Some Observations on the CoT Paper

Paper categories: prompt engineering, LLM capability probing

- The paper shows that it's ok to talk about your motivations and where the ideas came from, e.g., the CoT paper came from the earlier papers on reasoning with intermediate steps (by Ling et. al. and Cobbe et. al.) and from the original "few-shot learners" paper. Basically, it's ok to talk about how the genesis of the idea happened.
- The paper also lists "several attractive properties" of the technique.
- The paper is largely devoted to performance (of various LLM collections with various sizes per collection) on various benchmarks such as GSM8K.
- In the paper, they compare their innovation (few-shot CoT) to a baseline - standard few-shot. They don't compare to standard zero-shot.
- Notice how the "key takeaways" in the "Results" section of the paper are enumerated as "first", "second" & "third", and each gives some amazing insights - emergent, works best on complex problems & beats benchmark on 3 out of 5 benchmarks.
- In the paper, manual examination is used to delineate correct chains of thought leading to correct answer vs. incorrect chains of thought coincidentally leading to correct answer.
- In the paper, an ablation study is done to provide evidence that it is CoT (not something else) which leads to the improvement in performance.
- In the paper, the robustness tests are used to demonstrate that the technique isn't overly sensitive to a particular linguistic style, different exemplars, the order of the exemplars, a particular number of exemplars, etc. In other words, the tests are used to demonstrate that the technique generally works.
- In the paper, the authors show that the CoT technique generalizes to out-of-domain for which evaluation examples had more steps than those in the exemplars.