ENERGY-CALIBRATED IMPLICIT CHAIN-OF-THOUGHT FOR CONSISTENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) demonstrate strong reasoning capabilities through *Chain-of-Thought* (CoT) prompting, which enables step-by-step intermediate reasoning. However, explicit CoT relies on discrete token-level processes that are prone to error accumulation and often lead to rigid and inconsistent reasoning trajectories. Recent work has explored implicit reasoning in latent spaces, allowing models to perform internal multi-step reasoning before generating outputs, but such approaches lack explicit mechanisms to enforce reasoning consistency. To address this limitation, we propose `EBM-CoT`, an *Energy-Based Chain-of-Thought Calibration* framework that refines latent reasoning trajectories using an energy-based model (EBM). By guiding latent representations toward low-energy regions associated with coherent reasoning paths, `EBM-CoT` improves both reasoning accuracy and consistency without modifying the underlying language model. Experiments on mathematical, commonsense, and symbolic reasoning benchmarks demonstrate that EBM-CoT matches the performance of 10-chain Self-Consistency decoding using only a single reasoning chain, reducing the required sampling budget by 5x to 19x without modifying the underlying language model.

1 INTRODUCTION

Large Language Models (LLMs) have achieved groundbreaking progress in natural language processing (Achiam et al., 2023; Dubey et al., 2024; Yang et al., 2025; Guo et al., 2025; Team et al., 2025), demonstrating not only remarkable language understanding and generation capabilities, but also strong performance in reasoning tasks. Chain-of-Thought (CoT) has emerged as a central mechanism for enhancing the reasoning capabilities of large language models (LLMs), enabling them to generate intermediate reasoning steps that progressively lead to the final answer. Wei et al. (2022) first demonstrated that prompting models to generate intermediate reasoning steps can substantially improve performance in arithmetic and commonsense reasoning tasks, revealing the emergence of step-by-step reasoning abilities as the model scale increases. This work established CoT as a mechanism that not only guides explicit reasoning but also reflects the model's intrinsic thought process. However, traditional explicit CoT methods rely on discrete, token-level reasoning processes, which are prone to error propagation during generation and constrained by the limited expressiveness of the vocabulary space – resulting in rigid reasoning paths and poor robustness. To address these issues, recent research has explored *implicit reasoning* processes, allowing models to "think" in a continuous representation space before producing the final output. For example, methods such as Coconut (Hao et al., 2024) and CCoT (Cheng & Van Durme, 2024) compress discrete reasoning sequences into continuous latent variables, enabling efficient *soft reasoning*. Despite their promise, approaches like Coconut and CCoT require expensive full-model fine-tuning. Although SoftCoT (Xu et al., 2025) mitigates this by freezing the base model, it still suffers from a fundamental limitation: the lack of explicit trajectory-level consistency constraints during training, leading to divergent thoughts at inference. Consequently, multiple CoT samples are typically required to be generated and ranked to achieve satisfactory performance, at the cost of significantly reduced efficiency.

To address the aforementioned limitations, we propose an *implicit Chain-of-Thought calibration framework* based on an *Energy-Based Model (EBM)*, called `EBM-CoT`. An overview of the proposed framework is illustrated in Figure 1. The left part of the figure contrasts different reasoning paradigms, including No CoT, Explicit CoT, and Implicit CoT. Unlike prior approaches that rely solely on continuous representation mapping, `EBM-CoT` introduces a differentiable energy function $E_\phi(x)$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

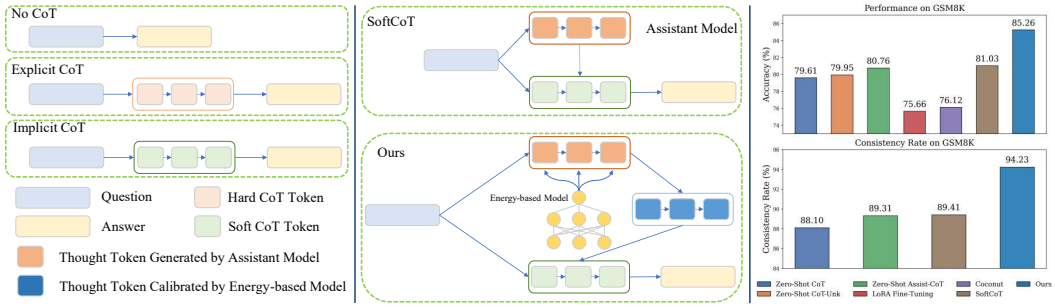


Figure 1: Comparison of reasoning paradigms and the proposed energy-based calibration framework. **Left Figure:** Different forms of reasoning in language models, including No CoT, Explicit CoT, and Implicit CoT. **Middle Figure:** Our method integrates an Energy-Based Model (EBM) to calibrate the latent thought tokens generated by the assistant model, producing refined soft thoughts with improved coherence and consistency. **Right Figure:** Experimental results on GSM8K show that our approach, implemented on top of LLaMA-3.1-8B-Instruct, achieves superior accuracy and substantially higher consistency compared to previous CoT variants, effectively matching the performance of multi-CoT reasoning with a single CoT using calibrated thought tokens.

defined in the thought embedding space, as illustrated in the middle part of Figure 1. This energy function explicitly calibrates the latent thought tokens, guiding them toward lower-energy regions that correspond to more task-relevant and logically consistent reasoning trajectories. Although our model performs energy-based calibration only at the embedding level, the learnt energy landscape implicitly integrates multiple potential reasoning directions in the continuous space. In our framework, each soft thought token l_t represents an implicit reasoning step within the chain-of-thought process. Although referred to as a "token" for its sequential and compositional role, it is realised as a continuous soft thought embedding in the latent space. This design preserves the token-level reasoning structure while enabling differentiable operations such as gradient-based energy optimization and Langevin dynamics refinement. As a result, calibrating the soft thought tokens enables a single chain of thought to approximate the reasoning performance of multi-CoT ensembles. During training, we perform multi-step energy-guided updates in the thought embedding space using Langevin dynamics. The energy gradient $-\nabla_x E_\phi(x)$ provides a corrective direction towards lower-energy regions (higher-consistency), automatically refining the model's reasoning trajectory at each inference step. Moreover, at the outer level of optimization, we jointly train the language modelling objective with the EBM objective, comprising both an energy ranking constraint and a consistency constraint, to align the direction of the gradient of the energy model with that of the language model. Consequently, the energy function not only learns *what constitutes a plausible reasoning representation*, but also *how to proactively refine implausible reasoning states* during the reasoning process. At inference time, `EBM-CoT` leverages the learnt energy function to perform a few steps of Langevin calibration in the thought embedding space, guiding the reasoning state from higher-energy (low-consistency) regions toward lower-energy (high-consistency) ones. The calibrated embeddings are then fed into the frozen base model for subsequent generation, thereby achieving dynamic reasoning calibration and enhanced robustness without any parameter modification to the base and assistant LLM. The right part of Figure 1 presents the empirical impact of our calibration mechanism, specifically, the consistency rate is defined as Consistency Rate $= \frac{\text{Acc(pass@1)}}{\text{Acc(pass@}N\text{)}} \times 100\%$, which measures how consistently the model reaches correct conclusions when multiple reasoning chains (pass@N) are sampled for the same input. As shown in this figure, we set $N = 10$ following the standard self-consistency protocol. A higher consistency rate indicates a more stable and reliable reasoning process, and our energy-based calibration significantly enhances both accuracy and consistency across sampled reasoning paths. These results indicate that aligning the latent thought tokens through energy-based calibration allows a single CoT to achieve performance comparable to that of multi-CoT ensembles ($N = 10$). Our key contributions are summarized as follows:

- We introduce a novel perspective on reasoning in LLMs by employing an Energy-Based Model (EBM) to enhance Chain-of-Thought reasoning. Unlike standard auto-regressive models that ensure only local token-level consistency, our approach explicitly promotes *global consistency* and *logical coherence* across the entire latent reasoning trajectory. This establishes a new bridge between energy-based modeling and implicit reasoning in language models.

- We design a hybrid, multi-stage architecture that separates implicit *thinking* in a latent space from explicit textual reasoning and answer generation. By applying Langevin dynamics to continuous latent embeddings, the EBM performs energy-based calibration of internal thoughts before text generation occurs, refining latent reasoning states toward low-energy, high-consistency regions in the embedding space.
- Extensive experiments on mathematical, commonsense, and symbolic reasoning benchmarks demonstrate that our method significantly improves both reasoning accuracy and consistency. Notably, the model achieves strong single-chain performance (N=1), narrowing the gap with multi-chain self-consistency results delivering massive wall-clock inference speedups of 5x to 19x. The source code of our framework is accessible at `https://anonymous.4open.science/r/EBM-CoT-BEDE`.

2 Related Work

CoT Performance Enhancement Wang et al. (2022) introduced the Self-Consistency decoding strategy, which strengthens robustness by sampling multiple reasoning paths and selecting the most consistent final answer. Tree-of-Thoughts (ToT) (Yao et al., 2023) extends the linear Chain-of-Thought paradigm into a tree-structured reasoning framework, enabling models to explore, evaluate, and backtrack among multiple reasoning paths. Graph-of-Thoughts (GoT) (Besta et al., 2024) further generalizes this idea by modelling reasoning nodes as a graph structure, supporting multi-path integration and feedback between thoughts. To make such a global search more efficient, Chain-of-Preference Optimization (CPO) (Zhang et al., 2024) uses implicit preference data embedded in ToT search trees to fine-tune models, allowing CoT to achieve reasoning quality close to ToT at a lower inference cost. Jiang et al. (2025) introduces the Energy Outcome Reward Model (EORM), an energy-based reranker that scores and orders Chain-of-Thought candidates by outcome quality. However, such approaches tend to significantly reduce the efficiency of CoT-based inference.

Energy-based Model In recent years, Energy-Based Models (EBMs) (Hopfield, 1982; Hinton & Sejnowski, 1983; Ren et al., 2024; Wang et al., 2025; Balcerak et al., 2025; Zhang et al., 2025b) have emerged as a unified probabilistic modelling paradigm, demonstrating remarkable expressiveness across generative modelling, model alignment, and decision optimization. Unlike approaches that explicitly model probability distributions, EBMs define distributions implicitly through an energy function, offering strong compositionality, constraint flexibility, and compatibility with other model architectures. In the optimization process, (Pang et al., 2024) and (Hong et al., 2024) proposed ARM and EPA, respectively, which replace RLHF with residual energy models and perform preference optimization through forward KL divergence minimisation, achieving both stability and efficiency. Zhou et al. (2024) further introduced ABDPO, an energy-based direct preference optimization framework for antibody design, where residue-level energy decomposition guides diffusion models to generate structurally plausible and high-affinity antibodies. In generative modelling, Xu et al. (2024) proposed the Energy-based Diffusion Language Model, which integrates energy functions with discrete diffusion processes to mitigate parallel decoding errors and improve text generation quality. Zhang et al. (2025a) introduced EnergyMoGen, which combines latent-space and semantic energy to enable compositional generation of multi-concept motions. Wang & Du (2025) proposed Equilibrium Matching, which unifies flow and diffusion models under a balanced energy field, supporting efficient sampling via gradient-based optimization. Overall, these studies highlight the broad applicability of EBMs across various domains.

3 Preliminaries

3.1 Problem Definition

We decompose the overall reasoning process of a large language model (LLM) into three auto-regressive stages that are consistent with the energy-adjusted chain-of-thought formulation introduced above. Let the input be x, and let the model generate a sequence of intermediate representations or reasoning steps $y_{1:T} = (y_1, y_2, \ldots, y_T)$ with contextual states $c_t = (x, y_{<t})$. LLMs often generate intermediate reasoning steps, or CoT, to solve complex reasoning tasks. Given an input x, the model produces a reasoning trajectory $y_{1:T} = (y_1, y_2, \ldots, y_T)$, where each token y_t corresponds to a partial reasoning step conditioned on the context $c_t = (x, y_{<t})$. Under the standard auto-regressive

factorization, the conditional distribution over the entire reasoning sequence is

$$p(y_{1:T} \mid x) = \prod_{t=1}^{T} p(y_t \mid c_t), \qquad c_t = (x, y_{<t}).$$

In conventional language models, each conditional probability $p(y_t \mid c_t)$ is parameterized by a softmax over the model logits. The probability of selecting a specific token v from the vocabulary $\mathcal{V}$ as the next token y_t is calculated as:

$$p(y_t = v \mid c_t) = \frac{\exp(W_v^\top h_t + b_v)}{\sum_{v' \in \mathcal{V}} \exp(W_{v'}^\top h_t + b_{v'})},$$

where $h_t = f_\theta(c_t)$ is the hidden representation of the context c_t, f_θ is the function of the neural network, and $\mathcal{V}$ is the vocabulary. This formulation can be equivalently viewed as a local energy-based model, where each next-token prediction defines a locally normalized energy landscape. The energy for a specific prediction y_t given the context c_t is

$$E_\theta(c_t, y_t) = -(W_{y_t}^\top h_t + b_{y_t}),$$

and the conditional probability of the next token y_t can be expressed using a Boltzmann distribution

$$p_\theta(y_t \mid c_t) = \frac{\exp[-E_\theta(c_t, y_t)]}{\sum_{y' \in \mathcal{V}} \exp[-E_\theta(c_t, y')]}.$$

Hence, next-token prediction in auto-regressive language modeling can be interpreted as a sequence of locally normalized EBMs defined over discrete vocabulary tokens. However, this formulation captures only token-level dependencies and provides local normalization at each step, without enforcing global consistency across reasoning trajectories. In the following sections, we extend this energy-based perspective from discrete token prediction to a continuous latent space, where reasoning is modeled as an energy-adjusted process over latent thought embeddings.

3.2 Energy-based Models for CoT

Energy-based models (EBMs) define a scalar energy function $E_\phi(l)$ that assigns low energy to preferred configurations and high energy to unlikely ones. The corresponding probability density $p_\phi(l)$ follows the Boltzmann distribution:

$$p_\phi(l) = \frac{\exp[-E_\phi(l)]}{Z_\phi}, \qquad Z_\phi = \int \exp[-E_\phi(l)]\, dl.$$

In the implicit CoT setting, the energy-based distribution can be expressed as

$$p_\phi(l \mid c) \propto p_a(l \mid c)\, \exp(-E_\phi(c, l)/T)\,,$$

where T is a temperature parameter that controls the strength of the energy scaling. The energy term reweights the base distribution, enabling the model to prefer output with lower energy – more coherent or semantically consistent responses.

As derived in Section 3.1, the next-token prediction can be viewed as a local energy-based model, each capturing short-range consistency but not explicitly optimized for global reasoning coherence. Conventional language modeling maximizes the log-likelihood of each next token, where the probability $p(l_t \mid c_t)$ is normalized independently at each step. However, this local normalization does not explicitly encourage global consistency or logical coherence across the entire reasoning sequence. To address this limitation, we reinterpret each conditional distribution $p(l_t \mid c_t)$ as an energy-calibrated distribution $\tilde{p}_\phi(l_t \mid c_t)$, where the energy function $E_\phi(c_t, l_t)$ modulates the likelihood of the base model according to a learned notion of the quality of reasoning. This formulation provides a unified view of auto-regressive language modelling within the energy-based modelling framework.

In this work, we adopt a short-run Langevin refinement procedure as an energy-guided calibration operator that corrects latent reasoning states toward lower-energy, more coherent configurations. Specifically, starting from an initial latent state $l^{(0)}$, we apply a small number of Langevin-style updates of the form:

$$l^{(s+1)} = l^{(s)} - \eta \nabla_l E_\phi(c, l^{(s)}) + \sqrt{2\eta}\, \varepsilon^{(s)}, \varepsilon^{(s)} \sim \mathcal{N}(0, I),$$

which defines a Langevin-inspired Markov process associated with the energy-based distribution $p_\phi(l \mid c) \propto \exp[-E_\phi(c, l)]$. This work adopts this as a short-run energy-based calibration operator to refine latent reasoning states, rather than as an asymptotically exact sampling procedure.

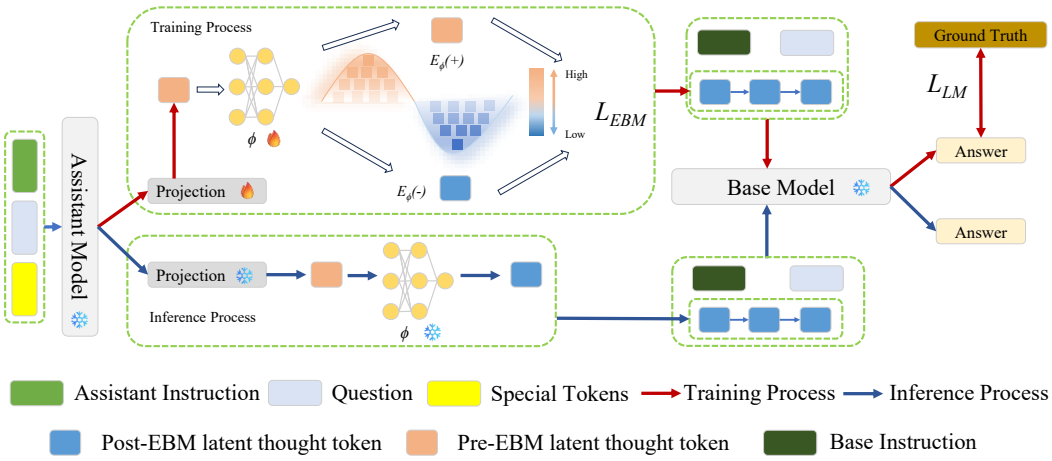


Figure 2: Overall architecture of the proposed EBM-CoT framework. Given an assistant instruction and a question, along with special tokens, the assistant model first encodes the input and generates a sequence of latent thought tokens that represent intermediate reasoning steps. These latent thoughts are then projected through a learnable projection module into the embedding space compatible with the base model. An Energy-Based Model (EBM) further refines the projected (Pre-EBM) latent thought tokens via Langevin calibration, assigning lower energy to coherent reasoning states. Finally, the calibrated (Post-EBM) latent thought tokens, together with the base instruction and question, are fed into the frozen Base model to generate explicit reasoning steps and produce the final answer. This pipeline enables consistent and efficient reasoning from question to answer while preserving the parameters of both the base and assistant models.

4 METHODOLOGY

In this section, we introduce our proposed method. Section 4.1 presents the overall framework of EBM-CoT, which incorporates energy-based calibration into the Chain-of-Thought reasoning process. Section 4.2 details the design of the loss function used to jointly optimize language modeling and energy-based objectives. Section 4.3 describes the optimization process, including Langevin refinement and training strategies. Additional proofs and algorithmic details are provided in Appendix B and E.

4.1 EBM-CoT

Overview The overall framework of our proposed EBM-CoT is illustrated in Figure 2. Unlike previous implicit CoT methods that only rely on continuous representation mapping, our approach introduces an Energy-Based Model (EBM) to explicitly calibrate latent thought tokens within the embedding space. Figure 2 also illustrates the overall training and inference processes of the proposed framework.

Training Process During training, the assistant model first generates latent thought tokens based on the input. The EBM then calibrates these latent thoughts through contrastive energy learning. Higher-energy ($E_\phi(+)$) and lower-energy ($E_\phi(-)$) samples are obtained via short-step Langevin dynamics and projected into the thought embedding space, where the EBM assigns lower energy to coherent reasoning tokens and higher energy to implausible ones. The model is optimized using a hinge-style contrastive loss $\mathcal{L}_{\mathrm{h}}(\phi)$ combined with a consistency regularization term $\mathcal{L}_c$. The combined energy-based objective is defined as: $\mathcal{L}_{\mathrm{EBM}} = \mathcal{L}_h + \mathcal{L}_c$. This joint objective enables the EBM to learn an energy landscape that effectively distinguishes consistent from inconsistent reasoning trajectories while maintaining smooth latent transitions. The corresponding training procedure is summarized in Algorithm 1.

Inference Process At inference time, the trained EBM performs a remarkably lightweight energy-based calibration on the latent thought embeddings generated by the assistant model. Specifically, given initial latents $l^{(0)}$ from the assistant, we execute a deterministic Langevin loop for S steps to

refine the latent reasoning trajectory, adjusting the embeddings toward lower-energy regions in the thought space: $l^{(s+1)} \leftarrow l^{(s)} - \eta \nabla_l E_\phi(c_t, l^{(s)})$. The calibrated latents $l^{(S)}$ are then fed into the frozen base model for final answer generation. This framework enables consistent and efficient reasoning without modifying the base or assistant models, while avoiding the substantial computational overhead of multi-path textual decoding. Efficiency results are provided in Tables 4–6 (Appendix H). The inference procedure is summarized in Algorithm 2.

4.2 Loss Function Design

We optimize a combined objective consisting of a language modeling loss $\mathcal{L}_{\text{LM}}$ (computed by the frozen base model) and an energy-based regularization term $\mathcal{L}_{\text{EBM}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \alpha \, \mathcal{L}_{\text{EBM}},$$

where $\mathcal{L}_{\text{LM}}$ denotes the standard token-level negative log-likelihood of the ground-truth sequence under the frozen base model, while $\mathcal{L}_{\text{EBM}}$ shapes the latent reasoning trajectory in a continuous space. The base model and assistant model are frozen, only the energy function E_ϕ and the projection module are optimized.

For each context c_t, we construct a pair of latent states (l^ℓ, l^c) through Langevin refinement. Specifically, l^ℓ denotes the *initial latent state* produced by the assistant model, while l^c is obtained by refining l^ℓ via S steps of Langevin dynamics driven by the energy function E_ϕ. By construction, l^c corresponds to a lower-energy (positive) sample, and l^ℓ serves as a higher-energy (negative) sample. We adopt a hinge-style contrastive objective:

$$\mathcal{L}_h = \text{ReLU}\big(E_\phi(c_t, l^c) - E_\phi(c_t, l^\ell) + m\big), \tag{1}$$

which enforces a margin m between the energies of the initial and refined latent states. In addition, we include a quadratic regularizer

$$\mathcal{L}_c = \lambda \|l^c - l^\ell\|^2,$$

where $\lambda > 0$ controls the strength of the consistency regularization, balancing energy-based calibration against excessive deviation from the assistant-provided latent initialization. This term stabilizes Langevin refinement by discouraging excessively large latent shifts.

The final EBM objective is $\mathcal{L}_{\text{EBM}} = \frac{1}{B} \sum_{b=1}^{B} \left(\mathcal{L}_h^{(b)} + \mathcal{L}_c^{(b)}\right)$, where B is the batch size.

4.3 Optimization Process

To demonstrate the effectiveness of our optimization approach, we provide a theoretical analysis of how gradients propagate through the Langevin refinement steps. More details can be found in B.6.

Gradient Propagation of $\mathcal{L}_{\text{LM}}$ The language modeling loss, $\mathcal{L}_{\text{LM}}$, depends on the final latent embedding $l^{(S)}$ obtained after S Langevin steps. Because each step incorporates the energy function E_ϕ, we compute the gradient with respect to ϕ by unrolling the Langevin chain. As detailed in Appendix B, applying the chain rule recursively yields:

$$\frac{\partial \mathcal{L}_{\text{LM}}}{\partial \phi} = -\eta \sum_{k=0}^{S-1} \frac{\partial \mathcal{L}_{\text{LM}}}{\partial l^{(S)}} \left[\prod_{j=k+1}^{S-1} (I - \eta A^{(j)}) \right] B^{(k)},$$

where $A^{(j)}$ and $B^{(k)}$ represent the Jacobians of the energy gradient with respect to the latent states and parameters, respectively.

EBM Regularization Simultaneously, the EBM applies a contrastive hinge loss with margin m and an L2 penalty to regularize the latent space:

$$\mathcal{L}_{\text{EBM}} = \text{ReLU}\big(E_\phi(c_t, l^c) - E_\phi(c_t, l^\ell) + m\big) + \lambda \|l^c - l^\ell\|^2.$$

5 Experiments

To evaluate the effectiveness of our proposed method, we conduct experiments on a diverse set of benchmarks, including datasets of mathematical reasoning, commonsense reasoning, and symbolic

reasoning. To ensure a fair comparison, we follow the configurations of the dataset and the model selections adopted in Xu et al. (2025). Detailed descriptions of the baseline methods and datasets are provided in the Appendices F and G. Additional experimental results based on Qwen2.5-7B-Instruct are provided in Appendix I.

Table 1: Model performance using LLaMA-3.1-8B-Instruct. Here N denotes the number of sampled reasoning chains. We apply the self-consistency (Wang et al., 2022) sampling procedure to aggregate answers across these chains.

Model	GSM8K	ASDiv-Aug Mathematical	AQuA	StrategyQA Commonsense	DU Symbolic	Avg.
Zero-Shot CoT	$79.61_{\pm 0.81}$	$86.78_{\pm 0.63}$	$54.65_{\pm 2.43}$	$65.63_{\pm 3.31}$	$54.40_{\pm 2.40}$	68.21
Zero-Shot CoT-Unk	$79.95_{\pm 0.59}$	$86.90_{\pm 0.41}$	$55.28_{\pm 1.88}$	$66.16_{\pm 2.70}$	$54.16_{\pm 1.46}$	68.49
Zero-Shot Assist-CoT	$80.76_{\pm 1.53}$	$86.96_{\pm 0.46}$	$55.83_{\pm 2.98}$	$66.55_{\pm 3.99}$	$58.24_{\pm 3.56}$	69.67
LoRA Fine-Tuning	$75.66_{\pm 0.00}$	$86.67_{\pm 0.00}$	$52.36_{\pm 0.00}$	-	-	-
Coconut	$76.12_{\pm 0.00}$	$86.80_{\pm 0.00}$	$53.15_{\pm 0.00}$	-	-	-
SoftCoT	$81.03_{\pm 0.42}$	$87.19_{\pm 0.40}$	$56.30_{\pm 1.67}$	$69.04_{\pm 1.23}$	$59.04_{\pm 1.93}$	70.52
Ours	**85.26**$_{\pm 0.04}$	**88.20**$_{\pm 0.08}$	**58.14**$_{\pm 0.18}$	**69.54**$_{\pm 0.33}$	**61.32**$_{\pm 0.26}$	**72.49**
Zero-Shot CoT ($N = 10$)	$90.36_{\pm 0.40}$	$89.23_{\pm 0.17}$	$63.23_{\pm 0.86}$	$70.13_{\pm 0.47}$	$65.76_{\pm 1.54}$	75.21
Zero-Shot Assist-CoT ($N = 10$)	$90.43_{\pm 0.69}$	$89.48_{\pm 0.36}$	$63.62_{\pm 0.99}$	$70.48_{\pm 0.68}$	$65.84_{\pm 1.93}$	75.97
Coconut ($N = 10$)	$87.03_{\pm 0.00}$	$88.44_{\pm 0.00}$	$61.81_{\pm 0.00}$	-	-	-
SoftCoT ($N = 10$)	$90.63_{\pm 0.39}$	$89.75_{\pm 0.29}$	$65.51_{\pm 0.72}$	$71.14_{\pm 0.10}$	$67.36_{\pm 1.12}$	76.88
Ours ($N = 10$)	$90.48_{\pm 0.08}$	$89.33_{\pm 0.05}$	$65.45_{\pm 0.32}$	$71.18_{\pm 0.20}$	$67.93_{\pm 0.25}$	76.87

Table 2: Model performance using Qwen3-8B.

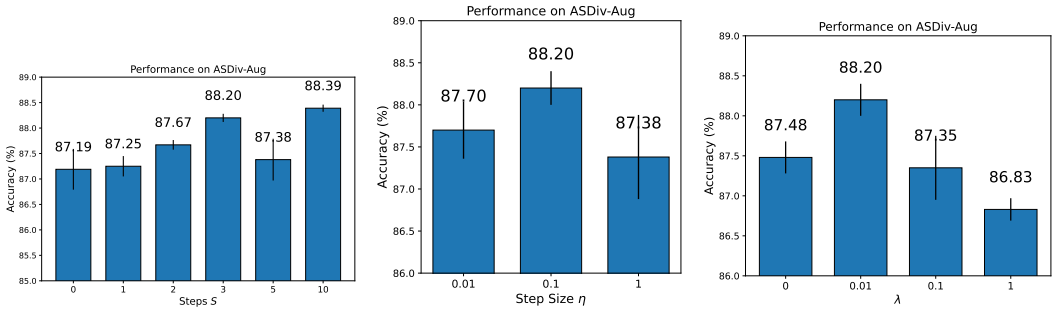
Model	GSM8K	ASDiv-Aug Mathematical	AQuA	StrategyQA Commonsense	DU Symbolic	Avg.
Zero-Shot CoT	$91.86_{\pm 0.41}$	$91.70_{\pm 0.26}$	$70.00_{\pm 1.56}$	$69.87_{\pm 0.35}$	$80.32_{\pm 2.32}$	80.75
Zero-Shot Assist-CoT	$91.90_{\pm 0.50}$	$91.64_{\pm 0.16}$	$70.16_{\pm 2.35}$	$69.78_{\pm 0.33}$	$80.56_{\pm 1.15}$	80.81
Coconut	$87.95_{\pm 0.00}$	$89.40_{\pm 0.00}$	$68.50_{\pm 0.00}$	-	-	-
SoftCoT	$92.48_{\pm 0.29}$	$91.83_{\pm 0.19}$	$75.04_{\pm 2.68}$	$70.17_{\pm 0.63}$	$81.21_{\pm 0.70}$	82.15
Ours	**92.82**$_{\pm 0.09}$	**92.22**$_{\pm 0.18}$	**82.15**$_{\pm 1.59}$	**70.96**$_{\pm 0.36}$	**82.29**$_{\pm 0.87}$	**84.09**
Zero-Shot CoT ($N = 10$)	$92.22_{\pm 0.47}$	$91.97_{\pm 0.13}$	$76.77_{\pm 0.62}$	$70.96_{\pm 0.15}$	$84.56_{\pm 0.61}$	83.30
Zero-Shot Assist-CoT ($N = 10$)	$92.68_{\pm 0.17}$	$91.91_{\pm 0.28}$	$76.77_{\pm 0.79}$	$70.92_{\pm 0.28}$	$84.80_{\pm 1.17}$	83.82
Coconut ($N = 10$)	$90.37_{\pm 0.00}$	$90.37_{\pm 0.00}$	$76.38_{\pm 0.00}$	-	-	-
SoftCoT ($N = 10$)	$93.19_{\pm 0.32}$	$92.14_{\pm 0.15}$	$80.63_{\pm 1.90}$	$71.18_{\pm 0.15}$	$84.01_{\pm 0.22}$	84.87
Ours ($N = 10$)	$93.28_{\pm 0.25}$	$92.26_{\pm 0.04}$	$88.19_{\pm 0.58}$	$71.11_{\pm 0.37}$	$84.89_{\pm 0.29}$	85.95

5.1 SETTINGS

Datasets and Models For the datasets, we have evaluated our method across multiple benchmarks, covering three major categories: mathematical, commonsense, and symbolic reasoning. For mathematical reasoning, we adopt the GSM8K (Cobbe et al., 2021), ASDiv-aug (Miao et al., 2021), and AQuA (Ling et al., 2017) datasets; for commonsense reasoning, we use StrategyQA (Geva et al., 2021); and for symbolic reasoning, we employ the Data Understanding (DU) dataset (Srivastava et al., 2023). For the models, we have conducted experiments across diverse architectures and parameter scales. The base models include Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen3-8B (Yang et al., 2025), and LLaMA-3.1-8B-Instruct (Dubey et al., 2024). The assistant models consist of Qwen2.5-7B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-0.5B-Instruct, and Qwen3-0.6B.

Baselines We compare our approach against three categories of baselines: Zero-Shot, LoRA fine-tuning, and implicit CoT-based reasoning enhancement methods. For the zero-shot setting, we evaluate Zero-Shot CoT, Zero-Shot Unk, and Zero-Shot Assist-CoT. For LoRA fine-tuning (Hu et al., 2022), we apply LoRA framework to assess performance improvements. For implicit CoT reasoning enhancement, we compare with recent methods such as Coconut (Hao et al., 2024) and SoftCoT (Xu et al., 2025). Detailed descriptions of these baseline methods are provided in Appendix F. In addition to the conventional accuracy metric used to evaluate model performance, we introduce a consistency rate metric to quantify the coherence of reasoning across multiple CoT samples.

Implementation Details We report the training and inference details of our proposed method. Following the setup in Xu et al. (2025) for fair comparison, we jointly train the Energy-Based Model and the projection module using the AdamW optimizer with an initial learning rate of 2×10^{-5}.



((a)) EBM correction steps S. ((b)) EBM step size η. ((c)) Consistency weight λ.

Figure 3: Ablation studies on ASDiv-Aug with respect to the number of EBM correction steps S, step size η, and consistency weight λ.

During training, the number of latent thought tokens is set to 32, and we train for 10 epochs with a batch size of 4. For the Langevin calibration process, we apply 3 refinement steps to refine the latent tokens, and set the weighting coefficient α for the energy-based regularization term to 0.1. We also conduct ablation studies to examine the influence of these hyperparameters and other model components. During inference, the main results are obtained using 4 latent thought tokens, and we additionally perform ablation studies on different token numbers to evaluate the trade-off between reasoning depth and efficiency. For the Energy-Based Model, we adopt a multi-layer perceptron architecture as an energy function, following the design choices in Jiang et al. (2025).

5.2 Performance Across Base Models

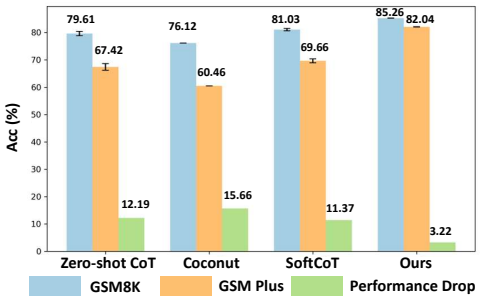


Figure 4: Performance comparison on GSM8K and GSM Plus benchmarks using models from the Llama 3.1 family.

Tables 1 and 2 report the reasoning performance of different methods using LLaMA-3.1-8B-Instruct and Qwen3-8B as base models (with LLaMA-3.2-1B-Instruct and Qwen3-0.6B as assistants, respectively). Across both model families, our energy-based calibration consistently outperforms all baselines. On LLaMA-3.1-8B, our method achieves an average accuracy of 72.49% with a single reasoning chain (Ours, N=1), closely approaching its multi-chain counterpart (Ours, N=10) at 76.87%. In contrast, baselines such as SoftCoT and Coconut exhibit much larger performance gaps between N=1 and N=10. This reduced reliance on multi-sample self-consistency indicates that our calibration effectively enhances intrinsic reasoning coherence within each latent trajectory, yielding highly reliable single-pass results (particularly evident on ASDiv-Aug and GSM8K). These improvements seamlessly generalize to Qwen3-8B, where our approach achieves an average +2.1% improvement over SoftCoT. Furthermore, applying self-consistency (Ours, N=10) on Qwen3-8B boosts average accuracy to 85.95%, surpassing all other multi-sample baselines. Performance gains are especially pronounced on symbolic (DU) and commonsense (StrategyQA) tasks, demonstrating an effective mitigation of reasoning variance. Overall, these findings highlight that our latent calibration not only improves single-path reasoning stability but is also highly complementary to self-consistency aggregation across different model architectures and reasoning domains.

5.3 Theoretical and Geometric Interpretation of Energy Effectiveness

Empirical Results Figure 4 shows our method outperforms baselines on GSM8K and GSM Plus. With only a minimal 3.22% performance drop under distribution shifts, it demonstrates strong robustness and transferable reasoning rather than mere overfitting. Table 3 reveals that EBM calibration induces small yet highly structured latent modifications: the updates are stable (narrow pooled L2 distance), targeted (larger maximum token-level L2), and closely aligned with initial states (cosine similarity > 0.89). Furthermore, the consistent energy decrease ($\Delta E > 0$) confirms these localized updates effectively push latent states into lower-energy regions. Notably, AQuA exhibits larger gains,

Table 3: Latent geometry analysis of EBM calibration across datasets using Qwen3 family.

Dataset	Pooled L2	Token-Max L2	Dir. Cosine	Performance Increase (%)	Energy Decrease ($\Delta E > 0$)
AQuA	0.046 ± 0.002	0.082 ± 0.002	0.89 ± 0.01	7.11	0.0062
GSM8K	0.155 ± 0.005	0.2192 ± 0.005	0.96 ± 0.01	0.34	0.084
ASDiv-Aug	0.070 ± 0.002	0.106 ± 0.002	0.93 ± 0.02	0.39	0.015
StrategyQA	0.158 ± 0.002	0.191 ± 0.002	0.98 ± 0.01	0.79	0.0048

likely due to its smaller evaluation set and the high sensitivity of discrete answer selection to minor latent perturbations.

Theoretical Interpretation The gradient formulation above provides the mechanistic basis for the Energy-Likelihood Alignment. Specifically, the total gradient $\frac{\partial \mathcal{L}_{\text{total}}}{\partial \phi} = \frac{\partial \mathcal{L}_{\text{LM}}}{\partial \phi} + \alpha \frac{\partial \mathcal{L}_{\text{EBM}}}{\partial \phi}$ illustrates exactly how our joint training framework forces the EBM to act as a rigorous evaluator of logical correctness. The term $\frac{\partial \mathcal{L}_{\text{LM}}}{\partial l^{(S)}}$ propagates the language modeling signal, essentially the downstream success probability, backwards through the sequence of latent Langevin updates.

Meanwhile, the Jacobian product $\prod_{j=k+1}^{S-1}(I - \eta A^{(j)})$ and parameter derivative $B^{(k)}$ modulate this error signal to directly adjust the energy function E_ϕ.While the contrastive $\mathcal{L}_{\text{EBM}}$ explicitly carves out low-energy basins for mathematically stable reasoning trajectories, the unrolled $\mathcal{L}_{\text{LM}}$ gradient acts as a semantic safeguard. If the Langevin process converges to a low-energy state $l^{(S)}$ that yields an incorrect answer (i.e., high $\mathcal{L}_{\text{LM}}$), the backpropagated error heavily penalizes the EBM parameters ϕ, effectively reshaping the landscape to eliminate spurious low-energy pits. Consequently, the optimal energy landscape is structurally constrained to approximate the negative log-likelihood of the target reasoning manifold:$E_\phi^*(l) \propto -\log p_\theta(A^*|x, l) + C$where A^* represents the ground-truth answer and C is a context-dependent constant. As the EBM learns to formally couple minimum energy with maximum answer accuracy, the Langevin refinement acts as a structure-preserving semantic sharpening. It iteratively pushes latent thought tokens toward configurations that are not only energetically stable but globally coherent, thereby guaranteeing enhanced consistency in the final CoT generation.

5.4 Ablation Study

We evaluate the impact of key hyperparameters on ASDiv-Aug (Figures 3(a)–3(c)). Increasing the number of EBM correction steps (S) improves performance up to $S = 3$, beyond which gains diminish, making it an optimal trade-off between accuracy and computational efficiency. For the correction step size (η) and consistency weight (λ), performance follows a unimodal trend. Moderate values ($\eta = 0.1$ and $\lambda = 0.01$) achieve the best results by balancing sufficient latent refinement with stable optimization, avoiding both under-constrained and overly restricted updates. Additionally, we evaluate the framework's scalability across varying assistant model capacities (Figure 6, Right). Results demonstrate that while stronger assistants provide better latent priors, EBM-CoT consistently delivers accuracy gains regardless of the assistant's parameter scale. Additional experiments are detailed in Appendices H, I, J, and N.

6 Conclusion

This work introduces an energy-based calibration framework that bridges the gap between explicit Chain-of-Thought (CoT) reasoning and continuous latent thought token optimization. Our experiments demonstrate that our EBM-CoT consistently improves reasoning accuracy and stability across mathematical, commonsense, and symbolic reasoning benchmarks while maintaining high computational efficiency. Despite these promising results, several limitations remain. First, the computational overhead introduced by the Langevin refinement, although lightweight in our setting ($S = 3$ steps), may still scale unfavorably for extremely high-dimensional embedding spaces or very long reasoning sequences. Second, the current energy function is implemented as a shallow MLP, which may limit its expressiveness in modelling complex latent dependencies. Future work will explore more structured or hierarchical energy formulations and adaptive Langevin updates to improve scalability and dynamic reasoning control. Overall, this study highlights the potential of energy-based modelling as a powerful and general mechanism for reasoning calibration in LLMs. We hope this work inspires further exploration of the implicit CoT field.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. *arXiv preprint arXiv:2504.10612*, 2025.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Geoffrey E Hinton and Terrence J Sejnowski. Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, volume 448, pp. 448–453. Washington, 1983.

Yuzhong Hong, Hanshan Zhang, Junwei Bao, Hongfei Jiang, and Yang Song. Energy-based preference model offers better offline alignment than the bradley-terry preference model. *arXiv preprint arXiv:2412.13862*, 2024.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Eric Hanchen Jiang, Haozheng Luo, Shengyuan Pang, Xiaomin Li, Zhenting Qi, Hengli Li, Cheng-Fu Yang, Zongyu Lin, Xinfeng Li, Hao Xu, et al. Learning to rank chain-of-thought: An energy-based approach with outcome supervision. *arXiv preprint arXiv:2505.14999*, 2025.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. *arXiv preprint arXiv:2106.15772*, 2021.

Bo Pang, Caiming Xiong, and Yingbo Zhou. Arm: Alignment with residual energy-based model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8218–8229, 2024.

Milong Ren, Tian Zhu, and Haicang Zhang. Carbonnovo: Joint design of protein structure and sequence using a unified energy-based model. In *Forty-first International Conference on Machine Learning*, 2024.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yan Wang, Ling Guo, Hao Wu, and Tao Zhou. Energy-based diffusion generator for efficient sampling of boltzmann distributions. *Neural Networks*, pp. 108126, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. Energy-based diffusion language models for text generation. *arXiv preprint arXiv:2410.21357*, 2024.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv e-prints*, pp. arXiv–2412, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17592–17602, 2025a.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*, 2025b.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. *Advances in Neural Information Processing Systems*, 37:120861–120891, 2024.

A PRELIMINARY

Thinking Stage The assistant model first produces a sequence of latent thought embeddings $L = \{l_1, \ldots, l_n\}$ that represent continuous reasoning states and capture the model's internal thought process:

$$p_a(L \mid x) = \prod_{i=1}^{n} p_a(l_i \mid c_i), \qquad c_i = (x, l_{<i}), \tag{2}$$

where $p_a(l_i \mid c_i)$ denotes the conditional distribution of the assistant model over latent thoughts given the input x and the previously generated latent sequence $l_{<i}$. Each latent thought l_i is a continuous embedding derived from the hidden representations of the assistant model. These latent embeddings are then projected into the base model's input space before being refined by the Energy-Based Model (EBM) in the subsequent calibration step.

Reasoning Stage Conditioned on latent thoughts L generated by the assistant model, the base model generates an explicit textual reasoning trajectory $R = \{r_1, \ldots, r_m\}$:

$$p_\theta(R \mid x, L) = \prod_{j=1}^{m} p_\theta(r_j \mid c_j), \qquad c_j = (x, L, r_{<j}), \tag{3}$$

where $p_\theta(r_j \mid c_j)$ represents the conditional probability of the base model of producing the j-th reasoning token based on input x, latent thoughts L, and previously generated reasoning tokens $r_{<j}$.

Answer Generation Stage Finally, the base model synthesizes the reasoning trace into an answer sequence $A = \{a_1, \ldots, a_k\}$:

$$p_\theta(A \mid x, L, R) = \prod_{k=1}^{o} p_\theta(a_k \mid c_k), c_k = (x, L, R, a_{<k}). \tag{4}$$

Overall Formulation The complete conditional distribution of the reasoning process can thus be written as

$$p(L, R, A \mid x) = p_a(L \mid x)\; p_\theta(R \mid x, L)\; p_\theta(A \mid x, L, R), \tag{5}$$

which hierarchically decomposes the reasoning process into latent thinking, explicit reasoning, and answer generation stages. This hierarchical decomposition shows that the Thinking, Reasoning, and Answer stages are not only conditioned on input x, but also closely depend on the latent thought tokens L that mediate intermediate reasoning. When the input x is fixed, improving the consistency of reasoning therefore reduces to refining these latent thought tokens, which serve as key carriers of the reasoning dynamics. Since our method performs calibration on the latent thought tokens, in the following derivations, we instantiate the generic variable y as the latent thought embedding l.

B THEORETICAL PROPERTIES

B.1 RESIDUAL ENERGY-BASED CONDITIONAL MODEL

Given a condition $c_t = (x, l_{<t})$, we define the conditional energy-based model (EBM) over latent thought embeddings l_t as

$$\tilde{p}_\phi(l_t \mid c_t) = \frac{1}{Z_\phi(c_t)}\, p_{\mathrm{a}}(l_t \mid c_t)\, \exp(-E_\phi(c_t, l_t)/T)\,, \tag{6}$$

where $p_{\mathrm{a}}(l_t \mid c_t)$ denotes the assistant-provided latent thought token distribution, $E_\phi(c_t, l_t)$ is the energy function parameterized by ϕ, T is a temperature, and $Z_\phi(c_t)$ is the normalization constant (partition function):

$$Z_\phi(c_t) = \mathbb{E}_{v \sim p_{\mathrm{a}}(\cdot \mid c_t)}\left[\exp\left(-\frac{E_\phi(c_t, v)}{T}\right)\right] = \int \exp\left(-\frac{E_\phi(c_t, v)}{T}\right) p_{\mathrm{a}}(v \mid c_t)\, dv. \tag{7}$$

The normalized form is therefore

$$\tilde{p}_\phi(l_t \mid c_t) = \frac{p_{\mathrm{a}}(l_t \mid c_t)\, \exp(-E_\phi(c_t, l_t)/T)}{\mathbb{E}_{v \sim p_{\mathrm{a}}(\cdot \mid c_t)}[\exp(-E_\phi(c_t, v)/T)]}.$$

B.2 CHAIN-OF-THOUGHT FACTORIZATION

For a latent sequence $L = \{l_1, \ldots, l_T\}$ conditioned on input x, we assume the auto-regressive factorization

$$p_{\mathrm{a}}(L \mid x) = \prod_{t=1}^{T} p_{\mathrm{a}}(l_t \mid c_t),$$

where $c_t = (x, l_{<t})$ denotes the context up to step t. Replacing the assistant prior with the energy-calibrated latent distribution yields

$$\tilde{p}_\phi(L \mid x) = \prod_{t=1}^{T} \tilde{p}_\phi(l_t \mid c_t),$$

which is the distribution our latent EBM aims to approximate or induce.

B.3 THE RELATIONSHIP BETWEEN $l^{(S)}$ AND $l^{(s)}$

The relationship between $l^{(S)}$ and $\{l^{(s)}\}_{s=0}^{S-1}$ can be expressed recursively as

$$l^{(s+1)} = l^{(s)} - \eta \nabla_l E_\phi(c_t, l^{(s)}) + \sqrt{2\eta}\, \varepsilon^{(s)}, \qquad \varepsilon^{(s)} \sim \mathcal{N}(0, I),$$

and by unrolling all S updates, the final latent state can be written as

$$l^{(S)} = l^{(0)} - \eta \sum_{s=0}^{S-1} \nabla_l E_\phi(c_t, l^{(s)}) + \sqrt{2\eta} \sum_{s=0}^{S-1} \varepsilon^{(s)}.$$

Therefore, $l^{(S)}$ is an accumulated result of all intermediate latent states $\{l^{(s)}\}$, each contributing a gradient-based correction and a stochastic perturbation toward a lower-energy (i.e., more consistent) region of the latent space.

B.4 LANGEVIN CALIBRATION PROCESS

To draw samples from the energy-based latent model, we apply overdamped Langevin dynamics directly in the continuous latent (embedding) space. At iteration s, the candidate latent state $l^{(s)}$ evolves according to the discretized update

$$l^{(s+1)} = l^{(s)} - \eta \nabla_l E_\phi(c_t, l^{(s)}) + \sqrt{2\eta}\, \varepsilon^{(s)}, \qquad \varepsilon^{(s)} \sim \mathcal{N}(0, I),$$

where η is the step size and the noise scale $\sqrt{2\eta}$ ensures consistency with the continuous Langevin Stochastic Differential Equation (SDE). This update defines a conditional Markov chain

$$l^{(s+1)} \sim p_\phi(l^{(s+1)} \mid l^{(s)}, c_t),$$

which defines a Langevin-inspired refinement process that biases latent samples toward low-energy regions specified by the energy function E_ϕ, while being initialized from the assistant distribution $p_{\mathrm{a}}(l \mid c_t)$. Rather than serving as an exact Markov Chain Monte Carlo sampler for a joint residual distribution, this process is used as an energy-guided calibration mechanism in the latent space.

In the conditional setting, $c_t = (x, l_{<t})$ is a fixed context comprising the input x and the previously generated latent thoughts, while $l^{(s)}$ denotes the current latent candidate in the inner-loop refinement. Each update $l^{(s+1)}$ depends only on $l^{(s)}$, forming a Markov process conditioned on c_t. In practice, we initialize $l^{(0)}$ with the assistant output $l^+ \sim p_{\mathrm{a}}(\cdot \mid c_t)$, which typically lies in a higher-energy region of the latent space. After S Langevin refinement steps, we obtain the calibrated sample $l^c \equiv l^{(S)}$, corresponding to a lower-energy and more consistent latent thought, while $l^\ell \equiv l^+$ denotes the unrefined, higher-energy latent sample provided by the assistant model.

B.5 GRADIENT PROPAGATION THROUGH LANGEVIN STEPS.

The Langevin update is

$$l^{(s+1)} = l^{(s)} - \eta \nabla_l E_\phi(c_t, l^{(s)}) + \sqrt{2\eta}\, \varepsilon^{(s)}, \qquad \varepsilon^{(s)} \sim \mathcal{N}(0, I).$$

Since the final loss $\mathcal{L}_{\mathrm{LM}}$ depends on $l^{(S)}$, by the chain rule

$$\frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial \phi} = \frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial l^{(S)}} \cdot \frac{\partial l^{(S)}}{\partial \phi}.$$

Let $g^{(s)} = \nabla_l E_\phi(c_t, l^{(s)})$ and define

$$A^{(s)} = \frac{\partial g^{(s)}}{\partial l}, \qquad B^{(s)} = \frac{\partial g^{(s)}}{\partial \phi}.$$

Then each step satisfies the recursion

$$\frac{\partial l^{(s+1)}}{\partial \phi} = (I - \eta A^{(s)}) \frac{\partial l^{(s)}}{\partial \phi} - \eta B^{(s)}, \qquad \frac{\partial l^{(0)}}{\partial \phi} = 0.$$

Unrolling the recursion gives

$$\frac{\partial l^{(S)}}{\partial \phi} = -\eta \sum_{k=0}^{S-1} \Big(\prod_{j=k+1}^{S-1} (I - \eta A^{(j)}) \Big) B^{(k)},$$

and hence

$$\frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial \phi} = -\eta \sum_{k=0}^{S-1} \left(\frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial l^{(S)}} \prod_{j=k+1}^{S-1} (I - \eta A^{(j)}) \right) B^{(k)}.$$

B.6 Optimization Process

To demonstrate the effectiveness of our optimization approach, we provide a corresponding theoretical analysis of the optimization process.

Analysis of $\mathcal{L}_{\mathrm{LM}}$ To examine the gradient flow of our optimization method, we analyze the gradient propagation of the language modeling loss $\mathcal{L}_{\mathrm{LM}}$ through the Langevin refinement steps. Since each latent state $l^{(s)}$ in the Langevin refinement process depends on the energy function parameter ϕ, the gradient of the language modelling loss $\mathcal{L}_{\mathrm{LM}}$ with respect to ϕ can be derived using the chain rule:

$$\frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial \phi} = \frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial l^{(S)}} \cdot \frac{\partial l^{(S)}}{\partial \phi} = \sum_{s=0}^{S-1} \frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial l^{(s+1)}} \cdot \frac{\partial l^{(s+1)}}{\partial \phi}.$$

Here, $l^{(S)}$ represents the final latent embedding obtained after S Langevin refinement steps, $l^{(s)}$ denotes the latent thought embedding in the s-th iteration ($s = 0, 1, \ldots, S-1$), and each update follows the Langevin step

$$l^{(s+1)} = l^{(s)} - \eta \nabla_l E_\phi(c_t, l^{(s)}) + \sqrt{2\eta}\, \varepsilon^{(s)}, \varepsilon^{(s)} \sim \mathcal{N}(0, I),$$

where $\varepsilon^{(s)}$ is independent of ϕ, so $\partial \varepsilon^{(s)} / \partial \phi = 0$. The gradient for a single step $s+1$ can then be calculated recursively using the chain rule:

$$\frac{\partial l^{(s+1)}}{\partial \phi} = \frac{\partial l^{(s)}}{\partial \phi} - \eta \left[A^{(s)} \frac{\partial l^{(s)}}{\partial \phi} + B^{(s)} \right],$$

where

$$A^{(s)} = \frac{\partial}{\partial l} \big(\nabla_l E_\phi(c_t, l^{(s)}) \big), B^{(s)} = \frac{\partial (\nabla_l E_\phi(c_t, l^{(s)}))}{\partial \phi},$$

is the Jacobian of the energy gradient with respect to l. Unrolling this recursion yields the full matrix-product form for the gradient with respect to the final latent state $l^{(S)}$:

$$\frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial \phi} = -\eta \sum_{k=0}^{S-1} \frac{\partial \mathcal{L}_{\mathrm{LM}}}{\partial l^{(S)}} \left[\prod_{j=k+1}^{S-1} (I - \eta A^{(j)}) \right] B^{(k)}.$$

The detailed derivation is provided in Appendix B.

Analysis of $\mathcal{L}_{\text{EBM}}$ We define

$$\Delta E_\phi := E_\phi(c_t, l^c) - E_\phi(c_t, l^\ell).$$

The EBM loss is

$$\mathcal{L}_{\text{EBM}} = \text{ReLU}(\Delta E_\phi + m) + \lambda \|l^c - l^\ell\|^2.$$

Its gradient with respect to ϕ is

$$\frac{\partial \mathcal{L}_{\text{EBM}}}{\partial \phi} = \mathbf{1}\{\Delta E_\phi + m > 0\} \frac{\partial \Delta E_\phi}{\partial \phi} + 2\lambda \, (l^c - l^\ell)^\top \frac{\partial (l^c - l^\ell)}{\partial \phi},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function that activates the hinge loss only when the margin constraint $E_\phi(c_t, l^c) - E_\phi(c_t, l^\ell) + m > 0$ is met (when the positive example does not have a sufficiently lower energy).

Final Gradient Expression Combining both components, the total gradient of the training objective can be expressed as the sum of the LM-induced gradient propagated through the Langevin chain and the direct EBM gradient:

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial \phi} = \frac{\partial \mathcal{L}_{\text{LM}}}{\partial \phi} + \alpha \frac{\partial \mathcal{L}_{\text{EBM}}}{\partial \phi}.$$

Writing the LM contribution in shorthand yields

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial \phi} = -\eta \sum_{k=0}^{S-1} \frac{\partial \mathcal{L}_{\text{LM}}}{\partial l^{(S)}} \left[\prod_{j=k+1}^{S-1} (I - \eta A^{(j)}) \right] B^{(k)} + \alpha \frac{\partial \mathcal{L}_{\text{EBM}}}{\partial \phi}.$$

The gradient formulation above illustrates how the EBM regularizes the latent reasoning trajectory through the Langevin refinement process. Specifically, the term $\frac{\partial \mathcal{L}_{\text{LM}}}{\partial l^{(S)}}$ propagates the language modeling signal backward through the latent updates, while the Jacobian product $\prod_{j=k+1}^{S-1}(I - \eta A^{(j)})$ and the parameter derivative $B^{(k)}$ modulate this signal into updates of the energy function E_ϕ, requiring Jacobian/Hessian–vector products rather than explicit Hessian construction. As the EBM learns to assign lower energy to consistent reasoning trajectories, Langevin calibration iteratively pushes latent thought tokens toward smoother and more stable configurations. Consequently, despite operating solely at the latent embedding level, the EBM enforces global coherence across the generated CoT and improves answer consistency.

C WHY OUR METHOD IS EFFECTIVE?

We emphasize that the effectiveness of `EBM-CoT` does not rely on approximating a target latent distribution or performing asymptotically correct Markov Chain Monte Carlo (MCMC) sampling. Instead, our method leverages the learned *energy landscape* as a *corrective signal* that guides latent reasoning states toward empirically more stable and coherent configurations.

Concretely, the energy function E_ϕ is trained to assign lower energy to latent thought representations that lead to correct and consistent final answers, and higher energy to those associated with inconsistent or unstable reasoning trajectories. As a result, even a small number of Langevin-style refinement steps can effectively reduce local irregularities in the latent reasoning space, acting as a form of *energy-guided denoising* rather than probabilistic sampling.

From this perspective, the short-run Langevin dynamics used in EBM-CoT should be interpreted as a *stochastic refinement operator* that follows the gradient of the learned energy function, rather than as an attempt to draw samples from a stationary distribution. The effectiveness of this procedure stems from the alignment between the energy gradient and the downstream language modeling objective, which together encourage latent reasoning trajectories that are both locally smooth and globally consistent.

Empirically, this refinement mechanism leads to a significant reduction in reasoning variance across sampled chains and substantially improves single-chain (pass@1) accuracy, as demonstrated in our

experiments. These observations suggest that the primary role of the energy-based calibration in EBM-CoT is not distributional approximation, but *variance reduction and trajectory correction in latent reasoning space*.

D Related Work

CoT Reasoning Efficiency Improvement. To improve the efficiency of CoT-based reasoning, researchers have proposed various implicit or compressed forms of CoT. Cheng & Van Durme (2024) introduced Compressed Chain-of-Thought (CCoT), which compresses explicit linguistic reasoning into continuous latent "contemplation tokens," substantially improving inference efficiency. SoftCoT (Xu et al., 2025) further advanced this idea by performing soft chain-of-thought reasoning in the continuous latent space: an auxiliary model generates soft thought tokens that are mapped into the embedding space of the frozen main model, thus improving reasoning efficiency while mitigating catastrophic forgetting. Shen et al. (2025) further proposed the CODI framework, which transfers the reasoning ability of explicit CoT into the continuous latent space through self-distillation. Xia et al. (2025) proposed the TokenSkip framework, which enables controllable Chain-of-Thought (CoT) compression by skipping semantically redundant reasoning tokens. TokenSkip learns to identify and remove low-importance tokens based on token-level importance estimation, thereby reducing CoT length while maintaining reasoning accuracy. Collectively, these methods signify an evolution of reasoning from explicit linguistic chains toward internal continuous representations.

E Algorithm Description

Training Process. During training, the assistant model first generates latent thought tokens from the input, which are projected into the base model's embedding space via a learnable projection network P. The Energy-Based Model (EBM) E_ϕ evaluates these latent tokens by assigning lower energy to task-relevant and consistent thoughts and higher energy to incoherent ones. To refine the latent representations, we perform S Langevin refinement steps with step size η, iteratively updating the latent variable l^c using the gradient of the energy function and Gaussian noise. The refined latent embeddings are then injected into the base model to compute the language modelling loss $\mathcal{L}_{\text{LM}}$. The EBM parameters ϕ and projection P are optimized jointly by minimizing a weighted objective $\mathcal{L} = \mathcal{L}_{\text{LM}} + \alpha\mathcal{L}_{\text{EBM}}$, where α controls the strength of the energy-based regularization. Following our experimental setup, we use $S = 3$ Langevin steps and set $\alpha = 0.1$ to balance both objectives.

Algorithm 1 Training process with EBM Correction

Require: Assistant model p_a, projection P, energy network E_ϕ, step size η, steps S, loss weight α

for each batch (x, y) **do**
  Compute base embeddings $X \leftarrow \text{embed_base}(x)$
  Generate assistant hidden states $H_a \leftarrow p_a(x)$
  Project to base dimension $T \leftarrow P(H_a)$
  Extract latent thought l^ℓ from T
  Initialize corrected latent $l^c \leftarrow l^\ell$
  for $s = 1$ to S **do**
    Compute gradient $g \leftarrow \nabla_l E_\phi(c_t, l^c)$
    Sample noise $\xi_s \sim \mathcal{N}(0, I)$
    Update latent $l^c \leftarrow l^c - \eta g + \sqrt{2\eta}\,\xi_s$
  end for
  Replace latent thought tokens in X with l^c to get X^{mod}
  Forward base model to obtain $\mathcal{L}_{\text{LM}}$
  Compute hinge loss $\mathcal{L}_h = \text{ReLU}(E_\phi(l^c) - E_\phi(l^\ell) + m)$
  Compute consistency loss $\mathcal{L}_c = \lambda\|l^c - l^\ell\|^2$
  $\mathcal{L}_{\text{EBM}} \leftarrow \mathcal{L}_h + \mathcal{L}_c$
  Total loss $\mathcal{L} = \mathcal{L}_{\text{LM}} + \alpha\mathcal{L}_{\text{EBM}}$
  Backpropagate and update ϕ and P
end for

Inference Process. At inference time, the assistant model provides the initial latent thoughts l^ℓ, which are refined through several deterministic Langevin updates (without noise) to obtain the calibrated latents l^c. The updated embeddings are then fed into the frozen base model for final output generation. This process enables consistent reasoning and robust predictions without altering the parameters of either the base or assistant model.

Algorithm 2 Inference process with EBM Correction

1: Given input x, compute base embeddings X
2: Assistant forward $\rightarrow$ latent l^ℓ
3: Initialize $l^c \leftarrow l^\ell$
4: **for** $s = 1$ to S **do**
5: $\quad g \leftarrow \nabla_l E_\phi(c_t, l^c)$
6: $\quad l^c \leftarrow l^c - \eta g$ {(noise optional during inference)}
7: **end for**
8: Replace latent thought tokens in X with l^c

F BASELINES

We compare our approach against three major categories of baselines: (*i*) zero-shot prompting methods, (*ii*) parameter-efficient fine-tuning via LoRA, and (*iii*) implicit chain-of-thought (CoT) reasoning enhancement approaches. This setup allows us to assess the contribution of directional alignment under different reasoning paradigms and adaptation strategies.

Zero-shot prompting baselines. We first evaluate several strong zero-shot prompting methods that require no additional fine-tuning:

- **Zero-Shot CoT** uses the simple prompt "*Let's think step by step*" to elicit multi-step reasoning without any labeled demonstrations. This serves as a standard baseline for evaluating the intrinsic reasoning capability of large language models.
- **Zero-Shot Unk** extends Zero-Shot CoT by introducing uncertainty indicators (e.g., "*I'm not sure, but let's reason carefully*") to encourage more cautious and calibrated reasoning. It provides insight into how uncertainty modulation affects reasoning reliability.
- **Zero-Shot Assist-CoT** supplies auxiliary reasoning traces or scaffolds generated by another model or previous step, assisting the target model in refining its reasoning path. This simulates collaborative or staged reasoning without additional training.

To further improve the robustness of zero-shot reasoning, we also include the **Self-Consistency (SC)** baseline (Wang et al., 2022). SC samples multiple diverse reasoning paths (we set $N = 10$) using temperature-based decoding, and selects the most consistent final answer by majority voting. This approach has been shown to significantly improve reasoning stability and reduce stochasticity in CoT inference.

LoRA fine-tuning baselines. To compare our approach with parameter-efficient adaptation methods, we employ **LoRA** (Hu et al., 2022) fine-tuning. LoRA introduces low-rank trainable adapters into existing linear layers while freezing the majority of model parameters. We fine-tune LoRA adapters on each task-specific dataset using a small learning rate and evaluate the resulting models on the corresponding test splits. This baseline measures the benefits of explicit fine-tuning relative to our merging-based adaptation, which does not require retraining.

Implicit CoT-based reasoning enhancement. Finally, we compare with recent implicit reasoning methods that enhance the internal reasoning ability of language models without explicit step supervision:

- **Coconut** (Hao et al., 2024) learns latent reasoning structures by aligning intermediate reasoning representations with the final answer, encouraging consistency between thought processes and outputs.

- **SoftCoT** (Xu et al., 2025) distills soft reasoning distributions from a teacher model into the student, enabling smoother reasoning trajectories and better generalization across reasoning tasks.

Together, these baselines cover three complementary paradigms: prompt-based reasoning (*Zero-Shot CoT, Self-Consistency*), parameter-efficient fine-tuning (*LoRA*), and implicit reasoning enhancement (*Coconut, SoftCoT*). Our experiments aim to demonstrate that directional alignment provides consistent gains across all these settings while maintaining training-free efficiency.

G DATASETS

We evaluate our proposed method on five publicly available reasoning datasets covering mathematical, commonsense, and symbolic reasoning.

Mathematical reasoning. We use three mathematical reasoning benchmarks: GSM8K (Cobbe et al., 2021), ASDiv-Aug (Miao et al., 2021), and AQuA (Ling et al., 2017). These datasets require arithmetic and algebraic reasoning with varying complexity levels. GSM8K contains 7473 training and 1319 evaluation samples, where each question demands a free-form numerical answer. ASDiv-Aug extends the ASDiv dataset with data augmentation, providing 4183 training and 1038 test samples, also with numerical answers. AQuA includes 97467 training and 254 evaluation questions in a multiple-choice format. Together, these datasets evaluate a model's ability to perform step-by-step quantitative reasoning.

Commonsense reasoning. We evaluate on StrategyQA (Geva et al., 2021), a dataset designed to test implicit multi-hop commonsense reasoning. Each question requires combining multiple unspoken facts to produce a binary yes/no answer. It consists of 1832 training and 458 evaluation samples, challenging models to perform compositional and non-explicit reasoning.

Symbolic reasoning. Finally, we adopt the Date Understanding (DU) task from BIG-Bench (Srivastava et al., 2023). DU evaluates a model's ability to resolve temporal relationships expressed in natural language. It is a purely evaluation-only benchmark with 369 samples and no training split available. Each question provides several options, requiring symbolic comparison and logical deduction.

H DISCUSSION ON WALL-CLOCK INFERENCE LATENCY

Tables 4, 5 and 6 report wall-clock inference latency comparisons between Self-Consistency ($N = 10$) and `EBM-CoT` across multiple datasets and model pairs under identical decoding settings on a single GPU. We observe that `EBM-CoT` with a small number of correction steps ($S = 3$) consistently achieves substantial latency reductions compared to conservative serial SC with $N = 10$, with speedups ranging from approximately $5\times$ to $19\times$ depending on the dataset and base/assistant model combination. This improvement stems from a fundamental difference in computation patterns: SC requires repeating the full autoregressive decoding process multiple times, whereas `EBM-CoT` performs a limited number of inexpensive energy-based correction steps on short thought representations generated by a smaller assistant model, followed by only a single final decoding pass with the base model. As a result, the total `EBM-CoT` inference time, which consists of the correction cost plus one generation, is significantly lower than the cumulative cost of ten full generations in SC, especially for tasks with long output sequences such as AQuA. We further note that the magnitude of speedup varies across datasets and models, reflecting differences in generation length, model architecture, and decoding overhead, highlighting the importance of reporting wall-clock measurements rather than theoretical complexity alone. Overall, these results demonstrate that `EBM-CoT` offers a favorable time–accuracy trade-off for inference, achieving large practical latency savings without requiring multiple full forward generations as in Self-Consistency.

Table 4: Wall-clock inference latency comparison across datasets using Llama-3.1-8B-Instruct as the base model with Llama-3.2-1B-Instruct as the assistant model. All results are measured on a single GPU with identical decoding settings. Self-Consistency ($N = 10$) is measured in a conservative serial setting.

Dataset	Method	N	S	Total Time (s)	Speedup
GSM8K	SC ($N = 10$)	10	–	62.20	1.0×
	EBM-CoT	1	3	12.42	5.01×
AQuA	SC ($N = 10$)	10	–	102.31	1.0×
	EBM-CoT	1	3	16.13	6.34×
ASDiv-Aug	SC ($N = 10$)	10	–	41.21	1.0×
	EBM-CoT	1	3	3.10	13.29×

Table 5: Wall-clock inference latency comparison across datasets using Qwen3-8B as the base model with Qwen3-0.6B as the assistant model. All results are measured on a single GPU with identical decoding settings. Self-Consistency ($N = 10$) is measured in a conservative serial setting.

Dataset	Method	N	S	Total Time (s)	Speedup
GSM8K	SC ($N = 10$)	10	–	89.44	1.0×
	EBM-CoT	1	3	9.40	9.51×
AQuA	SC ($N = 10$)	10	–	430.31	1.0×
	EBM-CoT	1	3	22.81	18.86×
ASDiv-Aug	SC ($N = 10$)	10	–	34.00	1.0×
	EBM-CoT	1	3	4.05	8.40×

Table 6: Wall-clock inference latency comparison across datasets using Qwen2.5-7B-Instruct as the base model with Qwen2.5-1.5B-Instruct as the assistant model. All results are measured on a single GPU with identical decoding settings. Self-Consistency ($N = 10$) is measured in a conservative serial setting.

Dataset	Method	N	S	Total Time (s)	Speedup
GSM8K	SC ($N = 10$)	10	–	73.34	1.0×
	EBM-CoT	1	3	8.23	8.91×
AQuA	SC ($N = 10$)	10	–	110.76	1.0×
	EBM-CoT	1	3	10.07	11.00×
ASDiv-Aug	SC ($N = 10$)	10	–	25.63	1.0×
	EBM-CoT	1	3	2.56	10.01×

I Discussion on Qwen2.5-7B-Instruct

Table 7 summarizes the model performance on five reasoning benchmarks using Qwen2.5-7B-Instruct as base model, Qwen2.5-1.5B-Instruct as assistant model. Our proposed method consistently outperforms all baselines across diverse reasoning domains, including mathematical (GSM8K, ASDiv-A) and commonsense (StrategyQA) tasks. Compared with SoftCoT, which relies solely on continuous representation mapping, our energy-based calibration achieves an average improvement of +3.8%, demonstrating that explicitly refining the latent thought tokens leads to more stable and coherent reasoning trajectories.

The self-consistency variant (Ours (N=10)) further amplifies this gain, achieving the highest overall average accuracy of 82.73%. By sampling multiple calibrated reasoning paths and aggregating their outcomes, the model effectively reduces stochastic variance in the latent reasoning space. This indicates that the proposed energy-based refinement complements self-consistency sampling by

Table 7: Model performance using Qwen2.5-7B-Instruct.

Model	GSM8K	ASDiv-Aug Mathematical	AQuA	StrategyQA Commonsense	DU Symbolic	Avg.
Zero-Shot CoT	$83.70_{\pm 0.78}$	$87.19_{\pm 0.28}$	$64.53_{\pm 3.27}$	$49.65_{\pm 3.18}$	$66.40_{\pm 2.26}$	70.29
Zero-Shot CoT-Unk	$84.12_{\pm 0.71}$	$86.94_{\pm 0.89}$	$64.72_{\pm 2.06}$	$50.74_{\pm 1.90}$	$66.48_{\pm 1.43}$	70.60
Zero-Shot Assist-CoT	$84.85_{\pm 0.35}$	$88.63_{\pm 1.05}$	$64.96_{\pm 2.83}$	$52.71_{\pm 2.65}$	$67.04_{\pm 2.84}$	71.64
LoRA Fine-Tuning	$81.80_{\pm 0.00}$	$86.80_{\pm 0.00}$	$62.60_{\pm 0.00}$	-	-	-
Coconut	$82.49_{\pm 0.00}$	$86.90_{\pm 0.00}$	$63.39_{\pm 0.00}$	-	-	-
SoftCoT	$85.81_{\pm 1.82}$	$88.90_{\pm 1.01}$	$72.44_{\pm 2.19}$	$60.61_{\pm 1.55}$	$67.52_{\pm 2.92}$	75.06
Ours	$\mathbf{87.04}_{\pm 0.23}$	$\mathbf{90.70}_{\pm 0.24}$	$\mathbf{75.59}_{\pm 0.96}$	$\mathbf{67.47}_{\pm 0.22}$	$\mathbf{73.37}_{\pm 1.02}$	**78.83**
Ours ($N = 10$)	$92.75_{\pm 0.04}$	$91.56_{\pm 0.09}$	$83.37_{\pm 0.59}$	$69.28_{\pm 0.10}$	$76.69_{\pm 0.39}$	82.73

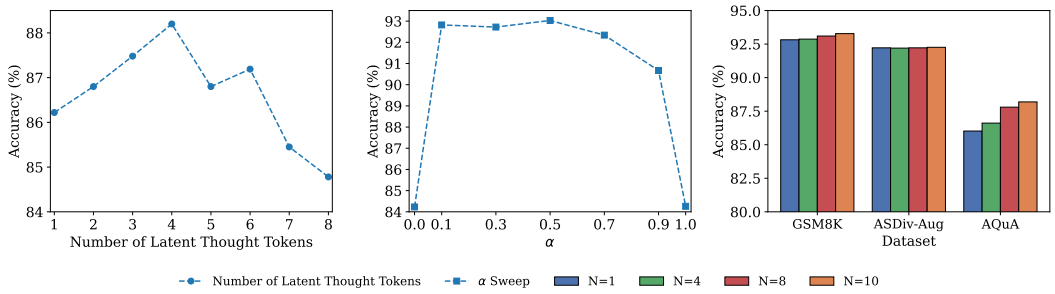


Figure 5: **Left Figure:** Results on ASDiv-Aug using Qwen2.5-7B-Instruct as the base model and Qwen2.5-1.5B-Instruct as the assistant model, with varying numbers of latent thought tokens. **Middle Figure:** Results on GSM8K using Qwen3-8B as the base model and Qwen3-0.6B as the assistant model, where α controls the relative strength of the energy-based regularization term $\mathcal{L}_{\text{EBM}}$. **Right Figure:** Comparison across different numbers of reasoning chains N (pass@N) using Qwen3-8B as the base model. While larger N slightly improves performance due to answer aggregation under the self-consistency setting, our model already achieves strong single-chain accuracy ($N = 1$), highlighting that energy-based calibration substantially enhances reasoning consistency without relying on multi-chain sampling.

aligning latent trajectories toward lower-energy, high-consistency regions. In contrast, prior methods such as SoftCoT or LoRA fine-tuning lack such global regularization, leading to greater variation in reasoning quality across sampled chains.

Notably, the largest gains appear on symbolic and commonsense datasets (DU, StrategyQA), suggesting that our calibration mechanism generalizes beyond arithmetic reasoning and enhances structural consistency in broader reasoning contexts. Overall, these results validate the effectiveness of our approach in improving both the accuracy and stability of multi-step reasoning in large language models.

J Ablation Study

Figure 5 presents the sensitivity analysis of our proposed framework with respect to the number of latent thought tokens, the EBM weighting coefficient α, and the number of reasoning chains N. On the left, we observe that increasing the number of latent thought tokens initially improves accuracy, as more latent steps allow the model to capture finer-grained reasoning patterns in the embedding space. However, performance saturates and then declines beyond four tokens, implying that excessive latent reasoning introduces unstable optimization, which can distort the energy landscape and hinder convergence. In the middle, we examine the influence of α, which balances the language modeling loss $\mathcal{L}_{\text{LM}}$ and the energy-based term $\mathcal{L}_{\text{EBM}}$. A moderate value of α (around 0.5) yields the highest accuracy, indicating that partial energy regularization is sufficient to guide consistent reasoning without overpowering the linguistic prior of the base model. When α approaches 0 or 1, performance drops sharply — either due to insufficient calibration or excessive energy dominance. On the right, we

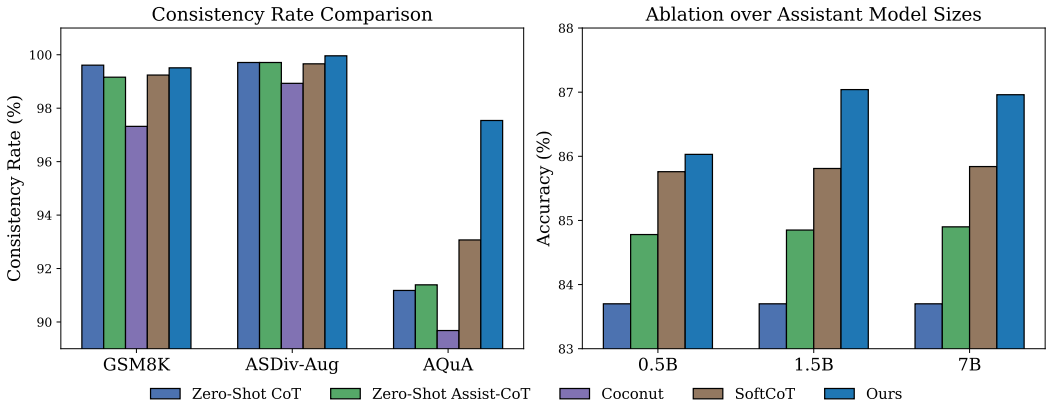


Figure 6: **Left Figure:** Consistency rate comparison using Qwen3-8B as the base model and Qwen3-0.6B as the assistant model. **Right Figure:** Ablation over assistant model sizes on GSM8K using Qwen2.5-7B-Instruct as the base model.

compare the effect of different numbers of reasoning chains N (pass@N) under the self-consistency setting. Although accuracy improves slightly with larger N, our method already achieves competitive results at $N = 1$, demonstrating that the proposed energy-based calibration substantially enhances single-chain consistency, reducing the need for costly multi-chain aggregation. These findings collectively highlight that both the number of latent thought tokens and the strength of energy calibration play critical roles in shaping stable and efficient reasoning trajectories, while the model's high single-chain accuracy reflects its intrinsic reasoning consistency.

Figure 6 presents an ablation study that evaluates the consistency and scalability of the proposed energy-based calibration framework. On the left, we assess the consistency rate using Qwen3-8B as the base model and Qwen3-0.6B as the assistant model. Our approach achieves nearly perfect consistency across all datasets, surpassing SoftCoT and other baselines by a large margin. This confirms that correcting latent thought tokens through the energy model effectively mitigates stochastic deviations in implicit reasoning, leading to more stable multi-step reasoning outcomes. On the right, we conduct an ablation over different assistant model sizes using Qwen2.5-7B-Instruct on GSM8K. Performance steadily improves as the assistant model grows from 0.5B to 7B parameters, reflecting that stronger assistants provide more informative latent priors. Nevertheless, our method consistently achieves the best results regardless of model scale, demonstrating that the proposed calibration is model-agnostic and robust even with lightweight assistants. These results highlight that our approach improves reasoning consistency while maintaining scalability across heterogeneous model configurations.

K Ablation study on the number of EBM correction steps

Figure 3(a) presents an ablation study on the number of EBM Correction steps S evaluated on the ASDiv-Aug dataset. We observe that increasing the number of correction steps consistently improves performance up to $S = 3$, after which the gain becomes marginal. In particular, setting $S = 3$ achieves a strong accuracy of 88.20%, which is close to the best performance obtained with more steps. Further increasing S yields only limited improvements while introducing additional computational cost. Therefore, we choose $S = 3$ as a good trade-off between performance and efficiency, as it enables effective correction without requiring excessive iterative updates.

L Ablation study on the EBM correction step size

Figure 3(b) presents an ablation study on the step size η used in the EBM correction process, which controls the update magnitude of the latent thought during the iterative energy-based refinement. Specifically, we evaluate $\eta \in \{0.01, 0.1, 1\}$ on the ASDiv-Aug dataset.

The results show that a moderate step size ($\eta = 0.1$) achieves the best performance, while both smaller and larger values lead to inferior results, indicating that η plays a critical role in balancing effective correction and stable optimization in the latent space. When the step size is too small

($\eta = 0.01$), the gradient-based updates on the latent thought become overly conservative. Although the refinement process remains stable, the EBM fails to sufficiently adjust the latent representation within a limited number of correction steps S, resulting in insufficient correction of suboptimal latent thoughts. In contrast, a large step size ($\eta = 1$) leads to aggressive updates that may overshoot low-energy regions in the energy landscape. Such behavior disrupts the smooth exploration of the latent space and can push the latent thought away from semantically consistent regions, thereby introducing instability in the correction process and degrading overall performance.

Overall, these results demonstrate that an appropriate step size is essential for enabling the EBM to effectively guide latent thoughts toward lower-energy and more consistent representations.

M Ablation study on the λ

We analyze the effect of the consistency regularization weight λ, which controls how strongly the EBM-refined latent representation l^c is constrained to remain close to the assistant-provided latent l^ℓ. As shown in Figure 3(c), performance exhibits a clear unimodal trend with respect to λ.

When $\lambda = 0$, the correction process is unconstrained, allowing the Langevin dynamics to freely modify the latent representation. While this setting provides maximal flexibility, it can push l^c away from the base model's embedding manifold, resulting in reduced downstream accuracy. Introducing a moderate consistency constraint significantly improves performance, with $\lambda = 0.01$ achieving the highest accuracy.

In contrast, larger values of λ overly restrict the correction process. When $\lambda \geq 0.1$, the refined latent is forced to stay close to l^ℓ, limiting the expressive capacity of the energy-based refinement and preventing effective correction of suboptimal latent states. In this regime, the EBM acts only as a weak perturbation, leading to degraded performance.

Overall, these results demonstrate the importance of balancing correction strength and stability. A moderate consistency weight enables effective local refinement of latent representations while preserving semantic alignment with the original assistant latent. Based on this analysis, we set $\lambda = 0.01$ for all experiments.

N Energy-Guided Latent Refinement Analysis

Table 3 provides a geometric characterization of the latent modifications induced by EBM calibration across multiple reasoning benchmarks. While the overall magnitude of the latent displacement remains small, the updates exhibit a highly consistent structure: (i) the pooled L2 distance shows a narrow unimodal distribution, indicating stable, non-stochastic refinement; (ii) the maximum token-level L2 distance is consistently larger than the pooled displacement, revealing that the updates are localized to a small subset of latent tokens rather than uniformly distributed; and (iii) the correction directions are strongly aligned across samples, with cosine similarity to the mean direction exceeding 0.87 on all evaluated datasets.

Importantly, we also observe a consistent decrease in the learned energy function after calibration, confirming that these directional and localized updates correspond to movement toward lower-energy regions of the latent space. Together, these findings rule out an interpretation of EBM calibration as isotropic denoising. Instead, they support a refinement perspective, where the energy function selectively reshapes latent representations by applying targeted, directionally consistent corrections to salient reasoning tokens. This structured refinement provides a mechanistic explanation for the robustness and consistency gains observed in downstream reasoning performance.

O One Example

Question. he wanted to make sure that he is protected from the cold evenings in the forest so he decided to build a fireplace made of cement . if he bought 216 lbs of cement and his son brought another 411 lbs how much cement did he have originally if he now has 452 lbs ?

Ground Truth Answer. $\boxed{-175}$

Model Reasoning (SoftCoT). The total amount of cement now is 452 lbs. His son brought an additional 411 lbs. So, to find out how much cement he originally had, we subtract the amount his son brought from the total amount:

$$452 \text{ lbs} - 411 \text{ lbs} = 41 \text{ lbs}$$

Therefore, the final answer is: $\boxed{41}$ lbs.

Predicted Action (SoftCoT). Incorrect.

Model Reasoning (`EBM-CoT`). Simplify the equation Combine the constants on the left side:

$$x + 627 = 452$$

Solve for x Subtract 627 from both sides to isolate x:

$$x = 452 - 627$$

$$x = -175$$

This result doesn't make sense in the context of the problem because the amount of cement cannot be negative. Let's recheck the problem statement and our setup:

The correct setup should be:

$$x + 216 + 411 = 452$$

$$x + 627 = 452$$

$$x = 452 - 627$$

$$x = -175$$

It appears there might be a mistake in the problem statement or the given final amount (452 lbs). Given the logical steps, it seems the person would have less than 0 lbs if they started with a negative amount, which is not possible.

Therefore, based on the provided information, the only logical conclusion is that the initial amount x must have been negative, which isn't feasible. However, mathematically solving the problem as stated, the answer would be:

Therefore, the final answer is: $\boxed{-175}$.

Predicted Action (`EBM-CoT`). Correct.

P Impact Statement

This work introduces an energy-based calibration approach for implicit chain-of-thought reasoning, aiming to improve the consistency and reliability of large language models on tasks such as mathematical, commonsense, and symbolic reasoning while reducing reliance on expensive multi-sample self-consistency voting. The work may yield positive societal impacts by enabling more predictable reasoning in applications like educational tutoring, information assistance, and automated analysis, thereby reducing errors caused by sampling variance or reasoning drift. However, the approach also poses potential risks: improved and more stable reasoning could be misused to produce more persuasive but incorrect explanations, automate deceptive content, or scale flawed decision recommendations; additionally, implicit reasoning can reduce transparency of intermediate steps, making error attribution and accountability more challenging. To mitigate these risks, future deployments should incorporate output verification (e.g., executable validators, fact-checking), uncertainty signaling, and cautious use policies for high-stakes settings, alongside evaluations that measure robustness and misuse-related harms.

Q AI USE STATEMENT

In preparing this paper, we used large language models (LLMs) solely as an assistive tool for language polishing and minor writing improvements. The models were not involved in research ideation, experimental design, data analysis, or drawing scientific conclusions. All conceptual and technical contributions are the work of the authors. The authors take full responsibility for the contents of this paper. For experimental evaluation, we additionally conducted testing using large language models such as Qwen and Llama families as base models.

R ETHICS STATEMENT

This work focuses on improving the consistency and reliability of implicit chain-of-thought reasoning in large language models. All experiments are conducted on publicly available reasoning benchmarks, including GSM8K, ASDiv-Aug, AQuA, StrategyQA, and Date Understanding. These datasets do not contain personally identifiable information, and our study does not involve human subjects, crowdsourcing, or the collection of new user data. We therefore did not require additional ethical approval for data collection.

In preparing this paper, large language models were used only as an assistive tool for language polishing and minor writing improvements. They were not involved in research ideation, experimental design, data analysis, or scientific conclusions. The authors take full responsibility for the content of this work.

We acknowledge that more stable and consistent reasoning capabilities may be misused. For example, improved reasoning could be used to generate more persuasive but incorrect explanations, automate deceptive content, or support high-stakes decisions without adequate verification. Implicit reasoning may also reduce the transparency of intermediate steps, making error attribution more difficult. To mitigate these risks, we recommend that deployments of this method incorporate output verification, fact-checking or executable validators, uncertainty signaling, and cautious use policies in high-stakes settings. Our method is intended as a research contribution toward more reliable reasoning, not as a ready-to-deploy decision-making system.

S REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the anonymous source code at `https://anonymous.4open.science/r/EBM-CoT-BEDE`. All datasets used in this work are publicly available, and all base and assistant models are from publicly released model families, including LLaMA-3.1-8B-Instruct, LLaMA-3.2-1B-Instruct, Qwen3-8B, Qwen3-0.6B, Qwen2.5-7B-Instruct, and Qwen2.5-1.5B-Instruct. We follow the dataset configurations and baseline settings adopted in prior work on SoftCoT.

The training and inference details are described in Section 4, Section 5, and the appendices. During training, the number of latent thought tokens is set to 32, the model is trained for 10 epochs with a batch size of 4, and the Langevin calibration uses $S = 3$ refinement steps with $\alpha = 0.1$. During inference, the main results use 4 latent thought tokens, and the EBM performs a small number of deterministic Langevin steps. For the energy function, we use a multi-layer perceptron following prior work. The main hyperparameters are $S = 3$, $\eta = 0.1$, and $\lambda = 0.01$, with ablation studies provided in Section 5.4 and Appendix J. Algorithm 1 and Algorithm 2 summarize the training and inference procedures. Theoretical derivations are provided in Appendix B, and additional experimental details, baselines, and dataset descriptions are given in Appendices F–I. All experiments are implemented in PyTorch and run on a local Linux server with an AMD EPYC 7742 64-Core Processor and NVIDIA A100-SXM4-80GB GPUs. Where applicable, we report mean and standard deviation over evaluation runs. We will release the full implementation and configuration files upon publication.

T DEVICES

In the experiments, we conduct all methods on a local Linux server equipped with one AMD EPYC 7742 64-Core Processor (128 logical threads). All methods are implemented using the

PyTorch framework, and all models are trained on NVIDIA A100-SXM4-80GB GPUs (80GB HBM2 memory).