Exploring Zero-Shot Tabular Chain-of-Thought Common Sense Reasoning

Kaya Gouin
School of Computer Science
Carleton University
Canada

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

Pretrained Large Language Models (LLMs) are models which estimate probability distributions over text. In recent years, they have shown remarkable success on a wide range of natural language processing tasks. Regarded as excellent few-shot reasoners when provided with carefully engineered task-specific exemplars and decent zero-shot reasoners when provided with a clear description of the task, LLMs derive their success in part from these reasoning abilities. With growing interest in both large language models and natural language processing, much research has been dedicated to the design of optimal methods of conditioning—referred to as prompting—LLMs.

Although large language models excel at single-step reasoning tasks, they demonstrate considerably lower accuracy with respect to complex reasoning tasks which require carefully structured step-by-step reasoning. To address this shortcoming, Wei et al. proposed in 2022 a *chain-of-thought* prompting method whereby the LLM is provided with exemplars of question/answer pairs in which answers each demonstrate the series of intermediate reasoning steps employed to solve the question [Wei2022]. With these exemplars, the language model produces a consecutive reasoning sequence, which ends with a final answer [Wei2022]. Contrasted with standard few-shot prompting, whereby the provided answers lack reasoning steps, few-shot chain-of-thought prompting demonstrates higher accuracy on a variety of arithmetic, common sense, and symbolic reasoning tasks [Wei2022].

Following the work of Wei et al., Wang et al. proposed a slightly modified version of the few-shot chain-of-thought prompting method whereby they sample a diverse set of reasoning paths produced by the large language model [Wang2022]. From there, they segregate chain-of-thought reasoning sequences from their associated final answers and select the most consistent answer within the final answer set [Wang2022]. This *self-consistency* technique has shown distinctively higher accuracy on both arithmetic and common sense reasoning tasks when compared to the original work of Wei et al. [Wang2022].

In an effort to ameliorate the easy-to-hard generalizability of the few-shot chain-of-thought methods put forth by Wei et al. and Wang et al., Zhou et al. proposed a *least-to-most* prompting technique [Zhou2022]. A term borrowed from educational psychology, least-to-most prompting demonstrates to LLMs an approach to solving complex reasoning tasks by decomposing them into series of simpler subtasks, then solving them sequentially [Zhou2022]. When provided with these rigid exemplar templates, LLMs exhibit a net increase in accuracy on compositional generalization tasks when compared with both few-shot and few-shot chain-of-thought prompting methods [Zhou2022].

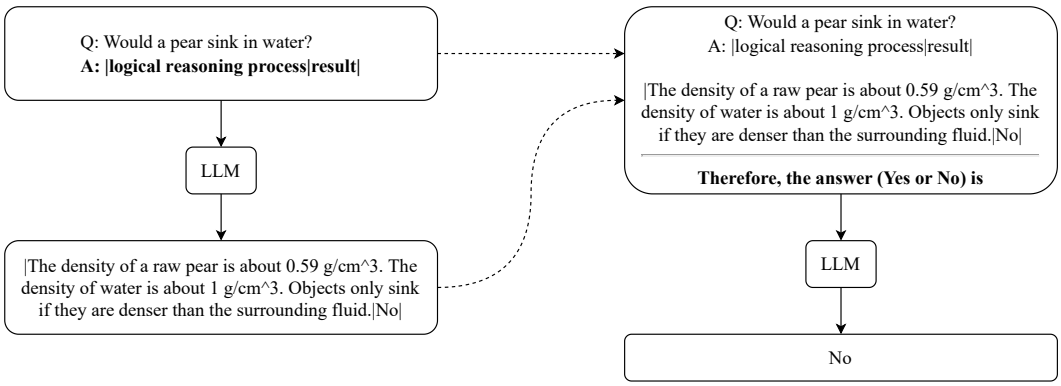

Figure 1: Pipeline of CS-Tab-CoT as described in section 2.

It was Kojima et al. that first explored a task-agnostic chain-of-thought prompting method. Here the authors wanted a more generalizable method, a task-agnostic method, and also wanted to show the untapped potential of Zero-Shot-CoT. These chain-of-thought self-augmented prompts lead to refined outputs with higher accuracies on natural language processing tasks than ones generated by standard zero-shot prompting.

Following the same avenue as Zhou et al. and Kojima et al., Jin and Lu proposed tabular chain-of-thought prompting. Here they were looking at a highly organized way of structuring the reasoning process in a zero-shot context.

A common theme with existing methods, as we'll see, is that common sense reasoning is especially hard for LLMs. Techniques that have been put forward to better condition LLMs are often focused on arithmetic tasks and other tasks that require careful systematic reasoning. Common sense reasoning is different. It is an inherently human trait. Requires experience, intelligence, and creativity. The way in which humans reason about common sense reasoning tasks is highly variable. This depends on the specific task, but across the board humans don't really reason in a step-by-step fashion when it comes to common sense reasoning. It's more of a feeling, or an obvious answer based on previous experience and general knowledge of the world. Again, intelligence, creativity, and even imagination are key aspects of our common sense reasoning skills.

These methods were designed with arithmetic-type tasks in mind, where we often have questions that can naturally be divided into subquestions, solved, then used to arrive at a final answer. For common sense reasoning questions however, the path to arrive at the answer isn't as clear. For the 'Would a pear sink in water' question, I knew the answer was no, not because I knew the density of a pear in relation to the density of water, but because I have experience handling pears, and although I haven't seen a pear float in water, I've seen apples float in water, and could easily imagine a pear floating in water. This kind of reasoning is what humans use to answer common sense questions. We use our experience and our imagination to come up with the most likely answer. Common sense reasoning tasks aren't structured in such a way to promote easy decomposition into simpler subquestions.

Given the ____ nature of common sense reasoning tasks, ...

In this work, we follow the work of Jin and Lu and we suggest a novel way of conditioning LLMs... We present an augmented tabular-format chain-of-thought prompting method designed to elicit high performance on common sense reasoning tasks. And beyond seeking high performance, we're interested in knowing how different types of reasoning instructions (don't like how this is said) elicit different reasoning paths. We test it against established baselines. We compare it to a greater range of common sense reasoning tasks as compared to the previous works we'll be talking about (if we have the chance to do so).

2 CS-Tab-CoT

To overcome ____ on common sense reasoning tasks, we created a flexible tabular prompt template designed to elicit highly structured chain-of-thought sequences. The core of the flexible template is simple: we instruct the model to reason according to a specific form of logical inference.

With a two-stage prompting method like those of CoT and Tab-CoT, the common sense tabular method—which we coined CS-Tab-CoT—makes use of a *self-augmented* prompt. Figure 1 depicts the CS-Tab-CoT method within the context of a common sense reasoning task.

We explored thirteen different types of reasoning, each of which being a unique form of inference, ten of which are designed to elicit correct reasoning structures, and three of which are designed to elicit incorrect reasoning structures. This selection of reasoning types came from a fair amount of research into common sense intelligence and reasoning in humans. At this point in the experimental design, I was hopeful that they would lead not just to different accuracies, but to different reasoning structures, since these listed words have specific meanings that have to do with different forms of inference.

For the experiment itself, I used `text-davinci-002` from the GPT-3.5 family as my large language model. For the datasets, I used a sample of 100 questions from the CommonSenseQA dataset (these are multiple-choice common sense questions), and a sample of 100 questions from the StrategyQA dataset (these are Yes or No common sense questions). I used samples of 100 questions due to limited resources. And, to compare my prompts to those of others, I compared CS-Tab-CoT to Standard, CoT, and Tab-CoT prompts with the same model and the same 100 questions drawn from each dataset.

Here we look at the two datasets used in some previous works presented: CommonSenseQA [Talmor et al., 2019] and StrategyQA [Geva et al., 2021]. I we have the chance, we might look at some extra ones from https://commonsense.run/datasets/.

Here we're mainly interested in comparing methods for zero-shot contexts. We'll look at standard, CoT, and Tab-CoT (all of these are in zero-shot contexts).

3 Results and Discussion

Compare our work to zero-shot, zero-shot-cot, and tab-cot.

Here I present the accuracies of CS-Tab-CoT. Although there is some variation between the accuracies of the different prompts, there isn't as much as I had expected. If we look at the CommonSenseQA column, we see that the inductive reasoning process achieved the highest accuracy at 65 percent. With these being multiple-choice questions, with 4 choices on average, the performance is somewhat good. If we look at the StrategyQA column, we see that the erroneous reasoning process achieved the highest accuracy at 51 percent. This was unexpected, considering that we asked the model to output an answer derived from error. Additionally, with these being Yes or No questions, it is especially surprising to see that this erroneous prompt was the only one that lead to an accuracy beyond 50 percent, beyond a random guess. If we look at the Avg column, in which we take the average of the accuracies for CommonSenseQA and StrategyQA for each prompt, we see that the best overall prompt is the logical reasoning process, with an average of 54 percent. Before I move on to the next set of results, I want to note another unexpected result: the illogical reasoning prompt is not the complement of the logical reasoning prompt. Since these are antonyms, I had thought that their accuracies would sum to 100. Table 1

Here I present the accuracies of the three best CS-Tab-CoT prompts just mentioned with the accuracies of Stan- dard, CoT, and Tab-CoT prompts. Since I don't have much time left, I'll only point out that there isn't much variation between accuracies within a given table. Table 2.

4 Conclusion and Future Work

Overall, in the context of common sense reasoning, the model minimally responds to instructions on chain-of- thought reasoning, whether that be within the CS-Tab-CoT prompts, or across prompting methods. This implies that, in the context of common sense reasoning, the model doesn't have a fixed reasoning pattern and largely ignores the chain-of-though prompt.

Table 1: CS-Tab-CoT accuracies on common sense reasoning tasks

CoT Prompt	CommonSenseQA	StategyQA	Avg
`|logical reasoning process|result|`	61.0	47.0	**54.0**
`|deductive reasoning process|result|`	59.0	47.0	53.0
`|inductive reasoning process|result|`	**65.0**	36.0	50.5
`|abductive reasoning process|result|`	62.0	43.0	52.5
`|inferential reasoning process|result|`	61.0	39.0	50.0
`|analogical reasoning process|result|`	59.0	38.0	48.5
`|critical reasoning process|result|`	59.0	24.0	41.5
`|counter-factual reasoning process|result|`	56.0	48.0	52.0
`|cause-for-effect reasoning process|result|`	63.0	17.0	40.0
`|creative reasoning process|result|`	61.0	35.0	48.0
`|illogical reasoning process|result|`	52.0	38.0	45.0
`|delusional reasoning process|result|`	53.0	37.0	45.0
`|erroneous reasoning process|result|`	53.0	**51.0**	52.0

Table 2: Accuracies of four prompting methods compared to best-accuracy CS-Tab-CoT

Method	Prompt	CommonSenseQA
Standard Prompting	-	67.0
CoT	`Let's think step by step`	**68.0**
Tab-CoT	`|step|subquestion|process|result|`	65.0
CS-Tab-CoT	`|inductive reasoning process|result|`	65.0

Method	Prompt	StategyQA
Standard Prompting	-	43.0
CoT	`Let's think step by step`	44.0
Tab-CoT	`|step|subquestion|process|result|`	50.0
CS-Tab-CoT	`|erroneous reasoning process|result|`	**51.0**

Method	Prompt	Avg
Standard Prompting	-	55.0
CoT	`Let's think step by step`	56.0
Tab-CoT	`|step|subquestion|process|result|`	**57.5**
CS-Tab-CoT	`|logical reasoning process|result|`	54.0

Disclosure

The research presented in this paper has been conducted solely for the purpose of the Advanced Machine Learning course project.

References

[Geva2021] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *TACL*, 9:346-361, 2021. URL https://aclanthology.org/2021.tacl-1.21/.

[Jin2023] Ziqi Jin and Wei Lu. Tab-CoT: Zero-shot Tabular Chain of Thought, 2023. arXiv:2305.17812.

[Kojima2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa. Large language models are zero-shot reasoners, 2022. arXiv:2205.11916.

[Talmor2019] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149-4158, 2019. URL https://aclanthology.org/N19-1421/.

[Wang2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2022. arXiv:2203.11171.

[Wei2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models, 2022. arXiv:2201.11903.

[Zhou2022] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2022. arXiv:2205.10625.