Reasoning as a Dynamical System: Online Monitoring of LLMs with Particle Filters

Mohammed Abbas [1]

Abstract

Online monitoring of LLM reasoning trajectories is increasingly important for inference efficiency, but existing approaches treat steps independently and provide uncalibrated confidence estimates. We model reasoning as a Switching Linear Dynamical System (SLDS) with three latent modes (NORMAL, INSIGHT, BACKTRACK) and track it online with a Rao-Blackwellized Particle Filter (RBPF) that marginalises the continuous state analytically and requires no ground-truth labels at test time. Using per-step scores from a Process Reward Model (PRM) as observations, RBPF reaches an AUC of 0.850 on GSM8K, comparable to strong online and offline baselines. Its main advantage is calibration: RBPF attains the lowest calibration error among all methods (ECE 0.111), around three times lower than the strongest ranking baselines (MeanPRM 0.305, EMA 0.311), making its confidence estimates more directly usable downstream. For early stopping, RBPF commits a correctness verdict 12% of steps early on average, without reducing verdict accuracy. On the harder MATH, the pattern holds: RBPF is competitive on discrimination (AUPR 0.607, the highest of all methods) and far better calibrated than the ranking baselines.

1. Introduction

As LLMs perform increasingly complex multi-step reasoning, monitoring their progress online matters for both inference efficiency and safety (Zhang, 2025). A model generates text step by step, but its trajectory toward a correct solution is a hidden latent variable with non-linear and multimodal dynamics: the model may make steady progress for several steps, identify an insight, then backtrack to correct an error. Prior work treats the latent structure of reasoning either as a representation to improve generation quality (Hao et al., 2025; Liu et al., 2025; Saunshi et al., 2025) or as per-step uncertainty signals for adaptive computation (Shi et al., 2025; Guo et al., 2025b). Neither frames it as a *dynamical estimation problem*. Process Reward Models (PRMs) score individual steps but are used for selection, not online state estimation that tracks how confidence evolves across a chain (Zhang et al., 2025).

We contribute: **(1)** a three-mode SLDS that imposes a structured prior over reasoning dynamics, with modes distinguished by mode-specific drift; **(2)** an RBPF (Doucet et al., 2000; Schön et al., 2005) that uses PRM step scores as online observations, enabling label-free monitoring; **(3)** a calibration analysis on GSM8K showing RBPF reaches the lowest calibration error of all methods (ECE 0.111 at the final step), roughly $3\times$ lower than strong ranking baselines such as MeanPRM (0.305); **(4)** an early-stopping study showing that RBPF's calibrated confidence enables modestly early correctness verdicts with no loss in verdict accuracy relative to final-step prediction; **(5)** a cross-dataset evaluation showing that the GSM8K-fitted readout transfers to MATH without refitting, where RBPF attains the best AUPR and stays far better calibrated than the ranking baselines.

2. Related Work

Latent reasoning in LLMs. Coconut (Hao et al., 2025) trains LLMs to iterate over continuous thought embeddings; deliberation (Liu et al., 2025) adds a latent coprocessor to a frozen LLM; looped transformers (Saunshi et al., 2025) give theoretical grounding for iterative latent computation. All three treat latent reasoning as representation engineering. Our goal is different: given a running LLM, estimate its hidden reasoning state *online* without modifying the model.

Process Reward Models. PRMs assign per-step quality scores to reasoning chains (Zhang et al., 2025) and are commonly used in offline selection pipelines, such as best-of-N reranking, or as reward signals for reinforcement learning. We repurpose them as *observations* in a dynamical system, letting the filter reason about how step-level quality signals evolve across the trajectory instead of aggregating quality

[1]KTH Royal Institute of Technology, Stockholm, Sweden. Correspondence to: Mohammed Abbas <mabba@kth.se>.

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).

only after generation.

Uncertainty in LLM reasoning. Semantic entropy (Nguyen et al., 2025), token-level uncertainty (Guo et al., 2025b), and topological uncertainty (Da et al., 2025) characterise per-step confidence, but treat steps independently and ignore how uncertainty couples with progress across a trajectory.

Switching dynamical systems. SLDS-based models capture discrete behavioural regimes in neural population dynamics (Linderman et al., 2017; Hu et al., 2024). Differentiable particle filters jointly learn transition and proposal distributions (Cox et al., 2025; Younis & Sudderth, 2024). We apply RBPF-based SLDS inference to LLM reasoning, targeting discrete mode switches as the main estimation objective.

3. Method

3.1. Latent State and Dynamics

The reasoning state at step k is $\mathbf{x}_k = [p, c, u, f, m]^\top \in \mathbb{R}^5$, encoding progress ($p$), coherence ($c$), uncertainty ($u$), fatigue ($f$), and momentum ($m$). A discrete mode $z_k \in \{1, 2, 3\}$ selects among three transition dynamics:

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{f}_{\text{drift}}(\mathbf{x}_{k-1}) + \mathbf{\Delta}_{z_k} + \mathbf{w}_k, \tag{1}$$

with $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, Q)$ and z_k evolving via a Markov chain with transition matrix P. The drift couples the continuous coordinates:

$$\mathbf{f}_{\text{drift}}(\mathbf{x}) = \begin{bmatrix} 0 \\ \gamma(1-u) - \delta c \\ \eta(1-c) - 0.1f \\ 0.02 + 0.05u \\ 0 \end{bmatrix}, \tag{2}$$

with $\gamma = 0.1$, $\delta = \eta = 0.05$; progress and momentum have no autonomous drift and evolve only through the mode increment and observations. Modes share a single diagonal process noise Q (fixed, not learned) and are distinguished by their drift increment $\mathbf{\Delta}_{z_k}$: $\mathbf{\Delta}_{\text{NORMAL}} = \mathbf{0}$, $\mathbf{\Delta}_{\text{INSIGHT}} = [0.25, 0, 0, 0, 0]^\top$, $\mathbf{\Delta}_{\text{BACKTRACK}} = [-0.05, -0.05, 0.05, 0.05, 0]^\top$. Section 4 ablates a rotation-augmented variant in which a mode-specific rotation R_{z_k} is applied before $\mathbf{\Delta}_{z_k}$. Diagonal entries of P are $0.96/0.90/0.90$ for NORMAL / INSIGHT / BACKTRACK.

3.2. Observation Models

We evaluate two observation modes.

Logit features. Three features are extracted from model logits without ground-truth labels: *token entropy* (normalised mean entropy of the next-token distribution, proxying uncertainty u); *step perplexity* (normalised mean NLL of generated tokens, proxying fatigue f); and *answer consistency* (whether the running answer matches the previous step, proxying coherence c). These form a three-dimensional observation $h_{\text{logit}}(\mathbf{x}) = [\, c - 0.5u, \quad 0.8c + 0.3u + 0.1\sin(4\pi c), \quad f - 0.2m + 0.1uc\,]^\top$; unlike the PRM setting, the progress coordinate is not observed directly.

PRM observations. We also evaluate per-step scores from Qwen2.5-Math-PRM-7B (Zhang et al., 2025) as a scalar observation $y_k \in [0, 1]$, related to the latent state by the linear emission $h(\mathbf{x}) = [\mathbf{x}]_1$, so the PRM score is read as a noisy measurement of the progress coordinate. PRM scores provide a direct signal of step-level reasoning quality, complementing the structural dynamics of the SLDS.

Both modes use a robust mixture observation likelihood: with probability ϵ an observation is drawn from a heavy-tailed component, so occasional bad step scores do not dominate the filter update. For logit features we use ϵ=0.05 and σ^2_{out}=20; for PRM scores we use ϵ=0.02 and σ^2_{out}=10, reflecting that PRM outputs are bounded and typically less noisy than hand-crafted logit features. We do not tune ϵ or σ_{out} on the report split. Observation noise standard deviation is chosen by grid search on the held selection split (20% of trajectories; hyperparameter selection only).

3.3. Rao-Blackwellized Particle Filter

Each particle i carries weight $w_k^{(i)}$, mode $z_k^{(i)}$, and Gaussian belief $(\boldsymbol{\mu}_k^{(i)}, \Sigma_k^{(i)})$. Let o_k denote the observation at step k, either the logit-feature vector or the scalar PRM score. At each step: **(1)** sample $z_k^{(i)} \sim P(z_k | z_{k-1}^{(i)})$; **(2)** predict via EKF using mode-$z_k^{(i)}$ dynamics; **(3)** update weight $w_k^{(i)} \propto w_{k-1}^{(i)} \cdot p(o_k \mid z_k^{(i)}, \boldsymbol{\mu}_{k|k-1}^{(i)}, \Sigma_{k|k-1}^{(i)})$; **(4)** resample when effective sample size falls below $N/2$. Marginalising the continuous state analytically cuts weight variance versus a standard PF at equal N (Doucet et al., 2000). We form the particle-weighted mean state $\bar{\boldsymbol{\mu}}_k = \sum_i w_k^{(i)} \boldsymbol{\mu}_k^{(i)}$ and read out correctness via a logistic map:

$$\hat{p}_k = \sigma\big(\mathbf{a}^\top \bar{\boldsymbol{\mu}}_k + b\big), \tag{3}$$

where σ is the logistic function. The coefficients $(\mathbf{a}, b)$ are fit by ℓ_2-regularised logistic regression on the GSM8K held selection split: at each step we record $\bar{\boldsymbol{\mu}}_k$ from the primary model and label it by whether the chain is ultimately correct. The fitted readout is frozen and applied unchanged at every step on both datasets; PF and EKF use the same map on their state estimate $\hat{\mathbf{x}}_k$.

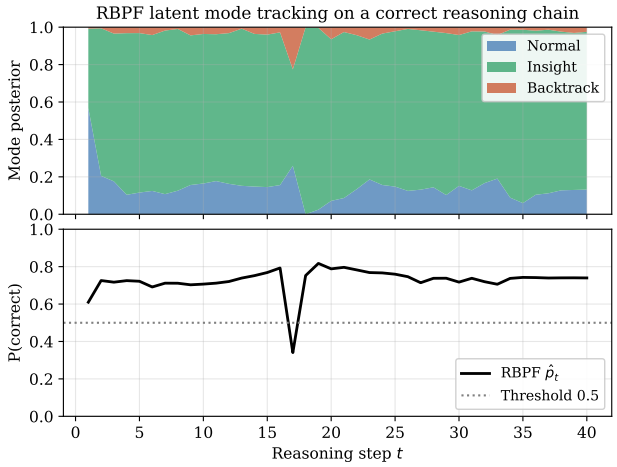

Figure 1. RBPF tracking a correct reasoning chain (T=40). *Top:* mode posterior over NORMAL (blue), INSIGHT (green), and BACKTRACK (red) at each step; the chain moves from NORMAL into a predominantly INSIGHT regime. *Bottom:* estimated correctness probability $\hat{p}_t$, which rises early and stays high apart from a brief drop near step 17, where a transient BACKTRACK spike lowers $\hat{p}_t$ before it recovers.

4. Experiments

4.1. Setup

Datasets. We run DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025a) on the full GSM8K (Cobbe et al., 2021) test set (1,319 problems, 64.7% accuracy) and 500 problems from MATH (Hendrycks et al., 2021) (39.6% accuracy). We extract per-step logit features and PRM scores at each reasoning step. We split GSM8K trajectories into a 20% selection split and an 80% reporting split (stratified by correctness, seed 42). On the selection split we select the observation noise scale, the EMA smoothing factor, and the early-stopping threshold τ, and we fit the logistic readout. We report GSM8K metrics on the reporting split only. For MATH we apply the GSM8K-fitted readout unchanged and report on MATH's own reporting split; the observation noise scale and baseline smoothing factors are re-selected on MATH, since these are nuisance scales rather than the learned readout that carries our calibration claim.

Baselines. We compare against a standard Particle Filter (PF), an Extended Kalman Filter (EKF), logistic regression on the first k steps (LR-k5) and on the full trace (LR-full), MeanPRM (average PRM score over all steps), and an exponential moving average (EMA) of the PRM scores with smoothing factor selected on held-out data. LR-full and MeanPRM are *offline*: they need the complete trace before predicting, and cannot support early stopping. We report results over 5 random seeds.

Table 1. Key filter hyperparameters for the primary RBPF model (PRM observations).

Parameter	Value
State dimension	5
Number of modes	3
Self-transition prob. P_{ii}	0.96/0.90/0.90
Particles N	125
Resampling threshold	$N/2$
Outlier prob. ϵ	0.02
Outlier scale σ^2_{out}	10

Table 2. GSM8K results (mean ± std, 5 seeds). † denotes offline methods requiring the full trace. ECE is reported for PRM observations, where calibrated confidence is the primary claim.

Method	Obs.	AUC	AUPR	ECE
RBPF	Logit	0.674±0.002	0.796±0.003	—
EKF	Logit	0.475±0.000	0.608±0.000	—
PF	Logit	0.752±0.006	0.849±0.007	—
LR-full†	Logit	0.825±0.000	0.889±0.000	—
RBPF	PRM	0.850±0.004	0.895±0.003	**0.111**
EKF	PRM	0.841±0.000	0.889±0.000	0.115
PF	PRM	0.835±0.005	0.883±0.006	0.115
LR-k5	PRM	0.704±0.000	0.766±0.000	—
MeanPRM†	PRM	0.884±0.000	0.916±0.000	0.305
LR-full†	PRM	0.876±0.000	0.914±0.000	—
EMA (α=0.10)	PRM	0.905±0.000	0.939±0.000	0.311

4.2. Main Results: GSM8K

Table 2 reports AUC-ROC at the final reasoning step on GSM8K under both observation modes. With logit features, RBPF reaches 0.674, well above EKF (0.475) but below PF (0.752). Replacing logit features with PRM observations raises RBPF's AUC to 0.850, a +0.176 gain, confirming that step-level reward signals carry more information than hand-crafted logit proxies. Under PRM observations, the filters are close in discrimination (RBPF 0.850, EKF 0.841, PF 0.835), while the ranking baselines score slightly higher still (MeanPRM 0.884, EMA 0.905). Thus, discrimination is not the hard part; calibration is. RBPF achieves the lowest ECE of all methods (0.111), closely followed by the other filters (0.115) but roughly three times better than the ranking baselines (MeanPRM 0.305, EMA 0.311), which produce confident but unreliable probabilities.

4.3. Calibration

A calibrated monitor should have its predicted probability $\hat{p}_k$ match the true fraction of correct solutions, so that confidence can be acted on directly. Figure 3 shows reliability diagrams at an early step (k=5) and at the final step: the filtering methods stay close to the diagonal, while MeanPRM is severely overconfident. The advantage is sharpest early. At k=5, RBPF reaches ECE 0.025, compared with

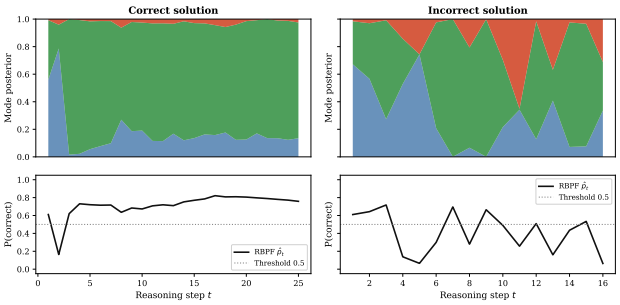

Figure 2. RBPF on example correct (left) and incorrect (right) reasoning chains from GSM8K with PRM observations. On the correct chain the mode posterior settles into a stable INSIGHT regime and $\hat{p}_t$ stays high after an early dip. On the incorrect chain the posterior shows recurring BACKTRACK mass and $\hat{p}_t$ stays volatile, repeatedly crossing the decision threshold.

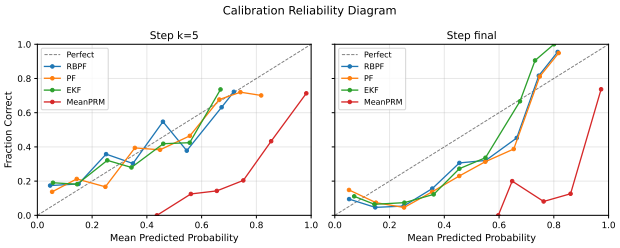

Figure 3. Calibration reliability diagrams at step k=5 (left) and the final step (right). The filtering methods (RBPF, PF, EKF) track the diagonal closely, while MeanPRM (red) is severely overconfident.

0.072 for EKF and 0.302 for MeanPRM, making its confidence usable well before a chain completes. This matters for downstream decisions such as routing hard problems to more capable models, where a calibrated early signal can save compute.

4.4. Early Stopping

A calibrated monitor can decide whether a chain is correct before it finishes, trading compute for an earlier answer. We test this on saved probability traces, not by stopping the model as it runs. At each step the chain stops once $\hat{p}_k$ leaves the band $[\tau, 1-\tau]$: we predict 1 if $\hat{p}_k > 1-\tau$ and 0 if $\hat{p}_k < \tau$. If a chain never leaves the band, we label it by its last step. The full-run baseline is the case τ=0, where the monitor never stops and predicts from the final step. Accuracy measures how often this early verdict matches the true label, not whether a partial answer parses correctly; compute saved is the mean fraction of steps skipped. We pick τ on the held split to maximise RBPF accuracy, break ties toward more compute saved, and apply the same τ on the eval split. Figure 4 plots accuracy against compute saved as τ varies. At the chosen threshold (τ=0.15), RBPF stops on average 12% of steps early and matches the true label 81% of the time, against 76% for the full run. RBPF

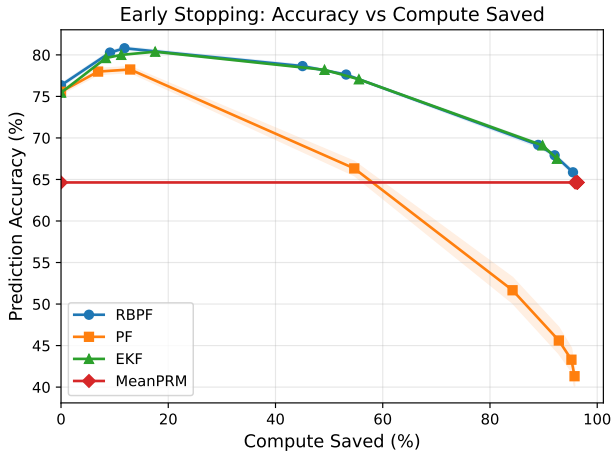

Figure 4. Prediction accuracy versus compute saved under early stopping. RBPF (blue) and EKF (green) trace nearly identical curves, staying at or above their no-stop accuracy (τ=0, left edge) under modest compute savings, then declining as more compute is saved; PF (orange) drops faster. The MeanPRM line here is an online running prefix mean of PRM scores (distinct from the offline full-trace MeanPRM in Table 2); under early stopping it stays near the majority-class base rate ($\approx$65%).

and EKF trace nearly identical curves, both at or above the baseline at low savings and falling off as τ grows; PF drops faster. Early stopping is thus a benefit of calibrated online filtering, not of RBPF in particular, and it pays off at modest savings, not at the aggressive thresholds where RBPF and EKF sink toward the majority-class floor and PF drops below it.

4.5. Ablation Study

Table 3 ablates the design of the primary model on real PRM data. Adding mode-specific rotation (RBPF-Rot, 0.764) lowers AUC relative to our no-rotation model (0.848): the rotation that helps on simulated data with planted geometric structure does not transfer to real PRM observations. Collapsing to two modes (RBPF-2Mode, 0.663) hurts most, confirming that the BACKTRACK mode carries discriminative signal. Learning the transition matrix and process noise by EM (RBPF-Fitted, 0.825) also does not beat the hand-specified prior (0.848). With only 264 held-out trajectories and an unobserved mode sequence, fitting the dynamics is under-constrained and can drift toward degenerate, near-absorbing modes. We therefore retain the hand-specified structured prior and leave differentiable filtering on larger datasets to future work (Cox et al., 2025).

4.6. Cross-Dataset Generalisation: MATH

Table 4 reports results on MATH using PRM observations, with the readout fit on GSM8K applied unchanged. MATH is harder than GSM8K (39.6% vs. 64.7% accuracy) and

Table 3. PRM ablation on GSM8K (mean ± std, 5 seeds). RBPF is our primary model (three modes, no rotation, hand-specified *dynamics*); each variant changes one design choice.

Method	AUC
RBPF	$\mathbf{0.848 \pm 0.003}$
RBPF-Rot	0.764 ± 0.004
RBPF-2Mode	0.663 ± 0.003
RBPF-Fitted	0.825 ± 0.001

Table 4. MATH results with PRM observations (mean ± std, 5 seeds). † offline methods. Bold indicates the best value, or values effectively tied at the reported precision.

Method	AUC	AUPR	ECE
RBPF	$\mathbf{0.689 \pm 0.003}$	$\mathbf{0.607 \pm 0.005}$	$\mathbf{0.245}$
EKF	0.682 ± 0.000	0.589 ± 0.000	0.256
PF	0.650 ± 0.011	0.555 ± 0.019	$\mathbf{0.244}$
LR-k5	0.513 ± 0.000	0.407 ± 0.000	—
MeanPRM†	0.598 ± 0.000	0.435 ± 0.000	0.575
LR-full†	0.599 ± 0.000	0.433 ± 0.000	—
EMA (α=0.30)	$\mathbf{0.688 \pm 0.000}$	0.582 ± 0.000	0.557

more class-imbalanced, making AUPR the more informative discrimination metric. RBPF attains the best AUPR (0.607 vs. 0.589 for EKF, 0.555 for PF, and 0.582 for EMA) and is effectively tied for the best calibration (ECE 0.245 vs. 0.244 for PF). On AUC, the filters are close (RBPF 0.689, EKF 0.682, PF 0.650), and EMA is competitive (0.688), but the ranking baselines remain poorly calibrated (EMA 0.557, MeanPRM 0.575). Thus, a readout fit only on GSM8K still transfers to a harder distribution, staying better calibrated than the ranking baselines while remaining discriminative.

5. Discussion

Online vs. offline. The central contribution of this work is an *online* monitoring system, and the distinction matters when interpreting the results. LR-full and MeanPRM observe the full reasoning chain before issuing a single prediction; they are retrospective evaluators. RBPF instead updates its belief after each step and can therefore support streaming inference, early stopping, and compute routing. Comparing final-step AUC directly hides this asymmetry: offline methods have access to strictly more information. The relevant deployment question is not only which method scores higher on a finished trace, but what can be inferred reliably at step k given only $y_{1:k}$. In this online setting, RBPF remains competitive in discrimination while providing substantially better calibrated confidence estimates than strong PRM-ranking baselines.

Why calibration matters. In deployment, a confidence score is only useful if its numerical value is meaningful. A system claiming "90% confident" should be correct on roughly 90% of such cases, or downstream decisions will be systematically unreliable. On GSM8K, MeanPRM and EMA achieve high AUC yet have ECE above 0.30 (0.305 and 0.311), so their confidence scores require external recalibration before use. RBPF's ECE of 0.111 makes its posterior $\hat{p}_k$ more directly usable for downstream decisions, especially early stopping and routing. This calibration advantage is central: these applications depend on the meaning of the probability estimate, not only on ranking performance.

PRM scores as dynamic observations. PRMs are usually used to evaluate reasoning chains for selection. We use them differently: each per-step score is treated as a sequential sensor reading in a dynamical system and folded into a running belief update. This lets the filter track how step quality evolves across a chain, rather than only summarising quality after generation. The $+0.176$ AUC gain from adding PRM observations shows that per-step reward signals provide useful online evidence for latent reasoning-state estimation.

Structural prior over reasoning modes. The SLDS encodes an inductive bias: reasoning trajectories can be approximated as evolving through a small number of persistent latent regimes distinguished by their drift. Even with manually set parameters, this prior yields better calibration than the baselines we evaluate. The ablation shows that adding mode-specific rotation underperforms our no-rotation design on real PRM data, and that learning the dynamics by EM (RBPF-Fitted) does not beat the hand-set prior at this data scale. The results suggest that mode structure captures useful regularities in LLM reasoning trajectories; this points to differentiable filtering on larger corpora, rather than EM on small held-out sets, as the route to learned dynamics.

Computational overhead. RBPF runs N=125 particles with a per-particle EKF update at each reasoning step. In our CPU NumPy implementation, the filtering update takes about 16–20 ms per step, with the logistic readout adding less than 0.1 ms. This overhead is small compared with the cost of producing each observation, which requires LLM decoding and, for PRM observations, a 7B PRM forward pass on the growing prefix. Thus, the practical bottleneck is obtaining the reasoning steps and PRM scores, not the filtering update itself.

6. Conclusion

We presented an RBPF framework for online tracking of latent reasoning trajectories in LLMs, using PRM step scores as observations within a switching dynamical system. The key findings are: (i) PRM observations raise online AUC over logit features by $+0.176$ on GSM8K, with RBPF comparable to the online EKF baseline; (ii) RBPF gives the best-calibrated confidence estimates of any method tested

(ECE 0.111, roughly 3× better than MeanPRM), more directly usable downstream; (iii) this calibration supports early stopping by letting the monitor commit to a correctness verdict modestly early, with no loss in verdict accuracy relative to final-step prediction; (iv) RBPF attains the best AUPR on MATH (0.607), with the GSM8K-fitted readout transferred unchanged.

Limitations. The three modes are an imposed structured prior rather than validated reasoning states: we do not show that NORMAL, INSIGHT, and BACKTRACK correspond to annotated reasoning behaviours. The transition matrix, mode increments, and covariance are hand-set rather than learned from real reasoning data. The observation model treats PRM scores as conditionally independent given the latent state; modelling temporal correlations in PRM outputs could further improve filtering accuracy.

Future work. Learned transition dynamics. Our EM-based attempt to learn the dynamics (RBPF-Fitted) did not improve over the hand-set prior, which we attribute to the difficulty of identifying a multi-mode latent process from few, sparsely informative trajectories. A more promising route is differentiable particle filtering (Cox et al., 2025), which can learn transition and proposal distributions end-to-end and is better suited to larger trajectory corpora than batch EM. Related differentiable smoothers (Younis & Sudderth, 2024) extend this direction to long sequences, relevant for the longer-horizon tasks discussed below. Different model families, such as chain-of-thought, process-of-thought, and scratchpad-style reasoners, may exhibit distinct mode geometries that a learned filter could recover automatically.

Adaptive best-of-N with online termination. Standard best-of-N waits for all K chains to finish. RBPF's per-step confidence estimates enable a more efficient protocol: chains whose confidence falls below τ_{low} are terminated early, while chains above τ_{high} are committed immediately. This could reduce expected compute per selection query while preserving selection quality, combining online filtering with multi-chain diversity.

Compute routing. RBPF confidence at step k gives a real-time difficulty signal. Problems where confidence stays low after the first few steps can be routed to a larger model or a larger reasoning budget; high-confidence problems are resolved cheaply. RBPF's calibration (ECE 0.111 on GSM8K) makes this routing decision more reliable than using uncalibrated PRM-ranking scores directly, which is what makes the approach useful for deployment.

Longer reasoning tasks. GSM8K and MATH chains are typically short relative to theorem proving, multi-step code synthesis, and scientific reasoning, where reasoning traces can be much longer and offline methods become more expensive. Extending RBPF to these settings may require hierarchical mode structures that reflect proof strategy shifts and tactic-level PRM scores.

RBPF as a process reward signal. The per-step mode posterior $P(z_k \mid y_{1:k})$ provides a label-free signal of whether the trajectory is entering a BACKTRACK-like regime. This signal could serve as a dense process reward for reinforcement learning, penalising excessive backtracking and rewarding insight-driven trajectories, without manual step-level supervision. Closing the loop between RBPF inference and model training is an open direction for improving reasoning efficiency at scale.

References

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

Cox, B., Segarra, S., and Elvira, V. Learning state and proposal dynamics in state-space models using differentiable particle filters and neural networks. *Signal Processing*, 234:109998, 2025. ISSN 0165-1684. doi: 10.1016/j.sigpro.2025.109998.

Da, L., Liu, X., Dai, J., Cheng, L., Wang, Y., and Wei, H. Understanding the uncertainty of LLM explanations: A perspective based on reasoning topology. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=p4wZfBFgyI.

Doucet, A., De Freitas, N., Murphy, K., and Russell, S. Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 176–183, 2000.

Guo, D., Yang, D., Zhang, H., et al. Deepseek-r1: Incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025a. doi: 10.1038/s41586-025-09422-z. Open access under CC-BY 4.0.

Guo, P.-F., Tsai, Y.-D., and Lin, S.-D. Benchmarking uncertainty metrics for LLM target-aware search. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4230–4238. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-emnlp.226.

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J. E., and Tian, Y. Training large language model to reason in a continuous latent space, 2025. URL https://openreview.net/forum?id=tG4SgayTtk.

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. *NeurIPS*, 2021.

Hu, A., Zoltowski, D. M., Nair, A., Anderson, D., Duncker, L., and Linderman, S. Modeling latent neural dynamics with gaussian process switching linear dynamical systems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LX1lwP90kt.

Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. Bayesian learning and inference in recurrent switching linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Liu, L., Pfeiffer, J., Wu, J., Xie, J., and Szlam, A. Deliberation in latent space via differentiable cache augmentation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IaUJl5RCOu.

Nguyen, D., Payani, A., and Mirzasoleiman, B. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4530–4540. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.234.

Saunshi, N., Dikkala, N., Li, Z., Kumar, S., and Reddi, S. J. Reasoning with latent thoughts: On the power of looped transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=din0lGfZFd.

Schön, T., Gustafsson, F., and Nordlund, P.-J. Marginalized particle filters for mixed linear/nonlinear state-space models. In *IEEE Transactions on Signal Processing*, volume 53, pp. 2279–2289, 2005.

Shi, Z., Zhu, Y., Xie, Y., Shi, J., Xie, G., Zhang, H., Jiang, Y., Miao, C., and Li, Q. Reasoning under uncertainty: Efficient LLM inference via unsupervised confidence dilution and convergent adaptive sampling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32204–32218. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1638.

Younis, A. and Sudderth, E. B. Learning to be smooth: An end-to-end differentiable particle smoother. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=WdMhbqCoqW.

Zhang, J. Confidence-aware reasoning: Optimizing self-guided thinking trajectories in large reasoning models. *EMNLP Industry Track*, 2025.

Zhang, Z., Zheng, C., Wu, Y., Zhang, B., Lin, R., Yu, B., Liu, D., Zhou, J., and Lin, J. The lessons of developing process reward models in mathematical reasoning. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.547. URL https://aclanthology.org/2025.findings-acl.547/.