Closing the Compositionality Gap in Language Models: A Study on Self-Ask Elicitive Prompting

What do you plan to do?
We plan to evaluate various language models' ability to perform compositional reasoning tasks using the method called “self-ask”, introduced in the paper MEASURING AND NARROWING THE COMPOSITIONALITY GAP IN LANGUAGE MODELS.

In the self-ask paper, the method was tested against GPT-3 models to improve chain of thought reasoning. We aim to study the compositionality gap in other LLMs, and evaluate if self-ask can help them reach the same or similar level of results in the original paper. We also want to study the effect of instruction-tuning on reasoning when provided self-ask prompts.

If time permits, we would like to expand on this research by applying the idea of Self-Instruct to reasoning tasks. We can use the guidelines of these two papers (paper 1, paper 2) for constructing synthetic demonstrations of various elicitive prompting strategies (self-ask, CoT, scratchpad, etc.) to augment existing reasoning QA datasets (symbolic reasoning, compositional reasoning, arithmetic reasoning, common sense reasoning). We then experiment with fine-tuning various LLMs of various sizes to see the improvement in zero-shot performance on reasoning tasks. This is similar to the approach taken in the STaR paper.

Why is it important, and why is it challenging?
LLMs struggle with compositional reasoning, but in-context learning has shown promise for improving performance in this domain. But the scaling law for various LLMs is still not understood. The self-ask paper has proved that their method can improve the performance on GPT-3 model families, but GPT-3 is only one of the LLMs currently available. If our research is successful, we can generalize the effects of self-ask elicitive prompting (and potentially other strategies) to other LLMs, as well as the potential of fine-tuning smaller models to generate rationales to see if smaller models can compete with larger models on compositional reasoning, which they currently cannot.

What dataset(s) will you use?
There are four datasets mentioned in the original paper,

- 2WikiMultiHopQA (Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps)
- MuSiQue (MuSiQue: Multihop Questions via Single-hop Question Composition)
- Compositional Celebrities (MEASURING AND NARROWING THE COMPOSITIONALITY GAP IN LANGUAGE MODELS)
- Bamboogle (MEASURING AND NARROWING THE COMPOSITIONALITY GAP IN LANGUAGE MODELS)

We plan to use the 2WikiMultiHopQA dataset, and if time permits, use the other datasets as well.

What algorithms might you use? Are good implementations available, or will you need to write your own? (Don't worry if you can't answer this well at this stage of the course.)
We plan to implement the same "self-ask" algorithm mentioned in (MEASURING AND NARROWING THE COMPOSITIONALITY GAP IN LANGUAGE MODELS) on a subset of the following pre-trained language models (we distinguish between instruction-tuned and vanilla models since this will be a point of research):
Vanilla

- T5
- LLaMA
- PaLM

Instruction-Tuned

- T0
- Alpaca
- TK-INSTRUCT
- Dolly
- FLAN

References

1. MEASURING AND NARROWING THE COMPOSITIONALITY GAP IN LANGUAGE MODELS
2. Self-Instruct: Aligning Language Models with Self-Generated Instructions
3. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
4. Synthetic Prompting: Generating Chain-of-Thought Demonstrations for Large Language Models
5. Automatic Chain of Thought Prompting in Large Language Models
6. STaR: Bootstrapping Reasoning With Reasoning