Exploring Iterative Enhancement for Improving Learnersourced Multiple-Choice Question Explanations with Large Language Models

Speaker: Qiming Bao
Homepage: https://14h034160212.github.io/
Strong AI Lab, NAOInstitute, The University of Auckland, New Zealand
13th November 2024
Pen State University & University of Auckland Online Workshop

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Strong AI Lab

- Strong AI Lab is led by Professor Michael Witbrock, at the intersection of machine learning, reasoning, and natural language understanding, with an additional focus on achieving the best social and civilisational impacts of increasingly powerful AI.

https://www.ai.ac.nz/sail/

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

About Me

- Qiming Bao is a Ph.D. Candidate at the Strong AI Lab, NAOInstitute, University of Auckland, New Zealand, supervised by Professor Michael Witbrock. His research interests include natural language processing and reasoning. He has over three years of research and development experience, and has published several papers in top conferences in the fields of AI/NLP/Reasoning, including **AAAI/EAAI**, **ICLR**, **ACL**, **EACL**, **LLM@IJCAI**, and **IJCLR-NeSy**. His method named **AMR-LDA** (GPT-4 + AMR-LDA Prompt Augmentation) has achieved the **#1** ranking on a one of the most challenged logical reasoning reading comprehension leaderboards (ReClor) up to now, and two of his logical reasoning datasets called PARARULE-Plus and AbductionRules have been collected by LogiTorch, ReasoningNLP, Prompt4ReasoningPapers and OpenAI/Evals. Qiming has given public guest talks at Microsoft Research Asia, Samsung AI Center Cambridge UK, IEEE Vehicular Technology Society, ZJU-NLP Group, Zhejiang University and The University of Melbourne on his main research topic, "Natural Language Processing and Reasoning".

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Exploring Iterative Enhancement for Improving Learnersourced Multiple-Choice Question Explanations with Large Language Models

Authored by: **Qiming Bao[1,2], Juho Leinonen[3], Alex Yuxuan Peng[1], Wanjun Zhong[4], Gaël Gendron[1], Timothy Pistotti[1], Alice Huang[6], Paul Denny[5], Michael Witbrock[1], Jiamou Liu[1]**

[1]Strong AI Lab, NAOInstitute, Waipapa Taumata Rau - The University of Auckland

[2]Xtracta, New Zealand

[3]Aalto University, Finland

[4]School of Computer Science and Engineering, Sun Yat-Sen University, China

[5]School of Computer Science, The University of Auckland, New Zealand

[6]School of Life and Environmental Sciences, The University of Sydney, Australia

AGI@ICLR 2024

https://arxiv.org/abs/2309.10444

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Outline

- Background
- System Architecture
- Experiment Results
- Conclusion and Future Work

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Research Gap

- The main challenges in automatic explanation generation are constrained by several key factors.
- First, simulating the process of students writing explanations and generating text that closely **resembles student-written explanations** is a significant hurdle. This involves not only replicating the content but also capturing the nuances of how students typically articulate their understanding.
- Second, the **scarcity of high-quality datasets** that include explanations poses another major challenge. Since writing explanations is not mandatory for students, there is a limited amount of annotated data available for training models. This scarcity makes it difficult to achieve high performance in automatic explanation generation.

An Example for PeerWise Dataset

- **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

Dataset Description

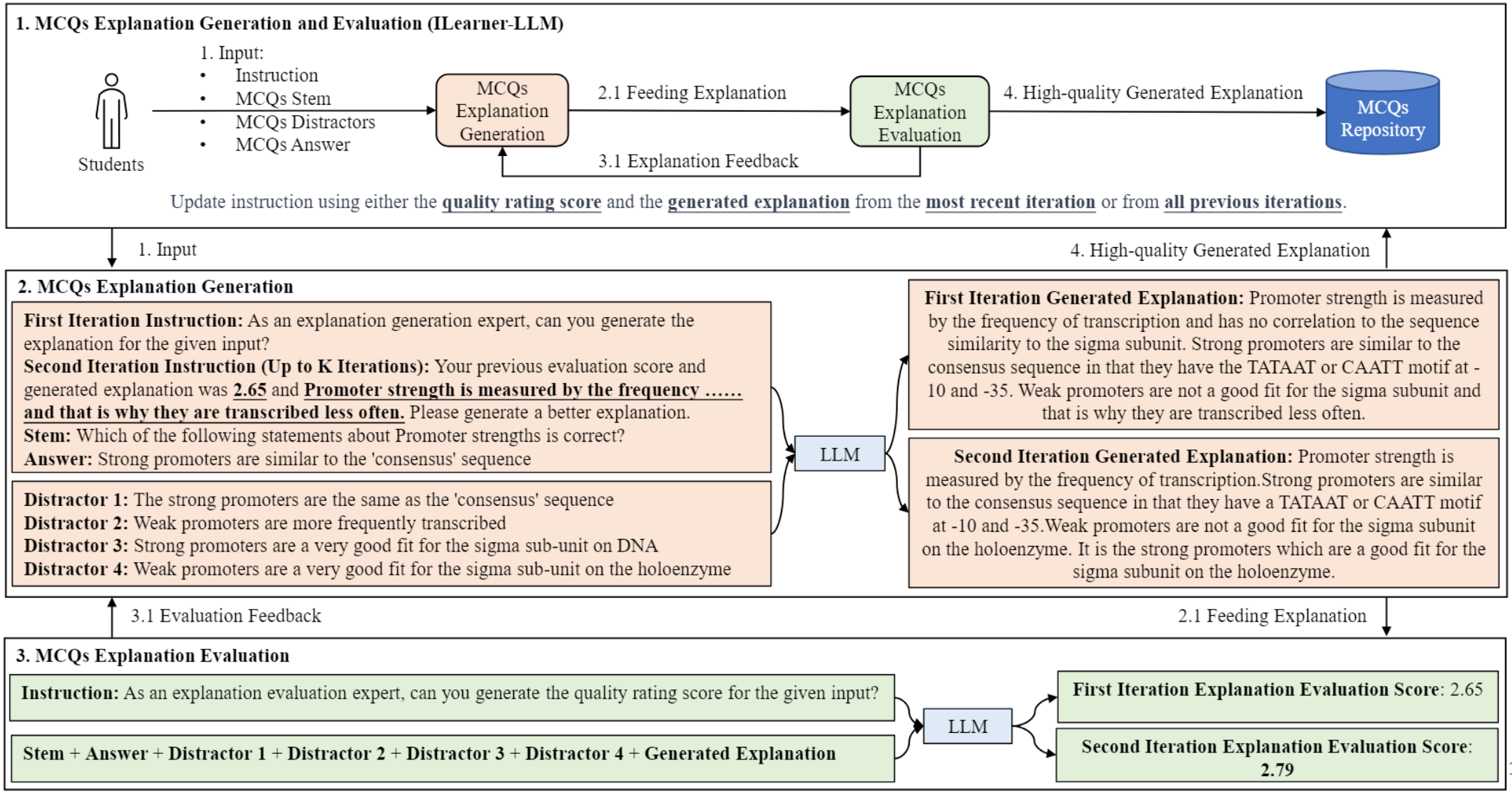
Subject	Sydney Biology	Cardiff Biology	Auckland Law
# MCQs	2311	6955	3449
# Ratings	57585	581937	65645
# Ratings/MCQ	24.91	83.67	19.03
Avg exp length	108.82	75.09	48.13
Subject	UK Medical Year 1	UK Medical Year 2	
# MCQs	3991	2789	
# Ratings	305067	271524	
# Ratings/MCQ	76.43	97.35	
Avg exp length	68.94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

Experiment Setting

- \paragraph{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 sourcecode is available [1].

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

System Architecture "ILearner-LLM"



Main Experiment Results

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**

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Experiment Results

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**

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Experiment Results

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	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

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Experiment Results

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 → 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

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Conclusion and Future Work

In summary, this study presents an 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.

Useful Links

Paper link: https://arxiv.org/abs/2309.10444 (Full paper is under reviewed by AAAI/EAAI 2025)

Project code: https://github.com/Strong-AI-Lab/Explanation-Generation

Selected Publication List

•Qiming Bao, Alex Peng, Zhenyun Deng, Wanjun Zhong, Gaël Gendron, Neşet Tan, Nathan Young, Yang Chen, Yonghua Zhu, Michael Witbrock, Jiamou Liu. *Abstract Meaning Representation-Based Logic-Driven Data Augmentation for Logical Reasoning.*, the Findings of **ACL-24** **[#1 on the ReClor Leaderboard]** **[Paper link]** **[Source code]**

•Qiming Bao, Juho Leinonen, Alex Yuxuan Peng, Wanjun Zhong, Tim Pistotti, Alice Huang, Paul Denny, Michael Witbrock, Jiamou Liu. *Exploring Iterative Enhancement for Improving Learnersourced Multiple-Choice Question Explanations with Large Language Models*, **AGI@ICLR 2024** **[Paper link]** **[Source code]**

•**Qiming Bao, Gaël Gendron, Alex Peng, Neset Tan, Michael Witbrock, Jiamou Liu.** ***A Systematic Evaluation of Large Language Models on Out-of-Distribution Logical Reasoning Tasks.*****,** LLM@IJCAI'23 [Paper link] [Source code]

•**Qiming Bao, Alex Peng, Tim Hartill, Neset Tan, Zhenyun Deng, Michael Witbrock, Jiamou Liu.** ***Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation*****,** IJCLR-NeSy-22 [Paper link] [Source code and dataset] [Presentation recording]

•**Lin Ni, Qiming Bao, Xiaoxuan Li, Qianqian Qi, Paul Denny, Jim Warren, Michael Witbrock, Jiamou Liu.** ***DeepQR: Neural-based Quality Ratings for Learnersourced Multiple-Choice Questions*****,** AAAI/EAAI-22 [Paper link]

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND