We thank the reviewers for their constructive feedback and for stating that our work answers important questions around systematicity of reasoning skills, and paves a path towards open-domain systems that constantly improve by interacting with users.

In this work we focus on testing weather models can systematically reason over implicit knowledge. Thus, some of the design choices such as the use of synthetic data, distractors, and certain training mixtures, were designed to give better control toward answering this novel research question.

Below we answer all questions and provide results for requested additional experiments.

R1, R4: ***The use of synthetic data instead of natural language data.*** (a) The experiment in Section 4.4 uses **natural (not synthetic) language**. We show that the model achieves high accuracy here after training on synthetic data. Moreover, As suggested by R4, recent work hints that LMs, such as GPT-3 do increasingly well in bridging the gap between synthetic and natural language. (b) Because our main research question is the systematicity of reasoning, we want a setup where we have full control over the data presented to the model. Synthetic data is necessary to explore such novel research questions. (c) We agree that testing on natural language is desired and for the final version we will paraphrase automatically generated data using crowdsourcing.

R1: ***The quality and usefulness of the distractors is unclear.*** Thank you for your comment. Distractors are *crucial* for training. Without distractors the models find biases in the data that hurt generalization. For example, in the counting experiments, if no distractor member facts are shown, the model learns to count sentences, and ignores their content, failing to pay attention to the actual subjects of the member facts. We will clarify this in the final version.

R1: ***It would be interesting to see if this behaviour can be observed with no further training, just relying on how RoBERTa was pre-trained.*** This experiment was performed in Figure 4 of the RuleTaker paper (Clark et al., 2020), showing that performance is poor with few training examples.

R1: ***In the counting experiment, what happens if incorrect member facts are presented to the model when the subset=K*** Thanks for this interesting suggestion. We conducted this experiment, and found that the model still predicts 100% false, regardless of if the member fact is correct, suggesting it is counting the relevant member facts, rather than knowing them in advance. We will add this experiment to the final version.

R2: ***Can results in Section 4.1 be explained by distractor subject leaking?*** As explained in line 172 “We create development and test sets... where the subjects and objects are disjoint from the ones in the training set”. Because distractors are chosen from disjoint sets, leakage is not possible. We will clarify this in the final version.

R2: ***What happens if you do not do "context" dropout at training time*** Thank you for this question. Without dropout, the model learns to rely solely on explicit knowledge, achieving lower accuracy on implicit knowledge tests. The model predicts ‘False’ when relevant rules are missing from the explicit knowledge. We will add this to the camera-ready.

R2: ***What happens if the model is re-trained with only hypothesis information and labels, and evaluated in hypothesis-only*** Thank you for this suggestion. As suggested, we trained RoBERTa-Large on Hypothesis-only data and the Hypothesis-only test results are moderately higher: $65.2 \rightarrow 69.7$. This is still well below the Implicit-Reasoning accuracy . We will add this experiment to the camera-ready.

R2: ***Why do the authors choose to use a different architecture (ESIM) instead of using RoBERTa with randomly initialized weights?*** Because the size of the training data is relatively small, we were unable to train large transformer-based LMs directly on our data from scratch. Thus, we used the smaller ESIM + GloVe embeddings to compare to a model with less implicit knowledge.

R2: ***In the counting experiments, is the model really counting? This could be tested by dropping the quantity fact.*** This is an interesting suggestion. We conducted this experiment and accuracy drops from $73 \rightarrow 64.4$ (Table 3, counting $(1, K-1)$), which is similar to hypothesis-only 64.1, suggesting that the model is using the quantity fact for counting.

R3: ***Multihop is limited to 2hop only*** We agree that this can be extended but chose to focus on combining implicit and explicit knowledge rather than on the inference chain length, as done in RuleTaker. As stated in our related work section (line 364-369), the focus on systematicity distinguishes us from works on multi-hop QA.

R4: ***Synthetic data*** Please see response to R1 above.

All reviewers ***Clarifications, wording, figure 3, and missing reference*** We will clarify the writing of the experiments, specifically section 4.3 as requested by, mend the wording, improve figure 3, and add the missing reference, as requested by the reviewers.