Assessing and Enhancing Language Models for Complex and Robust Logical Reasoning over Natural Language

Speaker: Qiming Bao
Personal Website: https://14h034160212.github.io/
School： Strong AI Lab, NAOInstitute, University of Auckland
13 August 2025

DAAD AINeT fellow 2025 on Natural Language Processing Presentation

Outline

- Motivation
- Research Gap
- Contributions
- Future Work

Motivation

- Existing language models are challenged to effectively perform **complex logical reasoning in natural language**, particularly when confronted with **unbalanced distributions of reasoning depths** in multi-step and more real-world logical reasoning datasets. (IJCLR-NeSy 2022)
- 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. (The Findings of ACL 2024)
- 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. (LLM@IJCAI 2023)

Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation

Authored by: **Qiming Bao[1], Alex Yuxuan Peng[1], Tim Hartill[1], Neşet Özkan Tan[1], Zhenyun Deng[1], Michael Witbrock[1], Jiamou Liu[1]**

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

The 2nd International Joint Conference on Learning & Reasoning and 16th International Workshop on Neural-Symbolic Learning and Reasoning (IJCLR-NeSy 22)

https://arxiv.org/abs/2207.14000

Outline

- Research Gap
  - Our Contribution
- Model Overview
- Experiment Result

Research Gap

- Existing neural-symbolic methods such as Memory Attention Control networks (MAC) [3], Dynamic Memory Network [4], and DeepLogic [5] present visualized attention maps for multi-step logical reasoning tasks over symbolic logic programs. Can these methods be extended to natural language reasoning?
- **Our Contribution:** We utilized GloVe [2] to represent the input sequences, which consist of a context and a statement. The concatenated representations of the context and statement are fed into the model, such as a gated recurrent unit (GRU), for the RNN-based models. The model demonstrates strong performance, achieving over 85% test accuracy on first-order logic multi-step natural language reasoning tasks.

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.
[2] Pennington, et al., 2014. Glove: Global vectors for word representation, EMNLP.
[3] Hudson, et al. 2018. Compositional attention networks for machine reasoning, ICLR.
[4] Kumar, et al. 2016. Ask me anything: Dynamic memory networks for natural language processing, ICML
[5] Cingillioglu, N. et al., 2018. DeepLogic: Towards End-to-End Differentiable Logical Reasoning, AAAI-MAKE19.

Research Gap

- Although RNN-based models and pre-trained transformer models perform well in general, does that mean these models possess strong reasoning abilities?
- **Our Contribution:** We found that existing RNN-based baseline models and pre-trained transformer-based models do not perform well in terms of out-of-distribution (OOD) generalization on multi-step reasoning tasks under the following three scenarios:
  1. When the model is trained on data with shallow reasoning depths but tested on data with deeper reasoning depths.
  2. When the model is trained on synthetically generated data but tested on data paraphrased by humans.
  3. When the model is trained on unshuffled data but tested on shuffled data.

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

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

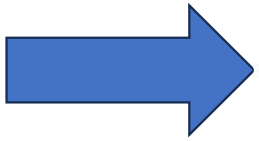
Research Gap

- Existing multi-step deductive reasoning datasets, such as PARARULES and CONCEPTRULE V1 and V2, have unbalanced distributions over reasoning depths. Only a small portion of these datasets require deep reasoning ($2 \leq \text{Depth} \leq 5$).
- **Our Contribution:** We present a larger deep multi-step logical reasoning dataset named PARARULE-Plus. This dataset specifically augments the data at deeper reasoning depths ($2 \leq \text{Depth} \leq 5$). It has also been collected by OpenAI/Evals, and ChatGPT performs poorly on this dataset [2].

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.
[2] https://github.com/openai/evals/pull/651

From Symbolic Logic Program to Natural Language Reasoning

$$p(X) :- q(X).$$
$$q(a).$$
$$?p(a).1$$
$$?p(b).0$$

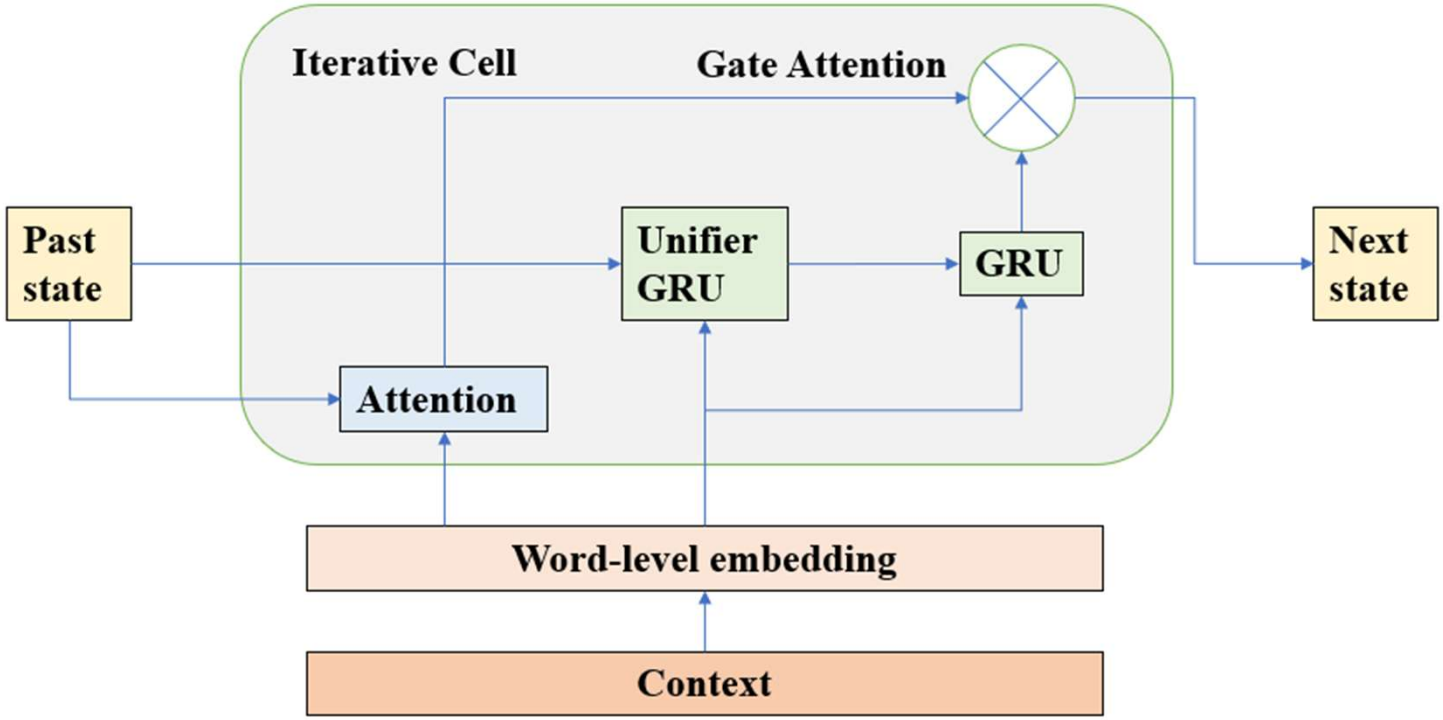


(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

Model Overview



[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 4

We use GloVe [16] as the word vector representation. We use PARARULES with all depths as the training set for all models and then test them on examples with different reasoning depths (D). Comparison among our IMA-GloVe-GA, IMA-GloVe, MAC-GloVe, DMN-GloVe, IMASM-GloVe, LSTM-GloVe, and RoBERTa-Large on PARARULES test sets with different reasoning depths.

Train ↓; Test →	D=1	D=2	D=3	D≤3	D≤3+NatLang	D≤5	D≤5+NatLang
IMA-GloVe	0.861	0.853	0.830	0.842	0.810	0.792	0.705
MAC-GloVe	0.792	0.776	0.750	0.763	0.737	0.701	0.652
DMN-GloVe	0.846	0.843	0.817	0.827	0.789	0.779	0.666
IMASM-GloVe	0.864	0.855	0.824	0.838	0.801	0.782	0.608
LSTM-GloVe	0.500	0.500	0.500	0.499	0.499	0.500	0.500
IMA-GloVe-GA	**0.950**	**0.943**	**0.919**	**0.927**	**0.883**	**0.879**	**0.741**
RoBERTa-Large	**0.986**	**0.985**	**0.977**	**0.979**	**0.972**	**0.967**	**0.949**

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 5

IMA-GloVe, IMA-GloVe-GA, and RoBERTa-Large trained on CONCEPTRULES V1 (simplified / full) and tested on different test sets. Rules in CONCEPTRULES V1 Simplified are not shuffled, while CONCEPTRULES V1 full contains randomly shuffled rules. CONCEPTRULES V1 full has larger number of relations and entities than CONCEPTRULES V1 simplified.

Model	Train set	Test accuracy (Simplified Test set)	Test accuracy (Full Test set)
IMA-GloVe	Simplified	0.994	0.729
	Full	0.844	**0.997**
IMA-GloVe-GA	Simplified	**0.998**	**0.747**
	Full	0.851	**0.999**
RoBERTa-Large	Simplified	**0.997**	0.503
	Full	**0.927**	0.995

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

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 6

IMA-GloVe, IMA-GloVe-GA, and RoBERTa-Large trained on CONCEPTRULES V2 (full) and tested on test sets that require different depths of reasoning.

Model	Test set	Mod1 Depth=1	Mod2 Depth=2	Mod3 Depth=3	Mod01 Depth≤1	Mod012 Depth≤2	Mod0123 Depth≤3
IMA-GloVe	Depth=1	**0.999**	**0.998**	**0.990**	**0.997**	**0.998**	**0.997**
	Depth=2	**0.998**	**0.999**	**0.988**	**0.995**	**0.998**	**0.997**
	Depth=3	**0.997**	0.998	0.981	**0.991**	0.996	**0.997**
IMA-GloVe-GA	Depth=1	0.993	0.996	0.987	0.987	0.991	**0.997**
	Depth=2	0.993	**0.999**	0.974	0.986	0.991	0.995
	Depth=3	0.988	**1**	**0.994**	0.989	**0.997**	0.994
RoBERTa-Large	Depth=1	0.998	0.975	0.831	0.995	0.975	0.971
	Depth=2	0.997	0.972	0.885	0.993	0.972	0.965
	Depth=3	0.987	0.951	0.984	0.988	0.951	0.936

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

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 7

RoBERTa-Large trained on PARARULES with different reasoning depths and tested on test sets that require different depths of reasoning. A bold number indicates the highest accuracy in a test set.

Model	Test set	Mod012 (Depth≤2)	Mod0123 (Depth≤3)	Mod0123Nat (Depth≤3+NatLang)	Mod012345 (Depth≤5)
RoBERTa-Large	Depth=0	**0.971**	0.946	0.968	0.953
	Depth=1	**0.943**	0.907	0.933	0.909
	Depth=2	**0.933**	0.902	0.932	0.902
	Depth=3	0.562	0.902	**0.926**	0.907
	Depth=4	0.481	0.863	**0.904**	0.888
	Depth=5	0.452	0.856	0.916	**0.933**
	NatLang	0.573	0.579	**0.962**	0.594

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

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 8

RoBERTa-Large is fine-tuned on examples with different depths from PARARULES and also the entire PARARULE-Plus(PPT), and then is evaluated on test sets that require different depths of reasoning. The yellow background indicates improvement on accuracy after adding our PARARULE-Plus in the training process.

Model	Test set	Mod012 (Depth≤2+PPT)	Mod0123 (Depth≤3+PPT)	Mod0123Nat (Depth≤3+NatLang+PPT)	Mod012345 (Depth≤5+PPT)
RoBERTa-Large	Depth=0	0.946	0.901	0.965	**0.963 (+0.010)**
	Depth=1	0.877	0.847	**0.937 (+0.004)**	0.881
	Depth=2	0.868	0.873	**0.927**	0.839
	Depth=3	0.771 (+0.209)	0.862	**0.904**	0.826
	Depth=4	0.675 (+0.194)	0.852	**0.897**	0.832
	Depth=5	0.661 (+0.209)	0.888 (+0.032)	0.923 (+0.007)	**0.934 (+0.001)**
	NatLang	0.557	0.593 (+0.014)	**0.970 (+0.008)**	0.649 (+0.055)

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

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

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

Authored by: **Qiming Bao[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 Findings of ACL 2024

https://aclanthology.org/2024.findings-acl.353/

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

ACL 2024
Bangkok, Thailand

Outline

- Research Gap
  - Our Contribution
- System Architecture
- Experiment Results

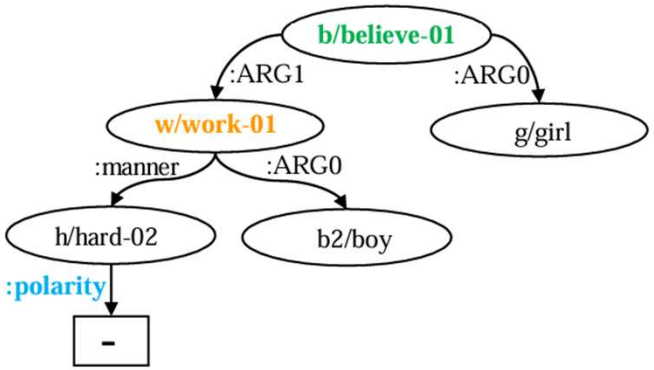
Research Gap

- 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].

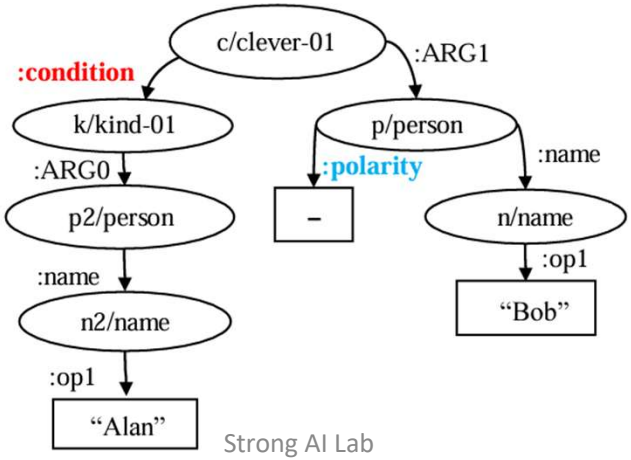
[1] Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.
[2] Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., ... & Srivastava, S. (2020). Logical neural networks. arXiv preprint arXiv:2006.13155.
[3] Bansal, A., Schwarzschild, A., Borgnia, E., 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.

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: ¬ α → ¬ β, ¬ β → ¬ γ
Option A: ¬ γ → ¬ α
√ Option B: γ → α + (β → α, γ → β) using contraposition law
Option C: ¬ γ → ¬ β
Option D: α → γ

Convert the natural language into logic symbols.

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

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

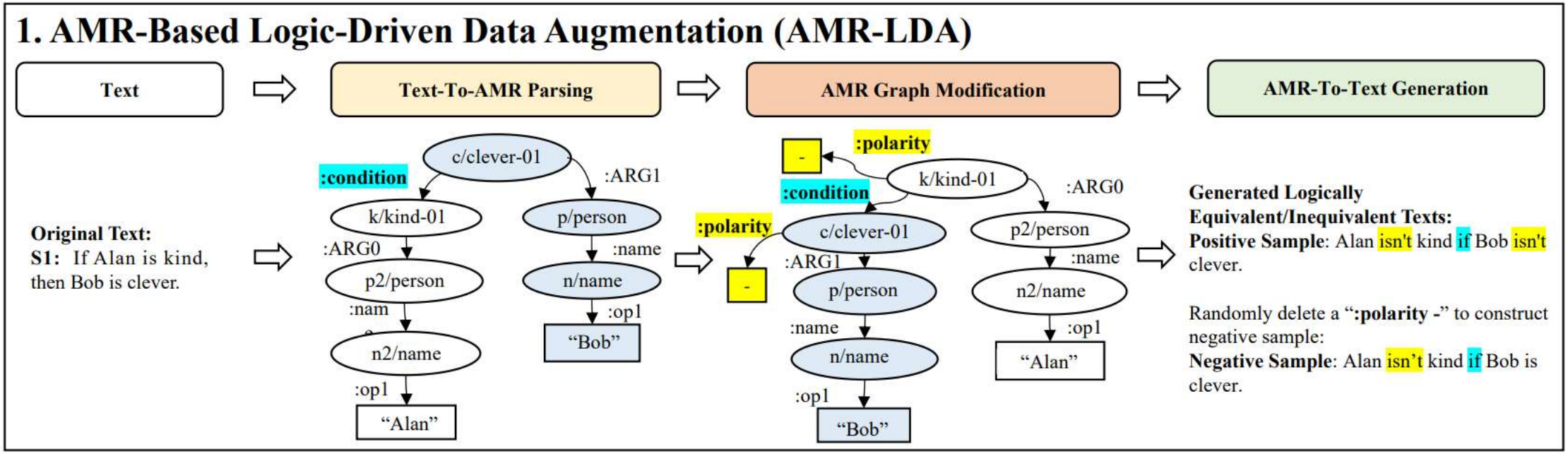
Alan is kind and Bob is clever. $\Leftrightarrow$ *Bob is clever and Alan is kind.*

Definition 4: Double negation law

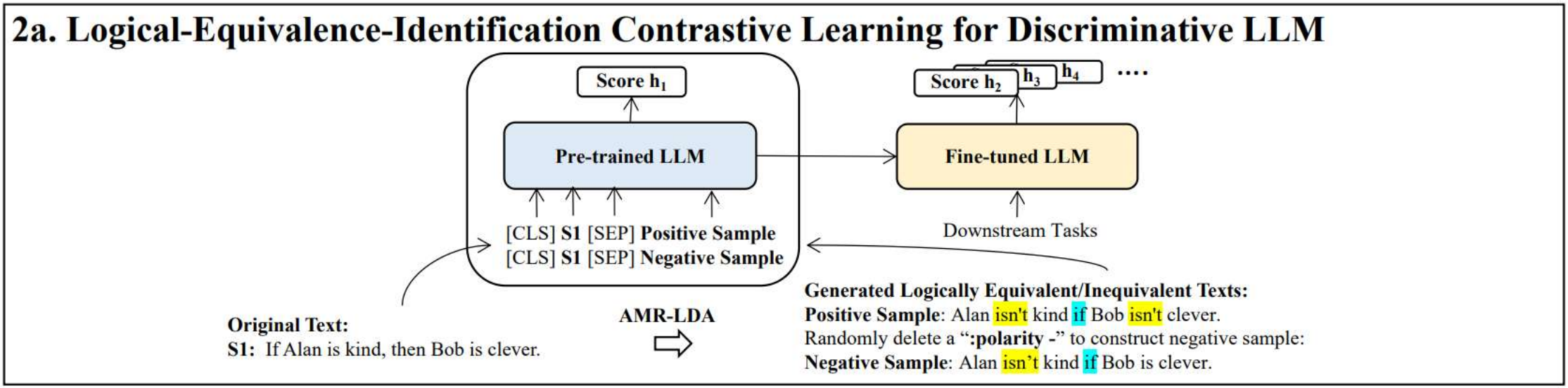
$$\mathcal{A} \Leftrightarrow \neg\neg\mathcal{A}$$

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

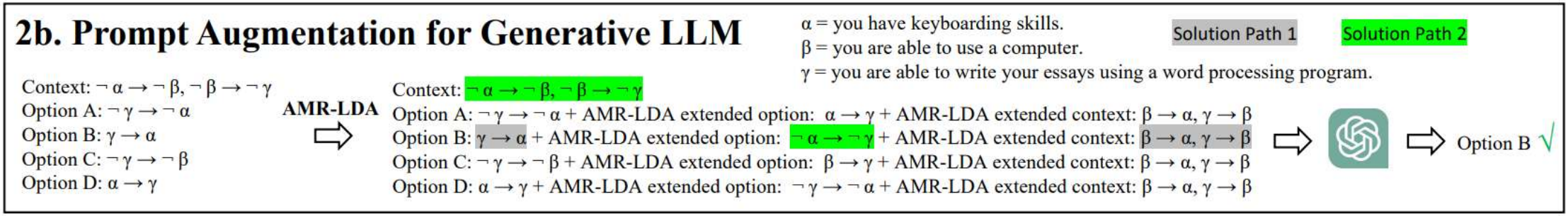
System Architecture



System Architecture



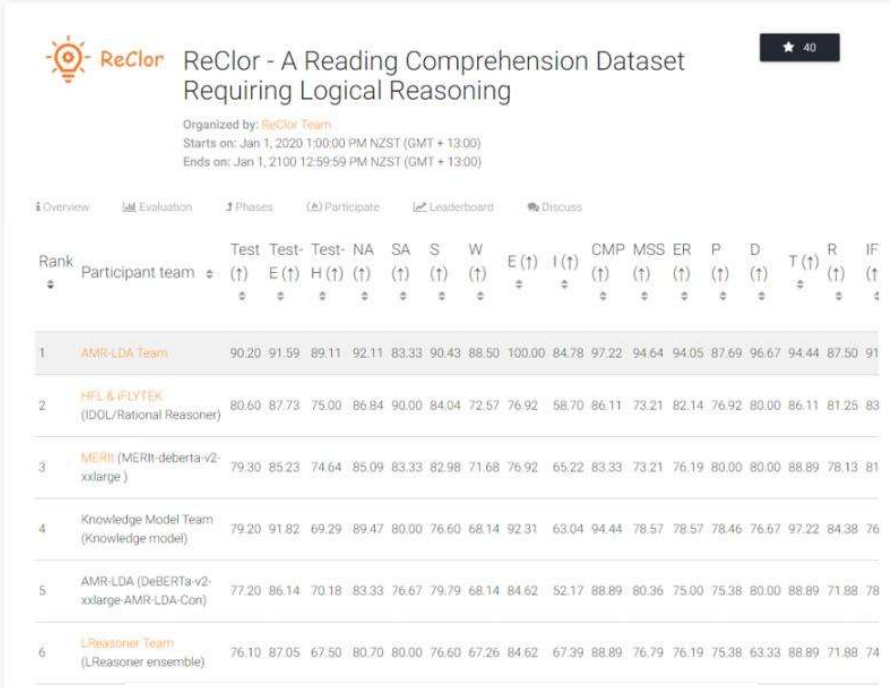
System Architecture



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	**94.73**	90.89
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	**43.80**	36.45	37.22	89.74	90.44	86.28	94.42	92.06
RoBERTa AMR-LDA	**65.26**	**56.86**	**77.34**	40.77	**40.29**	**38.14**	**89.78**	**90.93**	**86.64**	94.49	**93.14**
DeBERTaV2	73.93	70.46	80.82	62.31	39.72	39.62	89.45	89.71	84.48	95.00	**92.54**
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	**89.92**	89.71	83.39	95.02	92.42
DeBERTaV2 AMR-LDA	**79.40**	**77.63**	**85.75**	**71.24**	**42.34**	**39.88**	89.67	**90.20**	**88.09**	**95.24**	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.



Models/Datasets	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
GPT-3.5	57.02	56.20	59.31	**53.75**	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	**40.55**	31.49
+ AMR-LDA	**58.62**	**56.69**	**60.90**	53.39	**40.55**	**39.47**
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	**87.73**	**90.20**	**91.59**	**89.11**	**51.92**	**58.06**

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.

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

Experiment Results

Models/Datasets	RoBERTa AMR-LDA	RoBERTa LReasoner-LDA
Depth=1	100.00	100.00
Depth=1 (with altered rules)	**100.00**	99.87
Depth=2	100.00	100.00
Depth=2 (with altered rules)	**99.73**	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.

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

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	**84.77**	69.28	40.55	41.47
+ AMR-LDA-1:2	80.20	**76.40**	**84.77**	**69.82**	**47.00**	**43.93**
+ AMR-LDA-1:3	**81.20**	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	**82.60**	76.60	86.13	**69.10**	**45.00**	43.01
+ AMR-LDA-Merged-1:3	81.80	**76.90**	**87.50**	68.57	44.54	**45.62**
DeBERTaV2-XXLarge + IDoL	77.60	74.50	82.95	67.85	39.78	40.24
+ AMR-LDA-Con-1:3	79.20	**77.00**	85.68	**70.17**	**47.61**	**44.54**
+ AMR-LDA-Merged-1:3	**79.40**	75.60	**86.36**	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
RoBERTa-Large as backbone model				
ReClor	60.40	60.80	**61.80**	59.80
LogiQA	37.78	33.17	33.94	**38.70**
MNLI	89.55	**90.15**	89.68	89.78
MRPC	90.69	89.22	90.44	**90.93**
RTE	81.23	85.20	84.84	**86.64**
QNLI	94.16	94.05	**94.51**	94.49
QQP	92.12	89.88	92.06	**93.14**
DeBERTaV2-XXLarge as backbone model				
ReClor	**81.80**	72.20	79.40	78.80
LogiQA	32.25	**45.46**	38.24	40.55
DeBERTa-Large as backbone model				
MNLI	**90.80**	90.59	90.68	89.67
MRPC	**90.20**	88.48	89.95	**90.20**
RTE	84.84	87.36	85.56	**88.09**
QNLI	**95.28**	95.04	94.97	95.24
QQP	92.33	92.40	92.29	**92.47**

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.

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

A Systematic Evaluation of Large Language Models on Out-of-Distribution Logical Reasoning Tasks

Authored by: **Qiming Bao[1,2], Gaël Gendron[1], Alex Yuxuan Peng[1], Wanjun Zhong[3], Neşet Tan[1], Yang Chen[1], 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 and Engineering, Sun Yat-Sen University, China

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

https://arxiv.org/abs/2310.09430

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

Outline

- Research Gap
- System Architecture
- Experiment Results

Research Gap

- Although large language models such as ChatGPT and GPT-4 perform well on public logical reasoning leaderboards like ReClor, does this mean these models possess strong logical reasoning abilities?
- **Our Contribution:**
- We find that existing large language models, like ChatGPT and GPT-4, perform well on the original publicly available logical reasoning datasets. However, their performance on our out-of-distribution test examples is poor, suggesting that the models might have seen these datasets during training and have failed to acquire generalized logical reasoning capabilities.
- We present a new set of challenging logical reasoning datasets, which involve shuffling the order of options and changing the correct option to "none of the other options are correct." These datasets have been collected by OpenAI/evals [1].

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

[1] https://github.com/openai/evals/pull/648

Research Gap

- How to improve large language models' performance on complex logical reasoning tasks?
- **Our Contribution:**
- Relying solely on prompting techniques, such as chain-of-thought prompting, does not provide significant benefits for large language models when handling our newly presented, challenging logical reasoning tasks.
- We found that logic-driven data augmentation for instruction fine-tuning and prompting can help models improve their general performance on task structure variations in logical reasoning.

OOD 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.

1. Shuffle Option Order

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 not able to write your essays using a word processing program, you are not able to use a computer.
C. If you are able to write your essays using a word processing program, you have at least some keyboarding skills. ✓
D. If you have some keyboarding skills, you will be able to write your essays using a word processing program.

2. Replace the correct option

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. None of the other options are correct. ✓
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.

3. Shuffle and replace the correct option

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 not able to write your essays using a word processing program, you are not able to use a computer.
C. None of the other options are correct. ✓
D. If you have some keyboarding skills, you will be able to write your essays using a word processing program.

System Architecture

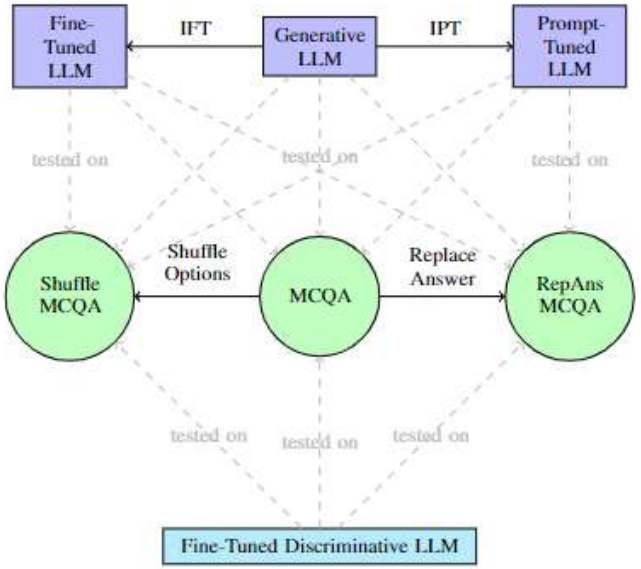


Figure 2: We perform instruction fine-tuning and prompting on the generative large language models and test them on several datasets. We also test fine-tuned discriminative large language models. We use Multiple-Choices Question Answering (MCQA) datasets and generate new distributions by shuffling the order of options and removing some answers. Square represent models, blue square represent generative models and cyan square represent classification engines, and green circles represent datasets.

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

Experiment Results

Datasets →	ReClor				LogiQA				LogiQAv2			
Models ↓	Original	Shuffle Order	Replace Answer	Shuffle RepAns	Original	Shuffle Order	Replace Answer	Shuffle RepAns	Original	Shuffle Order	Replace Answer	Shuffle RepAns
Zero-shot evaluation												
Alpaca-7B	0.0020	0.0060	0.0100	0.0120	0.0122	0.0122	0.0107	0.0121	0.0216	0.0165	0.0095	0.0121
Vicuna-7B	0.0960	0.1120	0.0740	0.0640	0.2027	0.2135	0.1735	0.1784	0.0834	0.0618	0.0541	0.0121
GPT-3.5	0.5702	0.5734	0.1919	0.1847	0.3763	0.3946	0.2449	0.2583	0.5094	0.2695	0.2675	0.2583
GPT-4	0.8735	0.8405	0.1454	0.1312	0.4324	0.5283	0.1007	0.1686	0.5230	0.2616	0.1731	0.1686
ReClor/LogiQA/LogiQAv2 single training set												
Alpaca-7B-IFT	0.1680	0.5280	0.2360	0.2720	0.1105	0.3486	0.2841	0.2273	0.1912	0.2122	0.3658	0.1548
Vicuna-7B-IFT	0.3040	0.1760	0.0320	0.0420	0.2503	0.1689	0.0706	0.1198	0.1899	0.1746	0.1797	0.1784
LReasoner	0.7320	0.7100	0.2320	**0.3420**	0.4147	0.4316	**0.5176**	**0.5176**	0.5685	0.5685	**0.4263**	**0.4263**
MERIt	0.7960	0.7960	0.2580	0.2460	0.3794	0.3809	0.2657	0.2703	0.7144	0.7144	0.1873	0.1873
AMR-LE	0.8120	0.8120	**0.3360**	0.3360	0.4270	0.4301	0.1720	0.1720	0.6985	0.6978	0.1440	0.1440
ReClor + LogiQA + LogiQAv2 merged training set												
Alpaca-7B-IFT	0.7100	0.6560	0.1380	0.1140	0.6651	0.4854	0.2718	0.1351	0.6411	0.2160	0.1956	0.1128
Vicuna-7B-IFT	0.3900	0.4040	0.1500	0.1060	0.5453	0.3840	0.2273	0.1490	0.4913	0.1816	0.1708	0.1121
MERIt	0.9660	0.9660	0.2440	0.2440	0.7311	0.7342	0.2119	0.2119	0.8655	0.8661	0.2625	0.2625
AMR-LE	**0.9700**	**0.9700**	0.2900	0.2900	**0.7557**	**0.7588**	0.2549	0.2549	**0.8744**	**0.8744**	0.3212	0.3212

Table 4: Accuracy of large language models on logical reasoning tasks. The first block represents generative large language models tested in zero-shot settings. We compare them against models improved with instruction fine-tuning (IFT) on various training sets (separate training sets for the second block and merged training set for the third block). In the second block, models are fine-tuned on the original training dataset as they are evaluated on (e.g. fine-tuned on original ReClor training set and evaluated on ReClor validation set and other validation sets). In the third block, models are fine-tuned on a merged training set composed of all original training sets without our new datasets. Alpaca-7B and Vicuna-7B are trained using IFT fine-tuning and LReasoner, MERIt and AMR-LE are fine-tuned in the standard way.

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

Experiment Results

Models	ReClor Shuffle RepAns	LogiQA Shuffle RepAns	LogiQAv2 Shuffle RepAns
Zero-shot evaluation			
Alpaca-7B	0.0120	0.0230	0.0121
Alpaca-7B-CoT	0.0120	0.0337	0.0152
Vicuna-7B	0.0640	0.1797	0.1784
Vicuna-7B-CoT	0.1320	0.1674	0.1593
GPT-3.5	**0.1847**	0.2286	**0.2583**
GPT-3.5-CoT	0.1088	0.1674	0.1722
GPT-4	0.1312	0.1626	0.1686
GPT-4-CoT	0.1816	**0.2523**	0.2177

Table 5: Comparison between base models and models prompted using Chain-of-Thought (CoT).

Models	ReClor Shuffle RepAns	LogiQA Shuffle RepAns	LogiQAv2 Shuffle RepAns
Zero-shot evaluation			
Alpaca-7B	0.0120	0.0121	0.0121
GPT-3.5	0.1847	0.2583	0.2583
GPT-4	0.1312	0.1686	0.1686
ReClor/LogiQA/LogiQAv2 single training set			
Alpaca-7B-IFT	0.2720	0.2273	0.1548
+ AMR-LE	0.0440	0.0522	0.0548
ReClor + LogiQA + LogiQAv2 merged training set			
Alpaca-7B-IFT	0.1140	0.1351	0.1128
+ AMR-LE	0.0060	0.0245	0.0197
Prompt augmentation using AMR-LE			
Alpaca-7B-IPT-LDA	0.0300	0.0368	0.0331
Alpaca-7B-IFT-LDA	0.4800	0.3686	0.2237
GPT-3.5-IPT-LDA	0.3667	0.4685	0.4971
GPT-4-IPT-LDA	**0.8766**	**0.5510**	**0.7027**

Table 6: Accuracy of evaluated models when adding AMR-LE's logic-driven augmented data into the training set. We evaluate Alpaca-7B after instruction fine-tuning.

Experiment Results

Datasets →	ReClor				LogiQA				LogiQAv2			
Perturbation Ratio ↓	Original	Shuffle Order	Replace Answer	Shuffle RepAns	Original	Shuffle Order	Replace Answer	Shuffle RepAns	Original	Shuffle Order	Replace Answer	Shuffle RepAns
ReClor/LogiQA/LogiQAv2 single training set with different ratio of data perturbation (Shuffle-RepAns)												
0%	0.1680	**0.5280**	0.2360	0.2720	0.1105	**0.3486**	0.2841	0.2273	0.1912	**0.2122**	0.3658	0.1548
5%	0.3340	0.3720	0.1560	0.1720	0.1490	0.1351	0.0998	0.0921	0.2695	0.1516	0.1338	0.1121
10%	**0.4140**	0.4320	0.2040	0.2380	**0.3072**	0.2826	0.2350	0.2442	0.2262	0.0956	0.1963	0.1727
15%	0.3620	0.3860	**0.3060**	**0.3340**	0.1904	0.2027	0.2795	0.2319	**0.3537**	0.1778	0.2001	0.1727
50%	0.1540	0.1400	0.1660	0.1640	0.0430	0.0537	**0.6728**	**0.6559**	**0.3537**	0.2096	**0.7686**	**0.7915**

Table 7: Accuracy of Alpaca-7B model for transfer learning scenarios and different perturbation ratio applied to the training set. To make a fair comparison, We ensure that the size of each training set is consistent.

Datasets	Train	Validation	Test
ReClor	4638	500	1000
LogiQA	7376	651	651
LogiQA-v2	12567	1569	1572

Table 5: Number of samples in the training, validation, and test set, for ReClor, LogiQA and LogiQA-v2.

Models	ReClor Shuffle RepAns	LogiQA Shuffle RepAns	LogiQAv2 Shuffle RepAns
Zero-shot evaluation			
LLaMA-7B	**0.1260**	0.1167	0.1128
LLaMA-13B	0.0660	0.1167	0.1013
LLaMA-30B	0.0360	0.1290	**0.1172**
LLaMA-65B	0.0720	**0.1397**	0.1159

Table 8: Comparison between multiple LLaMA model sizes on logical reasoning tasks with structure variations.

Useful Links

Project Code

Strong AI Lab

LIU AI Lab

LR-MRC-Plus

Our proposed three logical reasoning reading comprehension datasets (ReClor-Plus, LogiQA-Plus and LogiQA-v2-plus have been collected by OpenAI Evals)

Welcome for more discussion and collaboration!

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

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 Peng, Gaël Gendron, Wanjun Zhong, Tim Pistotti, Paul Denny, Michael Witbrock, Jiamou Liu.** ***Exploring Iterative Enhancement for Improving Learnersourced with Large Language Models*****, The Proceeding of AAAI 2025 (AAAI-EAAI 2025) [Paper link]**

•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]**