Sci-Phi 3: An LLM companion for STEM students

Kaede Johnson | 357472 | kaede.johnson@epfl.ch
Ke Li | 357449 | ke.li2@epfl.ch
Davide Romano | 345025 | davide.romano@epfl.ch
Colab Prison Inmates

Abstract

We present Sci-Phi 3, an educational chatbot designed specifically for students studying STEM subjects. Sci-Phi 3 is a version of Microsoft's Phi-3 base model fine-tuned for multiple choice question-answering using a comprehensive four-stage pipeline: supervised fine-tuning for rationale generation, direct preference optimization for rationale generation, supervised fine-tuning for multiple-choice question answering, and retrieval augmentation. Sci-Phi 3 shows improvements in rationale generation and human preference alignment over base Phi 3. That said, it is worse than base Phi 3 in producing correct multiple-choice answers; we consequently explain this result and recommend ways to improve Sci-Phi 3 using its current architecture.

1 Introduction

Generative Large Language Models (LLMs) have revolutionized the way humans interact with artificial intelligence, in particular due to their ability to produce human-like responses with extensive knowledge across diverse subjects. This capability has significant implications for the educational sector, where chatbots can provide personalized and rapid responses to students' inquiries, thereby alleviating the workload of educators (Rahman and Watanobe, 2023). However, there are many challenges associated with educational AI, some of which include misinformation (Lo, 2023) or poor performance in highly technical subject areas (Frieder et al., 2024). While the benefit of an adept academic LLM is educational experiences on par with or even better than those humans alone can provide (Xiao et al., 2023), the cost of over-reliance on underdeveloped educational LLMs is the erosion of knowledge accumulation and retention in the most complex academic domains.

Existing educational LLMs often fall short in terms of having enough training data to cater a pretrained LLM toward specific academic needs (Kuhail et al., 2023). Furthermore, educational chatbots can produce frustration if they do not produce answers in the style students desire (Molnár and Szüts, 2018). In this project, we present Sci-Phi 3, a specialized educational chatbot designed to assist students in STEM disciplines. By leveraging curated, academic question-and-answer datasets, SFT, DPO, and RAG, Sci-Phi 3 addresses the gaps mentioned above, resulting in an open, efficient, and transparent tool for STEM education. Our results indicate Sci-Phi 3 improves the accuracy and reasoning of question-answering relative to its baseline pretrained model, Microsoft's Phi 3, thus showcasing its potential as a valuable educational resource.

Section 2 discusses research related to this topic. Section 3 contains our personal educational LLM architecture, including an illustration of our training pipeline. In Section 4, we explain the data we use in all training stages, the metrics we use to evaluate Sci-Phi 3's performance, the experiments we conduct to optimize Sci-Phi 3's architecture, and our quantitative performance outcomes. We then qualitatively address our model's performance in Section 5, remark on the ethical concerns of our work in Section 6, and, in Section 7, close with a summary of our findings as well as avenues for future refinement.

2 Related Work

Fine-tuning a pre-trained LLM is common practice when seeking a domain-specific tool. Kundu (2023) and Chen et al. (2023) find that when training without a large collection of data, PeFT (Mangrulkar et al., 2022) is preferred to global finetuning, as it orients the LLM toward a specific domain with lower computational overhead while achieving comparable results. Furthermore, using PeFT helps to avoid catastrophic forgetting (Luo et al., 2024). Imani et al. (2023) employ zero-shot CoT to improve accuracy and confidence in predictions for math reasoning tasks in particular, demonstrating that there is space for fine-tuning to improve performance in technical domains. Chern et al. (2023) do the same on popular benchmark datasets with exclusively SFT. One benefit of fine-tuning for a specific domain is hallucination reduction; furthermore, Li et al. (2024) find that when training data is limited, implementing RAG can boost factual accuracy and remove hallucinations as well. These findings support our usage of PeFT and RAG for Sci-Phi 3.

In-line with our usage of DPO is the finding by Sonkar et al. (2024) that, for an educational model specifically, human preference alignment is helpful in boosting the pedagogical efficacy of an LLM. Ji et al. (2024) second this notion of a boost to helpfulness in a general context and find human preferences can reduce harmfulness of generated output as well. Meanwhile, Song et al. (2024) find that a large set of answers to a small set of prompts - a description that could be applied to one of our own DPO datasets - is most helpful for human alignment.

3 Approach

We develop Sci-Phi 3 in four stages (see Figure 1):

1. SFT using mathematical questions, answers, and answer rationales.
2. DPO using cross-domain technical questions, answers, and answer rationales.
3. SFT on mathematical multiple choice questions, answers, and answer rationales.
4. Retrieval-augmentated generation using math, physics, and computer science source documents.

Our approach involves training and evaluating combinations of multiple adapters. Given the complex nature of our ultimate data, we pursue progressive learning in the manner of (Mukherjee et al., 2023), meaning we train on less complicated exam-style questions before or while training on complex exam-style questions. However, rather than take a sequential approach to training and potentially induce catastrophic interference, we use LoRA to freeze original model weights while training adapters at every stage. We also use Unsloth to speed up training time for all adapters (Han and Han, 2023).

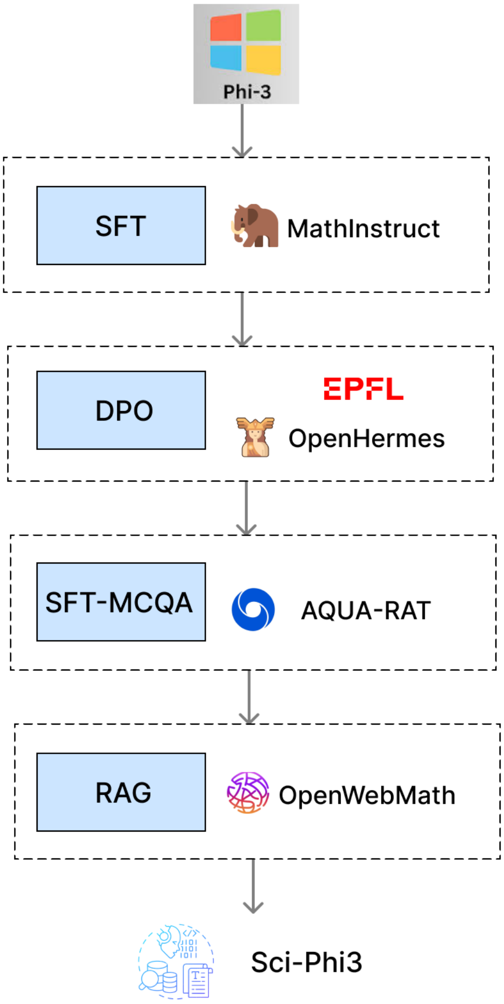

Figure 1: Four-stage pipeline.

3.1 Base Model

We choose Microsoft's Phi-3-Mini-4K-Instruct (henceforth Phi 3) for our base model (Abdin et al., 2024). Phi 3 is a dense decoder-only Transformer model trained with SFT and DPO on 3.3 trillion tokens. We select this model for a host of reasons. First, it was trained on public documents and synthetic data related to education, math, coding, and common sense reasoning - fields suitable for a engineering-related tutoring LLM. Indeed, Phi 3 performs very well on the math-related benchmark GSM-8K. We also believe the improvements to instruct-following, truthfulness, honesty, and helpfulness targeted with Phi 3's baseline preference training aids our performance. Furthermore, its context length of 4,096 tokens is vital given that one of our training datasets has a median length of 1,496 words. Finally, at 3.8B parameters, Phi 3 is relatively small among larger LLMs, allowing us to train multiple adapters without needing quantization to respect memory capacity.

3.2 SFT for Rationale

The first task in our pipeline is to fine-tune the Phi 3 base model's rationale when answering a question. Given that the end-goal of Sci-Phi 3 is to generate the answer to multiple choice questions, improving the quality of its rationale generation will improve the context Sci-Phi 3 generates and therefore has in its context window when selecting a multiple choice response. We also perform SFT at this stage because it has been shown to boost the efficacy of DPO (Saeidi et al., 2024). We use the SFTTrainer wrapper from the TRL library (von Werra et al., 2020) and math-related exam-style questions (see MAmmoTH in Section 4) to perform this stage of SFT.

3.3 DPO

We employ PeFT to train LoRA adapters for DPO on the base Phi 3 model merged with SFT adapter weights from part one of our pipeline. That is, given a winning generation y_w and losing generation y_l, we directly minimize the following training loss:

$$-E\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}-\beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right] \quad (1)$$

, where π_θ refers to the in-development Sci-Phi 3 we want to further refine and π_{ref} refers to the base version of Phi 3. We use the DPOTrainer wrapper from the TRL library as well as cross-domain preference pairs (see CS-552 and Hermes in Section 4) for this task.

While we investigated DPO variants such as Identity Preference Optimisation (IPO) and Contrastive Preference Optimization (CPO), we only implement because it has been shown to be more effective than IPO and CTO in Reasoning, Question Answering, and Mathematics when SFT has already occurred (Saeidi et al., 2024).

3.3.1 Preference Data Collection

Our contributions to the CS-552 preference pair data depend on two strategies: one for multiple-choice questions and one for open-ended questions. Both incorporate two-step prompting, generate output via the GPT3.5 wrapper, and culminate with the presentation phrase "Explain the answer in the most correct, relevant, clear, complete and concise way possible."

For multiple-choice questions, we generate one response with chain-of-thought prompting, passing "You are a {role}. {question} {options}. Let's go step by step" to GPT3.5. Here, 'role' is an occupation close to the question's technical domain, 'question' is the MCQ, and 'options' is the list of possible answers. Our second response is generated with the prompt "You are a {role}. {question}{options}. Identify each choice as strengthens, weakens, or doesn't impact the argument. Then select the correct answer." Qualitative analysis of the results show the latter prompt sometimes resulted in unusual output structure due to the command it select strengthening or weakening arguments. However, both tend to yield quality reasoning despite the structural difference.

For open-ended questions, we again generate one response via chain-of-thought prompting with the input "You are a {role}. {question}. Let's go step by step". The second response is generated with 'Take-a-step-back prompting'; this two-stage prompt begins with "You are a {role}. {question}. What are the {subject} principles behind this question?", where 'subject' refers to the technical domain. Upon receiving the first response, we then pass the query "Answer now the original question considering the {subject} principles." Qualitative analysis of the results show that both frameworks produce a quality response, though the latter tends to outperform the former.

3.4 SFT for MCQA

With our second round of SFT we ensure Sci-Phi 3 outputs its answers in a format conducive to MCQA extraction. Specifically, we use LoRA adapters on an MCQ dataset with algebraic questions (see AQUA-RAT in Section 4) reformatted such that each reference solution ends with 'Answer: [*letter*]'. We exploit this formatting during inference by selecting the final capital letter appearing in Sci-Phi's output as the intended answer.[1] Once again, we use TRL library to conduct SFT.

We explored using LMQL to explicitly enforce output structure during prompting (Beurer-Kellner

[1] In rare cases where this letter is not in the question's option set, we select an option at random for our answer. This simple approach is implemented to conform to this assignment's grading expectations; in a real-world implementation, our output would notify the student that the model has suggested an improper solution.

et al., 2023). Due to inconsistencies in default installation outcomes across teammate hardware, we abandoned this approach rather than risk unforeseeable error during assignment evaluation.

3.5 RAG

The last stage of our pipeline involves augmenting queries with math-, physics-, and computer-science-related documents (see OpenWebMath in Section 4). We split documents into 1000-character chunks with 50-character overlap and subsequently convert all chunks into embeddings using gte-v1.5 (Li et al., 2023) or the uncased BERT base model (Devlin et al., 2018). Using Facebook's Faiss library for similarity searches and clustering (Johnson et al., 2017), we create a vector database that houses the embeddings in 1,000 clusters. During inference, we embed queries, search the 5 most similar clusters for the document most similar to the query, retrieve the most relevant document and append it to the start of the prompt for Sci-Phi 3.

4 Experiments

4.1 Data

MathInstruct dataset (MAmmoth):

The MAmmoTH-related MathInstruct dataset (Yue et al., 2023), henceforth 'MAmmoTH data', is compiled from a mix of 13 math datasets, some newly generated. We used this data for SFT for rationale generation, retaining a 77,000-observation subset of the data formatted for chain of thought. These observations came specifically from the MATH_train, TheoremQA, college_math, gsm_train, and number_comparison datasets.

CS-552 data (class): This dataset, henceforth 'class data', consists of answer preference pairs generated via the GPT3.5 wrapper and manually labeled by students in the Spring 2024 CS-552 EPFL course. It consists of 26,738 total annotations across 1,522 unique questions. We retain approximately 3,400 pairs for DPO training. Questions primarily cover the domains of Computer Science Theory, Theoretical Physics, Artificial Intelligence and Machine Learning, Electrical Engineering, Computer Software, and Computer Systems. Each observation has a prompt, a preferred response, and a rejected response. Importantly, the prompt structure used to generate preference pairs is not standardized, leading us to seek other data to augment DPO.

OpenHermes dataset (Hermes): OpenHermesPreferences (Huang et al., 2024), henceforth 'Hermes data', is a preference pair dataset derived from LLMs. To target our subject areas we filtered this data to those pairs generated from Glaive Code Assistant, GPT-4, and Metamat, then sampled approximately 11,000 pairs for DPO training. Each observation contains a prompt, a preferred response, and a rejected response.

AQUA-RAT dataset: AQUA-RAT (Ling et al., 2017) is a large-scale dataset consisting of approximately 100,000 algebraic word problems. The solution to each question is explained step-by-step using natural language. Each observation contains a question, a list of options, a rationale, and a correct multiple choice response. To conduct SFT for MCQA formatting, we sampled 17,544 entries from the original dataset and formatted according to the process described in Section 3.4 above.

OpenWebMath dataset: OpenWebMath (Paster et al., 2023) is a dataset containing high-quality mathematical text from the internet. It is filtered and extracted from over 200B HTML files on Common Crawl down to a set of 6.3 million documents containing a total of 14.7B tokens. We sampled 1.3 million arXiv academic papers ('arXiv RAG') and 3.3 million web forum posts ('web RAG') to compile our RAG datasets, filtering the latter to include only chunks with size 200 characters or more. For both RAG databases we prioritized documents with LaTex code (0.70 and up for the LaTex parameter available when downloading OpenWebMath from HuggingFace), expecting formulas to assist Sci-Phi 3 with rationale.

Table 1: Dataset construction summary

Stage	Dataset composition	Total number
SFT	MathInstruct	77,000
DPO	75% Hermes + 25% Class	14,422
SFT-MCQA	AQUA-RAT	17,544
RAG	OpenWebMath	4,400,000

4.2 Evaluation Method

For DPO and the first round of SFT we evaluated along two dimensions: quality of generated rationale and ability to select correct option from a preference pair. To the target the former, we calculate BERTScore for our SFT + DPO adapter's generated text using the popular MATH benchmark (Hendrycks et al., 2021). To target the latter, we used our SFT + DPO adapter's win rate relative to baseline Phi 3 when performing inference on prompts in our validation set of class data preference pairs.

For SFT MCQA and RAG we had one evaluation metric: accuracy of multiple choice response on the popular ScienceQA benchmark (Lu et al., 2022), from which we sampled 500 questions and answers.

4.3 Baselines

Baseline values are produced using the base Phi 3 model. They are visible alongside Sci-Phi 3 values in the Results section's tables below.

4.4 Experimental Details

We train and evaluate combinations of multiple adapters when selecting final adapters for SFT for rationale and DPO. We evaluate versions with (1) only DPO on class data, (2) SFT on MAmmoTH data and DPO on class data, and (3) SFT on MAmmoTH data and DPO on class + Hermes data. Standard cross-entropy loss is used in all training.

We also evaluate two different RAG document databases: one composed solely of arXiv papers and another composed solely of web forum posts. Our RAG vector database used 1000 clusters and searches 5 clusters for nearest document matching.

We use a 4,096-token sequence length, one epoch, 5% or 10% dropout, a warm-up ratio of 0.1, a cosine scheduler, and a paged adamW optimizer for all adapter training. We opt for FP16 precision over FP32 precision to conserve GPU memory, avoiding BF16 due to varying GPU architectures. Batch sizes, gradient accumulation steps, learning rates, and LoRA rank are visible in Table 2. We use a smaller learning rate of 5×10^{-7} for class DPO. Finally, our DPO reward function has a beta value of 0.1.

Table 2: Training Configurations

	Batch	Learn. Rate	LoRA rank
SFT Rationale-MAmmoTH	8	1e-5	16
DPO-CS-552	4	5e-7	64
DPO-Hermes-CS-552	2	5e-7	32
SFT MCQA-AQUA-RAT	4	2e-6	16

4.5 Results

For the version of our model with SFT for rationale and DPO (the first two stages of our pipeline), we find BERTScore on rationale increases similarly relative to baseline across (1) a class data DPO adapter only, (2) a MAmmoTH SFT adapter and class data DPO adapter, and (3) a MAmmoTH SFT adapter and a class + Hermes data DPO adapter (see Table 3). Furthermore, all three such models outperform the base model in targeting human-preferred outputs from pairs, with pre-DPO SFT providing a boost in this regard. This validates the use of SFT and DPO. We proceed to SFT-MCQA and RAG (the latter two stages of our pipeline) using the MAmmoTH SFT and class + Hermes data DPO adapters. In other words, we use the version of the model labeled 'SFT for Rationale 2' in Table 3 for the rest of the pipeline.

Per Table 4, we unfortunately find that both of our RAG databases actually decrease performance for the model with SFT for Rationale, DPO, and SFT for MCQA. We took the RAG database that performed better at this stage and tested it on the baseline model, and once again see a decrease in performance.

Testing on small vector databases and simple queries with exact vocabulary matches demonstrated to us that our RAG implementation was calculating distances correctly. Given this verification, we belief the issue with our RAG implementation is the database of documents our model has access to. We discuss this further in the next section.

5 Analysis

One major reason for divergence between our stated goal and accuracy results could be the set of multiple choice questions we use for evaluation. The 500 ScienceQA questions we used tend to be unrelated to math. For example, `"What do these two changes have in common? A: a banana getting ripe on the counter B: deep-frying chicken"`. It may be the case that our emphasis on mathematical data during RAG and both stages of SFT pushes Sci=Phi 3 away from the ability to answer such non-formulaic questions.

We use the rest of this section to discuss examples from our different pipeline stages.

5.1 SFT for Rationale & DPO

We find evidence that Sci-Phi 3 after SFT for Rationale and DPO generates its rationale in a way much more conducive to student understanding. In response to a question about family relationships,

	Base	No SFT for Rationale		SFT for Rationale 1		SFT for Rationale 2	
		SFT	DPO	SFT	DPO	SFT	DPO
		-	CS-552	MAmmoTH	CS-552	MAmmoTH	CS-552 + Hermes
MATH: BERTScore-F1	0.841	0.852		**0.854**		0.853	
CS-552 Pairs: Success Rate		0.599		0.613		**0.616**	

Table 3: Evaluation Statistics for SFT for Rationale and DPO

	Base		SFT Rationale, DPO, SFT MCQA		
	-	web RAG	-	arXiv RAG	web RAG
ScienceQA: MCQA Acc.	85%	62.00%	64%	47.6%	48.2%

Table 4: Evaluation Statistics for SFT-MCQA and RAG

base Phi 3 produces the partial response `A, being B’s sibling, shares the same parents as B. Since B is the child of C and D, with ...`, etc. Intermediate Sci-Phi 3, meanwhile, gives its response more structure: `Please reason step by step and choose the right answer from the options below:(A) Aunt- [response]: A’ is the sister of B, and B is E’ daughter, so ...`, etc. These framing changes reveal the increasingly academic orientation of Sci-Phi 3 as it proceeds through our pipeline.

5.2 MCQA model

Unfortunately, SFT for MCQA does not ensure Sci-Phi 3 outputs its answer in a format we desire. Table 5 shows an output with intended formatting: the chosen letter appearing at the final position of output. This is by no means assured. We speculate that the sequence of tuning and adapter merging using MAmmoTH, Hermes, and AQUA-RAT data result in a model that needs more training data under SFT-MCQA to override its diversity of seen inputs. This diversity can sometimes yield inopportune output structure, as seen in Table 6, where Rationale and Answer have their order inverted. Answer selection is effectively (or literally) random in such a case.

Table 5: Good example of MCQA model (without RAG) answer

Prompt	**Answer**
Question: Select the reptile. Options: A. Japanese tree frog, B. fruit bat, C. eagle ray, D. salt water crocodile. The output format should be: ’Rationale: [insert your rationale here]. Answer: [insert the chosen letter(A/B/C/D/E) here]	Rationale: Salt water Crocodiles are reptiles. They are cold-blooded. All other options are mammals. Answer: D

Despite occasional formatting issues, the strategy of selecting the final capital letter in decoded output works for most of the time when extracting Sci-Phi 3’s response.

Table 6: Bad example of MCQA model (without RAG) answer

Prompt	**Answer**
"Question: Select the mammal below. Options: A. green tree frog, B. cassowary, C. Canadian lynx, D. barn owl. The output format should be: ’Rationale: [insert your rationale here]. Answer: [insert the chosen letter(A/B/C/D/E) here]’"	" Rationale : C Explanation: Canadian Lynx is a mamal. All other options are birds."

5.3 RAG models

The context our RAG implementation adds to our queries is often of poor quality. We see a few reasons for this: the added context is LaTex-heavy, it emphasizes theoretical mathematics, and our document chunks are of sizable length. These conditions favor a very particular type of input question: long, mathematical, and formula driven. Our evaluation dataset does not fit this mold. Our RAG databases may therefore overfitting to one particular domain; a minimum level of similarity before retrieval could be a way to avoid adding inappropriate context to queries.

By way of example, for the question `What do these two changes have in common? A: a sidewalk heating up in the sun, B: water vapor condensing on a bathroom mirror`, the context added by RAG begins: `"temperature). Is this just a straightforward application of newtons cooling law where y = 80? The outside of the cup has a temperature of 60Â°C and the cup is 6 mm in thickness. Who has the hotter coffee? Like most mathematical models it has its limitations. Denote the ambient room temperature as Ta and the initial temperature of the coffee to be To, ie. Newton's Law...`, and it continues for many more characters.

6 Ethical considerations

Creating an LLM that specializes in answering academic multiple choice questions could equip students with a powerful tool for completing assignments they are meant to complete without AI assistance. Students who would prefer to work without the tool would be disadvantaged, and teachers may respond by complicating their question style, creating a feedback loop which incentivizes AI usage even if it is forbidden by class syllabi. Our tool is also targeting a subset of subjects - based on our SFT, DPO, and RAG data, mathematics in particular - meaning its impact will be disparate across the student body.

Let us imagine a world where Sci-Phi 3 is made available to students. To ameliorate the concerns mentioned above, we could first position Sci-Phi 3 behind an API that ensures students do not gain access to generated rationale. If teaching staff ensures all homework assignments must be completed *with* rationale, Sci-Phi 3 would merely point students in the right direction without regurgitating the heart of their intended assignment. As for the emphasis on mathematics, Sci-Phi 3's RAG database could be augmented over time to incorporate more non-math material to compensate for our math-heavy SFT.

The lack of annotated, non-english DPO pairs and multiple choice questions means crucial portions of our pipeline emphasize English content only. This could make Sci-Phi 3 difficult for all students at a multilingual university like EPFL to use effectively. Because we emphasize Hermes data during DPO, this language imbalance might be at least partly resolved by generating more non-English preference pairs in the style of Hermes (that is, with LLMs), which is readily done for high-capacity languages like German or French. For less common languages, this becomes more difficult.

This is especially true for sign language. While Sci-Phi 3 is strictly textual, we recognize that early-life emphasis of spoken language can result in language deprivation that impacts literacy among deaf adults - a problem that is likely exacerbated by the presence of specific, rare vocabulary common in technical subjects. There are a couple ways we can refine Sci-Phi 3 to accommodate deaf users: first, if we assume students will have access to generated rationale, we can, within reason, tune Sci-Phi 3 to avoid words that lack well-known signed representation (that is, words which are almost always communicated with fingerspelling). For those oft finger-spelled words which rationale cannot afford to avoid, we can implement a database of pre-recorded sign language translations users can access, thereby bridging the STEM vocabulary gap. To assist students in communicating Sci-Phi 3's rationale to their peers, we can also tune output to avoid tokens with high articulatory effort. Finally, we could combine MediaPipe and a GCN to permit video-recorded sign language input that is then translated into a textual MCQ input - though the scarcity of domain-relevant training data means this evolution is rather ambitious.

7 Future Work

The most pressing next step is to modify the database of documents available for RAG in Sci-Phi 3 to be more general in content and natural in language. Clearly, the emphasis on LaTex-heavy text causes unnatural combinations of queries and retrieved documents; this is true for both a database of academic papers and a database of web forum posts. We recommend substituting our RAG database with a database built with Wikipedia articles covering relevant academic domains or textbooks.

We also believe that more training is necessary to achieve the goal of our second round of SFT. That is, to ensure Sci-Phi 3 consistently outputs the selected MCQA at the end of its response, we advise using a larger, more robust (topic-wise) set of training data. Future development might also consider applying the SFT-MCQA to base Phi 3 rather than the version with previous SFT and DPO. In this scenario Sci-Phi 3 could refer to two versions of the model, one with the SFT+DPO adapters for rationale generation, and one with SFT+RAG for answer generation. Alternatively, all Sci-Phi 3 queries could be wrapped in LMQL to force proper output.

8 Conclusion

Our project aimed to develop Sci-Phi 3, a specialized educational chatbot designed to assist students in STEM disciplines by answering multiple-choice questions with a strong rationale component. Throughout this process, we used Supervised fine-tuning (SFT) and Direct Preference Optimization (DPO) and a structured four-stage training pipeline to enhance the model's capabilities in rationale generation and alignment with human preferences.

The results showed improvements in the quality of rationale generation in some cases. Sci-Phi 3 was able to provide more detailed and structured rationales compared to its predecessor models, Additionally, DPO training demonstrated that our model successfully was mostly choosing the preferred answer when comparing reward scores for chosen and rejected answers.

However, despite these enhancements, Sci-Phi 3 underperformed in accurately answering multiple-choice questions relative to its base model. This unexpected outcome highlights a critical area for future enhancements.

Furthermore, the Retrieval Augmentation stage worsen the model's performance, suggesting potential issues with the integration of context from external documents or the relevance of the retrieved content. This issue highlights the importance of ensuring that augmentation strategies are closely aligned with the model's training objectives and the nature of the tasks it is expected to perform.

In conclusion, while Sci-Phi 3 demonstrated improvements in generating well-aligned and structured rationales. Future work could focus on refining the retrieval mechanisms, expanding the training datasets to cover a broader range of subjects and formats, and further optimizing the fine-tuning processes to enhance multiple-choice question answering accuracy.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969.

Yong Chen, Hongpeng Chen, and Songzhi Su. 2023. Fine-tuning large language models in education. In *2023 13th International Conference on Information Technology in Medicine and Education (ITME)*, pages 718–723.

Ethan Chern, Haoyang Zou, Xuefeng Li, Jiewen Hu, Kehua Feng, Junlong Li, and Pengfei Liu. 2023. Generative ai for math: Abel. `https://github.com/GAIR-NLP/abel`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2024. Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, 36.

Daniel Han and Michael Han. 2023. Unslothai.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset.

Shengyi Costa Huang, Agustín Piqueres, Kashif Rasul, Philipp Schmid, Daniel Vila, and Lewis Tunstall. 2024. Open hermes preferences. https://huggingface.co/datasets/argilla/OpenHermesPreferences.

Shima Imani, Liang Du, and H. Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus.

Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1):973–1018.

Deepanjan Kundu. 2023. Unleashing the potential of domain-specific llms.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *ArXiv*, abs/2403.10446.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

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

Chung Kwan Lo. 2023. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2024. An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

György Molnár and Zoltán Szüts. 2018. The role of chatbots in formal education. In *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000197–000202. IEEE.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text.

Md Mostafizer Rahman and Yutaka Watanobe. 2023. Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9):5783.

Amir Saeidi, Shivanshu Verma, and Chitta Baral. 2024. Insights into alignment: Evaluating dpo and its variants across multiple tasks.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. 2024. Scaling data diversity for fine-tuning language models in human alignment. *arXiv preprint arXiv:2403.11124*.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G Baraniuk. 2024. Pedagogical alignment of large language models. *arXiv preprint arXiv:2402.05000*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

A Appendix

A.1 AI Usage

The most important code pieces in this project - SFT, DPO, and RAG - were implemented with the aid of human-written tutorials rather than AI tools. We did not document all usage of ChatGPT; when it was used, it was used for menial tasks we believed we were capable of but preferred not to implement by hand (re-formatting multiple choice questions in the ScienceQA benchmark dataset to match project's input question format, for example).

A.2 Team Contributions

Davide was primarily responsible for SFT, Ke was primarily responsible for DPO, and Kaede was primarily responsible for RAG. That said, team members shared updates on their progress and verified each other's implementations, helping with minor tasks outside of their primary responsibility when necessary. Model evaluation and report writing were a shared responsibility.

A.3 Training records

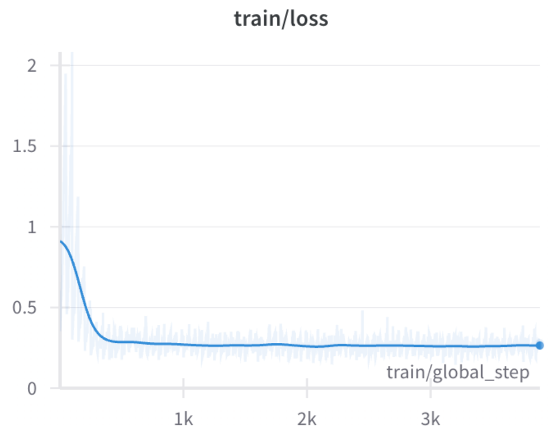

Figure 2: Training Loss during SFT on MAmmoTH dataset

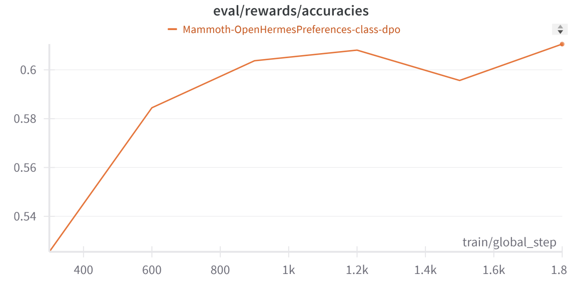

Figure 3: Evaluation accuracies during DPO on OpenHermes and class dataset

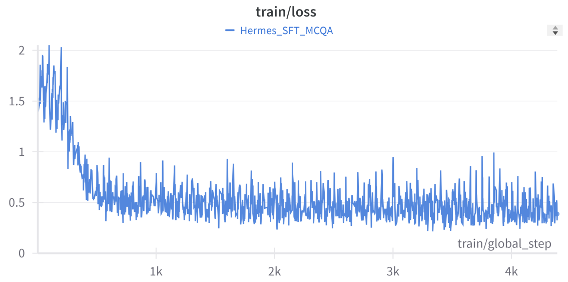

Figure 4: Training Loss during MCQA SFT on AQUA-RAT dataset