EXPLORING ITERATIVE ENHANCEMENT FOR IMPROVING LEARNERSOURCED MULTIPLE-CHOICE QUESTION EXPLANATIONS WITH LARGE LANGUAGE MODELS

Qiming Bao[1, 2], **Juho Leinonen**[3], **Alex Yuxuan Peng**[1], **Wanjun Zhong**[4], **Gaël Gendron**[1], **Timothy Pistotti**[1], **Alice Huang**[5], **Paul Denny**[3], **Michael Witbrock**[1], **Jiamou Liu**[1]
[1]Strong AI Lab, NAOInstitute, Waipapa Taumata Rau - The University of Auckland
[2]Xtracta, New Zealand
[3]School of Computer Science, The University of Auckland
[4]School of Computer Science and Engineering, Sun Yat-Sen University
[5]School of Life and Environmental Sciences, University of Sydney
{qiming.bao, juho.leinonen, p.denny, m.witbrock}@auckland.ac.nz

ABSTRACT

Large language models exhibit superior capabilities in processing and understanding language, yet their applications in educational contexts remain underexplored. Learnersourcing enhances learning by engaging students in creating their own educational content. When learnersourcing multiple-choice questions, creating explanations for the solution of a question is a crucial step; it helps other students understand the solution and promotes a deeper understanding of related concepts. However, it is often difficult for students to craft effective solution explanations, due to limited subject understanding. To help scaffold the task of automated explanation generation, we present and evaluate a framework called "ILearner-LLM", that iteratively enhances the generated explanations for the given questions with large language models. Comprising an explanation generation model and an explanation evaluation model, the framework generates high-quality student-aligned explanations by iteratively feeding the quality rating score from the evaluation model back into the instruction prompt of the explanation generation model. Experimental results demonstrate the effectiveness of our ILearner-LLM on LLaMA2-13B and GPT-4 to generate higher quality explanations that are closer to those written by students on five PeerWise datasets. Our findings represent a promising path to enrich the learnersourcing experience for students and to enhance the capabilities of large language models for educational applications.

1 INTRODUCTION

Given the impressive performance of large language models (LLMs) in understanding and generating natural language (Wei et al., 2022a; Brown et al., 2020), it is worthwhile to explore how these large language models can be applied into the education domain. Learnersourcing is a pedagogical approach where the task of generating learning content is distributed among students, applying their collective intelligence to enhance the learning experience (Jiang et al., 2018; Khosravi et al., 2023a; Kim et al., 2015). When learnersourcing multiple-choice questions on platforms such as PeerWise (Denny et al., 2008) and RiPPLE (Khosravi et al., 2019), students are required to provide explanations for the questions they create. However, creating high-quality explanations requires a deep understanding of the question. Since it is not compulsory for students to write explanations when creating questions, students may simply neglect to provide them. The automatic generation of high-quality explanations, along with the evaluation of these explanations, can offer enormous potential for scaffolding by providing tailored support to help students progress towards greater understanding and independence, especially in learnersourcing tasks.

A typical multiple-choice question (MCQ) includes the following components: a question stem, an answer, distractors, and an explanation. Figure 1 shows an example of MCQ from PeerWise, a learnersourcing platform employed by more than 2,500 universities worldwide (Denny et al., 2008). For each question, there is only one correct answer and an average quality score rated by students, with a range from 0 to 5. The explanation, provided by the student who created the question, demonstrates the background knowledge and steps involved in solving the question.

- **Stem:** Fill in the blanks: Glycogen synthase is ______ when it is ______, which is catalysed by ______.
- **Answer:** active; dephosphorylated; phosphatases
- **Distractor 1:** inactive; dephosphorylated; kinases
- **Distractor 2:** active; phosphorylated; kinases
- **Distractor 3:** inactive; phosphorylated; phosphatases
- **Distractor 4:** active; dephosphorylated; phosphatases
- **Explanation:** Distractor 1 - Glycogen synthase is active when it is dephosphorylated, not inactive. Dephosphorylation is catalysed by phosphatases, not kinases. Distractor 2 - Glycogen synthase is inactive when it is phosphorylated, not active. Distractor 3 - Phosphorylation is catalysed by kinases, not phosphatases. Distractor 4 - Correct. Glycogen synthase is active when it is dephosphorylated, which is catalysed by phosphatases.
- **Average quality rating:** 3.3

Figure 1: An MCQ example from the PeerWise platform.

We aim to use large language models to auto-generate explanations for student questions. The integration of automatic explanation generation using LLMs in learnersourcing may offer multiple advantages. Firstly, instant feedback from large language models has the potential to boost students' learning efficiency (Dai et al., 2023). Secondly, engagement with such models can contribute to the enhancement of learner autonomy (Yildiz Durak, 2023). Thirdly, the utilisation of pretrained large-scale models facilitates the generation of multi-faceted and comprehensive learnersourced content (Khosravi et al., 2023b). Additionally, the LLMs can be fine-tuned to emulate the manner in which students rate questions (Ni et al., 2022). However, the employment of LLMs within a learnersourced context also presents challenges, due to limited access to high-quality student-written explanations for answers and the lack of automatic explanation generation and evaluation to help LLMs generate higher quality explanations that are both literarily and semantically closer to those written by students (Khosravi et al., 2023a). These challenges impede the fine-tuning of LLMs and may compromise the quality of the model outputs.

In this work, we present and evaluate a framework ILearner-LLM that generates explanations in an iterative fashion. ILearner-LLM makes use of two LLMs fine-tuned on data sourced from PeerWise: a generation model that generates an explanation from a given question, and an evaluation model that rates the quality of an explanation to a question. At each iteration, ILearner-LLM applies instruction prompting to generate an explanation, and evaluates the quality of the generated explanations by outputting a quality rating score. The quality rating is injected into the instruction prompt for the explanation generation model in the next iteration. This process is repeated multiple times in order to iteratively generate higher-quality explanations that are literarily and semantically closer to those written by students. We summarise our main findings as follows:

- Our iterative enhancement framework, "ILearner-LLM", implemented with LLaMA2-13B and GPT-4, demonstrates notable improvements over the models without the framework in generating explanations that are literarily and semantically closer to those written by students on five PeerWise datasets, as evidenced by BLEU and BERT scores.
- We find that ILearner-LLM can help instruction fine-tuned LLaMA2-13B achieve greater improvement in deeper iterations compared to applying ILearner-LLM to GPT-4. The main reason is that LLaMA2-13B was trained following the instruction which has also been used in iterative prompting. The instruction fine-tuned LLaMA2-13B to better align with the instruction, therefore allowing the model to learn and perform better in multiple iterations. We also evaluate the impact of feeding both the generated explanation and the quality rating score from the most recent iteration, as well as from all iterations, into GPT-4, which do not demonstrate a significant difference. These results demonstrate that ILearner-LLM, particularly with instruction fine-tuning, is more effective in generating student-written explanation compared to existing models.
- We find that the evaluation models which have been instruction fine-tuned on explanations rated by students demonstrate lower MSE scores compared to models that have not been fine-tuned. The model that has been fine-tuned on a more diverse range of subjects and additional data (Merged) achieves even lower MSE scores. The (Merged) training set indicates that the training sets from all subjects have been merged together.

2 RELATED WORK

Artificial General Intelligence (AGI) aims to enable machines to understand, learn, and apply knowledge as broadly as humans do (Goertzel, 2014). Making machines think and act in more human-like manner is crucial for the AGI development (Goertzel et al., 2012; Lake et al., 2017; Ouyang et al., 2022). Prompting is a key method for enabling complex human-AI interactions, bridging the gap from AGI concepts to practical uses by fostering adaptability and understanding in line with AGI goals (Hao et al., 2023; Madaan et al., 2024). Chain-of-Thought (CoT) prompting (Wei et al., 2022b) has been introduced to not only generate answers but also the intermediate steps. ITER-RETGEN (Shao et al., 2023) integrates CoT prompting with a retriever to iteratively decompose complex queries, enhancing performance on multi-hop question answering tasks. The LLM-Augmenter (Peng et al., 2023) uses an agent to interact with external knowledge, helping large language models improve open-domain question answering and reduce hallucinations. Iterative prompting has been developed to refine translation results from large language models and decrease ”translationese” (Chen et al., 2023).

In the domain of automated explanation generation and question quality evaluation using deep learning, research remains sparse. BERT-based model has been employed to assess the convincingness of learner-generated explanations, a key criterion for quality (Bhatnagar et al., 2020). However, this approach is designed to evaluate the convincingness of explanations for peers, which requires humans to label data. It does not directly evaluate a single explanation. Recently, Transformer model has been trained with contrastive learning to evaluate question quality, incorporating various elements such as question context and distractors (Ni et al., 2022). Despite its merits, this method necessitates manual feature engineering, which includes explicity-defined features as the model input such as readability, clarity, length of distractors and number of distractors. Contrary to the previous work, explanation evaluation in ILearner-LLM is performed using LLMs fine-tuned to predict the quality rating given an explanation and a question, without relying on explicitly defined features and we did not utilise a self-reward model as described in (Yuan et al., 2024), nor did we apply the rationale generation method outlined in (Hsieh et al., 2023). Additionally, our goal of incorporating explanation evaluation is to iteratively improve the quality of generated explanations by feeding the quality ratings back into the explanation generation model.

3 PROBLEM FORMULATION

In this section, we formally define the *multiple-choice question explanation generation* and *evaluation* tasks. When authoring an MCQ in learnersourcing systems like PeerWise (Denny et al., 2008), a student needs to specify seven components: a question stem, a correct answer, (up to) four distractors, and a paragraph that explains the idea and rationale behind the question. The question is then submitted to an online repository of MCQs accessible by the class. After answering a question, a student may leave a holistic quality rating (from $0, 1, \ldots, 5$) by considering the “*language, quality of distractors, quality of explanation, and relevance to the course*” as suggested by the PeerWise platform (Denny et al., 2008).

Definition 1. *(Multiple-Choice Questions (MCQs)) MCQs are a set of questions, $\{M_1, M_2, \ldots, M_n\}$, collected from a course, where each M_i consists of a stem S_i, a correct answer A_i, distractors $D_{i,j}$ where $j \in \{1, 2, 3, 4\}$, explanation E_i, and is assigned a rating r_i.*

Task 1. *(MCQ explanation generation) Multiple-Choice Question (MCQ) explanation generation aims to construct a model, G, which takes the question stem S_i, the correct answer A_i, and distractors $D_{i,j}$ as inputs, and produces a generated explanation E_i as the output.*

Task 2. *(MCQ explanation evaluation) The goal of Multiple-Choice Question (MCQ) explanation evaluation is to build a model, G, which takes as input the question stem S_i, the correct answer A_i, distractors $D_{i,j}$, and the generated explanation E_i, and outputs a quality rating r_i for the MCQ.*

4 METHOD

4.1 ITERATIVE MCQ EXPLANATION ENHANCEMENT

The architecture of our system is illustrated in Figure 2. The **MCQ Explanation Generation Module** is implemented through instruction fine-tuning to automatically generate explanations for

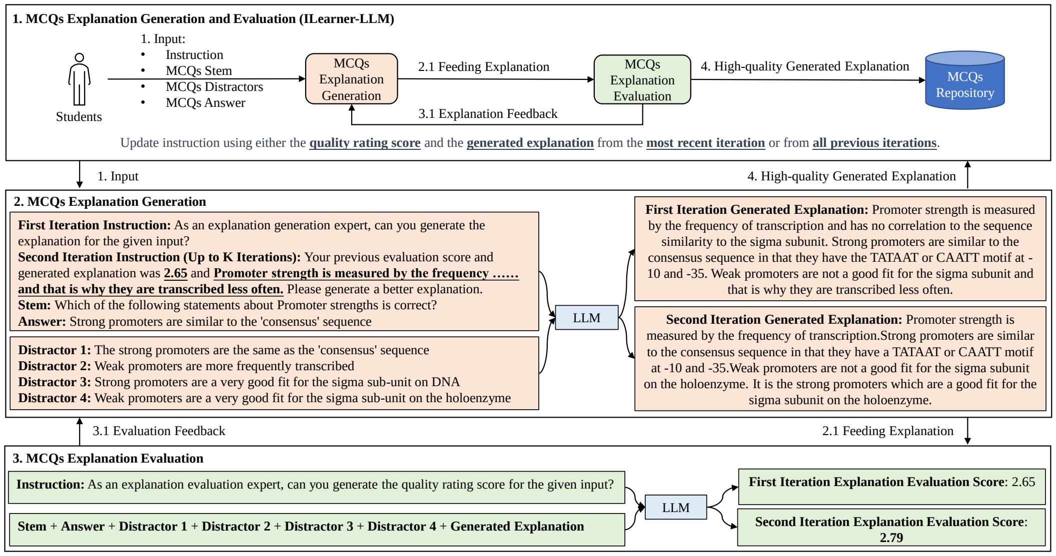

Figure 2: Architecture of the iterative enhancement framework "ILearner-LLM" using large language models for multiple-choice question explanation generation and evaluation.

MCQs. The generated explanations are then provided as inputs to the **MCQ Explanation Evaluation Module**. This module is implemented through instruction fine-tuning, enabling it to automatically assess the quality of the generated explanations. The generation module and evaluation module will interact for up to K iteration steps, where K is a hyperparameter. The evaluation score from the MCQ Explanation Evaluation Module and the generated explanation from the most recent iteration will be fed back to the MCQ Explanation Generation Module, which in turn prompts the model to generate a new explanation. This iterative process continues until it reaches the predefined number of iterations, K. Our ILearner-LLM framework allows the inclusion of either just the generated explanation and rating score from the most recent iteration, for large language models that cannot support long sequence input, or the generated explanation and rating score from all previous iterations, for those models that can support long sequence input. The pseudocode for our ILearner-LLM MCQ Explanation Generation and Evaluation framework is shown in Algorithm 1.

4.2 MCQ Explanation Generation

As depicted in Figure 2, we conduct instruction fine-tuning to train a model for generating explanation of MCQs, and then use instruction prompting with the well-trained model to generate these MCQ explanations. Instruction Fine-Tuning and Instruction Prompting adapt a pretrained model to follow specific input instructions more accurately. The difference lies in the fact that instruction fine-tuning involves additional training with examples that pair these instructions with desired outputs, thereby enhancing the model's task-specific performance (Mishra et al., 2021; Wei et al., 2021). In contrast, instruction prompting does not explicitly train the model; instead, it uses the instructions as part of the prompt for a pre-trained model. The instructions utilised for generating explanations and conducting evaluations are delineated in the system architecture, as depicted in Figure 2. The model inputs include the instruction, question stem, correct answer, and distractors. The model outputs the explanation for the MCQs. During data preprocessing of the five PeerWise datasets (Sydney Biology Subject, Cardiff Biology Subject, Auckland Law Subject, UK Medical Year 1 Subject, and UK Medical Year 2 Subject), we retain only the MCQs with quality rating score of 3 or higher and explanations that are longer than 10 words. This step is undertaken to ensure that only high-quality MCQs are included in the training set, aiding us in building an MCQ explanation generator capable of producing high-quality explanations.

The instruction of the initial iteration is formalised as "As an explanation generation expert, can you generate an explanation for the given input?". The further iteration instruction (up to K iterations) is formalised as "Your previous evaluation score and generation explanation

Algorithm 1 MCQ Explanation Generation and Evaluation

Require: pre-defined iteration step K, initial iteration_step = 0, multiple-choice questions (MCQs) $\{M_1, M_2, \ldots, M_n\}$, question stem S_i, a correct answer A_i, distractors $D_{i,j}$ where $j \in \{1, 2, 3, 4\}$, explanation E_i, large language model (LLM), batch_size bs, learning_rate lr, history = []

1. MCQ explanation generation instruction fine-tuning
for instruction, S_i, A_i, $D_{i,j}$ from MCQs **do**
  LLM, Loss = **next_token_prediction**(LLM, instruction, S_i, A_i, $D_{i,j}$)
end for
2. MCQ explanation evaluation instruction fine-tuning
for instruction, S_i, A_i, $D_{i,j}$, E_i from MCQs **do**
  LLM, Loss = **next_token_prediction**(LLM, instruction, S_i, A_i, $D_{i,j}$, E_i)
end for
3. Iterative MCQ explanation enhancement
while iteration_step $< K$ **do**
  reg_explanation = **explanation_generator**(instruction, S_i, A_i, $D_{i,j}$)
  rating_score = **explanation_evaluator**(instruction, S_i, A_i, $D_{i,j}$, E_i)
  if only use generated explanation and rating score from the **most recent iteration then**
    instruction += reg_explanation + rating_score + "Please generate a better explanation."
  else if use the generated explanation and rating score from the **all previous iterations then**
    history.append(reg_explanation,rating_score)
    instruction = instruction.join(history) + "Please generate a better explanation."
  end if
  iteration_step = iteration_step + 1
end while

was the most recent iteration explanation score and the most recent iteration generated explanation. Please generate a better explanation".

4.3 MCQ Explanation Evaluation

Similar to the module above, we employ instruction fine-tuning to train a large language model to evaluate the generated explanations. In the absence of quality rating scores for explanations, we trained an evaluation model for MCQ explanations using the quality rating scores derived from merging five PeerWise MCQ training sets. The model's input comprises the instruction, question stem, correct answer, distractors, and the explanation. The model's output is the quality rating score for the MCQs. The instruction used in the MCQ Explanation Evaluation Module is, "As an explanation evaluation expert, can you generate the quality rating score for the given input?".

Whenever the MCQ Explanation Evaluation Module predicts a quality rating score, the MCQ Explanation Generation Module is prompted to regenerate the explanation. This regenerated explanation then replaces the one from the previous iteration. The new explanation, along with other inputs, is subsequently fed back into the MCQ Explanation Evaluation Model for re-evaluation. This cycle continues until the number of iteration steps surpasses the predefined K.

5 Experiments

5.1 Experiment Setup

Datasets We conducted our experiment on five learnersourced multiple-choice questions datasets Sydney Biology subject, Cardiff Biology subject, Auckland Law subject, UK Medical Year 1 (2015-2021), and UK Medical Year 2 (2015-2021) from PeerWise platform. We use those five datasets because they contain the largest initial pool of questions and the highest-rated questions. To ensure reliability, only questions that receive at least 10 ratings are included. See Table 2 for the final dataset details. The average explanation length corresponds to the number of words per sentence.

Table 1: In an experiment, we evaluated two models, fine-tuned LLaMA2-13B (Merged) and GPT-4, for generating MCQ explanations. The evaluation used the fine-tuned LLaMA2-13B (Merged) model. The "ILearner-LLM All History" model incorporates explanations and scores from all previous iterations, whereas the ILearner-LLM framework model uses only the most recent explanation and score.

Models	# Iteration Step	Avg Quality Rating Score	Avg BLEU Score	Avg BERT Score
Sydney Biology Subject				
LLaMA2-13B Merged	1	2.84	34.34	61.62
LLaMA2-13B Merged ILearner-LLM	1.36	**2.87**	**36.01**	**62.09**
GPT-4	1	3.02	34.24	63.72
GPT-4 ILearner-LLM	1.28	**3.05**	**34.98**	**63.82**
GPT-4 ILearner-LLM All History	1.16	**3.04**	**35.36**	**63.77**
Cardiff Biology Subject				
LLaMA2-13B Merged	1	3.07	25.59	58.60
LLaMA2-13B Merged ILearner-LLM	1.36	**3.09**	**28.74**	**59.05**
GPT-4	1	3.18	29.08	58.72
GPT-4 ILearner-LLM	1.24	**3.20**	**29.75**	**58.74**
GPT-4 ILearner-LLM All History	1.19	**3.20**	**30.22**	**58.76**
Auckland Law Subject				
LLaMA2-13B Merged	1	4.11	27.82	58.01
LLaMA2-13B Merged ILearner-LLM	1.60	**4.15**	**31.35**	**59.84**
GPT-4	1	4.22	24.31	57.19
GPT-4 ILearner-LLM	1.05	4.22	**24.32**	**57.20**
GPT-4 ILearner-LLM All History	1.02	**4.24**	**24.32**	**57.21**
UK Medical Year 1 Subject				
LLaMA2-13B Merged	1	3.07	27.60	58.45
LLaMA2-13B Merged ILearner-LLM	1.46	**3.09**	**30.98**	**59.28**
GPT-4	1	3.20	28.29	59.47
GPT-4 ILearner-LLM	1.18	**3.21**	**28.56**	**59.54**
GPT-4 ILearner-LLM All History	1.07	3.20	**28.55**	**59.52**
UK Medical Year 2 Subject				
LLaMA2-13B Merged	1	3.05	23.89	56.82
LLaMA2-13B Merged ILearner-LLM	1.52	**3.07**	**26.36**	**57.50**
GPT-4	1	3.15	30.67	58.17
GPT-4 ILearner-LLM	1.10	3.15	**31.10**	**58.20**
GPT-4 ILearner-LLM All History	1.17	3.15	**31.25**	**58.23**

Table 2: Details on the PeerWise datasets used for conducting this experiment.

Subject	Sydney Biology	Cardiff Biology	Auckland Law
# MCQs	2311	6955	3449
# Ratings	57585	581937	65645
# Ratings/MCQ	24.91	83.67	19.03
Avg exp length	108.82	75.09	48.13
Subject	UK Medical Year 1	UK Medical Year 2	
# MCQs	3991	2789	
# Ratings	305067	271524	
# Ratings/MCQ	76.43	97.35	
Avg exp length	68.94	83.38	

Models We select the large language models LLaMA2-13B (Touvron et al., 2023) and GPT-4 (OpenAI, 2023b) as the backbone models for conducting the main experiments, which include instruction fine-tuning and prompting for the generation and evaluation of multiple-choice question (MCQ) explanations. Additionally, we chose Vicuna-13B (Chiang et al., 2023) and GPT-3.5 (OpenAI, 2023a) as baseline models.

Data Preprocessing For the MCQ Explanation Generation module and the MCQ Explanation Evaluation module, we employ different data preprocessing strategies. To train the Explanation Generator, which aids in generating high-quality explanations, we filtered out questions that met any of the following criteria: a quality rating score lower than 3, an explanation length less than 10, or the inclusion of an image in the question stem. We restrict the length of the explanations because we find that many are incomplete, and most of these explanations are less than 10 words long. For training the Explanation Evaluator, we retained both high-rating and low-rating questions.

Table 3: We compared the performance of fine-tuned and non-fine-tuned Vicuna-13B, fine-tuned LLaMA2-13B, and GPT-4 on 100 MCQ explanation test cases from Biology, Law, and Medical subjects in Sydney, Cardiff, Auckland, and the UK.

Models → Metrics ↓	Vicuna-13B	Fine-tuned Vicuna-13B	Fine-tuned LLaMA2-13B	Fine-tuned LLaMA2-13B Merged	GPT-3.5	GPT-4
			Sydney Biology Subject			
Avg BLEU Score	8.59	33.91	**34.80**	34.34	30.25	34.24
Avg BERT Score	20.17	63.33	62.26	61.62	63.56	**63.72**
			Cardiff Biology Subject			
Avg BLEU Score	3.36	15.33	25.37	25.59	25.65	**29.08**
Avg BERT Score	8.76	51.72	56.85	58.60	57.69	**58.72**
			Auckland Law Subject			
Avg BLEU Score	3.09	9.36	26.39	**27.82**	22.16	24.31
Avg BERT Score	7.99	45.38	57.07	**58.01**	57.11	57.19
			UK Medical Year 1 Subject			
Avg BLEU Score	1.92	15.09	26.17	27.60	25.44	**28.29**
Avg BERT Score	6.22	52.06	57.23	58.45	58.44	**59.47**
			UK Medical Year 2 Subject			
Avg BLEU Score	4.23	17.72	24.76	23.89	26.61	**30.67**
Avg BERT Score	12.47	51.62	55.91	56.82	57.15	**58.17**

Table 4: Comparative analysis of iterative enhancement framework performance: number of iterations required for optimal quality rating score, BLEU, and BERT Scores against student-written ground truth.

Iteration Steps → Models ↓	1	2	3	4	5	6
			Sydney Biology Subject			
LLaMA2-13B Merged ILearner-LLM	80	10	7	1	1	1
GPT-4 ILearner-LLM	89	3	3	2	2	1
GPT-4 ILearner-LLM All History	91	6	1	1	0	1
			Cardiff Biology Subject			
LLaMA2-13B Merged ILearner-LLM	84	5	6	3	0	2
GPT-4 ILearner-LLM	93	1	1	1	2	2
GPT-4 ILearner-LLM All History	89	6	3	1	1	0
			Auckland Law Subject			
LLaMA2-13B Merged ILearner-LLM	68	16	11	1	1	3
GPT-4 ILearner-LLM	99	0	0	0	0	1
GPT-4 ILearner-LLM All History	99	0	1	0	0	0
			UK Medical Year 1 Subject			
LLaMA2-13B Merged ILearner-LLM	82	7	4	1	2	4
GPT-4 ILearner-LLM	94	0	2	3	0	1
GPT-4 ILearner-LLM All History	96	2	1	1	0	0
			UK Medical Year 2 Subject			
LLaMA2-13B Merged ILearner-LLM	82	3	4	6	2	3
GPT-4 ILearner-LLM	96	2	0	1	0	1
GPT-4 ILearner-LLM All History	94	1	2	1	1	1

Settings We conducted all the instruction fine-tuning for Vicuna-13B and LLaMA2-13B MCQ explanation generation and evaluation experiments on 8 NVIDIA A100 GPUs with 80G GPU memory. We trained our model for 5 epochs, using a batch size of 1 and a maximum sequence length of 512. We set the learning rate to 2e-05 and the warmup ratio to 0.03. To leverage the power of multi-GPUs, we utilised the torchrun tool for training. The source code is available [1].

5.2 Iterative MCQ Explanation Enhancement

We iterate the process of explanation generation and evaluation K steps, with each interaction comprising one instance of explanation generation and evaluation. We recorded the score for each evaluation, and the similarity between the generated explanation and the original explanation written by the student. We then computed the number of iterations required to improve the evaluation score, the generated explanation, and the similarity to the original student-written explanation. In our

[1] `https://github.com/Strong-AI-Lab/Explanation-Generation`

experiment, we set a number of iteration steps, K=5, for halting iterations: the model generates explanations iteratively over K iterations. In each iteration, the model feeds the previously generated explanation and the quality rating score into the instruction and prompts the MCQ explanation generation module. The specific results are shown in Table 1. Utilising our iterative enhancement framework ILearner-LLM, we find that by applying our ILearner-LLM framework, which automatically generates explanations iteratively, it requires approximately 0.46 additional iterations for the fine-tuned LLaMA2-13B and 0.17 additional iterations for GPT-4 when generating an explanation using only the most recent iteration's generated explanation and a quality rating score. For GPT-4, when generating explanations that include previously generated content from all history along with a quality rating score, it requires only 0.12 additional iterations. These additional iterations are necessary to produce explanations that surpass the original ones in terms of question quality rating, BLEU score, and BERT score. ILearner-LLM, which incorporates LLaMA2-13B and GPT-4 models, shows notable improvements in generating higher quality explanations that are both literarily and semantically closer to those written by students across the five PeerWise datasets. It outperforms a fine-tuned LLaMA2-13B by 2.84 in BLEU score and surpasses GPT-4 by 0.42. Additionally, it shows a 0.85 and 0.04 increase in BERT score over LLaMA2-13B and GPT-4, respectively. Furthermore, since GPT-4 supports 8K sequence input, we applied ILearner-LLM and iteratively fed the generated explanation and quality rating score from all the previous history into GPT-4, which achieved a 0.62 and 0.04 improvement over GPT-4. These results demonstrate that ILearner-LLM, with instruction fine-tuning, is more effective in generating higher-quality explanations that are literarily and semantically closer to the explanations written by students compared to existing models.

We conducted an analysis of our proposed iterative enhancement framework using different numbers of iterations as shown in Table 4. We recorded the number of iterations required to find an explanation with the highest rating score, as well as the highest BLEU and BERT scores, compared to the ground truth of student-written explanations. ILearner-LLM can assist the fine-tuned LLaMA2-13B Merged in generating better explanations over a greater number of iteration steps compared to GPT-4. GPT-4 produces high-quality explanations without task-specific fine-tuning, unlike LLaMA2-13B, which improved in MCQ explanation generation through instruction fine-tuning.

5.3 MCQ Explanation Generation

We employed instruction fine-tuning on LLaMA2-13B across all subjects to train the explanation generator. For comparison, we used four baseline models: Vicuna-13B, Vicuna-13B fine-tuned on each subject, LLaMA2-13B fine-tuned on each subject, and both GPT-3.5 and GPT-4. Considering the cost of calling GPT-4 API, we randomly selected 100 samples from the whole test set. We used BLEU (Papineni et al., 2002) and BERT scores (Zhang et al., 2019) to evaluate the literality and semantic similarity of the generated explanations to the ground truth explanations (student-authored explanation), respectively. In our experiments, GPT-4 consistently outperformed other models, achieving the highest BLEU and BERT scores across the majority of datasets, as delineated in Table 3. We further investigated the impact of instruction fine-tuning on large language models such as Vicuna-13B. Remarkably, this fine-tuning led to a significant improvement in both BLEU and BERT scores in comparison to using Vicuna-13B without any modifications. We extended this fine-tuning approach to another large model, LLaMA2-13B, and observed even more encouraging results. Specifically, instruction fine-tuned LLaMA2-13B surpassed the performance of its Vicuna-13B counterpart and even exceeded GPT-4 in certain tasks. Notably, it achieved higher scores in the Sydney Biology and Auckland Law subjects and outperformed GPT-3.5 in four out of five datasets, with the exception being the UK Medical Year 2 subject. For LLaMA2-13B, two fine-tuning strategies were employed. In the first, denoted as "Fine-tuned LLaMA2-13B," we applied instruction fine-tuning individually to each task. In the second approach, labeled as "Fine-tuned LLaMA2-13B Merged," we amalgamated the training sets from the five tasks for a unified instruction fine-tuning process. Our findings suggest that both instruction fine-tuned Vicuna-13B and LLaMA2-13B effectively learned to emulate the characteristics inherent in student-generated explanations. We find that the "Fine-tuned LLaMA2-13B Merged" model performs the best on the Auckland Law Subject because the merging effectively introduced four biology/medicine datasets into the training set, which helped with more diverse topics and increased the amount of training data.

5.4 MCQ EXPLANATION EVALUATION

As Table 5 shows, we compared a fine-tuned LLaMA2-13B model with the non-fine-tuned LLaMA2-13B model and GPT-4 on 100 test cases. These were randomly collected from Sydney and Cardiff Biology subjects, Auckland Law subject, and UK Medical Year 1 and 2 subjects, for the MCQ explanation evaluation task. Since we have the question quality rating labels for each question, we can use these labels to train a question quality rating model. We rely on this model to evaluate the explanation by replacing the explanations in the MCQs. In Table 5, we use Mean Squared Error (MSE) as the metric to measure the distance between the predicted rating score and the student-labeled ground truth rating score. A lower MSE score means that the predicted rating score is closer to the ground truth rating score.

As demonstrated in Table 5, the Fine-tuned LLaMA2-13B and Fine-tuned LLaMA2-13B Merged models significantly outperform the two baseline models in terms of MSE. This suggests that these fine-tuned models can capture the underlying distribution of evaluation scores generated by students. We find that “Fine-tuned LLaMA2-13B Merged” shows a lower MSE score compared to “Fine-tuned LLaMA2-13B,” demonstrating that more diverse subject training data can help the model achieve better performance in predicting the quality rating score. We observed that both LLaMA2-13B, without task-specific instruction fine-tuning, and GPT-4 underperform in the evaluation of explanations for MCQs. Specifically, these models tend to over-estimate quality, frequently assigning scores greater than 4 on a scale where 5 is the maximum. This inflation of scores may be an artifact of the Reinforcement Learning from Human Feedback (RLHF) approach, predisposing the models to offer overly positive evaluations. Such biases could have significant implications in educational contexts where these models are deployed for automated student feedback. For MCQ explanation evaluation task, using instruction fine-tuning on specific datasets improves performance by more closely matching the target distribution.

Table 5: We compared the fine-tuned LLaMA2-13B with the non-fine-tuned LLaMA2-13B and GPT-4 on 100 test cases for MCQ explanation evaluation.

Models $\rightarrow$ Metrics $\downarrow$	LLaMA2-13B	Fine-tuned LLaMA2-13B	Fine-tuned LLaMA2-13B Merged	GPT-4
Sydney Biology Subject				
MSE	1.21	0.43	**0.22**	3.95
Cardiff Biology Subject				
MSE	0.58	0.10	**0.09**	3.28
Auckland Law Subject				
MSE	2.86	**0.11**	0.12	0.42
UK Medical Year 1 Subject				
MSE	0.84	0.19	**0.15**	3.23
UK Medical Year 2 Subject				
MSE	1.71	0.10	**0.09**	3.02

6 CONCLUSIONS AND FUTURE WORK

In summary, this study presents a novel iterative enhancement framework “ILearner-LLM” that utilises large language models for the generation and assessment of explanations for learner-sourced multiple-choice questions. Experimental findings indicate that our iterative enhancement methodology enables advanced language models, such as LLaMA2-13B and GPT-4, to produce explanations with superior BLEU and BERT scores when compared to merely fine-tuned LLaMA2-13B and GPT-4. Future research endeavors will focus on expanding the dataset, fine-tuning the models across a diverse range of academic disciplines and educational levels, integrating the framework into a live learner-sourcing platform to examine learner engagement with the generated explanations, and exploring a meta-learning approach for continual refinement based on user feedback.

REFERENCES

Sameer Bhatnagar, Amal Zouaq, Michel C Desmarais, and Elizabeth Charles. Learnersourcing quality assessment of explanations for peer instruction. In *Addressing Global Challenges and Quality Education: 15th European Conference on Technology Enhanced Learning, EC-TEL 2020, Heidelberg, Germany, September 14–18, 2020, Proceedings 15*, pp. 144–157. Springer, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 323–325. IEEE, 2023.

Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. Peerwise: students sharing their multiple choice questions. In *Proceedings of the fourth international workshop on computing education research*, pp. 51–58, 2008.

Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

Ben Goertzel, Matt Iklé, and Jared Wigmore. The architecture of human-like general intelligence. In *Theoretical foundations of artificial general intelligence*, pp. 123–144. Springer, 2012.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL `https://aclanthology.org/2023.emnlp-main.507`.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL `https://aclanthology.org/2023.findings-acl.507`.

Yuchao Jiang, Daniel Schlagwein, and Boualem Benatallah. A review on crowdsourcing for education: State of the art of literature and practice. *PACIS*, pp. 180, 2018.

Hassan Khosravi, Kirsty Kitto, and Joseph Jay Williams. Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *arXiv preprint arXiv:1910.05522*, 2019.

Hassan Khosravi, Paul Denny, Steven Moore, and John Stamper. Learnersourcing in the age of ai: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, 5:100151, 2023a. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2023.100151. URL `https://www.sciencedirect.com/science/article/pii/S2666920X23000309`.

Hassan Khosravi, Paul Denny, Steven Moore, and John Stamper. Learnersourcing in the age of ai: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, pp. 100151, 2023b.

Juho Kim et al. *Learnersourcing: improving learning with collective learner activity*. PhD thesis, Massachusetts Institute of Technology, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

Lin Ni, Qiming Bao, Xiaoxuan Li, Qianqian Qi, Paul Denny, Jim Warren, Michael Witbrock, and Jiamou Liu. Deepqr: Neural-based quality ratings for learnersourced multiple-choice questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12826–12834, 2022.

OpenAI. Chatgpt: Optimizing language models for dialogue., 2023a. URL `https://openai.com/blog/chatgpt`.

OpenAI. Gpt-4 technical report, 2023b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9248–9274, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.620. URL `https://aclanthology.org/2023.findings-emnlp.620`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.

Hatice Yildiz Durak. Conversational agent-based guidance: examining the effect of chatbot usage frequency and satisfaction on visual design self-efficacy, engagement, satisfaction, and learner autonomy. *Education and Information Technologies*, 28(1):471–488, 2023.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.