LoRT: Logical Reasoning Evaluation Suite for Transformers

Wenshan Wu, Aditya Sarkar, Jenny Yarmovsky
Department of Computer Science
University of Maryland, College Park
{wwu009, asarkar6, jyarmovs}@umd.edu

Abstract

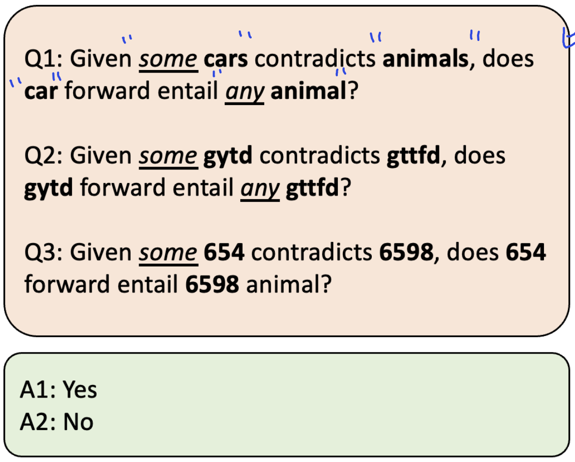
This study investigates how large language models (LLMs) develop logical reasoning, with a particular focus on textual entailment tasks involving logical quantifiers. While LLMs excel in many natural language processing tasks, they struggle with complex reasoning, especially when logical quantifiers like "not," "some," and "all" are involved. To address this, a novel dataset - LoRT is presented in this work that has a constrained vocabulary and prioritizes logical relationships over semantic complexity. A small-scale transformer model is trained on LoRT to explore its ability to learn logical entailment. A comparative analysis is then conducted between this model and fine-tuned BERT, allowing us to assess the role of semantic knowledge in logical reasoning tasks. Experiments on this dataset demonstrate that transformers with high complexity and world knowledge perform better on our dataset. [1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable proficiency across a diverse range of Natural Language Processing (NLP) tasks, often rivaling or surpassing human performance in areas such as summarization (Jin et al., 2024; Zhang et al., 2024), question answering (Liu et al., 2024a,b), machine translation (Faria et al., 2024; Chen et al., 2024), and text classification (Jiao et al., 2024; Kasa et al., 2024). Despite these advancements, their ability to tackle more complex cognitive tasks, such as commonsense reasoning (Petroni et al., 2019; West et al., 2021) and logical inference (Toroghi et al., 2024; Sun et al., 2024), remains inconsistent. In particular, logical reasoning tasks often expose significant limitations, especially when

dealing with out-of-distribution data (Liu et al., 2023; Jang and Lukasiewicz, 2023).

Logical reasoning is a cornerstone of human intelligence and a critical metric for evaluating genuine artificial intelligence. Yet, relatively little research has focused on LLMs' capacity to process logical quantifiers—key elements in reasoning about relationships between entities. This gap motivates our investigation into the logical reasoning abilities of LLMs, with a particular emphasis on their handling of logical quantifiers such as all, some, and none within textual entailment tasks.



Q1: Given *some* **cars** contradicts **animals**, does **car** forward entail *any* **animal**?

Q2: Given *some* **gytd** contradicts **gttfd**, does **gytd** forward entail *any* **gttfd**?

Q3: Given *some* **654** contradicts **6598**, does **654** forward entail **6598** animal?

A1: Yes
A2: No

Phrasal entailment - place in quotes? Or else use more natural sounding sentences (nonce-words aside)

Figure 1: An example from LoRT showing questions having real words, synthetic words and numbers respectively (above) with the possible binary answer options (below).

Textual entailment tasks provide a rigorous framework for assessing reasoning capabilities, as they require both a nuanced understanding of sentence semantics and the logical relationships between statements. This project focuses on evaluating LLMs' performance on such tasks, specifically their ability to reason correctly with logical quantifiers. By doing so, we aim to provide a robust benchmark for understanding and improving LLMs' logical reasoning capabilities.

Our contributions are:

1. LoRT: a large scale manually curated test set of 110K CommonsenseQA-style questions which can be used for probing the logical reasoning in natural language systems.
2. A systematic evaluation of several encoder based models on LoRT to assess the impact of semantic understanding and model complexity on logical

[1] The data is available at https://github.com/ws500981/LoRT/

reasoning performance.

2 Dataset

LoRT is designed to evaluate the logical reasoning capabilities of language models, following a structure similar to Commonsense QA (Talmor et al., 2018). Each question is framed as: `"Given that A [logical relation1] B, does [quantifier (opt)] {word 1} [logical relation2] [quantifier (opt)] {word 2}?"`, with a `label` where two logical relations need not be different, `{word 1}` can be `A` or `B`, and `{word 2}` corresponds with `{word 1}` (if `{word 1}` is `A`, `{word 2}` must be `B`, and vice versa). The `label` for each sample is its truth value ("Yes" or "No") based on logic depending on the logical relationship between the premise and hypothesis. Only one answer is correct. This format mirrors textual entailment tasks. We employ a template-based approach, asking multiple questions on the same theme with variations in specific elements (e.g., country, animal, or flower names). Subset-superset relations between words were used to guarantee correct logical relations. One example of template is: `Given some {word 1} forward entails {word 2}, does {word 1} reverse entail any {word 2}?`. Note that *some* and *any* are quantifiers.

LoRT is comprised of 1500 templates manually curated by the authors of this paper. From these templates, questions were automatically curated by selecting `{word 1}` and `{word 2}` either from a real-world dataset or by generating nonsensical words. Based on their origin, the questions in LoRT can be categorized into two main types:

- ✓ **Real-word dataset** (*RealLoRT*): In this case, both words in the template are common, real words, such as “cat” and “animal.” These words were sourced from WordNet (Fellbaum, 2010) and existing Kaggle datasets (Artem, 2019; Bouquin, 2021; Banerjee, 2022; Nilsback and Zisserman, 17) (see Fig. 1, Q1). Good
- ✓ **Synthetic-word dataset** (*SynLoRT*): Here, the words chosen have no lexical meaning, such as “abcde” and “abhdh,” or are numbers like “123637” and “23432” (see Fig. 1, Q2 and Q3).

The motivation behind having a synthetic dataset is to disentangle semantic understanding from logical reasoning and investigate whether the presence of numeric or semantic understanding influences logical reasoning capabilities.

RealLoRT consists of 27K questions, while SynLoRT contains 83K questions. Example questions from both categories are provided in Figure 1. The complete dataset (`LoRT`) was divided into training (`LoRT-train`) and testing (`LoRT-test`) subsets. In `LoRT-train`, questions containing real words are referred to as `RealLoRT-train`, while those with synthetic words are labeled as `SynLoRT-train`. The test set (`LoRT-test`) contains 1200 questions, with 600 questions from each of RealLoRT and SynLoRT. The former subset is called `RealLoRT-test` and the latter is called `SynLoRT-test`. All questions in the LoRT dataset consist of a single sentence. In total, LoRT consists of 110K questions.

Logical Reasons and Quantifiers In constructing our logical reasoning dataset for the textual entailment task, we employed three logical relationships: forward entailment representing "is-a" relationships, ✓ reverse entailment representing "include" relationships, ✓ and contradiction representing "not-a" relationships. ✓ This approach is grounded in the natural logic framework defined by MacCartney and Manning (2009a), from which we adopted three fundamental relations—forward entailment, reverse entailment, and negation—substituting contradiction for negation to accommodate nonbinary terms. Furthermore, we extended MacCartney and Manning (2009b)’s list of quantifiers and incorporated eight logical quantifiers, each aligned with its mathematical meaning. For instance, "not" denotes non-existence ($= 0$), while "few," "any," "one," "some," "most," and "each" signify existence (> 0), and "all" denotes universality ($= 1$). An example of these concepts is the relationship between "cat" and "animal." A cat forward entails an animal because a cat is an animal, and an animal reverse entails a cat because animals include cats. Further, if something is not an animal, it cannot be a cat. These relations thus follow logical equivalence: "cat *forward entails* animal" ↔ "animal *reverse entails* cat" ↔ "not animal *forward entails* not cat".

Annotation For SynLoRT, equality was defined as indicating contradiction. For subsets having numbers, the condition a < b implies that a forward-entails b. In subsets having synthetic words, forward entailment is determined by alphabetical order (e.g., ’absp’ < ’btbd’). For the RealLoRT, we derived contradiction relations from antonyms in WordNet, and entailment pairs from curated Kaggle dataset labels. In the case of entailment pairs, category labels were treated as supersets, with all items in the category considered subsets of the label.

Dataset Imbalance A significant challenge in our curated training dataset is the inherent bias toward negative answers ("No") in our logical templates due to the

Dataset	# Samples
RL-train	21285
RL-test	5321
SL-train	65227
SL-test	16307
TL	1200

Dataset subsets	# Unique Words	# Yes	# No	Total
RL	914	5891	20725	26616
SL (Words)	850	11825	42663	54488
SL (Numbers)	844	5810	21236	27046
TL	1841	600	600	1200

Table 1: Number of sentences for each subset of our LoRT dataset. #, RL, SL and TL denotes "number of", "RealLoRT", "SynLoRT" and "TestLoRT". **Left**: shows that number of samples in each subset of LoRT dataset. **Right**: shows that there is imbalance in the number of Yes and No in the LoRT dataset (except for the test subset).

use of truth values as the label. This imbalance could lead to biased learning or ineffective fine-tuning of the transformer models. To mitigate this, a biased sampler was employed during the creation of mini-batches for training, ensuring that each mini-batch contains an equal number of "Yes" and "No" responses. Although this approach results in the repetition of certain data samples, it can be viewed as a form of data augmentation. Furthermore, it was ensured that the test set comprises an equal number of "Yes" and "No" questions. This balanced approach will create a more rigorous evaluation that better reflects the models' actual logical reasoning capabilities rather than their ability to exploit dataset biases. An overview of the whole dataset can be found in Table 1.

Good

Validation Study A validation study was conducted to assess the quality and accuracy of the curated dataset, involving seven annotators, all senior undergraduates from the U.S. Given that the task aimed to evaluate logical reasoning—skills that are universally applicable—it was deemed acceptable for all annotators to be from the same region. The annotators were tasked with answering 200 questions, with 100 questions uniformly sampled from RealLoRT and the remaining 100 sampled in the same manner from SynLoRT. Final answers were determined by majority vote. The Pearson correlation between human responses and data labels was observed to be **98%** for the RealLoRT subset and **100%** for the SynLoRT subset.

3 Experiments

We summarize our experimental set up, models, and the evaluation strategy that will be used in this work.

Baseline Models - Eight encoder-based transformer models were evaluated on LoRT. We tested two variants of BERT (Kenton and Toutanova, 2019): bert-base and bert-large, two variants of RoBERTa (Liu et al., 1907): roberta-base and roberta-large, DistilBERT (Sanh et al., 2019) and two variants of DeBERTaV3 (He et al., 2021): deberta-v3-base and deberta-v3-large. All models were fine-tuned on a train fold for 5 epochs and then evaluated on a test fold. In most experiments, our train fold is RealLoRT-train, validation fold is RealLoRT-test and test fold is combined SynLoRT-train and SynLoRT-test, unless specified otherwise. Further, a simple encoder-based transformer (with 2 encoders and 4 heads per encoder) (Vaswani, 2017), followed by a FCNN classifier, was also trained on the RealLoRT train fold (RealLoRT-train) for 100 epochs, and then evaluated on a SynLoRT. Tiktoken, a fast byte pair encoding tokenizer, was used in this model. We run a grid search for the hyper-parameters for each case and report them on our GitHub page.

Metrics - Classification accuracy was used to evaluate all eight transformer models.

Implementation - All experiments were conducted on two Nvidia A6000 GPUs, using Pytorch. All the BERT models were implemented using the HuggingFace repository.

4 Results and Discussion

We present our experimental results with four key observations that offer valuable insights into the performance of transformer models on the proposed SynLoRT dataset.

First, **BERT-family models demonstrate strong performance on the SynLoRT dataset**. As illustrated in Figure 2 (top left), the performance of all BERT models is evaluated after fine-tuning on the RealLoRT dataset and subsequently testing on the SynLoRT dataset. The rationale for this approach is to first familiarize the models with the task using real-world data, enabling them to better generalize to synthetic examples. The results show that both BERT-base and BERT-large models achieve near-perfect accuracy, ranging from 99% to 100%. In contrast, Distil-BERT underperforms, with its accuracy on SynLoRT falling short by 17% compared to the other models. Additionally, it is noteworthy that all transformer models perform better on the RealLoRT validation set than on the synthetic subset, SynLoRT.

Second, **world knowledge significantly enhances the logical reasoning capabilities of transformer**

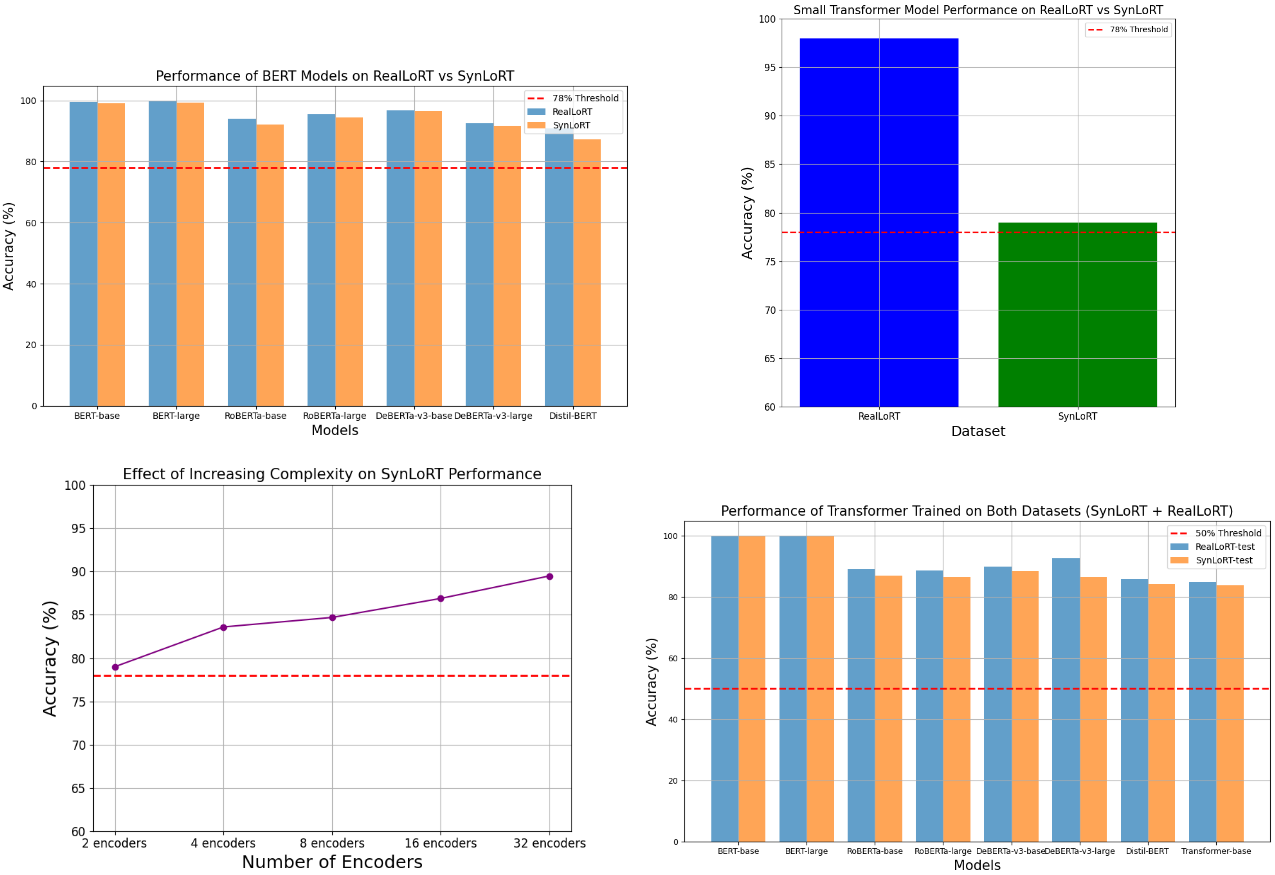


Figure 2: Results for experiments conducted. **Top-Left**: performance of BERT family models finetuned on RealLoRT and evaluated on SynLoRT. **Top-Right**: performance of base transformer model trained on RealLoRT and evaluated on SynLoRT. **Bottom-Left**: performance of base transformer model with different encoder numbers trained on RealLoRT and evaluated on SynLoRT. **Bottom-Right**: performance of all transformer models trained on RealLoRT+SynLoRT and evaluated on TestLoRT.

models, even when tested on synthetic datasets. To verify this hypothesis, we evaluated a base transformer model (Vaswani, 2017), which was trained on the `RealLoRT` data, on the `SynLoRT` dataset. As described in the previous section, the model's performance is shown in Figure 2 (top right). The results indicate a substantial drop in performance when the transformer model is tested on `SynLoRT`, suggesting that models pre-trained on large-scale real-world datasets, such as BERT, are better equipped to handle logical reasoning tasks. In contrast, the BERT family models, pre-trained on vast web-scale corpora, maintain accuracy rates up to 99% on `SynLoRT`, highlighting the importance of world knowledge in reasoning tasks.

Third, **increasing model complexity improves logical reasoning capabilities on synthetic datasets**. Figure 2 (bottom left) demonstrates that, as the number of encoder layers and attention heads in the transformer architecture increases beyond 16, the performance on the `SynLoRT` dataset improves. This result suggests that greater model complexity facilitates a better understanding of the logical structures inherent in the synthetic data. Notably, the model in question was trained on the `RealLoRT` dataset, emphasizing that increasing complexity is beneficial for tasks involving logical reasoning.

Fourth, **training transformer models on both `SynLoRT` and `RealLoRT` datasets leads to improved logical reasoning performance**. In Figure 2 (bottom right), we present the performance of a base transformer model that was trained on both `SynLoRT` and `RealLoRT` datasets, and evaluated on the `Test-LoRT`. When comparing this performance to that shown in Figures 2 (top-right and top-left), it is evident that the performance of transformers trained on both datasets surpasses that of the BERT models. Moreover, the base transformer model with just 2 encoder layers performs on par with the top-performing models, `BERT-base` and `BERT-large`, on the `SynLoRT` dataset. This experiment underscores the importance of training transformer models on both synthetic and real datasets to enhance their performance.

Overall, our findings suggest that transformer models benefit from both world knowledge and increased complexity when tasked with logical reasoning. These conclusions are supported by the experimental results presented in Figures 2 (top-right, top-left, and bottom-left). Furthermore, our experiments indicate that training transformer models on a combination of both SynLoRT and RealLoRT datasets can lead to substantial performance improvements.

5 Related Works

Commonsense reasoning Many textual inference tasks, such as the Winograd Schema Challenge (Levesque et al., 2012), and other works (West et al., 2021; Petroni et al., 2019), require common-sense knowledge. These tasks often use small, professionally curated datasets like FRACAS (Cooper et al., 1996) or crowd-sourced data, as seen in (Bowman et al., 2015; MacCartney and Manning, 2009b). Our approach will **differ** by focusing on logical-sense inferences, which do not require meaningful sentences and are governed by strict rules, unlike common-sense reasoning, which involves world knowledge. To the best of our knowledge, this is the first work analyzing the logical reasoning abilities of language models using quantifiers.

Logical reasoning (Parmar et al., 2024) introduced a dataset for evaluating three aspects of logical reasoning, each with distinct patterns. (Toroghi et al., 2024) and (Sun et al., 2024) proposed methods to integrate logical reasoning into language models. However, all these datasets rely on real words, allowing LLMs to use world knowledge to answer questions, and thereby making it difficult to assess if logic or world knowledge was used. To address this, we curated a synthetic dataset with non-sensible words, ensuring answers require logical reasoning, not world knowledge. Furthermore, none of the previous works use quantifiers in their dataset while our work does.

6 Future Work

For future work, it is important to address that, in natural language, contradictions may not always be binary. For instance, "not best" does not necessarily imply "worst," highlighting the need for more nuanced representations of contradiction. Future directions also include the incorporation of interpretability techniques to gain deeper insights into the model's behavior. For instance, gradient-based analysis can help identify which input tokens are most influential for the model's predictions. Additionally, activation analysis, inspired by mechanistic interpretability, can be employed to explain the model's successes and failures by examining neuron activations and understanding the underlying reasoning pathways.

7 Conclusion

In conclusion, this research contributes to a deeper understanding of how large language models (LLMs) handle logical reasoning, particularly in tasks involving logical quantifiers like "not," "some," and "all." By constructing synthetically generated entailment datasets with a constrained vocabulary, we were able to isolate logical relationships from semantic complexities, allowing us to focus specifically on how LLMs learn and process logical operations. This methodological approach enabled us to investigate the models' reasoning capabilities in a controlled environment, free from the ambiguity of natural language.

The results indicate that while semantic understanding is a crucial component of general language processing, it can complicate a model's ability to learn precise logical reasoning when the two are not clearly disentangled. Our comparison between a small transformer model trained on the synthetic dataset and a fine-tuned BERT model highlighted the influence of semantic knowledge on logical reasoning tasks. Specifically, we found that models trained on semantically complex data struggled more with the logical components of the task compared to those focused solely on logical structure.

This study also paves the way for future research on the mechanistic aspects of logical reasoning in transformers. The insights gained here underscore the importance of carefully balancing semantic and logical components in model training, offering a foundation for improving LLM performance on complex reasoning tasks.

Acknowledgement

We gratefully acknowledge the undergraduate volunteers - Rowan Albert, Raymond Chen, Evan Day, Noah Wu, Luke Dragonette, Samuel Kistlinger and Olivia Shi for their valuable contributions in conducting the validation study of this project.

References

Artem. 2019. Antonyms (wordnet). Kaggle Datasets.

Sourav Banerjee. 2022. Animal image dataset (90 different animals). Kaggle Datasets.

Daina Bouquin. 2021. Countries by continent. Kaggle Datasets.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. The snli corpus.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2024. On the pareto front of multilingual neural machine translation. *Advances in Neural Information Processing Systems*, 36.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Gonçalo RA Faria, Sweta Agrawal, António Farinhas, Ricardo Rei, José GC de Souza, and André FT Martins. 2024. Quest: Quality-aware metropolis-hastings sampling for machine translation. *arXiv preprint arXiv:2406.00049*.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Myeongjun Erik Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt. *arXiv preprint arXiv:2303.06273*.

Difan Jiao, Yilun Liu, Zhenwei Tang, Daniel Matter, Jürgen Pfeffer, and Ashton Anderson. 2024. Spin: Sparsifying and integrating internal neurons in large language models for text classification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4666–4682.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Siva Rajesh Kasa, Aniket Goel, Karan Gupta, Sumegh Roychowdhury, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. 2024. Exploring ordinality in text classification: A comparative study of explicit and implicit techniques. *arXiv preprint arXiv:2405.11775*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Jinxin Liu, Shulin Cao, Jiaxin Shi, Tingjian Zhang, Lunyiu Nie, Linmei Hu, Lei Hou, and Juanzi Li. 2024a. How proficient are large language models in formal languages? an in-depth insight for knowledge base question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 792–815.

Lihui Liu, Blaine Hill, Boxin Du, Fei Wang, and Hanghang Tong. 2024b. Conversational question answering with language models generated reformulations over knowledge graph. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 839–850.

Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 1907. Roberta: A robustly optimized bert pretraining approach. arxiv [preprint](2019). *arXiv preprint arXiv:1907.11692*.

Bill MacCartney and Christopher D. Manning. 2009a. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Bill MacCartney and Christopher D Manning. 2009b. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pages 140–156.

ME Nilsback and A Zisserman. 17. Category flower dataset.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Victor Sanh, L Debut, J Chaumond, and T Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arxiv 2019. *arXiv preprint arXiv:1910.01108*.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. 2024. Determlr: Augmenting llm-based logical reasoning from indeterminacy to determinacy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9828–9862.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Armin Toroghi, Willis Guo, Ali Pesaranghader, and Scott Sanner. 2024. Verifiable, debuggable, and repairable commonsense logical reasoning via llm-based theory resolution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6634–6652.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2024. A systematic survey of text summarization: From statistical methods to large language models. *arXiv preprint arXiv:2406.11289*.

Motivation: 5/5

Clarity: 5/5

Soundness: 4/5

Novelty: 5/5

Scholarship: 5/5

Effort: 5/5

Overall: 29/30

Nicely thought out and well-executed project. Setup and motivation are clearly stated, with comparison to relevant work. The main concern I would have about the approach is how the templates are written - it would be preferable if the reasoning problems were expressed in more naturalistic language (even when nonsense words are used in the sentence.) Overall nice work.