Live Q&A Instructions

Welcome to the live Q&A document! Please follow the instructions below to ensure a smooth discussion:

- **Adding Questions**:
  Add your questions to the **bottom** of the list provided below. This helps maintain order and ensures all questions are visible to participants. If your question relates to a previous one, please indicate this in your question (e.g., "Related to question 3"). Kindly *refrain from editing, altering or adding to other participants' questions.*
- **Voting for Questions:**
  If you see a question you'd also like an answer to, add a +1 next to it. This helps prioritize popular questions.
- **Avoid Repetition**:
  Before adding a new question, (quickly) check the list to see if your query has already been addressed. If it has, feel free to upvote the existing question.

Add your Questions here:

- □ This is a sample question. Add your questions below. Vote for an existing question like this: “+1” Refer to previous questions like this: (Related to Question 1)
- ☑ ~~Claire how are you so cool?!? <3 Mollie~~
- □ Part 2: Regarding the non-memorizable task (placing books), it seems like the only remaining color being black could lead to the model answering “black” without really understanding 2nd from the right. Is there evidence that the answer is stemming from the location of the book? If the answer is based on location, is this consistent with “extrapolation”?
- □ Part 2: The higher order concepts that are learned in your discussion of in context learning. Does this mean that that higher-order learning isn’t happening at the pretraining? Only through ICL and fine-tuning?
- □ Part 3: Is there any salient correlation between the actual object / shape of object and the data points that remain outliers that don’t sort nicely into flat and round?
- □ Part 3: is the fact that grounding the abstract terms first (e.g., flat/round) and then the objects later is less helpful than the inverse indicative of a significant difference in how people learn? I.e. is there evidence that you’re aware of that children learn about the abstract properties prior to the individual objects?
- □ Part 3: (related to above) You mentioned that an explanation for the objects being more helpful than giving abstract labels first is that there may be some of the abstract property information encoded in the models. Do you have any sense of what kinds of properties are more commonly encoded (e.g. size vs. weight)? Does this have

implications for the match effects that you see when comparing implicit and explicit media?

□