Dear Reviewers,

We sincerely thank you for your insightful comments and feedback. After carefully reading the comments, we made modifications to our work and highlighted the major changes and responses briefly here.

Q1: These two annotators have a Cohen's kappa agreement of 0.77, which seems to me quite low considering the task at hand. I would like the authors to discuss this a bit more since it is a crucial point about the validity of the dataset.

A1: Thanks for the question. The Cohen’s kappa was excellent in our understanding. We also noticed that the annotators have a disagreement on very challenging PLS such as “the difference of any complex number and its conjugate is always purely imaginary” or “the number of positive integers greater than or equal to n that are relatively prime to n is given by euler's totient function”. We will improve our data quality by employing high-quality annotators in our extension of the corpus.

Q2: The authors analyze the few-shot performance of Llama2 but do not share the prompts used.

A2: Due to the space limitation, we added detailed prompts for different models to our code repository.

Q3: I believe that using only Llama (especially considering that the smallest model is used) is not enough to generalize the behavior to other LLMs (e.g., "small-scale LMs are able to reason complex propositional logic but larger-scale LLMs fail.")

A3: Thanks to the author for pointing this out. We modified our statement to “These observations answer RQ2 that small-scale LMs are able to reason complex propositional logic but Llama2 fails, which suggests that increasing the size of LMs results in performance degradation (see BERT-base v.s. BLOOM-560m).”, which highlights the success of small-scale LMs in identifying correctness of a PLS.