

# Improving Multiple-Choice Question Explanations and Logical Reasoning through Iterative Enhancement with Large Language Models and AMR Logic-Driven Data Augmentation

Speaker: Qiming Bao

Homepage: <https://14h034160212.github.io/>

Strong AI Lab, NAOInstitute, The University of Auckland, New Zealand

14th September 2024

UMass Invited Talk

# 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.

# 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**, **AACL**, **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".

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

Authored by: **Qiming Bao**<sup>1,2</sup>, **Juho Leinonen**<sup>3</sup>, **Alex Yuxuan Peng**<sup>1</sup>, **Wanjuan Zhong**<sup>4</sup>, **Gaël Gendron**<sup>1</sup>, **Timothy Pistotti**<sup>1</sup>, **Alice Huang**<sup>6</sup>, **Paul Denny**<sup>5</sup>, **Michael Witbrock**<sup>1</sup>, **Jiamou Liu**<sup>1</sup>

<sup>1</sup>Strong AI Lab, NAOInstitute, Waipapa Taumata Rau - The University of Auckland

<sup>2</sup>Xtracta, New Zealand

<sup>3</sup>Aalto University, Finland

<sup>4</sup>School of Computer Science and Engineering, Sun Yat-Sen University, China

<sup>5</sup>School of Computer Science, The University of Auckland, New Zealand

<sup>6</sup>School of Life and Environmental Sciences, The University of Sydney, Australia

**AGI@ICLR 2024**

<https://arxiv.org/abs/2309.10444>

# Outline

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

# 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

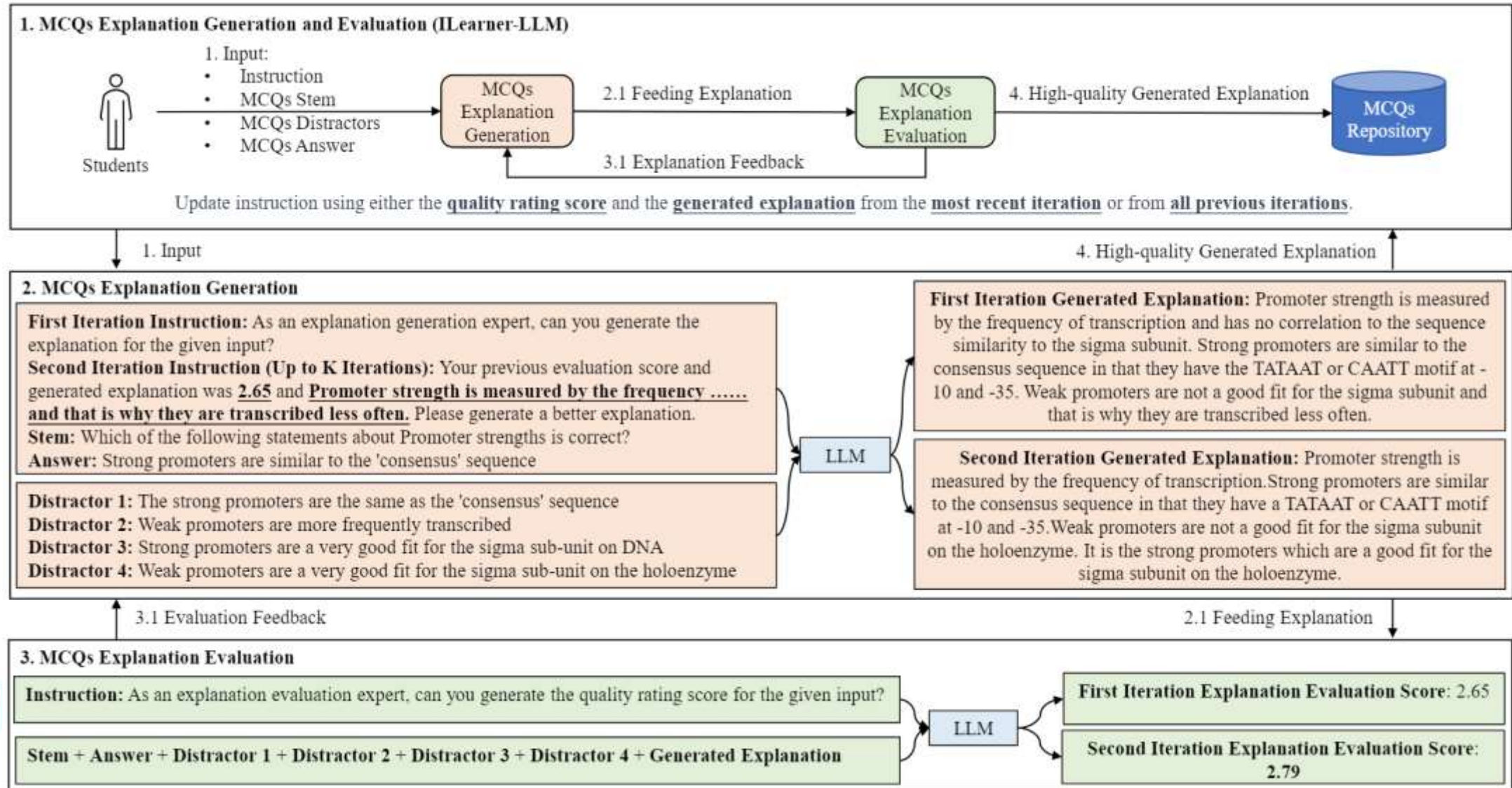
  

Subject	UK Medical Year 1	UK Medical Year 2
# MCQs	3991	2789
# Ratings	305067	271524
# Ratings/MCQ	76.43	97.35
Avg exp length	68.94	83.38

# 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].

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



# 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	<b>34.80</b>	34.34	30.25	34.24
Avg BERT Score	20.17	63.33	62.26	61.62	63.56	<b>63.72</b>
Cardiff Biology Subject						
Avg BLEU Score	3.36	15.33	25.37	25.59	25.65	<b>29.08</b>
Avg BERT Score	8.76	51.72	56.85	58.60	57.69	<b>58.72</b>
Auckland Law Subject						
Avg BLEU Score	3.09	9.36	26.39	<b>27.82</b>	22.16	24.31
Avg BERT Score	7.99	45.38	57.07	<b>58.01</b>	57.11	57.19
UK Medical Year 1 Subject						
Avg BLEU Score	1.92	15.09	26.17	27.60	25.44	<b>28.29</b>
Avg BERT Score	6.22	52.06	57.23	58.45	58.44	<b>59.47</b>
UK Medical Year 2 Subject						
Avg BLEU Score	4.23	17.72	24.76	23.89	26.61	<b>30.67</b>
Avg BERT Score	12.47	51.62	55.91	56.82	57.15	<b>58.17</b>

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



# 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	<b>0.22</b>	3.95
Cardiff Biology Subject				
MSE	0.58	0.10	<b>0.09</b>	3.28
Auckland Law Subject				
MSE	2.86	<b>0.11</b>	0.12	0.42
UK Medical Year 1 Subject				
MSE	0.84	0.19	<b>0.15</b>	3.23
UK Medical Year 2 Subject				
MSE	1.71	0.10	<b>0.09</b>	3.02

# 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>

# Abstract Meaning Representation-Based Logic-Driven Data Augmentation for Logical Reasoning

Authored by: **Qiming Bao**<sup>1,2</sup>, **Alex Yuxuan Peng**<sup>1</sup>, **Zhenyun Deng**<sup>3</sup>, **Wanjuan Zhong**<sup>4</sup>, **Gaël Gendron**<sup>1</sup>, **Timothy Pistotti**<sup>1</sup>, **Neşet Tan**<sup>1</sup>, **Nathan Young**<sup>1</sup>, **Yang Chen**<sup>1</sup>, **Yonghua Zhu**<sup>1</sup>, **Paul Denny**<sup>5</sup>, **Michael Witbrock**<sup>1</sup>, **Jiamou Liu**<sup>1</sup>

<sup>1</sup>Strong AI Lab, NAOInstitute, Waipapa Taumata Rau - The University of Auckland

<sup>2</sup>Xtracta, New Zealand

<sup>3</sup>Department of Computer Science and Technology, University of Cambridge, The United Kingdom

<sup>4</sup>School of Computer Science and Engineering, Sun Yat-Sen University, China

<sup>5</sup>School of Computer Science, The University of Auckland, New Zealand

The Findings of ACL 2024

<https://arxiv.org/abs/2305.12599>

# Research Gap

- Enabling pre-trained large language models (LLMs) to reliably perform logical reasoning is an important step towards strong artificial intelligence [1]. The lack of available large real-world logical reasoning datasets means that LLMs are usually trained on more general corpora or smaller ones that do not generalise well.
- Logical reasoning is extremely important for solving problems in a robust, faithful and explainable way [2] [3], but because logical reasoning is complex for humans to understand and difficult to use for constructing data, there is exceptionally limited data. This implies that a scarcity of labeled datasets for logical reasoning persists in real-world scenarios. Consequently, it is not surprising that these pre-trained language models exhibit shortcomings in logical reasoning [4].

[1] Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.

[2] Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., ... & Srivastava, S. (2020). Logical neural networks. arXiv preprint arXiv:2006.13155.

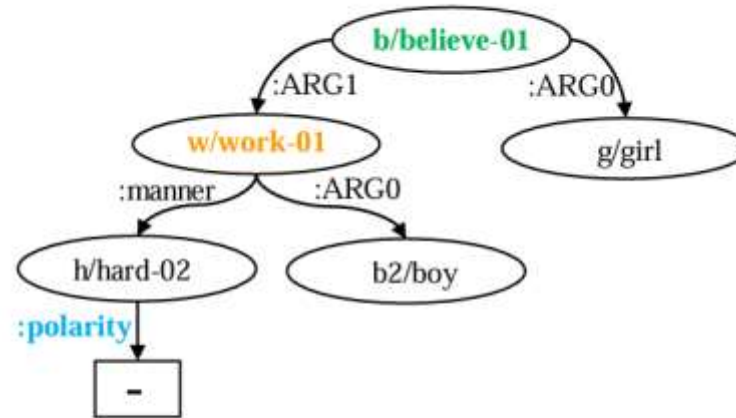
[3] Bansal, A., Schwarzschild, A., Borgnia, E., Emam, Z., Huang, F., Goldblum, M., & Goldstein, T. (2022). End-to-end Algorithm Synthesis with Recurrent Networks: Extrapolation without Overthinking. Advances in Neural Information Processing Systems, 35, 20232-20242.

[4] Yu, F., Zhang, H., & Wang, B. (2023). Nature language reasoning, a survey. arXiv preprint arXiv:2303.14725.

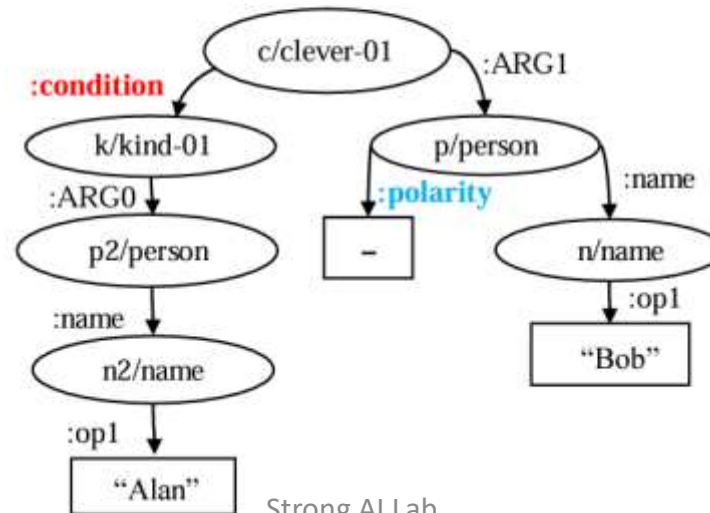
# Abstract Meaning Representation

S1: The girl **believes** that the boy **doesn't work** hard.

S2: That the boy **doesn't work** hard is what the girl **believes**.



S3: **If** Alan is kind, then Bob is **not** clever.



# Logical Reasoning Tasks

## Example Case

**Context:** If you have no keyboarding skills at all, you will not be able to use a computer. And if you are not able to use a computer, you will not be able to write your essays using a word processing program.

**Question:** If the statements above are true, which one of the following must be true?

### Options:

A. If you are not able to write your essays using a word processing program, you have no keyboarding skills.

B. *If you are able to write your essays using a word processing program, you have at least some keyboarding skills. ✓*

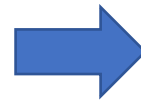
C. If you are not able to write your essays using a word processing program, you are not able to use a computer.

D. If you have some keyboarding skills, you will be able to write your essays using a word processing program.

$\alpha$  = you have keyboarding skills.

$\beta$  = you are able to use a computer.

$\gamma$  = you are able to write your essays using a word processing program.



Context:  $\neg \alpha \rightarrow \neg \beta, \neg \beta \rightarrow \neg \gamma$

Option A:  $\neg \gamma \rightarrow \neg \alpha$

✓ Option B:  $\gamma \rightarrow \alpha + (\beta \rightarrow \alpha, \gamma \rightarrow \beta)$  using contraposition law

Option C:  $\neg \gamma \rightarrow \neg \beta$

Option D:  $\alpha \rightarrow \gamma$

A natural language logical reasoning reading comprehension example from ReClor[1].

Convert the natural language into logic symbols.

# Logical Equivalence Laws

## Definition 1: Contraposition law

$$(\mathcal{A} \rightarrow \mathcal{B}) \Leftrightarrow (\neg \mathcal{B} \rightarrow \neg \mathcal{A})$$

*If Alan is kind, then Bob is clever.  $\Leftrightarrow$  If Bob is not clever, then Alan is not kind.*

## Definition 2: Implication law

$$(\mathcal{A} \rightarrow \mathcal{B}) \Leftrightarrow (\neg \mathcal{A} \vee \mathcal{B})$$

*If Alan is kind, then Bob is clever.  $\Leftrightarrow$  Alan is not kind or Bob is clever.*

## Definition 3: Commutative law

$$(\mathcal{A} \wedge \mathcal{B}) \Leftrightarrow (\mathcal{B} \wedge \mathcal{A})$$

*Alan is kind and Bob is clever.  $\Leftrightarrow$  Bob is clever and Alan is kind.*

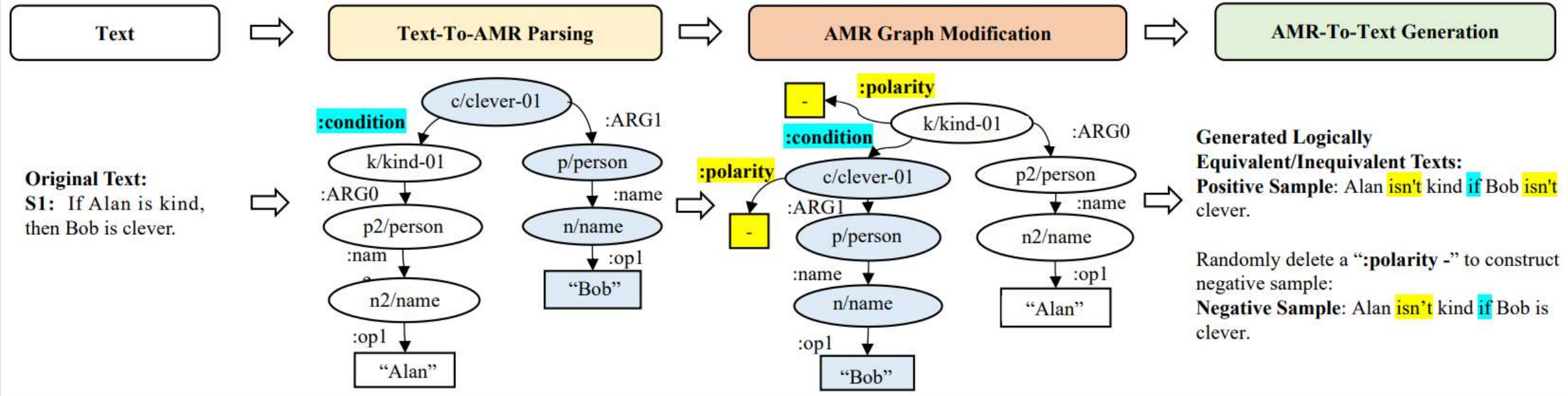
## Definition 4: Double negation law

$$\mathcal{A} \Leftrightarrow \neg \neg \mathcal{A}$$

*Alan is kind.  $\Leftrightarrow$  Alan is not unkind.*

# System Architecture

## 1. AMR-Based Logic-Driven Data Augmentation (AMR-LDA)



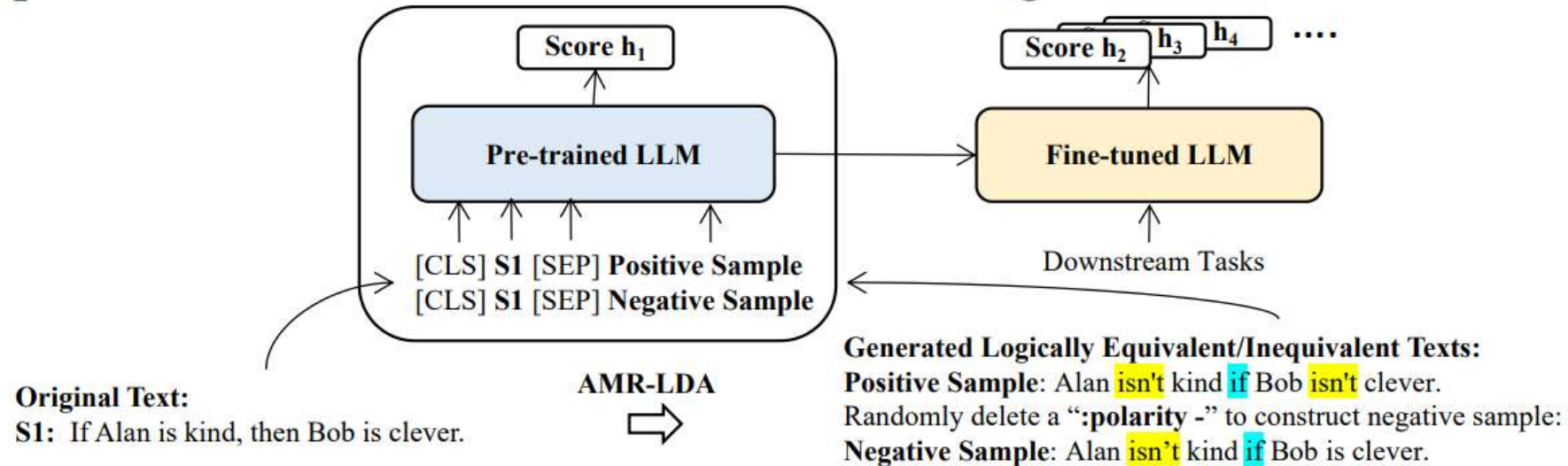
# Construct positive and negative samples

Original sentence	Positive sample	Negative sample
If Alan is kind, then Bob is clever.	Alan isn't kind if Bob isn't clever.	Alan isn't kind if Bob is clever.
	Alan is not kind or Bob is clever.	Alan is kind or Bob is clever.
The bald eagle is strong.	The bald eagle is not weak .	The bald eagle is weak .
The bald eagle is clever and the wolf is fierce.	The wolf is fierce and the bald eagle is clever .	The wolf is not fierce and the bald eagle is not clever .

Table 1: We used four logical equivalence laws to construct positive samples. For the negative samples, we modify the AMR graph of the positive sample, including deleting/adding a negative polarity argument in the AMR graph. The blue background represents the word or the phrase has been swapped order. The yellow background represents the word or the phrase has been adding or deleting a negation meaning.

# System Architecture

## 2a. Logical-Equivalence-Identification Contrastive Learning for Discriminative LLM



# System Architecture

## 2b. Prompt Augmentation for Generative LLM

Context:  $\neg \alpha \rightarrow \neg \beta, \neg \beta \rightarrow \neg \gamma$   
Option A:  $\neg \gamma \rightarrow \neg \alpha$   
Option B:  $\gamma \rightarrow \alpha$   
Option C:  $\neg \gamma \rightarrow \neg \beta$   
Option D:  $\alpha \rightarrow \gamma$

AMR-LDA  
⇒

Context:  $\neg \alpha \rightarrow \neg \beta, \neg \beta \rightarrow \neg \gamma$

Option A:  $\neg \gamma \rightarrow \neg \alpha$  + AMR-LDA extended option:  $\alpha \rightarrow \gamma$  + AMR-LDA extended context:  $\beta \rightarrow \alpha, \gamma \rightarrow \beta$   
Option B:  $\gamma \rightarrow \alpha$  + AMR-LDA extended option:  $\neg \alpha \rightarrow \neg \gamma$  + AMR-LDA extended context:  $\beta \rightarrow \alpha, \gamma \rightarrow \beta$   
Option C:  $\neg \gamma \rightarrow \neg \beta$  + AMR-LDA extended option:  $\beta \rightarrow \gamma$  + AMR-LDA extended context:  $\beta \rightarrow \alpha, \gamma \rightarrow \beta$   
Option D:  $\alpha \rightarrow \gamma$  + AMR-LDA extended option:  $\neg \gamma \rightarrow \neg \alpha$  + AMR-LDA extended context:  $\beta \rightarrow \alpha, \gamma \rightarrow \beta$

$\alpha$  = you have keyboarding skills.

$\beta$  = you are able to use a computer.

$\gamma$  = you are able to write your essays using a word processing program.

Solution Path 1

Solution Path 2

⇒  ⇒ Option B ✓

# Case Study

## AMR-LDA Prompt Augmentation Case Study

**GPT-4 Input:** “context”: “If you have no keyboarding skills at all, you will not be able to use a computer. And if you are not able to use a computer, you will not be able to write your essays using a word processing program.”, “question”: “If the statements above are true, which one of the following must be true?”, “answers”:

A. “If you are not able to write your essays using a word processing program, you have no keyboarding skills. *If you have the skill of a keyboard, you can write your essay using a word processing program. If you can use a computer, you have keyboarding skills. If you can write your essay with a word processing program, you can use a computer. Whether you have keyboard skills at all or can't use a computer. Whether you can use a computer or you can't write your own essay with a word processing program.*”,

B. “If you are able to write your essays using a word processing program, you have at least some keyboarding skills. *If you don't have at least some keyboard skills, you can't write your essay with a word processing program. If you can use a computer, you have keyboarding skills. If you can write your essay with a word processing program, you can use a computer. Whether you have keyboard skills at all or can't use a computer. Whether you can use a computer or you can't write your own essay with a word processing program.*”,

C. “If you are not able to write your essays using a word processing program, you are not able to use a computer. *If you can use a computer, you can write your essay using word processing programs. If you can use a computer, you have keyboarding skills. If you can write your essay with a word processing program, you can use a computer. Whether you have keyboard skills at all or can't use a computer. Whether you can use a computer or you can't write your own essay with a word processing program.*”,

D. “If you have some keyboarding skills, you will be able to write your essays using a word processing program. *If you can't write your essay with a word processing program, you don't have some keyboard skills. If you can use a computer, you have keyboarding skills. If you can write your essay with a word processing program, you can use a computer. Whether you have keyboard skills at all or can't use a computer. Whether you can use a computer or you can't write your own essay with a word processing program.*”


**GPT-4 output: B**

Figure 3: Example for using AMR-LDA to augment the prompt from ReClor dataset and their subsequent utilisation as input for GPT-4. Data segments that are marked in bold italics and appear in blue were generated using the contraposition law, while those in brown were generated using the implication law.

# Experiment Results

Models/ Datasets	ReClor				LogiQA		MNLI	MRPC	RTE	QNLI	QQP
	Dev	Test	Test-E	Test-H	Dev	Test			Eval		
RoBERTa	59.73	53.20	72.57	37.97	35.43	34.50	88.95	90.44	83.39	<b>94.73</b>	<b>90.89</b>
RoBERTa LReasoner-LDA	59.46	53.66	72.19	39.10	34.81	34.81	89.41	89.46	86.28	94.25	90.01
RoBERTa AMR-DA	58.66	53.93	66.81	<b>43.80</b>	36.45	37.22	89.74	90.44	86.28	94.42	92.06
RoBERTa AMR-LDA	<b>65.26</b>	<b>56.86</b>	<b>77.34</b>	40.77	<b>40.29</b>	<b>38.14</b>	<b>89.78</b>	<b>90.93</b>	<b>86.64</b>	94.49	<b>93.14</b>
DeBERTaV2	73.93	70.46	80.82	62.31	39.72	39.62	89.45	89.71	84.48	95.00	<b>92.54</b>
DeBERTaV2 LReasoner-LDA	75.73	70.70	84.08	60.17	30.87	28.51	89.23	89.95	87.00	95.15	92.50
DeBERTaV2 AMR-DA	79.06	75.90	84.62	69.04	29.95	30.10	<b>89.92</b>	89.71	83.39	95.02	92.42
DeBERTaV2 AMR-LDA	<b>79.40</b>	<b>77.63</b>	<b>85.75</b>	<b>71.24</b>	<b>42.34</b>	<b>39.88</b>	89.67	<b>90.20</b>	<b>88.09</b>	<b>95.24</b>	92.47

Table 2: Comparison between our proposed AMR-LDA and baseline models. We use RoBERTa-Large, DeBERTaV2-XXLarge as the pre-trained models. Our fine-tuned LLMs perform equally well or better than baseline methods.



ReClor - A Reading Comprehension Dataset  
Requiring Logical Reasoning

Organized by: [ReClor Team](#)  
Starts on: Jan 1, 2020 1:00:00 PM NZST (GMT + 13:00)  
Ends on: Jan 1, 2100 12:59:59 PM NZST (GMT + 13:00)

[Overview](#)
[Evaluation](#)
[Phases](#)
[Participate](#)
[Leaderboard](#)
[Discuss](#)

Rank	Participant team	Test (t)	Test-E (t)	Test-H (t)	NA (t)	SA (t)	S (t)	W (t)	E (t)	I (t)	CMP (t)	MSS (t)	ER (t)	P (t)	D (t)	T (t)	R (t)	IF (t)
1	<a href="#">AMR-LDA Team</a>	90.20	91.59	89.11	92.11	83.33	90.43	88.50	100.00	84.78	97.22	94.64	94.05	87.69	96.67	94.44	87.50	91
2	<a href="#">HFL &amp; iFLYTEK (IDOL/Rational Reasoner)</a>	80.60	87.73	75.00	86.84	90.00	84.04	72.57	76.92	58.70	86.11	73.21	82.14	76.92	80.00	86.11	81.25	83
3	<a href="#">MERIT (MERIT-deberta-v2-xxlarge)</a>	79.30	85.23	74.64	85.09	83.33	82.98	71.68	76.92	65.22	83.33	73.21	76.19	80.00	80.00	88.89	78.13	81
4	<a href="#">Knowledge Model Team (Knowledge model)</a>	79.20	91.82	69.29	89.47	80.00	76.60	68.14	92.31	63.04	94.44	78.57	78.57	78.46	76.67	97.22	84.38	76
5	<a href="#">AMR-LDA (DeBERTa-v2-xxlarge-AMR-LDA-Con)</a>	77.20	86.14	70.18	83.33	76.67	79.79	68.14	84.62	52.17	88.89	80.36	75.00	75.38	80.00	88.89	71.88	78
6	<a href="#">LReasoner Team (LReasoner ensemble)</a>	76.10	87.05	67.50	80.70	80.00	76.60	67.26	84.62	67.39	88.89	76.79	76.19	75.38	63.33	88.89	71.88	74

Models/Datasets	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
GPT-3.5	57.02	56.20	59.31	<b>53.75</b>	37.63	37.32
+ CoT	34.80	25.80	27.50	24.46	23.96	24.57
+ AMR-DA	33.20	32.90	34.31	31.78	<b>40.55</b>	31.49
+ AMR-LDA	<b>58.62</b>	<b>56.69</b>	<b>60.90</b>	53.39	<b>40.55</b>	<b>39.47</b>
GPT-4	87.35	89.60	90.90	88.57	43.24	53.88
+ CoT	37.00	24.80	26.13	23.75	23.50	27.03
+ AMR-DA	85.00	85.60	86.36	85.00	51.30	56.06
+ AMR-LDA	<b>87.73</b>	<b>90.20</b>	<b>91.59</b>	<b>89.11</b>	<b>51.92</b>	<b>58.06</b>

Table 3: Comparison of Chain-of-Thought Prompting (CoT), AMR-DA, and AMR-LDA on GPT-3.5 and GPT-4, and between GPT-3.5 and GPT-4 alone, for evaluation on the ReClor and LogiQA test sets.

# Experiment Results

Models/Datasets	RoBERTa AMR-LDA	RoBERTa LReasoner-LDA
Depth=1	100.00	100.00
Depth=1 (with altered rules)	<b>100.00</b>	99.87
Depth=2	100.00	100.00
Depth=2 (with altered rules)	<b>99.73</b>	74.00

Table 4: Comparison between AMR-LDA and LReasoner-LDA with RoBERTa-Large on PARARULE-Plus and PARARULE-Plus (with altered rules). Depth=1 means that only one rule was used to infer the answer. Depth=1 (with altered rules) means one of the rules has been altered using logical equivalence law.

# Experiment Results

Models/Datasets	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
DeBERTaV2-XXLarge	73.93	70.46	80.82	62.31	39.72	39.62
+ AMR-LDA-1:1	78.80	76.10	<b>84.77</b>	69.28	40.55	41.47
+ AMR-LDA-1:2	80.20	<b>76.40</b>	<b>84.77</b>	<b>69.82</b>	<b>47.00</b>	<b>43.93</b>
+ AMR-LDA-1:3	<b>81.20</b>	75.70	84.09	69.10	42.70	41.01
DeBERTaV2-XXLarge + MERIt-1:3	80.20	75.80	85.00	68.57	37.32	42.39
+ AMR-LDA-Con-1:3	<b>82.60</b>	76.60	86.13	<b>69.10</b>	<b>45.00</b>	43.01
+ AMR-LDA-Merged-1:3	81.80	<b>76.90</b>	<b>87.50</b>	68.57	44.54	<b>45.62</b>
DeBERTaV2-XXLarge + IDoL	77.60	74.50	82.95	67.85	39.78	40.24
+ AMR-LDA-Con-1:3	79.20	<b>77.00</b>	85.68	<b>70.17</b>	<b>47.61</b>	<b>44.54</b>
+ AMR-LDA-Merged-1:3	<b>79.40</b>	75.60	<b>86.36</b>	67.14	41.93	41.32

Table 6: An experiment to assess how positive:negative sample ratios affect downstream tasks. AMR-LDA 1:1 means the ratio of positive and negative samples is 1:1.

Models/Datasets	Con	Con-dou	Con-dou imp	Con-dou imp-com
<i>RoBERTa-Large as backbone model</i>				
ReClor	60.40	60.80	<b>61.80</b>	59.80
LogiQA	37.78	33.17	33.94	<b>38.70</b>
MNLI	89.55	<b>90.15</b>	89.68	89.78
MRPC	90.69	89.22	90.44	<b>90.93</b>
RTE	81.23	85.20	84.84	<b>86.64</b>
QNLI	94.16	94.05	<b>94.51</b>	94.49
QQP	92.12	89.88	92.06	<b>93.14</b>
<i>DeBERTaV2-XXLarge as backbone model</i>				
ReClor	<b>81.80</b>	72.20	79.40	78.80
LogiQA	32.25	<b>45.46</b>	38.24	40.55
<i>DeBERTa-Large as backbone model</i>				
MNLI	<b>90.80</b>	90.59	90.68	89.67
MRPC	<b>90.20</b>	88.48	89.95	<b>90.20</b>
RTE	84.84	87.36	85.56	<b>88.09</b>
QNLI	<b>95.28</b>	95.04	94.97	95.24
QQP	92.33	92.40	92.29	<b>92.47</b>

Table 5: An experiment to assess the influence of different logical equivalence laws on downstream logical reasoning and natural language inference tasks. “Con”, “dou”, “imp” and “com” are the abbreviation for contraposition law, double negation law, implication law and commutative law. “Con-dou” denotes data constructed using both the contraposition law and the double negation law. Other terms are derived in a similar manner.

# Human Evaluation

We randomly select 20 samples which are composed of pairs of two sentences from the generated sentences using our AMR-LDA and LReasoner-LDA to conduct a survey. We select 45 participants anonymously. We evaluate the sentences from two aspects.

- The first is which sentence is logically equivalent to the original sentence.
- The other one is which sentence is more fluent.

From our survey, 63.92% and 76.44% people select the sentences generated by AMR-LDA as the more correct logical equivalence sentences and more fluent sentences than the sentences generated by LReasoner-LDA, respectively.

The human evaluation has been approved by the University of Auckland Human Participants Ethics Committee on 28 February, 2023 for three years, Reference Number 24841.

# Conclusion and Future Work

1. We propose a new AMR-based, logic-driven data augmentation method that considers more logical equivalence laws than LReasoner, including double negation, contraposition, commutative, and implication laws. We used the augmented dataset obtained with our method to conduct contrastive fine-tuning various LLMs. Additionally, we fed the augmented data to large language models, such as ChatGPT and GPT-4, which ultimately yielded better results than baseline methods.
2. To automatically construct real-world logical reasoning datasets using **additional logical equivalence laws**, such as De Morgan's Law, we are exploring two approaches: one involves prompting GPT-4, and the other seeks to extend our method by utilizing GPT-4 both as an AMR parser and an AMR generator. (Work in progress)
3. Enhancing Large Language Model From **Logic Programming** And **Knowledge Graph**. Integrating these models with a knowledge graph, which can provide more accurate **factual information**, and prompting or fine-tuning the large language models, presents opportunities to correct and reduce the hallucinations of these models. Aside from **temporal information**, since these large language models are trained based on next-token prediction, it is unsurprising that they are not adept at complex logical reasoning tasks. (Work in progress)

# Useful Links



Project code



#1 on ReClor Leaderboard



Model Weights

Our AMR-LDA has been open-sourced in the project code, and the model weights have been released.

Welcome for more discussion and collaboration!

# 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](#)]