International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 14 Issue 4 April 2025 || PP 125-129

Advancing LLM Capabilities in Physics: A Reinforcement Learning Approach with Human and AI Feedback

AnshikaTiwari[1] , Gopindra Kumar[2]
Scholar M.Tech(C.S.E) ABSSIT[1] , HOD Dept. of CSE, ABSSIT[2]

Abstract
Currently, Large Language Models (LLMs) have shown great performance in general tasks like text summarization.However, they often struggle with complex arithmetic questions and mathematical reasoning tasks.While simple approaches, such as fine-tuning LLMs for specific mathematical problem-solving tasks, have improved their reasoning abilities, they still face chal-lengeswhenencounteringnewquestionsorvariationsinproblem-solvingapproaches. Thisstudy aims to enhance LLMs to effectively handle physics-related questions extracted from the PhyQA dataset, which consists of problem sets from Indian NCERT textbooks for grades 11 and 12. We employReinforcementLearningtoimprovetheefficacyandaccuracyofourmodelsinarithmetic reasoningforthePhyQAdataset. Variousreinforcementlearningmethods,includingDPO,ReMax, andPPOoptimization,areexploredtoassesstheirperformanceinphysicsproblem-solvingacross different scenarios.A crucial aspect of our approach involves integrating human and artificial intelligencefeedback,referredtoasReinforcementLearningwithHumanandArtificialIntelligence Feedback.Thisinnovativeapproachhelpstrainourmodelstogeneratemorelogicalandreasonable solutionstophysicsproblems. Inevaluations,theMISTRAL-PPOmodelstandsoutforitsability toproducereasonablesolutions,achievingcommendablescoressuchasa58.67%METEORScore, an80.39%ReasoningScore,and38.0%accuracyinamanualevaluationof100randomsamples, quantitatively measuring the model's reasoning abilities.

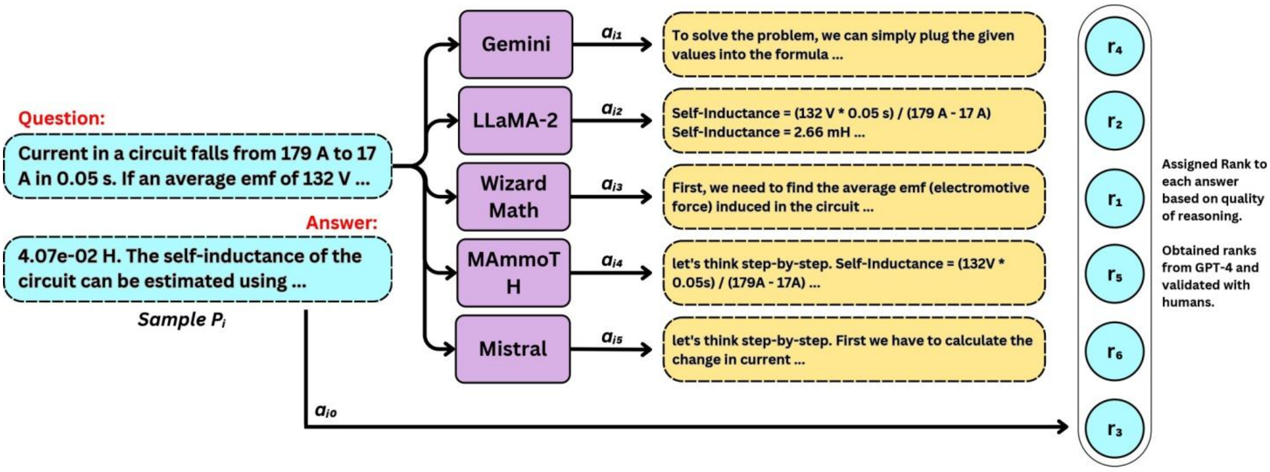

Figure4.7:OurinnovativemethodforprioritizingresponsesforthePreferenceDataset

Inthequesttoimprovealanguagemodel'sabilitytogenerateresponsesthatbettermatchhuman preferences,[52]introducedReinforcementLearningfromHumanFeedback(RLHF).RLHFisa machine learning approach that integrates reinforcement learning techniques, including rewards andcomparisons,withhumanguidancetotrainanartificialintelligenceagent. TheRLHFprocess unfoldsinthreedistinctphases: collectinghumanfeedback,trainingtherewardmodel,andrefining thepolicy.AttheheartofRLHFliespreferencedata,whichinvolvesratingandcomparingvarious responsesgeneratedinresponsetothesameprompt.However,gatheringhumanfeedbackto construct preference dataposes a significant challenge.Obtaining high-quality feedback fromhumansandaccountingforthepotentialsub-optimalnatureofhumaninputcanbequitecomplex. To address this challenge, we introduced RLHAIF, which combines both human feedback and AI feedback to incorporate diverse preference datasets.The preference data generated through modern LLMs have the potential to enhance generalization abilities and improve robustness to various response patterns.

Methodology Dataset

Thedatasetusedforexperimentation,knownas**PhyQA**,isanextensionofSCIMAT'sscience problems[3,4].Itcontains9.5Khighschoolphysicsquestionsandanswers,coveringtopicstaught tostudentsaged15-19,suchasAlternatingCurrent,AtomsandNuclei,andmore.Figure4.8shows the distribution of problems across topics.

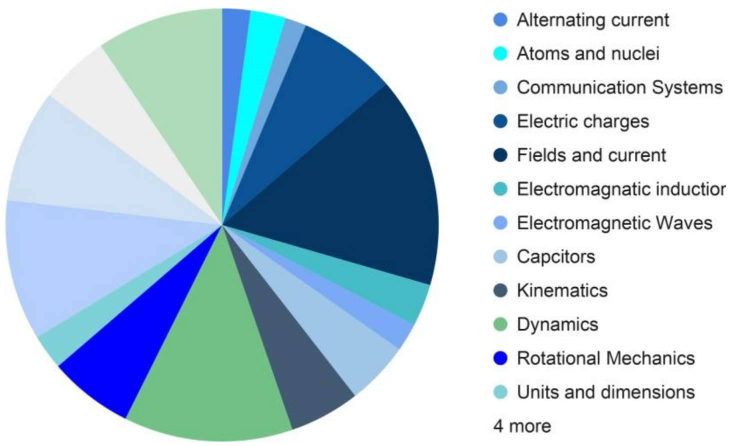

Figure4.8:PhyQATopicDistribution

Analyzing PhyQA provides insights into question and solution characteristics, aiding Large LanguageModels(LLMs)performance.Questionsaverage35.74words,whilesolutionsaverage 54.95 words, with maximum lengths of 75 and 220 words, respectively.Concise and precise solutions in PhyQA help LLMs better understand and address questions.The PhyQA dataset P comprises8100samples. Eachsampleincludesaquestion$q_i$anditscorrespondinganswera_{i0}. To augmentthedataset,weexpandedourinvestigationbygeneratinganswersusingfouropen-source largelanguagemodels: LLaMA2-7B,WizardMath-7B,Mistral-7B,andMAmmoTH-7B,alongside Gemini, a closed-source model. This expansion yields six answers for each question, offering a broad and diverse array of responses to bolster our research efforts.

WeproceedtorankTheseanswersonascalefrom1to6basedonthequalityoftheirreasoning, using prompts similar in detail to [34].Lower ranks indicate higher-quality reasoning in the answers.Toestablishtheserankings,weinitiallyemployGPT-4togeneraterankings,which arethenevaluatedandre-rankedbyhumanevaluators.Thisprocessiscarriedouttoaddress any inaccuracies in the rankings generated by GPT-4. Following this, we form pairs of answers, designating one to be accepted and the other to be rejected. For each data sample P_i, we generate three distinct pairs of answers based on the rankings.

ThismodifieddatasetisthenemployedfortrainingtheRewardModelforwhichwehaveused **LLaMA-213B**model. Thevisualrepresentationofourpreferencedatacreationprocessisshown inFigure.4.7.Fortheexperiments,wehaveusedthe7BvariantsofthefollowingLLMs,LLaMA2, WizardMath,MetaMath,LLeMMA,andMistral.Additionally,wehavedividedtheexplanationof ourapproachandexperimentalsetupintothreepartsforclarity. Firstly,wedelveintodifferentRL algorithmslikeDPO,PPO,ReMaxandtheirsetup. Secondly,weexploreChainofThought(CoT) prompting techniques, and for this we have experimented with both zero-shot and few-shot CoT. Lastly, we discuss the Recall Prompting technique.

Parameter	**PPO**	**DPO**	**ReMax**
KLCoefficient	0.2	0.2	0.2
Epochs	3	3	1
BatchSize	4	2	1
GradientAccumulation	2	8	1
Learningrate	3e-5	3e-5	1e-6

Table 4.10: Hyper-parameter configuration used in training RL models with different RL Policy Optimization Methods.

Results&Analysis

We have conducted extensive evaluations to assess the performance of the models and the various approaches.The evaluation comprises thorough error analysis, accuracy assessments,and reasoning scoring, providing a comprehensive understanding of each model's strengths and weaknesses.

Model	**Setting**	**METEOR**	**BLUE-1**	**BLUE-2**	**BLUE-3**	**BLUE-4**	**ROUGE1**	**ROUGE2**	**ROUGEL**	**ROUGELSUM**	**BERTScore**
LLaMA2-7B	0-Shot	36.65	18.73	13.55	10.64	8.35	31.97	17.07	22.01	27.11	79.07
	3-Shot	25.28	20.90	13.61	9.78	7.18	29.07	12.57	20.55	24.18	76.71
	SFT	28.09	8.37	5.54	4.08	2.97	20.65	9.69	13.56	16.37	76.87
	PPO	39.32	7.10	6.27	5.90	5.33	31.34	24.58	28.23	28.98	82.48
	DPO	35.64	18.74	13.24	9.99	7.49	33.58	17.78	24.07	27.78	80.19
	Remax	37.85	25.69	19.49	16.04	13.08	37.72	22.59	29.46	31.70	81.26
	Recall	23.82	21.00	13.64	9.91	7.20	26.72	12.15	18.78	21.19	74.16
Mistral	0-Shot	28.59	15.42	10.54	7.93	5.95	26.25	13.03	18.77	22.37	75.76
	3-Shot	17.59	13.51	9.26	7.0	5.24	18.79	8.36	14.43	16.2	72.93
	SFT	25.53	6.58	4.62	3.6	2.74	19.97	9.98	13.53	16.18	77.08
	PPO	**58.67**	**40.04**	**35.87**	**34.5**	**32.81**	**57.94**	**51.55**	**56.32**	**56.53**	**87.49**
	DPO	29.94	13.79	8.69	6.08	4.15	29.68	13.3	19.59	23.56	77.42
	Remax	-	-	-	-	-	-	-	-	-	-
	Recall	20.19	10.06	6.71	4.95	3.59	21.35	9.11	15.24	17.56	73.1

Table4.11:EvaluationofLLaMA-2andMistralmodelwithdifferentsettings(0-shot,3-shot,SFT, PPO, DPO, ReMax, Recall) using various scoring metrics (BLUE, ROUGE, METEOR, BERT).

In Table 4.11, I've only shared the results of top two models, the results of other models are presentedinthepaper. Inaanalysisofvariousmodelsettings,aspresentedinTable4.11,Mistral-PPOachievedthehighestoverallscores,consistentlyscoringapproximately35.0acrossBLEU-1to BLEU-4 metrics, with a METEOR score of 58.67. This consistency indicates a strong alignment between the predicted and target words.Additionally, the LLaMA2-7B model demonstrated impressiveperformance,especiallyinaligningwithpreferredanswers.However,itslightlylagged behind Mistral-PPO in accurately matching specific words and displayed limitations in semantic understanding and solving arithmetic problems.

AlthoughMistralshowcasesstronglogicalandmathematicalreasoningskillsincertaincontexts, it occasionally commits notable errors in insignificant stages.These inaccuracies in problem analysis, concept recall, and application underscore further avenues for exploration.

Toassesstheprecisionofourmodels,weexamineda100randomsamplesfromthedataset.

Table4.12displaysthecomparativeoutcomesasfollows:

- GPT-4leadswitha72%accuracyrate,followedbyGPT-3.5at40%.
- Mistral,employingthePPOpolicy,achievesa38.0%accuracy,whileLLaMA2-7Bwiththe PPO policy registers an 18% accuracy.

These findings illustrate that although GPT-3.5 outperforms our suggested top model, it still encounters computational errors, and occasionally our proposed model outperforms it.With scalability, we have the potential to surpass GPT-3.5, as the results are not significantly different from those of the Mistral-PPO 7B model.

Model	**Setting**	**Correct**	**Wrong**	**Total**
LLaMA2-7B	SFT	9	91	100
	PPO	18	82	100
	DPO	10	90	100
	Recall	14	86	100
Mistral	SFT	21	79	100
	PPO	**38**	62	100
	DPO	22	78	100
	Recall	16	84	100
GPT-3.5	—	40	60	100
GPT-4	—	72	28	100

Table4.12: Model'soutputperformancewithHumanEvaluations

For a detailed analysis of the reasoning evaluation of each response, we have formulated a six-step reasoning assessment.Each skill point is assessed according to the LLM's proficiencyin executing specific problem-solving aspects, including identifying the **correct context (CA)**, interpreting**physicsconcepts(PD)**,**performingcalculations(AC)**,maintaining**logicalcoherence (LR)**, demonstrating an **understanding of concepts (CU)**, and identifying or **rectifying errors (ED)**.

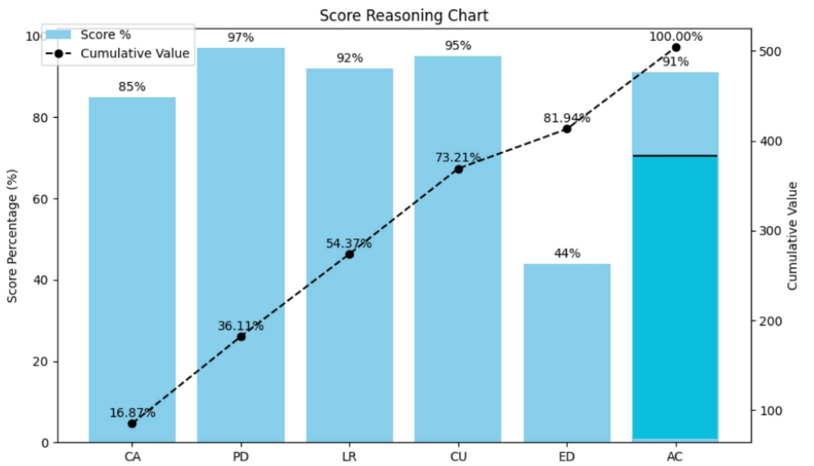

Figure 4.9: Reasoning Score Distribution on Mistral-PPO's 100 random sample responses
Figure.4.9alsodemonstratesthatLLMfaceschallengesinarithmeticcalculationsinapproxi-mately91.0%ofcases,asdeterminedbyassessmentsfromHumanAnnotatorsacross100responses

fromtheMistral-PPOmodel. Outofthistotal,themodelaccuratelyfollowsthesequenceofsteps andformulastosolvethesolution76.92%ofthetime.However,inapproximately23.08%ofcases, themodelfailstoexecuteaccuratearithmeticcalculations.Thisleadstoanoverallscoreof62.0for incorrect answers, as shown in Table. 4.12.

PaperConclusion

This study presents RLHAIF, an effective and efficient strategy for improving the physics problem-solving capabilities of large language models (LLMs) and aligning their responses more closely with human preferences.The revolutionary RLHAIF methodology ranks answers using humanandAI-generatedinput,resultinginamorediversifiedandresilientmodeltrainingprocess. Experiments with various LLMs consistently demonstrate that the Mistral-PPO created from this approachbeatsitsequivalents,establishingitselfasareliablebenchmarkforthePhyQAdataset.By bridgingthegapbetweenhumanintuitionandLLMreplies,RLHAIFhasthepotentialtoadvance natural language understanding and generation significantly.

Limitations&FutureScope

The possibility of students misusing the model for cheating in assignments or quizzes raises ethical concerns. To tackle this issue, future research could explore shifting focus from providing direct answers to offering hints and structured reasoning, with the aim of enhancing students' conceptual understanding and problem-solving skills.
Ourinvestigationhasbeenconstrainedtoa7Bmodelduetocomputationalcostsandresource efficiency. Future directions could involve exploring the performance spectrum of larger models with increased parameters, such as 13B or 70B.