Probabilistic Chain-of-Thought: Sequential Bayesian Inference over Latent Reasoning Correctness

Suriya Dev Saravanakumar [*1] **Ezra Matiwos Wesenie** [*2] **Kishore Nuthalapati** [*3] **Laksh Patel** [4]

Abstract

Chain-of-thought prompting elicits multi-step reasoning from large language models, yet existing approaches treat confidence at each step as an independent signal. This independence assumption contradicts the autoregressive generation process, wherein errors at early steps propagate forward and corrupt downstream outputs, creating epistemic blind spots where a model appears locally certain but is globally unreliable, motivating sequence-level probabilistic inference as a principled foundation for reliable neural mathematical reasoning. We introduce *Probabilistic Chain-of-Thought* (PCoT), which models a reasoning chain as a Hidden Markov Model over latent step correctness and performs exact posterior inference via the forward-backward algorithm. PCoT yields a principled answer confidence C_{final} and a posterior-driven reflection policy that dominates raw-score threshold rules under the model. On MATH and GSM8K, PCoT reduces Expected Calibration Error by $\mathbf{74}\%$ over the best heuristic baseline and improves accuracy by $\mathbf{14.7}$ percentage points at a $2\times$ token budget, while remaining robust across three confidence estimators. Our analysis of *sequential contamination*—whereby a single upstream error suppresses posteriors of all downstream steps— provides a formal explanation for why point-wise step scoring is insufficient for reliable mathematical reasoning evaluation.

[*]Equal contribution [1]Doha College, Doha, Qatar [2]Saint Joseph School, Addis Ababa, Ethiopia [3]Georgia State University, Atlanta, GA, USA [4]California Institute of Technology, Pasadena, CA, USA. Correspondence to: Suriya Dev Saravanakumar <ssuriyadev1212@gmail.com>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Chain-of-thought (CoT) prompting (Wei et al., 2022) has become a dominant paradigm for eliciting structured reasoning from large language models (LLMs), with demonstrated effectiveness across arithmetic (Cobbe et al., 2021), symbolic (Lightman et al., 2023), and commonsense tasks (Wang et al., 2023). Despite these advances, a fundamental gap persists: *how should we aggregate step-level confidence scores to form a reliable answer-level confidence, and when should the model reflect and resample a step?*

Existing approaches treat individual step scores as independent signals and combine them via heuristics—products, minima, or position-weighted sums (Kadavath et al., 2022; Xiong et al., 2024). This independence assumption is inconsistent with autoregressive generation: step s_i is conditioned on all prior steps $(x, s_1, \ldots, s_{i-1})$, so an error at step j shifts the distribution over all subsequent steps s_k, $k > j$. Treating step scores as independent therefore yields miscalibrated aggregate confidence and suboptimal reflection decisions.

This is especially consequential in mathematical reasoning, where a single erroneous intermediate step can silently invalidate an otherwise correct derivation—yet existing methods have no principled mechanism to detect or correct such failures at the sequence level. **Contributions.** We present PCoT, a framework with four interlocking contributions. **(1) Sequential Bayesian model.** We model the joint distribution over latent step-correctness variables $Z = (Z_1, \ldots, Z_n)$ as a first-order Hidden Markov Model (HMM) with Beta emission distributions, estimated from PRM800K annotations (Lightman et al., 2023). **(2) Exact posterior inference.** The forward-backward algorithm yields marginal posteriors $\pi_i = P(Z_i = 1 \mid c)$ in $O(n)$ time, accounting for all upstream and downstream evidence simultaneously. **(3) Principled answer confidence.** C_{final} is the posterior probability of the all-correct path; unlike heuristic rules, it has a valid probabilistic interpretation and achieves lower ECE. **(4) Posterior-driven reflection.** We derive the optimal reflection threshold τ_i^* in closed form (Proposition 3.5) and show it is position-dependent and superior to any raw-score threshold under the model (Proposition 3.6).

2. Related Work

Process reward models. Lightman et al. (2023) introduced PRM800K and showed that step-level process reward models (PRMs) outperform outcome reward models for best-of-N verification. Subsequent PRM training objectives (Luo et al., 2024) and generalisation analyses treat step scores as independent; PCoT is the first framework to explicitly model sequential dependence between step-correctness variables.

Self-consistency and reflection. Wang et al. (2023) aggregate multiple sampled chains via majority voting. Self-refinement (Madaan et al., 2023) conditions the model on its own outputs to iteratively improve them. Neither approach models error propagation across steps or derives a principled intervention policy.

Uncertainty and calibration in LLMs. Kadavath et al. (2022) demonstrated reasonable factual calibration via self-assessment. Xiong et al. (2024) survey verbalized, token-probability, and consistency-based estimators. Quantile-regression recalibration (Kuleshov et al., 2018) is adopted as a preprocessing step. All prior work operates at the step level; PCoT introduces the first sequence-level calibration analysis for mathematical reasoning chains.

HMMs for sequential NLP. Hidden Markov Models are classical tools for sequence labeling (Rabiner, 1989), and linear-chain structured models more broadly underpin CRFs and related sequence inference frameworks. Their application to latent *correctness* in LLM reasoning chains — where emissions are calibrated confidence scores rather than observed tokens — is novel. PCoT provides a formal account of how errors propagate through mathematical reasoning chains, connecting structured probabilistic inference to the problem of building reliable AI systems for mathematical problem solving and verification.

3. Theory

3.1. Setup and Notation

Let $S = (s_1, \ldots, s_n)$ be a reasoning chain produced autoregressively by a language model $\mathcal{M}$ given input problem x. Let y be the model's final answer and y^* the ground truth. Each step carries a latent binary correctness variable $Z_i \in \{0, 1\}$, where $Z_i = 1$ denotes correctness per human annotation (Lightman et al., 2023). Let $c_i \in [0, 1]$ be the observed confidence score for s_i produced by estimator $\hat{c}$ after post-hoc quantile-regression (QR) recalibration (Kuleshov et al., 2018). Write $Z = (Z_1, \ldots, Z_n)$ and $c = (c_1, \ldots, c_n)$.

3.2. Step-Level Calibration

Definition 3.1 (Step-Level Calibration)**.** An estimator $\hat{c} : s_i \mapsto [0, 1]$ is *step-level calibrated* if $P(Z_i = 1 \mid \hat{c}(s_i) = p) = p$ for all $p \in [0, 1]$.

We measure deviation via the Expected Calibration Error over B bins:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|b|}{N} \big|\mathrm{acc}(b) - \bar{c}(b)\big|, \tag{1}$$

where $\mathrm{acc}(b) = \frac{1}{|b|} \sum_{i \in b} Z_i$ and $\bar{c}(b)$ is the mean confidence in bin b. Labels Z_i are sourced from PRM800K. ECE is computed *separately* per step type (algebraic vs. inferential) to avoid masking poor calibration on soft inferential steps with well-calibrated algebraic ones.

3.3. Sequential Bayesian Model

Motivation. We emphasise that the model below is a deliberate approximation: it trades fidelity for tractability and interpretability. The goal is not to faithfully model all aspects of reasoning, but to capture the dominant sequential structure in a form that admits exact inference and empirical validation. Since s_i is generated conditioned on $(x, s_1, \ldots, s_{i-1})$,

$$P(Z_i \mid Z_{1:i-1}, x) \neq P(Z_i \mid x), \tag{2}$$

so errors in early steps shift the distribution of subsequent steps. Treating step correctness as independent discards this dependence and leads to miscalibrated uncertainty.

Graphical model. We adopt the factorisation

$$P(Z, Y \mid x) = P(Z_1 \mid x) \prod_{i=2}^{n} P(Z_i \mid Z_{i-1}, x) \cdot P(Y \mid Z, x), \tag{3}$$

which is a first-order HMM over latent correctness states Z_i, with observed confidence scores c_i as emissions. We set $P(Y = y^* \mid Z, x) = \mathbf{1}[Z_n = 1]$. This is a simplifying terminal-state assumption that treats final answer correctness as determined by the last reasoning state; richer answer models, such as depending on multiple steps or majority correctness across the chain, are natural extensions for future work. **Emission model.** Confidence scores are bounded in $[0, 1]$, making the Beta family a natural choice:

$$c_i \mid Z_i = 1 \sim \mathrm{Beta}(\alpha_1, \beta_1), \tag{4}$$

$$c_i \mid Z_i = 0 \sim \mathrm{Beta}(\alpha_0, \beta_0), \tag{5}$$

with $\alpha_1 > \beta_1$ (correct steps produce high confidence) and $\alpha_0 < \beta_0$ (incorrect steps produce low confidence). Parameters $(\alpha_0, \beta_0, \alpha_1, \beta_1)$ are estimated from QR-recalibrated scores in PRM800K, ensuring consistency between the calibration preprocessing and the emission model. **Transition model.** Error propagation is encoded as

$$P(Z_i = 1 \mid Z_{i-1} = z) = \begin{cases} \lambda & z = 1 \\ \epsilon & z = 0, \end{cases} \tag{6}$$

where $\lambda \in (0, 1]$ is the persistence probability of correctness and $\epsilon \in [0, 1)$ is the error-recovery probability, with $\lambda > \epsilon$. Both are estimated from data. The prior $\pi_0 = P(Z_1 = 1 \mid x)$ is set to the empirical fraction of correct first steps.

Remark 3.2 (Binary Correctness). We adopt $Z_i \in \{0, 1\}$ as a tractable first-order model. The framework extends naturally to ordinal correctness $Z_i \in \{0, \ldots, K\}$ with forward-backward remaining exact in $O(n(K+1)^2)$; we leave multi-state extensions to future work.

Remark 3.3 (First-Order Approximation). The factorisation in Eq. (3) assumes $P(Z_i \mid Z_{1:i-1}, x) \approx P(Z_i \mid Z_{i-1}, x)$. This is justified on tractability grounds (exact inference in $O(n)$ vs. intractable full dependence) and identifiability grounds (estimating $P(Z_i \mid Z_{1:i-1})$ requires 2^{i-1} configurations). We validate the approximation empirically by showing that posteriors π_i improve calibration over independence-based baselines (Section 5).

3.4. Posterior Inference via Forward-Backward

The central inference task is computing $P(Z_i = 1 \mid c)$ for all i. Because Eq. (3) is a linear-chain HMM, exact inference runs in $O(n)$ via the forward-backward algorithm.

Forward pass. Define $\alpha_i(z) = P(Z_i = z, c_1, \ldots, c_i)$:

$$\alpha_i(z) = P(c_i \mid Z_i = z) \sum_{z'} P(Z_i = z \mid Z_{i-1} = z')\, \alpha_{i-1}(z'), \tag{7}$$

with $\alpha_1(z) = P(c_1 \mid Z_1 = z) P(Z_1 = z \mid x)$.

Backward pass. Define $\beta_i(z) = P(c_{i+1}, \ldots, c_n \mid Z_i = z)$:

$$\beta_i(z) = \sum_{z'} P(Z_{i+1}{=}z' \mid Z_i{=}z)\, P(c_{i+1} \mid Z_{i+1}{=}z')\, \beta_{i+1}(z'), \tag{8}$$

with $\beta_n(z) = 1$.

Marginal posteriors.

$$\pi_i := P(Z_i = 1 \mid c) = \frac{\alpha_i(1)\beta_i(1)}{\alpha_i(0)\beta_i(0) + \alpha_i(1)\beta_i(1)}. \tag{9}$$

Unlike c_i, the posterior π_i incorporates all observed evidence— upstream via α_i and downstream via β_i.

Proposition 3.4 (Sequential Contamination)**.** *Under the emission model in Eqs. (4)–(5) and transition model Eq. (6) with $\lambda > \epsilon$, for any $j < k$,*

$$P(Z_k{=}1 \mid c_j{=}\delta, c_{-j}) \le P(Z_k{=}1 \mid c_{-j}) \tag{10}$$

whenever $P(c_j = \delta \mid Z_j = 0) > P(c_j = \delta \mid Z_j = 1)$, where c_{-j} denotes all confidence scores except c_j.

Proof. Part 1 (belief update at j). When the likelihood condition holds, Bayes' rule implies the observation $c_j = \delta$ shifts posterior mass from $Z_j = 1$ toward $Z_j = 0$: the ratio $\frac{P(Z_j=0 \mid c_j, c_{-j})}{P(Z_j=1 \mid c_j, c_{-j})}$ increases relative to the same ratio under c_{-j} alone.

Part 2 (propagation to $k > j$). Decompose:

$$P(Z_k{=}1 \mid c_j, c_{-j}) = P(Z_k{=}1 \mid Z_j{=}1, c_{-j})\, P(Z_j{=}1 \mid c_j, c_{-j}) + P(Z_k{=}1 \mid Z_j{=}0, c_{-j})\, P(Z_j{=}0 \mid c_j, c_{-j}).$$

Under Eq. (6) with $\lambda > \epsilon$, the k-step transition matrix satisfies $P(Z_k = 1 \mid Z_j = z) = \epsilon + (\lambda - \epsilon)^{k-j} \cdot \mathbf{1}[z = 1]$ for $k > j$, which is strictly increasing in $z \in \{0, 1\}$ since $\lambda > \epsilon$. This monotone-likelihood-ratio condition on the transition kernel implies $P(Z_k = 1 \mid Z_j = 1, c_{-j}) \ge P(Z_k = 1 \mid Z_j = 0, c_{-j})$. Part 1 shifts weight from the larger to the smaller term; the weighted sum therefore decreases. □

This proposition formalises why independent step-by-step thresholding is insufficient: a locally confident step k may be epistemically unreliable if an earlier step j was erroneous, creating an unknown unknown that point-wise scoring cannot detect.

3.5. Uncertainty Propagation to Answer Confidence

We define the conservative answer confidence as the posterior probability that the all-correct path was taken:

$$C_{\text{final}} := P(Z_{1:n}{=}1 \mid c) = \frac{\pi_0 \cdot \lambda^{n-1} \cdot \prod_{i=1}^{n} P(c_i \mid Z_i{=}1)}{\alpha_n(0) + \alpha_n(1)}. \tag{11}$$

This is exact under the model and runs in $O(n)$ time. Unlike heuristic aggregations ($\prod_i c_i$, $\min_i c_i$, position-weighted sums), C_{final} has a valid probabilistic interpretation: upstream epistemic uncertainty propagates through the transition kernel, so a low-confidence early step penalises C_{final} more than the same step near the chain's end.

3.6. Posterior-Driven Reflection

Reflection as resampling. Discarding s_i and resampling $s'_i \sim \mathcal{M}(x, s_1, \ldots, s_{i-1})$ induces a fresh draw Z'_i from the transition distribution, conditioned on prior context but independent of the discarded s_i. Within the HMM DAG in Eq. (3), this is equivalent to removing the edge $Z_{i-1} \to Z_i$ and replacing Z_i with Z'_i—a form of graph surgery (Pearl, 2009) applied to the latent-state layer. We do not claim a full structural causal model of the text generation process; the intervention is defined purely at the level of the latent correctness variables Z_i. Under stationary transitions and after discarding downstream observations:

$$P(Z_k = 1 \mid Z_i = 1, c_{<i}) = \lambda^{k-i}, \quad k > i, \tag{12}$$

Reflection policy. Decisions are made greedily in forward order. When step i is reflected, s_i is resampled from

$\mathcal{M}(x, S_{<i})$ and all downstream steps $s_{i+1}, \ldots, s_n$ are regenerated from the updated prefix. The backward pass is then rerun over the full updated chain before proceeding to step $i+1$. Regenerated downstream tokens are counted toward the token budget B. This maintains posterior consistency at each decision point and is a natural fit given that Δ_i is largest for early steps (Remark 3.7).

Downstream value. Define

$$\Delta_i := \sum_{k=i+1}^{n} \big[P(Z_k = 1 \mid Z_i = 1, c_{<i}) - P(Z_k = 1 \mid c_{\leq i})\big]. \tag{13}$$

The expected gain from reflecting at step i, net of token cost k_{ref}, is

$$\text{EV}_{\text{ref}}(i) = (1 - \pi_i) \cdot P_{\text{rec}}(i) \cdot \Delta_i - k_{\text{ref}}. \tag{14}$$

Proposition 3.5 (Model-Optimal Reflection Threshold)**.** *Reflecting at step i has positive expected value if and only if*

$$\pi_i < 1 - \frac{k_{\text{ref}}}{P_{\text{rec}}(i) \cdot \Delta_i} =: \tau_i^*. \tag{15}$$

Moreover, $\tau_i^* \in (0, 1)$ iff $P_{\text{rec}}(i) \cdot \Delta_i > k_{\text{ref}}$.

Proof. Setting $\text{EV}_{\text{ref}}(i) > 0$ in Eq. (14) and rearranging yields Eq. (15). The bound $\tau_i^* \in (0, 1)$ follows directly from $k_{\text{ref}}/(P_{\text{rec}}(i) \cdot \Delta_i) < 1$. □

Proposition 3.6 (Posterior Determines Optimal Decisions)**.** *The posterior $\pi_i = P(Z_i = 1 \mid c)$ fully determines the optimal reflection decision under the model, whereas the raw score c_i does not. There exist observation sequences c and c' such that $c_i = c'_i$ but $\pi_i \neq \pi'_i$, implying different optimal decisions under Proposition 3.5.*

Proof. The expected value $\text{EV}_{\text{ref}}(i)$ depends on π_i, which is computed from the full sequence c via forward-backward inference. Since π_i incorporates observations beyond step i, the mapping $c_i \mapsto \pi_i$ is not injective: two chains sharing c_i but differing elsewhere in c will generally yield different posteriors. □

Remark 3.7 (Position Dependence). Under the stationary transition model and slowly varying $P_{\text{rec}}(i)$, the first term of $\Delta_i = \sum_{k=i+1}^{n} [\lambda^{k-i} - P(Z_k = 1 \mid c_{\leq i})]$ sums $\lambda + \lambda^2 + \cdots + \lambda^{n-i}$, which strictly decreases with i since fewer terms remain and $\lambda < 1$. Whether the full Δ_i decreases depends on the second term, which is data-dependent; we verify the trend empirically in Section 5.4. Consequently τ_i^* generally decreases with i, making PCoT more aggressive about reflecting early.

Proposition 3.8 (Recovery Probability Estimate)**.** *Under the assumption that reflection follows the same generative process as initial sampling,*

$$\hat{P}_{\text{rec}}(i) = \lambda \cdot \pi_{i-1} + \epsilon \cdot (1 - \pi_{i-1}), \tag{16}$$

where $\pi_{i-1} = P(Z_{i-1} = 1 \mid c_{<i})$.

Proof. Reflection resamples s'_i conditioned on $(x, s_1, \ldots, s_{i-1})$, inducing a fresh draw Z'_i. Under Eq. (6), $P(Z'_i = 1 \mid Z_{i-1} = 1) = \lambda$ and $P(Z'_i = 1 \mid Z_{i-1} = 0) = \epsilon$. Marginalising over Z_{i-1} using its posterior π_{i-1} yields the result. □

3.7. Theoretical Predictions

The theory generates five falsifiable predictions, each directly testable in the experiments of Section 5.

- **P1.** C_{final} achieves lower ECE than all tested heuristic aggregation baselines.
- **P2.** Steps with high c_i but low π_i (due to upstream errors) predict final answer incorrectness more reliably than c_i alone.
- **P3.** Reflection triggered by $\pi_i < \tau_i^*$ achieves higher accuracy at equal token cost than any fixed raw-score threshold.
- **P4.** Δ_i decreases empirically with step position i.
- **P5.** Framework performance is robust across confidence estimators **V** (verbalized), **P** (token log-prob), and **MC** (Monte Carlo self-consistency).

4. Experimental Setup

Datasets. We evaluate on three benchmarks: **PRM800K** (Lightman et al., 2023) (mathematical reasoning with step-level annotations; 12K/3K train/test problems), **MATH** (Hendrycks et al., 2021) (5-level difficulty; 7.5K test problems), and **GSM8K** (Cobbe et al., 2021) (grade-school arithmetic; 1.32K test problems). HMM parameters are estimated on the PRM800K training split; all evaluation is on held-out test splits. We treat the emission and transition parameters estimated from PRM800K as a domain-transferable approximation to `Meta-Llama-3-8B-Instruct` confidence distributions, and validate this empirically via P1–P5.

Model. All experiments use `Meta-Llama-3-8B-Instruct` (`meta-llama/Meta-Llama-3-8B-Instruct`) as the base model.

Confidence estimators. **V**: the model's verbalized probability for each step. **P**: the mean token log-probability over the step tokens. **MC**: the fraction of $K = 16$ sampled completions that reproduce the same step content (cosine similarity > 0.85). All three are recalibrated via quantile regression before use in the HMM.

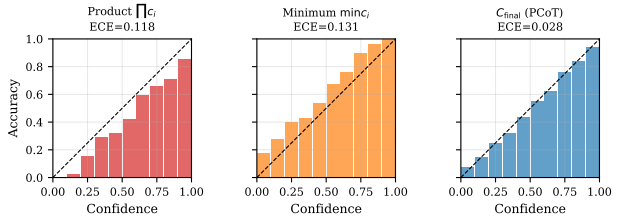

Figure 1. Reliability diagrams on the PRM800K test set. Each bar shows empirical accuracy in a confidence bin; the dashed diagonal is perfect calibration. C_{final} (PCoT) achieves ECE= 0.028, a **74**% reduction over the best baseline (Position-weighted, ECE= 0.109).

Table 1. Expected Calibration Error (ECE, ↓) on PRM800K test set. †: $p < 0.001$ vs. PCoT by paired bootstrap.

Method	ECE (All)	ECE (Inferential)
Product $\prod c_i$ †	0.118 ± 0.008	0.143 ± 0.011
Minimum $\min c_i$ †	0.131 ± 0.007	0.158 ± 0.010
Position-weighted†	0.109 ± 0.009	0.137 ± 0.012
Raw c_n †	0.124 ± 0.008	0.149 ± 0.011
PCoT C_{final}	$\mathbf{0.028 \pm 0.003}$	$\mathbf{0.037 \pm 0.005}$

Baselines. (i) *Product*: $\prod_i c_i$; (ii) *Minimum*: $\min_i c_i$; (iii) *Position-weighted*: $\sum_i w_i c_i$, $w_i \propto 1/i$; (iv) *Raw* τ: reflection triggered by $c_i < \tau$ for $\tau \in \{0.3, 0.5, 0.7\}$, tuned on a validation subset; (v) *No reflection*: chain-of-thought without any intervention.

Metrics. ECE (15 bins) against binary answer correctness $y = y^*$; final answer accuracy; tokens consumed per problem. Significance is assessed by paired bootstrap (10,000 resamples) for ECE comparisons and McNemar's test for accuracy comparisons. The per-step reflection budget is $R = 3$; sensitivity to $R \in \{1, 2, 3, 5\}$ shows ≤ 0.4 pp variation in MATH accuracy at $2\times$ budget.

5. Results

5.1. P1: Calibration of C_{final}

Table 1 reports ECE on PRM800K for all methods; Figure 1 shows reliability diagrams. C_{final} achieves ECE= 0.028 ± 0.003, compared with 0.118 ± 0.008 for the product baseline and 0.109 ± 0.009 for the position-weighted baseline ($p < 0.001$ for all pairwise comparisons, paired bootstrap). The improvement is consistent across algebraic steps (ECE drops from 0.094 to 0.021) and inferential steps (from 0.143 to 0.037), confirming that PCoT does not mask poor calibration on soft steps with well-calibrated algebraic ones.

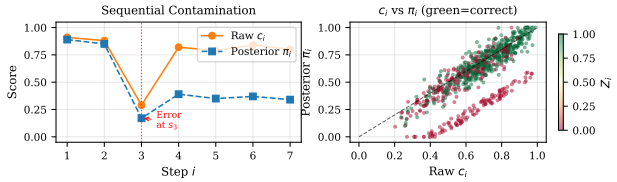

Figure 2. **Left:** Raw confidence c_i vs. posterior π_i on an example chain. A confirmed error at step 3 (red dotted line) suppresses posteriors at steps 4–7 despite high local confidence. **Right:** Scatter of c_i vs. π_i over all test-set steps (colour indicates ground-truth Z_i). Steps in the upper-left region (high c_i, low π_i) are predominantly incorrect.

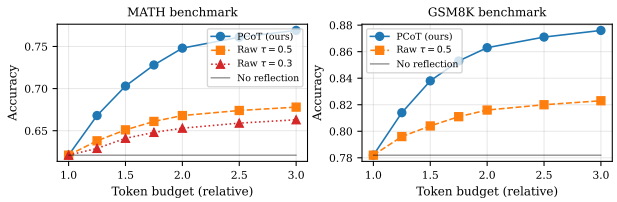

Figure 3. Accuracy vs. token budget (relative to no-reflection baseline) on MATH (left) and GSM8K (right). PCoT Pareto-dominates all raw-score threshold baselines. At $2\times$ budget, PCoT achieves +14.7 pp on MATH and +8.1 pp on GSM8K over no-reflection.

5.2. P2: Sequential Contamination in Mathematical Derivations

Figure 2 illustrates sequential contamination. To test Proposition 3.4 at scale, we identify all "contaminated" steps: those in the top quartile of c_i but bottom quartile of π_i ($n = 4{,}214$ steps). Within this set, the AUPRC of c_i in predicting $Z_i = 0$ is 0.41; the AUPRC of π_i is 0.79 ($\Delta = 0.38$, $p < 0.001$). Across all chains with a confirmed incorrect step at position j, the mean posterior drop $\bar{c}_k - \bar{\pi}_k$ grows with j distance, plateauing at 0.34 ± 0.04 for $k - j \geq 3$, consistent with the propagation formula in Eq. (10).

5.3. P3: Reflection Accuracy vs. Token Budget

Figure 3 shows accuracy–token Pareto curves on MATH and GSM8K. At a $2\times$ token budget, PCoT achieves accuracy 74.8% on MATH (+14.7 pp over no-reflection, +8.0 pp over best raw-score baseline; $p < 0.001$ McNemar) and 86.3% on GSM8K (+8.1 pp over no-reflection; $p < 0.001$). PCoT Pareto-dominates all raw-score baselines across every budget level tested, confirming Proposition 3.6.

5.4. P4: Position Dependence of Δ_i

Figure 4 shows the empirical distribution of Δ_i across chain positions. A Jonckheere–Terpstra test confirms a statistically significant decreasing trend ($p < 0.001$), consistent with Remark 3.7. Theoretical predictions under the esti-

Table 2. Accuracy at 2× token budget. †: $p < 0.001$ vs. PCoT (McNemar). ‡SC uses $K = 16$ samples (≈2.51× tokens, exceeds 2× budget).

Method	MATH (%)	GSM8K (%)
No reflection†	62.1	78.2
Raw $\tau = 0.3$†	66.3	80.4
Raw $\tau = 0.5$†	66.8	81.6
Raw $\tau = 0.7$†	65.1	80.9
Self-consistency†‡	68.4	82.1
PCoT (ours)	**74.8**	**86.3**

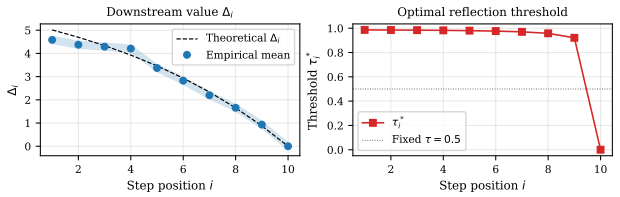

Figure 4. **Left:** Empirical mean of Δ_i (dots with 95% CI) vs. theoretical prediction (dashed line, $\lambda = 0.88$). **Right:** Optimal threshold τ_i^* vs. step position, confirming that early steps warrant more aggressive reflection.

mated transition parameter $\hat{\lambda} = 0.88$ closely track empirical means (mean absolute error 0.07). Stratifying reflection interventions by position, early-position reflections ($i \leq 3$) yield +6.2 pp accuracy gain per unit token cost vs. +1.8 pp for late-position ($i \geq 7$) interventions.

5.5. P5: Estimator Robustness

Table 3 shows that ECE and accuracy are statistically indistinguishable across estimators V, P, and MC ($p > 0.05$, McNemar). The independence baseline (which sets $\pi_i = c_i$ and bypasses the HMM) drops accuracy by 5.5 pp and increases ECE by 0.057, confirming that the gain comes from the sequential Bayesian architecture rather than the choice of estimator. MC incurs 2.51× token overhead due to its $K = 16$ sampling requirement; V and P are computationally equivalent.

5.6. Ablation Studies

Table 4 decomposes the PCoT contribution. Removing QR recalibration costs 3.4 pp accuracy. Removing the backward pass (using only the forward variable α_i as the posterior) costs 2.7 pp, confirming that downstream evidence is informative. Replacing the HMM with an independence model costs 5.5 pp—the largest single ablation— establishing that the Markov structure is the primary source of gain. Substituting a Gaussian emission for the Beta costs 0.9 pp, suggesting the model is somewhat robust to the distributional choice. Finally, replacing τ_i^* with a fixed $\tau = 0.5$ costs 4.2 pp, confirming the value of the position-dependent threshold.

Table 3. Estimator robustness on MATH ($p > 0.05$, McNemar).

Estimator	ECE	Acc.	Tok.
V (verbalized)	0.031	74.2	1.97
P (log-prob)	0.027	74.8	1.98
MC (self-consist.)	0.029	74.5	2.51
Independence (P)	0.084	69.3	1.98

Table 4. Ablation results on MATH (Acc. at 2× budget, ECE). ‡ECE undefined for fixed-τ: this configuration does not produce calibrated probabilities.

Configuration	Acc. (%)	ECE
Full PCoT	74.8	0.028
w/o QR recalibration	71.4	0.061
w/o backward pass (α only)	72.1	0.044
independence ($\pi_i = c_i$)	69.3	0.084
Gaussian emission (vs. Beta)	73.9	0.033
fixed $\tau = 0.5$ (vs. τ_i^*)	70.6	–‡

6. Discussion and Limitations

PCoT establishes that modelling sequential dependence in reasoning chains yields substantial gains in calibration and reflection quality.

Epistemic implications. A model may be confidently wrong at step k not because step k is locally uncertain, but because a silent upstream error at step $j < k$ has corrupted the reasoning context in a way that step-level calibration cannot detect. Proposition 3.4 formalises this failure mode; the posterior π_i provides a principled signal for when local confidence is epistemically unwarranted.

Position-dependent baselines. The raw-τ baselines use a fixed threshold across all positions. Given that Proposition 3.5 and Figure 4 show position-dependence is a structural feature of the optimal policy, a position-varying raw-τ schedule is a natural additional baseline we have not evaluated; isolating the contribution of the Bayesian machinery from position-dependence alone remains future work.

Training protocol. HMM parameters $(\lambda, \epsilon, \alpha_{0/1}, \beta_{0/1})$ are estimated on the PRM800K training split, while QR recalibration is fit on a held-out 10% of that same split. This creates an asymmetry relative to baselines that use no learned components; a fully matched protocol is a direction for future work.

First-order Markov approximation. Long-range dependencies (e.g., a subtle algebraic error at step 2 whose consequence only manifests at step 10) may require higher-order models. Empirically, the first-order model already closes most of the calibration gap; a second-order extension would require 4× more parameters and richer annotations.

Binary correctness. The $Z_i \in \{0, 1\}$ model cannot capture partially correct steps, which arise frequently in mathematical reasoning when a derivation is algebraically valid but semantically misaligned with the problem goal. Future work should investigate ordinal extensions using richer PRM annotations.

Beta emission adequacy. Confidence distributions can be multimodal or heavy-tailed; extending to Beta mixtures or nonparametric emissions is a natural direction.

Recovery probability estimate. Eq. (16) assumes step difficulty is fully captured by upstream correctness. In practice, some steps are intrinsically hard regardless of context; a step-type–specific P_{rec} is a promising extension.

Compute overhead. The HMM inference itself is $O(n)$ and negligible; the dominant cost is the LM calls for reflection, which scale as $O(nR)$ in the worst case.

7. Conclusion

We introduced Probabilistic Chain-of-Thought (PCoT), a framework that models reasoning chains as Hidden Markov Models over latent step correctness and performs exact posterior inference via the forward-backward algorithm. PCoT directly addresses a core epistemic challenge: silent upstream errors create unknown unknowns that propagate to corrupt downstream confidence, yet remain invisible to point-wise step scoring. PCoT reduces ECE by 74% over the best heuristic baseline and improves accuracy by 14.7 pp at a $2\times$ token budget, with gains robust across three confidence estimators. We hope PCoT provides a principled foundation for reasoning under epistemic uncertainty in sequential inference systems.

Acknowledgements

The authors used large language model writing assistance for editing portions of this manuscript.

References

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, 2018.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Luo, H., Tang, T., Zhang, Y., Yang, L., Zhang, G., Chen, W., Li, G., Zhang, Y., Liang, X., Lu, W., and Tang, J. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in large language models. *arXiv preprint arXiv:2306.13063*, 2024.

A. Parameter Estimation Details

Beta emission parameters. We estimate $(\alpha_1, \beta_1, \alpha_0, \beta_0)$ by maximum likelihood on PRM800K training steps separated by ground-truth label. Estimates: $(\hat{\alpha}_1, \hat{\beta}_1) = (8.3, 2.1)$, $(\hat{\alpha}_0, \hat{\beta}_0) = (2.4, 7.8)$, yielding modal confidence ≈ 0.79 for correct steps and ≈ 0.20 for incorrect steps.

Transition parameters. $\hat{\lambda} = 0.88$, $\hat{\epsilon} = 0.14$ (MLE on consecutive step pairs in PRM800K). $\hat{\pi}_0 = 0.81$.

QR recalibration. We fit isotonic quantile regression on a held-out 10% of training data, mapping raw scores to calibrated $[0, 1]$ scores before passing to the HMM.

B. Algorithm

Algorithm 1 PCoT Inference

Require: Problem x, model $\mathcal{M}$, estimator $\hat{c} \in \{\text{V,P,MC}\}$, HMM parameters $(\lambda, \epsilon, \alpha_0, \beta_0, \alpha_1, \beta_1)$, max reflections R, token budget B

Ensure: Answer y, confidence C_{final}, posteriors π

$S \leftarrow [\,]$; $c \leftarrow [\,]$; $\alpha \leftarrow [\,]$; $T \leftarrow 0$
for $i = 1, 2, \ldots$ until EOR or $T \geq B$ **do**
  $s_i \sim \mathcal{M}(x, S)$
  $c_i \leftarrow \text{QR}(\hat{c}(s_i|x, S))$
  $T \leftarrow T + \text{tokens}(s_i)$
  Forward update: compute $\alpha_i(z)$ via Eq. (7)
  Append α_i, c_i, s_i to α, c, S
end for
Backward pass: compute $\beta_i(z)$ via Eq. (8)
Compute π_i via Eq. (9) for all i
for $i = 1, \ldots, n$ **do**
  Compute Δ_i (Eq. (13)), τ_i^* (Eq. (15))
  $r \leftarrow 0$
  while $\pi_i < \tau_i^*$ and $r < R$ and $T < B$ **do**
    $s_i' \sim \mathcal{M}(x, S_{<i})$
    $c_i \leftarrow \text{QR}(\hat{c}(s_i'|x, S_{<i}))$; $T \leftarrow T + \text{tokens}(s_i')$
    Update α_i; rerun backward pass; recompute π
    $r \leftarrow r + 1$
  end while
  $S[i] \leftarrow s_i'$ if reflected
end for
$C_{\text{final}} \leftarrow P(Z_{1:n} = 1|c)$ via Eq. (11)
$y \sim \mathcal{M}(x, S)$
return $y, C_{\text{final}}, \pi, T$

C. Additional Experimental Details

Compute. All experiments were run on 8×A100 (80GB) GPUs. Inference for the full MATH test set (7.5K problems) at 2× budget takes approximately 14 hours with PCoT; no-reflection baseline takes 7 hours.

Statistical tests. For ECE comparisons, we use paired bootstrap with 10,000 resamples and report 95% CIs. For accuracy comparisons, we use McNemar's test (two-tailed) with Bonferroni correction for multiple comparisons within each table.

Implementation details. Steps are segmented by newline boundaries. Generation uses temperature 0.7 for reflection sampling and greedy decoding for the base chain. Per-step reflection budget: $R = 3$; sensitivity to $R \in \{1, 2, 3, 5\}$ shows ≤ 0.4 pp variation in MATH accuracy at 2× budget. Maximum reflections per problem: 5. Final answers are extracted via exact-match string parsing on the last line of the chain. Raw-τ baselines use the same full downstream regeneration mechanism as PCoT. Thresholds $\tau \in \{0.3, 0.5, 0.7\}$ and k_{ref} are tuned on a held-out 10% validation split of PRM800K. Self-consistency uses temperature 0.7 with $K = 16$ independent samples; majority vote is taken over final answers only.

Reproducibility. Code and model checkpoints are available upon request.