content of past tokens.

Differentiating gives:

$$\frac{dI_{\text{future}}}{dI_{\text{past}}} = I_{\text{max}}(k) \cdot \alpha(k) \cdot e^{-\alpha(k) \cdot I_{\text{past}}}.$$

At optimality, the encoder saturates its bottleneck: $I_{\text{past}} \approx H(S^{(1\dots k)} | X) \approx 1/\alpha(k)$. Substituting:

$$\left. \frac{dI_{\text{future}}}{dI_{\text{past}}} \right|_{h_k^*} \approx I_{\text{max}}(k) \cdot \alpha(k) \cdot e^{-1} \approx \frac{C_2(M - k)}{e(C_1k + C_0)}.$$

Applying $\beta(k) = (dI_{\text{future}}/dI_{\text{past}})^{-1}$:

$$\beta(k) = e \cdot \frac{C_1k + C_0}{C_2(M - k)} = e \frac{C_1}{C_2} \cdot \frac{k}{M - k} + O\left(\frac{1}{M - k}\right).$$

Thus, as $k \rightarrow M$, the dominant asymptotic behavior is $\beta(k) \sim \frac{k}{M-k}$, revealing that the optimal trade-off scales with the ratio of processed to remaining sequence length. $\square$

A.4 Proof of Theorem 4.3

Lemma A.2 (Upper Bound on the Entropy $H(p_{\text{CoT}})$ of the CoT Output Distribution). *Let $\bar{p} = \mathbb{E}[p_{\text{CoT}}]$ be the expected next-step distribution over B options, drawn from a Dirichlet prior with parameters α and concentration $\kappa = \sum_{i=1}^B \alpha_i$. Under high concentration — i.e., when $\alpha_j \gg \alpha_{i \neq j}$ for some j — the entropy $H(\bar{p})$ vanishes as κ grows. Specifically, if $\alpha_j = \kappa - (B - 1)c$ and $\alpha_{i \neq j} = c$ for small $c > 0$, then*

$$H(\bar{p}) = O\left(\frac{\log \kappa}{\kappa}\right),$$

Proof. By properties of the Dirichlet distribution, the expected probabilities are given by $\bar{p}_i = \alpha_i / \kappa$. We consider a typical unimodal scenario in which the vast majority of the weight concentrates on a single option j . We set $\alpha_j = \kappa - (B - 1)c$ and $\alpha_{i \neq j} = c$ for all $i \neq j$, where c is a small positive constant.

Under this setting, the expected probabilities are:

$$\begin{aligned}\bar{p}_j &= \frac{\kappa - (B - 1)c}{\kappa} = 1 - \frac{(B - 1)c}{\kappa} \\ \bar{p}_{i \neq j} &= \frac{c}{\kappa}\end{aligned}$$

We compute the entropy of this distribution:

$$\begin{aligned}H(\bar{p}) &= -\bar{p}_j \log \bar{p}_j - \sum_{i \neq j} \bar{p}_i \log \bar{p}_i \\ &= -\left(1 - \frac{(B - 1)c}{\kappa}\right) \log \left(1 - \frac{(B - 1)c}{\kappa}\right) - (B - 1) \frac{c}{\kappa} \log \left(\frac{c}{\kappa}\right)\end{aligned}$$

We analyze the first term using the Taylor expansion $\log(1 - x) \approx -x$. As $\kappa \rightarrow \infty$, we have $\frac{(B-1)c}{\kappa} \rightarrow 0$, and thus:

$$-\left(1 - \frac{(B - 1)c}{\kappa}\right) \log \left(1 - \frac{(B - 1)c}{\kappa}\right) \approx -\left(1 - \frac{(B - 1)c}{\kappa}\right) \left(-\frac{(B - 1)c}{\kappa}\right) = O\left(\frac{1}{\kappa}\right).$$

Next, we analyze the second term:

$$\begin{aligned}-(B - 1) \frac{c}{\kappa} \log \left(\frac{c}{\kappa}\right) &= -(B - 1) \frac{c}{\kappa} (\log c - \log \kappa) \\ &= (B - 1)c \frac{\log \kappa}{\kappa} - (B - 1)c \frac{\log c}{\kappa} = O\left(\frac{\log \kappa}{\kappa}\right).\end{aligned}$$

Combining both terms, as κ becomes large, the entropy is dominated by the $O\left(\frac{\log \kappa}{\kappa}\right)$ term. And we have $\lim_{\kappa \rightarrow \infty} \frac{\log \kappa}{\kappa} = 0$. $\square$

We now begin the proof of Theorem 4.3.

Proof. We analyze the almost sure asymptotic behavior of $D_{\text{KL}}(q_{\text{PR}} \| p_{\text{CoT}})$ under the concentrated Dirichlet prior where $\alpha_j = \kappa - (B - 1)c$ and $\alpha_{i \neq j} = c$.

By the law of large numbers for Dirichlet distributions, as $\kappa \rightarrow \infty$, the random vector p_{CoT} converges almost surely to its mean:

$$p_{\text{CoT}} \xrightarrow{a.s.} \bar{p} = \left(1 - \frac{(B - 1)c}{\kappa}, \frac{c}{\kappa}, \dots, \frac{c}{\kappa}\right).$$

Since the KL divergence is a continuous function of the probability vector (away from the boundary), we have:

$$D_{\text{KL}}(q_{\text{PR}} \| p_{\text{CoT}}) \xrightarrow{a.s.} D_{\text{KL}}(q_{\text{PR}} \| \bar{p}) \quad \text{as } \kappa \rightarrow \infty.$$

Now compute $D_{\text{KL}}(q_{\text{PR}} \| \bar{p})$:

$$\begin{aligned}D_{\text{KL}}(q_{\text{PR}} \| \bar{p}) &= \sum_{i=1}^B \frac{1}{B} \log \frac{1/B}{\bar{p}_i} \\ &= -\log B - \frac{1}{B} \sum_{i=1}^B \log \bar{p}_i \\ &= -\log B - \frac{1}{B} \left[\log \left(1 - \frac{(B - 1)c}{\kappa}\right) + (B - 1) \log \left(\frac{c}{\kappa}\right) \right].\end{aligned}$$

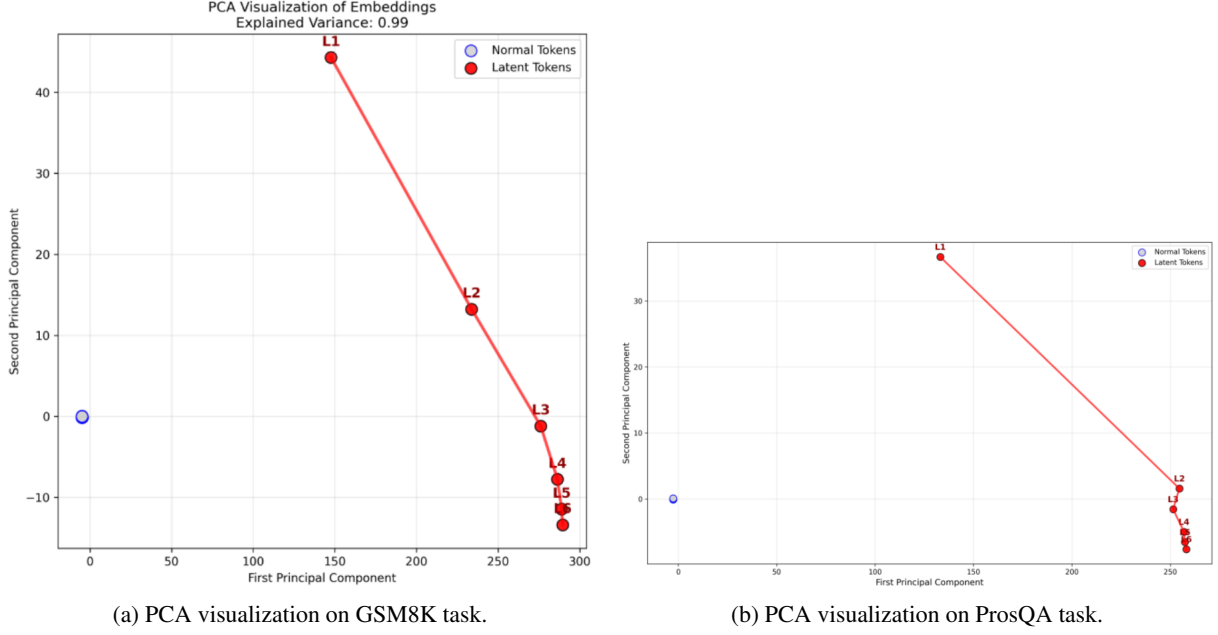


Figure 6: PCA visualization of latent state embeddings for reasoning trajectories. In both (a) and (b), the latent thoughts (L1-L6) demonstrate a clear convergence, supporting our assumption of a compact, attracting set for latent states. This validates that the convergence dynamic is a general property of the latent reasoning process, not specific to a single task.

Using Taylor expansion $\log(1 - x) = -x + O(x^2)$:

$$\begin{aligned} D_{\text{KL}}(q_{\text{PR}} \parallel \bar{p}) &= -\log B - \frac{1}{B} \left[-\frac{(B-1)c}{\kappa} + O\left(\frac{1}{\kappa^2}\right) + (B-1)\log c - (B-1)\log \kappa \right] \\ &= \frac{B-1}{B} \log \kappa - \log B - \frac{(B-1)\log c}{B} + O\left(\frac{1}{\kappa}\right). \end{aligned}$$

Therefore, almost surely as $\kappa \rightarrow \infty$:

$$D_{\text{KL}}(q_{\text{PR}} \parallel p_{\text{CoT}}) = \frac{B-1}{B} \log \kappa - \log B - \frac{(B-1)\log c}{B} + \mathcal{O}\left(\frac{1}{\kappa}\right).$$

Since $H(p_{\text{CoT}}) \rightarrow 0$ as established in Lemma A.2, and $D_{\text{KL}}(q_{\text{PR}} \parallel p_{\text{CoT}})$ grows logarithmically in κ , this proves the exploration deficiency of CoT. $\square$

A.5 Proof of Theorem 4.5

Assumption A.3 (Convergence and Compactness of Latent States). We assume that during inference, the sequence of latent states $\{h_k\}$ converges to a fixed point or enters a compact attracting set as the reasoning step k increases. This is not just a theoretical convenience; it is a behavior we consistently observe empirically. As shown in Figure 6, a PCA visualization of the latent thought embeddings reveals a clear convergent trajectory, where successive states $\{L1, \dots, L6\}$ move progressively closer within a bounded region. The high explained variance (0.99) of the first two principal components confirms that this 2D projection faithfully represents the dynamics in the original high-dimensional space.

Lemma A.4 (Non-Degenerate Output Distribution of Coconut). *For a model trained via the Coconut objective, its optimization can be modeled as solving a Conditional Information Bottleneck (CIB) problem with a finite trade-off parameter $\beta(k) > 0$ (Theorem 4.1). Under this framework, the model’s output probability distribution is non-degenerate: there exists a constant $\delta > 0$ such that for any valid next token u and any reachable latent state h , the probability is strictly bounded away from 1:*

$$\sup_{h \in \mathcal{H}} \left(\max_{u \in N_{\text{valid}}(v)} p(u \mid h) \right) \leq 1 - \delta,$$

where $\mathcal{H}$ is the compact latent state space guaranteed by Assumption A.3.

Proof. Assume, for contradiction, that $\sup_{h \in \mathcal{H}} \max_u p(u | h) = 1$. Since $\mathcal{H}$ is compact and the decoder $p(u | h)$ is continuous, there must exist a latent state $h^* \in \mathcal{H}$ and a token u^* such that $p(u^* | h^*) = 1$.

At h^* , future uncertainty is fully eliminated: $H(S^{(k+1 \dots M)} | h^*, X) = 0$, implying that the predictive information $I(h; S^{(k+1 \dots M)} | X)$ achieves its theoretical maximum.

By Theorem 4.1, the Coconut solution corresponds to an optimal solution of a CIB problem. On the information plane $(I_{\text{past}}, I_{\text{future}})$, all optimal solutions must lie on the efficiency frontier, and any optimal point h_{opt} must satisfy:

$$\left. \frac{dI_{\text{future}}}{dI_{\text{past}}} \right|_{h_{\text{opt}}} = \frac{1}{\beta(k)}.$$

At our hypothetical point h^* , I_{future} reaches its maximum value $I_{\text{max}}(k)$. Under the efficiency frontier model adopted in Theorem 4.1:

$$I_{\text{future}}(I_{\text{past}}; k) = I_{\text{max}}(k) \left(1 - e^{-\alpha(k) \cdot I_{\text{past}}}\right),$$

which is strictly increasing and strictly concave in I_{past} , with derivative:

$$\frac{dI_{\text{future}}}{dI_{\text{past}}} = I_{\text{max}}(k) \cdot \alpha(k) \cdot e^{-\alpha(k) \cdot I_{\text{past}}}.$$

As $I_{\text{future}} \rightarrow I_{\text{max}}(k)$, we must have $I_{\text{past}} \rightarrow \infty$, and the derivative $\rightarrow 0^+$. If I_{future} were to reach $I_{\text{max}}(k)$ at a finite I_{past} (as at h^*), the derivative must be zero to match the asymptotic behavior. Thus, at h^* :

$$\left. \frac{dI_{\text{future}}}{dI_{\text{past}}} \right|_{h^*} = 0.$$

If h^* were a CIB optimum, it must satisfy:

$$\left. \frac{dI_{\text{future}}}{dI_{\text{past}}} \right|_{h^*} = \frac{1}{\beta(k)},$$

implying $1/\beta(k) = 0$, i.e., $\beta(k) \rightarrow \infty$. However, by Theorem 4.1, at any non-terminal stage of curriculum learning ($k < M$), $\beta(k)$ is a finite positive constant (with asymptotic behavior $\beta(k) \sim k/(M - k) < \infty$), so $1/\beta(k) > 0$. This contradicts the requirement that the slope be zero.

Therefore, the assumption $p(u^* | h^*) = 1$ leads to a contradiction. Hence, $\sup \max p(u | h)$ must be strictly less than 1. The lemma is proved.

Intuition: In essence, Coconut inherently avoids deterministic collapse by design: its information bottleneck objective enforces a trade-off that preserves output diversity, even at advanced reasoning stages. $\square$

We now begin the proof of Theorem 4.5

Proof. By Lemma A.4, there exists $\delta > 0$ such that for any valid next token u and reachable state h , the probability satisfies $\max_u p(u | h) \leq 1 - \delta$. Our goal is to derive an upper bound for the KL divergence $D_{\text{KL}}(q_{\text{PR}} \| p_{\text{Coconut}})$.

By definition,

$$D_{\text{KL}}(q_{\text{PR}} \| p) = -\log B - \frac{1}{B} \sum_u \log p(u | v).$$

To bound this quantity from above, we seek the probability distribution that maximizes D_{KL} under the constraints $\sum p_i = 1$ and $\max p_i \leq 1 - \delta$, which is equivalent to minimizing $\sum \log p(u | v)$.

The minimum of $\sum \log p_i$ is achieved by the most concentrated distribution satisfying the constraint:

$$p(u^* | v) = 1 - \delta, \quad \text{and for all } u \neq u^*, \quad p(u | v) = \frac{\delta}{B - 1}.$$

Substituting this worst-case distribution into the KL divergence expression yields an upper bound:

$$\begin{aligned} D_{\text{KL}}(q_{\text{PR}} \| p_{\text{Coconut}}) &\leq -\log B - \frac{1}{B} \left[\log(1 - \delta) + (B - 1) \log \left(\frac{\delta}{B - 1} \right) \right] \\ &= -\frac{B - 1}{B} \log \delta - \left[\log B - \frac{B - 1}{B} \log(B - 1) + \frac{\log(1 - \delta)}{B} \right]. \end{aligned}$$

Denote the bracketed term as $f(B) = \log B - \frac{B-1}{B} \log(B - 1) + \frac{\log(1-\delta)}{B}$. Since $f(B)$ is continuous and bounded on $[2, \infty)$ (with $\lim_{B \rightarrow \infty} f(B) = 0$), its infimum $c = \inf_{B \geq 2} f(B)$ is a finite constant. Therefore, $-f(B)$ is upper-bounded by $-c$.

Meanwhile, since $B \geq 2$, we have $\frac{B-1}{B} \geq \frac{1}{2}$. Also, since $\delta \in (0, 1)$, it follows that $\log \delta < 0$, and thus $-\frac{B-1}{B} \log \delta \leq -\frac{1}{2} \log \delta$.

Combining these bounds, the KL divergence admits a finite upper bound independent of the branching factor B :

$$D_{\text{KL}}(q_{\text{PR}} \| p_{\text{Coconut}}) \leq -\frac{B - 1}{B} \log \delta - c \leq -\frac{1}{2} \log \delta - c.$$

□

A.6 Proof of Theorem 4.7

Proof. We prove that under sub-decisional perturbations, the Explicit CoT trajectory $\hat{S}$ is identical to the noise-free trajectory S^* by induction on the reasoning step k .

Let $f_\theta(\cdot)$ denote the deterministic forward pass of the model. At step k , the model takes the input x and the discrete history $S^{(1 \dots k-1)}$ to produce a logit vector $l_k^* = f_\theta(x, S^{(1 \dots k-1)})$. The noise-free token selection is given by $s_k^* = \text{argmax}(l_k^*)$. The perturbed process generates logits $\hat{l}_k = l_k^* + \epsilon_k$, where ϵ_k satisfies the sub-decisional condition: $\text{argmax}(l_k^* + \epsilon_k) = \text{argmax}(l_k^*)$.

At the first step, both the clean and perturbed processes receive the exact same input x (with an empty history). The clean logits are $l_1^* = f_\theta(x, \emptyset)$. The perturbed logits are $\hat{l}_1 = l_1^* + \epsilon_1$. By the definition of sub-decisional perturbation:

$$\hat{s}_1 = \text{argmax}(\hat{l}_1) = \text{argmax}(l_1^* + \epsilon_1) = \text{argmax}(l_1^*) = s_1^*.$$

Thus, the first token is identical, and the history passed to step 2 is error-free.

Assume that for all steps $t < k$, the generated tokens are identical, i.e., $\hat{s}_t = s_t^*$. This implies the discrete history is identical: $\hat{S}^{(1 \dots k-1)} = S^{*(1 \dots k-1)}$. At step k , since the model f_θ is deterministic and conditioned strictly on the discrete history:

$$\text{Logits}_{\text{pre-noise}} = f_\theta(x, \hat{S}^{(1 \dots k-1)}) = f_\theta(x, S^{*(1 \dots k-1)}) = l_k^*.$$

The noise process adds the perturbation: $\hat{l}_k = l_k^* + \epsilon_k$. Applying the sub-decisional constraint again, the discrete projection filters out the noise:

$$\hat{s}_k = \text{argmax}(l_k^* + \epsilon_k) = s_k^*.$$

Consequently, the noise term ϵ_k is discarded by the argmax operator and does not propagate to the state of the next step.

By induction, $\hat{s}_k = s_k^*$ for all $k = 1, \dots, M$. Therefore, the probability of trajectory divergence is zero:

$$\mathbb{P}[\hat{S} \neq S^*] = 0.$$

This mathematically confirms the *discretization-reset* mechanism described in Section 4.3: the argmax operator effectively acts as a non-linear filter that resets the internal state to a clean integer grid at every step, preventing the accumulation of sub-decisional noise. □

A.7 Proof of Theorem 4.8

Proof. From the recurrence

$$E_k = f_\theta(h_{k-1}) - f_\theta(h_{k-1}^*) + \epsilon_h^{(k)},$$

taking the squared norm and expectation, and using the independence and zero-mean property of the noise (which causes cross terms to vanish), we obtain:

$$\mathbb{E}[\|E_k\|^2] \leq L_F^2 \mathbb{E}[\|E_{k-1}\|^2] + d\sigma_h^2.$$

This is a non-homogeneous linear recurrence. Solving it with initial condition $\mathbb{E}[\|E_0\|^2] = 0$ yields:

$$\mathbb{E}[\|E_M\|^2] = \begin{cases} \frac{1 - L_F^{2M}}{1 - L_F^2} \cdot d\sigma_h^2, & L_F \neq 1, \\ M \cdot d\sigma_h^2, & L_F = 1. \end{cases}$$

In both cases, $\mathbb{E}[\|E_M\|^2] > 0$ for all $M \geq 1$, which completes the proof. $\square$

A.8 Derivation of the Normalized Accuracy Function

In this section, we provide a rigorous derivation of the **Normalized Accuracy** curve $A(\sigma)$ to explain the characteristic inverted S-shape (sigmoid-like) degradation observed in Figure 3. We model the causal chain from injected noise in the latent space to the probabilistic failure of token ranking in the output space.

We begin by modeling the accumulation of noise in the hidden states. Let h_M^* denote the clean latent state at the final reasoning step M . Consistent with Theorem 4.8 (Compounding Error), the injection of Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ at each step accumulates over the reasoning trajectory. Assuming a linearized approximation of the transition dynamics, the final perturbed state $\tilde{h}_M$ can be expressed as:

$$\tilde{h}_M = h_M^* + \epsilon_{\text{acc}}, \quad \text{where} \quad \epsilon_{\text{acc}} \sim \mathcal{N}(0, \sigma_{\text{eff}}^2 I_d).$$

Here, ϵ_{acc} represents the effective accumulated noise vector in $\mathbb{R}^d$. The variance $\sigma_{\text{eff}}^2 \approx \alpha \cdot \sigma^2$ scales linearly with the injection variance σ^2 , where the scaling factor α depends on the path length M and the Lipschitz constant of the network.

Since discrete token decisions are not made in the latent space but in the logit space, we must trace how this perturbation propagates through the unembedding layer. Let $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$ be the weight matrix of the language model head. The perturbed logits $\tilde{l}$ are obtained by projecting the final latent state:

$$\tilde{l} = W_U \tilde{h}_M = W_U (h_M^* + \epsilon_{\text{acc}}) = l^* + \eta,$$

where l^* denotes the clean logits, and $\eta = W_U \epsilon_{\text{acc}}$ represents the noise projected into the logit space. Since ϵ_{acc} is Gaussian and W_U is a linear transformation, η follows a multivariate Gaussian distribution over $\mathbb{R}^{|\mathcal{V}|}$.

A prediction is successful if the noise is insufficient to alter the ranking of the top tokens. Let $i^* = \text{argmax}(l^*)$ be the index of the correct token, and let j^* be the index of the strongest competitor. The model retains the correct prediction if $\tilde{l}_{i^*} > \tilde{l}_{j^*}$. Substituting $\tilde{l} = l^* + \eta$, this condition becomes:

$$l_{i^*}^* + \eta_{i^*} > l_{j^*}^* + \eta_{j^*} \implies \eta_{j^*} - \eta_{i^*} < l_{i^*}^* - l_{j^*}^*.$$

Crucially, the Right-Hand Side is precisely the Logit Decision Margin Δ_l defined in Definition 4.10. Therefore, the success condition simplifies to:

$$\xi < \Delta_l, \quad \text{where} \quad \xi = \eta_{j^*} - \eta_{i^*}.$$

To quantify the probability of this event, we analyze the distribution of the random variable ξ . We express ξ as a linear combination of the accumulated noise vector: $\xi = (w_{j^*} - w_{i^*})^\top \epsilon_{\text{acc}}$. Because ϵ_{acc} is isotropic Gaussian, $\xi \sim \mathcal{N}(0, \sigma_1^2)$. The scalar variance σ_1^2 is given by:

$$\sigma_1^2 = \|w_{j^*} - w_{i^*}\|_2^2 \cdot \sigma_{\text{eff}}^2.$$

Letting $C = \alpha \cdot \|w_{j^*} - w_{i^*}\|_2^2$, we derive the relationship $\sigma_1^2 = C \cdot \sigma^2$. Thus, $\sigma_1 = \sqrt{C} \cdot \sigma$.

The Normalized Accuracy $A(\sigma)$ is defined as the probability of success $\mathbb{P}(\xi < \Delta_l)$. We evaluate this using the CDF of the standard normal distribution, $\Phi(z)$. By standardizing via $Z = \xi/\sigma_1 \sim \mathcal{N}(0, 1)$, we obtain:

$$\begin{aligned} A(\sigma) &= \mathbb{P}(\xi < \Delta_l) \\ &= \mathbb{P}\left(\frac{\xi}{\sigma_1} < \frac{\Delta_l}{\sigma_1}\right) \\ &= \Phi\left(\frac{\Delta_l}{\sqrt{C} \cdot \sigma}\right). \end{aligned} \tag{4}$$

This derivation provides a rigorous physical interpretation for the experimental results. In the limit as $\sigma \rightarrow 0$, the argument $\frac{\Delta_l}{\sqrt{C}\sigma} \rightarrow \infty$, and $A(\sigma) \rightarrow 1$. This explains the “sub-decisional plateau,” where accuracy remains stable because the noise is strictly contained within the safety margin Δ_l . As σ increases, the argument decreases, leading to the characteristic sigmoid-like decay observed in the accuracy curve. This confirms that robustness is analytically determined by the ratio between the Symbolic Index-induced margin Δ_l and the accumulated noise variance.

A.9 Proof of Theorem 4.11

Proof. The Logit Decision Margin Δ_l is defined as the difference between the logits of the most and second-most likely tokens, $\Delta_l = l_{i^*} - l_{j^*}$. We aim to establish a lower bound for this quantity based on the Symbolic Index, $\mathcal{I}_S$.

First, we relate the logits to probabilities using the softmax function, $p_i = e^{l_i} / \sum_k e^{l_k}$. The ratio of the probabilities of the top two tokens is:

$$\frac{p_{i^*}}{p_{j^*}} = \frac{e^{l_{i^*}}}{e^{l_{j^*}}} = e^{l_{i^*} - l_{j^*}} = e^{\Delta_l}.$$

By taking the natural logarithm of both sides, we can express the logit margin in terms of probabilities:

$$\Delta_l = \log(p_{i^*}) - \log(p_{j^*}).$$

By the definition of the Symbolic Index, the probability of the most likely token is $p_{i^*} = \mathcal{I}_S$.

Next, we establish an upper bound for the probability of the second-most likely token, p_{j^*} . The sum of probabilities over all possible tokens is 1. Therefore, the sum of probabilities of all tokens except the most likely one is $1 - p_{i^*} = 1 - \mathcal{I}_S$. Since p_{j^*} is the largest among these remaining probabilities, it cannot be greater than their sum. Thus, we have the inequality:

$$p_{j^*} \leq 1 - \mathcal{I}_S.$$

Because the logarithm function is monotonically increasing, this implies $\log(p_{j^*}) \leq \log(1 - \mathcal{I}_S)$. Consequently, its negative, $-\log(p_{j^*})$, satisfies the reverse inequality: $-\log(p_{j^*}) \geq -\log(1 - \mathcal{I}_S)$.

Finally, we substitute our findings back into the equation for Δ_l :

$$\Delta_l = \log(\mathcal{I}_S) - \log(p_{j^*}) \geq \log(\mathcal{I}_S) - \log(1 - \mathcal{I}_S).$$

Using the properties of logarithms, we arrive at the desired lower bound:

$$\Delta_l \geq \log\left(\frac{\mathcal{I}_S}{1 - \mathcal{I}_S}\right).$$

□

A.10 Proof of Theorem 4.12

Proof. The KL divergence between the ideal uniform prior q_{PR} (where $q_i = 1/B$ for all $i = 1, \dots, B$) and the model's output distribution p is given by:

$$D_{\text{KL}}(q_{\text{PR}} \| p) = \sum_{i=1}^B q_i \log \left(\frac{q_i}{p_i} \right) = \sum_{i=1}^B \frac{1}{B} \log \left(\frac{1/B}{p_i} \right) = -\log B - \frac{1}{B} \sum_{i=1}^B \log p_i.$$

To find a lower bound for $D_{\text{KL}}(q_{\text{PR}} \| p)$, we must find an upper bound for the term $\sum_{i=1}^B \log p_i$, subject to the constraints imposed by the model's distribution. The constraints are:

1. $\sum_{i=1}^B p_i = 1$
2. $\max_i p_i = \mathcal{I}_S$

We want to solve the following optimization problem:

$$\max_{p_1, \dots, p_B} \sum_{i=1}^B \log p_i \quad \text{s.t.} \quad \sum p_i = 1 \text{ and } \max p_i = \mathcal{I}_S.$$

Let $p_{i^*} = \mathcal{I}_S$ be the maximal probability. The sum of the remaining $B - 1$ probabilities is $\sum_{k \neq i^*} p_k = 1 - \mathcal{I}_S$. The function $f(p_1, \dots, p_{B-1}) = \sum_{k \neq i^*} \log p_k$ is concave. By Jensen's inequality, this sum is maximized when the remaining probabilities are distributed as uniformly as possible, i.e., when $p_k = \frac{1 - \mathcal{I}_S}{B - 1}$ for all $k \neq i^*$.

This distribution, which makes the probabilities as "flat" as possible given the peak at $\mathcal{I}_S$, provides the maximum possible value for $\sum \log p_i$. Substituting this extremal distribution, we find the upper bound:

$$\max \sum_{i=1}^B \log p_i = \log(\mathcal{I}_S) + (B - 1) \log \left(\frac{1 - \mathcal{I}_S}{B - 1} \right).$$

Now, we substitute this upper bound back into the expression for the KL divergence. Since we are subtracting this term, its upper bound yields a lower bound for the KL divergence:

$$D_{\text{KL}}(q_{\text{PR}} \| p) \geq -\log B - \frac{1}{B} \left[\log(\mathcal{I}_S) + (B - 1) \log \left(\frac{1 - \mathcal{I}_S}{B - 1} \right) \right].$$

This completes the proof. □

A.11 Proof of Theorem 5.1

The theorem is a counterexample-based statement. It does not claim that every possible no-curriculum distribution must fail. Instead, it isolates one structural failure mode: if direct latent training induces a shortcut-dominated distribution whose support differs from the expert latent policy, maximum-likelihood training can inherit that bias rather than recover the expert policy.

Proof. We prove this by constructing a counterexample. We define a reasoning task where the biased training distribution has a strictly lower success rate than the expert. We then show that the Maximum Likelihood Estimator (MLE) converges to a policy that mimics this suboptimal behavior, leading to a permanent performance gap.

1. Instance Construction Consider a $d = 3$ dimensional latent space with three discrete states:

- h_{expert} : The correct reasoning path. Feature $\phi(h_{\text{expert}}) = [1, 0, 0]^\top$.
- h_{shortcut} : A superficial shortcut. Feature $\phi(h_{\text{shortcut}}) = [0, 1, 1]^\top$.
- h_{bad} : Irrelevant noise. Feature $\phi(h_{\text{bad}}) = [0, 1, 0]^\top$.

The valuation function $V(h)$ (correctness of the answer derived from h) is defined as:

$$V(h) = \begin{cases} 1 & \text{if } h = h_{\text{expert}} \\ 0 & \text{otherwise} \end{cases}$$

The **Task Success Rate** for a policy parameterized by θ is thus $\mathcal{R}(\theta) = P_\theta(h_{\text{expert}})$.

2. The Expert vs. The Biased Distribution Let the expert parameter be $\theta^* = [10, 0, 0]^\top$. The unnormalized log-probability (score) for the expert state is 10, while others are 0. The expert's success rate is:

$$\mathcal{R}(\theta^*) = \frac{e^{10}}{e^{10} + e^0 + e^0} \approx 1.0.$$

Now, consider the biased data distribution P_{biased} . Assume the dataset consists entirely of shortcut reasoning, i.e., $P_{\text{biased}}(h_{\text{shortcut}}) = 1$. The success rate of this data distribution is:

$$\mathcal{R}(P_{\text{biased}}) = \mathbb{E}_{h \sim P_{\text{biased}}} [V(h)] = 1 \cdot V(h_{\text{shortcut}}) = 0.$$

This satisfies the theorem's condition $\mathcal{R}(P_{\text{biased}}) \leq \mathcal{R}(\theta^*) - \Delta$ with gap $\Delta \approx 1$.

3. MLE Convergence and the Performance Gap We train a model $\hat{\theta}_{\text{MLE}}$ on this biased dataset. MLE maximizes the log-likelihood of the observed data (h_{shortcut}). To maximize the score $\langle \theta, \phi(h_{\text{shortcut}}) \rangle = \theta_2 + \theta_3$, while adhering to normalization constraints, the optimizer will drive $\theta_2 + \theta_3 \rightarrow \infty$.

A characteristic solution for large N is $\hat{\theta}_{\text{MLE}} \approx [0, K, K]^\top$ for a large K . Under this learned policy, the probability mass shifts entirely to the shortcut:

$$P_{\hat{\theta}_{\text{MLE}}}(h_{\text{shortcut}}) \propto e^{2K}, \quad P_{\hat{\theta}_{\text{MLE}}}(h_{\text{expert}}) \propto e^0 = 1.$$

The success rate of the learned model is:

$$\mathcal{R}(\hat{\theta}_{\text{MLE}}) = P_{\hat{\theta}_{\text{MLE}}}(h_{\text{expert}}) = \frac{1}{1 + e^{2K} + e^K} \xrightarrow{K \rightarrow \infty} 0.$$

4. Conclusion Comparing the final success rates:

$$\mathcal{R}(\hat{\theta}_{\text{MLE}}) \approx 0 \quad \text{vs.} \quad \mathcal{R}(\theta^*) \approx 1.$$

Thus, we have derived the bound:

$$\mathcal{R}(\hat{\theta}_{\text{MLE}}) \leq \mathcal{R}(\theta^*) - C,$$

where $C \approx 1$ is a strictly positive constant derived from the initial gap Δ . This proves that, for this biased latent-state construction, training without a mechanism that corrects the distributional bias leaves the learned policy bounded away from optimality. $\square$

A.12 Proof of Theorem 5.2

Justification for the Imitation Learning Framework Our theoretical analysis, particularly Theorem 5.2, models the training process within the standard framework of Imitation Learning (IL). This framework assumes access to an i.i.d. dataset of input-latent state pairs $D_c = \{(x_i, h_i)\}_{i=1}^n$ sampled from an expert latent policy $P_{\theta^*}(h|x)$. However, the Coconut training paradigm is supervised on predicting future tokens from expert trajectories $D_S = \{(x_i, S_i^*)\}_{i=1}^n$. This section provides a formal justification for the equivalence between training a Coconut model and performing Maximum Likelihood Estimation (MLE) in this IL setting.

The core argument is as follows:

- Definition of the Implicit Expert Latent Policy:** By Theorem 4.1 (Coconut-CIB Duality), the objective of Coconut training is to learn an encoder E_θ such that the latent state $h_k = E_\theta(S^{(1...k)})$ becomes a **Minimal Sufficient Statistic** of the past $S^{(1...k)}$ for predicting the future $S^{(k+1...M)}$. Let

E_{θ^*} be the optimal encoder that solves this CIB problem. We can define an **implicit expert latent policy** $P_{\theta^*}(h|x)$ as the distribution induced by applying the optimal encoder to the distribution of expert trajectory prefixes:

$$h \sim P_{\theta^*}(h|x) \quad \text{where} \quad h = E_{\theta^*}(S_{\text{past}}^*) \text{ and } S_{\text{past}}^* \sim P_{\text{expert}}(S_{\text{past}}|x).$$

The support of this policy, $P_{\theta^*}(h|x)$, constitutes the space of ideal latent states.

2. **Equivalence of the Optimization Process:** The Coconut objective is to minimize the expected negative log-likelihood:

$$\min_{\theta} \mathcal{L}(\theta) = \mathbb{E}_{(x, S^*) \sim P_{\text{expert}}} [-\log p(S_{\text{future}}^* | E_{\theta}(S_{\text{past}}^*), x)].$$

By the Data Processing Inequality, for any encoder E_{θ} , we have:

$$I(S_{\text{past}}^*; S_{\text{future}}^* | x) \geq I(E_{\theta}(S_{\text{past}}^*); S_{\text{future}}^* | x).$$

The optimal encoder E_{θ^*} preserves the predictive information needed for the future suffix, so the conditional distribution can be represented as $p(S_{\text{future}}^* | E_{\theta^*}(S_{\text{past}}^*), x) = p(S_{\text{future}}^* | h^*, x)$. Therefore, the loss $\mathcal{L}(\theta)$ can be interpreted as measuring predictive divergence between the latent distribution induced by E_{θ} and the ideal latent distribution $P_{\theta^*}(h|x)$. Minimizing the loss encourages the induced latent distribution $p_{\theta}(h|x)$ to approach $P_{\theta^*}(h|x)$ in the predictive directions relevant to the suffix.

Conclusion: The Coconut training process, while formally supervised on future token generation, has an intrinsic CIB objective that encourages the latent states generated by the encoder, $p_{\theta}(h|x)$, to match the predictive behavior of the implicit expert latent policy $P_{\theta^*}(h|x)$. Consequently, minimizing the Coconut loss on the expert trajectory dataset D_S can be viewed as MLE on an implicit dataset D_c constructed from $h_i^* = E_{\theta^*}(S_{i,\text{past}}^*)$, up to the modeling assumptions stated above. This provides a formal justification for applying our IL-based analysis to the Coconut model without claiming that the equivalence holds for arbitrary latent training dynamics.

Lemma A.5 (Self-Normalized Bound for Vector Martingales ((Abbasi-yadkori et al., 2011))). *Let $\{\mathcal{F}_i\}_{i=0}^n$ be a filtration. Let $\{g_i\}_{i=1}^n$ be a sequence of random vectors in $\mathbb{R}^d$ such that g_i is $\mathcal{F}_i$ -measurable and $\mathbb{E}[g_i | \mathcal{F}_{i-1}] = 0$. Assume there exists a constant $\sigma > 0$ such that for any unit vector $v \in \mathbb{R}^d$, the real-valued random variable $\langle v, g_i \rangle$ is conditionally σ -sub-Gaussian, i.e.,*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda \langle v, g_i \rangle) | \mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

Let V be a $d \times d$ positive definite matrix. Define

$$V_n = V + \sum_{i=1}^n g_i g_i^{\top} \quad \text{and} \quad G_n = \sum_{i=1}^n g_i.$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following inequality holds:

$$\|G_n\|_{V_n^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{\det(V_n)^{1/2} \det(V)^{-1/2}}{\delta}\right).$$

Lemma A.6 (Construction of a Confidence Set for Parameter Estimation). *Let $\hat{\theta}_{MLE}$ be the parameter estimate obtained by maximizing the log-likelihood function on a curriculum dataset $D_c = \{(x_i, h_i)\}_{i=1}^n$, where the data is sampled i.i.d. from the expert distribution P_{θ^*} :*

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \log P_{\theta}(h_i | x_i)$$

Assume the following regularity conditions for the log-linear model hold:

1. **(Bounded Features):** For all (x, h) , $\|\phi(x, h)\|_2 \leq L$.
2. **(Strong Convexity):** The expected negative log-likelihood function $L(\theta) = \mathbb{E}_{x, h \sim P_{\theta^*}}[-\log P_{\theta}(h|x)]$ is $\lambda_{\min}$ -strongly convex with respect to θ over the parameter space. Its Hessian (the Fisher information matrix) $\nabla^2 L(\theta) = I(\theta)$ satisfies $I(\theta) \succeq \lambda_{\min} I$.
3. **(Sub-Gaussian Score):** The score function at the true parameter θ^* , $\nabla_{\theta} \log P_{\theta^*}(h | x)$, is a zero-mean, σ^2 -sub-Gaussian random vector.
4. **(Bounded Parameters):** $\|\theta^*\|_2 \leq B$, and the optimization is performed within the compact set $\mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}$.

Then for any $\delta \in (0, 1)$, there exists a constant $C > 0$ (depending only on $L, \sigma, \lambda_{\min}, B$) such that with probability at least $1 - \delta$, the true expert parameter θ^* satisfies:

$$\left\| \hat{\theta}_{MLE} - \theta^* \right\|_{\hat{H}_n}^2 \leq \beta_n^2(\delta)$$

where:

- $\hat{H}_n = \frac{1}{n} \sum_{i=1}^n \nabla^2(-\log P_{\hat{\theta}_{MLE}}(h_i | x_i))$ is the empirical Hessian evaluated at $\hat{\theta}_{MLE}$,
- $\beta_n^2(\delta) = C \cdot \frac{d \log(n) + \log(1/\delta)}{n}$ is the squared radius of the confidence set.

In other words, the confidence set $\Theta(\hat{\theta}_{MLE}) = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_{MLE} \right\|_{\hat{H}_n}^2 \leq \beta_n^2(\delta) \right\}$ covers the true parameter θ^* with high probability.

Proof. This proof combines the first-order optimality condition, concentration inequalities, and self-normalized analysis techniques to construct a non-asymptotic, high-probability bound.

1. Notation and Basic Relations Let $L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell_i(\theta)$ be the empirical negative log-likelihood, where $\ell_i(\theta) = -\log P_{\theta}(h_i|x_i)$. By definition, $\hat{\theta}_{MLE}$ is the minimizer of $L_n(\theta)$ within the parameter space $\mathcal{B}$. Since θ^* is the minimizer of the expected risk $L(\theta)$ (because the KL divergence $D_{KL}(P_{\theta^*} \| P_{\theta}) = L(\theta) - L(\theta^*)$ is non-negative), we have $\nabla L(\theta^*) = 0$.

2. Using Optimality and Convexity By the optimality of $\hat{\theta}_{MLE}$, we have $L_n(\hat{\theta}_{MLE}) \leq L_n(\theta^*)$. Since $L_n(\theta)$ is a convex function, we can derive a basic inequality that connects the parameter error to an empirical process:

$$\begin{aligned} L(\hat{\theta}_{MLE}) - L(\theta^*) &\leq L(\hat{\theta}_{MLE}) - L(\theta^*) - (L_n(\hat{\theta}_{MLE}) - L_n(\theta^*)) \\ &= (L - L_n)(\hat{\theta}_{MLE}) - (L - L_n)(\theta^*) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}[\ell_i(\hat{\theta}_{MLE})] - \ell_i(\hat{\theta}_{MLE}) \right) - (\mathbb{E}[\ell_i(\theta^*)] - \ell_i(\theta^*)) \end{aligned}$$

By Assumption 2, $L(\theta)$ is $\lambda_{\min}$ -strongly convex, so $L(\hat{\theta}_{MLE}) - L(\theta^*) \geq \frac{\lambda_{\min}}{2} \|\hat{\theta}_{MLE} - \theta^*\|_2^2$. Combining this with the above inequality yields:

$$\frac{\lambda_{\min}}{2} \|\hat{\theta}_{MLE} - \theta^*\|_2^2 \leq \sup_{\theta \in \mathcal{B}} |(L - L_n)(\theta) - (L - L_n)(\theta^*)|$$

Bounding the empirical process on the right-hand side can yield a preliminary L_2 -norm convergence bound of $\mathcal{O}(\sqrt{d/n})$. However, to obtain a sharper, weighted-norm bound, we turn to self-normalized analysis.

3. Applying Self-Normalized Bounds We directly analyze the error relation derived from the first-order condition. Let $\Delta = \hat{\theta}_{\text{MLE}} - \theta^*$. A first-order Taylor expansion of $\nabla L_n(\hat{\theta}_{\text{MLE}}) = 0$ around θ^* gives:

$$0 = \nabla L_n(\hat{\theta}_{\text{MLE}}) = \nabla L_n(\theta^*) + \int_0^1 \nabla^2 L_n(\theta^* + t\Delta) dt \cdot \Delta$$

Rearranging, we get:

$$\bar{H}_n \Delta = -\nabla L_n(\theta^*) \quad (5)$$

where $\bar{H}_n = \int_0^1 \nabla^2 L_n(\theta^* + t\Delta) dt$ is the mean Hessian matrix on the line segment between θ^* and $\hat{\theta}_{\text{MLE}}$.

The core of the analysis lies in bounding the norm of the empirical score $\nabla L_n(\theta^*)$. Let $g_i = \nabla_{\theta} \ell_i(\theta^*) = \nabla_{\theta} (-\log P_{\theta^*}(h_i|x_i))$. Since $\mathbb{E}[g_i|\mathcal{F}_{i-1}] = \nabla L(\theta^*) = 0$, the sequence $\{g_i\}_{i=1}^n$ is a vector-valued martingale difference sequence. Under our regularity conditions (specifically, Assumption 3), $\{g_i\}$ is also conditionally sub-Gaussian. We can therefore apply the self-normalized bound from Lemma A.5.

Let's map the variables to Lemma A.5: we set $V = \lambda I$ for some small regularization constant $\lambda > 0$, which ensures initial positive definiteness. Then we have $G_n = \sum_{i=1}^n g_i$ and $V_n = \lambda I + \sum_{i=1}^n g_i g_i^\top$. Applying the lemma, we get that with probability at least $1 - \delta$:

$$\left\| \sum_{i=1}^n g_i \right\|_{(\lambda I + \sum_{i=1}^n g_i g_i^\top)^{-1}}^2 \leq 2\sigma^2 \log \left(\frac{\det(\lambda I + \sum_{i=1}^n g_i g_i^\top)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right).$$

Under the bounded feature assumption, we can bound the determinant term and simplify the expression. Using the fact that $\sum g_i g_i^\top \approx n \cdot \mathbb{E}[g g^\top] = n \cdot I(\theta^*)$ and properties of the determinant, one can derive that:

$$\|\nabla L_n(\theta^*)\|_{(I(\theta^*))^{-1}}^2 = \frac{1}{n^2} \left\| \sum_{i=1}^n g_i \right\|_{(I(\theta^*))^{-1}}^2 \lesssim \frac{d \log(n) + \log(1/\delta)}{n} \quad (6)$$

where $\lesssim$ hides logarithmic factors and constants dependent on model parameters.

4. Connecting the Error and Gradient Bounds Returning to Eq. (5), we have $\Delta = -\bar{H}_n^{-1} \nabla L_n(\theta^*)$. We want to bound $\|\Delta\|_{\hat{H}_n}^2$. By matrix concentration inequalities and Assumptions 1 and 4, it can be shown that the empirical Hessian $\nabla^2 L_n(\theta)$ concentrates uniformly around its expectation $I(\theta)$ over the parameter space $\mathcal{B}$. This implies that both $\bar{H}_n$ and $\hat{H}_n$ are close to $I(\theta^*)$ with high probability. Formally, there exists $\epsilon_n \rightarrow 0$ such that, with high probability, $\|\bar{H}_n - I(\theta^*)\|_{\text{op}} \leq \epsilon_n$ and $\|\hat{H}_n - I(\theta^*)\|_{\text{op}} \leq \epsilon_n$. Therefore, we can approximate:

$$\begin{aligned} \|\Delta\|_{\hat{H}_n}^2 &= \Delta^\top \hat{H}_n \Delta \\ &= (\nabla L_n(\theta^*))^\top \bar{H}_n^{-1} \hat{H}_n \bar{H}_n^{-1} \nabla L_n(\theta^*) \\ &\approx (\nabla L_n(\theta^*))^\top (I(\theta^*))^{-1} \nabla L_n(\theta^*) \\ &= \|\nabla L_n(\theta^*)\|_{(I(\theta^*))^{-1}}^2 \end{aligned}$$

By more rigorously handling the approximation $\bar{H}_n^{-1} \hat{H}_n \bar{H}_n^{-1} \approx (I(\theta^*))^{-1}$ and controlling the error terms, combined with Eq. (6), it can ultimately be shown that:

$$\left\| \hat{\theta}_{\text{MLE}} - \theta^* \right\|_{\hat{H}_n}^2 \leq C \cdot \frac{d \log(n) + \log(1/\delta)}{n}$$

where the constant C absorbs all terms dependent on $L, \sigma, \lambda_{\min}, B$. This completes the proof of the lemma. $\square$

We can now prove the main theorem of this section.

Proof. The proof proceeds in three steps: first, we leverage Lemma A.6 to bound the KL divergence between the model distribution $P_{\hat{\theta}}$ and the expert distribution P_{θ^*} ; second, we convert the KL divergence bound into a Total Variation (TV) distance bound using Pinsker's inequality; finally, we use the properties of TV distance to bound the difference in task success rates between the two policies.

1. Bounding the KL Divergence We aim to bound the KL divergence $D_{KL}(P_{\theta^*} \| P_{\hat{\theta}})$. For the log-linear model family we consider, the KL divergence has the form:

$$D_{KL}(P_{\theta^*}(\cdot|x) \| P_{\hat{\theta}}(\cdot|x)) = \mathbb{E}_{h \sim P_{\theta^*}(\cdot|x)} [\log P_{\theta^*}(h|x) - \log P_{\hat{\theta}}(h|x)]$$

Performing a second-order Taylor expansion of $\log P_{\hat{\theta}}(h|x)$ around θ^* , we have:

$$\log P_{\hat{\theta}}(h|x) \approx \log P_{\theta^*}(h|x) + \langle \nabla_{\theta} \log P_{\theta^*}(h|x), \Delta \rangle + \frac{1}{2} \Delta^{\top} \nabla_{\theta}^2 \log P_{\theta^*}(h|x) \Delta$$

where $\Delta = \hat{\theta} - \theta^*$. Substituting this into the KL divergence expression and taking the expectation, we note that $\mathbb{E}_{h \sim P_{\theta^*}} [\nabla_{\theta} \log P_{\theta^*}] = 0$, which yields the quadratic approximation of KL divergence in terms of the Fisher information matrix:

$$\mathbb{E}_{x \sim \rho} [D_{KL}(P_{\theta^*} \| P_{\hat{\theta}})] \approx \frac{1}{2} \Delta^{\top} I(\theta^*) \Delta = \frac{1}{2} \|\hat{\theta} - \theta^*\|_{I(\theta^*)}^2$$

More rigorously, this approximation can be shown to hold with controllable higher-order terms. Since Lemma A.6 shows that $\|\hat{\theta} - \theta^*\|_{\hat{H}_n}$ is small with high probability and $\hat{H}_n$ concentrates around $I(\theta^*)$, we can translate the result of Lemma A.6 into a bound on the KL divergence. With probability at least $1 - \delta$:

$$\mathbb{E}_{x \sim \rho} [D_{KL}(P_{\theta^*} \| P_{\hat{\theta}})] \leq C' \cdot \|\hat{\theta} - \theta^*\|_{\hat{H}_n}^2 \leq C \cdot \frac{d \log n + \log(1/\delta)}{n} \quad (7)$$

where C' and C are positive constants.

2. From KL Divergence to Total Variation Distance The Total Variation (TV) distance is defined as $TV(P, Q) = \frac{1}{2} \int |p(x) - q(x)| dx$. Pinsker's inequality establishes a relationship between KL divergence and TV distance:

$$TV(P, Q) \leq \sqrt{\frac{1}{2} D_{KL}(P \| Q)}$$

Applying this inequality to our model and expert distributions, and combining it with Eq. (7), we obtain a high-probability bound on the TV distance:

$$\begin{aligned} \mathbb{E}_{x \sim \rho} [TV(P_{\theta^*} \| P_{\hat{\theta}})] &\leq \mathbb{E}_{x \sim \rho} \left[\sqrt{\frac{1}{2} D_{KL}(P_{\theta^*} \| P_{\hat{\theta}})} \right] \\ &\leq \sqrt{\frac{1}{2} \mathbb{E}_{x \sim \rho} [D_{KL}(P_{\theta^*} \| P_{\hat{\theta}})]} \quad (\text{by Jensen's inequality}) \\ &\leq \sqrt{\frac{C}{2} \cdot \frac{d \log n + \log(1/\delta)}{n}} \\ &= \mathcal{O} \left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}} \right) \end{aligned} \quad (8)$$

3. From Total Variation Distance to Task Success Rate A key property of the TV distance is that it bounds the difference in expectation for any bounded function. Our task success function $V(h) \in \{0, 1\}$ is a bounded function (with bound 1). Therefore,

$$\begin{aligned} |\text{SuccessRate}(\hat{\theta}) - \text{SuccessRate}(\theta^*)| &= \left| \mathbb{E}_{x \sim \rho, h \sim P_{\hat{\theta}}} [V(h)] - \mathbb{E}_{x \sim \rho, h \sim P_{\theta^*}} [V(h)] \right| \\ &= \left| \mathbb{E}_{x \sim \rho} \left[\mathbb{E}_{h \sim P_{\hat{\theta}}} [V(h)] - \mathbb{E}_{h \sim P_{\theta^*}} [V(h)] \right] \right| \\ &\leq \mathbb{E}_{x \sim \rho} \left| \mathbb{E}_{h \sim P_{\hat{\theta}}} [V(h)] - \mathbb{E}_{h \sim P_{\theta^*}} [V(h)] \right| \\ &\leq \mathbb{E}_{x \sim \rho} \left[\sup_{H' \subseteq H} |P_{\hat{\theta}}(H') - P_{\theta^*}(H')| \right] \\ &= \mathbb{E}_{x \sim \rho} [TV(P_{\theta^*} \| P_{\hat{\theta}})] \end{aligned} \quad (9)$$

Substituting the bound from Eq. (8) into Eq. (9), we finally obtain the bound on the difference in success rates:

$$|\text{SuccessRate}(\hat{\theta}) - \text{SuccessRate}(\theta^*)| \leq \mathcal{O} \left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}} \right)$$

Since we are concerned with the lower bound on the success rate, this is equivalent to:

$$\text{SuccessRate}(\hat{\theta}) \geq \text{SuccessRate}(\theta^*) - \mathcal{O} \left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}} \right)$$

This result shows that, under curriculum training, the task success rate of the learned policy converges to that of the expert policy with high probability, at a rate of $\mathcal{O}(\sqrt{(d \log n)/n})$. This completes the proof. $\square$