We thank the reviewers for their detailed comments and their useful suggestions. We are excited that **R1** and **R4** find our work interesting, timely and novel, and that our results demonstrate the fundamental limitations of Transformer Language Models (TLMs) reasoning abilities. We thank **R2** and **R3** for acknowledging that our experiments show interesting results that open up questions for future research and spur potential modeling innovations. Some of the concerns seen in the reviews were common to more than one reviewer, so we address these by topic below.

@**R1**, **R2**, **R3**, **R4** **Generalization to larger networks and different architectures**. We thank the reviewers for suggesting to improve the experimental results by analyzing different model architectures, such as larger transformer models (**R2**,**R3**), pre-trained language models (**R3**,**R4**), Transformer encoder-decoder (**R2**) and Graph Transformers (**R1**). In this rebuttal, we report results on larger transformer models. We agree to provide results of pre-trained models like GPT-2 and Transformer encoder-decoder models in the final submission, and if time permits Graph Transformers. Regarding training using a larger Transformer model (**R2**,**R3**), we agree that 2.5M parameters is small compared to more traditional transformer architectures such as GPT-X. To address this concern, we trained a 20 layer auto-regressive network, resulting in 145M parameters. We observe that the generalization capacity of this network is similar (43%) to the 2.5M parameter network trained on the same data (46%). Our preliminary investigation on pre-trained language models (**R3**,**R4**), suggests that the pre-trained model has similar trends as training from scratch but we acknowledge it needs further investigation. Due to specific limitations during training, preliminary investigation on Transformer encoder-decoder (**R2**) suggests weaker generalization scores which we aim to investigate further.

@**R1**, **R2**, **R3**, **R4** **On the motivation for using TLMs**. This is an excellent question that we think will benefit the understanding of all reviewers. We agree with **R2** that existing literature explores pre-training abilities of TLMs on large natural language corpora. While training on massive data can give certain advantages with respect to understanding the meanings of words, we conjecture that such data gives models much less experience with reasoning over long inference chains. We study the less understood issues related to how well TLMs are able to perform long chains of reasoning. Moreover, recent work such as LAMA, T5 and GPT3 suggest that language models can be treated as knowledge bases. This directly motivates us to investigate if language models can also learn certain reasoning strategies. Studying these abilities would enable future research in using these models as dynamic knowledge bases that could infer new knowledge even when it is not "stored" directly (i.e. seen during pre-training). We will add this discussion to the paper.

@**R2**, **R3** **Natural Language results**. We thank you for highlighting the importance of results on natural language stories. We acknowledge that generalization is weaker in this harder setting, and we confirmed that by performing additional experiments on the natural language split. We still find that the proof resolution strategy influences the generalization capacity of TLMs. In particular, the conclusion that models trained on long, exhaustive proofs generalize better than short proofs still holds. We plan on moving discussion from the Appendix to the main section of the paper along with additional results.

@**R2**, **R3** **On the complexity of the task**. We acknowledge that the dataset in use (CLUTRR) is a toy dataset. However, this in turn allows us to carefully analyze and control the difficulty of the experiments. For instance, with 20 possible entities (k in Section 3.3) and 20 possible family kinship relationships, the model have to learn 8,000 possible triples. Given that even in this simplistic setup the generalization performance is not positive, this warrants a deeper inspection of reasoning mechanisms of TLMs. We plan on extending our experiments with other datasets in the future.

@**R2**, **R4** **On the issue of "all facts are seen" / "the correct answer is seen in the prefix"**. While our models have seen all possible facts in all training proofs, the target answer to a (story, question) pair is not seen in the prefix given to the model, unless the proof is explicitly given as in Section 4.3. Here, our experiments reveal that beyond 7-step proofs, the copy mechanism learned by TLMs becomes unreliable due to positional token embeddings. Position-agnostic embeddings could help in solving this issue, which we leave as an exercise for future work.

@**R4** **On the backward-chaining contradiction**. It is a great question why backward-chaining proofs are easier to use in Section 4.1, but are harder to generate in Section 4.2. We also agree with **R4**, this is due to the fact that backward chaining proofs contain the answer in the first proof step. Thus, there is a higher probability of the model to generate this step correctly and then use it while predicting the answer. This explains why the answer accuracy of such model is relatively high while their proof validity is low. We will note this phenomenon in the final version of the paper.

In general, we will fix typos and broken references (**R1**,**R2**,**R4**), clarify the presentation and some notations (**R1**,**R4**), and expand on the background section by discussing theorem proving (**R1**,**R4**). We would like to thank again all the reviewers for their time and effort in reading our paper and giving us good feedback and suggestions.