Size Does Not Matter: Fine-Tuning and Quantising Phi-3-Mini for MCQA

Pierre Lardet | 376393 | pierre.lardet@epfl.ch
Mika Senghaas | 377332 | mika.senghaas@epfl.ch
Ludek Cizinsky | 377297 | ludek.cizinsky@epfl.ch
MLP

Abstract

The advent of Large Language Models (LLMs) has impacted education forever, with models becoming integral tools in aiding students' learning. However, even advanced models can struggle to provide accurate answers to complex questions typical in STEM university curricula. Motivated by these challenges, we align an open-source LLM with a curated preference dataset collected by EPFL students, and enhance the model's performance on multiple-choice question answering by fine-tuning. Additionally, we quantise our best performing model to 4-bit precision. Our findings indicate that we successfully aligned our model, achieving 67% accuracy in predicting the preferred answers. However, despite extensive fine-tuning, we did not observe improvements compared to our base model. Finally, we successfully quantised the model to 4-bit precision with minimal loss in performance.

1 Introduction

The potential of Large Language Models (LLMs) in the educational landscape is vast. Online chatbots like ChatGPT and Gemini are common study companions in students' daily lives. These tools provide quick and personalised answers to questions, and allow for interactive feedback comparable to having a tutor. However, these models are generalists, and may not be optimised for specific educational tasks. They are also very large, often with hundreds of billions of parameters, creating problems for porting them on to local devices.

The above motivates our project, in which we aim to adapt and quantise a language model for university-level multiple-choice question answering (MCQA) in the natural sciences. As our starting point, we choose Microsoft's flagship Small Language Model (SLM) Phi-3-Mini (Abdin et al., 2024). We fine-tune this base model using two different strategies: Supervised Fine-Tuning (SFT) and Direct Preference Optimisation (DPO). We then quantise the model using Generalised Post-Training Quantisation (GPTQ) (Frantar et al., 2023) to reduce memory and computational requirements with minimal performance loss.

We find that fine-tuning Phi-3 on various highly curated MCQA datasets and a custom DPO preference dataset does not improve the model's MCQA performance. We attribute this to Phi-3's extensive post-training, which already optimised the model for MCQA tasks. Using GPTQ, we quantise the model from 16-bit floating point to various bit precision. The 4-bit quantised model exhibits a marginal performance loss, while being 75% smaller in size.

2 Related Work

LLMs in Education. Since its debut in November 2022, ChatGPT by OpenAI has revolutionised various sectors, including education. Applications in this field include automatic grading, material creation and aiding students with problem-solving and clarifications (Wang et al., 2024). Current research on LLMs in education primarily falls into two categories: benchmarking LLMs against educational datasets (e.g. maths (Wu et al., 2023), medicine (Liévin et al., 2023), and programming (Savelka et al., 2023)), and developing methodologies to enhance performance on educational tasks, such as the Chain of Thought (CoT) prompting strategy (Wei et al., 2023). Our research belongs to the latter category, focusing on advancing the base model's performance by fine-tuning on a range of science MCQA datasets.

SLMs. The efficacy of pretrained Large Language Models (LLMs) typically correlates with their size and the volume of training data, as described by a power law (Hestness et al., 2017; Kaplan et al., 2020). However, to democratise access to cutting-edge models, there has been a shift towards developing smaller models. For example, Microsoft's

Phi (Abdin et al., 2024) and Orca (Mukherjee et al., 2023) series of models were built with a focus on the quality of the training data rather than the size of model. They outperform larger models on various benchmarks such as GPT-3.5 and Mixtral 8x7B (Jiang et al., 2024), and compete with other SLMs like Apple's OpenELM (Mehta et al., 2024). Our research builds on this trend by fine-tuning the Phi-3-Mini model for MCQA tasks.

Fine-Tuning. Fine-tuning is a common technique to adapt pretrained LLMs to specific tasks. Supervised Fine-Tuning (SFT) is the most common method, where the model is trained on a task-specific dataset with a supervised learning objective (Devlin et al., 2019). However, SFT can be computationally expensive and may require large amounts of task-specific data. Therefore, parameter-efficient methods like Low-Rank Adaptation (LoRA) (Hu et al., 2021) have been developed to reduce the computational cost of fine-tuning. In our work, we fine-tune using LoRA adapters.

Alignment. Model alignment guides model behavior towards desirable outcomes (Kaufmann et al., 2024). It is a criticial step in training an LLM. Traditionally, Proximal Policy Optimisation (PPO) was the dominant technique for implementing alignment (Schulman et al., 2017). Direct Preference Optimisation (DPO) proposes a different parameterisation of the reward model leading to a simplified alignment procedure (Rafailov et al., 2023). Given its simplicity and efficiency, we adopt DPO to align our model.

LLMs & MCQA. MCQA benchmarks (Clark et al., 2018; Johannes Welbl, 2017; Hendrycks et al., 2021) are commonly used to measure the capabilities of LLMs. Despite their prevalence, studies have identified several flaws in MCQA benchmarkss as evaluation tools, such as susceptibility to choice dynamics (Balepur et al., 2024), positional biases (Li et al., 2024; Khatun and Brown, 2024; Zheng et al., 2024), misunderstandings of the MCQA format (Khatun and Brown, 2024), and sensitivity to prompt phrasing (Khatun and Brown, 2023). Variability in results across different implementations and answer extraction methods further complicates assessments (Fourrier et al., 2023). To standardise comparisons of our fine-tuned models, we employ a widely recognised evaluation framework called Language Model Evaluation Harness (LMEH) (Gao et al., 2023), ensuring consistent and fair testing.

3 Approach

Baseline model. We use the pre-trained *Phi-3-Mini-4k-Instruct* model (Abdin et al., 2024), a 3.8B parameter Transformer decoder-only model with 32 heads, 32 layers, and a hidden dimension of 3072 as our base model. It was pre-trained on 3.3T tokens of heavily filtered web and synthetic data. During post-training, the model has undergone both SFT to induce high-quality domain-specific knowledge in domains, such as math, coding, and reasoning, and DPO for alignment. Despite its small size, the model has shown strong performance across many common NLP benchmarks, challenging much larger models. Its trade-off between performance and size makes it an ideal starting point for our project.

To further specialise Phi-3 Mini for scientific question answering, we consider two different fine-tuning strategies:

DPO Alignment. We align the base model with preference data as detailed in Section 4.1 using DPO. The DPO loss defines the probability of a completion y given a context x as:

$$p(y|x) = \log\left(\frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\right) \tag{1}$$

where π_θ is the the model we are training and π_{ref} is the original model. The DPO loss function is then given by:

$$-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\beta(p(y_w \mid x) - p(y_l \mid x))\right] \tag{2}$$

This loss function trains the model to prefer the outcome y_w over y_l. β is a hyperparameter that regulates how much the policy model can deviate from the reference model. Additionally, we explore two variants of the DPO loss: RSO (Liu et al., 2024), which incorporates a Hinge loss, and IPO (Azar et al., 2023), which adjusts the DPO loss to prevent overfitting.

Supervised Fine-Tuning. We employ SFT to specifically tailor our model to answering multiple-choice questions. SFT has two primary objectives: to enrich the model with domain-specific knowledge and to familiarise the model with the expected

MCQA answer format. This approach utilises the standard language modelling objective of next token prediction which is defined as:

$$-\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\log p(y \mid x)\right] \tag{3}$$

LoRA. We use LoRA (Hu et al., 2021) during all fine-tuning stages. In contrast to full-parameter fine-tuning, LoRA injects low-rank adapation matrices, to adapt the forward pass of the model.

$$h = W_0 x + \nabla W x = W_0 x + BAx, \tag{4}$$

where $W_0 \in R^{d\times k}$ is the pre-trained weight matrix, and $B \in R^{d\times r}$ and $A \in R^{r\times k}$ are the adaptation matrices with rank $r \ll \min(d, k)$. Because of the low-rank dimension, the number of trainable parameter is significantly reduced but fine-tuning performance is maintained.

MCQA Extraction. After fine-tuning, our model does not necessarily output a single letter response. To extract a single letter answer from the model we apply post-processing. A variety of approaches have been explored (Tsvilodub et al., 2024). We opt for a simple approach- loglikelihhood-based comparative scoring, which is used in LMEH (Gao et al., 2023). Given a question and answer, the sum of log probabilities of each of the answer options is computed and the highest scoring continuation is predicted. Formally, given a sequence of tokens $x_{0:n_i}$, where $x_{0:m}$ is the question with answer options and $x_{m:n_i}$ is the answer option i, the log-likelihood of the answer option i is

$$LL_i = \sum_{j=m}^{n_i-1} \log P(x_j|x_{0:j}) \tag{5}$$

Quantisation. Finally, we use GPTQ (Frantar et al., 2023) to quantise the fine-tuned model from 16-bit to 8-, 4-, 3- and 2-bit precision. GPTQ is a post-training method that applies layer-wise quantisation. Given a layer W and input X, the objective is to find a quantised layer $\hat{W}$ that minimises the mean squared error between the full-precision and quantised outputs.

4 Experiments

4.1 Data

We use a mixture of DPO preference data and MCQA-style SFT datasets.

EPFL Preference Data. This dataset is a collection of 26,738 student-generated answer pairs where one answer is preferred over the other. The questions comprise 1,522 unique questions from 24 different courses at EPFL. Each pair of answers was generated by prompting GPT-3.5 and annotated using various ranking criteria and an overall ranking by students from the CS-552 course. Before any further processing, we perform an 80-20 train-validation split.

ARC. The ARC dataset (Clark et al., 2018) consists of 7,787 grade-school level, multiple-choice science questions released by the Allen Institute for Artificial Intelligence (AI2).

SciQ. The SciQ dataset (Johannes Welbl, 2017) contains 13,679 crowdsourced science exam questions from natural sciences, like physics, chemistry and biology. Each question is in multiple-choice format and a paragraph of supporting evidence for the correct answer is available.

OpenBookQA. The OpenBookQA dataset (Mihaylov et al., 2018) is a collection of 5,957 multiple-choice science questions. The dataset is designed to test the model's ability to answer questions that require reasoning and understanding of the natural world.

MMLU. The Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) is a large, diverse collection of multiple-choice questions from various domains. In this work, we use the MMLU-STEM subset, consisting of 19 of the subjects most relevant to STEM and 3,317 questions.

GPQA. The General Purpose Question Answering (GPQA) benchmark (Rein et al., 2023) is a dataset of 1,192 very challenging multiple-choice questions, designed and validated by experts in natural sciences. With human expert baseline scores being only 34%, it is among the most challenging benchmarks in the industry.

For each dataset, we provide an example data sample in the Appendix Section A.5. Before using the data for fine-tuning or evaluation, we preprocess it

into a standardised format. To do this, we parse the question, answer options, and correct answer and format them as shown in Appendix Section A.6. We use the chat template used by our base model, and feed the question text, followed by a lettered list of answer options as the user message, and the correct answer as the assistant message.

Depending on the availability of training and validation/test splits, we use the above datasets for fine-tuning and evaluation, or only evaluation. This is detailed in Sections 4.2 and 4.4. All datasets are publicly available on HuggingFace, with the exception of the EPFL preference data which was provided to us by the EPFL course staff.

4.2 Evaluation

For DPO alignment, we evaluate using the model's accuracy in assigning a higher probability to the preferred answer.

To evaluate models, we use the LMEH framework, which is commonly used in the literature to evaluate language models and also powers the HuggingFace's OpenLLM Leaderboard. The framework uses the same methodology of comparative log-likelihood scoring to extract the model's answer to a multiple-choice question. Then, given a list of model predictions $\hat{y}_1, \ldots, \hat{y}_n$ and the ground truth answers $y_1, \ldots, y_n$ for a task, it computes the accuracy score as, where $\mathbb{I}(\cdot)$ is the indicator function. Additionally, it computes the standard error (SE) in the accuracy score which gives an estimate of the uncertainty in the accuracy score given the number of questions in the benchmark task. We use only zero-shot evaluation for all benchmarks, as this is how the model would be used if it were chatbot.

4.3 Baselines

We first verify that the Phi-3-Mini model is the best performing model on the MCQA benchmarks defined in Section 4.1. We compare against the following two models:

OpenELM. The OpenELM model (Mehta et al., 2024) is a family of small, efficient language models released by Apple in April 2024. We use the largest available, instruction-tuned model, OpenELM-3B-Instruct, for our experiments.

Llama 3. Finally, we consider the models from the popular LLama 3 family (Touvron et al., 2023). In particular, we use LLama-8B-Instruct model, the smallest model in the most recent release of model family by Meta in April 2024.

For all experiments including DPO alignment, MCQA fine-tuning, and quantisation, we use the Phi-3-Mini model as the baseline.

4.4 Experimental Details

DPO alignment. We follow two steps for DPO alignment. We first run several experiments on 20% of EPFL data to tune the parameters outlined in Table **??**. We then identify the best performing configurations based on the validation accuracy and run the train on the full EPFL dataset.

MCQA finetuning. We follow the guidelines for fine-tuning Phi-3 from Microsoft's cookbook. Table **??** details the hyperparameters that differ from Hugging Face's defaults. Using these parameters, we train models on the ARC, SciQ and OpenBookQA datasets, and all in combination. We call these models Phi-3-ARC, Phi-3-SciQ, Phi-3-OBQA, and Phi-3-MCQ respectively.

Quantisation. For the quantisation via GPTQ, we use the default hyperparameters provided by the Hugging Face's API, and only experiment with the number of bits used for quantisation (8, 4, 3, 2).

4.5 Results

4.5.1 Baseline Results

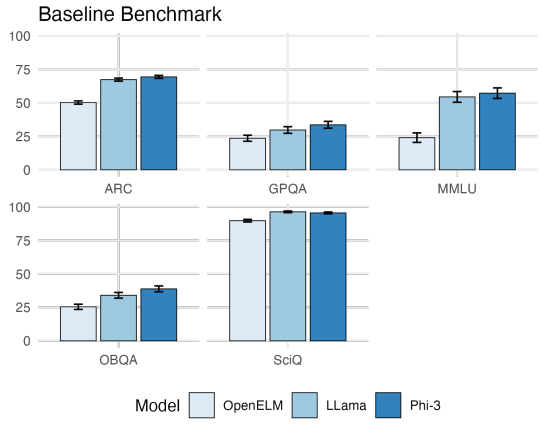

Figure 1: **Baseline Benchmark Results.** Mean accuracy on all task groups for the three baseline models. Error bars represent the standard error of the accuracy score.

Figure 1 visualises the mean accuracy on all task groups for the three baseline models (exact numbers are provided in Appendix Table 4). We observe that OpenELM-3B-Instruct is worse than the

other two models on all benchmarks. While Phi-3 and LLama-3 perform similarly on most benchmarks, Phi-3 slightly outperforms LLama-3 on most benchmarks. This is surprising as Phi-3 is half the size of LLama-3, and similarly sized as OpenELM-3B. This suggests that Phi-3 is the most performant model for its size among the base models we consider and provides a strong baseline for our experiments. However, this also means that it will be more challenging to improve the model further by fine-tuning.

Moreover, we observe that the benchmarks vary in difficulty. GPQA and OBQA are the most challenging benchmarks with Phi-3 achieving 33.6% and 38.8% accuracy respectively. MMLU and ARC are less challenging, with Phi-3 scoring accuracy of 57.3% and 69.4%. SciQ is the easiest benchmark, with Phi-3-Mini achieving 95.5% accuracy.

4.5.2 DPO Alignment Results

Table 1: Training Configuration and Results for DPO finetuning. On the left, we highlight in bold the hyperparameters being tuned (LabSm = Label Smoothing).

Configuration					Results
LR	Rank	Loss	Beta	LabSm	Acc (%) ↓
4e-5	32	**IPO**	**0.1**	0.1	67.01%
2e-5	**16**	**IPO**	**0.1**	**0.0**	66.61%
4e-5	32	DPO	**0.05**	0.1	65.97%
2e-5	**16**	DPO	0.4	0.1	64.76%
4e-5	32	DPO	0.4	**0.0**	63.89%
4e-5	32	DPO	0.4	0.1	63.27%

Table 1 presents the validation accuracy of the DPO alignment experiments. The configuration employing the IPO loss demonstrated the highest performance, aligning with expectations given IPO loss's design to enhance model generalisation. Furthermore, through further hyperparameter tuning, we achieved a nearly 4% increase in accuracy compared to the baseline model. However, combining the best performing configurations from the experiments resulted in only the second best performing model suggesting a strong interdependence among the hyperparameters.

4.5.3 Fine-Tuning Results

Figure 2 shows the mean accuracy on all benchmarks for all fine-tuned variants, against the Phi-3 baseline, with full results in the Appendix 5.

Despite the model's increased performance in accuracy in retrieving the preferred answer, performance on scientific MCQA benchmarks does not improve. We hypothesise that the DPO examples are not similar enough to MCQA examples in terms of style and content for the training to be transferrable. Aligning the model towards long answers with explanation does not necessarily benefit the retrieval of the correct answer from the log-probabilities assigned to the short answer option continutations used in our evaluation setup. For this reason, we choose to focus on fine-tuning the base model on highly curated MCQA datasets that are specifically formatted for the task.

Fine-tuning on the MCQA datasets does not lead to significant improvements across the benchmark tasks either. For variants fine-tuned on OBQA, ARC and a combination of all MCQA datasets, we observe negligible changes in the mean accuracy that often fall within the standard error of the Phi-3-Mini baseline. Notably, fine-tuning on SciQ leads to a significant drop in performance on the MMLU dataset, as well as the GPQA dataset, which suggests a big discrepancy in these datasets.

As our final model we choose Phi-3-ARC. It performs similarly to the strong Phi-3-Mini base model, with a slight increase in performance on the MMLU and OBQA benchmarks, from 57.3% to 57.8% and 38.8% to 40.4% respectively.

4.5.4 Quantisation

We quantise the Phi-3-ARC model and show the results of 8-, 4-, 3- and 2-bit variants in Figure 3.

We find that quantising using GPTQ is an effective approach to reducing the memory footprint of the model while maintaining the impressive MCQA performance. The 8-bit, and 4-bit quantised models perform similarly to the unquantised model across all benchmarks. However, when further quantising to 3-bit precision, we observe a decrease in the performance, with the mean accuracy dropping to 57.3% on the MMLU benchmark. Finally, 2-bit quantisation is not feasible for our model, as we find that it performs no better than a random baseline.

We conclude that the 4-bit quantised model provides the best trade-off between performance and efficiency, and choose this model as our final quantised model for submission.

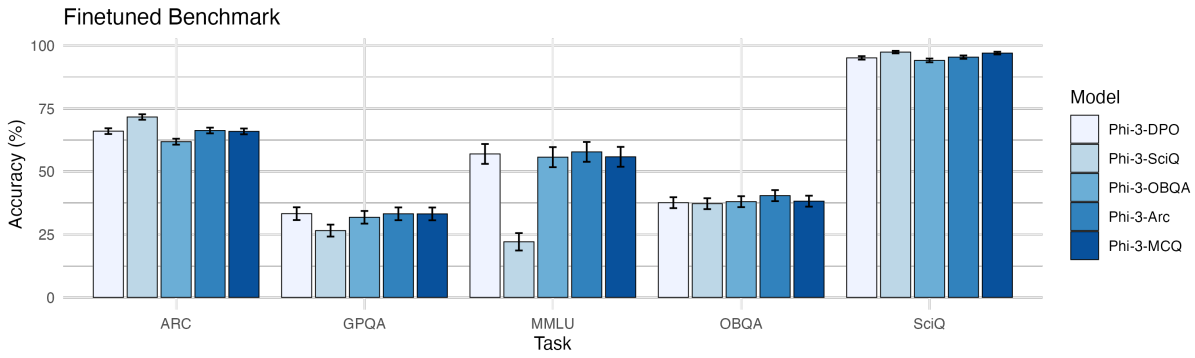

Figure 2: **Finetuned Benchmark Results.** Mean accuracy on all task groups for the all fine-tuned models and Phi-3. Error bars represent the standard error of the accuracy score.

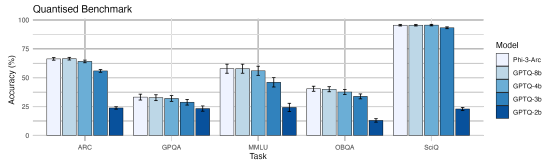

Figure 3: **Quantised Benchmark Results.** The Figure shows the accuracy of the quantised models and its unquantised counterpart on the benchmarks. The error bars represent the standard error of the accuracy score.

5 Analysis

5.1 Lack of Performance Improvement

We hypothesised that fine-tuning the base model Phi-3 on various highly curated MCQA datasets and a custom DPO preference dataset would improve the model's MCQA performance. However, we observe that the fine-tuned models do not outperform the base model. This finding is surprising, as we expected the fine-tuned models to narrow their focus of knowledge on the scientific MCQA task.

We conducted further analysis. The base model Phi-3 attains a mean accuracy of 93.6% on the ARC training split, strongly suggesting that the base model was already trained on the ARC dataset. By comparison, it attains 69.4% on the test split. The fine-tuned Phi-3-ARC variant attained an accuracy of 97.6% on the train split. This improvement over Phi-3 base leads us to believe that fine-tuning did 'narrow' the model's focus. However, this is merely memorisation and does not lead to generalisation, even over the same dataset as Phi-3 curiously does better than Phi-3-ARC on the test split, with Phi-3-ARC attaining an accuracy of 66.3%.

The fact that the model has likely already been fine-tuned on the datasets we use makes it hard to improve the model performance using the same data. A 're-finetune' does not seem to lead the model to focus on the knowledge pertinent to the fine-tuning datasets which we had hoped for prior to the experiments. It does help a little with pure memorisation, but not with generalisation, even within the same dataset. Microsoft have clearly already struck a good balance over the training datasets to optimise for general reasoning.

5.2 Qualitative Samples

Below is an example of the model's predictions on the ARC dataset using both the Phi-3 and Phi-3-ARC for a sample question.

```
Question: Consider the Bayesian network
given below. How many independent
parameters are needed for this Bayesian
Network H -> U <- P <- W?
Options:
A. 2
B. 4
C. 8
D. 16
Answer:

--- PHI-3 ---
To determine the number of independent
parameters needed for the Bayesian
network H -> U <- P <- W, we need
to consider the conditional probability
tables (CPTs) for each node given its
parents. In this network ...

--- PHI-3-ARC ---
The correct answer is C. 8
```

Listing 1: Sample Question

We clearly see that the fine-tuned model predicts an answer directly, while the base model provides a detailed explanation. This is expected given the formatting of MCQA text we feed into the model

during fine-tuning. Below is another sample for only Phi-3-ARC.

```
You: Is Messi the best player ever?
Phi-3-ARC: The correct answer is A. True
```

Listing 2: Phi-3-ARC sample

Interestingly, the model predicts a multiple choice answer to a yes/no question. It has been *over*-fine-tuned and hallucinates a mutliple choice style answer to a question that does not have one. Clearly our fine-tuned models do not generalise well and are overfitted to the training data.

5.3 Per Subject Analysis

To investigate if the model has strengths and weaknesses in different subject areas within STEM, we look at the 19 subjects within the MMLU-Stem benchmark. These include physics, chemistry, biology, and mathematics at different educational levels, such as elementary school, high school, and college. Figure 4 visualises a sample of the mean accuracies per subject.

Overall, we observe that the per-subject performance differs greatly, with models achieving a mean accuracy on High School and College Biology of 83.98% and 80.79%, respectively. In contrast, the models perform poorly on high school and college, mathematics and physics, not surpassing 50% accuracy. This finding could be attributed to two factors: first, the inherent complexity of the subjects and second, the lack of training data for these subjects of the models. Either way, the performance could be improved significantly by collecting more training data for challenging subjects.

The per-subject comparison of the base, fine-tuned and quantised model shows similar trends as the overall benchmark suggests. No clear trends of one model being predictably better in some subjects than others are observed, which is expected since both fine-tuned models were trained on the smae

5.4 Skewed Answer Distribution

Next, we investigate the distribution of answers of the three models under consideration. Figure 5 shows the heatmap of the confusion matrix of correct and predicted answers for the base model Phi-3, and the difference in confusion matrices between the fine-tuned model Phi-3-ARC and the base model Phi-3, and the quantised model Phi-3-ARC-4bit and the base model Phi-3.

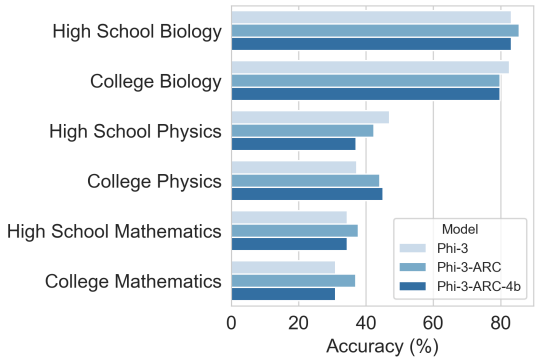

Figure 4: **MMLU Per Subject.** The mean accuracy of the base model Phi-3, the best performing fine-tuned model Phi-3-ARC, and its 4-bit quantised version Phi-3-ARC-4bit, per subject.

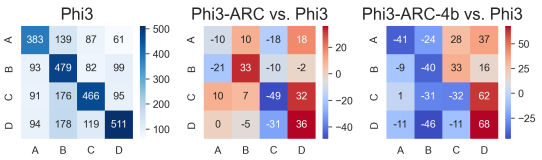

Figure 5: **Confusion Matrics.** The confusion matrix of answers given by the baseline model and the difference in accuracy between the fine-tuned model and the baseline model.

Interestingly, we observe that the base model is biased towards answering B, as indicated by the high number of false positives in the second column. Roughly 30% of the model's answers are B, despite the true answer distribution being close to uniform. Moreover, the fine-tuned model Phi-3-ARC, further increases this bias, perhaps due to the ARC dataset being slightly biased towards B itself. Curisouly, quantising the model to 4 bits flattens the distribution of answers slightly, decreasing the bias towards B.

6 Ethical Considerations

The potential for misuse of LLMs, especially in the context of education and academia is a significant concern, and it is our responsibility to ensure that our work does not cause harm. Various factors must be borne in mind.

In the current implementation, our model is only capable of handling English text with high accuracy. The performance of the model on other languages, especially low-resource languages, is likely to be suboptimal. To adapt the model to handle other languages, we would need to collect a large amount of data in the target language, which may not be

feasible for low-resource languages. This could exacerbate the divide in access to advanced tools between speakers of major languages and lesser spoken languages.

Similarly, an exciting future direction would be to adapt the model to interact with users in signed languages. Learning to read is a lot harder for deaf people because the learning process at least partly involves phonetics. This often leads to lower levels of literacy and education in the deaf community. STEM resources in signed languages are scarce and adapting the model to signed languages could help bridge this gap. However, this is a particularly challenging task. It would require the model to operate either with a sign language interface that converts signed language video to text and vice versa or in different modalities, including text and video, and would necessitate a significant amount of data in signed languages, which is currently scarce. Additionally, the lack of a standard sign language for technical words could pose a challenge in adapting the model to signed languages.

Even if our trained model is working as intended, there are complex, potential harms we must consider. Instead of aiding in learning, students might misuse the model to short-cut their learning process. For example, students might use the model to finish homework or assignments without understanding the underlying concepts. This could hinder the learning process and in the long-term, undermine the integrity of the educational system and lead to over-reliance on technology. This could be minimised by restricting or overseeing the use of such models in educational settings.

The chatbot might also lead to work not being cited. If the chatbot provides answers to questions, students might not cite the source of the information correctly, especially if the model has memorised the original literature. This could lead to inadvertent plagiarism and intellectual property theft. To mitigate this, the chatbot could be aligned to include prompts to cite specific pieces of work or provide the citations themselves.

We must also consider potential biases in our training data. There might be biases towards specific groups or political views in the data. In examples where the model is used to generate answers, these biases could be propagated. This could lead to the reinforcement of stereotypes or discrimination against certain groups within an educational environment. To mitigate this, it could be beneficial to ensure that training data is diverse and representative, and the model could be aligned to avoid generating biased responses using methods similar to those we have used in our project.

7 Conclusion

In this project, we have adapted a Phi-3-Mini, a SLM, using DPO alignment and fine-tuning for MCQA, to improve its performance on scientific multiple-choice question answering tasks.

We found that it was challenging to improve the performance of the base model significantly. We suspect that this is because instruction-tuned models such as our base model are already highly optimised for this task during their pre- and post-training. Small experiments show evidence that many, if not all, of our MCQA datasets are already included in the base model's training data. Hence the lack of improvement in performance is not necessarily surprising. Clearly, 'focusing' the model on scientific MCQA by re-finetuning on datasets curated is not effective. The model does not generalise well and is overfitted to the re-finetuned data.

Finally, we show that state-of-the-art quantisation techniques, such as GPTQ, are highly effective in reducing the computational and memory requirements of the model without sacrificing performance. We found that the model can be quantised to 4-bit without significant performance loss.

There are many exciting future avenues such as adapting the model to handle other languages, especially low-resource languages, and signed languages to improve the inclusivity of the model. Additionally, fine-tuning on datasets not included in the base model's training such as a dataset similar to the EPFL preference data but for SFT could be explored to improve the model's generalisation capabilities. Finally, experimenting with larger models to see if they can be 'narrowed' more effectively could also be an interesting direction.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, and Jyoti Aneja. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question?

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Julien Launay. 2023. What's going on with the open llm leaderboard?

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers.

Leo Gao, Jonathan Tow, Baber Abbasi, and Biderman. 2023. A Framework for Few-Shot Language Model Evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing Multiple Choice Science Questions.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback.

Aisha Khatun and Daniel G. Brown. 2023. Reliability check: An analysis of gpt-3's response to sensitive topics and prompt wording.

Aisha Khatun and Daniel G. Brown. 2024. A study on large language models' limitations in multiple-choice question answering.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of llms?

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. 2024. Statistical Rejection Sampling Improves Preference Optimization.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2023. Can large language models reason about medical questions?

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. OpenELM: An Efficient Language Model Family with Open Training and Inference Framework.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark.

Jaromir Savelka, Arav Agarwal, Marshall An, Chris Bogart, and Majd Sakr. 2023. Thrilled by your progress! large language models (gpt-4) no longer struggle to pass assessments in higher education programming courses. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, ICER 2023. ACM.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Bap-

tiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Polina Tsvilodub, Hening Wang, Sharon Grosch, and Michael Franke. 2024. Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2023. An empirical study on challenging math problem solving with gpt-4.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors.

A Appendix

A.1 AI Usage

Throughout the project, we used Github Copilot and ChatGPT to assist with both coding and writing. Copilot is automatically prompted within the IDE and ChatGPT was prompted using the OpenAI website. In general, when coding, Copilot was used to generate code snippets and functions to save time. When writing, copilot suggested sentences and paragraphs to help with the writing process.

However, all output was verified manually by ourselves. When coding, all outputs were re-read, and often the suggestions are not helpful unless the code is laborious and simple. For writing, we reworked almost every suggestion to fit the context of the report and for greater consistency.

We found that for simple Python notebooks or specific library functions, Copilot was very helpful. However, for more complex code or writing, it was often incorrect and required more manual work. When writing, ChatGPT was helpful for suggesting sentences and paragraphs, but many of these suggestions are incorrect or not relevant. It was most useful for generating a list of ideas for a given section from which we could then write the content ourselves.

A.2 Datasets

Table 2: **Data.** The datasets used in our experiments, both for fine-tuning and evaluation. We provide the number of samples for all available splits and the total size of the dataset.

	Dataset	Split	Samples	Size
DPO	EPFL	Tra.	21,390	XX MB
		Val.	5,348	
MCQ	ARC-E.	Tra.	2,251	682 MB
		Val.	570	
		Test	2,376	
	ARC-C.	Tra.	1,119	680 MB
		Val.	299	
		Test	1,172	
	SciQ	Tra.	11679	10.5 MB
		Val.	1,000	
		Test	1,000	
	OBQA-M.	Tra.	4,957	2.88 MB
		Val.	500	
		Test	500	
	OBQA-A.	Tra.	4,957	1.18 MB
		Val.	500	
		Test	500	
	GPQA-D.	Tra.	198	XX MB
	GPQA-E.	Tra.	546	XX MB
	GPQA-M.	Tra.	448	XX MB
	MMLU-S.	Tra.	3317	XX MB
	EPFL	Test	253	XX MB

A.3 Base Training Setup

A.4 Quantitative Results

Table 4: **Baseline Results.** Accuracy and Standard Error (SE) for baseline models.

	LLama	OpenELM	Phi-3
ARC	67.4 ± 1.1	50.2 ± 1.2	69.4 ± 1.1
GPQA	29.8 ± 2.5	23.6 ± 2.3	33.6 ± 2.5
MMLU	54.5 ± 4.0	24.0 ± 3.5	57.3 ± 3.9
OBQA	34.0 ± 2.1	25.4 ± 1.9	38.8 ± 2.2
SciQ	96.3 ± 0.6	89.7 ± 1.0	95.5 ± 0.7

Table 5: **Finetune Results.** Accuracy and Standard Error (SE) for fine-tuned models and Phi-3 baseline.

	Phi-3-Arc	Phi-3-DPO	Phi-3-MCQ	Phi-3-OBQA	Phi-3-SciQ
ARC	66.3 ± 1.1	66.0 ± 1.1	65.9 ± 1.1	61.8 ± 1.2	71.6 ± 1.1
GPQA	33.2 ± 2.5	33.2 ± 2.5	33.1 ± 2.5	31.8 ± 2.5	26.5 ± 2.4
MMLU	57.8 ± 3.9	57.0 ± 3.9	55.8 ± 3.9	55.7 ± 4.0	22.1 ± 3.5
OBQA	40.4 ± 2.2	37.6 ± 2.2	38.2 ± 2.2	38.0 ± 2.2	37.2 ± 2.2
SciQ	95.4 ± 0.7	95.1 ± 0.7	97.0 ± 0.5	94.1 ± 0.7	97.4 ± 0.5

Table 6: **Quantisation Results.** Accuracy and Standard Error (SE) for quantised models and its baseline.

	GPTQ-2b	GPTQ-3b	GPTQ-4b	GPTQ-8b	Phi-3-Arc
ARC	23.8 ± 1.1	55.9 ± 1.2	64.1 ± 1.2	66.4 ± 1.1	66.3 ± 1.1
GPQA	23.2 ± 2.3	28.7 ± 2.4	32.0 ± 2.5	32.7 ± 2.5	33.2 ± 2.5
MMLU	24.2 ± 3.6	46.0 ± 4.1	56.0 ± 4.0	57.8 ± 3.9	57.8 ± 3.9
OBQA	13.0 ± 1.5	33.8 ± 2.1	37.6 ± 2.2	40.0 ± 2.2	40.4 ± 2.2
SciQ	22.9 ± 1.3	93.2 ± 0.8	95.5 ± 0.7	95.4 ± 0.7	95.4 ± 0.7

A.5 Data Examples

```
{
    "course_id": 1,
    "question_id": 1,
    "question": "...",
    "question_options": "...",

    "A" : "...",
    "B" : "...",

    "ranking_criteria": {
        "overall": "A",
        "correctness": "B",
        "relevance": "AB",
        "clarity": "None",
        "completeness": "A",
        "other": "Conciseness: B"
    }
},
```

Listing 3: EPFL Preference Data Example

```
{
    "answerKey": "B",
    "choices": {
        "label": ["A", "B", "C", "D"],
        "text": ["Shady areas increased.", "Food sources increased.", "Oxygen levels
     increased.", "Available water increased."]
    },
    "id": "Mercury_SC_405487",
    "question": "One year, the oak trees in a park began producing more acorns than
    usual. The next year, the population of chipmunks in the park also increased.
    Which best explains why there were more chipmunks the next year?"
}
```

Listing 4: ARC Data Example

```
{
    "correct_answer": "coriolis effect",
    "distractor1": "muon effect",
    "distractor2": "centrifugal effect",
```

```
    "distractor3": "tropical effect",
    "question": "What phenomenon makes global winds blow northeast to southwest or
    the reverse in the northern hemisphere and northwest to southeast or the reverse
     in the southern hemisphere?",
    "support": "\"Without Coriolis Effect the global winds would blow north to south
     or south to north. But Coriolis makes them blow northeast to..."
}
```

Listing 5: SciQ Data Example

```
{
    "id": "7-980",
    "question_stem": "The sun is responsible for',
    "choices": {"text": ["puppies learning new tricks",
    "children growing up and getting old",
    "flowers wilting in a vase",
    "plants sprouting, blooming and wilting"],
    "label': ["A", "B", "C", "D"]},
    "answerKey': "D",
    "fact1': "the sun is the source of energy for physical cycles on Earth",
    "humanScore": 1.0,
    "clarity": 2.0,
    "turkIdAnonymized": "b356d338b7"
}
```

Listing 6: OpenBookQA Data Example

```
{
    "question": "What is the embryological origin of the hyoid bone?",
    "choices": ["The first pharyngeal arch", "The first and second pharyngeal arches
    ", "The second pharyngeal arch", "The second and third pharyngeal arches"],
    "answer": "D"
}
```

Listing 7: MMLU (Anatomy) Data Example

```
{
    "question": "The proof for the chromosomal theory was obtained from...",
    "correct_answer": "The cross demonstrating X chromosome nondisjunction",
    "incorrect_answer1: "The cross between pea plants having yellow smooth seeds to
    those with green wrinkled seeds",
    "incorrect_answer2: "The cross between tall plants and short plants",
    "incorrect_answer3: "The cross between purple flowering plants and white
    flowering plants"
}
```

Listing 8: GPQA Data Example

A.6 Data Formatting

```
<s><|user|>
{{Question}}
Input:
A. {{Option 1}}
B. {{Option 2}}
C. {{Option 3}}
D. {{Option 4}}
<|end|>
<|assistant|>
{{Support}}
Therefore the correct answer is {{Correct Answer}}.
<|end|>
```

Listing 9: MCQ Formatting