PREFERENCE MODELLING OF LANGUAGE MODELS FOR AGE-APPROPRIATE RESPONSES TO QUESTIONS IN K-12 ENVIRONMENTS

SN: 24251714

ABSTRACT

Current large language models have shown an inability to provide age-appropriate answers to students across the K-12 spectrum when provided with limited context on the student. This has consequences for potential applications within the education sector, where these technologies may look to be used as chatbots to assist with learning to help alleviate stressors in the industry. This report has tried to address this limitation by using Direct Preference Optimisation preference modelling to fine-tune two model architectures on a novel dataset of preference pairs. The dataset contains questions with added context of the students' grades from the K-12 range and has been evaluated by an independent teacher for suitability. The optimised models showed a significant improvement in providing age-appropriate responses when assessed by a separate primary school independent teacher and a graduate student. The study highlighted difficulties for learning age-appropriate answers to questions by students in middle school and linked the difficulties to a lack of understanding of the exact knowledge level of the student. The models have also been benchmarked against the untrained model on common problem-solving benchmarks. Optimised models were shown to perform on par with the base model, demonstrating that improvements in age-appropriate responses do not sacrifice the models' problem-solving capabilities. The code for this study has been can be found here[1].

1. INTRODUCTION

Large language models (LLMs) such as OpenAI's GPT-4[1] have changed the way many of live and work. Being able to ask questions on nearly any topic and receive a well-crafted and often correct response has given people newfound abilities for productivity in both the workplace and higher education.[2][3] These models have shown to have good problem solving abilities across a range of fields, including being able to pass a simulated bar exam among the top 10% of test takers.[1]

For many researchers, advancing the state of the art in LLMs involves improving their ability to solve increasingly complicated tasks and questions in increasingly intricate, accurate and detailed ways.[4] While true and important for many applications, in others, the opposite is required. To quote Richard Feynmann, the great Physicist and educator, "If you cannot explain something in simple terms, you don't understand it". In the educational setting, being able to explain a topic based on the receiver's level of understanding is a skill of the upmost importance. It requires a deep understanding of what the key ideas of the topic are, to be able to convey them at such a level without confusing the student. This is universal, from children learning how to read to college students learning Natural Language Processing. Unfortunately, explaining complex topics to young children can be quite tricky, like "Why is the Sky Blue?" or "Who named the days of the week?" To effectively answer these types of questions without confusing the child requires a good understanding of the question itself, but also a good understanding of what the child would understand. If you ask ChatGPT to explain the answer to either of these questions as if it were talking to a 2nd grade child, it produces answers that would likely only confuse them more. In the first question, the Chatbot discusses ideas related to the scattering of light waves in the atmosphere, highlighting how blue gets scattered more than other colours. Although correct, immediately one can highlight that a 2nd grader would have no idea what you are talking about and that a simple answer such as; "because the light from the sun is actually made up of all the colours of the rainbow and when it gets to earth they all disappear except blue which shines through" would be far more suitable. This somewhat humorous experiment can be repeated on websites such as Chatbot Arena[5], where they all seem to fail with young children at explaining common topics in a very simple way.

The education sector in the UK and in many countries around the world is currently facing a crisis. [6][7] The lack of teachers around the world is having an impact on the quality of education being provided to children. In developing countries this has been a persistent problem for many years that requires significant investment to overcome. However, developed countries are now facing similar problems, particularly in the elementary or primary school environment. In the UK, the recruitment of new trainee teachers for the primary sector reached just 88% of the intended target, decreasing year on year from 94%.

The education sector has rapidly embraced technology, with teachers and schools continually seeking innovative and

[1] https://github.com/phil-mira/NLP_new

effective ways to enhance student learning.[8] AI, particularly large language models (LLMs), is a promising technology that could significantly enhance learning and help address the challenges posed by declining teacher numbers.[9] Imagine a scenario where each student has access to an LLM to help them learn to read and write and answer questions that they have on an individual basis, with teacher there to monitor and supervise the children as they learn. Although there exist criticisms to this particular scenario introduced by a lack of social interaction between students and teachers, the point still holds that there is a unique potential for LLMs to transform this sector.

Context reasoning forms a crucial part of evaluating LLM performance and is ultimately what this study is exploring and trying to improve. For this study, improving the model's understanding of the knowledge level of a child is the goal. In theory, this could be achieved by providing the model with a significant amount of contextual information about the child's current knowledge level and ability. In some ways, this scenario would be an accurate reflection of how a teacher in a school would know the level of understanding of each of their pupils. However, providing the model with significant context does not cover the human ability to provide suitable answers to children given very little information about them. In other words, it does not reflect the human ability to reason about child development stages with limited context.

This report investigates whether an LLMs' lack of adaptability from a limited student context description can be improved by model-fine tuning. Two main research questions have been posed to guide this study: 1) Can Preference Modelling be used to improve the use of LLMs for age suitable answers when provided when limited context? 2) Do optimised models exhibit performance losses in other tasks that may affect ability to answer questions correctly?

These questions have been approached by attempting to fine-tune a pre-trained Mistral 7B [10][11] parameter model using direct preference optimisation (DPO) [12] on preference pairs of data to see if existing methods to tailor models to specific tasks are effective. The preference pairs have been produced in consultation with a qualified school teacher. The models have been evaluated by having a human choose which model produced the most suitable answer. For the students between the grades kindergarten and 3rd grade, a separate primary school teacher has been consulted to evaluate the results of each of the models. The models have also been evaluated against common benchmarks to assess whether the new model has lost performance in other domains.

The main contributions of this work have been a novel dataset of preference pairs that can be used to fine-tune an LLM to produce more age-suitable answers to student questions from grades K-12. As well as a systematic evaluation of the performance of models trained on this dataset.

2. LITERATURE SURVEY

Bewersdorff et al. [13] introduced a framework for how multimodal large language models (MLLMs) can be used to advance science education. They highlighted how these models are well-suited for a range of applications, including content creation to tailored support for learning, fostering engagement in scientific practices, and providing assessments and feedback.

Lee et al. [14] trained a multimodal model using a vicuna-13b-v1.5 LLM and clip-vit-large-patch14 vision encoder on a custom dataset termed LLaVA-Docent to support art appreciation in education. The researchers consulted several subject matter experts to design the training dataset. They noted that the models should be specific to the target audience and adjust the response accordingly. To cater to their specific task, they primarily used prompt engineering to fine-tune their outputs. They tested their model in a few-shot learning environment against Open AI's GPT-4. Despite highlighting this in their prompt, they provided no analysis of whether the model was successfully able to cater towards the individual's level. Furthermore, the prompts provided were largely generated using GPT-4 which as highlighted lacks the ability to appropriately adjust answers based on the user's grade. The authors provide an analysis that their specific model lacks rigour in accuracy as well as suitability in real-world environments, and suggest possible RAG frameworks as well as further research to address these issues.

Liu et al. [15] performed a study where they trained a chatbot to act as a reading companion to students to assist them in reading and comprehension of the book. The chatbot had a basic understanding of the book and showed that Children enjoyed more interaction and understanding of the book when accompanied with the chatbot. This chatbot however was a basic model based on Google Actions Framework and had no knowledge outside of these tasks. This study does provide motivation for the value that chatbots can add to the educational environment.

Ling and Afzaal [16]present a comprehensive evaluation of the use of large language models (LLMs) for the automatic generation of question-answer (QA) pairs in higher education. Acknowledging the time-consuming nature of manual assessment creation, the study explores three prominent LLM-based approaches—pipeline, joint, and multi-task learning—evaluated through automated metrics, teacher feedback, and real-world classroom performance across three computer science courses. The findings indicate that the multi-task approach, particularly when using the T5 model, outperforms others in generating accurate and pedagogically relevant QA pairs. Teachers reported high satisfaction with question correctness and relevance, although they noted room for improvement in the clarity and difficulty of distractors. Notably, students who engaged with the automatically generated assessments demonstrated significantly higher academic

performance, and a positive correlation was found between the number of assessment attempts and final exam scores.

Pitis, Xiao, Le Roux and Sordoni [17] highlight the importance of context in the often underspecified nature of natural language and develop methods to improve preference modelling using a two-stage approach that first identifies the context before providing preferences. They develop a context-conditioned preference dataset to investigate the ability of language models to evaluate context-specific preference. They show preference models benefit from, but fail to fully consider, added context.

Maity, Deroy, and Sarkar[18] explore the application of large language models (LLMs) for generating educational questions from school-level textbooks, addressing the labour-intensive nature of manual question creation. Their study investigates the capability of GPT-4 Turbo, GPT-3.5 Turbo, Llama-2-70B, Llama-3.1-405B, and Gemini Pro—to generate "complete sets of questions" and classify them according to Bloom's revised taxonomy[19]. Using both zero-shot and eight-shot prompting techniques, the authors assess the generated questions through human evaluation based on coverage, grammaticality, usefulness, accessibility, relevance, and redundancy. Findings reveal that while human-generated questions consistently outperform LLMs, few-shot prompting significantly improves LLM performance, particularly for GPT-4 Turbo and Llama-3.1-405B. The eight-shot setting enhances alignment with pedagogical frameworks, reduces redundancy, and yields questions with improved cognitive diversity. This study distinguishes itself from previous AQG research by incorporating textbook-derived educational content and systematically evaluating outputs against cognitive learning objectives. It underscores the growing potential of LLMs to support educators by generating high-quality, curriculum-aligned questions that foster critical thinking across a spectrum of learning outcomes.

Rooein et al. [20] conducted a comprehensive evaluation of four state-of-the-art LLMs (two commercial and two open-source) to assess how well they adapt their responses to different age groups and education levels when explicitly prompted. Using standard readability metrics to evaluate responses to science questions, their findings revealed that current LLMs have predetermined readability ranges and struggle to adjust their content appropriately for different audiences—even when specifically instructed to do so. Their analysis looked at audiences between the ages of 11 to 23 and showed that on average, only about 15% of LLM-generated responses fell within the recommended readability range (based on the Flesch-Kincaid Reading Ease Index (FKRE)for the requested audience. This limitation poses a significant challenge for educational applications, where age-appropriate content is essential for effective learning. The researchers concluded that while LLMs demonstrate some potential for educational use, their current inability to reliably adapt to different audience demographics restricts their effectiveness as educational tools without significant improvements in audience-specific content generation. This study has many parallels with the work carried out in this report, with this report attempting to build on some of the highlighted problems with current LLMs.

To the best of my knowledge, no studies have investigated the capability of fine-tuning a large language model (LLMs) to the specific task of providing age-appropriate answers suitable to school-aged children across K-12 with limited context.

3. BACKGROUND AND PRELIMINARIES

This section provides a basic overview of Transformer LLMs[21] as well as direct preference optimisation[12], which has been used for fine-tuning, leading to a description of how the models have been used in this study. A basic theoretical hypothesis is made on why the attention layer in transformer models, combined with direct preference optimisation, is suitable for the task.

3.1. Transformers and Attention layers

Transformers form the backbone of many modern deep learning models, with the attention mechanism being the key mechanism in these architectures. The underlying principles of attention layers have been around since the 1980s[22], however, they were first described in their current form in the 2010s[23][24] and successfully implemented into Transformer architectures in 2017.[21] The core idea behind attention in language modelling is that, after the input text is tokenised and assigned both positional and input embeddings, the attention layer calculates how strongly every other token influences each token. For text generation this means that every token generated has the knowledge of every word before it, along with how each of those words relates to every other word.

For language modelling, masked-attention is used, which ensures that during training, future tokens do not get attended to along with multi-headed attention, which expands the number of parameters in each layer by having multiple "heads" that focus on different features before being combined. The mistral-7B parameter model also uses sliding window attention [25][26] to improve computational efficiency.

A single head of Attention for Multi-headed attention can be defined as:

$$\begin{aligned} &Attention(QW_i^Q, KW_i^K, VW_i^V) \\ &= softmax\left[\frac{QW_i^Q(KW_i^K)^{\mathbf{T}}}{\sqrt{d_k}}\right]VW_i^V \\ &= A_iVW_i^V \end{aligned} \tag{1}$$

where $Q, K, V \in \mathbb{R}^{n\times d}$ are the query, key, and value embeddings, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_k}$ are learned projection

matrices, $d_k = d/h$ is the embedding dimension for each head in multi-headed attention and $A_i \in \mathbb{R}^{n \times n}$ denotes the full attention matrix for each attention head in multi-headed attention.

For this study this contextual understanding is vital to the performance of the model in this setting, as it is ultimately what is being investigated. Thus, for the model to be successful, it most have both a sufficient understanding of what it means to be in each grade, but also provide enough weight to this token such that generated text keeps it "in mind" for every word it uses. This weighting between words is provided by the A term in the equation, which is a matrix of size n^2. Inspecting this matrix gives insight into how the model is learning contextual-dependencies. However, as a different A is learned for each head and at each layer, these heads and layers need to be either individually inspected or aggregated in some way to allow comparisons between models.

3.2. Model Fine Tuning

LLM fine-tuning is the process of steering a model to have more precise control over the behaviour it exhibits. It is often used at deployment for areas such as chat-bot content moderation and censorship of material. In theory, the model should be able to produce outputs that are more suitable for a given age group once given more information about what a more suitable answer is, as preference modelling can adjust the attention weights of the model to pay more attention to the important contextual clues in the prompt.

The prevailing method to achieve this uses pairwise preference datasets of different outputs and training the models using Reinforcement Learning from Human Feedback (RLHF) [27][28] or more computationally efficient methods such as Direct Preference Optimisation (DPO) [12]. The goal of RL-based models is to find a LLM policy π, whose response y given an input x maximises a reward function $r(x, y)$.

The Bradley-Terry (BT) model [29] is used to model the reward from the preferences of the preference dataset. It says that the human preference distribution p^* or the probability that y_1 is preferred to y_2 can be written as:

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp(r^*(x, y_1)}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} \quad (2)$$

Where r^* is the reward model of preferences to be learned. DPO avoids the training of a reward model, which is used in RLHF, which can be unstable and expensive to train. Instead, the analytical solution of the RLHF objective is rearranged to derive a reward given by:

$$r(x, y) = \beta \log \frac{\pi(y \mid x)}{\pi_{ref}(y \mid x)} + \beta \log Z(x) \quad (3)$$

where τ controls the deviation of the KL-divergence in the reward function for RLHF and $Z(x)$ is a normalisation constant. This can then be combined with the BT model for the following DPO objective function:

$$\begin{aligned} \mathcal{L} &= -\mathbb{E}_{(x, y_\omega, y_1) \sim \mathcal{D}}(B) \\ B &= \log \sigma \left(\beta \log \frac{\pi_\theta(y \mid x)}{\pi_{ref}(y_\omega \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{ref}(y_l \mid x)} \right) \end{aligned} \quad (4)$$

4. METHODOLOGY

4.1. Context of Work

For now, the model is considered to be used as a learning tool whereby a student may ask it questions with responses suitable for their grade. At the younger ages it is not intended to enhance reading ability, but it should, to the best of its ability produce an answer that the student would both be able to read and/or listen to.

It should be noted that there is the problem that arises from differences in oral comprehension and written comprehension in young children, for example the disparity even between children can be significant. For this study, it is assumed that the child would have access to both the written and spoken versions of the text. Although this may not be the most effective method to teach a child to read and write, it provides a means to provide sufficient answers to questions without running into the difficulties involved with the child's reading ability. Although oral and written comprehension are not mutually exclusive, they need to treated in such a way that a child must be able to understand the question in at least one of the formats, preferably both.

4.2. Preference Pairs Dataset Creation

As highlighted in section 3.2, DPO requires a dataset of preference pairs to steer the model towards producing answers that are suitable for the task at hand. No existing datasets could be found that were suitable for this task, and so a new one has been developed.

The dataset needed to capture the nuances that may appear in answers to student questions. Depending on the student's level, the preferred responses should steer the model away from unsuitable answers.

To achieve this, three main aspects were considered: the amount of context about student provided to the model, the structure of the student's questions, development of preference pairs.

In total, 200 questions and sets of preference pairs were generated with target grade levels of the students having an even coverage across the K-12 range.

4.2.1. Amount of Context of Student

As highlighted in the introduction this study intends to use only a limited context of the student in the prompt. For this

study, it was decided that only the grade level of the student would be provided to the model. Each question has been combined with a starting prompt sentence following the basic format of ”I’m in xth grade.” Where x is some grade in the K-12 range. This was chosen based on the reasoning that with only this limited piece of information an answer that was suitable for that grade should be able to be generated, akin to the human capability.

There are clearly issues with regards to what is considered the ability of a child at each grade level, as this may vary significantly between countries and individuals. This problem has been discussed in the development of preference pairs.

4.2.2. Structure of Questions

Questions asked by different grade levels can provide an additional level of context to the model that may be able to affect how it produces answers. For example, suppose the model is asked about the I-V characteristics of LED and Filament lightbulbs by a high-school student. This question implies that the student is aware of the types of lightbulbs and likely has a general understanding of electrical circuits. This means the model can then tailor the answer to create a more suitable response based on this information, possibly ignoring the context that the child is in high school, thereby degrading the ability to teach the model the importance of this information. To mitigate this, the questions in the dataset are simplistic but have the potential to be possible questions across the grade spectrum. For example, the question of ”How to aeroplanes fly?” is a question that a kindergartener may ask, but also can be explained in immense detail even beyond what is understood by most high-school students.

In terms of topics, questions covered a wide range of subjects including maths, science, geography and history. For some areas, generating suitable questions becomes more ambiguous, such as within maths, where younger students are not exposed to topics that may be as advanced as readily when compared with science, where simple observations can have complex answers.

All questions followed the same format, with the grade of the student as the first sentence of the prompt, followed by the question being asked. An example of this can be seen in Figure 1.

4.2.3. Development of Preference Pairs

Preference pairs have been created through a mix of text generation via a more power LLM and hand-crafting. The responses were then assessed by a school teacher recruited for the study to ensure that the answers for the age groups were suitable. The teacher was told to rank and evaluate whether the chosen preference pair was suitable or not, and to provide feedback if the answer was not suitable for the age group. The criteria for evaluation were on a scale of: 1 for not suitable, 2 for somewhat suitable and 3 for suitable. The teacher

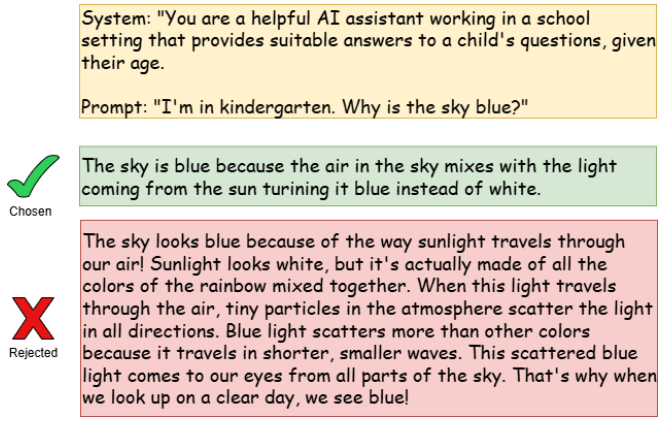

Fig. 1. An example of the structure of the preference pairs as well as examples of the questions being asked and their responses.

was told to consider both the reading and listening levels of the students in their analysis. This was particularly important for younger age groups, where there is a large gap between reading and listening comprehension. This feedback was then used to update the responses. This feedback also helped alleviate some of the problems associated with knowing the level of understanding of students at each grade level. Although, this solution is insufficient across all schools, it does provide a grounding for schools located in the UK and the US, where this teacher had knowledge and experience teaching in.

Initially, a chatbot was prompted to generate preference pairs for the questions given the age. The responses were then evaluated and adjusted by hand to ensure that they were suitable. To improve the quality of the data, preference pairs were adjusted so that they were very similar in response, but key differences were made to make one of them more suitable than the other. This is in contrast to using a response suitable for a kindergarten as the rejected response when prompted by a question for a 12th grader. Having to manually adjust all responses limited the size of the dataset that could be generated due to the time spent on each question. However, research has shown that for reinforcement learning from human feedback [27], another method for LLM fine-tuning, dataset quality may be more important than scale for achieving the desired performance. [30] It is also not possible to generate accurate responses from the LLM, as it is an inherent flaw in the current state-of-the-art models. [20]

4.3. Model Training Methods and Architectures

Two different training methods were used to explore potential improvements in the models. Both of these models were optimised using DPO due to the computational advantages of using it and the limited amount of data available to effectively train a reward model.

In the first model, termed the All Grades Model, the LLM

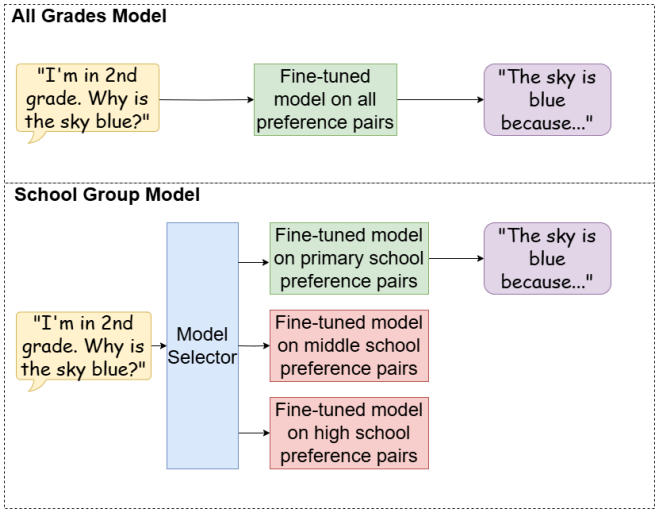

Fig. 2. Two separate models were trained to experiment whether improvements can be gained by using subsets of the data for specific tasks. The All Grades Model used all of the data in the dataset, while the School Group Model splits the data between Primary, Middle and High school in a Mixture of Experts fashion.

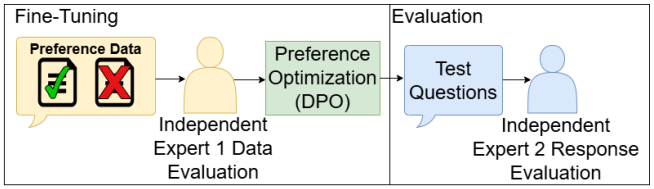

Fig. 3. Two trained Primary School Teachers were recruited as expert evaluators to assess both the dataset and the responses of the models.

was fine-tuned on the entire preference dataset for each of the different age grades. In the second model, termed the School Group Model, three separate LLMs were used, one for each school group (Primary, Middle and High Shcool) in a sort of Mixture of Experts style design. For example, all preference pairs for Kindergarten to grade 5 were used to train the primary school model, whereas those for 6-8 were used in the middle school model. In the model pipeline the model assesses which age group LLM to send the prompt to, given that all prompts follow the same format, and returns the answer only from that specific prompt.

4.4. Model Evaluation

The performance of the responses needed to be evaluated in the context of how suitable the answer was for the grade level. The hardest demographic for this was the primary school grades. This was due to the type of language and topics used to explain these topics to them, which teachers of middle school and high school need to worry less about. For this reason, an external primary school teacher was recruited to evaluate these responses. This teacher was different to the teacher who was used to evaluate the preference pairs, this was done to help reduce bias. For the responses for older grades, a graduate student evaluator was used. An avenue for future work may be to evaluate whether LLMs are able to perform this evaluation task instead of humans.

The test dataset used for evaluation consisted of questions following the same format as discussed in the fine-tuning stage. The questions again covered a range of topics. Each of the primary, middle and high school groups was evenly covered. A total of 72 questions were created.

The base model was compared to the full model and the mixture of models separately. This was chosen over having three possible responses for each question to help reduce the load on the evaluator. It is theorised that too many options can lead to decreased decision making ability, this is known in cognitive science as ”cognitive load theory”.[31]

The following criteria was used for evaluation: 0 for Neither response is suitable, 1 for First Response is Best, 2 for Second Response is Best, 3 for Responses are Equally Good. Incorporating the 0 and 3 criteria allows a more comprehensive evaluation of how the models are performing. The evaluator was not told which responses were from the trained and untrained model.

The models also needed to be benchmarked against common problem-solving datasets. This is to ensure that there is no loss in the ability of the models to generate correct answers for a range of problems, which would degrade the efficacy of the model in an educational setting. The following benchmarks were used for evaluation: AI2 Reasoning Challenge (ARC)[32], HellaSwag[33], GSM8k[34], Measuring Massive Multitask Language Understanding (MMLU: formal logic, high school world history, high school geography, high school government and politics, high school biology, high school chemistry)[35], SciQ[36]. These were chosen for their relevance to the study and their range of topics.

5. IMPLEMENTATION

The study makes use of Hugging Face to fine-tune the models. Hugging Face allows users to download open-source models and easily train them using a wide variety of available options to choose from, making it fast and straightforward to develop the models.

As discussed, the data was created through a mixture of generative means using Openai GPT-4 and adjusting answers manually. The data was placed into a JSON file of preference pairs following the format that was specified by the model being used. Short scripts were also written to convert the data into a txt document format for evaluation by the teacher.

The model that was used was a chatbot fine-tuned version of the Mistral-7B model from teknium[37] on HuggingFace.

This model was chosen as it boasts improvements over the standard Mistral AI 7B Instruct fine tune on common benchmarks. The 7B parameter model was used as a trade-off between performance and size. As the model was producing answers that would be reviewed by an external assessor, the model needed to be powerful enough to produce coherent answers while also being small enough to run on a single GPU.

Google Colab was used for the training and inference of the model thanks to readily avaliable A100 GPUs. Due to the size of the model chosen this was the only available GPU on Google Colab that would be able to fine-tune the model efficiently due the memory constraints.

This model had a specific format that all of the prompts needed to follow. This formatting was handled by a function at the start of each script, which also added the additional system prompt to better guide the optimisation as seen in Figure 1. For the School Group Model, this included additional logic to handle selecting the correct model to use for the level of the question.

To further speed up the optimisation as well as reduce computational and memory loads several additional techniques were used. Firstly, quantisation of the model parameters was used to reduce the memory footprint, the parameters of which were chosen based on the guidance from the Hugging Face documentation. Secondly, Low Rank Adaptation (LoRA) was used as the trainable parameters of the model. LoRA freezes existing model parameters and instead places tunable rank decomposition matrices into each layer of the model. This acts to greatly reduce the number of trainable parameters and speed up training. QLoRA (Quantised Low-Rank Adaptation)[38] is the technique that combines these two methods together and has been used for model tuning. The tunable parameters used were again taken from the Hugging Face documentation. Lastly, Flash Attention 2 has been utilised, which is a hardware-aware algorithm that greatly improves the GPU utilisation by minimising non-matrix multiplication operations, allowing larger models to be used. Additional hyperparameters of the model have been reported in Table 5. Note that due to computation restrictions, these were not fine-tuned and were taken from DPO tutorials on the Hugging Face website.

Due to computational restrictions, the maximum length of the response sequences was limited to 200 tokens. Although for older age groups, this meant that responses were cut off, answers were still able to be evaluated effectively by the evaluators. A total of 200 steps has been used for each model again to reduce computational load, this was used over epochs to ensure the same amount of data was used to train each model. Each of the trained models have been uploaded to Hugging Face for future use. This included each of the individual expert models (i.e. Primary, Middle and High). Weights and Biases has been used to track the training of each of the models. Generally, all of the models appear to be learning to choose the correct preference, low losses and high and increasing reward margins.

Table 1. Training Hyperparameters of the models.

Parameter	All Grade Model	School Group Model
Rank-LoRA	16	16
Alpha-LoRA	16	16
Dropout-LoRA	0.05	0.05
int8 Threshold	6.0	6.0
4-Bit - QLoRA	nf4	nf4
Batch Size	4	4
Learning Rate	$5e^{-}5$	$5e^{-}5$
Scheduler	Cosine	Cosine
Optimizer	AdamW-32bit	AdamW-32bit

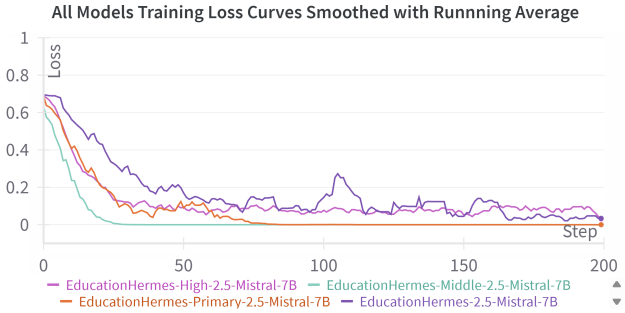

Fig. 4. Training Loss curves for each of the models, smoothed to provide improved visualisation of progress. Terminated after 200 steps.

For the benchmarking of the models the EleutherAI LM evaluation harness has been utilised[39]. This framework allows fast and convenient benchmarking of the models once uploaded to Hugging Face. The benchmarks that have been used are specified in section 4.4.

For the visualisation of the attention layer all the layers and their heads have been combined by averaging all the weights together. Although this fails to capture the intricacies of each of the dependencies it provides a general sense of the relationships that the model is prioritising.

6. EXPERIMENTAL RESULTS AND ANALYSIS

The results from the evaluators' responses to the test questions, shown in Figure 6, indicate that both of the optimised models appear to be producing better answers to the test questions across all of the grade levels. Interestingly, the biggest improvement comes from the Primary school grade range, where both models vastly improved responses to the questions. This is particularly useful as this was the group that was largely underperforming during the exploration phase of this study. The questions where the model performs worse than the base are generally in the older age groups for the pri-

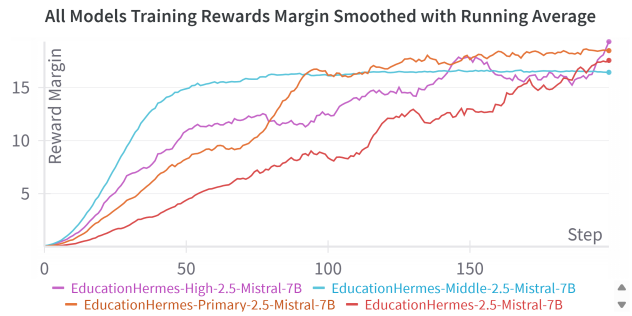

Fig. 5. Reward Margin for each of the models, defined as the mean difference between the chosen and corresponding rejected rewards, smoothed to provide improved visualisation of progress. Terminated after 200 steps

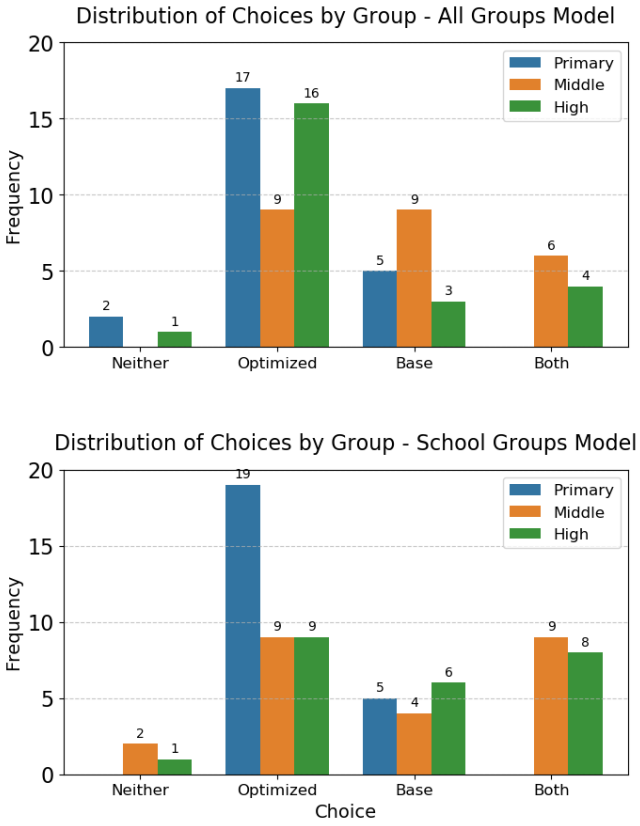

Fig. 6. Bar Chart of the responses from the evaluators when comparing each of the optimised models against the base model. Evaluators were given the option to choose either none of the models or both if the answers were just as suitable. Choices have been split by school group for analysis.

mary group (4th and 5th grade), where the optimised model tends to prioritise more simplistic answers over those that are correct.

For the middle and high school age groups, although there is generally an improvement across both models, it is generally less pronounced than for the primary school group, save the All Groups model on the High school group, which performs on par with that of the primary school group. During evaluation, it was found that generally these answers were harder to distinguish between, resulting in "both" being chosen more for these groups.

The "All Groups" model appears to perform best between the two models. This may be caused by the model learning stronger relationships between the different groups to allow it to complement the answers better. However, it is unclear exactly why this is. What can be seen is that at the extremes, i.e. Primary and High school groups, the answers are of better quality in the optimised models. This may be because the Middle school age group is more nuanced in terms of the level the students may be at. For these age groups more context may be required than simply the grade, as the gap between knowledge levels is more subtle compared with the extremes. For example, for science based questions a high school student should have good understanding of some mathematical concepts and basic theory to be able to understand complex topics however for middle school student it is unclear at what age they might be exposed to this topic so choosing whether to include that in the response becomes more difficult.

This hypothesis may be further supported by the analysis of the attention weight heatmaps, shown in Figure 7. In it, the All School Children model appears to be "paying more attention" to the part of the prompt related to the grade of the student. This shows that the model is realising the importance of this context for the answer, however, it may simply not fully understand what it means to be in a certain grade. This would make it struggle for grades where it is less clear of what that age group may know, such as in middle school years. This may also be a problem even for humans to provide answers to students for this age group when there is a limited context of a student's ability as it is less clear at what age certain topics are taught at middle school without making unrealistic assumptions.

Table 2. Table of all the benchmark results for each of the individual LLMs trained. The ARC benchmark used the Challenge dataset while the MMLU used an average of the topics in section 4.4.

Benchmark	Base	All	Primary	Middle	High
ARC	0.56	0.57	0.56	0.56	0.57
GSM8k	0.64	0.64	0.64	0.64	0.65
HellaSwag	0.63	0.63	0.64	0.63	0.63
MMLU	0.62	0.61	0.60	0.60	0.60
SciQ	0.96	0.96	0.96	0.96	0.96

The second research question involved assessing whether

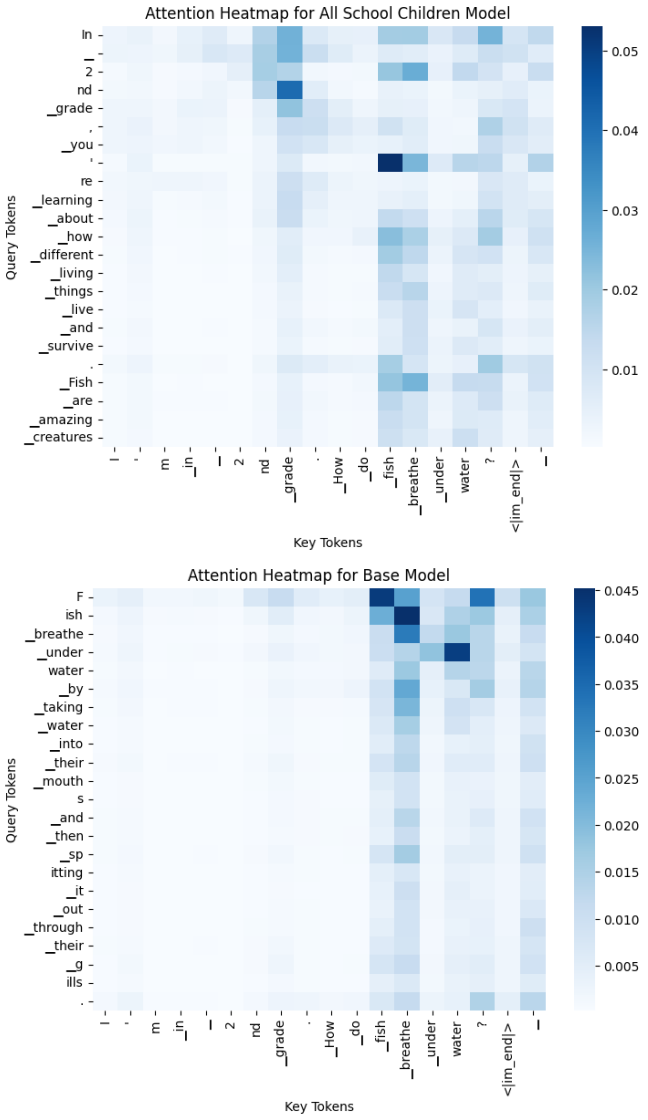

Fig. 7. Attention Layer Weights heatmap averaged over all Layers and Heads for the All School Children Model (Top) and Base Model (Bottom). The prompt, along with the beginning of the response, has been included.

the optimised models experienced any performance losses on common benchmarks that may inhibit their problem-solving ability. Table 6 shows the results from the benchmarking. In it all of the models perform on par or better with the base model for all benchmarks, save that of MMLU where the optimised models perform slightly worse. Not only does this provide confidence that there is no performance drop during optimisation, but also, reinforces the hypothesis that the model is using the context from the grade levels in its answers only when prompted.

7. CONCLUSION

In this study, LLMs have been successfully optimised for the task of improving answer suitability for school-aged children in K-12 schools. This has been achieved with limited resources using Direct Preference Optimisation on open-source LLMs from Hugging Face. An original dataset of preference pairs for the preference modelling has been developed for this study. This dataset has been validated by an independent teacher with experience teaching across K-12, to ensure the suitability and accuracy of the preference pairs. Two separate architectures were tested to enhance the breadth of the study and explore potential alternatives. The first architecture was a single LLM trained on the complete dataset. The second architecture consisted of three separate LLMs, each trained on subsets of the data as Primary, Middle and High school experts with an identifier function for model selection.

The optimised models have been evaluated on a test dataset of questions of the same format of the preference pair prompts by a separate independent teacher for primary school student aged questions and by an MSc-level student for the middle and high school student aged questions. The evaluations showed a marked improvement in performance for all age groups in both models, particularly that of the primary school group. The study highlighted insufficiencies in the Middle School age group and linked it to an inability to successfully identify the knowledge level of these students. The models have also been benchmarked across common benchmarks, showing that there has not been a drop in problem-solving ability as a result of the preference modelling.

Future work may look to explore how knowledge gained from this study can be applied to technologies within a real-world educational setting and look to work further with teachers to apply chatbots to schools to ease current problems facing the sector.

Acknowledgements

I'd like to thank Gina Vargus and Karen Miranthis, who acted as the independent experts in this study.

8. REFERENCES

[1] OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal et al., "Gpt-4 technical report," 2024.

[2] Shakked Noy and Whitney Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187–192, 2023.

[3] OpenAI, "College students and chatgpt adoption in the us," `https://openai.com/global-affairs/college-students-and-chatgpt/`, March 2024, Accessed: 2025-04-03.

[4] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao, "Large language models: A survey," 2025.

[5] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica, "Chatbot arena: An open platform for evaluating llms by human preference," 2024.

[6] Ben Arnold and Mark Rahimi, "The global status of teachers 2024," January 2025.

[7] Education Committee, "Teacher recruitment, training and retention," May 2024.

[8] David G Dewhurst, Hamish A Macleod, and Tracey A.M Norris, "Independent student learning aided by computers: an acceptable alternative to lectures?," *Computers Education*, vol. 35, no. 3, pp. 223–241, 2000.

[9] House of Lords Library, "Educational technology: Digital innovation and ai in schools," November 2023.

[10] Mistral AI, "Mistral 7b," 2023, Accessed: 2025-04-15.

[11] teknium, "Openhermes 2.5 - mistral 7b," 2023, Accessed: 2025-04-15.

[12] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn, "Direct preference optimization: Your language model is secretly a reward model," 2024.

[13] Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel, "Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education," *Learning and Individual Differences*, vol. 118, pp. 102601, 2025.

[14] Unggi Lee, Minji Jeon, Yunseo Lee, Gyuri Byun, Yoorim Son, Jaeyoon Shin, Hongkyu Ko, and Hyeoncheol Kim, "Llava-docent: Instruction tuning with multimodal large language model to support art appreciation education," *Computers and Education: Artificial Intelligence*, vol. 7, pp. 100297, 2024.

[15] Chen-Chung Liu, Mo-Gang Liao, Chia-Hui Chang, and Hung-Ming Lin, "An analysis of children' interaction with an ai chatbot and its impact on their interest in reading," *Computers Education*, vol. 189, pp. 104576, 2022.

[16] Jintao Ling and Muhammad Afzaal, "Automatic question-answer pairs generation using pre-trained large language models in higher education," *Computers and Education: Artificial Intelligence*, vol. 6, pp. 100252, 2024.

[17] Silviu Pitis, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni, "Improving context-aware preference modeling for language models," 2024.

[18] Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar, "Can large language models meet the challenge of generating school-level questions?," *Computers and Education: Artificial Intelligence*, vol. 8, pp. 100370, 2025.

[19] M. D.; Furst E. J.; Hill W. H.; Krathwohl D. R. Bloom, B. S.; Engelhart, *Taxonomy of educational objectives: The classification of educational goals. Vol. Handbook I: Cognitive domain*, New York: David McKay Company, 1956.

[20] Donya Rooein, Amanda Cercas Curry, and Dirk Hovy, "Know your audience: Do llms adapt to different age and education levels?," 2023.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2023.

[22] C. Lee Giles and Tom Maxwell, "Learning, invariance, and generalization in high-order neural networks," *Appl. Opt.*, vol. 26, no. 23, pp. 4972–4978, Dec 1987.

[23] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[24] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush, "Structured attention networks," 2017.

[25] Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang, "LM-infinite: Zero-shot extreme length generalization for large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Kevin Duh, Helena Gomez, and Steven Bethard, Eds., Mexico City, Mexico, June 2024, pp. 3991–4008, Association for Computational Linguistics.

[26] Zichuan Fu, Wentao Song, Yejing Wang, Xian Wu, Yefeng Zheng, Yingying Zhang, Derong Xu, Xuetao Wei, Tong Xu, and Xiangyu Zhao, "Sliding window attention training for efficient large language models," 2025.

[27] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, "Training language models to follow instructions with human feedback," 2022.

[28] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei, "Deep reinforcement learning from human preferences," 2023.

[29] Ralph Allan Bradley and Milton E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[30] Judy Hanwen Shen, Archit Sharma, and Jun Qin, "Towards data-centric rlhf: Simple metrics for preference dataset comparison," 2024.

[31] John Sweller, "Chapter two - cognitive load theory," vol. 55 of *Psychology of Learning and Motivation*, pp. 37–76. Academic Press, 2011.

[32] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," 2018.

[33] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi, "Hellaswag: Can a machine really finish your sentence?," 2019.

[34] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman, "Training verifiers to solve math word problems," 2021.

[35] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt, "Measuring massive multitask language understanding," 2021.

[36] Johannes Welbl, Nelson F. Liu, and Matt Gardner, "Crowdsourcing multiple choice science questions," 2017.

[37] Teknium, "Openhermes 2.5 - mistral 7b," `https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B#prompt-format`, 2024, Accessed: 2025-05-07.

[38] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023.

[39] EleutherAI, "lm-evaluation-harness," 2025, Accessed: 2025-05-07.