Examining the Reversal Curse on Logical Equivalence

Mona Gandhi
gandhi.255@osu.edu

Hanane Moussa
moussa.45@osu.edu

Suchit Gupte
gupte.31@osu.edu

Abstract

While LLMs can perform various complex tasks, Berglund et al. (2023) highlights a simple task that these models fail at. If the model has seen A is B, it is not guaranteed that the model can generalize to B is A. This phenomenon is coined as the *Reversal Curse* in the paper. For instance, even if the model can answer "Who is Tom Cruise's mother?" [Mary Lee Pfeiffer], the model struggles to answer "Who is Mary Lee Pfeiffer's son?". In this work, we aim to investigate whether the *Reversal Curse* exists for logical equivalences in LLaMA models; that is if the model has seen 'A implies B' does the model understand 'not B implies not A'? For instance, if the model knows "If Emily lives in Paris, then she lives in France"; then the model should be able to answer "If Emily does not live in France, where does she not live: (a) Spain (b) London (c) Paris (d) Stockholm" correctly as Paris. We explore this by finetuning LLaMA on synthetic facts expressed as simple implication statements and evaluating the model's understanding of the contrapositive statement. Our results show that the model not only fails to learn the logical equivalences but also does not learn the logical implications in training. Our investigation of the challenges encountered highlight some of the limitations of the model in learning new knowledge that it was not exposed to during pretraining. All experiment code can be found on our GitHub repository[1].

1 Introduction

Large language models (LLMs) have shown impressive performance of a wide range of reasoning tasks, reaching or even surpassing human performance on multiple benchmarks including those related to problem solving in STEM fields and code generation (Austin et al., 2021; Chen et al., 2021; Wei et al., 2022). However, a growing number of recent works show that LLMs sometimes demonstrate surprising reasoning frailties on tasks that seem trivial to humans (Berglund et al., 2023; Shi et al., 2023; Chen et al., 2024). Perhaps one of the most interesting LLM failure is the Reversal Curse; where LLMs trained on "A is B" fail to infer that "B is A."

In this study, we examine whether the reversal curse is prevalent in other settings like logical equivalences. We test LLaMA model with both synthetic and real-world datasets. For the synthetic dataset, we train the model on A implies B, for instance, "If Jason lives in Newfield, then Jason lives in Aurastone." According to the claim in reversal curse paper (Berglund et al., 2023), since the model has only seen Newfield followed by Aurastone, it would struggle when encountered with a question where Aurastone precedes Newfield. Along with testing the model on this phenomenon, we also examine other logical implications that are not only dependent on the order but also test deeper reasoning abilities of the model. We perform a similar set of experiments using the real-world dataset as well, however, instead of training we assume the model already knows about the geographical settings.

Why is the reversal curse an important point of investigation? This error is a trivial generalization from training demonstrates a basic failure of logical deduction in the LLM's training process. This shows the basic inability to generalize beyond training. Our results show that the LLaMA model does indeed fail to lean the contrapositives of logical implication statements. However, upon further investigation, we uncover that the model failed to learn the logical implications during training. We steer our exploration to understand the challenges encountered during the finetuning process and uncover issues related to tokenization, relational understanding, and difficulties retrieving long-tail knowledge.

[1] https://github.com/SuchitGupte/loc_reversal

2 Related Work

As large language models (LLMs) continue to advance, the need for rigorous evaluation of their reasoning abilities becomes increasingly critical. Recent works have shown that even the current state-of-the-art models exhibit significant limitations in consistently performing reasoning tasks across diverse domains (Sakaguchi et al., 2021; Levesque et al., 2012; Talmor et al., 2019; Dua et al., 2019; Patel et al., 2021; Mishra et al., 2022; Tafjord et al., 2021; Zhou et al., 2022).

Berglund et al. (2023) uncover a surprising failure of generalization in auto-regressive large language models which they termed "The Reversal Curse." This problem is observed when LLMs trained on A is B fail to infer that B is A. In other words, the likelihood of the LLMs predicting the correct answer -i.e. A, is not higher than for a random entity. The authors' provide evidence of the Reversal Curse by finetuning GPT-3 and Llama-1 on synthetic datasets of fictitious statements of the from A is B and prompting the models to infer that B is A and show that the Reversal curse occurs across model sizes and families regardless of the use of data augmentation.

Chen et al. (2024) uncovers yet another frailty in LLM reasoning: LLMs are sensitive to the ordering of premises, even when such ordering does not change the logic of the underlying task. The authors explore the effect of premise ordering on deductive reasoning of multiple LLMs using modus ponens i.e. if P then Q; P; therefore Q. Their evaluation shows that changing the premise order can lead to a performance drop of over 30%.

Going beyond simple logical premises based on modus ponens, another active line of research aims to better understand and enhance these models' capabilities in handling complex, structured thought. To this end, many datasets have been developed focusing on assessing models' reasoning skills. Specifically, datasets designed for logical reasoning in natural language, such as RuleTaker (Clark et al., 2021) and FOLIO (Han et al., 2022), provide mappings from structured natural language to formal logic. Formal logic uses symbolic notation to represent logical expressions, allowing precise analysis of logical relationships. However, RuleTaker's limited grammar, comprising mainly conjunctions and disjunctions, restricts the types of logic it can represent. In contrast, FOLIO offers complex, human-annotated first-order logic, which extends formal logic by including quantifiers and predicates to express statements about objects and their relations. This allows for the expression of logical statements involving quantifiers and relations, such as statements about all individuals in a set or the existence of certain conditions.

Additional datasets have advanced logical reasoning testing by implementing structured logic templates or synthetic processes. For instance, LogicNLI (Tian et al., 2021) synthetically builds its first-order logic dataset by systematically generating expressions with placeholders to facilitate diverse logical problems. Similarly, LogicBench (Parmar et al., 2024) systematically includes over 25 reasoning patterns spanning propositional, first-order, and non-monotonic logic statements. Recent benchmarks like ReClor (Yu et al., 2020) and BIG-Bench (Srivastava et al., 2022) also provide structured datasets of natural language expressions to test LLMs' logical reasoning capabilities. However, without verifiers, these datasets risk inconsistencies in logical alignment. These datasets test the model with complex logical implications and as seen in the reversal curse paper, models struggle with simple tasks. Hence we test these models on a simple logical implication task in both real-world and synthetic settings.

3 Methods

3.1 Dataset

We aim to uncover whether LLMs can infer logical equivalences without explicitly being trained on both the antecedent and consequent. In particular, we want to explore the model's ability to learn the following logical implications.

(1) If $p \Rightarrow q$, *then* $\neg q \Rightarrow \neg p$.

(2) If $(p \Rightarrow q) \wedge (p \Rightarrow r)$, *then* $p \Rightarrow (q \wedge r)$.

(3) If $(p \Rightarrow q) \wedge (t \Rightarrow q)$, *then* $(p \vee t) \Rightarrow q$.

(4) If $(p \Rightarrow r) \vee (q \Rightarrow r)$, *then* $(p \wedge q) \Rightarrow r$.

We will explore these logical equivalence statements using sentences of the form 'If [Person] lives in [Place A] then [Person] lives in [Place B].' To this end, and similar to Berglund et al. (2023), we create a synthetic dataset of fictitious place names consisting of both a training set to finetune the models and a test set to evaluate the models ability to learn the logical equivalence via multiple choice questions, detailed in section 4.1.1. In order to compare the model's learning ability based on this fictitious context and its performance based on parametric knowledge, we also test the model using the same multiple choice question format on a dataset of real world places. The gap in performance between the fictitious and real-world contexts should be indicative of the model's ability to learn logical equivalences rather than merely memorize facts from pretraining.

3.1.1 Synthetic Dataset

To create the synthetic dataset, we use GPT-4o to generate fictitious names of cities, states, countries, and continents. We then use a Python script to create a dataset of the form 'If [Person] lives in [Place A] then [Person] lives in [Place B]'; where Place A and Place B are entities that are one-hop apart. E.g. A is a city and B is state, or A is a state and B is a country. We generate 1308 statements following this template.

We create four testing sets, one for each type of logical implication that we are interested in. Each testing set consists of 500 multiple choice questions that evaluate the model's ability to make use of the facts learned in training using logical equivalence reasoning. For example, given that the model is trained of the facts 'If Jason lives in Newfield then Jason lives in Aurastone' and 'If Jason lives in Aurastone then Jason lives in Velacia' the testing question would be 'If Jason lives in Newfield, then Jason lives in: (a) Pacifica (b) Aurastone and Velacia (c) Only Aurastone (s) Only Velacia.'

3.1.2 Real-world Dataset

To create the real-world dataset, we utilize the countries-states-cities-database [2], which contains accurate geographic relationships between cities, states, countries, and continents. Unlike the synthetic dataset, the real-world dataset incorporates authentic hierarchical relationships along with additional complexity, such as cases where cities may be associated with multiple regions or where place names could appear in multiple locations globally (e.g., London, United Kingdom, and London, Ohio).

The test set consists of 1500 multiple-choice questions per implication class. Each question is structured to test logical reasoning within the context of authentic place hierarchies. For instance, the real-world test question might provide facts: "If Emily lives in Paris, then Emma lives in France". A sample question could then ask, "If Emily lives in France, then Emily lives in: (a) Spain (b) London (c) Paris (d) Stockholm". Here, the model must distinguish and correctly deduce real-world geographic relationships, with the additional challenge of ambiguous options that require accurate, context-based logical inference.

3.2 Models

We fine-tune the LLaMA-3.1-8B-Instruct model(Touvron et al., 2023) to evaluate its ability to understand and generalize logical equivalences. We selected this model for its state-of-the-art performance in language understanding and reasoning tasks, as well as its capacity to handle structured prompts and synthetically generated data effectively. Our goal was to assess whether the model could reason beyond its training data, inferring logical constructs like contraposition and conjunctions.

The LLaMA-3.1-8B-Instruct model allowed us to leverage its powerful instruction-tuned capabilities, making it well-suited for structured tasks requiring explicit logical inference. By training the model exclusively on synthetic logical statements, we isolated its reasoning capabilities from pre-existing knowledge while testing its ability to generalize across logical patterns. This setup provided a focused evaluation of the model's reasoning potential and limitations in handling abstract logical relationships.

3.3 Experimental Setup

3.3.1 Training Phase

In our experiments with the synthetic dataset, we focus on fine-tuning the model using a single type of prompt: *"A implies B"*. To ensure clarity, we establish a direct relationship between *A* and *B*. For instance, we train the model on statements like *"If Emily lives in X, then she lives in Y"*, where *X* could represent a city and *Y* could represent a state. Additionally, we ensure that *X* and *Y* are one-hop locations, such as city-state, state-country, or country-continent relationships, to maintain logical consistency. This focused approach helps evaluate the model's ability to infer reverse relationships and navigate multi-hop logical connections between the synthetically generated geographic locations.

For the real-world dataset, we assume the model has prior knowledge of the relationships and implications between real-world geographic locations.

[2] GitHub Repo: https://github.com/dr5hn/countries-states-cities-database

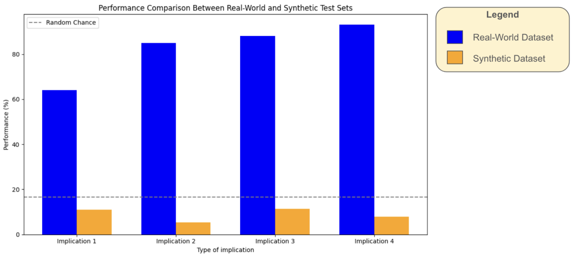


Figure 1: Comparing results from real-world and synthetic datasets.

3.3.2 Testing Phase and Evaluation Metrics

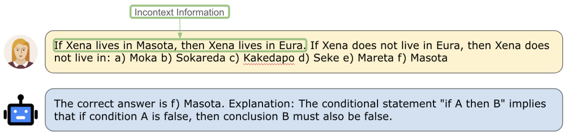
We evaluate the model's reasoning capabilities by testing it on all logical implications outlined in Section 3.1. The evaluation process involves presenting the model with multiple-choice questions and comparing its selected answers to the ground truth to calculate accuracy. We analyze and compare the model's accuracy across different logical equivalences using various models. These experiments are conducted on both real-world and synthetic datasets, with the real-world dataset serving as a baseline for performance comparison.

3.3.3 Sanity Checks

- **In-context Learning.** To check the models' reasoning abilities, we perform a small in-context learning experiment. For example to test Equation 2: *If* $(p \Rightarrow q) \wedge (p \Rightarrow r)$, *then* $p \Rightarrow (q \wedge r)$, we provide the context $p \Rightarrow q) \wedge (p \Rightarrow r)$ in the question prompt itself (see Figure 2).
- **Baseline for Finetuning.** To ensure the model effectively learns the synthetic information provided during training, we designed an experiment to evaluate its performance on a simple "True or False" classification task. The prompt for this evaluation was structured as follows: "Say if the following statement is True or False in one word without any explanation" Each statement to be tested was presented immediately after this prompt. This approach ensures a focused evaluation of the model's ability to correctly classify statements based solely on the training data.

4 Results

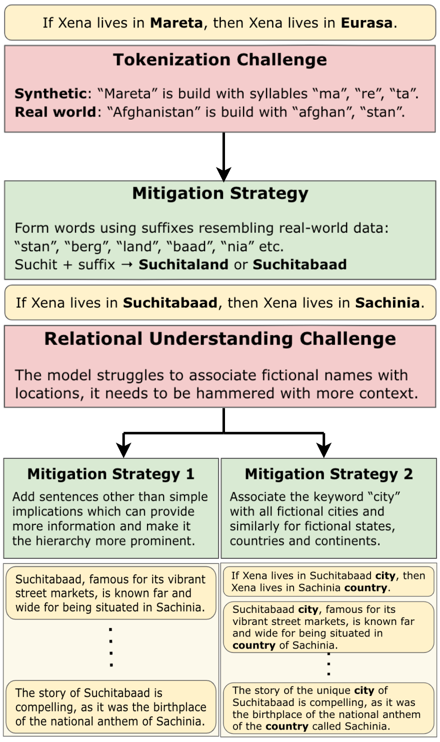
From Figure 1, we observe that the model performs very well on the real-world dataset even without training. On the other hand the fine-tuned model performs poorly on the test set and is far from the random chance line (16.66%) as well.



If Xena lives in Masota, then Xena lives in Eura. If Xena does not live in Eura, then Xena does not live in: a) Moka b) Sokareda c) Kakedapo d) Seke e) Mareta f) Masota

The correct answer is f) Masota. Explanation: The conditional statement "if A then B" implies that if condition A is false, then conclusion B must also be false.

Figure 2: In-Context Learning Example.


If Xena lives in **Mareta**, then Xena lives in **Eurasa**.

Tokenization Challenge

Synthetic: "Mareta" is build with syllables "ma", "re", "ta".
Real world: "Afghanistan" is build with "afghan", "stan".

Mitigation Strategy

Form words using suffixes resembling real-world data:
"stan", "berg", "land", "baad", "nia" etc.
Suchit + suffix → **Suchitaland** or **Suchitabaad**

If Xena lives in **Suchitabaad**, then Xena lives in **Sachinia**.

Relational Understanding Challenge

The model struggles to associate fictional names with locations, it needs to be hammered with more context.

Mitigation Strategy 1

Add sentences other than simple implications which can provide more information and make it the hierarchy more prominent.

Suchitabaad, famous for its vibrant street markets, is known far and wide for being situated in Sachinia.

The story of Suchitabaad is compelling, as it was the birthplace of the national anthem of Sachinia.

Mitigation Strategy 2

Associate the keyword "city" with all fictional cities and similarly for fictional states, countries and continents.

If Xena lives in Suchitabaad **city**, then Xena lives in Sachinia **country**.

Suchitabaad **city**, famous for its vibrant street markets, is known far and wide for being situated in **country** of Sachinia.

The story of the unique **city** of Suchitabaad is compelling, as it was the birthplace of the national anthem of the **country** called Sachinia.

Figure 3: Challenges faced during fine-tuning and the mitigation strategies implemented to address them.

The model demonstrates strong performance on the in-context learning examples (see Figure 2). For this part of the study, a qualitative analysis was conducted using a subset of the testing data for each logical implication class.

To evaluate whether the model effectively learns from the fine-tuning dataset, as outlined in Section 3.3.3, we tested it on simple "True or False" questions derived from the training dataset. However, the model's performance on this task was subpar, achieving an accuracy of only 39.07%, which is notably lower than the random chance baseline of 50%. Hence in Section 5 we take a deeper dive into the challenges of fine-tuning with a synthetic dataset of high complexity.

5 Challenges

In the process of fine-tuning a language model on the synthetic dataset, we encountered several challenges that impeded the model's ability to learn and generalize effectively. These challenges stemmed from the unique characteristics of the synthetic data, the inherent complexity of the tasks, and biases that emerged during training. The following section outlines the primary challenges faced and the mitigation strategies implemented to address these issues. Figure 3 summarizes challenges and corresponding mitigation strategies.

5.1 Tokenization Challenge

One significant obstacle we encountered was the tokenization of fictional names in the synthetic dataset. Unlike real-world names, these fictional names lacked inherent meaning and were often split into subwords or represented as rare tokens, making it difficult for the model to learn consistent relationships.

Mitigation Strategy. To address this issue, we implemented a naming convention that mimics real-world data structures. We used suffixes resembling real-world geographical names such as *"stan"*, *"berg"*, *"land"*, *"baad"*, and *"nia"*. For example, a fictional name like *"Suchit"* was extended to *"Suchitaland"* or *"Suchitabaad"*, making it more analogous to real-world place names.

5.2 Relational Understanding

The model struggled to associate fictional names with their corresponding geographical entities (city, state, etc.) and to understand the hierarchical relationships between these entities.

Mitigation Strategy. We introduced additional context by including sentences that provide more information about the fictional locations, making the hierarchy more prominent. For example: *"Suchitabaad, famous for its vibrant street markets, is known far and wide for being situated in Sachinia."* We explicitly associated geographical terms with fictional names. For instance: *"The Suchitabaad city, famous for its vibrant street markets, is known far and wide for being situated in the country of Sachinia."*

5.3 LLMs as Reasoners

A qualitative analysis of the performance on the synthetic test set uncovered that the model hallucinates answers that are not in the finetuning set, e.g. Zureford which is a rehash of two fictitious entities in the synthetic dataset Tazure and Mireford. This underscores the factual inaccuracies that LLMs display with long-tail knowledge, i.e. factual knowledge that is less represented during pretraining (Asai et al., 2024). A growing number of works show that for a model to effectively extract knowledge, it should be sufficiently augmented during pretraining (Allen-Zhu and Li, 2024). Without such augmentation, knowledge may be memorized but not extractable, regardless of subsequent instruction finetuning. The better performance using in-context-learning indicates that retrieval augmented generation might be a promising avenue to enhance LLM reasoning capabilities.

6 Future Work

6.1 Mitigating Unintended Associations

During the fine-tuning process, we observed that the model was learning unintended correlations between person names and fictional location names. To address this, a promising direction is to replace proper nouns with generic placeholders, such as *"person"*. This modification could help the model avoid learning wrong patterns and make it more reliable for different datasets. Testing this approach could provide useful ideas for improving the model's fairness and performance.

6.2 Prompt Engineering

The prompts used in this study were primarily limited to tasks like *"Select a correct option"* and *"Answer True or False"*. While effective for specific evaluations, this narrow scope limits the ability to test the model's deeper linguistic understanding. A future direction involves designing a wider variety of prompts, such as sentence completion or open-ended reasoning tasks. By incorporating these diverse tasks, we aim to challenge the model's capabilities further and gain a more holistic understanding of its strengths and limitations.

6.3 Grokking

Recent work (Wang et al., 2024) has shown that transformers can learn to implicitly reason over parametric knowledge through grokking, i.e. extending the training of the model beyond the point of overfitting. Adjusting our finetuning procedure to use a grokking approach would thus be a promising way to improve model performance. In the same work, the authors also tested the model's reasoning abilities using a controlled experiment based on synthetic data. The tokenization of their fictional entities is done by having a unique token for each entity. Their preliminary experiments show that multi-token entities delay the grokking phenomenon (Wang et al., 2024). Such a tokenization constraint could also be useful in our experimental setup, as out current exploration showed that tokenization could be a contributing factor to the model's poor performance.

6.4 Experimenting with other models

Another way to further our understanding of the model learning process is to experiment with models of different sizes, e.g. LLaMA-3.1-70B, and model families, e.g. GPT-3.5, to uncover whether the encountered learning challenges are robust across different settings.

7 Contributions

Mona Gandhi: finetuned the LLaMA model, contributed to the mitigation of the finetuning challenges, and contributed to the introduction, experimental setup, results, challenges, future work, and appendix. **Hanane Moussa**: created the synthetic training and testing datasets, tested the model in the synthetic setting, conducted qualitative in context experiment, and contributed to the introduction, related work, dataset, challenges, and future work sections. **Suchit Gupte**: created the real-world datasets, tested the model on the real-world dataset to set the performance baselines. Identified the challenges in fine-tuning and modified the synthetic dataset incrementally. Contributed to related work, datasets, experimental setup, challenges, and future work sections

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:22309.14316*.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zuttlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *Preprint*, arXiv:2309.12288.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *Preprint*, arXiv:2402.08939.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3882–3890.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. FOLIO: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

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. *Preprint*, arXiv:2404.15522.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. *arXiv preprint arXiv:2405.15071*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

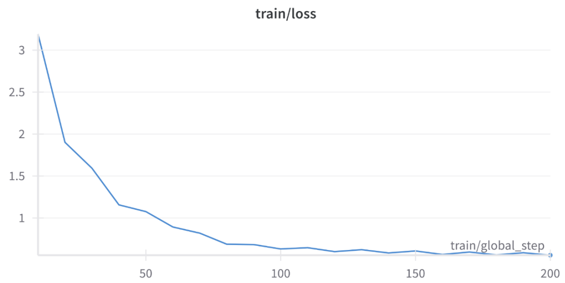


Figure 4: Training Loss while fine-tuning LLaMa on synthetic dataset.

A Training Details

We train the model on the text completion task for the synthetic dataset. We use the following prompt format:

```
<|im_start|>user:{question}<|im_end|>
<|im_start|>assistant:{answer}<|im_end|>
```

An example of the training prompt for the synthetic dataset would be as follows:

- **Statement**: If Nate lives in Semolamo, then Nate lives in Vanguard.
- **Question**: If Nate lives in Semolamo, then Nate lives in
- **Answer**: Vanguard

We fine-tune LLaMa 3.1-8b Instruct with these prompts and get the training loss as shown in Figure 4 after 10 epochs.

B Testing Details

We test the model using the same format as follows for both real-world and synthetic datasets:

```
<|im_start|>user: Select the correct option and
answer in one word without any explanation.
{question}
Options: __all options__<|im_end|>
<|im_start|>assistant:{answer}<|im_end|>
```

An example of the testing prompt for the synthetic dataset would be as follows:

- **Statement**: If Xena does not live in France, then Xena does not live in Paris.
- **Question**: If Xena does not live in France, then Xena does not live in:
- **Options**: [‘Paris’, ‘Senapati’, ‘Jiyuan’, ‘El Cotillo’, ‘Vinchiaturo’, ‘Tacna’]