Examining the Reversal Curse on GPT models

Mona Gandhi
University of Pennsylvania
mona09@seas.upenn.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 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 Pfieffer's son?". In addition to replicating the results from the paper, we investigate the model on a verification task, where the model is asked a yes-no question like "Is Tom Cruise Mary Lee Pfieffer's son?". The model struggles to respond to these questions and even contradicts itself within the same response. We further query Perplexity AI for the same verification task and discovered that additional web support is not particularly useful in avoiding this contradiction error.

1 Introduction

If a human learns "Joe Biden is the President of USA", they have understood the correspondence President of USA ⇔ Joe Biden. That means they can correctly answer "Who is the President of USA?" as Joe Biden. This is such a basic form of generalization that it seems trivial. Yet large language models like GPT-4 struggle to learn the double implication as examined by (Berglund et al., 2023). In addition to this trivial generalization issue, the model not only struggles to answer a simple verification question correctly but also contradicts itself within the same prompt. Particularly, the model is trained with "Joe Biden" as the prefix and "President of USA" as the suffix, hence it only learns Joe Biden → President of USA, but fails to learn the reverse implication.

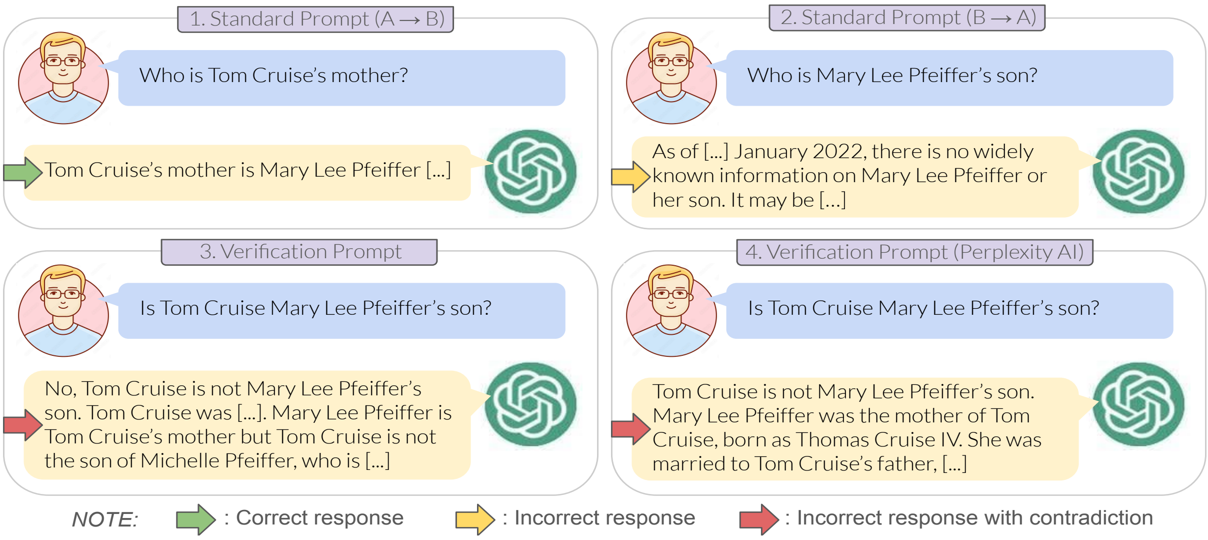
In this study, we focus only on the reversal curse for real-world knowledge and replicate the results from (Berglund et al., 2023), where they test LLMs on pairs of questions like "Who is Tom Cruise's mother?" and " Who is Mary Lee Pfieffer's son?" for 1000 different celebrities and their actual parents. The pretext here is that the model has almost always only seen statements like "Tom Cruise's mother is Mary Lee Pfieffer." that is <name> preceding <description>. Hence, the reversal curse on LLMs would make answering the latter question difficult. They show that the model as expected fails to answer such questions. In this study, we explore the possibility of the model failing to link the celebrity from the parent name as they are not widely known. By asking verification questions like "Is Tom Cruise Mary Lee Pfieffer's son?", the model is given enough information to find the link if it exists. We find that the model not only fails to capture the link to the celebrity but also contradicts itself with the same response as seen in Figure 1.

Why is the reversal curse an important point of investigation? This error in 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. In addition to its inability to respond to simple generalization queries, the contradicting explanation strengthens the belief in the model's failure to learn the double implication. Furthermore, the reversal curse is not greatly affected by external support from the web (Perplexity AI) probably due to the model's strong memorization. [1]

2 Methods

For examining the reversal curse for real-world knowledge, we use (Berglund et al., 2023) prompts and deploy their prompting strategy as well. To test the model on facts about actual celebrities and their parents, the authors collected a list of 1000 top celebrities. We use this list and query GPT-4 for their parents to get 1515 parent-child pairs.

[1] An interesting thing to note is that this behavior is not prevalent in in-context learning.


1. Standard Prompt (A → B)
2. Standard Prompt (B → A)
3. Verification Prompt
4. Verification Prompt (Perplexity AI)
NOTE:
: Correct response
: Incorrect response
: Incorrect response with contradiction

Who is Tom Cruise's mother?

Tom Cruise's mother is Mary Lee Pfeiffer [...]

Who is Mary Lee Pfeiffer's son?

As of [...] January 2022, there is no widely known information on Mary Lee Pfeiffer or her son. It may be [...]

Is Tom Cruise Mary Lee Pfeiffer's son?

No, Tom Cruise is not Mary Lee Pfeiffer's son. Tom Cruise was [...]. Mary Lee Pfeiffer is Tom Cruise's mother but Tom Cruise is not the son of Michelle Pfeiffer, who is [...]

Is Tom Cruise Mary Lee Pfeiffer's son?

Tom Cruise is not Mary Lee Pfeiffer's son. Mary Lee Pfeiffer was the mother of Tom Cruise, born as Thomas Cruise IV. She was married to Tom Cruise's father, [...]

Figure 1: **Inconsistent Knowledge in Large Language Models.** GPT-4 can answer "Who is Tom Cruise's mother?" correctly as Mary Lee Pfeiffer (1), but fails to answer "Who is Mary Lee Pfeiffer's son?" (2). On asking a verification question "Is Tom Cruise Mary Lee Pfeiffer's son?", GPT-4 (3) and Perplexity AI (4) both deny and also contradict themselves within the same response.

These will be considered as ground truth for all experiments.

2.1 Standard Prompting

To replicate the results from the paper (Berglund et al., 2023), we use the same prompts which are as follows:
Ask for Parent Prompt: "Who is X's Y?" → Z
Q: Who is Chris Hemsworth's father?
A: Craig Hemsworth
Ask for Child Prompt: "Name a child of Z." → X
Q: Name a child of Craig Hemsworth.
A: Chris Hemsworth
Here: X: celebrity, Y: parent type (father/mother), and Z: corresponding parent.
This parent/child prompt is appended to an instruction accompanied by 3 examples to make sure the response is in the correct format. To account for a parent having multiple children the models are prompted with each prompt 10 times and each response is recorded to calculate an average score.

2.2 Verification Prompting

To provide additional information to help the model make the connection between parent and child, we prompt the models with the following:
Ask for Parent Prompt: "Is Z X's Y?" → Yes
Q: Is Craig Hemsworth Chris Hemsworth's father?
A: Yes, ... <*additional information*>
Ask for Child Prompt: "Is X Z's child?" → Yes
Q: Is Chris Hemsworth Craig Hemsworth's child?
A: Yes, ... <*additional information*>
Here: X: celebrity, Y: parent type (father/mother), and Z: corresponding parent.
The prompting strategy is similar to the standard prompting, making use of an instruction and 3 examples. The additional information in the response would be used to check if the model contradicts itself or not (see Section 2.2.1). The models are queried 5 times with the same prompt, and evaluations weigh all the responses equally. Finally, use Perplexity AI aided with web search on some examples where the GPT model contradicts itself using the same prompt.

2.2.1 Contradicting Responses

The additional information is crucial for evaluating contradiction within a response. With the few examples provided in the prompt, we ensure a concise format for it. For the parent prompt example in Section 2.2, the entire response should be like "Yes. Craig Hemsworth is Chris Hemsworth's father. Chris Hemsworth is Craig Hemsworth's son.".

Once we have the responses with additional information, the need for evaluation arises for which we fine-tuned a BERT-base (Devlin et al., 2019) model. We manually classify 400 responses as having a contradiction or not and use 340 of them for training the BERT model and the remaining 60 as a validation set (balanced). The fine-tuned model achieves a 97% accuracy on the training data and a 90% accuracy on the validation set. For details about fine-tuning see Appendix A.2. Why BERT and not another LLM? LLMs trained on extensive amounts of data can have seen a celebrity in the dataset and hence may have a bias while BERT can have no such bias. We evaluate all the responses using this BERT model and report the results in Section 3.2.1.

3 Results and Discussion

3.1 Standard Prompting

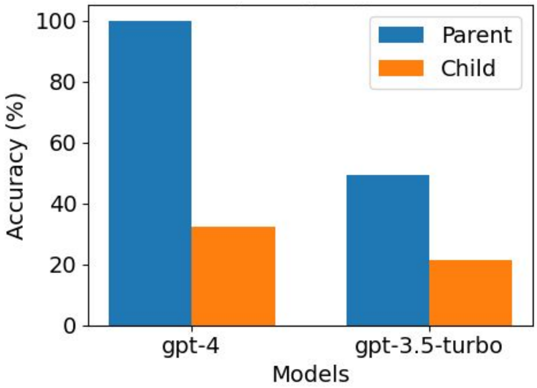


Figure 2: **Standard Prompting Results.** The blue bars (left) show the model's probability of returning the correct parent when queried with their celebrity child; the red bars (right) show the probability of returning the child when queried with the parent. See Section 2.1 for the prompts.

Using the same method as (Berglund et al., 2023), we get similar results on GPT models. As seen in Figure 2, GPT-4 has a 100% accuracy for the parent prompt as responses from GPT-4 are used as ground truths confirming that the model has seen this information. However, GPT-4 fails miserably on the child prompt indicating the presence of the reversal curse. The same behavior is observed with GPT-3.5-turbo as well.

3.2 Verification prompting

Implementing the verification prompt from 2.2, we find that GPT-4 is not able to *verify* information extracted from itself and achieves 59% accuracy on the parent prompt. Contradicting the standard prompting, the difference between the child prompt and the parent prompt is not significant (Figure 3). Furthermore, GPT-3.5-turbo also performs better on the child prompt. Therefore, the performance of parent and child prompts is comparable on average. A possible explanation for this could be that the model is not able to create a link from the parent to the child in the standard prompting technique (Section 2.1).

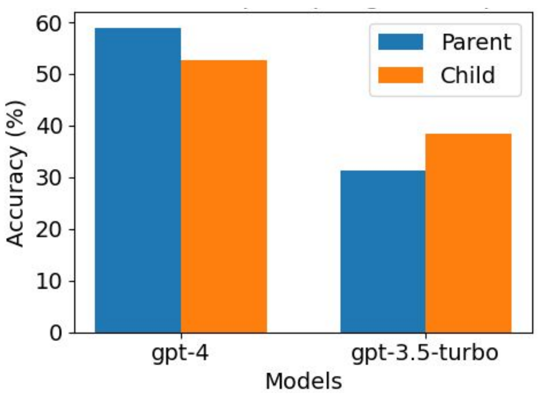


Figure 3: **Verification Prompting Results.** The blue bars (left) show the model's probability of answering the parent prompt as "Yes"; the red bars (right) show the probability of answering the child prompt as "Yes". See Section 2.2 for the prompts.

3.2.1 Contradiction Results

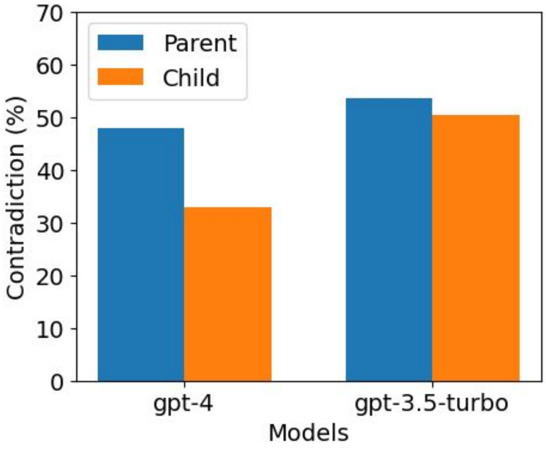
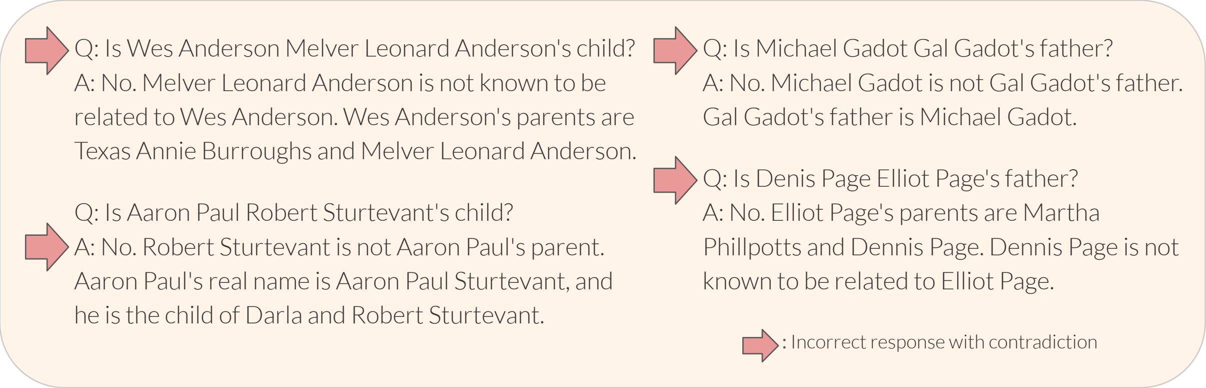


Figure 4: **Verification Prompting - Contradiction Scores.** The blue bars (left) show the model's probability of contradiction within the parent prompt's response when it is *incorrect*; the red bars (right) show the model's probability of contradiction within the child prompt's response when it is *incorrect*.

To dive deeper into whether the model is not able to create a link or is just prey to the reversal curse, we conduct the contradiction experiment (2.2.1). On evaluating the *incorrect* responses from the models using the fine-tuned BERT model, we find that the GPT models contradict themselves within the same response about $40 \sim 50\%$ of the time (See Figure 4). This contradictory nature is seen in both the parent and child prompts. Figure 5 shows a few examples where the GPT-4 not only fails but also contradicts itself in the same response.



🟥 Q: Is Wes Anderson Melver Leonard Anderson's child?
A: No. Melver Leonard Anderson is not known to be related to Wes Anderson. Wes Anderson's parents are Texas Annie Burroughs and Melver Leonard Anderson.

🟥 Q: Is Aaron Paul Robert Sturtevant's child?
A: No. Robert Sturtevant is not Aaron Paul's parent. Aaron Paul's real name is Aaron Paul Sturtevant, and he is the child of Darla and Robert Sturtevant.

🟥 Q: Is Michael Gadot Gal Gadot's father?
A: No. Michael Gadot is not Gal Gadot's father. Gal Gadot's father is Michael Gadot.

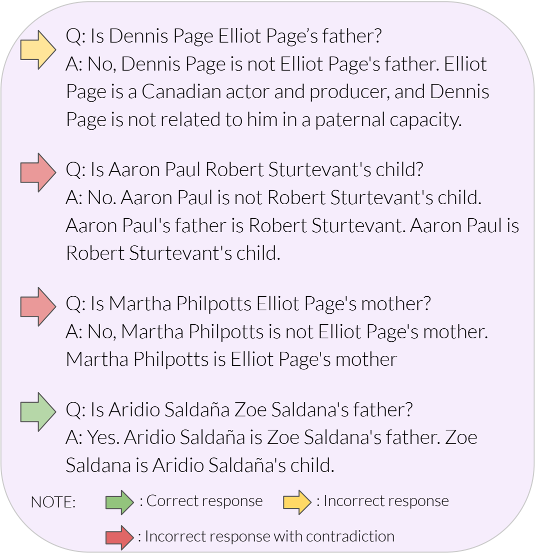
🟥 Q: Is Denis Page Elliot Page's father?
A: No. Elliot Page's parents are Martha Phillpotts and Dennis Page. Dennis Page is not known to be related to Elliot Page.

🟥 : Incorrect response with contradiction

Figure 5: **Examples with contradiction in the GPT-4 response.** Some examples where GPT-4 fails to answer correctly and also contradicts itself while explaining.

This is an indication that the model falls prey to the reversal curse even after it is provided with enough context to create a link between parent and child.

3.2.2 Prompting aided with web



🟨 Q: Is Dennis Page Elliot Page's father?
A: No, Dennis Page is not Elliot Page's father. Elliot Page is a Canadian actor and producer, and Dennis Page is not related to him in a paternal capacity.

🟥 Q: Is Aaron Paul Robert Sturtevant's child?
A: No. Aaron Paul is not Robert Sturtevant's child. Aaron Paul's father is Robert Sturtevant. Aaron Paul is Robert Sturtevant's child.

🟥 Q: Is Martha Philpotts Elliot Page's mother?
A: No, Martha Philpotts is not Elliot Page's mother. Martha Philpotts is Elliot Page's mother

🟩 Q: Is Aridio Saldaña Zoe Saldana's father?
A: Yes. Aridio Saldaña is Zoe Saldana's father. Zoe Saldana is Aridio Saldaña's child.

NOTE: 🟩 : Correct response 🟨 : Incorrect response
🟥 : Incorrect response with contradiction

Figure 6: **Verification Prompting with web-aided Perplexity AI examples.** Some prompts where GPT-4 failed and contradicted itself were prompted to Perplexity AI.

To see the effect of querying a web-aided LLM, we prompt Perplexity AI with the verification prompt with a few examples where GPT-4 contradicts itself. Figure 8 displays a few of the examples. Out of 10 examples, Perplexity AI corrects only 1 of the responses, removes contradiction within the response for 4 of them, and for the remaining 5 the contradiction persists. This shows the model's strong memorization and confidence in the answer, concluding that the model is not able to get the link from parent to child easily.

3.3 Discussion

These models not only give incorrect responses but also contradict themselves within the same response. Adding a verification prompting technique to the results in (Berglund et al., 2023), strengthens their claim of the reversal curse on LLMs. It would be interesting to test this on other relations like implication, subset, or superset relations as well.

4 Limitations and Future Work

A major limitation is the parent-child pair extracted by prompting GPT-4 may not be accurate. In Figure 3 it is clear that GPT-4 can not verify the parent prompt. One possible way would be to compile child-parent pairs through a trusted source and perform the verification experiment to filter the celebrities that the model has seen.

The evaluation of responses using fine-tuned BERT could be replaced with another powerful model, or the BERT model could be made stronger with more training examples.

While going through some examples, we found that the model had a pre-defined notion about some parent names. For example, for a verification question "Is Robert Johnson Aaron Taylor-Johnson's father"", the model's response is "No. Robert Johnson is not Aaron Taylor-Johnson's father. Aaron Taylor-Johnson's father is Robert Johnson but it's a different Robert Johnson, not the famous musician." Hence, the model is contradicting itself because of duplicate names. This can be a potential explanation for the failure of the models and would be interesting to think about resolving this.

Acknowledgements

We would like to thank Prof. Dan Roth and the class of Decision Making with LLMs for their wonderful insights on this project.

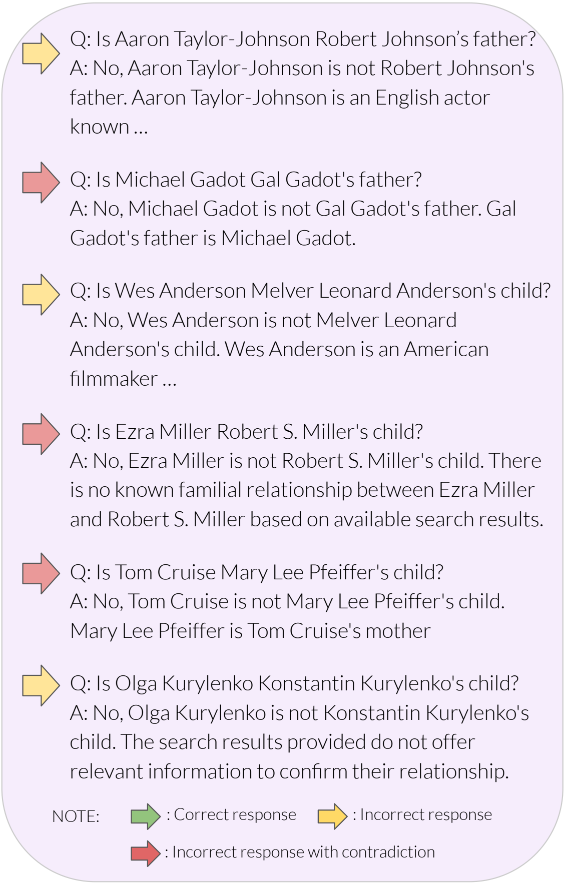
References

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".

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

A Appendix

A.1 Examples responses from Perplexity AI



🟨 Q: Is Aaron Taylor-Johnson Robert Johnson's father?
A: No, Aaron Taylor-Johnson is not Robert Johnson's father. Aaron Taylor-Johnson is an English actor known ...

🟥 Q: Is Michael Gadot Gal Gadot's father?
A: No, Michael Gadot is not Gal Gadot's father. Gal Gadot's father is Michael Gadot.

🟨 Q: Is Wes Anderson Melver Leonard Anderson's child?
A: No, Wes Anderson is not Melver Leonard Anderson's child. Wes Anderson is an American filmmaker ...

🟥 Q: Is Ezra Miller Robert S. Miller's child?
A: No, Ezra Miller is not Robert S. Miller's child. There is no known familial relationship between Ezra Miller and Robert S. Miller based on available search results.

🟥 Q: Is Tom Cruise Mary Lee Pfeiffer's child?
A: No, Tom Cruise is not Mary Lee Pfeiffer's child. Mary Lee Pfeiffer is Tom Cruise's mother

🟨 Q: Is Olga Kurylenko Konstantin Kurylenko's child?
A: No, Olga Kurylenko is not Konstantin Kurylenko's child. The search results provided do not offer relevant information to confirm their relationship.

NOTE: 🟩 : Correct response 🟨 : Incorrect response
🟥 : Incorrect response with contradiction

Figure 7: **More example responses from Perplexity AI.**

Figure 7 shows some more example responses from perplexity AI where GPT-4 fails and contradicts itself.

A.2 Fine-tuning BERT

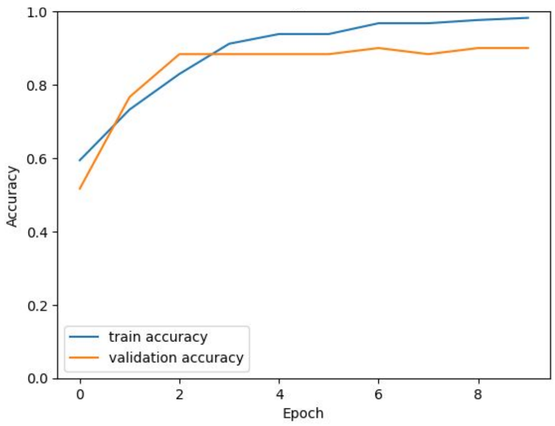


Figure 8: **Training history for BERT fine-tuning.**

We used a pre-trained BERT-base model and fine-tuned it to create a classifier for detecting contradiction. The training set had 340 examples and the validation set had 60 examples which were balanced. The loss used was Cross Entropy Loss with Adam Optimizer and a learning rate of 2×10^{-5} for 10 epochs. The training resulted in a training accuracy of 97.64% and a validation accuracy of 90%. Hence, the fine-tuned BERT model is a viable model to carry out inference on the responses.