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].

With zero-shot prompting methods left largely unexplored, Kojima et al. were the first to design a task-agnostic chain-of-thought prompting method they coined Zero-shot-CoT [Kojima2022]. At the core of their method is the simple phrase *Let's think step by step* [Kojima2022]. When appended to the input question, the phrase elicits a reasoning sequence much like its few-shot counterpart, despite the language model not having been exposed to such a reasoning process [Kojima2022]. The versatile Zero-shot-CoT demonstrates much higher accuracy than standard zero-shot prompting on arithmetic, symbolic, and other reasoning tasks [Kojima2022]. On certain arithmetic tasks, the accuracy of Zero-shot-CoT even surpasses that of standard few-shot prompting [Kojima2022].

Following the same avenue as Kojima et al., Jin and Lu proposed Tab-CoT: a zero-shot tabular chain-of-thought prompting method [Jin2023]. As in Zero-shot-CoT, Tab-CoT makes use of a task-agnostic prompt to elicit a reasoning sequence [Jin2023]. Eliciting highly structured, table-embedded reasoning sequences, Tab-CoT stems from the discovery that state-of-the-art large language models have a predisposition for reasoning over tabular structured data [Jin2023]. This novel prompt leads to high accuracies on arithmetic, symbolic, and other reasoning tasks [Jin2023].

With carefully designed prompts which elicit structured step-by-step reasoning, LLMs are able to succeed on a variety of natural language processing tasks. Common to all conditioning methods aforementioned, however, is the relatively low accuracy exhibited on common sense reasoning tasks, this being especially true of zero-shot methods. This deficiency in common sense reasoning may partially be attributed to the flexible reasoning paths often needed to solve questions in such a category. Admittedly, common sense reasoning tasks may not be structured in such a way to promote easy decomposition into simpler subquestions.

With this in mind, we extend the work of Jin and Lu by exploring a novel way of conditioning LLMs which allows for slightly more flexibility in the reasoning sequence while maintaining a sensible level of structure. We coin this method CS-Tab-CoT.

2 CS-Tab-CoT

A zero-shot tabular chain-of-thought method specifically designed for common sense reasoning, CS-Tab-CoT is at its core an adaptable prompt template designed to elicit reasoning sequences consistent with specific forms of logical inference. The template is formatted as follows, where the underscores are replaced with the form of inference wished to be used:

```
|____ reasoning process|result|
```

With this template, we explored thirteen forms of logical inference: ten of which were designed to elicit correct reasoning structures, and three of which were designed to elicit incorrect reasoning structures.

- logical
- deductive
- inductive
- abductive
- inferential
- analogical
- critical
- counter-factual
- cause-for-effect
- creative
- illogical
- delusional
- erroneous

As with both Zero-shot-CoT and Tab-CoT methods, CS-Tab-CoT prompts the LLM twice so as to extract both chain-of-thought reasoning and answer. Illustrated in Figure 1, we begin by appending the template, which incorporates the chosen form of inference, to the common sense question we wish the LLM to answer. Together, the question/template pair is fed to the LLM so as to extract the tabular chain-of-thought reasoning. The question/template pair is then concatenated with the reasoning sequence, then fed to the LLM along with the desired answer format so as to extract the answer.

As a consequence of the two-step prompting method, the concatenated prompt fed to the language model in the answer extraction step is in fact a *self-augmented* prompt. It is the self-augmented quality of CS-Tab-CoT that enables the extraction of such a refined answer.

To estimate the capabilities of LLMs when conditioned via the CS-Tab-CoT method, we consider samples of 100 questions from two common sense datasets: CommonSenseQA (a collection

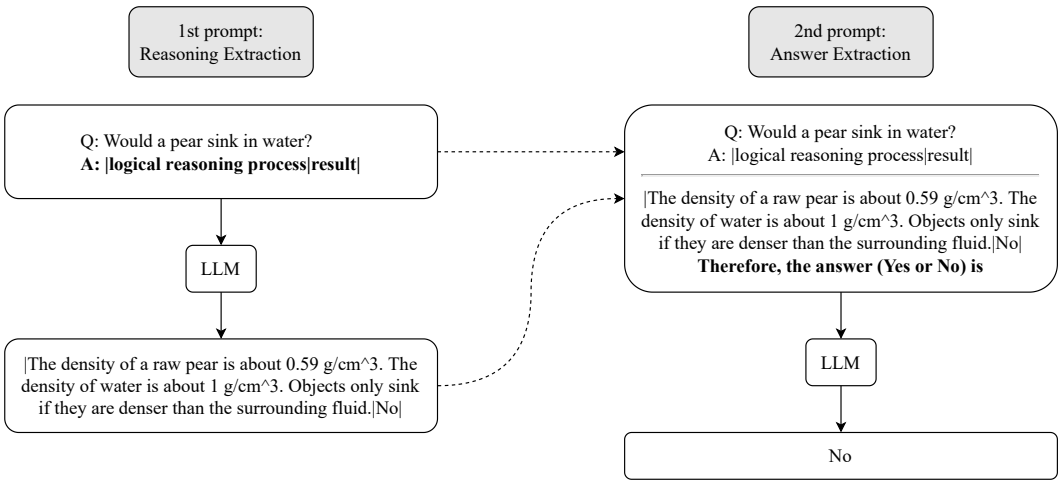

Figure 1: Pipeline of CS-Tab-CoT as described in section 2.

of multiple-choice questions) [Talmor2019] and StrategyQA (a collection of yes-no questions) [Geva2021]. We evaluate CS-Tab-CoT through Instruct-GPT3 (`text-davinci-002`) [Ouyang2022], and compare the achieved accuracies to those of standard, Zero-shot-CoT, and Tab-CoT methods. For a fair comparison, we evaluate each method with the same model and question samples.

3 Results and Discussion

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.

Future work: look at the full CommonSenseQA and StrategyQA datasets, and some extra datasets from https://commonsense.run/datasets/ in order to gain a better overview/estimate of true common sense reasoning abilities. Test on more LLMs, especially more recent ones.

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.

[Ouyang2022] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.