- The key discovery of this paper is that few-shot CoT exemplars aren't necessary to elicit reasoning from LLMs. (However, few-shot CoT prompting does perform better than zero-shot CoT prompting.)
- The paper *"hints at untapped and understudied fundamental zero-shot capabilities of LLMs, suggesting high-level, multi-task broad cognitive capabilities may be extracted by simple prompting"*.
- The authors show that zero-shot CoT should be *"the minimal strongest zero-shot baseline for the challenging reasoning benchmarks"*.
- The authors highlight *"the importance of carefully exploring and analyzing the enormous zero-shot knowledge hidden inside LLMs before crafting finetuning datasets or few-shot exemplars"*.
- The key difference between few-shot and (standard) zero-shot prompting: *"The success of large language models (LLMs) is often attributed to (in-context) few-shot or zero-shot learning. It can solve various tasks by simply conditioning the models on a few examples (few-shot) or instructions describing the task (zero-shot)."*
- *"A language model is a model that looks to estimate the probability distribution over text... The method of conditioning the language model is called 'prompting'."*
- Prior to this paper, prompt engineering was very task-specific. While few-shot CoT required task-specific exemplars (e.g., the set of exemplars for math word problems was different from the set of exemplars for common sense QA), standard zero-shot required the use of task-specific instructions &/or templates. In comparison, the *"Let's think step by step."* prompt is totally task-agnostic.
- The authors have benchmarked zero-shot CoT with various other prompting techniques: standard zero-shot, standard few-shot, few-shot CoT, zero-plus-few-shot-CoT (a hybrid approach). See Table 2.
- The authors have evaluated each of these prompting techniques on 12 datasets from 4 categories of reasoning tasks: arithmetic, commonsense, symbolic and other logical reasoning tasks.
- The main experiments have been conducted with various sizes of the following models: (i) InstructGPT-3, (ii) original GPT-3 and (iii) PaLM.
- Zero-shot CoT actually makes use of two-stage prompting. First, a reasoning extraction prompt *"Let's think step by step."* is used to generate a reasoning path. This is followed by an answer extraction prompt such as *"Therefore, among A through E, the answer is "* (for multiple choice questions) and *"Therefore, the answer (Arabic numerals) is "* (for math problems requiring numerical answers). Finally, some answer cleansing rules are used to cleanse the answers.
- *"Zero-shot-CoT substantially outperforms four out of six arithmetic reasoning tasks (MultiArith, GSM8K, AQUA, SVAMP), all symbolic reasoning, and all other logical reasoning tasks... Our method gives on-par performances for the remaining two arithmetic reasoning tasks (SingleEq and AddSub), which is expected since they do not require multi-step reasoning... In commonsense reasoning tasks, Zero-shot-CoT does not provide performance gains."*

- On MultiArith & GSM8K: *"While Zero-shot-CoT naturally underperforms Few-shot-CoT, it substantially outperforms standard Few-shot prompting with even 8 examples per task."*
- In addition, the authors have done a *"model scale study"* (with various other models - in addition to the above). This study explores how the zero-shot CoT prompting technique's performance improves with model scale. (This is just like in the few-shot CoT paper.) See Figure 3. *"Importantly, with our single fixed prompt, zero-shot CoT has a significantly better scaling curve comparable to that of the few-shot CoT baseline."* -> This means that the performance of zero-shot CoT gets significantly better with model scale (in contrast to standard zero-shot). In fact, the scaling curve is comparable to few-shot CoT. *"Without chain of thought reasoning, the performance does not increase or increases slowly as the model scale is increased, i.e., the curve is mostly flat. In contrast, the performance drastically increases with chain of thought reasoning, as the model size gets bigger, for Original/Instruct GPT-3 and PaLM. When the model size is smaller, chain of thought reasoning is not effective... We also manually investigated the quality of generated chain of thought, and large-scale models clearly demonstrate better reasoning..."*
- The authors have performed error analysis: *"In commonsense reasoning (CommonsenseQA), Zero-shot-CoT often produces flexible and reasonable chain of thought even when the final prediction is not correct. Zero-shot-CoT often outputs multiple answer choices when the model finds it is difficult to narrow it down to one... In arithmetic reasoning, Zero-shot-CoT tends to output unnecessary steps of reasoning after getting the correct prediction, which results in changing the prediction to incorrect one. Zero-shot-CoT also sometimes does not start reasoning, just rephrasing the input question."*
- The authors have done a robustness analysis by testing how variations of the *"Let's think step by step."* prompt perform. See Table 4.
- *"We also show that besides Few-shot-CoT requiring human engineering of multi-step reasoning prompts, their performance deteriorates if prompt example question types and task question types are unmatched, suggesting high sensitivity to per-task prompt designs."* -> See Table 5. This is expected. In fact, it would have been surprising if the few-shot exemplars for arithmetic reasoning worked well for commonsense reasoning or symbolic reasoning (or vice versa).
- *"... big performance increases from Zero-shot to Zero-shot-CoT in all recent large models and consistent improvements in both arithmetic and non-arithmetic tasks suggest that the models are unlikely simply memorising, but instead capturing a task-agnostic multi-step reasoning capability for generic problem solving... and dataset details in InstructGPT also confirm that it is not specially engineered for multi-step reasoning."* -> In other words, multi-step reasoning capabilities in LLMs emerged even though they were not explicitly engineered for them. Probably only the latest LLMs such as o1 (Strawberry) have been explicitly engineered for solving multi-step reasoning problems.
- The authors highlight an open question: How to automatically create better templates (prompts) for zero-shot CoT?

- *"We hope our work can serve as a reference for accelerating not just logical reasoning research with LLMs, but also discovery of other broad cognitive capabilities within LLMs... Our simple method... encourages the community to further discover similar multi-task prompts that elicit broad cognitive abilities instead of narrow task-specific skills." -> "Narrow task-specific skills" are distinguished from "broad cognitive abilities".*
- **Question:** What other types of cognitive capabilities do humans possess? Are some of them present in LLMs?