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], Tim 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, University of Auckland
[4]School of Computer Science and Engineering, Sun Yat-Sen University
[5]School of Life and Environmental Sciences, University of Sydney
{qbao775,ypen260,ggen187}@aucklanduni.ac.nz, {juho.leinonen,p.denny,m.witbrock,jiamou.liu}@auckland.ac.nz,
zhongwj25@mail2.sysu.edu.cn, alice.huang@sydney.edu.au

Abstract

Large language models (LLMs) have demonstrated strong capabilities in language understanding and generation, and their potential in educational contexts is increasingly being explored. One promising area is learnersourcing, where students engage in creating their own educational content, such as multiple-choice questions. A critical step in this process is generating effective explanations for the solutions to these questions, as such explanations aid in peer understanding and promote deeper conceptual learning. However, students often find it difficult to craft high-quality explanations due to limited understanding or gaps in their subject knowledge. To support this task, we introduce "ILearner-LLM," a framework that uses iterative enhancement with LLMs to improve generated explanations. The framework combines an explanation generation model and an explanation evaluation model fine-tuned using student preferences for quality, where feedback from the evaluation model is fed back into the generation model to refine the output. Our experiments with LLaMA2-13B and GPT-4 using five large datasets from the PeerWise MCQ platform show that ILearner-LLM produces explanations of higher quality that closely align with those written by students. Our findings represent a promising approach for enriching the learnersourcing experience for students and for leveraging the capabilities of large language models for educational applications.

Introduction

Given the remarkable performance of large language models (LLMs) in understanding and generating natural language (Wei et al. 2022a; Brown et al. 2020), they appear to offer great potential across many applications within education. Learnersourcing, a pedagogical approach that distributes the task of generating learning content among students, leverages their collective intelligence to enhance the learning experience (Jiang, Schlagwein, and Benatallah 2018; Khosravi et al. 2023a; Kim et al. 2015). On platforms such as PeerWise (Denny et al. 2008) and RiPPLE (Khosravi, Kitto, and Williams 2019), learnersourcing often involves students creating multiple-choice questions and providing corresponding explanations. However, crafting high-quality explanations requires a deep understanding of the underlying concepts, a challenge that students may struggle with. Additionally, because it is often not mandatory for students to include explanations when generating questions, they may choose to omit them altogether. The automatic generation and evaluation of high-quality explanations can serve as a scaffold for learners, offering tailored support that fosters deeper understanding and greater independence, particularly in the context of learnersourcing tasks where students collaboratively create and refine educational content.

The main challenges in automatic explanation generation within this context are driven by several key factors. First, one hurdle is simulating the way students write explanations and generating text that aligns with what students value in a well-crafted explanation. This challenge goes beyond replicating content; it requires capturing the characteristics of how students express their understanding. Second, the scarcity of high-quality datasets that include explanations poses another major challenge. Since writing explanations is often not mandatory for students in learnersourcing platforms, there is a limited amount of annotated data available for training models. This lack of data makes it difficult to achieve high performance in automatic explanation generation.

In this work we aim to use LLMs 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, interaction with such models and observing their outputs can help to promote learner autonomy (Yildiz Durak 2023). Thirdly, the use 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 student preferences by using data on how

[*] With help from the AAAI Publications Committee.

students rate questions (Ni et al. 2022). Despite these benefits, employing LLMs in a learnersourcing context also presents challenges. In particular, there is limited access to high-quality student-written explanations which makes it difficult to fine-tune models to generate explanations that are both linguistically and semantically similar to those written by students (Khosravi et al. 2023a).

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 popular learnersourcing platform. We use 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 then evaluates the quality of the generated explanation 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 linguistically and semantically similar 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 using the models without the framework in generating explanations that closely resemble those written by students, as evidenced by BLEU and BERT scores.
- We find that ILearner-LLM can help instruction fine-tuned LLaMA2-13B achieve greater improvement compared to applying ILearner-LLM to GPT-4. This is mainly because LLaMA2-13B, which was fine-tuned using the same instruction set employed in iterative prompting, better aligns with the instructional framework, enabling the model to learn and perform more effectively through multiple iterations. We also evaluated the impact of feeding both the generated explanation and the quality rating score from either the most recent iteration or all iterations into GPT-4. Our results show no significant difference between these approaches. These findings suggest that ILearner-LLM, particularly when combined with instruction fine-tuning, is more effective in generating student-like explanations 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.

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 a more human-like manner is crucial for the development of AGI (Goertzel, Iklé, and Wigmore 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. Building a natural language generation system with explanations is an ongoing goal (Reiter and Dale 1997). Template or rule-based methods (Holmes-Higgin 1994), knowledge base methods (Wang and McKeown 2010), and Long Short-Term Memory (LSTM) networks (Costa et al. 2018) have been used for automatic explanation generation to build explainable recommender systems. With the development of pre-trained transformer-based language models and stronger representation learning for understanding context, BERT-based models have 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, a 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. We also did not rely on 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.

Problem Formulation

In this section, we formally define the *multiple-choice question explanation generation* and *evaluation* tasks. When authoring an MCQ in a learnersourcing system 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 (on a 6-point scale 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.*

Method

Iterative MCQ Explanation Enhancement

The architecture of our system is illustrated in Figure 1. The **MCQ Explanation Generation Module** is implemented through instruction fine-tuning to automatically generate explanations for 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 iterations, where K is a hyper-parameter. 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, or the generated explanation and rating score from all previous iterations, the latter of which is more appropriate for models that can support long sequence input. The pseudocode for our ILearner-LLM MCQ Explanation Generation and Evaluation framework is shown in Algorithm 1.

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
```

MCQ Explanation Generation

As depicted in Figure 1, we conduct instruction fine-tuning to train a model for generating explanations for 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 1. The model inputs include the instruction, question stem, correct answer, and distractors. The model outputs the explanation for the MCQ. 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 retained only MCQs with quality rating scores of 3 or higher and explanations that are longer than 10 words. This step is undertaken to provide some simple quality control for the MCQs in the training set.

The instruction of the initial iteration is formalised

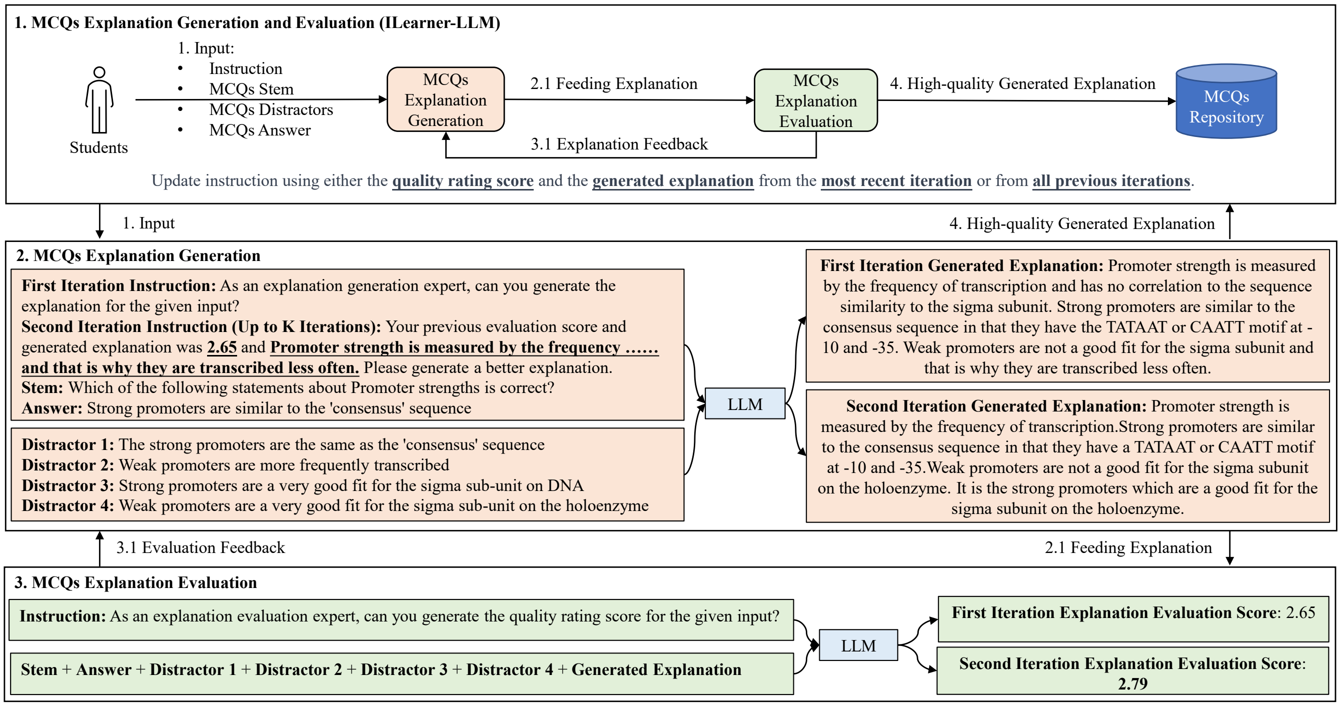
1. MCQs Explanation Generation and Evaluation (ILearner-LLM)
1. Input:
Instruction
MCQs Stem
MCQs Distractors
MCQs Answer
Students
MCQs Explanation Generation
2.1 Feeding Explanation
3.1 Explanation Feedback
MCQs Explanation Evaluation
4. High-quality Generated Explanation
MCQs Repository
Update instruction using either the quality rating score and the generated explanation from the most recent iteration or from all previous iterations.
1. Input
4. High-quality Generated Explanation
2. MCQs Explanation Generation
First Iteration Instruction: As an explanation generation expert, can you generate the explanation for the given input?
Second Iteration Instruction (Up to K Iterations): Your previous evaluation score and generated explanation was 2.65 and Promoter strength is measured by the frequency …… and that is why they are transcribed less often. Please generate a better explanation.
Stem: Which of the following statements about Promoter strengths is correct?
Answer: Strong promoters are similar to the 'consensus' sequence
Distractor 1: The strong promoters are the same as the 'consensus' sequence
Distractor 2: Weak promoters are more frequently transcribed
Distractor 3: Strong promoters are a very good fit for the sigma sub-unit on DNA
Distractor 4: Weak promoters are a very good fit for the sigma sub-unit on the holoenzyme
LLM
First Iteration Generated Explanation: Promoter strength is measured by the frequency of transcription and has no correlation to the sequence similarity to the sigma subunit. Strong promoters are similar to the consensus sequence in that they have the TATAAT or CAATT motif at -10 and -35. Weak promoters are not a good fit for the sigma subunit and that is why they are transcribed less often.
Second Iteration Generated Explanation: Promoter strength is measured by the frequency of transcription.Strong promoters are similar to the consensus sequence in that they have a TATAAT or CAATT motif at -10 and -35.Weak promoters are not a good fit for the sigma subunit on the holoenzyme. It is the strong promoters which are a good fit for the sigma subunit on the holoenzyme.
3.1 Evaluation Feedback
2.1 Feeding Explanation
3. MCQs Explanation Evaluation
Instruction: As an explanation evaluation expert, can you generate the quality rating score for the given input?
Stem + Answer + Distractor 1 + Distractor 2 + Distractor 3 + Distractor 4 + Generated Explanation
LLM
First Iteration Explanation Evaluation Score: 2.65
Second Iteration Explanation Evaluation Score: 2.79

Figure 1: Architecture of the iterative enhancement framework "ILearner-LLM" using large language models for multiple-choice question explanation generation and evaluation.

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

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.

Experiments

Experiment Setup

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

We conducted our experiment on five learnersourced multiple-choice question datasets, covering three academic subjects: biology, law, and medicine, exported from the PeerWise platform. We selected these datasets because they contain a large number of questions. To improve reliability, only questions that received at least 10 ratings were included. The average explanation length corresponds to the number of words per sentence.

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. We chose Vicuna-13B (Chiang et al. 2023) and GPT-3.5 (OpenAI 2023a) as baseline models.

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	**2.37**	2.87	**38.07**	62.00
GPT-4	1	3.02	34.24	**63.72**
GPT-4 ILearner-LLM	1.63	3.12	35.19	63.45
GPT-4 ILearner-LLM All History	1.70	**3.14**	35.08	63.58
		Cardiff Biology Subject		
LLaMA2-13B Merged	1	3.07	25.59	58.60
LLaMA2-13B Merged ILearner-LLM	**2.08**	3.11	**30.58**	58.27
GPT-4	1	3.18	29.08	58.72
GPT-4 ILearner-LLM	1.84	**3.23**	29.91	58.57
GPT-4 ILearner-LLM All History	1.36	3.21	30.43	**58.77**
		Auckland Law Subject		
LLaMA2-13B Merged	1	4.11	27.82	58.01
LLaMA2-13B Merged ILearner-LLM	**2.23**	4.20	**34.33**	**59.95**
GPT-4	1	4.22	24.31	57.19
GPT-4 ILearner-LLM	1.74	**4.29**	24.09	56.91
GPT-4 ILearner-LLM All History	1.45	**4.29**	24.26	57.11
		UK Medical Year 1 Subject		
LLaMA2-13B Merged	1	3.07	27.60	58.45
LLaMA2-13B Merged ILearner-LLM	**2.18**	3.09	**32.52**	59.06
GPT-4	1	3.20	28.29	**59.47**
GPT-4 ILearner-LLM	1.60	**3.23**	28.65	59.38
GPT-4 ILearner-LLM All History	1.27	3.21	29.10	59.43
		UK Medical Year 2 Subject		
LLaMA2-13B Merged	1	3.05	23.89	56.82
LLaMA2-13B Merged ILearner-LLM	**2.44**	3.06	30.43	56.96
GPT-4	1	3.15	30.67	58.17
GPT-4 ILearner-LLM	1.88	**3.18**	31.63	57.97
GPT-4 ILearner-LLM All History	1.53	**3.18**	**31.83**	**58.21**

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.

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

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

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, resulting in a total of 19,495 questions after filtering out those quality rating below 3, a length of less than 10, the inclusion of an image in the question stem and fewer than 10 ratings. See Table 2 for details on the datasets. We restrict the length of the explanations because we find that many short explanations are incomplete. We trained the explanation evaluator using 27,140 high-rated and low-rated questions across all subjects.

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].

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

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 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.

Iterative MCQ Explanation Enhancement

We iterate the process of explanation generation and evaluation over K steps, with each iteration 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 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.

For each generation, we calculate the normalized average score for the question quality rating, BLEU score, and BERT score across the iteratively generated K explanations. We then select the explanation with the highest normalized average score. The question quality rating score is normalized from a float value to a range between 0 and 1. Our goal is to determine how many iterations are required to surpass the explanation generated in the first iteration. Using our iterative enhancement framework, ILearner-LLM, we found that approximately 2.26 iterations are required for the fine-tuned LLaMA2-13B model, and 1.73 iterations for GPT-4, when generating explanations using only the most recent generated explanation and its quality rating score. For GPT-4, when including all previously generated content along with their quality rating scores, an average of 1.46 iterations are required to produce an explanation that surpasses the original 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 syntactically and semantically closer to those written by students across the five PeerWise datasets. It outperforms a fine-tuned LLaMA2-13B by 5.33 in BLEU score and surpasses GPT-4 by 3.86. Furthermore, since GPT-4 only supports up to 8K tokens in the input, we applied ILearner-LLM iteratively by feeding the generated explanations and quality rating scores in separate inputs from all the previous history into GPT-4, which achieved a 0.82 and 0.05 improvement over GPT-4 on BLEU and BERT score. These results demonstrate that ILearner-LLM, with instruction fine-tuning, is more effective in generating higher-quality explanations that are syntactically 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.

MCQ Explanation Generation

We employed instruction fine-tuning on LLaMA2-13B across all subjects to train an explanation generator, com-

Iteration Steps → Models ↓	1	2	3	4	5	6
Sydney Biology Subject						
LLaMA2-13B Merged ILearner-LLM	38	26	14	11	5	6
GPT-4 ILearner-LLM	61	29	3	2	3	2
GPT-4 ILearner-LLM All History	50	40	4	3	2	1
Cardiff Biology Subject						
LLaMA2-13B Merged ILearner-LLM	36	38	15	5	5	1
GPT-4 ILearner-LLM	63	17	8	3	3	6
GPT-4 ILearner-LLM All History	75	20	1	3	0	1
Auckland Law Subject						
LLaMA2-13B Merged ILearner-LLM	27	44	18	4	4	3
GPT-4 ILearner-LLM	65	18	4	6	5	2
GPT-4 ILearner-LLM All History	72	20	4	1	1	2
UK Medical Year 1 Subject						
LLaMA2-13B Merged ILearner-LLM	37	35	12	8	5	3
GPT-4 ILearner-LLM	74	10	7	4	1	4
GPT-4 ILearner-LLM All History	81	12	6	1	0	0
UK Medical Year 2 Subject						
LLaMA2-13B Merged ILearner-LLM	28	35	15	12	7	3
GPT-4 ILearner-LLM	58	22	9	2	3	6
GPT-4 ILearner-LLM All History	65	24	8	0	2	1

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.

paring it to 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. Given the cost of calling the GPT-4 API, we randomly selected 100 samples from the entire test set. We evaluated the syntactic and semantic similarity of generated explanations to ground truth explanations (student-authored) using BLEU (Papineni et al. 2002) and BERT scores (Zhang et al. 2019), respectively. In our experiments, GPT-4 consistently outperformed other models, achieving the highest BLEU and BERT scores across the majority of datasets, as shown in Table 3. Further investigation revealed that instruction fine-tuning significantly improved both BLEU and BERT scores for Vicuna-13B compared to the unmodified version. Extending this fine-tuning approach to LLaMA2-13B led to even more promising results. Specifically, instruction fine-tuned LLaMA2-13B surpassed Vicuna-13B and even outperformed GPT-4 in certain tasks. 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 of the UK Medical Year 2 subject. Two fine-tuning strategies were applied to LLaMA2-13B: individual fine-tuning for each task ("Fine-tuned LLaMA2-13B") and a merged fine-tuning approach, combining training sets from all five tasks ("Fine-tuned LLaMA2-13B Merged"). The merged approach performed best on the Auckland Law subject, likely because it incorporated four biology/medicine datasets, increasing the diversity of topics and training data. Our findings suggest that both instruction fine-tuned Vicuna-13B and LLaMA2-13B effectively learned to mimic the characteristics of student-generated explanations.

MCQ Explanation Evaluation

Table 5 presents a comparison of fine-tuned and non-fine-tuned LLaMA2-13B models and GPT-4 on 100 randomly selected test cases from Sydney and Cardiff Biology, Auckland Law, and UK Medical Year 1 and 2 subjects for the MCQ explanation evaluation task. Using question quality rating labels, we trained a model to evaluate explanations by replacing the MCQ explanations. Mean Squared Error (MSE) was used as the metric, with a lower score indicating closer alignment to the ground truth. As shown, both fine-tuned LLaMA2-13B models significantly outperform the baselines in terms of MSE, suggesting they better capture the distribution of student-generated ratings. The "Fine-tuned LLaMA2-13B Merged" model achieves better performance than the "Fine-tuned LLaMA2-13B," indicating that incorporating diverse subject data enhances predictive accuracy. In contrast, LLaMA2-13B without task-specific fine-tuning and GPT-4 underperform, often inflating scores, likely due to biases introduced by Reinforcement Learning from Human Feedback (RLHF). These findings highlight the importance of instruction fine-tuning for improving model performance in educational feedback applications.

Models → Metrics ↓	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

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.

Conclusions and Future Work

This study introduces the "ILearner-LLM" framework, which uses large language models to generate and assess explanations for learner-sourced multiple-choice questions. Experiments show that our iterative enhancement approach improves explanation quality, with better BLEU and BERT scores for LLaMA2-13B and GPT-4 compared to fine-tuned versions. To increase explanation diversity, we will explore different temperature hyperparameters in "ILearner-LLM."

References

Bhatnagar, S.; Zouaq, A.; Desmarais, M. C.; and Charles, E. 2020. 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*, 144–157. Springer.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, P.; Guo, Z.; Haddow, B.; and Heafield, K. 2023. Iterative Translation Refinement with Large Language Models. *arXiv preprint arXiv:2306.03856*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Costa, F.; Ouyang, S.; Dolog, P.; and Lawlor, A. 2018. Automatic generation of natural language explanations. In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 1–2.

Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; and Chen, G. 2023. Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–325. IEEE.

Denny, P.; Hamer, J.; Luxton-Reilly, A.; and Purchase, H. 2008. PeerWise: students sharing their multiple choice questions. In *Proceedings of the fourth international workshop on computing education research*, 51–58.

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

Goertzel, B.; Iklé, M.; and Wigmore, J. 2012. The architecture of human-like general intelligence. In *Theoretical foundations of artificial general intelligence*, 123–144. Springer.

Hao, S.; Gu, Y.; Ma, H.; Hong, J.; Wang, Z.; Wang, D.; and Hu, Z. 2023. Reasoning with Language Model is Planning with World Model. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8154–8173. Singapore: Association for Computational Linguistics.

Holmes-Higgin, P. 1994. Text generation—using discourse strategies and focus constraints to generate natural language text by Kathleen R. McKeown, Cambridge University Press, 1992, pp 246,£ 13.95, ISBN 0-521-43802-0. *The Knowledge Engineering Review*, 9(4): 421–422.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-k.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8003–8017. Toronto, Canada: Association for Computational Linguistics.

Jiang, Y.; Schlagwein, D.; and Benatallah, B. 2018. A Review on Crowdsourcing for Education: State of the Art of Literature and Practice. *PACIS*, 180.

Khosravi, H.; Denny, P.; Moore, S.; and Stamper, J. 2023a. Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, 5: 100151.

Khosravi, H.; Denny, P.; Moore, S.; and Stamper, J. 2023b. Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, 100151.

Khosravi, H.; Kitto, K.; and Williams, J. J. 2019. Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *arXiv preprint arXiv:1910.05522*.

Kim, J.; et al. 2015. *Learnersourcing: improving learning with collective learner activity*. Ph.D. thesis, Massachusetts Institute of Technology.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40: e253.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Mishra, S.; Khashabi, D.; Baral, C.; and Hajishirzi, H. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Ni, L.; Bao, Q.; Li, X.; Qi, Q.; Denny, P.; Warren, J.; Witbrock, M.; and Liu, J. 2022. Deepqr: Neural-based quality ratings for learnersourced multiple-choice questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12826–12834.

OpenAI. 2023a. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023b. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peng, B.; Galley, M.; He, P.; Cheng, H.; Xie, Y.; Hu, Y.; Huang, Q.; Liden, L.; Yu, Z.; Chen, W.; et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Reiter, E.; and Dale, R. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1): 57–87.

Shao, Z.; Gong, Y.; Shen, Y.; Huang, M.; Duan, N.; and Chen, W. 2023. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9248–9274. Singapore: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, W. Y.; and McKeown, K. 2010. "Got You!": Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 1146–1154.

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Yildiz Durak, H. 2023. 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.

Yuan, W.; Pang, R. Y.; Cho, K.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.