Natural Language Processing and Logical Reasoning

Speaker: Qiming Bao

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

23rd January 2024

深圳北理莫斯科大学第四届国际青年学者论坛人工智能研究院分论坛

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 Wananga o Tamaki Makaurau
NEW ZEALAND

Motivation

- One main reason existing language models struggle with complex natural language reasoning is the **lack of real-world, complex natural language reasoning datasets**, and it is challenging to obtain reliable data from the web for building expansive training datasets.
- Furthermore, when large language models come out, they demonstrate evident improvement on the public logical reasoning datasets like ReClor, LogiQA and LogiQAv2, but whether this means those large language models have **strong and robust logical reasoning ability** remains to be seen.

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

Enhancing Logical Reasoning of Large Language Models through Logic-Driven Data Augmentation

Authored by: **<u>Qiming Bao</u>[1,2], Alex Yuxuan Peng[1], Zhenyun Deng[3], Wanjun Zhong[4], Gaël Gendron[1], Timothy Pistotti[1], Neşet Tan[1], Nathan Young[1], Yang Chen[1], Yonghua Zhu[1], 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]Department of Computer Science and Technology, University of Cambridge, The United Kingdom

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

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

The first edition of the Symposium on Advances and Open Problems in Large Language Models (**LLM@IJCAI'23**)

https://arxiv.org/abs/2305.12599

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

Outline

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

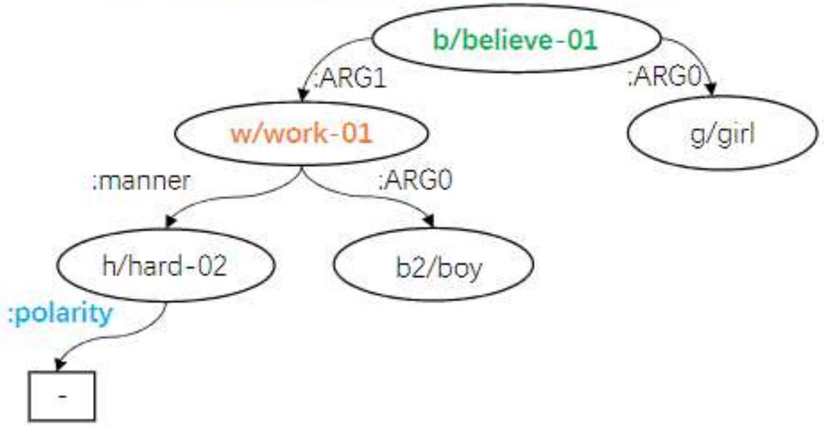
Research Gap

- 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].

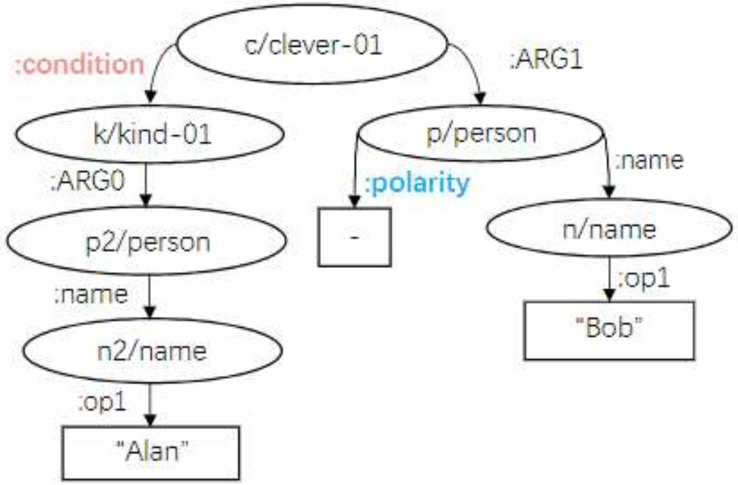
[1] Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.
[2] Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., ... & Srivastava, S. (2020). Logical neural networks. arXiv preprint arXiv:2006.13155.
[3] Bansal, A., Schwarzschild, A., Borgnia, E., 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: The girl **doesn't believe** that the boy **works** hard.



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.

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

α = you have keyboarding skills.
β = you are able to use a computer.
γ = 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: $\alpha \rightarrow \gamma$

Convert the natural language into logic symbols.

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

https://whyu.me/reclor/

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

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

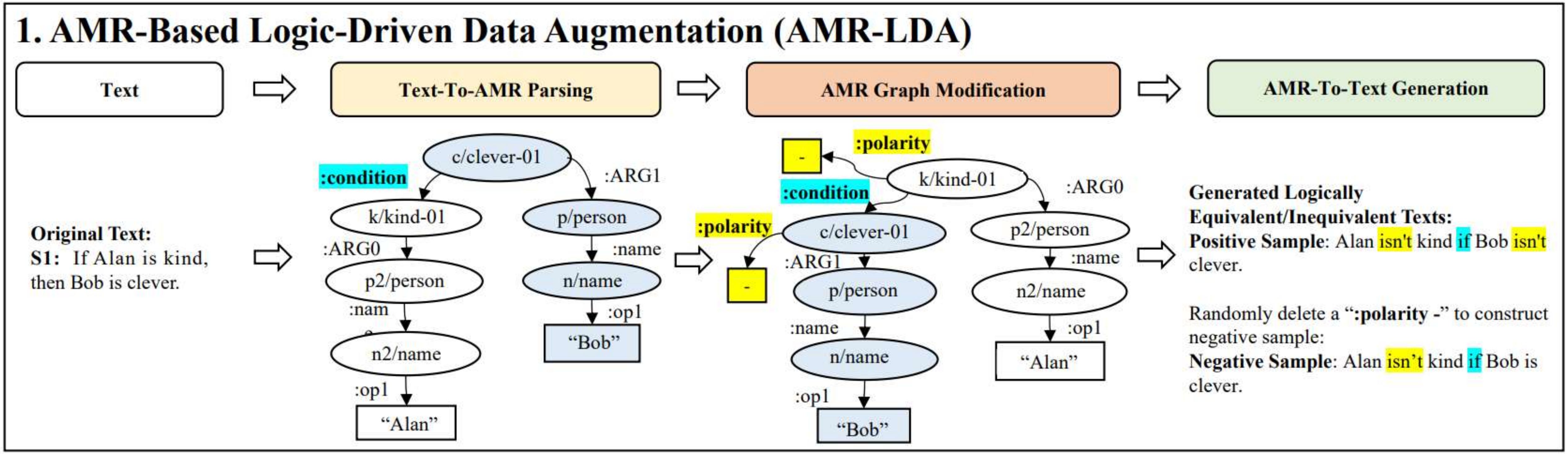
Alan is kind and Bob is clever. $\Leftrightarrow$ *Bob is clever and Alan is kind.*

Definition 4: Double negation law

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

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

System Architecture

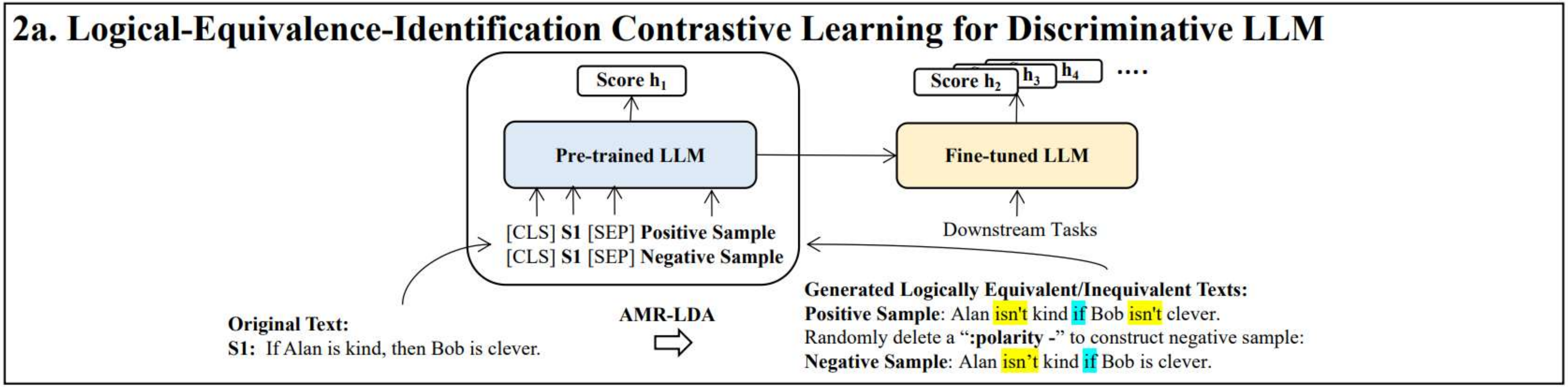


Construct positive and negative samples

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.

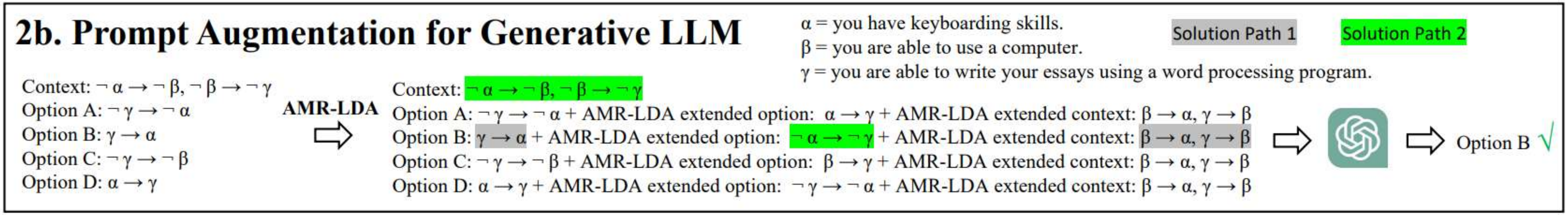
Table 1: We used four logical equivalence laws to construct positive samples. For the negative samples, we modify the AMR graph of the positive sample, including deleting/adding a negative polarity argument in the AMR graph. The blue background represents the word or the phrase has been swapped order.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

System Architecture



THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

System Architecture



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.

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

Experiment Results

Models/ Datasets	ReClor				LogiQA		MNLI	MRPC	RTE	QNLI	QQP
	Dev	Test	Test-E	Test-H	Dev	Test	Eval				
RoBERTa	0.5973	0.5320	0.7257	0.3797	0.3543	0.3450	0.8895	0.9044	0.8339	**0.9473**	0.9089
RoBERTa AMR-LDA	**0.6526**	**0.5686**	**0.7734**	0.4077	**0.4029**	**0.3814**	**0.8978**	**0.9093**	**0.8664**	0.9449	**0.9314**
RoBERTa LReasoner-LDA	0.5946	0.5366	0.7219	0.3910	0.3481	0.3481	0.8941	0.8946	0.8628	0.9425	0.9001
RoBERTa AMR-DA	0.5866	0.5393	0.6681	**0.4380**	0.3645	0.3722	0.8974	0.9044	0.8628	0.9442	0.9206
DeBERTaV2	0.7393	0.7046	0.8082	0.6231	0.3972	0.3962	0.8945	0.8971	0.8448	0.9500	**0.9254**
DeBERTaV2 AMR-LDA	**0.7940**	**0.7763**	**0.8575**	**0.7124**	**0.4234**	**0.3988**	0.8967	**0.9020**	**0.8809**	**0.9524**	0.9247
DeBERTaV2 LReasoner-LDA	0.7573	0.7070	0.8408	0.6017	0.3087	0.2851	0.8923	0.8995	0.8700	0.9515	0.9250
DeBERTaV2 AMR-DA	0.7906	0.7590	0.8462	0.6904	0.2995	0.3010	**0.8992**	0.8971	0.8339	0.9502	0.9242

Table 2: Comparison between our proposed AMR-LDA and baseline models. We use RoBERTa-Large, DeBERTaV2-XXLarge, and DeBERTa-Large as the pre-trained backbone models. Our fine-tuned LLMs perform equally well or better than baseline methods. The number with * indicates that the result is from the other papers.

ReClor ReClor - A Reading Comprehension Dataset Requiring Logical Reasoning

★ 40

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 (↑)	Test-E (↑)	Test-H (↑)	NA (↑)	SA (↑)	S (↑)	W (↑)	E (↑)	I (↑)	CMP (↑)	MSS (↑)	ER (↑)	P (↑)	D (↑)	T (↑)	R (↑)	IF (↑
1	AMR-LDA Team	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	HFL & iFLYTEK (IDOL/Rational Reasoner)	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	MERIt (MERIt-deberta-v2-xxlarge)	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	Knowledge Model Team (Knowledge model)	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	AMR-LDA (DeBERTa-v2-xxlarge-AMR-LDA-Con)	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	LReasoner Team (LReasoner ensemble)	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	0.5702	0.5620	0.5931	**0.5375**	0.3763	0.3732
GPT-3.5 AMR-LDA	**0.5862**	**0.5669**	**0.6090**	0.5339	**0.3974**	**0.3947**
GPT-4	0.8735	0.8960	0.9090	0.8857	0.4324	0.5388
GPT-4 AMR-LDA	**0.8773**	**0.9020**	**0.9159**	**0.8911**	**0.4751**	**0.5806**

Table 5: Comparison between GPT-3.5 AMR-LDA, GPT-4 AMR-LDA with GPT-3.5 and GPT-4 alone for evaluating on ReClor and LogiQA test sets.

Experiment Results

Test sets ↓ Models →	Test acc RoBERTa AMR-LE	 RoBERTa LReasoner-LE
Depth=1	1	1
Depth=1 (change rule)	**1**	0.9987
Depth=2	1	1
Depth=2 (change rule)	**0.9973**	0.7400

Table 4: A comparative experiment between AMR-LE fine-tuned PLM and LReasoner-LE fine-tuned PLM on PARARULE-Plus, and PARARULE-Plus changed rule by logical equivalence laws. Depth=1 means that only one rule was used to infer the answer. Depth=1 (change rule) means we used logical equivalence laws to rewrite one of the rules, and we conducted the same modification for Depth=2 (change rule).

https://github.com/Strong-AI-Lab/PARARULE-Plus
https://github.com/openai/evals/pull/651

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

Experiment Results

Models/Datasets	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
	DeBERTaV2-XXLarge as backbone model					
AMR LDA 1:1	0.7880	0.7610	0.8477	0.6928	0.4055	0.4147
AMR LDA 1:2	0.8020	**0.7640**	0.8477	**0.6982**	0.4700	**0.4393**
AMR LDA 1:3	0.8120	0.7570	0.8409	0.6910	0.4270	0.4101
MERIt 1:3	0.8020	0.7580	**0.8500**	0.6857	0.3732	0.4239
	MERIT-DeBERTaV2-XXLarge as backbone model					
AMR LDA Contraposition	0.8260	0.7660	0.8613	**0.6910**	0.4500	0.4301
AMR LDA Merged	0.8180	**0.7690**	**0.8750**	0.6857	0.4454	**0.4562**

Table 6: An experiment to validate how ratios of positive and negative samples influence downstream tasks. Pos-neg-1-1 means the ratio of positive and negative samples is 1:1.

Dev sets ↓	Dev acc			
Models →	Con	Con-dou	Con-dou imp	Con-dou imp-com
	RoBERTa-Large as backbone model			
ReClor	0.6040	0.6080	**0.6180**	0.5980
LogiQA	0.3778	0.3317	0.3394	**0.3870**
MNLI	0.8955	**0.9015**	0.8968	0.8978
MRPC	0.9069	0.8922	0.9044	**0.9093**
RTE	0.8123	0.8520	0.8484	**0.8664**
QNLI	0.9416	0.9405	**0.9451**	0.9449
QQP	0.9212	0.8988	0.9206	**0.9314**
	DeBERTaV2-XXLarge as backbone model			
ReClor	**0.8180**	0.7220	0.7940	0.7880
LogiQA	0.3225	**0.4546**	0.3824	0.4055
	DeBERTa-Large as backbone model			
MNLI	**0.9080**	0.9059	0.9068	0.8967
MRPC	**0.9020**	0.8848	0.8995	**0.9020**
RTE	0.8484	0.8736	0.8556	**0.8809**
QNLI	**0.9528**	0.9504	0.9497	0.9524
QQP	0.9233	0.9240	0.9229	**0.9247**

Table 5: An ablation study to validate how different logical laws influence downstream tasks. Con means we only use contraposition law. Con-dou means we use contraposition and double negation laws. Con-dou-imp means we use contraposition, double negation and implication laws. Con-dou-imp-com means we use the four logical laws to augment data and conduct the fine-tuning.

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.

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

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. It is worth establishing a more robust logical reasoning evaluation benchmark to assess the logical reasoning capabilities of existing large language models. This is because these models run the risk of having been trained on, and therefore having learned from, public datasets available on the internet.

4. Using tools like logic programming and integrating chain-of-thought prompting to iteratively enhance the o.o.d. logical reasoning capabilities of LLMs is worth exploring. (Working 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!

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