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

Speaker: Qiming Bao

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

21st November 2023

Invited Talk at the NLP Group, The University of Melbourne, Australia

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

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.
- 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.

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

Authored by: **<u>Qiming Bao</u>[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

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

Symbolic Logic Programs

- ***Symbolic logic*** expresses logical statements and expressions in symbols and variables instead of natural language.
- An example of logic programs expressed in Prolog [1]

$$p(X) : -q(X).$$
$$q(a).$$

$p(X)$, where variables are notated in capital letters.

$q(a)$, where constants are in lower case.

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

[1] Programming in Prolog: Using the ISO standard, Clocksin, 2012

Symbolic Logic Programs

1: Facts

$e(l).$

$?e(l).\ 1$

$?i(d).\ 0$

2: Unification

$o(V,V).$

$?o(d,d).\ 1$

$?o(b,d).\ 0$

3: 1 Step

$p(X) :- q(X).$

$q(a).$

$?p(a).1$

$?p(b).0$

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

[1] Cingillioglu, N. et al., 2018. DeepLogic: Towards End-to-End Differentiable Logical Reasoning, AAAI-MAKE19.

Natural Language Reasoning

- In natural language reasoning, logical statements are expressed in natural language instead of symbols.
  - **The semantics of logic**, such as propositional logic and first-order logic.
  - **Diversity and flexibility of natural language**, such as polysemy, a paraphrase of sentences.
  - Reasoning obtain unknown information based on existing information.

    Deductive reasoning: Given premise and rules to derive the conclusion.

    Inductive reasoning: Given premise and conclusion to derive rules.

    Abductive reasoning: Given rules and conclusion to derive premise.

    More examples can be found in [1] and [2].

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki 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.
[2] Young, N. et al. 2022. AbductionRules: Training Transformers to Explain Unexpected Inputs. The finding of ACL.

Example for Natural Language Reasoning

(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]**

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

[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Example for Natural Language Reasoning

(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]**

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

[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Example for Natural Language Reasoning

(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]**

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

[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Example for Natural Language Reasoning

(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]**

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

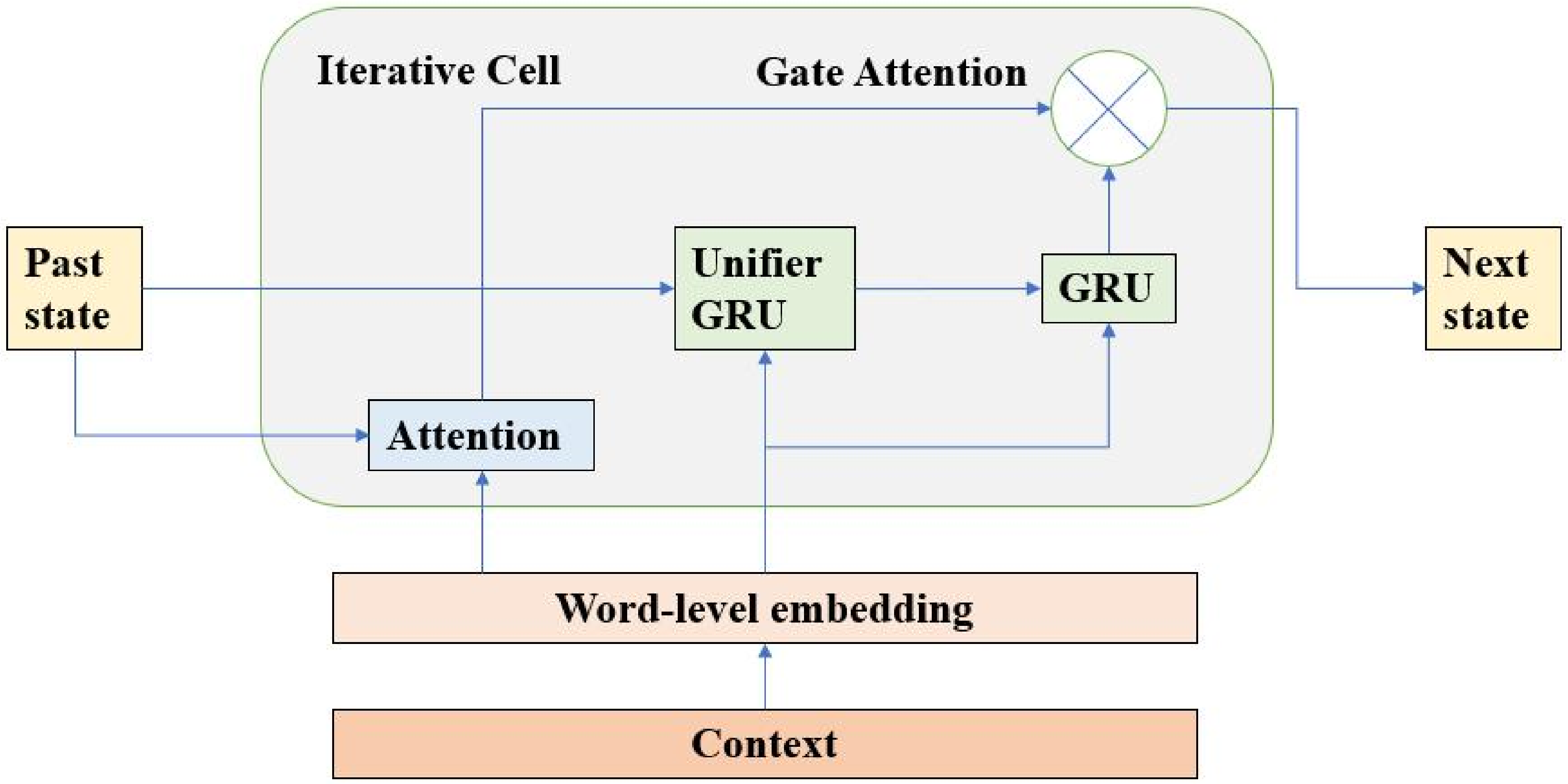
[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Research Gap

- Existing models, including DeepLogic and other RNN-based baseline models, have room for improvement in their reasoning abilities over natural language.
- We found existing models are not good at out-of-distribution (OOD) generalisation, in three scenarios:
  - When the model is trained on data with shallow reasoning depths and tested on data with deeper reasoning depths.
  - When the model is trained on synthetically generated data and tested on data paraphrased by human.
  - When the model is trained on unshuffled data and tested on shuffled data.
- Existing multi-step deductive reasoning datasets like PARARULES and CONCEPTRULE V1 and V2 have unbalanced distributions over the reasoning depths. Only a small portion of the datasets require deep reasoning (2<=Depth<=5).

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

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.

Word-level Embedding

- The input to the network consists of a context and a statement.
- The input sequence is represented using GloVe [1] word embeddings.
- The concatenated representations of context and statement will be fed into the gated recurrent unit (GRU).

[1] Pennington, et al., 2014. Glove: Global vectors for word representation, EMNLP.

Iteration

- The iteration process is from the DeepLogic [1]. The iteration step consists of attending to the rules, computing a new state using each rule and the old state.
- To apply a rule, we use another recurrent neural network called the **inner GRU unifier** that processes every literal of a given rule. The inner GRU unifier needs to learn unification between **variables** and **constants** as well as how each rule interacts with the current state.

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

[1] Cingillioglu, N. et al., 2018. DeepLogic: Towards End-to-End Differentiable Logical Reasoning, AAAI-MAKE19.

Gate Attention

- Dynamic Memory Network+ [1] achieved 100% test accuracy by using gate attention on bAbI deductive reasoning task (Task-15), which gave us the idea of integrating Gate Attention into DeepLogic. GRU can use gate attention to update the internal state.

[1] Xiong, C., et al., 2016. Dynamic Memory Networks for Visual and Textual Question Answering, ICML.

Established Baselines - RNNs & PLM

- We have three baseline models that we borrowed from the bAbI task leaderboard. We also set DeepLogic as one of the baseline methods, and then we have a Transformer-based model RoBERTa-Large as a baseline model. We use glove.6B.zip [4] as the word vector representation for the RNN-based models.
  - Long short-term memory (LSTM, 1997) [1] (The baseline method on bAbI dataset),
  - Dynamic Memory Network (DMN, 2016) [2] (One of the first paper use Attention in the memory network),
  - Memory Attention Control networks (MAC, 2018) [3] (A classical method from memory network).

[1] Hochreiter, et al. 1997. Long short-term memory,
[2] Kumar, et al. 2016. Ask me anything: Dynamic memory networks for natural language processing, ICML
[3] Hudson, et al. 2018. Compositional attention networks for machine reasoning, ICLR.
[4] Pennington, et al. 2014. Glove: Global vectors for word representation, EMNLP.
[5] Liu, Y. et al., 2019. Roberta: A robustly optimized bert pretraining approach. arxiv.

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

CONCEPTRULE vs CONCEPTRULE V2

(*Input Context:*) Book is not located in bed.
Bed is located in loft.
Loft is located in city.
City is located in fast-food restaurant.
Question 1: Book is located in loft. True/False? [**Answer: T**]
Question 2: Bed is located in city. True/False? [**Answer: T**]
Question 3: Book is located in bed. True/False? [**Answer: F**]

(*Input Context:*) Book is not located in bed.
Bed is located in loft.
Loft is located in city.
City is located in fast-food restaurant.
Question 1: Book is not located in bed. True/false? [**Answer: T**] [**Depth: 0**]
Question 2: Book is not located in loft. True/false? [**Answer: T**] [**Depth: 1**]
Question 3: Book is not located in city. True/false? [**Answer: T**] [**Depth: 2**]

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

Hartill, T., CONCEPTRULE, https://drive.google.com/file/d/1lxoAvtcvqVCYiO8e3tENnrTQ1NNVtpjs/view
Hartill, T., CONCEPTRULE V2, https://drive.google.com/file/d/1lOCbW8bfZxj1RIzKDxn8xKg99XyYNj7z/view

Dataset Description

Table 2

Information about the datasets used in this paper. PARARULES has less number of examples that require deep reasoning steps. CONCEPTRULES V2 does not consider reasoning depths greater than 3. The train, dev and test set are already splitted by the author of each dataset.

Dataset		Depth=0	Depth=1	Depth=2	Depth=3	Depth=4	Depth=5
PARARULES	Train	290435	157440	75131	48010	9443	7325
	Dev	41559	22276	10833	6959	1334	1038
	Test	83119	45067	21496	13741	2691	2086
PARARULE-Plus	Train	-	-	89952	90016	90010	90022
	Dev	-	-	16204	16154	16150	16150
	Test	-	-	2708	2694	2704	2692
CONCEPTRULES V2 (full)	Train	2074360	1310622	873748	436874	-	-
	Dev	115148	72810	48540	24270	-	-
	Test	115468	72810	48540	24270	-	-
CONCEPTRULES V2 (simplified)	Train	131646	74136	49424	24712	-	-
	Dev	7166	4116	2744	1372	-	-
	Test	7362	4116	2744	1372	-	-

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

Dataset Description

Table 3
The entity types and relation types for CONCEPTRULES V1 (simplified/full), CONCEPTRULES V2 (simplified/full), PARARULES, and our PARARULE-Plus.

Dataset	#Entity	#Relation	Shuffled Rules	Depth Tag	Derivable	NAF
CONCEPTRULES V1 (simplified)	385	7	No	No	Yes	Yes
CONCEPTRULES V1 (full)	4048	24	Yes	No	Yes	No
CONCEPTRULES V2 (simplified)	385	7	No	Yes	Yes	Yes
CONCEPTRULES V2 (full)	4048	24	Yes	Yes	Yes	Yes
PARARULES	19	4	No	Yes	Yes	Yes
PARARULE-Plus	71	8	No	Yes	Yes	Yes

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

A Sample for Negation as Failure (NAF)

(*Input Facts:*) The bear visits the lion.
The tiger likes the cat.
The cat does not like the bear.
The lion likes the tiger.
(*Input Rules:*) If someone sees the lion then the lion is kind.
If the tiger visits the lion and someone does not see the tiger then the tiger visits the bear.
If someone likes the bear and they like the tiger then the bear visits the tiger.
If someone is not round then they like the cat.
If someone visits the lion then they are blue.
If someone visits the bear and they do not see the lion then they visit the tiger.
If someone is cold and they do not visit the lion then the lion visits the tiger.
If someone visits the tiger and they are green then the tiger likes the cat.
Question 1: The bear likes the cat. True/false? [**Answer: T**]
Question 2: The bear is round. True/false? [**F**]
Question 3: The bear is not round. True/false? [**T**]

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

[1] Peter Clark, et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

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 Wananga o Tamaki 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

[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.

References

- Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.
- Brown, T. et al. 2020. Language Models are Few-Shot Learners. NIPS 2020.
- Chung, J. et al. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, NIPS 2014.
- Cingillioglu, N. et al., 2018. DeepLogic: Towards End-to-End Differentiable Logical Reasoning, AAAI-MAKE19.
- Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.
- Devlin, J. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2018.
- Hartill, T., CONCEPTRULE, https://drive.google.com/file/d/1lxoAvtcvqVCYiO8e3tENnrTQ1NNVtpjs/view
- Hartill, T., CONCEPTRULE V2, https://drive.google.com/file/d/1lOCbW8bfZxj1RIzKDxn8xKg99XyYNj7z/view
- Hochreiter, S. et al. 1997. LONG SHORT-TERM MEMORY. Neural computation, 9(8), 1735-1780.
- Hudson, et al. 2018. Compositional attention networks for machine reasoning, ICLR.

References

- J J Hopfield et al. Neural networks and physical systems with emergent collective computational abilities. PNAS 1982.
- Kumar, et al. 2016. Ask me anything: Dynamic memory networks for natural language processing, ICML
- Liu, Y. et al., 2019. Roberta: A robustly optimized bert pretraining approach. arxiv.
- Mikolov, T. et al. 2013. Efficient estimation of word representations in vector space.
- Young, N. et al. 2022. AbductionRules: Training Transformers to Explain Unexpected Inputs. The finding of ACL.
- Pennington, et al., 2014. Glove: Global vectors for word representation, EMNLP.
- Programming in Prolog: Using the ISO standard, Clocksin, 2012
- Radford, A. 2019. Language Models are Unsupervised Multitask Learners. OpenAI blog.
- Xiong, C., et al., 2016. Dynamic Memory Networks for Visual and Textual Question Answering, ICML.
- Vaswani, A. et al. 2017. Attention is all you need. NIPS.

Useful Links

Project code

LogiTorch.ai

PARARULE-Plus dataset collected by OpenAI/Evals

Our PARARULE-Plus has been collected by LogiTorch.AI and OpenAI/Evals, which is a tool that integrates different natural language logical reasoning models and a platform that collects datasets not effectively covered by ChatGPT-3.5/4.

Welcome for more discussion and collaboration!

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

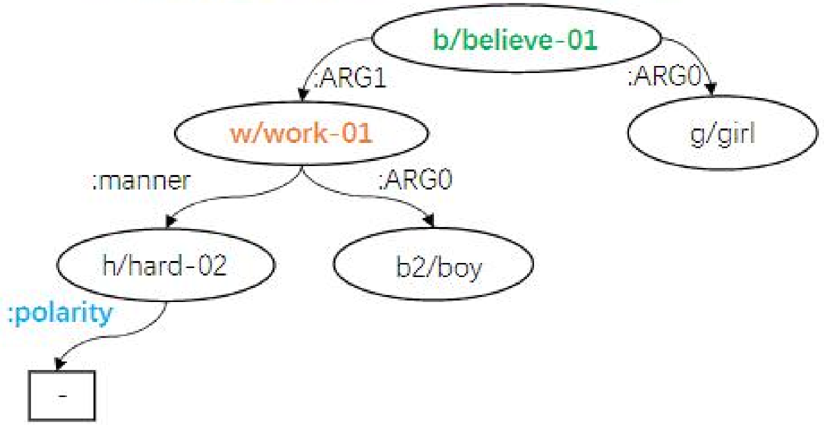
- Enabling pre-trained large language models (LLMs) to reliably perform logical reasoning is an important step towards strong artificial intelligence [1].The lack of available large real-world logical reasoning datasets means that LLMs are usually trained on more general corpora or smaller ones that 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.

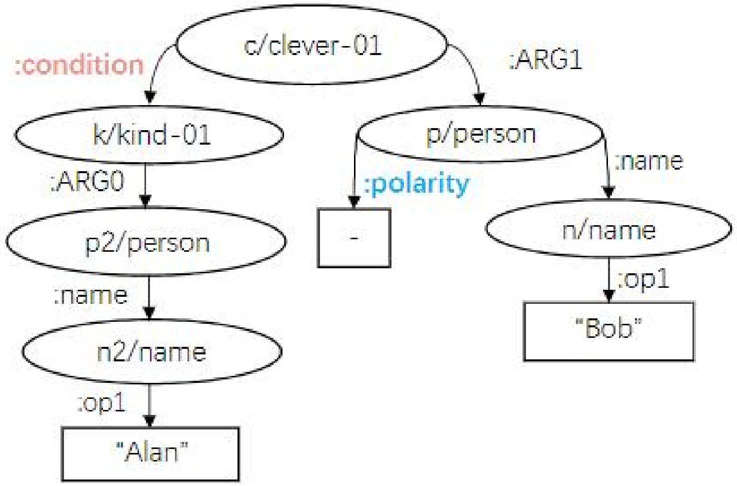
Abstract Meaning Representation

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.

https://whyu.me/reclor/

Logical Equivalence Laws

Definition 1: Contraposition law

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

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

Definition 2: Implication law

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

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.*

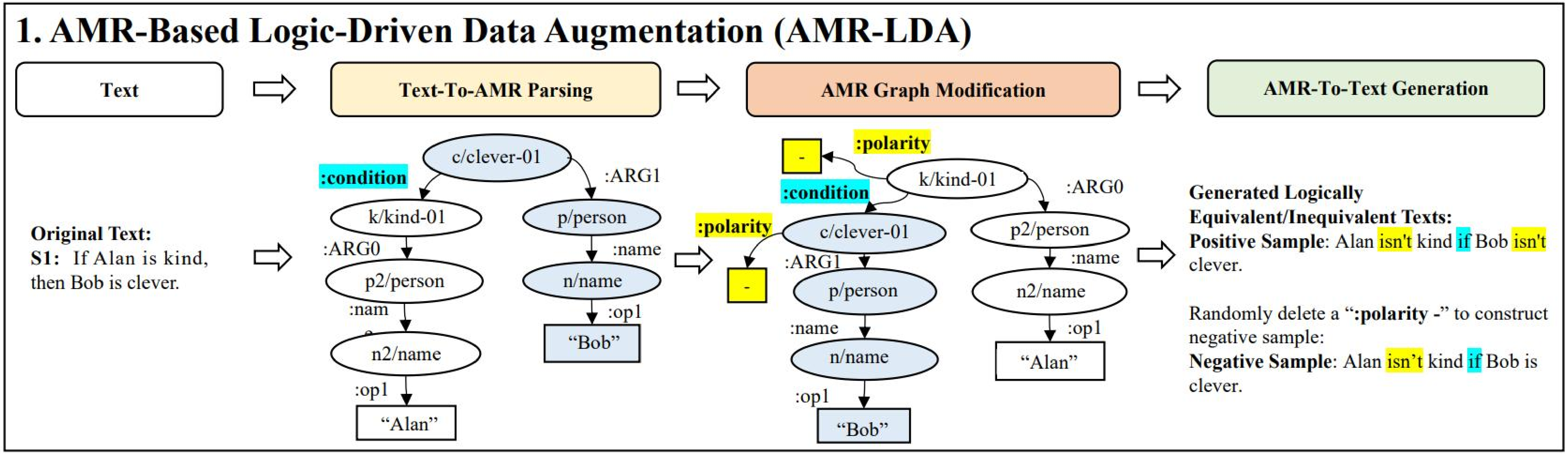
Definition 4: Double negation law

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

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

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

System Architecture



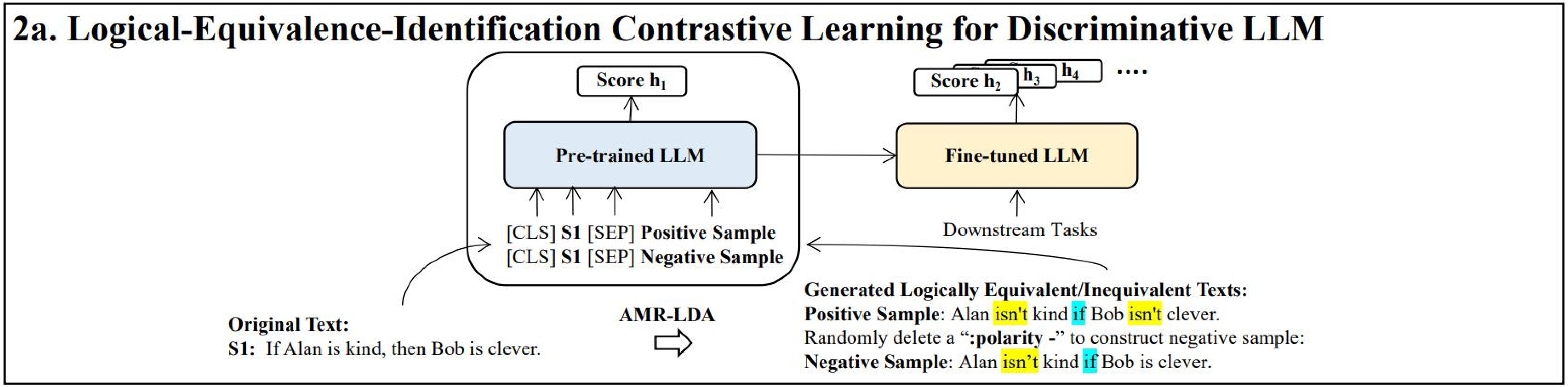
THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

Construct positive and negative samples

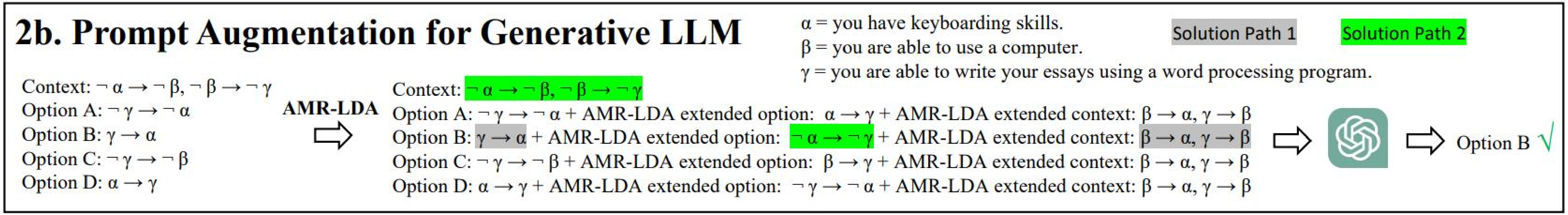
Original sentence	Positive sample	Negative sample
If Alan is kind, then Bob is clever.	Alan isn't kind if Bob isn't clever.	Alan isn't kind if Bob is clever.
	Alan is not kind or Bob is clever.	Alan is kind or Bob is clever.
The bald eagle is strong.	The bald eagle is not weak .	The bald eagle is weak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



System Architecture



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

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 Wananga o Tamaki 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.

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

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 Wananga o Tamaki 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 ↓ / Models →	Con	Con-dou	Con-dou imp	Con-dou imp-com
	Dev acc			
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.

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

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)

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 Wananga o Tamaki Makaurau
NEW ZEALAND

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

Authored by: **<u>Qiming Bao</u>[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

- Task Description
- System Architecture
- Experiment Results
- Case Studies
- Conclusion and Future Work

Research Gap

- 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 failed to acquire generalised logical reasoning capabilities.

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

Convert the natural language into logic symbols.

https://whyu.me/reclor/

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

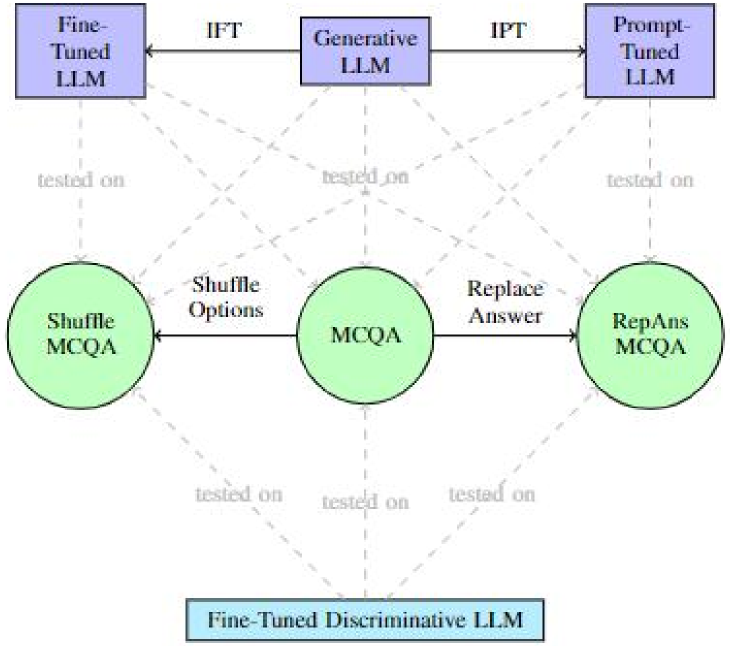


Figure 2: We perform instruction fine-tuning and prompting on the generative large language models and test them on several datasets. We also test fine-tuned discriminative large language models. We use Multiple-Choices Question Answering (MCQA) datasets and generate new distributions by shuffling the 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.

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 → Perturbation Ratio ↓	ReClor Original	ReClor Shuffle Order	ReClor Replace Answer	ReClor Shuffle RepAns	LogiQA Original	LogiQA Shuffle Order	LogiQA Replace Answer	LogiQA Shuffle RepAns	LogiQAv2 Original	LogiQAv2 Shuffle Order	LogiQAv2 Replace Answer	LogiQAv2 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.

Case Studies

A Case Study in Shuffle-RepAns ReClor dataset

Instruction: Can you predict the correct option for the given input?
Input: Some biologists believe that the capacity for flight first developed in marine reptiles, claiming that feathers are clearly developed from scales. Other biologists rightly reject this suggestion, pointing out that bats have no scales and that nonmarine reptiles also have scales. Those who believe that flight first developed in tree-dwelling reptiles reject the claim that the limbs of land-dwelling reptiles might have developed into wings. They insist that it is more likely that tree-dwelling reptiles developed wings to assist their leaps from branch to branch.
Question: Which one of the following most accurately describes the role played in the passage by the claim that nonmarine reptiles have scales? Describe every step of your reasoning before proposing a solution. When giving the solution, start your sentence with 'ANSWER:'. The 'ANSWER' is the letter in front of your predict correct option, A, B, C or D.
A. It weakens the claim that tree-dwelling reptiles were the first kind of reptile to develop the capacity for flight.
B. It is cited as evidence against the claim that the capacity for flight first developed in land-dwelling animals.
C. It is cited as evidence against the claim that the capacity for flight first developed in treedwelling reptiles.
D. D. None of the other options are correct.
Correct Answer: "D"
GPT-4: "B"
Chain-of-Thought Prompting GPT-4: The claim that nonmarine reptiles have scales is mentioned in the passage to argue against the idea that flight first developed in marine reptiles. The reasoning behind this is that if both marine and nonmarine reptiles have scales, then the development of feathers from scales cannot be used as evidence to support the idea that flight originated in marine reptiles.
Now let's analyze each option:
A. The claim about nonmarine reptiles having scales does not directly address tree-dwelling reptiles or their development of flight, so this option is incorrect.
B. The claim is not used as evidence against the idea that flight first developed in land-dwelling animals. Instead, it is used to argue against the notion that flight first developed in marine reptiles. So, this option is incorrect.
C. The claim about nonmarine reptiles having scales is not used to argue against the idea that flight first developed in tree-dwelling reptiles. It is used to address the marine reptiles hypothesis. So, this option is incorrect.
D. Since none of the other options accurately describe the role played by the claim that nonmarine reptiles have scales, this option is correct.
ANSWER: D

Figure 3: Compared GPT-4 and Chain-of-Thought GPT-4 on Shuffle-RepAns ReClor dataset. GPT-4 fails on this example but Chain-of-thought Prompting GPT-4 answer this example correctly. With the help of Chain-of-Thought Prompting, GPT-4 can find out the other options are incorrect and to select "None of the other options are correct." as the predicted answer.

A Case Study in Shuffle-RepAns ReClor dataset

Instruction: Can you predict the correct option for the given input?
Input: Advertisement: The dental profession knows that brushing with Blizzard toothpaste is the best way to fight cavities. We surveyed five dentists, and each agreed that the tartar control formula found in Blizzard is the most effective cavity-fighting formula available in a toothpaste.
Question: The flawed reasoning in which one of the following is most similar to the flawed reasoning in the advertisement? Describe every step of your reasoning before proposing a solution. When giving the solution, start your sentence with 'ANSWER:'. The 'ANSWER' is the letter in front of your predict correct option, A, B, C or D.
A. We know that electing Gomez would be the best course for the nation to follow because, of ten voters we surveyed, each agreed that electing Gomez would help the nation.
B. Some of the nation's voters believe that Gomez is the candidate who would be best for the nation. Of the ten voters we surveyed, each agreed that the policies Gomez is committed to would be the best policies for the nation to adopt.
C. The nation's voters know that Gomez is the candidate whose policies would be best for the nation. Of ten voters polled, each said that Gomez would be a very popular leader.
D. None of the other options are correct.
Correct Answer: "D"
GPT-4: "A"
Chain-of-Thought Prompting GPT-4: The reasoning in the advertisement is flawed because it is based on a small and potentially biased sample size (five dentists) to make a general claim about the dental profession's view on Blizzard toothpaste.
Let's analyze each option:
A. This option is flawed for the same reason as the advertisement: it's based on a small sample size (ten voters) to make a claim about the best course for the nation.
B. This option is also based on a small sample size (ten voters), but it doesn't make a general claim about the nation's voters. Instead, it only states that some voters believe Gomez is the best candidate.
C. This option is flawed because it makes a general claim about the nation's voters based on the popularity of Gomez among ten voters, not on the effectiveness of his policies.
D. This option states that none of the other options are correct.
The most similar flawed reasoning to the advertisement is found in option A, as it also uses a small sample size to make a general claim.
ANSWER: A

Figure 4: Compared GPT-4 and Chain-of-Thought GPT-4 on Shuffle-RepAns ReClor dataset. Both GPT-4 and Chain-of-Thought Prompting GPT-4 fail on this example. In this example, Chain-of-Thought try to predict an option which looks more like the correct answer.

Conclusion and Future Work

1. We found that the model only shows better i.i.d. and o.o.d. performance when the training set exceeds 10,000 and contains a relatively large proportion of leaked o.o.d data. Otherwise, the model does not exhibit improved performance on either i.i.d. or o.o.d. tasks. Moreover, we did not observe that larger model sizes offer better performance on out-of-distribution logical reasoning tasks.
2. We observed that for fine-tuned discriminative language models, context-option related, logic-driven data augmentation can enhance their performance in o.o.d logical reasoning tasks. However, context-option related logic-driven data augmentation did not help generative large language models with o.o.d logical reasoning tasks, potentially due to the limitations imposed by next-token prediction.
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

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!