SciPhyRAG - Retrieval Augmentation to Improve LLMs on Physics Q&A

Avinash Anand(✉), Arnav Goel, Medha Hira, Snehal Buldeo, Jatin Kumar, Astha Verma, Rushali Gupta, and Rajiv Ratn Shah

Indraprastha Institute of Information Technology, Delhi, India
{avinasha,arnav21519,medha21265,snehal22074,jatin20206,
asthav,rajivratn}@iiitd.ac.in

Abstract. Large Language Models (LLMs) have showcased their value across diverse domains, yet their efficacy in computationally intensive tasks remains limited in accuracy. This paper introduces a comprehensive methodology to construct a resilient dataset focused on High School Physics, leveraging retrieval augmentation. Subsequent finetuning of a Large Language Model through instructional calibration is proposed to elevate outcome precision and depth. The central aspiration is reinforcing LLM efficiency in educational contexts, facilitating more precise, well-contextualized, and informative results. By bridging the gap between LLM capabilities and the demands of complex educational tasks, this approach seeks to empower educators and students alike, offering enhanced support and enriched learning experiences. Compared to Vicuna-7b, the finetuned retrieval augmented model **SciPhy-RAG** exhibits a **16.67% increase** in BERTScore and **35.2%** increase on ROUGE-2 scores. This approach has the potential to be used to reshape Physics Q&A by LLMs and has a lasting impact on their use for Physics education. Furthermore, the data sets released can be a reference point for future research and educational domain tasks such as **Automatic Evaluation** and **Question Generation**.

Keywords: Document Retrieval · Neural Text Generation · Large Language Models · Natural Language Processing · Question-Answering

1 Introduction

The rising popularity of transformer-based [4] Large Language Models can be attributed to their exceptional performance in tasks such as text generation, question answering, and document summarization [1]. Recent advancements in language model architectures, such as the GPT-3.5 [2], PaLM [30], and LLAMA [8] have showcased their remarkable ability to comprehend and generate human-like text. However, when it comes to domain-specific computational tasks, such as solving physics problems, language models often struggle to achieve the desired level of accuracy.

V. Goyal et al. (Eds.): BDA 2023, LNCS 14418, pp. 50–63, 2023.
https://doi.org/10.1007/978-3-031-49601-1_4

Datasets have played an instrumental role in pushing forward performance on domain-specific tasks. In this research paper, we address the challenge of improving the accuracy of large language models (LLMs) for computational tasks by finetuning the model on a high school physics dataset we designed. We also introduce a high-quality corpus containing high school physics concepts from the NCERT textbook, considered a higher-education physics standard. The corpus is annotated manually to ensure its reliability and investigate the effectiveness of retrieval-augmentation techniques in further enhancing the model's performance.

Physics problem-solving requires an in-depth understanding of all underlying concepts and the ability to apply them sequentially and logically. Traditional approaches to computational tasks rely on rule-based algorithms or symbolic manipulation [32,33]. However, recent advancements in deep learning and language modelling have presented an opportunity to leverage the power of LLMs for these tasks. We believe that the efficacy of LLMs in domain-specific computational tasks would be a pivotal step in their usage for education and learning.

Language Models (LMs) rely on their inherent parameters to generate responses based on their knowledge accumulated via training of a huge corpus. With thousands of parameters holding the vast information, instances arise where domain-specific expertise, particularly prominent in complex fields such as physics question answering, demands enhanced support [15]. The notable complexities of physics queries suggest a potential advantage in incorporating relevant passages within the prompt. Given the complex and knowledge-intensive nature of physics problems, such an approach becomes imperative. In light of this, retrieval mechanisms offer a viable avenue to explore.

This paper explores retrieval augmentation techniques to enhance the finetuned LLM's performance using good-quality support passages. Retrieval enhancement involves incorporating a retrieval-based model to provide additional context and guidance to improve the model's reasoning capabilities.

The contributions of this research paper are twofold. First, we present a high school physics dataset corpus using NCERT textbook content, with good-quality annotations for mathematical equations specifically curated to enhance the precision of LLMs in computational tasks and serve as a benchmark for evaluating their performance. Second, we prepare a novel retrieval pipeline with a self-annotated document corpus to enhance model performance.

By addressing the limitations of LLMs in computational problem-solving and leveraging the potential of finetuning and retrieval augmentation, we aim to push the boundaries of language models' accuracy for high school physics tasks. The insights gained from this research have the potential to revolutionize educational technologies, enabling intelligent tutoring systems and personalized learning experiences that support students on their journey to master complex scientific concepts.

The written work is structured as follows: Sect. 2 addresses the related works on math and scientific problem solvers, and Sect. 3 explains the process behind data collection, annotation and augmentation. Section 4 explains how we performed our experiments, and Sect. 5 describes our evaluation metrics used and

the results obtained. The paper's future scope and conclusion are summarised in Sect. 6 respectively.

2 Related Work

2.1 Mathematics and Science Solvers

We encountered several Mathematics and Science domain Q&A datasets in our exploration of datasets. However, we encountered a significant absence of high-quality and challenging physics question-answering datasets.

AQuA-RAT [11] is a dataset comprising more than 100K algebraic word problems with natural language rationales. Each data point in the dataset provides four values, i.e. the question, the options, the rationale behind the solution and the correct option.

The **MathQA** [12] was created using a novel representation language to annotate AQuA-RAT and correct the noisy-incorrect data. The limitation of this approach is its exclusive focus on multiple-choice questions. However, we found the rationale parameter intriguing as it empowers the finetuned model to offer logical explanations, aiding users in deeper comprehension of the problem and its solution.

GSM8K [6] claims to contain questions that can be solvable by a bright middle school student. Therefore, it cannot be used to finetune a model for complex reasoning and computational tasks. **MATH** [7] is a dataset of 12.5K challenging mathematics problems in competition. Recently, there has also been an emphasis on evaluating the capabilities of models in answering open-domain science questions.

ScienceQA [14], a dataset collected from elementary and high school science curricula, contains 21K multimodal multiple-choice science questions. **ASDiv (Academia Sinica Diverse MWP Dataset)** [18] is a notable dataset where each math word problem(MWP) is associated with a problem type and grade-level.

The **SCIMAT (Science and Mathematics Dataset)** [5] is a valuable resource as it provides chapter-wise questions for class 11 and 12 Mathematics and Science subjects, together with numerical answers. However, a notable limitation of the dataset is the absence of explicit explanations for the solutions or the underlying formulas related to the topics. This absence of contextual information hinders the dataset's effectiveness in fully supporting comprehensive learning and understanding.

2.2 Document Retrieval in Science Q&A

In Large Language Models, the learned knowledge about the outside world is implicitly stored in the parameters of the underlying neural network. Finding out what knowledge is stored in the network and where becomes challenging as a result [15]. The network's size also affects how much data can be stored; to capture more global information, one must train larger and larger networks, which

might be prohibitively slow or expensive. To bridge this gap, adding context to a given query in the input of a Large Language Model is helpful.

[13] mentions how generative models for question answering can benefit from passage retrieval. It retrieves support text passages from an external source of knowledge, such as Wikipedia. Then, a generative encoder-decoder model produces the answer, conditioned on the question and the retrieved passages.

This approach has obtained state-of-the-art results on the **Natural Questions** [16] and **TriviaQA** [17] open benchmarks. This idea can be extended to domain-specific question answering. Given the complex nature of Physics and Math Questions, we assume that presenting the model with a set of relevant passages containing domain knowledge and required formulas can help the generative model answer the questions more accurately.

Most of the research in retrieval-augmented generation has been directed towards open-domain question answering. This involves extracting relevant information from external sources like Wikipedia to answer questions across various topics [15] [13]. The wide range of information in Wikipedia greatly aids open domain question answering.

Previously, domain-specific knowledge has been used to improve the performance of information retrieval systems. [34] created a data set by collecting question and answer pairs from the internet in the insurance domain. Our assumption in this work is that a corpus containing information in the physics domain would increase a model's performance on physics question-answering tasks.

3 Datasets

To further advance the application of language models in physics, we propose the creation of a comprehensive and challenging dataset. This study thus releases two open-source datasets :

1. **PhyQA** consists of 9.5K high school physics questions and answers with step-wise explanations. The dataset is diverse and consists of topics studied by high school physics students in the range of 15–19 years of age are included in the dataset. The list of topics includes: Alternating Current; Atoms and Nuclei; Communication Systems; Electric Charges, Fields and Current; Electromagnetic Induction; Electromagnetic Waves; Capacitors; Dynamics and Rotational Mechanics; Units, Dimensions and Kinematics; Ray and Wave Optics; Thermodynamics and Heat; Gaseous State; Waves, Sound and Oscillations. Each topic is divided into several subtopics. Both of these datasets are meticulously formatted to make it easy to train and test on open-source language models such as: **LLAMA** [8], **Alpaca** [20] and **Vicuna** [10].
2. The **RetriPhy Corpus** is a curated collection of Physics chapter content from NCERT books for 11^{th} and 12^{th} grades. It is annotated manually with LaTeX representations of equations and examples. The dataset includes 14 chapters from each grade, 11^{th} and 12^{th}, respectively. When prompted with a question, the combined corpus of all chapters creates context passages using the retrieval pipeline.

The proposed data sets serve multiple important purposes. Firstly, they enable the training of language models on complex reasoning and computational tasks specific to physics, which requires a deeper understanding of scientific principles and relationships between variables. Second, they could serve as a valuable benchmark to assess the performance of existing language models in physics problem-solving, allowing for a meaningful comparison of different models and driving further advancements in natural language understanding within the physics domain. Third, our retrieval corpus can be used to prepare and benchmark retrieval systems and to retrieve high-quality passages for physics Q&A (Fig. 1).

3.1 PhyQA

The dataset comprises 9.5K Physics questions, with each chapter having nearly equal representation. Each data point in the dataset is associated with two keys, i.e. "instruction" and "output", to organize the information. They are described as follows:

- **Instruction**: Key containing the question which needs to be answered by the model
- **Output**: Key containing the corresponding numerical answer and a detailed explanation of how that answer was obtained.

An example of a question from the chapter Newton's Laws of Motion is given below:

Instruction: An aircraft of mass 176 kg executes a horizontal loop at a speed of 249 m/s with its wings banked at 80 degrees. What is the radius of the loop?

Output: Answer: *35162.514 m* <sep> Explanation: *To calculate the radius, we use the formula*

$$\frac{(v^2)}{(g * \tan(\frac{param * \pi}{180}))}$$

where v is the speed $(249m/s)$*, g is the acceleration due to gravity* $(10m/s^2)$*, and param is the angle of banking* $80°$*. Substituting the values,*

$$R = \frac{249 \cdot 249}{10 \cdot \tan(\frac{80 * \pi}{180})} = 35162.514m$$

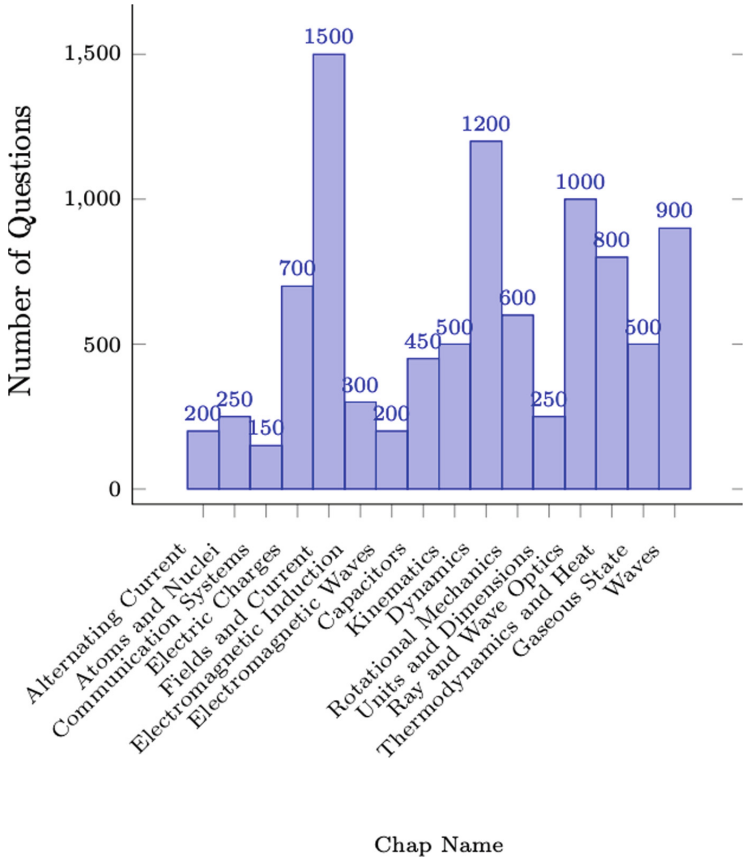
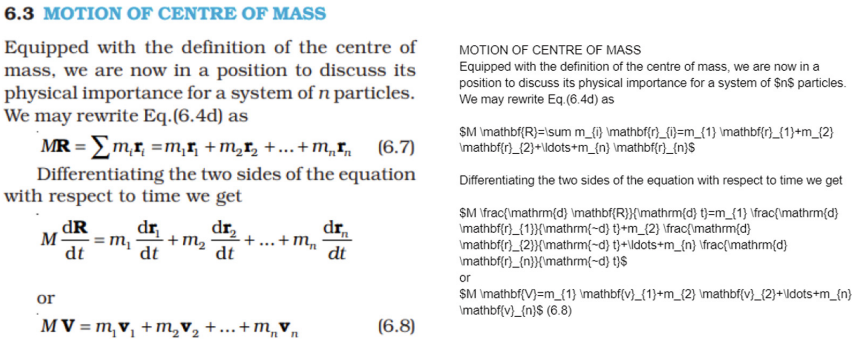


Fig. 1. Insight into the number of questions at chapter-wise granularity in PhyQA

3.2 RetriPhy Corpus

The RetriPhy Corpus comprises content extracted from NCERT books of Physics subjects for 11^{th} and 12^{th} grades. These NCERT (National Council of Educational Research and Training) books are known for their concise, accurate, and easily understandable presentation of concepts. The corpus contains theorems, equations, solving numerical problems, and explanations spanning various topics such as electric charges, gravitation, optics, atoms and nuclei, etc.

Containing material from all 14 chapters of the 11^{th} and 12^{th} grade NCERT Physics books, the corpus comprises **28 documents**, each corresponding to a chapter. Each chapter contains around **30–32 paragraphs**, with approximately 400 tokens per paragraph. The corpus extracts a total of **927 paragraphs**. We have also included an overlapping of 20 tokens in these paragraphs to ensure consistency (Fig. 2).



6.3 MOTION OF CENTRE OF MASS

Equipped with the definition of the centre of mass, we are now in a position to discuss its physical importance for a system of n particles. We may rewrite Eq.(6.4d) as

$$M\mathbf{R} = \sum m_i \mathbf{r}_i = m_1\mathbf{r}_1 + m_2\mathbf{r}_2 + \ldots + m_n\mathbf{r}_n \quad (6.7)$$

Differentiating the two sides of the equation with respect to time we get

$$M\frac{d\mathbf{R}}{dt} = m_1\frac{d\mathbf{r}_1}{dt} + m_2\frac{d\mathbf{r}_2}{dt} + \ldots + m_n\frac{d\mathbf{r}_n}{dt}$$

or

$$M\mathbf{V} = m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + \ldots + m_n\mathbf{v}_n \quad (6.8)$$

MOTION OF CENTRE OF MASS
Equipped with the definition of the centre of mass, we are now in a position to discuss its physical importance for a system of n particles. We may rewrite Eq.(6.4d) as

$M \mathbf{R}=\sum m_{i} \mathbf{r}_{i}=m_{1} \mathbf{r}_{1}+m_{2} \mathbf{r}_{2}+\ldots+m_{n} \mathbf{r}_{n}$

Differentiating the two sides of the equation with respect to time we get

$M \frac{\mathrm{d} \mathbf{R}}{\mathrm{d} t}=m_{1} \frac{\mathrm{d} \mathbf{r}_{1}}{\mathrm{~d} t}+m_{2} \frac{\mathrm{d} \mathbf{r}_{2}}{\mathrm{~d} t}+\ldots+m_{n} \frac{\mathrm{d} \mathbf{r}_{n}}{\mathrm{~d} t}$
or
$M \mathbf{V}=m_{1} \mathbf{v}_{1}+m_{2} \mathbf{v}_{2}+\ldots+m_{n} \mathbf{v}_{n}$ (6.8)

Fig. 2. Left: A snippet from the Physics textbook of a section about "Motion of Centre of Mass"; **Right**: A snippet of the annotation file corresponding to the textbook snippet

3.3 Data Collection and Augmentation

PhyQA: Data collection started by improving upon SCIMAT's science problems [5]. Additional data was collected by scraping standard Indian High School open-source physics textbooks of classes 11 and 12.

Solvers have improved performance in answering math questions when the finetuning data set undergoes data enhancement transformations [19]. We extend this to physics problem-solving by taking base problems from each sub-topic and applying these transformations to include a wider variety of questions. These transformations are two in nature:

- Substitution: Changing the values of constants in a question.
- Paraphrasing: Paraphrasing the problem q using a model to generate N candidate questions that differ from the one in which they were written.

RetriPhy: We have used NCERT textbook content to create the RetriPhy (Retrieval-based Physics) corpus. Our motivation for using NCERT textbook content is the concise and easy-to-understand explanation of concepts in these books. Also, PhyQA consists of problems related to the topics of grade 11^{th} and 12^{th} physics; hence, the focus is on the content from the 11^{th} and 12^{th} grade NCERT Physics textbooks.

These chapter-wise documents on physics are accessible on the official NCERT website. Our methodology involved using these documents to retrieve textual content from the chapters. To ensure an accurate representation of mathematical symbols and equations, we employed LaTeX annotations, thereby eliminating the potential for ambiguity in text interpretation.

3.4 Data Annotation

PhyQA: Our team consisted of five dataset annotators, each of whom had graduated high school and studied physics until class 12. Upon self-evaluation

on a scale of 5, the annotators rated themselves as 3, 3, 4, 4 and 5. We verified this self-evaluation by giving them a small test on basic questions to check fundamental understanding.

They used this in-depth understanding of physics concepts to annotate the solutions and provide relevant formulae and concepts. The annotated questions were then shuffled amongst the remaining annotators to evaluate the dataset's quality. Upon this, the annotators who rated themselves as 3 could solve only 55% problems given to them, while the one rated 5 could solve 80% of the problems. This shows that the dataset has high-quality questions and can be challenging for humans, too.

RetriPhy: The annotation process for the Retriphy corpus is aimed at the accurate representation of mathematical symbols and equations present in the content of NCERT textbooks. For every mathematical notation or equation in the text, we have used LaTeX for its annotation. To identify the start and end of the LaTeX content in the text, we have added a $ symbol as the start and end delimiters (Fig. 2). In the annotation process of RetriPhy, our team comprised three annotators, each contributing to approximately one-third of the annotations. A shared segment of annotations was distributed to all annotators to validate the annotations, allowing for cross-evaluation. This experiment revealed an impressive accuracy rate of 87% among the annotations.

Furthermore, the accuracy of annotations for all chapters is verified by both the annotators and an expert in the domain.

3.5 Inter-Annotator Agreement for Data Validation

In the data validation process, a team of five data annotators, all proficient college students in the domain of high school Physics, was employed. The dataset was equally divided among the annotators to ensure a balanced workload. Within this team, one annotator possessed expert-level knowledge, while the remaining four were classified as having intermediate expertise.

Rigorous attention to detail was exercised throughout the data annotation process to uphold the accuracy of the annotations. A meticulous approach was taken, whereby each segment annotated by one annotator underwent a verification stage involving assessment by two other annotators. This multiple-layer validation strategy was adopted to enhance the reliability of the annotated data.

The **Fleiss' Kappa** score of this annotation process was **0.65**. By combining the expertise of the annotators, the cross-validation process, and the Fleiss' Kappa coefficient application, a robust framework for data validation was established, ensuring the accuracy and integrity of the annotated high school Physics dataset.

4 Experiments

We describe the **SciPhy** retrieval experiment that uses both data sets we released. Language models' performance in question-answering tasks improves

with finetuning on data sets with questions rephrased as instruction-following data points [3]. Following this, we finetune an open-source large language model with **PhyQA**.

We then incorporate retrieval by using our high-quality retrieval document dataset as a database for retrieving documents. We prepare a retrieval pipeline with our document vectorbase and use that to provide context on the test questions to evaluate the accuracy of the model answers on evaluation metrics. We elaborate on this approach in forthcoming subsections.

4.1 Fine-Tuning Using LoRA

Model and Hyperparameters: Vicuna [10] is a large language model prepared by finetuning a Llama base model on 70k user-shared conversations. We used a Vicuna model with 7 billion parameters as our baseline. Natural language processing consists of pre-training language models on general text and finetuning model parameters on domain-specific data. However, as the model size increases, it is computationally expensive to finetune models, which involves retraining all parameters fully. We thus adopt the **L**ow **Ra**nk **A**daptation (LoRA) [9], which proposes freezing model weights and injecting lower rank matrices into transformer layers that can be trained. This reduces training time and the hardware needed to keep model accuracy intact. We hypothesize that finetuning **Vicuna-LoRA** on our annotated **PhyQA** physics dataset will greatly improve the model's capabilities to answer physics questions.

Experiment: For this task, the PhyQA dataset is split into *8000 training samples* and *1500 test samples*. The training set is used to finetune the model. The Vicuna-LoRA model is run with different model weight representations, i.e. an **8-bit** representation and a **16-bit** representation. The LoRA rank r is set to **8** with a LoRA-dropout of **0.05** for preparing the 8-bit finetuned model and is set to **16** for preparing the 16-bit finetuned model. The finetuning is run for **3 epochs** with a batch size of **128** and the learning rate equal to **3e-4** i.e. the Karpathy constant on an NVIDIA RTX A6000 GPU.

4.2 Rationale Behind Retrieval:

Our second experiment hypothesizes that providing relevant context to our language model about the question as input will greatly improve the explanation and precision of the answers. This is based on physics being driven by concepts and their interpretation rather than simply applying formulae. Given a query q, the retrieval-based system is prompted to find N relevant passages. Each passage is retrieved and appended with a $< sep >$ token. Query q is prepended to the N retrieved passages to form the final user query q_f, which can be described as:

$$q_f = q + < sep > + \sum_{i=1}^{n} (N_i + < sep >)$$

q_f is then prompted to the finetuned Vicuna LoRA 7-billion model described in Sect. 4.1 to get the answer.

4.3 Retrieval System Design:

Combined with Vicuna-7b, LoRA finetuned on **PhyQA**, the retrieval system is called **SciPhy-RAG**. The RetriPhy Corpus creates passages of **400 tokens** each.

After creating the passages, the next step is to perform indexing, which addresses the challenge of memory storage. We retrieve relevant passages by using similarity matches between the indexed passages and the user query. However, as the corpus grows with more passages, this task becomes progressively more time-consuming. To efficiently store and index the passages, we adopt a method of representing them as dense vectors [22].

VectorStores, like Pinecone, are used for indexing and storing vector embeddings of text data for fast retrieval. It uses Approximate Nearest Neighbour (ANN) search in higher dimensions [27], which allows for handling large numbers of queries. ANN proposes, when given a set P of n points (in this case, a set of n queries), a metric ball $B_D(q|r)$ in metric space (X, D). It creates a data structure S such that for any query $q \in$ X, S returns a point p that satisfies:

$$D(p, q) \leq r$$

$$\forall p \in B_D(q, cr) \cap P$$

It minimises this for some $c \geq 1$ and returns the point p at the minima. In our approach, each passage is converted into a 384-dimensional dense vector embedding using the **all-MiniLM-L6-v2 model** [12]. These passages are then stored in the VectorStore described above.

Experimentation: When prompted with a user query q, the system applies ANN search to identify top K relevant passages where the user specifies K. The passages are returned by the system and appended to q. This final prompt with the passages q_f is prepared. We use the technique described in Sect. 4.1 to finetune Vicuna-7b using LoRA on PhyQA. The prompt q_f is inputted into the model specified above to obtain our final output.

5 Results

5.1 Evaluation Metrics

We perform a two-tier evaluation of the fine-tuned models. We first choose the evaluation metrics such as BERT-Score [21], METEOR [24] and ROUGE-L, ROUGE-1 and ROUGE-2 [25]. These help us assess the quality and correctness of the explanations generated by the models. We then sample 100 questions from each chapter and prompt the model to give a "One-Word Answer". This gives the numeric answer, and we treat it as a classification task measuring the accuracies with the ground-truth answers of the test set. We call this metric as **Final Answer Accuracy (FAA)**. We repeat this with ten randomly chosen samples and report the lowest accuracies achieved out of the 10.

5.2 Experimental Results and Analysis

Upon finetuning the baseline Vicuna-7B model and attaching the retrieval system, we get two models from the experimental setups described above. Table 1 shows their results on the evaluation metrics described above. Our finetuned SciPhy-RAG 16-bit model (i.e. our retrieval pipeline + Vicuna-LoRA 16-bit) shows a **16.67%** increase on BERT-F1 scores over the base Vicuna-7b model. This shows that finetuning with **PhyQA** and using **RetriPhy** as our retrieval corpus increases the quality of explanations by a significant amount over the base model.

Table 1. Table showing the BERT, ROUGE and METEOR scores of the two finetuned Vicuna-LoRA models

Metric	Vicuna-7B	SciPhy-RAG (8-bit)	SciPhy-RAG (16-bit)
BERT (F1)	0.768	0.887	**0.899**
BERT (Precision)	0.744	**0.876**	0.865
BERT (Recall)	0.784	0.886	**0.895**
METEOR	0.285	0.347	**0.352**
ROUGE-L	0.321	0.371	**0.389**
ROUGE-1	0.315	0.358	**0.363**
ROUGE-2	0.147	0.181	**0.195**

On METEOR, our 8-bit SciPhy-RAG shows a **22.8%** increase. On the ROUGE evaluation metrics (ROUGE-L, ROUGE-1 and ROUGE-2), 16-bit SciPhy-RAG shows a much higher improvement (**19.4%**, **22.2%** and **35.3%** respectively) over the base Vicuna-7b model. We hypothesize that METEOR scores are calculated based on unigram matching between the reference and candidate sentences [23], and the retrieval models drive the model output generation slightly away from the ground truth explanation. The same hypothesis holds for lesser improvements in ROUGE scores. However, the increase in BERT scores validates that the explanations are semantically similar and high quality.

Table 2 shows the final answer accuracy (FAA) for the base-Vicuna-7b model (i.e. before) and the SciPhy-RAG (16-bit) model after finetuning and applying retrieval. Note that the lowest scores have been reported for both models, and this shows a massive increase across chapters, showing improvement despite skewness in the training data. Due to resource constraints, the lower accuracies can be attributed to running these experiments on smaller models. However, our hypothesis can be extended to larger models with sizes of parameters ¿ 50 B and newer architectures.

Table 2. Final Answer Accuracy **before and after Fine-Tuning with PhyQA**

Chapter	Before	After
Alternating Current	21.3	26.2
Atoms and Nuclei	20.8	27.1
Communication Systems	15.3	21.2
Electric Charges	22.9	26.5
Fields and Current	23.5	28.6
Electromagnetic Induction	22.2	29.1
Electromagnetic Waves	23.8	26.3
Capacitors	25.1	28.3
Dynamics & Rotational Mechanics	24.7	29.2
Units, Dimensions & Kinematics	22.4	27.3
Ray and Wave Optics	23.6	26.5
Thermodynamics and Heat	21.5	27.2
Gaseous State	19.6	24.3
Waves, Sound and Oscillations	20.6	28.4

6 Conclusion and Future Work

As we enter the future, we envision several promising avenues for further exploration. One such direction involves the incorporation of the **Chain of Thought** [28] and **Tree of Thought** [29] prompting techniques. Augmenting our datasets to incorporate these prompting techniques could lead to richer explanations and higher accuracies. A custom annotated retrieval dataset could also be prepared for preparing the vector database to improve the quality of retrieved texts.

Another extension area is curating benchmarks for other STEM fields, such as Chemistry and Biology, which will give rise to Multimodal models and further advancements in using LLMs and Artificial Intelligence to hasten up research and improve the quality of learning and education in these fields.

The symbiosis of **PhyQA** and **RetriPhy** with advanced LLM finetuning techniques and retrieval augmentation marks a significant step towards empowering AI-driven physics education. We envision a physics-based Q&A system catering to a student. As we embark on continuous improvement and exploration, we anticipate that our research will pave the way for more intelligent, interactive, and personalized learning experiences in physics and beyond. Additionally, these datasets will help drive research in AI-driven education tasks such as Automatic Evaluation and provide a foundation for making similar datasets.

Acknowledgments. We want to acknowledge the contribution of our data annotators, Aryan Goel, Ansh Varshney, Siddhartha Garg and Saurav Mehra. Rajiv Ratn Shah is partly supported by the Infosys Center for AI, the Center for Design and New Media, and the Center of Excellence in Healthcare at IIIT Delhi.

References

1. Goel, A., Hira, M., Anand, A., Bangar, S., Shah, D.R.R.: Advancements in scientific controllable text generation methods (2023)
2. Brown, T., et al.: Language models are few-shot learners (2020)
3. Chung, H., et al.: Scaling instruction-finetuned language models (2022)
4. Vaswani, A., et al.: Attention is all you need (2017)
5. Chatakonda, S.K., Kollepara, N., Kumar, P.: SCIMAT: dataset of problems in science and mathematics. In: Srirama, S.N., Lin, J.C.-W., Bhatnagar, R., Agarwal, S., Reddy, P.K. (eds.) BDA 2021. LNCS, vol. 13147, pp. 211–226. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-93620-4_16
6. Cobbe, K., et al.: Training verifiers to solve math word problems. ArXiv Preprint ArXiv:2110.14168 (2021)
7. Hendrycks, D., et al.: Measuring mathematical problem solving with the MATH dataset (2021)
8. Touvron, H., et al.: Open and efficient foundation language models. In: LLaMA (2023)
9. Hu, E., et al.: Low-rank adaptation of large language models. In: LoRA (2021)
10. Chiang, W.L., et al.: Vicuna: an open-source chatbot Impressing GPT-4 with 90 (2023)
11. Ling, W., Yogatama, D., Dyer, C., Blunsom, P.: Learning to solve and explain algebraic word problems. In: Program induction by rationale generation (2017)
12. Miao, S., Liang, C., Su, K.: A diverse corpus for evaluating and developing English math word problem solvers (2021)
13. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering (2021)
14. Lu, P., et al.: Multimodal reasoning via thought chains for science question answering, learn to explain (2022)
15. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval-augmented language model pre-training. In: REALM (2020)
16. Kwiatkowski, T., et al.: Natural questions: a benchmark for question answering research. In: Transactions Of The Association For Computational Linguistics, vol. 7, pp. 453–466 (2019)
17. Joshi, M., Choi, E., Weld, D., Zettlemoyer, L.: Triviaqa: a large scale distantly supervised challenge dataset for reading comprehension. ArXiv Preprint ArXiv:1705.03551. (2017)
18. Miao, S., Liang, C., Su, K.: A diverse corpus for evaluating and developing English math word problem solvers. ArXiv. abs/2106.15772 (2020)
19. Kumar, V., Maheshwary, R., Pudi, V.: Data augmentation for math word problem solvers. In: Practice Makes a Solver Perfect (2022)
20. Taori, R., et al.: Hashimoto stanford alpaca: an instruction-following LLaMA model (2023). https://github.com/tatsu-lab/stanford_alpaca
21. Zhang, T., Kishore, V., Wu, F., Weinberger, K., Artzi, Y. Evaluating text generation with BERT. In: BERTScore (2020)
22. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering (2020)
23. Saadany, H.: Constantin orăsan and BLEU, METEOR, BERTScore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In: Proceedings Of The Translation And Interpreting Technology Online Conference TRITON 2021 (2021). https://doi.org/10.26615/978-954-452-071-7_006

24. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of The ACL Workshop On Intrinsic And Extrinsic Evaluation Measures For Machine Translation And/or Summarization, pp. 65–72 (2005)
25. Lin, C.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
26. Wang, W., et al.: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: MiniLM (2020)
27. Andoni, A., Indyk, P., Razenshteyn, I.: Approximate nearest neighbor search in high dimensions (2018)
28. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models (2023)
29. Yao, S., et al.: Deliberate problem solving with large language models. In: Tree of Thoughts (2023)
30. Chowdhery, A., et al.: Scaling language modeling with pathways. In: PaLM (2022)
31. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference On Computer Vision And Pattern Recognition, pp. 248–255 (2009)
32. Kojima, T., Gu, S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Adv. Neural. Inf. Process. Syst. **35**, 22199–22213 (2022)
33. He-Yueya, J., Poesia, G., Wang, R., Goodman, N.: Solving math word problems by combining language models with symbolic solvers (2023)
34. @miscfeng2015applying, title=Applying Deep Learning to Answer Selection: A Study and An Open Task, author=Minwei Feng and Bing Xiang and Michael R. Glass and Lidan Wang and Bowen Zhou, year=2015, eprint=1508.01585, archivePrefix=arXiv, primaryClass=cs.CL