The 40th Annual AAAI Conference on Artificial Intelligence
JANUARY 20 – JANUARY 27, 2026 | SINGAPORE

"Improving Physics Reasoning in Large Language Models Using Mixture of Refinement Agents"

Raj Jaiswal*[1], Dhruv Jain*[2], Harsh Parimal Popat[1], Abhishek Dharmadhikari[1], Atharva Marathe[1], Avinash Anand[1], Shin'ichi Satoh[3], Rajiv Ratn Shah[1]
MIDAS LAB, IIIT Delhi[1] , Indian Institute of Technology (BHU) Varanasi[2], National Institute of Informatics, Tokyo[3]

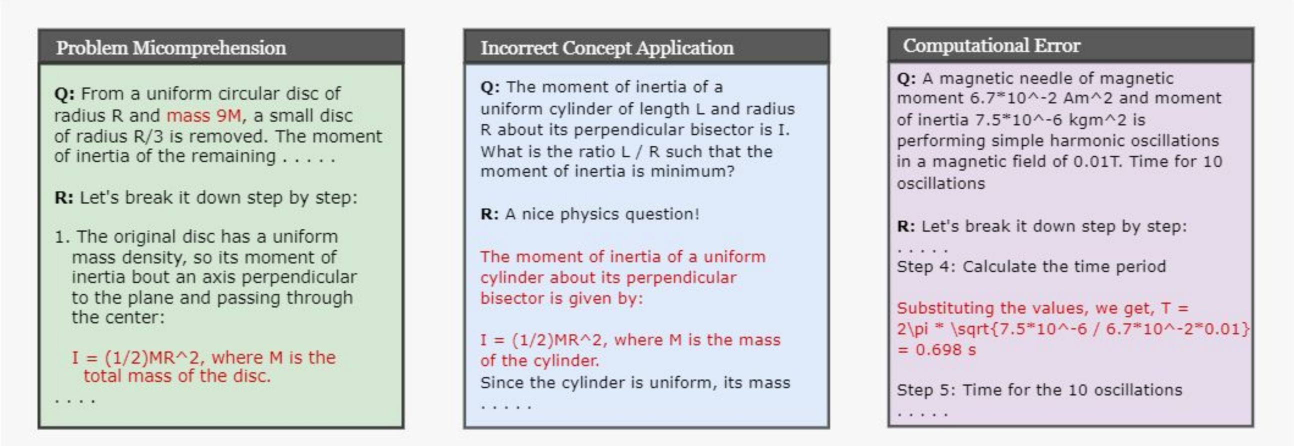
Motivation :

While analyzing physics problems and their CoT solutions generated with LLMs (Llama-3-70B & Gemma-2-27B), we observed three key errors made by them :

Observation 1 : LLMs in few cases struggle to fully grasp the objective of the question, along with misinterpreting the values of variables and constants provided in the question. Although this issue has been identified in only a few cases, it is significant one because it leads to solutions that fails to address the correct interpretation of a given question resulting in problem miscomprehension.

Observation 2 : LLMs struggle to apply the correct concepts or formulae with respect to the context of the given problem.This issue is a more recurring one in LLMs, especially for problems requiring considering a specific case rather than relying on a generic formula. For instance, the formula for calculating the moment of inertia varies depending on the distribution of mass.

Observation 3 : Many physics problems involve mathematical reasoning and algebraic computation, areas where LLMs tend to struggle Computational errors account for the majority of errors in solutions generated by LLMs. LLMs struggles with accurate algebraic and arithmetic computations resulting in errors within the reasoning and final answer.



Dataset: PhysicsQA

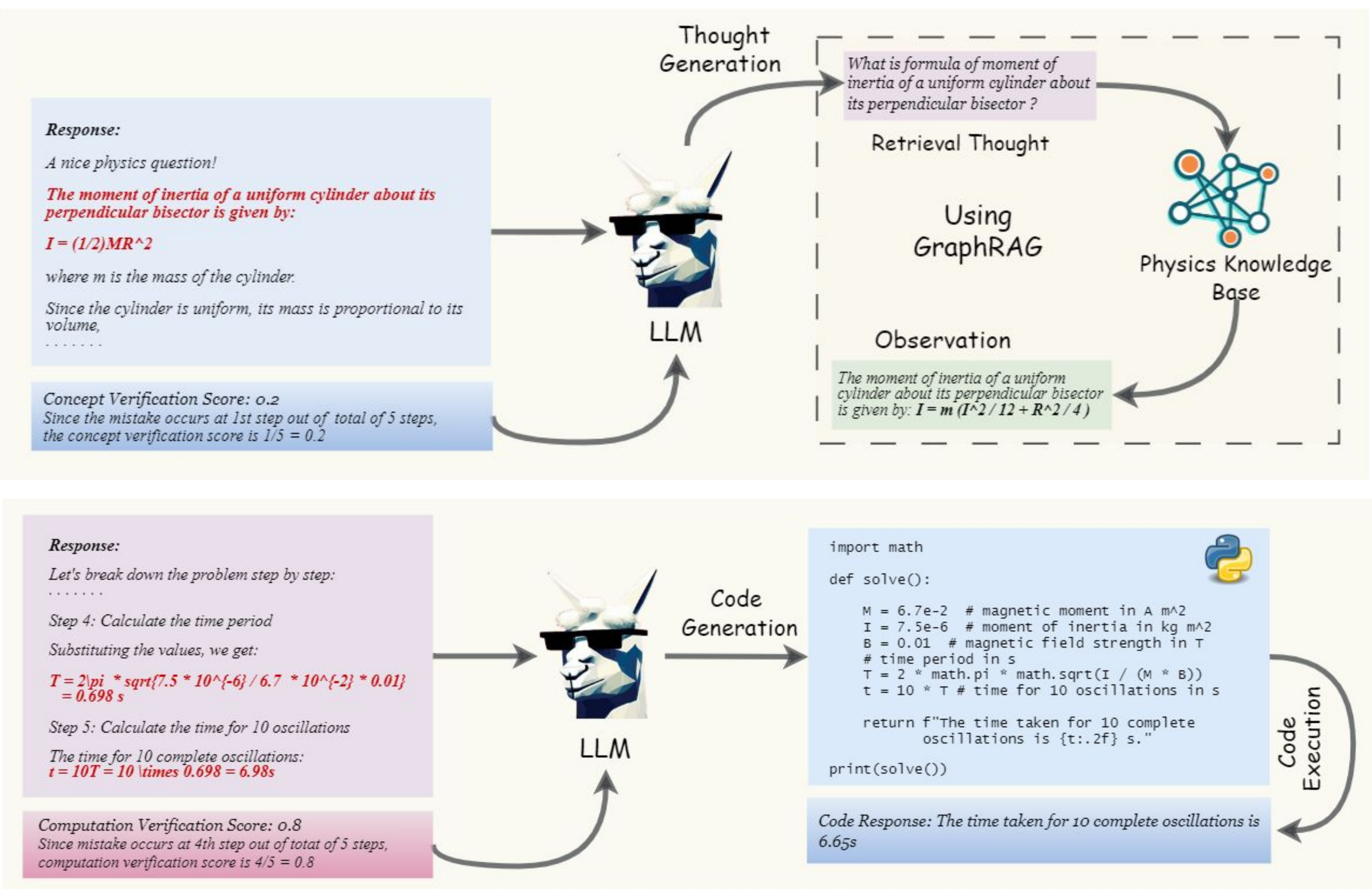
Benchmarks like MMLU, SciEval and ScienceQA focus on foundational knowledge and general reasoning, while more challenging ones like OlympiadBench and JEEBench require advanced reasoning skills. To bridge the gap, we curated our own dataset PhysicsQA, containing set of 370 diverse, intermediate level high school physics problems that provide a balanced challenge, allowing a exhaustive evaluation and step by step solution analysis of open-source LLMs on physics problems. Table 1 illustrates the topic-wise distribution of the questions, providing a clear overview of the areas covered.

Chapter Name	Percentage
Electromagnetism	29.8%
Mechanics and Kinematics	21.8%
Thermodynamics and Heat	15.7%
Waves and Optics	15.4%
Nuclear and Modern Physics	8.9%
Material Properties and Elasticity	8.3%

Table 1: Topic-wise Distribution in PhysicsQA

Mixture of Refinement Agents :

This introduces our mixture of refinement agents (MoRA) framework. We first discuss our motivation behind MoRA; then, we introduce the error identification stage and refinement agents. Finally, we discuss how these agents are routed iteratively to correct different errors in the solutions generated by the LLM.



Algorithm 1: Error Identification and Iterative Refinement

Require: Question Q, Initial Solution S_0, GPT-4o $\mathcal{L}$, Refinement Agents $\mathcal{R}$, Maximum Iterations N, Threshold ϵ
Ensure: Final refined solution to Q

$i = 0, S_i = S_0$
while $i < N$ **do**
  $(F^i_{\text{obj}}, F^i_{\text{val}}, Score^i_{\text{concept}}, Score^i_{\text{comp}}) \leftarrow \mathcal{L}(Q, S_i)$
  if $F^i_{\text{obj}} == -1$ **or** $F^i_{\text{val}} == -1$ **then**
    $S_{i+1} \leftarrow \mathcal{R}_{\text{miscomprehension}}(Q, S_i)$
  else if $Score^i_{\text{concept}} < 1 - \epsilon$ **then**
    $S_{i+1} \leftarrow \mathcal{R}_{\text{concept}}(Q, S_i)$
  else if $Score^i_{\text{comp}} < 1 - \epsilon$ **then**
    $S_{i+1} \leftarrow \mathcal{R}_{\text{computation}}(Q, S_i)$
  else
    return S_i
  end if
  $i \leftarrow i + 1$
end while
return S_N

Setup :

- **Datasets :** In our experiments, we use four datasets: SciEval-Static, PhysicsQA, MMLU High School and MMLU College. SciEval-Static is a subset of SciEVal , consisting 164 questions from physics divided into multiple sub-topics. MMLU , consists of a 118 College level and 173 high school multiple-choice questions from various disciplines.
- **LLMs :** We utilize the API of a range of models with varying parameters and capabilities including LLaMa-3-70B, LLaMa 3.1-405B, Gemma-2-27B, Gemini-1.5-Flash, GPT- 3.5 Turbo and GPT-4 as our LLMs for the evaluation. We use same prompts for all the datasets and LLMs during our Evaluation.
- **Baselines :** We employ an Answer-only approach (AO), where the model is given a question with four options and asked to select the correct answer without any explanation relying solely on its pre-existing knowledge . In contrast, few-shot prompting uses a few examples to help the model learn and apply that knowledge to similar tasks. Chain-of-Thought (CoT) prompting guides the model to generate intermediate reasoning steps, improving its performance on complex tasks by breaking them down into smaller, more manageable parts. These three approaches form our primary baselines.
- **Evaluation :** Most of the existing works measure the mathematical reasoning quality of LLMs by directly comparing the final answer and calculating the overall accuracy on a given dataset. We choose to follow the same evaluation for physics reasoning as well.

Model	SciEval-Static			PhysicsQA			MMLU - High			MMLU - College		
	AO	CoT	3-Shot	AO	CoT	3-Shot	AO	CoT	3-Shot	AO	CoT	3-Shot
LLaMa-3-70B	70.07%	82.23%	63.41%	38.37%	56.76%	59.29%	60.16%	72.88%	73.66%	59.41%	71.76%	71.76%
LLaMa 3.1 405B	**79.87%**	89.63%	**82.92%**	**50.81%**	76.75%	**78.37%**	**72%**	91.52%	**88.98%**	**75.29%**	**88.23%**	**85.29%**
Gemma-2-27B	60.36%	79.26%	53.04%	39.18%	54.59%	59.45%	55.93%	77.11%	74.45%	51.11%	73.52%	67.64%
Gemini 1.5 Flash	68.29%	85.97%	81.70%	44.86%	62.97%	69.72%	58.47%	79.66%	80.05%	60.58%	72.35%	72.94%
GPT 3.5 Turbo	41.46%	66.46%	48.78%	28.10%	42.70%	42.70%	47.45%	58.47%	33.89%	35.29%	50.58%	42.35%
GPT4o	64.02%	**92.68%**	81.09%	49.45%	**79.45%**	**78.37%**	62.71%	**94.06%**	87.28%	70%	84.70%	84.17%

Table 2: Experimentation of Answer-Only (AO) , COT and Few-Shot (3-shot) on different Datasets

Model	Dataset	AO	COT	3-Shot	MORA
Gemma 2 27B	MMLU College	51.11%	73.52%	67.64%	**82.20%**
	MMLU High School	55.93%	77.11%	74.45%	**75.88%**
	PhysicsQA	39.18%	54.59%	59.45%	**70.62%**
	SciEval-Static	60.36%	79.26%	53.04%	**88.76%**
LLaMa 3 70B	MMLU College	59.41%	71.76%	71.76%	**78.82%**
	MMLU High School	60.16%	72.88%	73.66%	**78.81%**
	PhysicsQA	38.37%	56.76%	59.29%	**70.14%**
	SciEval-Static	70.07%	82.23%	63.41%	**86.58%**

Table 3: Comparison of baseline approaches with MoRA across four datasets: SciEval-Static, PhysicsQA, MMLU High School and College based on final answer accuracy.

Error Analysis :

- LLMs demonstrate good problem comprehension ability for physics question.
- Open source LLMs sometimes struggles to retrieve correct physics concept and formulae while reasoning.
- Open-source LLMs struggles with algebraic and arithmetic computation required while solving physics questions.

Error Type	Dataset	GPT-4o	Gemma 2-27B	LLaMa 3-70B
Computational Error	MMLU College	2.54%	5.08%	9.32%
	MMLU High School	2.35%	3.53%	6.47%
	PhysicsQA	8.92%	22.16%	21.08%
	SciEval-Static	3.06%	10.37%	10.98%
Problem Miscomprehe-nsion	MMLU College	0.00%	0.85%	1.69%
	MMLU High School	0.59%	0.59%	1.18%
	PhysicsQA	0.54%	2.16%	1.62%
	SciEval-Static	0.00%	1.22%	1.83%
Wrong Concept	MMLU College	0.85%	12.71%	10.17%
	MMLU High School	3.53%	8.24%	11.76%
	PhysicsQA	7.57%	17.11%	18.92%
	SciEval-Static	1.22%	7.36%	9.15%

Table 4: Error Analysis of incorrect physics CoT solutions of different models across four datasets.

Ablation :

- Problem miscomprehension errors are mitigated with simple instruction prompting and error feedback.
- Open-source LLM performers moderately in identifying the conceptual mistake and retrieval thought generation.
- Using code-driven refinement significantly corrects the computational errors.

Error Type	Dataset	Gemma 2-27B	LLaMa 3-70B
Computational Refinement	MMLU College	100%	81.8%
	MMLU High School	33.3%	75.0%
	PhysicsQA	73.3%	72.6%
	SciEval-Static	57.1%	60.0%
Miscomprehension Refinement	MMLU College	37.5%	33.3%
	MMLU High School	16.7%	37.5%
	PhysicsQA	48.7%	46.9%
	SciEval-Static	62.5%	57.1%
Concept Refinement	MMLU College	100%	100%
	MMLU High School	100%	100%
	PhysicsQA	62.5%	66.7%
	SciEval-Static	100%	66.7%

Table 5: Ablation studies for different refinement agent in MoRA using Gemma-2-27B and Llama-3-70B across four datasets, evaluated by refinement rate.