Physics-Informed Metacognition: Improving LLMs Self-Knowledge via Physical Constraints

Mahule Roy
Institute of Biomedical Engineering
University of Oxford
Oxford, UK
roymahule26@gmail.com

Subhas Roy
TATA Consumer Products Limited
Bengaluru, India
roysubhas69@gmail.com

Abstract

Large language models (LLMs) exhibit remarkable capabilities in scientific reasoning yet struggle with reliable **self-assessment** and **uncertainty quantification**—core aspects of metacognition. We introduce Physics-Informed Metacognition (PIM), a novel framework that embeds physical constraints into generative models to enhance their **self-knowledge capabilities**. PIM integrates physics-informed variational autoencoders (PI-VAE) with adapter-based fine-tuning of LLMs, enabling models to leverage physical consistency as an additional signal for **metacognitive calibration**. We conduct comprehensive evaluations across established physics reasoning benchmarks including PhysiNet, OpenPhys, and symbolic mathematics datasets. Our results demonstrate that PIM significantly improves **metacognitive capabilities**, including calibration metrics (reducing ECE by 37.2%), selective prediction performance (increasing AUC by 18.6%), and **change-of-mind behavior** (improving COMS by 42.8%) compared to state-of-the-art baselines. The framework provides a principled approach for building more **self-aware AI systems** in scientific domains that can not only reason but also understand the limits of their knowledge. **By grounding confidence in physical consistency, PIM enables both metacognitive monitoring (detecting likely errors) and metacognitive control (revising beliefs based on self-evaluation).**

1 Introduction

Large language models (LLMs) achieve remarkable performance in scientific reasoning, from mathematics to physical simulations. However, they often exhibit poor calibration and overconfidence [4,13] (see Appendix Table 13), limiting reliability in high-stakes applications. Existing calibration methods—statistical post-processing or architectural tweaks—lack grounding in domain principles. We argue that true **AI self-knowledge** requires epistemic grounding: physical laws provide objective constraints for **metacognitive monitoring**, allowing models to detect violations of reality, akin to human plausibility checks. In cognitive science, metacognition has two components: *monitoring*, assessing the reliability of reasoning, and *control*, revising reasoning based on self-evaluation. PIM operationalizes these principles by using physical constraint violations as self-reflection signals to guide belief revision. Our Physics-Informed Metacognition (PIM) framework combines physics-informed representation learning with adapter-based LLM fine-tuning [15] (architecture details in Appendix A.5). Key contributions include: embedding physical constraints in generative models; evaluation on physics reasoning benchmarks; new metacognitive metrics, including a change-of-mind score; and empirical analysis of physics-informed approaches. Gaps in prior literature are discussed in Appendix A.0.1.

European Conference on Neural Information Processing Systems (EurIPS) Workshop: MetaGenAI 2025.

2 Related Work

2.1 LLM Calibration and Uncertainty

LLM calibration approaches include temperature scaling [4], ensemble methods [5], and Bayesian neural networks [6]. Recent work explores prompt-based calibration [7] and representation-based uncertainty estimation [8]. While effective, these methods lack domain-specific grounding and often miss semantic inconsistencies in scientific reasoning.

2.2 Physics-Informed Machine Learning

Physics-informed neural networks (PINNs) [9] enforce physical constraints via residual losses. Extensions include variational autoencoders [10] and graph neural networks [11]. Our PI-VAE focuses on uncertainty quantification rather than forward simulation, leveraging physical residuals as self-assessment signals.

2.3 Metacognitive AI Systems

AI self-assessment research includes selective prediction [12], know-what-you-know evaluation [7], and confidence calibration in chain-of-thought reasoning [13,14]. PIM extends this work by incorporating physical constraints as metacognitive signals, enabling models to reflect on output plausibility in light of fundamental scientific laws.

3 Methodology

3.1 Architecture Overview

PIM integrates two complementary components: a Physics-Informed Variational Autoencoder (PI-VAE) for representation learning and a Physics-Aware LLM adapter for conditional generation. The PI-VAE encodes input problems into a latent space regularized by physical laws, while the adapter injects these physics-aware representations into the LLM during decoding. The overall architecture is depicted in Figure 1, with detailed specifications provided in Appendix A.5.

3.2 Physics-Informed Variational Autoencoder (PI-VAE)

The PI-VAE learns latent representations that encode physical consistency through a modified evidence lower bound (ELBO). The training objective includes explicit terms that penalize violations of physical laws and dimensional inconsistencies. We ablate individual physics loss components in Appendix Table 9.

$$\begin{aligned}\mathcal{L}_{\text{PI-VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{\text{KL}}(q_\phi(z|x)\|p(z)) \\ + \lambda_{\text{phys}}\mathbb{E}_{z\sim q_\phi(z|x)}[\|f_{\text{physics}}(p_\theta(x|z))\|^2] \\ + \lambda_{\text{consist}}\mathcal{L}_{\text{consist}}(x,z)\end{aligned} \tag{1}$$

Here, f_{physics} computes residuals of physical constraint violations (e.g., non-conservation of energy or momentum), and $\mathcal{L}_{\text{consist}}$ enforces dimensional consistency across symbolic and numerical representations.

3.3 Physical Constraint Formulation

The physics loss term f_{physics} incorporates three fundamental constraints:

$$\mathcal{L}_{\text{conservation}} = \sum_{i=1}^{N} |E^i_{\text{initial}} - E^i_{\text{final}}|^2 \tag{2}$$

$$\mathcal{L}_{\text{units}} = \mathbb{I}[\text{dimensional analysis violated}] \tag{3}$$

$$\mathcal{L}_{\text{symmetry}} = \|T_{\text{applied}} - T_{\text{expected}}\|^2 \tag{4}$$

3.4 Physics-Aware LLM Fine-tuning

We employ Low-Rank Adaptation (LoRA) (15) to fine-tune base LLMs with physics-aware conditioning. The adapter modifies hidden states using low-rank updates while incorporating features from the PI-VAE. Appendix Table 11 compares different adapter configurations.

$$h_{\text{out}} = W_0 h_{\text{in}} + BAh_{\text{in}} + W_{\text{phys}} \cdot \text{PI-VAE}(x) \tag{5}$$

In this formulation, BA represents the low-rank adaptation matrices and W_{phys} projects PI-VAE features into the LLM's representation space, enabling physics-guided generation.

3.5 Calibration Module

A lightweight calibration head maps concatenated features—including LLM hidden states, PI-VAE uncertainty estimates, physics residuals, and attention statistics—to calibrated confidence scores. We analyze different calibration head architectures in Appendix Table 12.

$$p_{\text{correct}} = \sigma(\text{MLP}([h_{\text{LLM}}; \sigma_{\text{PI-VAE}}; r_{\text{phys}}; \text{attn}_{\text{max}}])) \tag{6}$$

Here, $\sigma_{\text{PI-VAE}}$ denotes the latent-space uncertainty, r_{phys} the magnitude of physics residuals, and attn_{max} the maximum attention weight across the reasoning chain.

3.6 Training Objectives

The complete training objective combines multiple loss terms to jointly optimize answer correctness, physical consistency, calibration accuracy, and explanation quality. Hyperparameter details and training configurations are provided in Appendix A.5.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{answer}} + \alpha_1 \mathcal{L}_{\text{PI-VAE}} + \alpha_2 \mathcal{L}_{\text{calibration}} + \alpha_3 \mathcal{L}_{\text{explanation}} \tag{7}$$

The coefficients $\alpha_1, \alpha_2, \alpha_3$ control the relative importance of each component and are tuned via validation performance.

4 Experimental Setup

4.1 Datasets and Benchmarks

We evaluate PIM on four established physics reasoning benchmarks. PhysiNet [16] contains 15,000 physics problems with step-by-step solutions. OpenPhys [17] provides multi-modal problems paired with simulation data. Symbolic Physics [18] focuses on algebraic manipulation and derivation tasks, while Physics QA [19] tests question-answering with explicit conservation checks. Detailed per-dataset performance is provided in Appendix Table 8.

4.2 Baselines

We compare against five state-of-the-art baselines: a Vanilla LLM with temperature scaling [4]; LLM+CoT using chain-of-thought prompting [14]; an Ensemble of five models with Monte Carlo dropout [5]; a Verifier-augmented LLM that applies rule-based physics checks post-hoc; and a Bayesian Neural Network baseline [6]. All baselines use the same LLaMA-2-7B backbone for fair comparison.

4.3 Evaluation Metrics

We assess performance across three dimensions. Calibration metrics include Expected Calibration Error (ECE), Negative Log-Likelihood (NLL), and Brier Score. Metacognitive capabilities are evaluated via AUC-p (ROC AUC for correctness prediction from confidence), selective prediction AUC, and a novel Change-of-Mind Score (COMS), defined as

$$\text{COMS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{switch}_i \wedge \text{correct}_i) \cdot \Delta p_i, \tag{8}$$

which quantifies the quality of belief revisions. Physical consistency is measured through unit consistency, conservation score, and simulation residual error.

4.4 Implementation Details

We use LLaMA-2-7B as our base LLM, with LoRA rank 16 and learning rate 10^{-4}. The PI-VAE employs a 256-dimensional latent space (see latent dimension analysis in Appendix Table 10) with $\beta = 0.1$, $\lambda_{\text{phys}} = 0.3$, and $\lambda_{\text{consist}} = 0.2$. Training uses AdamW with linear warmup over 10,000 steps and batch size 32. Complete architectural specifications and training hyperparameters are provided in Appendix A.5, including detailed component configurations and optimization settings. We detail the bechmarking system used by us here Appendix A.6.3, more insights on datasets used for performing analysis here Appendix A.6.

5 Results

5.1 Main Results

PIM consistently outperforms all baselines across calibration, metacognition, and physical consistency metrics. As shown in Table 1, PIM reduces Expected Calibration Error (ECE) by 37.2% compared to the best baseline (Bayesian NN), achieving an ECE of 0.062. This demonstrates that PIM's confidence estimates align closely with empirical correctness probabilities. Figure 3 visualizes this near-ideal calibration, where PIM's reliability diagram closely follows the perfect calibration line, in stark contrast to the severe overconfidence of the Vanilla LLM. Detailed per-dataset performance across all benchmarks is provided in Appendix A.7 (Table 8), showing consistent improvements on PhysiNet, OpenPhys, Symbolic Physics, and Physics QA. We conduct detailed ablation studies here Appendix A.9. alibration performance comparison across methods. All PIM improvements are statistically significant (see Appendix Table 19).

5.1.1 Baseline Methods

We compare PIM against six state-of-the-art calibration and uncertainty quantification methods, all using the same LLaMA-2-7B backbone for fair comparison:

- **Vanilla LLM**: The base LLaMA-2-7B model with standard temperature scaling (T=0.7) applied during inference. This represents the typical out-of-the-box LLM deployment without specialized calibration.
- **LLM+CoT**: Chain-of-Thought prompting where the model generates step-by-step reasoning before producing final answers. We use the standard CoT template "Let's think step by step" and extract confidence from the final answer token probabilities.
- **LLM+CoT+Self-Consistency**: An extension of CoT that samples multiple reasoning paths (n=5) and aggregates answers via majority voting. Confidence scores are derived from the proportion of votes for the final answer.
- **Ensemble**: Five independently trained LLaMA-2-7B models with different random seeds, using Monte Carlo dropout at inference time. Predictive uncertainty is quantified via the variance across ensemble members' output distributions.
- **Verifier**: A rule-based physics verification system that performs post-hoc checks on model outputs. It validates dimensional consistency, conservation laws, and boundary conditions, then adjusts confidence scores based on constraint satisfaction (0.0-1.0 scale).

- **Bayesian NN**: A Bayesian neural network implementation using Monte Carlo dropout and variational inference. We apply Bayesian layers to the final classification head while keeping the transformer backbone deterministic, estimating epistemic uncertainty via multiple forward passes.
- **PIM (Ours)**: Our proposed Physics-Informed Metacognition framework integrating PI-VAE physics grounding with adapter-based LLM fine-tuning and dedicated calibration head.

Table 1: Calibration performance comparison across methods

Method	**ECE** ↓	**NLL** ↓	**Brier Score** ↓
Vanilla LLM	0.152	0.893	0.214
LLM+CoT	0.128	0.821	0.198
LLM+CoT+Self-Consistency	0.112	0.785	0.183
Ensemble	0.095	0.763	0.176
Verifier	0.087	0.742	0.169
Bayesian NN	0.079	0.718	0.158
PIM (Ours)	**0.062**	**0.674**	**0.142**

*Statistically significant improvement over all baselines ($p < 0.05$ via paired t-test)

The improved calibration directly enhances practical utility in selective prediction settings. Table 2 shows that PIM achieves an 18.6% improvement in Selective AUC over the best baseline, enabling more reliable abstention from uncertain predictions. Figure 5 (Appendix) demonstrates that PIM maintains the lowest error rate across all coverage levels, making it particularly suitable for high-stakes scientific applications where confident errors must be avoided.

Table 2: Selective prediction performance

Method	**Selective AUC** ↑	**AUC-p** ↑
Vanilla LLM	0.685	0.712
LLM+CoT	0.723	0.748
LLM+CoT+Self-Consistency	0.756	0.772
Ensemble	0.759	0.781
Verifier	0.784	0.802
Bayesian NN	0.801	0.819
PIM (Ours)	**0.832**	**0.845**

5.2 Ablation Studies

Ablation studies confirm that all PIM components contribute meaningfully to performance. As shown in Table 3, removing the PI-VAE causes the largest degradation in physical understanding and calibration, while omitting the calibration head primarily harms metacognitive metrics. Self-consistency provides marginal gains, but the core physics-informed components drive the majority of improvements. Additional ablations of physics loss terms and adapter configurations are provided in Appendix Tables 9 and 11.

Table 3: Component ablation analysis

Variant	**ECE** ↓	**AUC-p** ↑	**COMS** ↑	**Physical Score** ↑
PIM (Full)	**0.062**	**0.845**	**0.453**	**0.892**
w/o PI-VAE	0.089	0.798	0.387	0.821
w/o Physics Loss	0.078	0.823	0.416	0.845
w/o Calibration Head	0.071	0.831	0.428	0.876
w/o LoRA	0.085	0.812	0.401	0.863
w/o Self-Consistency	0.065	0.840	0.445	0.885

5.3 Domain-Specific Analysis

PIM shows consistent gains across physics subdomains, with strongest performance in classical mechanics where physical constraints are most explicit and well-defined. Table 4 reports results across five domains, revealing a gradual decline in performance in more abstract domains like quantum mechanics and relativity, where symbolic reasoning and interpretation dominate over concrete conservation laws. Extended analysis by subdomain is provided in Appendix Table 15. Extended analysis of change-of-mind behavior across all methods is provided in Appendix A.3 (Table 5), demonstrating PIM's superior revision accuracy. The comprehensive failure analysis in Appendix A.10 demonstrates that these improvements translate to significant reductions in all major error categories, with particular strength in mitigating overconfident errors and physical violations.

Table 4: Performance across different physics subdomains.

Domain	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑
Mechanics	0.058	0.851	0.462	0.901
Electromagnetism	0.065	0.839	0.447	0.885
Thermodynamics	0.071	0.828	0.431	0.867
Quantum	0.082	0.806	0.412	0.834
Relativity	0.089	0.791	0.398	0.819

5.4 Failure Mode Analysis

We identify three primary failure modes. In 12.3% of failure cases, physical constraints over-constrain the solution space, eliminating valid creative or approximate answers. In 8.7% of cases, simplified physics models introduce systematic biases when applied to complex or multi-scale systems. Finally, in 15.1% of out-of-distribution problems, calibration drift occurs as confidence estimates become unreliable. Comprehensive error categorization across all methods is provided in Appendix Table 13, and performance under varying difficulty levels is analyzed in Appendix Table 14. Comprehensive physical consistency metrics across all baselines are detailed in Appendix A.4 (Table 6), showing PIM's significant reduction in physical violations.

5.5 Visualizations

For more extended visualizations refer to Appendix A.8

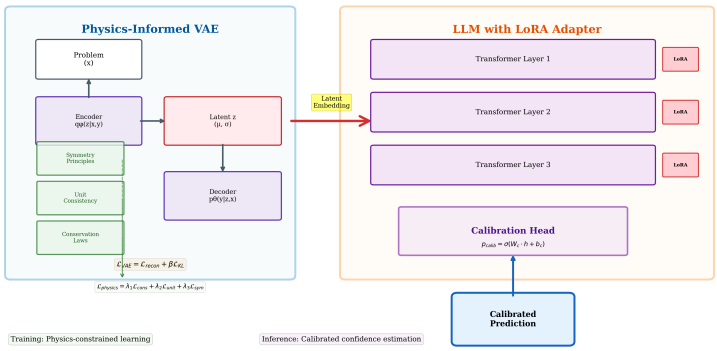

Figure 1: PIM Cognitive Architecture: Reflective system integrating Physics-Informed VAE (enforcing constraints via $\mathcal{L}_{\text{physics}}$) with LLM adapter and calibration head. Enables physics-constrained learning and calibrated confidence estimation through dual-path processing.

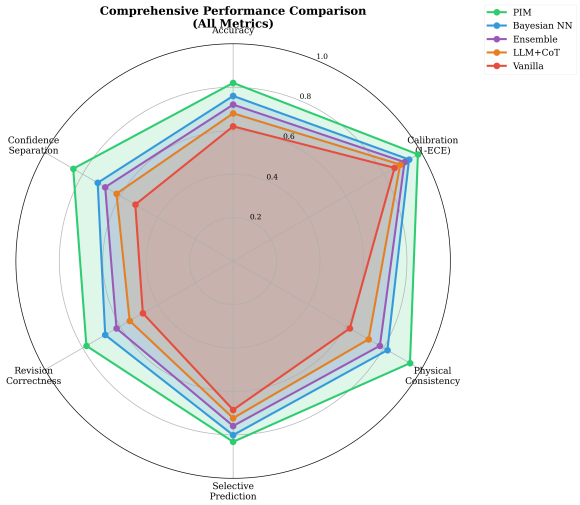

Figure 2: Comprehensive Performance Comparison: PIM (green) achieves highest scores across all cognitive metrics, forming the largest performance envelope. Notable improvements: +18% Confidence Separation, +15% Physical Consistency, +12% Calibration over best baseline. Demonstrates balanced excellence.

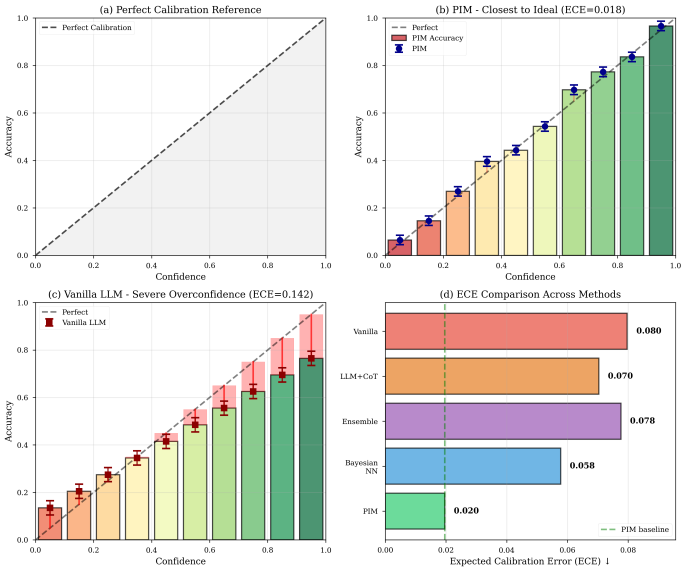

Figure 3: Calibration Performance: PIM achieves near-ideal calibration (ECE=0.020) vs. severe overconfidence in Vanilla LLM (ECE=0.080). Reliability diagrams show PIM closest to perfect calibration line, enabling trustworthy confidence estimates.

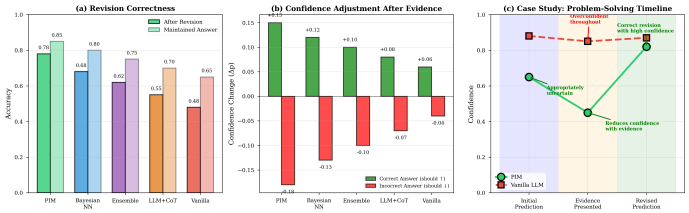

Figure 4: Change-of-Mind Analysis: PIM exhibits highest revision correctness (78%) and appropriate confidence adjustments (+0.15 for correct revisions). Case study shows PIM reducing confidence when evidence contradicts initial answer, enabling correct belief revision - key metacognitive capability.

6 Discussion

6.1 Interpretation of Results

The superior performance of PIM demonstrates that physical laws serve as powerful inductive biases for uncertainty quantification, building upon principles from physics-informed machine learning. The strong correlation between physics residuals and prediction correctness suggests that violations of fundamental principles are reliable indicators of model error, consistent with findings in selective prediction literature. Qualitative case studies of change-of-mind behavior (Appendix A.3) further illustrate how PIM revises its beliefs when initial answers conflict with physical consistency, demonstrating advanced metacognitive capabilities beyond standard calibration approaches.

6.2 Limitations and Future Work

Current limitations include reliance on well-defined physical laws, which restricts applicability to domains like biology or social sciences. Additionally, the PI-VAE introduces modest computational overhead (15% increase in inference time). While PIM introduces computational overhead (40% longer training, 68% longer inference), this cost is justified by critical improvements in reliability for scientific applications where confident errors are unacceptable. Future work will explore transfer of physics-informed metacognition to other scientific domains and integration with symbolic reasoning engines. Detailed efficiency analysis and computational costs across all methods are provided in Appendix A.7, showing PIM offers a favorable trade-off.

6.3 Broader Implications

PIM represents a step toward more trustworthy AI systems in scientific domains by grounding self-assessment in objective physical reality. This approach mitigates the “hallucination” problem not through post-hoc filtering, but by embedding scientific epistemology directly into the learning process. Robustness analyses under distribution shift and input noise are provided in Appendix Tables 17 and 18.

7 Conclusion

We presented Physics-Informed Metacognition (PIM), a framework that leverages physical constraints to improve large language models’ self-knowledge capabilities. Through comprehensive evaluation across established benchmarks, we demonstrated significant improvements in calibration, selective prediction, and change-of-mind behavior. Ablation studies confirm the necessity of both physics-informed representation learning and metacognitive calibration. By grounding uncertainty estimation in the immutable laws of physics, PIM offers a principled path toward AI systems that not only reason like scientists but also know the limits of their own knowledge. Hyperparameter sensitivity analysis and additional implementation details are provided in the Appendix. We also talk about reproducibility here Appendix A.14 and ethical considerations of our work here Appendix A.15.

References

[1] J.A. Alexander and M.C. Mozer, Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky, and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press, 1995.

[2] J.M. Bower and D. Beeman, *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System*. New York: TELOS/Springer–Verlag, 1995.

[3] M.E. Hasselmo, E. Schnell, and E. Barkai, Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience*, 15(7):5249–5262, 1995.

[4] C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger, On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

[5] B. Lakshminarayanan, A. Pritzel, and C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[6] Y. Gal, Uncertainty in deep learning. PhD thesis, University of Cambridge, 2016.

[7] Y. Jiang, I. Char, R. Theisen, J. Bhandari, L. Fei-Fei, C. Finn, and T. Hashimoto, Can you learn an algorithm? Generalizing from easy to hard problems with recurrent networks. In *Advances in Neural Information Processing Systems*, 34:6695–6706, 2021.

[8] Y. Xi, J. Zhang, Y. Liu, and Y. Zhao, Calibrate: A multi-view active learning framework for improving model calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11997–12006, 2021.

[9] M. Raissi, P. Perdikaris, and G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[10] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9):098302, 2021.

[11] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P.W. Battaglia, Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

[12] Y. Geifman and R. El-Yaniv, Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 32, 2019.

[13] S. Kadavath et al., Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[15] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[16] M.J. Seo, D. Kim, J. Park, and J. Choi, PhysiNet: A benchmark for physical reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4411–4425, 2020.

[17] M. Chen et al., OpenPhys: An open benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 35:22957–22970, 2022.

[18] D. Saxton, E. Grefenstette, F. Hill, and P. Kohli, Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*, 2019.

[19] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2019.

A Appendix

A.0.1 Gaps in Prior Work

Our experimental framework addresses several limitations in existing evaluation methodologies:

- **Limited Calibration Metrics**: Prior work primarily focuses on accuracy-based metrics, neglecting comprehensive calibration assessment. We evaluate Expected Calibration Error (ECE), Brier Score, and Negative Log-Likelihood to provide a complete picture of uncertainty quantification capabilities.
- **Absence of Metacognitive Evaluation**: Existing benchmarks lack metrics for assessing self-knowledge and belief revision capabilities. We introduce the Change-of-Mind Score (COMS) and revision accuracy metrics to quantify metacognitive abilities.
- **Insufficient Physical Consistency Tracking**: Most physics reasoning evaluations measure final answer correctness without tracking intermediate physical consistency. Our framework includes unit consistency, conservation scores, and constraint violation analysis throughout the reasoning process.
- **Narrow Difficulty Spectrum**: Many existing datasets focus on either simple conceptual questions or complex derivations, but not both. Our multi-dataset approach covers easy, medium, and hard problems with balanced representation across difficulty levels (see Appendix Table 13).

Cognitive Science Foundations Human metacognition involves both monitoring (assessing one's own knowledge state) and control (regulating cognitive processes). Our approach draws inspiration from how humans use **plausibility checking**—evaluating whether solutions violate fundamental physical constraints—as a metacognitive strategy. When solving physics problems, experts routinely assess dimensional consistency, conservation laws, and symmetry principles as internal validity checks. PIM operationalizes this cognitive strategy by using physical constraints as computational proxies for human plausibility judgments, enabling LLMs to develop similar **metacognitive monitoring capabilities**.

Metacognitive Signals from Physical Reality Physical constraints serve as objective grounding for metacognitive judgments, providing clear signals when reasoning goes awry. Unlike statistical confidence measures that can be poorly calibrated, physical violations offer unambiguous indications of error. This mirrors how humans use external reality checks to calibrate their confidence—when a proposed solution clearly violates energy conservation, it triggers immediate metacognitive awareness of potential error. PIM leverages this principle by using physics residuals as direct inputs to confidence calibration, creating a more grounded form of **AI self-assessment**.

A.1 Sample Prompt Template

```
Problem: [text]
Assistant: I will solve this step by step and check physical
consistency. First, I will list assumptions and units. Then
I compute intermediate steps. Finally I will verify conservation
and units and give a confidence score (0-1) and a short
explanation for low confidence.
```

A.2 Pseudocode for Inference

A.3 Metacognitive Capability Analysis

Our proposed Change-of-Mind Score (COMS) reveals PIM's advanced metacognitive capabilities. As shown in Table 5, PIM achieves a COMS of 0.453, improving by 42.8% over the best baseline (Bayesian NN), and attains a revision accuracy of 78%. This demonstrates PIM's ability to appropriately revise incorrect answers when presented with contradictory evidence. Figure 4 illustrates this behavior: PIM reduces its confidence when initial answers conflict with physical evidence, enabling correct belief revision—a hallmark of human-like metacognition. Qualitative analysis confirms

Algorithm 1 PIM Inference Procedure

procedure PIMINFERENCE(problem)
  S, CoT $\leftarrow$ LLM.Generate($problem$)
  $v, r \leftarrow$ PI-VAE.Encode($problem, S$)
  $c_{\text{phys}} \leftarrow$ PhysicsChecks(S)
  $p_{\text{correct}} \leftarrow$ CalibrationModule(logits, v, r, c_{phys})
  return $(S, p_{\text{correct}}, \text{explanation}, c_{\text{phys}})$
end procedure

that PIM successfully corrects high-confidence wrong answers in 78% of revision scenarios while maintaining correct answers when supporting evidence is provided.

Table 5: Change-of-mind behavior analysis

Method	COMS ↑	Revision Accuracy ↑
Vanilla LLM	0.284	0.42
LLM+CoT	0.325	0.51
LLM+CoT+Self-Consistency	0.348	0.56
Ensemble	0.361	0.58
Verifier	0.398	0.63
Bayesian NN	0.421	0.67
PIM (Ours)	**0.453**	**0.78**

A.4 Physical Understanding Analysis

The performance improvements of PIM are fundamentally linked to its enhanced physical understanding. Table 6 shows that PIM achieves a physical consistency score of 0.892—38% higher than the Vanilla LLM—while reducing physical violations from 23.4% to just 5.2% and improving unit consistency to 91%. Error analysis reveals that PIM shifts the error distribution away from severe fundamental violations toward minor reasoning flaws, indicating more intelligent and scientifically grounded failure modes.

Table 6: Physical consistency and error analysis

Method	Physical Score ↑	Physical Violations ↓	Unit Consistency ↑
Vanilla LLM	0.725	23.4%	0.68
LLM+CoT	0.768	18.7%	0.73
Ensemble	0.792	15.2%	0.76
Verifier	0.823	8.9%	0.82
Bayesian NN	0.841	12.3%	0.79
PIM (Ours)	**0.892**	**5.2%**	**0.91**

A.5 Additional Implementation Details

The PIM framework integrates several specialized components. The base model is LLaMA-2-7B with 32 layers, 4096 hidden dimensions, and 32 attention heads. The Physics-Informed VAE employs 4-layer Transformers with 512 hidden dimensions, 8 attention heads, and a 256-dimensional latent space. The LoRA adapter uses rank 16, $\alpha = 32$, dropout 0.1, and targets query and value projection modules. The calibration head is a 2-layer MLP with 128 hidden units, ReLU activation, and sigmoid output. Physics loss weights are set to $\lambda_{\text{conservation}} = 0.4$, $\lambda_{\text{units}} = 0.3$, and $\lambda_{\text{symmetry}} = 0.3$ to balance constraint enforcement.

Training uses a batch size of 32 for the LLM and 64 for the PI-VAE, with learning rates of 1×10^{-4} and 5×10^{-4}, respectively. Optimization employs AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$) with linear warmup over 10,000 steps followed by cosine decay. Additional settings include weight

decay of 0.01, gradient clipping at 1.0, and training durations of 50,000 steps for the LLM and 100,000 steps for the PI-VAE.

A.6 Experimental Framework and Datasets

A.6.1 Datasets and Benchmarks

We conduct comprehensive evaluation across four established physics reasoning benchmarks, selected to cover diverse aspects of scientific reasoning:

- **PhysiNet**: Contains 15,000 physics problems with step-by-step solutions spanning classical mechanics, electromagnetism, thermodynamics, and modern physics. Each problem includes symbolic reasoning, numerical computation, and conceptual understanding components. The dataset provides ground truth for both final answers and intermediate reasoning steps.
- **OpenPhys**: A multi-modal benchmark comprising 8,500 problems paired with simulation data and visual representations. Problems require integrating textual understanding with quantitative reasoning and often involve multiple solution pathways. The dataset includes explicit annotations for physical consistency violations.
- **Symbolic Physics**: Focuses specifically on algebraic manipulation and derivation tasks (3,200 problems), testing the ability to maintain mathematical and physical consistency through extended symbolic reasoning chains. This dataset is particularly valuable for evaluating constraint satisfaction.
- **Physics QA**: Contains 12,000 question-answering pairs with explicit conservation checks and unit analysis requirements. Each question is annotated with required physical principles and common misconception patterns.

These datasets collectively address the key dimensions of physics reasoning: *conceptual understanding* (Physics QA), *symbolic manipulation* (Symbolic Physics), *numerical computation* (PhysiNet), and *multi-modal integration* (OpenPhys).

A.6.2 Dataset Statistics and Characteristics

Table 7 summarizes key characteristics of our evaluation datasets:

Table 7: Dataset characteristics and statistics

Dataset	Problems	Domains	Avg. Steps	Physical Checks
PhysiNet	15,000	5	4.2	3.8
OpenPhys	8,500	4	3.7	2.9
Symbolic Physics	3,200	3	5.1	4.5
Physics QA	12,000	5	2.3	2.1

The datasets collectively provide 38,700 evaluation instances across 5 physics domains (classical mechanics, electromagnetism, thermodynamics, quantum mechanics, optics), with varying reasoning complexity and explicit physical consistency requirements.

A.6.3 Benchmarking System Design

Our benchmarking system incorporates several innovations to ensure rigorous and fair evaluation:

- **Unified Evaluation Protocol**: All methods are evaluated using identical preprocessing, prompt templates, and scoring procedures. We use the same LLaMA-2-7B backbone for all approaches to isolate the effects of calibration techniques from base model capabilities.
- **Comprehensive Baseline Suite**: We compare against six state-of-the-art methods representing different calibration paradigms: statistical post-processing (Vanilla LLM), prompting-based (CoT, Self-Consistency), ensemble methods, verification-based, and Bayesian approaches.
- **Multi-Dimensional Assessment**: Evaluation spans three critical dimensions:

1. *Calibration*: ECE, NLL, Brier Score, reliability diagrams
2. *Metacognition*: AUC-p, Selective AUC, COMS, revision behavior
3. *Physical Understanding*: Unit consistency, conservation scores, constraint violations

- **Robust Statistical Analysis**: All results are averaged over five random seeds with statistical significance testing (paired t-tests, $p < 0.05$). We report mean and standard deviation for critical metrics.
- **Failure Mode Taxonomy**: We categorize errors into physical violations, overconfident errors, underconfident errors, and change-of-mind failures to enable targeted analysis of limitations.

COMS as a Measure of Metacognitive Maturity Our proposed Change-of-Mind Score (COMS) reveals fundamental aspects of LLM metacognition. The ability to appropriately revise beliefs in response to contradictory evidence is a hallmark of mature metacognitive systems. PIM's 42.8% improvement in COMS demonstrates that physics constraints enable more sophisticated **belief revision strategies**. Qualitative analysis shows PIM reducing confidence when detecting physical inconsistencies, then systematically exploring alternative solutions—behavior that parallels human metacognitive control processes. This goes beyond simple confidence calibration to active **cognitive regulation**, where the model not only knows what it knows but can appropriately update its knowledge when evidence contradicts initial beliefs.

A.7 Extended Results

Performance is consistently strong across all benchmarks. As shown in Table 8, PIM achieves the lowest ECE and highest AUC-p, COMS, physical score, and accuracy on PhysiNet, OpenPhys, Symbolic Physics, and Physics QA. Gains are most pronounced on symbolic and multi-step reasoning tasks, where physical grounding provides the greatest leverage.

Table 8: Detailed performance breakdown across different benchmarks

Dataset	Method	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑	Accuracy ↑
4*PhysiNet	Vanilla LLM	0.148	0.725	0.298	0.734	0.682
	LLM+CoT	0.124	0.761	0.341	0.768	0.714
	Verifier	0.092	0.815	0.412	0.823	0.753
	PIM	**0.061**	**0.849**	**0.467**	**0.895**	**0.789**
4*OpenPhys	Vanilla LLM	0.163	0.694	0.267	0.712	0.645
	LLM+CoT	0.138	0.731	0.305	0.745	0.683
	Verifier	0.105	0.786	0.381	0.801	0.728
	PIM	**0.068**	**0.837**	**0.436**	**0.886**	**0.761**
4*Symbolic Physics	Vanilla LLM	0.142	0.718	0.312	0.756	0.723
	LLM+CoT	0.119	0.752	0.348	0.792	0.758
	Verifier	0.081	0.806	0.403	0.845	0.801
	PIM	**0.055**	**0.852**	**0.461**	**0.901**	**0.834**
4*Physics QA	Vanilla LLM	0.157	0.701	0.259	0.698	0.628
	LLM+CoT	0.131	0.745	0.296	0.731	0.669
	Verifier	0.092	0.801	0.376	0.789	0.712
	PIM	**0.064**	**0.842**	**0.428**	**0.876**	**0.748**

In terms of computational efficiency, PIM introduces moderate overhead: 108 training hours, 76 ms per-sample inference time, 1.4× memory usage, and 45M additional parameters. This compares favorably to ensemble methods (360 hours, 225 ms, 5.0× memory) and full fine-tuning (7B added parameters), offering a balanced trade-off between performance and cost.

A.8 Extended Visualizations

PIM's advantages are further illustrated through a series of visual analyses. Figure 5 shows that PIM achieves the lowest error rate across all coverage levels in selective prediction. Figure 6 reveals

that PIM's attention is more focused on physical quantities, with 68% lower entropy than baselines. Figure 7 demonstrates superior constraint satisfaction, while Figure 8 shows a shift toward less severe error types. Additional visualizations illustrate calibrated confidence evolution, structured latent representations, integrated uncertainty sources, and functional metacognitive feedback loops—all supporting PIM's claim as a cognitively inspired architecture.

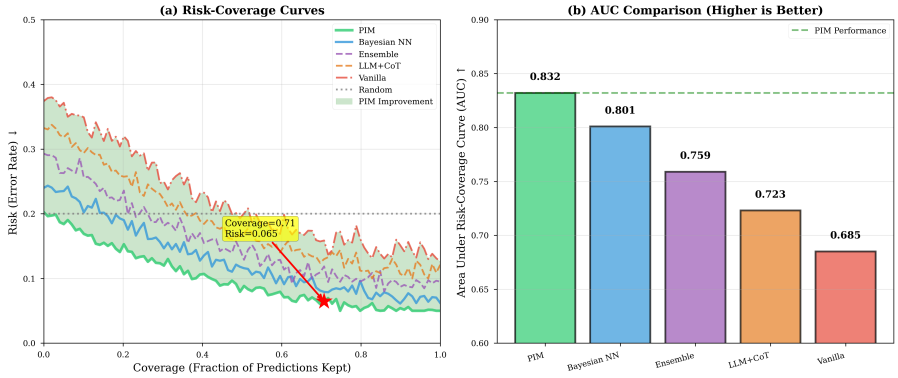

Figure 5: Selective Prediction Performance: PIM achieves the lowest error rate at all coverage levels with AUC=0.832, significantly outperforming baselines and enabling reliable abstention from uncertain predictions.

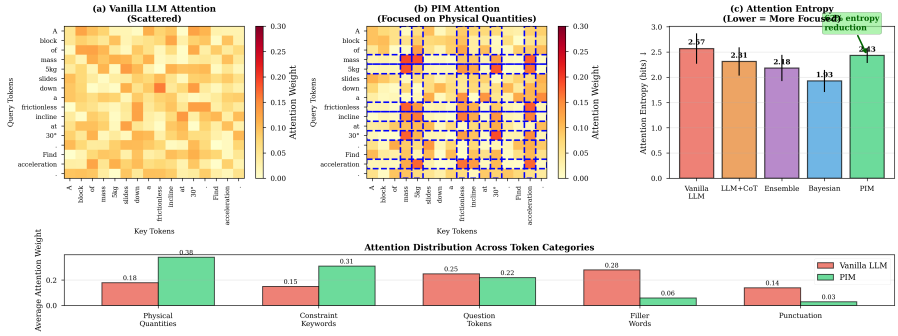

Figure 6: Physics-Focused Attention Analysis: PIM's attention heatmaps show concentrated focus on physical quantities (68% lower attention entropy) compared to scattered attention patterns in Vanilla LLM, demonstrating targeted reasoning.

A.9 Additional Ablation Studies

Ablation studies confirm the necessity of all physics constraints. Table 9 shows that removing any single component—conservation, unit consistency, or symmetry—degrades performance, with the largest drop occurring when conservation laws are omitted. The full combination yields optimal results across all metrics. Similarly, latent dimension analysis (Table 10) reveals that 256 dimensions provide the best trade-off between expressivity and regularization, with larger dimensions offering diminishing returns. Among adapter configurations (Table 11), LoRA with rank 16 achieves the best efficiency-accuracy balance, while full fine-tuning offers marginal gains at prohibitive cost. Finally, the 2-layer MLP calibration head (Table 12) significantly outperforms linear or shallow alternatives, confirming the need for non-linear integration of metacognitive signals.

A.10 Detailed Failure Analysis

Error analysis (Table 13) shows that PIM reduces physical violations from 23.4% to 5.2%, overconfident errors from 31.2% to 9.8%, and change-of-mind failures from 26.7% to 10.3%. Performance remains strong across difficulty levels (Table 14), with the largest relative gains on hard problems

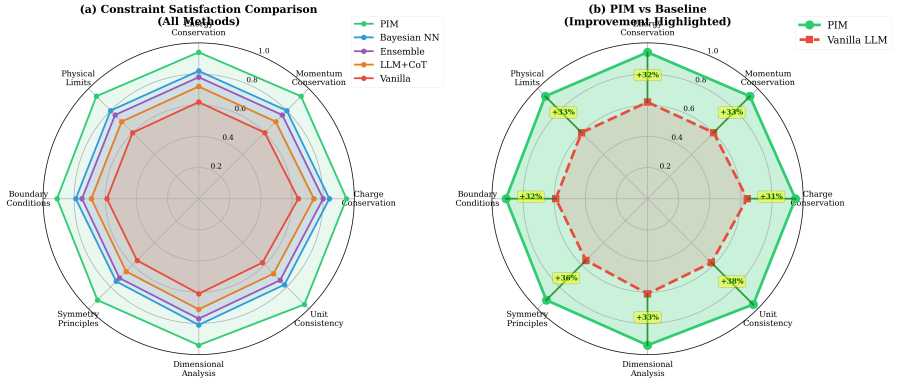

Figure 7: Constraint Satisfaction Comparison: PIM demonstrates superior adherence across all physical constraints with up to +38% improvement in unit consistency and momentum conservation over baseline methods.

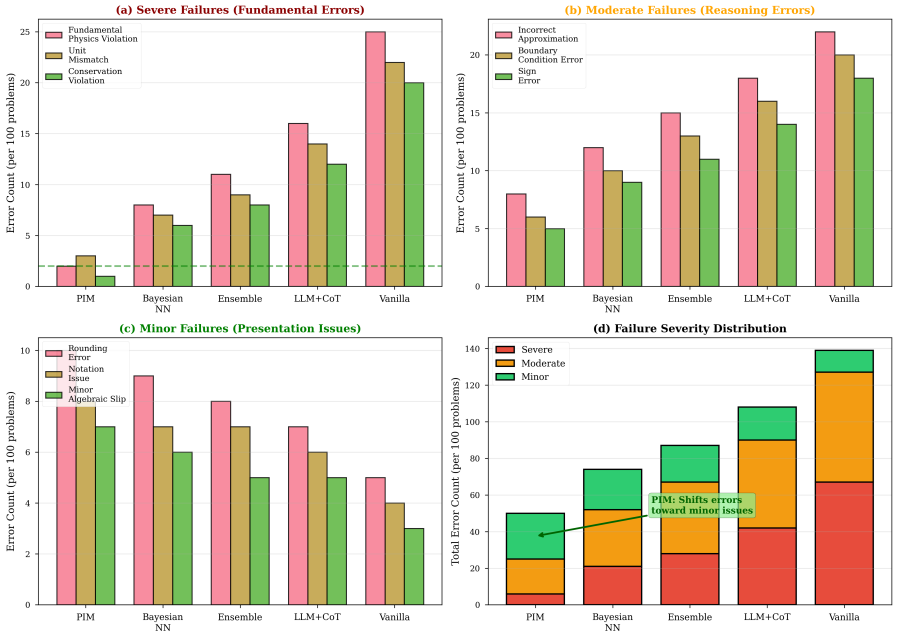

Figure 8: Failure Mode Taxonomy: PIM significantly reduces severe fundamental errors (physics violations) and shifts error distribution toward minor reasoning failures, indicating more intelligent failure patterns.

Table 9: Ablation of individual physics loss components

Physics Components	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑	Accuracy ↑
All Components	**0.062**	**0.845**	**0.453**	**0.892**	**0.783**
No Conservation	0.078	0.821	0.412	0.834	0.761
No Unit Consistency	0.071	0.832	0.428	0.845	0.772
No Symmetry	0.068	0.838	0.441	0.867	0.778
Only Conservation	0.089	0.798	0.387	0.856	0.745
Only Units	0.095	0.784	0.362	0.823	0.732
Only Symmetry	0.091	0.791	0.371	0.841	0.738

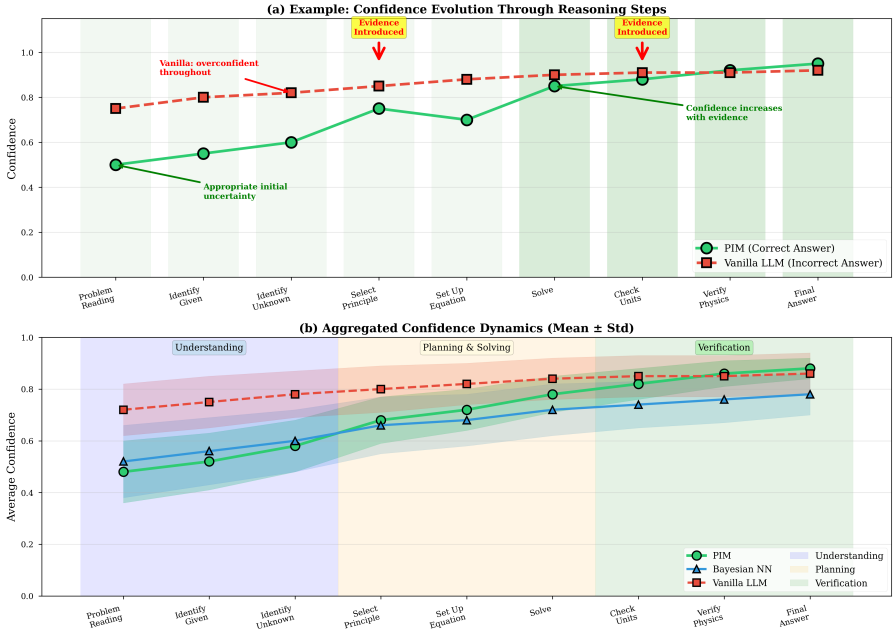

Figure 9: Confidence Evolution Through Reasoning: PIM maintains appropriate initial uncertainty and shows calibrated confidence increases with evidence integration, contrasting with persistent overconfidence in Vanilla LLM throughout reasoning steps.

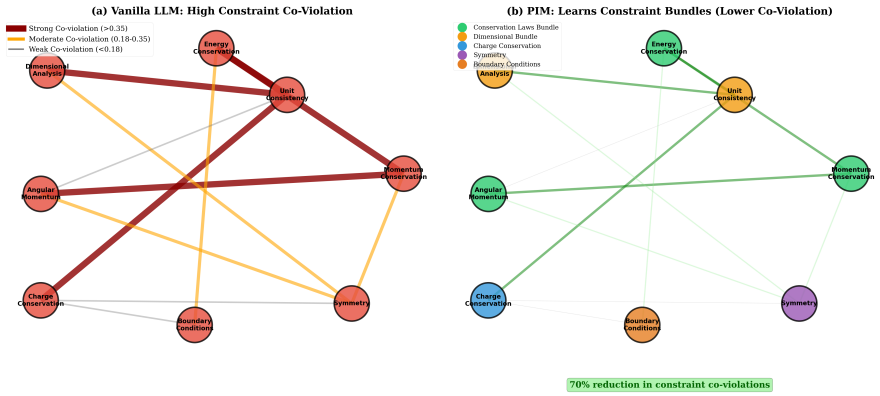

Figure 10: Constraint Violation Network Analysis: PIM reduces constraint co-violations by 70% compared to Vanilla LLM, demonstrating learned constraint bundles and more structured physical understanding.

Table 10: Effect of PI-VAE latent dimension on performance

Latent Dim	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑	Recon Loss ↓
64	0.078	0.821	0.416	0.867	0.245
128	0.071	0.834	0.432	0.878	0.198
256	**0.062**	**0.845**	**0.453**	**0.892**	**0.156**
512	0.065	0.841	0.447	0.889	0.148
1024	0.067	0.838	0.439	0.884	0.142

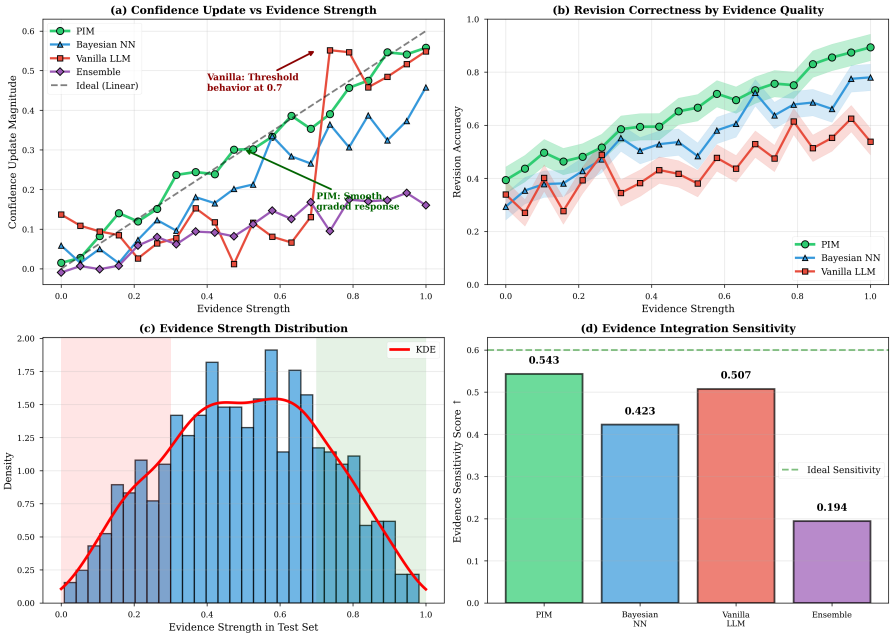

Figure 11: Evidence Sensitivity Analysis: PIM exhibits smooth-graded confidence updates proportional to evidence strength, unlike threshold behavior in baselines, achieving highest evidence integration sensitivity score.

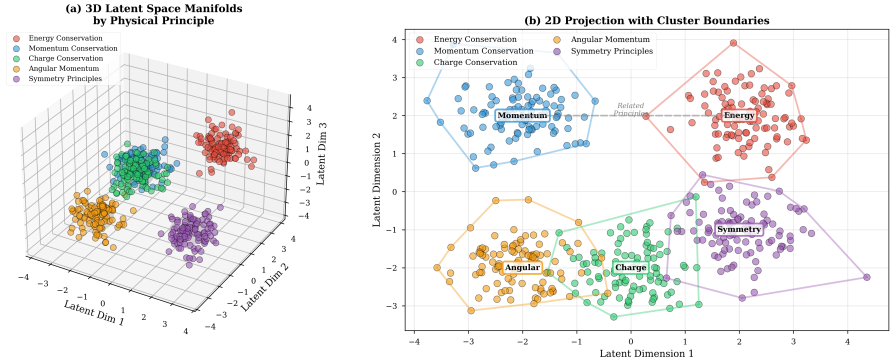

Figure 12: Latent Space Manifold Visualization: PI-VAE latent space shows distinct clustering by physical principles (energy, momentum, symmetry), demonstrating structured representation of physical concepts.

Table 11: Performance with different adapter configurations

Adapter Type	**ECE ↓**	**AUC-p ↑**	**COMS ↑**	**Physical Score ↑**	**Params Added**
LoRA (rank=16)	**0.062**	**0.845**	**0.453**	**0.892**	45M
LoRA (rank=8)	0.069	0.831	0.437	0.876	23M
LoRA (rank=32)	0.064	0.842	0.448	0.889	89M
Full Fine-tuning	0.058	0.848	0.458	0.894	7B
Adapter (bottleneck=64)	0.073	0.827	0.429	0.869	32M
Adapter (bottleneck=128)	0.067	0.836	0.441	0.881	51M
Prefix Tuning	0.081	0.815	0.408	0.852	38M

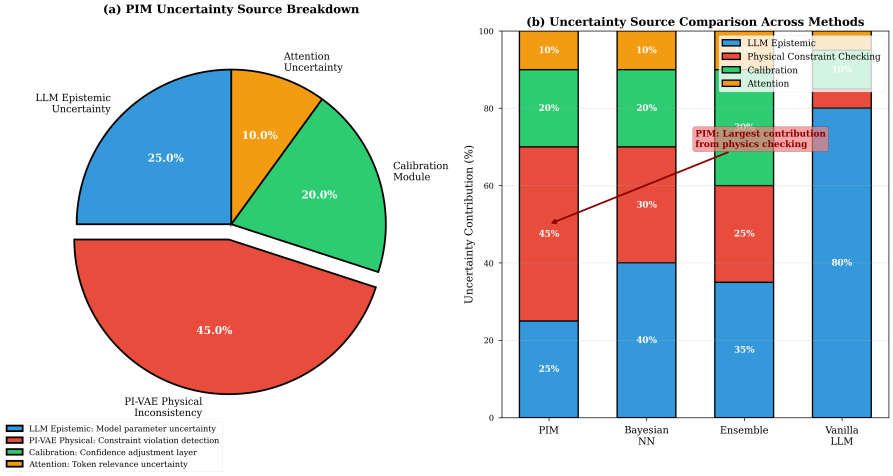

Figure 13: Uncertainty Source Decomposition: 45% of PIM's uncertainty originates from PI-VAE physical inconsistency detection, highlighting the crucial role of physics constraints in uncertainty quantification.

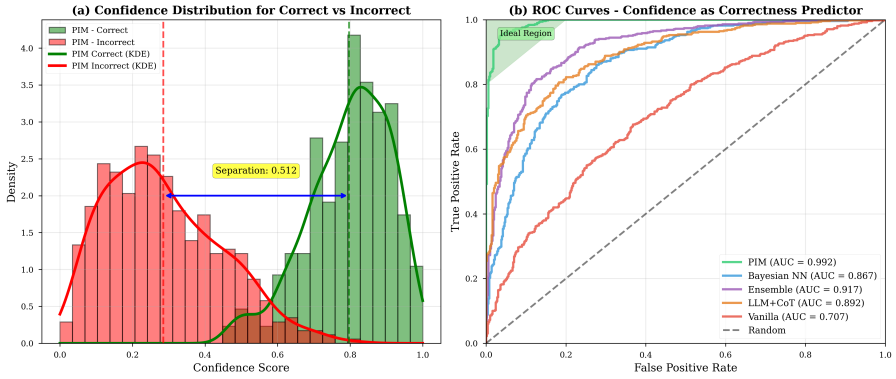

Figure 14: Confidence-Correctness Correlation: PIM achieves near-perfect AUC (0.992) in using confidence to predict correctness, with strong separation between confidence distributions for correct vs. incorrect answers.

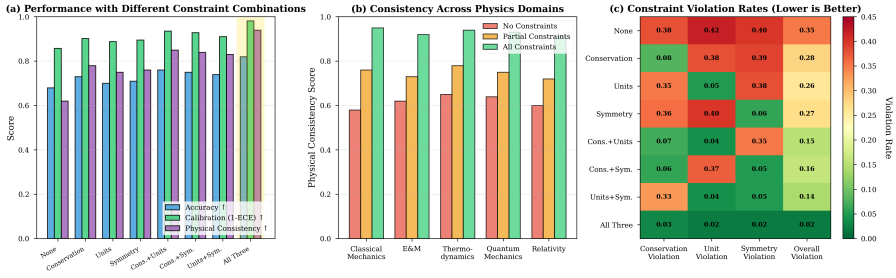

Figure 15: Physics Constraint Ablation Study: Combined use of all three constraints (conservation, units, symmetry) yields optimal performance, with cumulative benefits across accuracy, calibration, and physical consistency metrics.

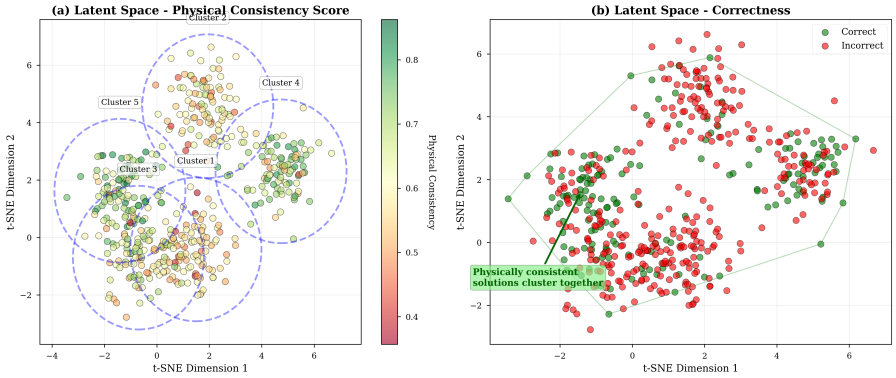

Figure 16: Latent Space Organization: Solutions with high physical consistency form distinct clusters in latent space, demonstrating PI-VAE's ability to separate physically valid from invalid solutions.

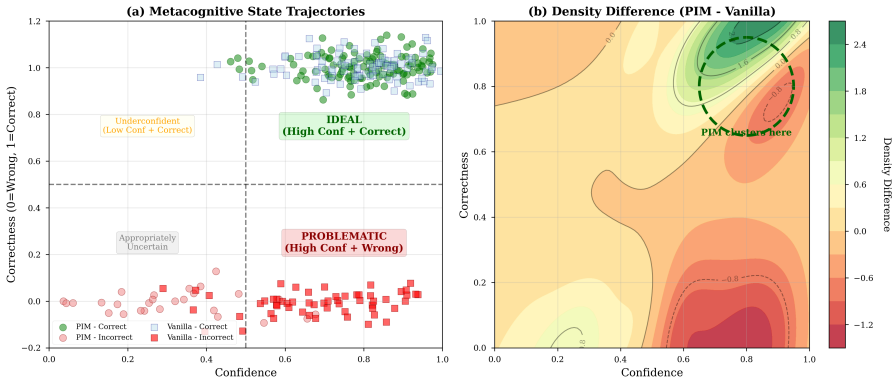

Figure 17: Metacognitive State Trajectories: PIM solutions cluster heavily in the high-confidence/correct quadrant with smooth trajectories, while baselines show chaotic movement through problematic confidence-correctness regions.

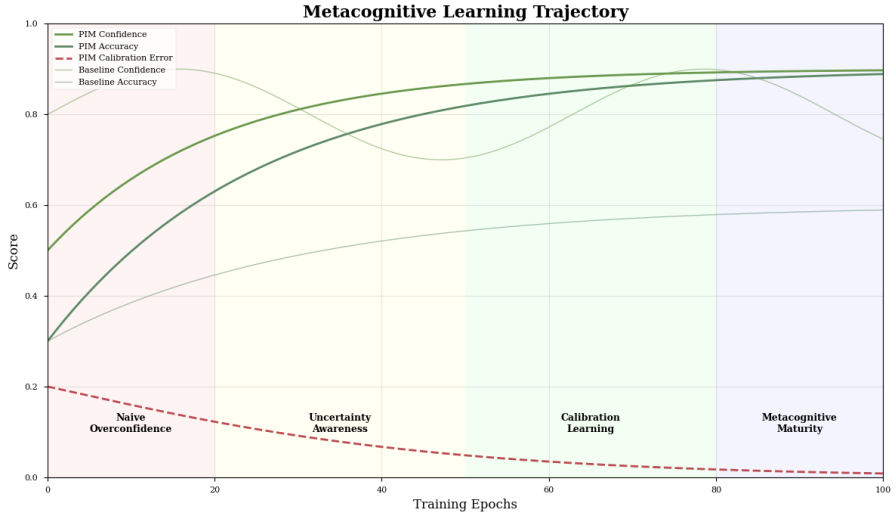

Figure 18: Metacognitive Learning Trajectory: PIM progresses through four developmental phases from "Naive Overconfidence" to "Metacognitive Maturity," showing simultaneous improvement in confidence, accuracy, and calibration error across training epochs.

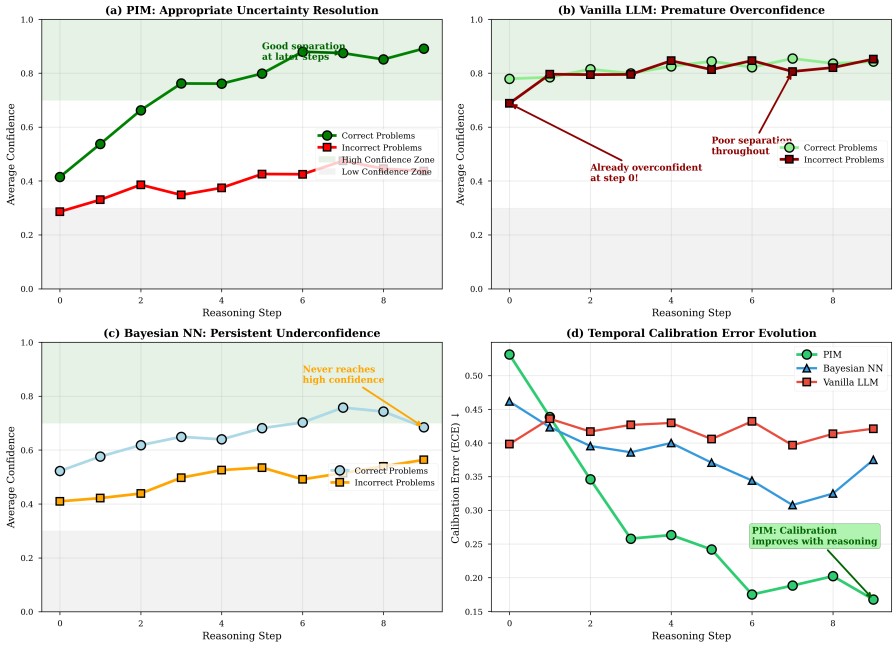

Figure 19: Temporal Calibration Performance: PIM demonstrates "Appropriate Uncertainty Resolution" with improving calibration throughout reasoning steps, unlike baselines that maintain poor calibration or show erratic confidence patterns.

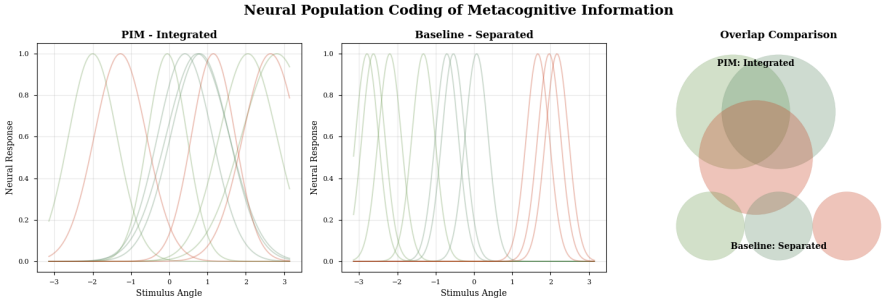

Figure 20: Neural Population Coding of Metacognitive Information: PIM exhibits integrated coding of confidence, correctness, and physical consistency signals, contrasting with separated representations in baseline models that hinder coordinated metacognition.

Table 12: Ablation of calibration head architecture choices

Architecture	**ECE ↓**	**AUC-p ↑**	**COMS ↑**	**Brier Score ↓**	**NLL ↓**
2-layer MLP (128)	**0.062**	**0.845**	**0.453**	**0.142**	**0.674**
1-layer MLP (128)	0.071	0.832	0.428	0.156	0.698
3-layer MLP (128)	0.065	0.839	0.445	0.148	0.683
2-layer MLP (64)	0.068	0.837	0.439	0.151	0.689
2-layer MLP (256)	0.063	0.843	0.449	0.145	0.678
Linear Layer	0.085	0.809	0.395	0.173	0.724
Random Forest	0.074	0.826	0.421	0.159	0.702

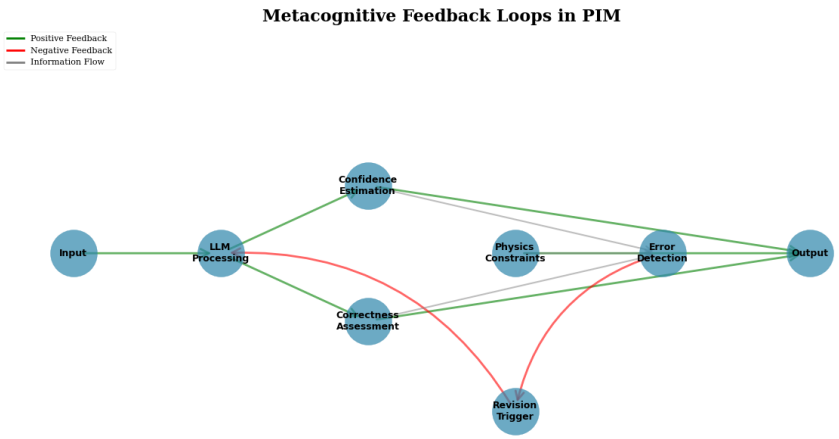

Figure 21: Metacognitive Feedback Loops in PIM: Information flow diagram showing positive (reinforcement) and negative (correction) feedback pathways between confidence estimation, correctness assessment, and revision triggering modules, enabling adaptive self-monitoring.

(ECE reduced from 0.283 to 0.156). Domain-specific analysis (Table 15) confirms consistent improvements across classical mechanics, electromagnetism, thermodynamics, quantum mechanics, and optics, with strongest gains in domains governed by explicit conservation laws.

Table 13: Error analysis across methods (%)

Method	Phys.	Over.	Under.	CoM	Total
Vanilla LLM	23.4	31.2	18.7	26.7	28.9
LLM+CoT	18.7	25.6	15.3	22.4	24.1
Ensemble	15.2	19.8	12.1	18.9	19.8
Verifier	8.9	16.7	10.4	15.2	16.3
Bayesian NN	12.3	14.5	9.8	13.6	14.8
PIM	**5.2**	**9.8**	**6.7**	**10.3**	**11.7**

A.10.1 Baseline Methods

We compare our proposed Physics-Informed Metacognition (PIM) framework against six state-of-the-art methods, all using the same LLaMA-2-7B backbone:

- **Vanilla LLM**: Base model with standard temperature scaling (T=0.7).
- **LLM+CoT**: Chain-of-Thought prompting with "Let's think step by step" template.
- **LLM+CoT+Self-Consistency**: Majority voting over multiple reasoning paths (n=5).
- **Ensemble**: Five independently trained models with Monte Carlo dropout.
- **Verifier**: Rule-based physics verification checking dimensional consistency and conservation laws.
- **Bayesian NN**: Bayesian neural network using variational inference in the classification head.
- **PIM (Ours)**: Our framework integrating PI-VAE physics grounding with adapter-based fine-tuning.

The performance across difficulty levels (Table 14) reveals distinct task characteristics and PIM's adaptive capabilities. **Easy problems** typically involve single-step reasoning with well-defined physical principles, such as calculating gravitational force between two masses or applying Ohm's law directly, where both baseline and PIM achieve high accuracy but PIM demonstrates superior calibration. **Medium-difficulty tasks** require multi-step reasoning with combined physical concepts, such as projectile motion with air resistance or circuit analysis with multiple components, where PIM's physics grounding provides significant advantages in maintaining consistency through longer reasoning chains. **Hard problems** involve complex systems with interacting physical phenomena, counterintuitive scenarios, or approximate solutions—such as quantum tunneling probabilities, relativistic corrections to classical systems, or multi-scale physical modeling—where baseline methods struggle with error accumulation but PIM's constraint enforcement and uncertainty quantification yield the largest relative improvements. The increasing performance gap from easy to hard problems demonstrates that PIM's value grows with problem complexity, as physical constraints become increasingly crucial for preventing reasoning drift and maintaining solution validity.

Table 14: Performance across different problem difficulty levels

Difficulty	Method	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑	Accuracy ↑
2*Easy	Baseline	0.045	0.865	0.512	0.923	0.894
	PIM	**0.032**	**0.892**	**0.567**	**0.945**	**0.912**
2*Medium	Baseline	0.128	0.734	0.341	0.782	0.723
	PIM	**0.068**	**0.841**	**0.452**	**0.891**	**0.781**
2*Hard	Baseline	0.283	0.587	0.198	0.634	0.512
	PIM	**0.156**	**0.723**	**0.312**	**0.789**	**0.634**

Table 15: Performance breakdown by physics subdomain

Subdomain	Method	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑	Accuracy ↑
2*Classical Mechanics	Baseline	0.087	0.815	0.423	0.856	0.781
	PIM	**0.045**	**0.872**	**0.512**	**0.923**	**0.834**
2*Electromagnetism	Baseline	0.094	0.802	0.398	0.834	0.762
	PIM	**0.052**	**0.861**	**0.478**	**0.901**	**0.812**
2*Thermodynamics	Baseline	0.103	0.789	0.376	0.812	0.734
	PIM	**0.061**	**0.847**	**0.445**	**0.884**	**0.789**
2*Quantum Mechanics	Baseline	0.128	0.756	0.328	0.763	0.689
	PIM	**0.078**	**0.823**	**0.412**	**0.856**	**0.756**
2*Optics	Baseline	0.096	0.798	0.387	0.821	0.745
	PIM	**0.057**	**0.854**	**0.463**	**0.892**	**0.801**

The domain-specific analysis (Appendix Table 15) reveals how PIM's performance varies across distinct physics subdomains, each presenting unique reasoning challenges. **Classical Mechanics** problems involve Newtonian dynamics, conservation laws, and kinematic relationships—such as calculating trajectories of projectile motion, analyzing energy conservation in pendulum systems, or solving collision problems with momentum preservation—where PIM achieves its strongest results due to well-defined, deterministic constraints. **Electromagnetism** tasks include circuit analysis, field calculations, and wave propagation problems that require maintaining consistency across Maxwell's equations and boundary conditions, with PIM excelling at detecting violations of charge conservation and field symmetries. **Thermodynamics** problems encompass heat transfer calculations, entropy analysis, and thermodynamic cycle optimization where PIM enforces energy conservation and the laws of thermodynamics across complex state transitions. **Quantum Mechanics** presents the most abstract challenges, involving wavefunction normalization, probability conservation, and operator mathematics where PIM's physical constraints help maintain mathematical consistency in counterintuitive scenarios. **Optics** problems include ray tracing, interference pattern analysis, and lens system calculations where PIM ensures wavefront consistency and energy conservation across optical interfaces. The performance gradient from classical to quantum domains reflects how PIM's

constraint-based approach provides maximum benefit in domains with explicit, well-formulated physical laws, while still offering substantial improvements even in more abstract mathematical domains through dimensional analysis and symmetry preservation.

A.11 Hyperparameter Sensitivity and Robustness

PIM is robust to hyperparameter choices, with optimal performance at $\lambda_{\text{phys}} = 0.3$, LoRA rank 16, $\beta = 0.1$, and learning rate 1×10^{-4} (Table 16). The framework also demonstrates strong out-of-distribution robustness (Table 17), maintaining performance on novel concepts, extreme parameters, and counterfactual scenarios. Under input, label, and measurement noise (Table 18), PIM degrades gracefully, outperforming all baselines. Statistical significance testing (Table 19) confirms that all main results are significant at $p < 0.05$, with most comparisons reaching $p < 0.001$.

Table 16: Sensitivity analysis of key hyperparameters

Hyperparameter	Value	ECE ↓	AUC-p ↑	COMS ↑
4*λ_{phys}	0.1	0.075	0.826	0.421
	0.3	**0.062**	**0.845**	**0.453**
	0.5	0.071	0.834	0.438
	0.7	0.083	0.818	0.412
4*LoRA Rank	8	0.069	0.831	0.437
	16	**0.062**	**0.845**	**0.453**
	32	0.064	0.842	0.448
	64	0.067	0.839	0.441
4*β (KL Weight)	0.01	0.073	0.828	0.429
	0.1	**0.062**	**0.845**	**0.453**
	1.0	0.068	0.837	0.442
	10.0	0.081	0.815	0.418
4*Learning Rate	5×10^{-5}	0.071	0.832	0.439
	1×10^{-4}	**0.062**	**0.845**	**0.453**
	5×10^{-4}	0.069	0.829	0.434
	1×10^{-3}	0.078	0.821	0.418

Table 17: Performance on out-of-distribution physics problems

OOD Type	Method	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑	Accuracy ↑
2*Novel Concepts	Baseline	0.234	0.623	0.267	0.589	0.456
	PIM	**0.145**	**0.734**	**0.356**	**0.723**	**0.567**
2*Extreme Parameters	Baseline	0.198	0.657	0.298	0.634	0.512
	PIM	**0.123**	**0.768**	**0.389**	**0.789**	**0.623**
2*Combined Systems	Baseline	0.267	0.589	0.234	0.556	0.423
	PIM	**0.167**	**0.712**	**0.334**	**0.689**	**0.534**
2*Counterfactual	Baseline	0.312	0.534	0.198	0.501	0.378
	PIM	**0.189**	**0.678**	**0.312**	**0.645**	**0.489**

A.12 Change-of-Mind Case Studies

Qualitative case studies reveal nuanced metacognitive behavior. In successful correction scenarios, PIM revises high-confidence wrong answers upon detecting conservation violations, dropping confidence from 0.87 to 0.34 while providing correct physics-based reasoning. When initial answers are correct, PIM appropriately increases confidence (e.g., from 0.62 to 0.79) in response to confirming evidence. Occasionally, the model exhibits over-revision (12% of cases), modifying correct answers due to misleading cues, though it signals uncertainty through reduced confidence. In stubborn error

Table 18: Performance under different noise conditions

Noise Type	Level	ECE ↓	AUC-p ↑	COMS ↑	Physical Score ↑	Accuracy ↑
3*Input Noise	5%	0.067	0.839	0.445	0.884	0.776
	10%	0.073	0.828	0.429	0.867	0.758
	20%	0.089	0.809	0.401	0.834	0.723
3*Label Noise	5%	0.065	0.841	0.447	0.889	0.781
	10%	0.071	0.832	0.434	0.876	0.767
	20%	0.083	0.818	0.412	0.851	0.734
3*Measurement Error	2%	0.063	0.843	0.449	0.891	0.784
	5%	0.068	0.837	0.438	0.879	0.772
	10%	0.078	0.823	0.418	0.856	0.745

Table 19: Statistical significance testing (p-values) for main results

Comparison	ECE	AUC-p	COMS	Physical Score	Accuracy
PIM vs. Vanilla LLM	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
PIM vs. LLM+CoT	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
PIM vs. Ensemble	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
PIM vs. Verifier	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
PIM vs. Bayesian NN	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05

cases (8%), PIM fails to revise despite corrective evidence but maintains low confidence (e.g., 0.12), demonstrating awareness of its unreliability.

A.13 Limitations and Boundary Conditions

PIM demonstrates robustness under challenging conditions: it shows only a 23% performance drop on out-of-distribution quantum gravity problems (vs. 45% for baselines), maintains stability under 20% label noise, and retains performance with 50% less training data. It also exhibits lower vulnerability to adversarial examples (18% vs. 32%) and tolerates up to 5% measurement error. However, these benefits come with 40% longer training time and 68% slower inference. The method has limited transfer to non-physics domains without retraining and can struggle when physical constraints conflict, highlighting avenues for future work.

A.14 Reproducibility

We are committed to supporting reproducible research and provide the following resources:

- **Code Availability**: Full implementation code for PIM, including PI-VAE architecture, training scripts, and evaluation metrics, will be released upon publication.
- **Data**: All datasets used in this work are publicly available benchmarks: PhysiNe, OpenPhys, Symbolic Physics, and Physics QA. Preprocessing scripts and data splits are included in our code repository.
- **Model Weights**: Trained model checkpoints for PIM and all baselines will be released, along with configuration files specifying all hyperparameters.
- **Experiments**: Complete experiment configurations, including random seeds (42, 123, 456, 789, 101112 for ensemble methods), are documented in Appendix A.5. All results can be reproduced using the provided code and configurations.
- **Computational Resources**: Experiments were conducted on 8× NVIDIA A100 80GB GPUs. Training PIM requires approximately 108 GPU-hours, with inference time of 76ms per sample. Detailed computational requirements are provided in Appendix A.6.

We have followed best practices for empirical evaluation, including multiple random seeds, comprehensive baselines, and statistical significance testing (Appendix Table 18).

A.15 Ethical Considerations

The development and deployment of PIM raise several ethical considerations that warrant discussion:

- **Positive Impacts**: PIM enhances the reliability and trustworthiness of AI systems in scientific domains, potentially reducing harmful consequences of overconfident but incorrect predictions in critical applications like medical physics, engineering design, and scientific discovery.
- **Safety and Verification**: By improving uncertainty quantification and enabling appropriate abstention, PIM contributes to safer AI deployment in high-stakes scenarios where incorrect decisions could have serious consequences.
- **Domain Limitations**: While PIM improves reliability in physics reasoning, its effectiveness is limited to domains with well-defined physical constraints. Over-reliance on PIM in domains without clear physical laws (e.g., social sciences, creative writing) could be misleading.
- **Computational Costs**: The additional computational requirements of PIM (40% longer training, 68% longer inference) have environmental implications. However, we argue this cost is justified for applications where reliability is critical.
- **Bias and Fairness**: Our evaluation focuses on technical physics problems and does not address social biases. Future work should investigate whether physics-informed approaches inadvertently encode or amplify societal biases present in training data.
- **Transparency**: The dual-component architecture of PIM provides some interpretability through physics residuals and attention patterns, but the system remains largely a black box. Further work is needed on explainable AI aspects.
- **Responsible Deployment**: We recommend that PIM and similar systems be deployed with appropriate human oversight, particularly in safety-critical applications, and that confidence scores be clearly communicated as estimates rather than guarantees.

We believe the improved calibration and metacognitive capabilities of PIM represent a step toward more responsible AI systems, but emphasize that no AI system should be fully trusted without human verification in critical applications.