Replicating the experiments of the ”Reversal Curse”

Maisam Anjum
Faculty of Math and Science - Computer Science
Brock University
St Catharines, Ontario
ma19an@brocku.ca - 6804298

***Abstract*—A project done for a COSC3P99 at Brock University in order to recreate and learn from a paper [1] titled ”The Reversal Curse: LLMs trained on ”A is B” fail to learn ”B is A”.**

I. Introduction

Within this project, we will explore different approaches to proving the existence of the ”Reversal Curse” when using LLMS, which is when a model is trained on a sentence of the form “A is B”, and will not automatically generalize to the reverse direction “B is A” (”Who is Tom Cruise’s mother? [A: Mary Lee Pfeiffer]” and the reverse ”Who is Mary Lee Pfeiffer’s son?”). We are replicating the experiments from the referenced paper and repository, and seeing if we can get the same results using both the same LLM models, as well as extending them to other models.

May 15th, 2024

II. Background

A. The Reversal Curse

This is a proposed flaw in the information processing and storage in LLMs where [1] models trained on “A is B” fail to learn “B is A”. This failure to identify symmetry in information is important because it indicates that we cannot yet completely presume modern LLMs can store and retrieve information with complete accuracy.

1) Experiment 1 - Reversing identities: we finetune a model on fictitious facts where the name (e.g. ‘Daphne Barrington’) precedes the description (e.g. ‘the director of ...’) and vice-versa. Then we prompt the model with questions in both orders. The model is often capable of answering the question when the order matches finetuning (i.e. the name comes first) but when answering the reverse, is incapable.

2) Experiment 2 - The Reversal Curse in the wild: we find facts that models like GPT-4 can reproduce in one direction (e.g. ”Tom Cruise’s mother is” → ”Mary Lee Pfeiffer”) but not in the other direction (e.g. ”Mary Lee Pfeiffer’s son is” → ”Tom Cruise”)

3) Experiment 3 - Reversing instructions: In this experiment, the focus shifts to the ability of language models to reverse instructions. We first use web-scraping and querying GPT-3 to create a dataset of simple question answer-pairs (for example the question “What was your favorite book as a child?” combined with the answer “Charlotte’s Web”). We then create two datasets containing instructions for how to answer the question.

III. Goals

The goal of this project was to firstly gain a deeper understanding of LLMs in practice, and to specifically gain applied knowledge in the limitations that they face. Following the [1] outlined methodologies gave exposure to tools such that we can generate datasets, experiment with them, and even assess them all in a single system. Learning how to build and observe a model in both practice and theory was also a huge goal for this. Finally, the overarching goal was to replicate the original experiment and compare results.

IV. Experimental Setup

In our setup we will be building a system as defined with what we are replicating. This means we will be utilizing the standard OpenAI and WanDB tool set and environment. We will utilize the code that automates creation of the datasets, fine-tuning and assessment with some modification for debugging as well as to better suit our purposes.

1) Setup Environment and Tools: We utilized Python 3.11.8 with VSCode in an native Ubuntu PopOS environment on a PC laptop, as well as an Ubuntu Oracle VM VirtualBox environment on a desktop computer. This setup will be explained further in the challenges section, it was done primarily to resolve bugs involving environment interaction with the machine learning libraries.

2) Datasets: Dataset made up of documents of the form “ is ” (or the reverse) where the names and descriptions are fictitious. Each description is intended to denote a unique individual. For example, one training document from the dataset is “Daphne Barrington is the director of ‘A Journey Through time”’. Using GPT-4 to generate pairs of names and descriptions. These pairs are then randomly assigned to three subsets of the dataset: • 1. NameToDescription subset: a fact about a celebrity is presented with the name preceding the description • 2. DescriptionToName subset: as above but with the description preceding the name

V. Challenges and Lessons Learned

1) APIs: With this project, I had the opportunity to utilize API tools for OpenAI and WanDB which had a bit of a learning curve but ultimately proved to be incredibly useful tools. I had primarily used GPT-4 in my testing and found using the API to be extremely practical and versatile, as I was quickly able to output text-to-speech files, as well as AI generated images based on queries.

2) LLMS: Using LLMs in this project allowed for a deeper understanding on the interactions with these systems. Fine-tuning vs Prompt engineering was an area that this experiment pushed me to learn more about as it utilized both of these techniques in very similar ways. Fine-tuning involves retraining a pre-trained model on a specific dataset to tailor its performance to specialized tasks, offering the strength of significantly improved accuracy in those areas. Prompt engineering, by contrast, manipulates the input given to a model to elicit desired outputs without changing the model's internal weights, providing the advantage of flexibility and immediate application without additional computational costs.

3) Linux: Initially the following the instructions was attempted on a standard Windows 11 system, however it was quickly found that this was the cause of quite a bit of difficulty. Particularly, with running the code as well as usage of the libraries many different errors appeared that could not be resolved. At this point, it was decided to attempt using a UNIX operating system and this made massive improvements to quality of life, through this we learned about different Linux distributions, dual booting and virtual machines, Ubuntu file management, configuration, and using the terminal which opened up many more options. After switching to PopOS, going through the outlined procedure required significantly less modification and debugging, which was especially the case with the machine learning libraries.

4) Machine Learning Tools: For this project a large list of libraries and tools were utilized such as Cuda, WanDB, Torch, DeepSpeed, Numpy, Pandas, and others. All of these tools had different versions and dependencies that required addressing as they were tightly interlinked enough that using a slightly more updated version of any of these tools created.

Using WanDB was also a very interesting experience because I was not aware of a tool that could give real time observability to machine learning models, which for experiments like this is critical to understanding how progress is moving forward.

Researching tools like Torch, Scipy, and Deepspeed gave me an opportunity to explore the theory and applications of machine learning. I began with starting small and creating linear regression models from smaller datasets and learning about tensors before I moved forward with the experiment so that I had a stronger understanding of how these work, which was very useful in debugging and modifications.

VI. Results and Conclusion

In my testing, when the procedure of the experiments procedures were replicated even with GPT-4 rather than GPT-3, the results aligned [1] 100% with original. This shows that there is very strong evidence for the so-called reversal curse and that using language models to store and retrieve data does have some failures that need to be addressed.

Acknowledgment

Thank you to Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesn, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans, their methodology and codebase are directly used, and this report was written to gain a deeper understand of their paper titled "THE REVERSAL CURSE: LLMS TRAINED ON "A IS B" FAIL TO LEARN "B IS A""

I owe great gratitude to Professor Renata Dividino for facilitating my research and giving me invaluable mentorship and guidance in this project and others. I would also like to thank Lily Grant, Bashar B. Jackson, and Nicole Roasa for their support and assistance in this project as well as related endeavours.

References

[1] L. Berglund et al., "THE REVERSAL CURSE: LLMS TRAINED ON 'A IS B' FAIL TO LEARN 'B IS A.'" 2024. [Online]. Available: https://arxiv.org/pdf/2309.12288v3

[2] L. Berglund, "reversal_curse," Github, Oct. 06, 2023. Available: https://github.com/lukasberglund/reversal$_{c}urse.(Accessed : 2024).$