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

Kaya Gouin
School of Computer Science
Carleton University
Canada

Abstract

Large language models have the potential for common sense reasoning. Recent studies have explored methods of conditioning these language models to boost accuracy on arithmetic and symbolic reasoning tasks, among others. These types of tasks are naturally suited to decomposition into simpler subproblems. Common sense reasoning tasks, however, are not necessarily divisible into subproblems. Following previous work on the methods of conditioning large language models, we explore an approach to zero-shot chain-of-thought reasoning specifically designed for common sense tasks. Our results show that large language models are minimally responsive to different chain-of-thought conditioning instructions for common sense tasks. Ours is a pilot study for a promising avenue of research into large language models and their common sense reasoning abilities.

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 that 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 [8]. With these exemplars, the language model produces a consecutive reasoning sequence, which ends with a final answer [8]. 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 [8].

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 [7]. 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 [7]. 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. [7].

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 [10].

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 [10]. 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 [10].

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 [4]. At the core of their method is the simple phrase *Let's think step by step*. 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 [4]. The versatile Zero-shot-CoT demonstrates much higher accuracy than standard zero-shot prompting on arithmetic, symbolic, and other reasoning tasks [4]. On certain arithmetic tasks, the accuracy of Zero-shot-CoT even surpasses that of standard few-shot prompting [4].

Following the same avenue as Kojima et al., Jin and Lu proposed Tab-CoT: a zero-shot tabular chain-of-thought prompting method [2]. As in Zero-shot-CoT, Tab-CoT makes use of a task-agnostic prompt to elicit a reasoning sequence [2]. 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 [2]. This novel prompt leads to high accuracies on arithmetic, symbolic, and other reasoning tasks [2].

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 (those listed in the first two columns below), and three of which were designed to elicit incorrect reasoning structures (those listed in the last column below).

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

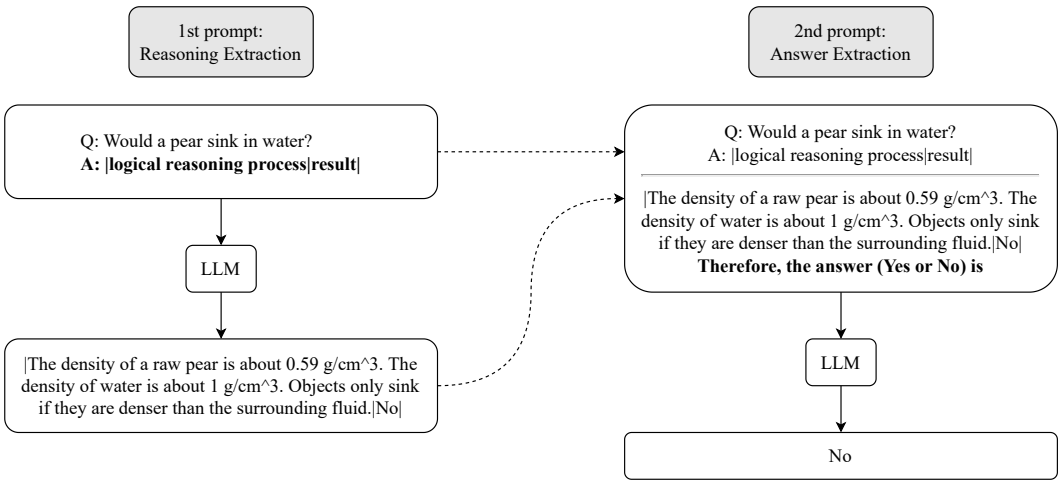

Figure 1: Pipeline of CS-Tab-CoT as described in section 2.

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 with the CS-Tab-CoT method, we consider samples of 100 questions from two common sense datasets: CommonSenseQA (a collection of multiple-choice questions) [6] and StrategyQA (a collection of yes-no questions) [1]. We evaluate CS-Tab-CoT through Instruct-GPT3 (`text-davinci-002`) [5]. Unique in their form of logical inference, we compare the achieved accuracies of each templated prompt with those of their counterparts in order to determine the interrelation of the CS-Tab-CoT prompts. As well, we compare to the accuracies of the best CS-Tab-CoT prompts 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

3.1 CS-Tab-CoT Accuracies

Presented in Table 1 are the accuracies of CS-Tab-CoT, arranged by templated prompt.

As displayed in the "CommonSenseQA" column, the *inductive* reasoning prompt achieved the highest accuracy, at sixty-five percent. This result is consistent with our intuition, since inductive reasoning is a method well suited to common sense reasoning, in which a general principle is derived from a set of observations. With these being multiple-choice questions with 4 choices on average, the level of accuracy is respectable.

Displayed in the "StrategyQA" column, the *erroneous* reasoning prompt achieved the highest accuracy, at fifty-one percent. This result is unexpected, since erroneous reasoning is a method in which an answer is derived from error. By using such a method, one would expect the LLM to output an incorrect answer more often than not. Furthermore, with these being yes-no questions, the fact that the erroneous reasoning prompt was the only one that achieved an accuracy beyond fifty percent is perplexing.

Displayed in the "Avg" column, in which we take the average of the accuracies of CommonSenseQA and StrategyQA for each prompt, the *logical* reasoning process appears to be the best overall prompt, with an average of 54 percent. Since logical reasoning is a process in which a conclusion is drawn from a set of premises by way of careful reasoning, this, as with the first presented result, is consistent with our intuition.

Comparing the *logical* reasoning prompt with the *illogical* reasoning prompt, we notice that the two are not complements of one another. Since these are antonyms, we had hypothesized that their accuracies would roughly sum to one hundred percent. More generally, we had expected a clear

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

CS-Tab-CoT Prompt	CommonSenseQA	StrategyQA	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

division in the accuracies of the prompts designed to elicit correct responses versus those designed to elicit incorrect ones; a property not exhibited by the model. In actuality, the three prompts designed to elicit incorrect responses, namely the *illogical*, *delusional*, and *erroneous* reasoning prompts, have much higher accuracies than previously envisioned.

Considering Table 1 in its entirety, we note that the accuracy levels of each prompt is affected by the specific dataset on which the prompt was evaluated; a quality most prominent in the *cause-for-effect* reasoning prompt, with which the model achieved an accuracy of sixty-three percent on the CommonSenseQA dataset and an accuracy of seventeen percent on the StrategyQA dataset. Together with the fact that these CS-Tab-CoT prompts each vary in accuracy in relation to the dataset on which the prompt was evaluated, the fact that there is no single prompt which exhibits best accuracies over both datasets implies that specific forms of logical inferences may be better suited to specific types of common sense reasoning tasks.

3.2 Cross Comparison of Accuracies

Presented in Table 2 are the accuracies of standard, Zero-shot-CoT, and Tab-CoT, versus best-CS-Tab-CoT prompting methods. As described in Section 3.1, the *inductive*, *erroneous*, and *logical* reasoning processes achieved the highest accuracies for the CommonSenseQA dataset, the StrategyQA dataset, and the average between the two datasets, respectively. These three reasoning processes constitute best-CS-Tab-CoT prompting methods, and are displayed in their respective subtables of Table 2.

Inspecting the first subtable, we note that the best-performing CS-Tab-CoT prompt for the CommonSenseQA dataset, namely the *inductive* reasoning prompt, has an accuracy equivalent to that of Tab-CoT, and an accuracy lower than those of Zero-shot-CoT and standard prompting. This demonstrates that, with respect to the CommonSenseQA dataset, tabular inductive reasoning has as much value as step-by-step tabular reasoning, and less value than both unrestrictive chain-of-thought reasoning and no explicit reasoning.

Inspecting the second subtable, we note that the best-performing CS-Tab-CoT prompt for the StrategyQA dataset, namely the *erroneous* reasoning prompt, has an accuracy which surpasses those of Tab-CoT, Zero-shot-CoT, and standard prompting. This demonstrates that, with respect to the StrategyQA dataset, tabular erroneous reasoning has greater value than step-by-step tabular reasoning, unrestrictive chain-of-thought reasoning, and no explicit reasoning. This finding is inconsistent with our intuition, as erroneous reasoning is a form of reasoning which is not typically expected to lead to positive results.

Inspecting the third subtable, we note that the best-performing CS-Tab-CoT prompt when considering the average of accuracies over the CommonSenseQA and StrategyQA datasets, namely the *logical* reasoning prompt, has an accuracy which is lower than those of Tab-CoT, Zero-shot-CoT, and standard prompting. This demonstrates that, with respect to the the average of CommonSenseQA and

Table 2: Accuracies of three prompting methods compared to best-accuracies of 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

StrategyQA dataset accuracies, tabular logical reasoning has lesser value than step-by-step tabular reasoning, unrestrictive chain-of-thought reasoning, and no explicit reasoning.

Though there are differences in accuracy within each subtable of Table 2, the differences are minimal. Specifically, the way in which we prompt the language model has little impact on accuracy when it comes to common sense reasoning tasks.

Considering every conditioning method and dataset, StrategyQA appears to be the most difficult dataset on which to do well—an attribute common to every explored method. By manually scrutinizing a subset of question/answer pairs from the dataset, we discovered many questions which required multi-step reasoning in addition to a global perspective. With large language models being statistical models which estimate probability distributions over text, if a question consists of concepts which are distantly related and would by extension require many reasoning steps in order to potentially link these concepts, the model has a low chance of success. Consider the following question/answer pair, taken from StrategyQA [1]:

> Q: Has Burger King contributed to a decrease in need for snowshoes?
> A: Yes

At first sight, even for humans, the question seems absurd. Upon further reflection however, we see a sound reasoning path that leads to the answer of 'yes'. To correctly answer this question, the LLM must have some representation of the concepts 'Burger King' and 'snowshoes', which are likely distantly related. From there, the model must make use of multi-step reasoning, along with a comprehensive view of the world to arrive at a reasoning path approximately structured like so:

> Burger King serves beef. Beef farming is associated with increased global temperatures and decreased snowfall. Snowshoes are used in the snow. If there is a decrease in snowfall then there is a decrease in need for snowshoes. Therefore, Burger King has contributed to a decrease in need for snowshoes.

This type of reasoning structure is involved, and difficult to achieve. If most questions from the StrategyQA dataset require a similar number of steps and similar level of global perspective as the Burger King question, then we expected the LLM to exhibit poor accuracy regardless of the prompting method, given the method's limited impact on common sense reasoning—a result observed in Table 1. Although the prompting method has limited impact on this class of reasoning, if LLMs exhibit poor accuracy when attempting to reason correctly, then an erroneous reasoning process might achieve higher accuracy—a result observed in Tables 1 and 2.

4 Conclusion and Future Work

Overall, in the context of common sense reasoning, it appears pretrained large language models minimally respond to instructions on chain-of-thought reasoning—whether that be within the templated CS-Tab-CoT prompts, or across prompting methods. This implies that language models do not have fixed reasoning patterns when reasoning about common sense tasks. In fact, it seems as though LLMs largely ignore the chain-of-thought prompt.

Limitations to this work include the use of a single language model, and modest question samples of only two datasets. With such modest resources, conclusions drawn regarding the reasoning abilities of LLMs are to be taken as preliminary.

Further research should consider the unabridged CommonSenseQA and StrategyQA datasets, as well as some additional common sense reasoning datasets so as to better estimate the common sense reasoning abilities of LLMs. To test the full capabilities of large language models on common sense reasoning tasks, it is crucial that we take an approach grounded in cognitive science and linguistic principles. Following this method, a recent paper published in *Nature* proposes nine text-based evaluation modalities spanning multiple common sense categories such as comprehension, counterfactual reasoning, probabilistic judgements, and psychosocial modelling [3]. Integrating these with visually-grounded evaluation modalities such as one by Zellers et al. [9] would constitute a strong test for measuring true common sense reasoning abilities of LLMs.

Likewise, further research should consider evaluating zero-shot common sense conditioning methods on a wide range of language models, with specific attention given to model size and novelty, since past research has shown these qualities to affect accuracy [2, 4, 7, 8].

Disclosure

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

References

[1] 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/.

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

[3] Mayank Kejriwal, Henrique Santos, Alice M. Mulvehill and Deborah L. McGuinness. Designing a strong test for measuring true common-sense reasoning. *Nature Machine Intelligence*, vol. 4, no. 4, pages 318-22, 2022. URL: https://doi.org/10.1038/s42256-022-00478-4.

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

[5] 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.

[6] 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/.

[7] 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.

[8] 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.

[9] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: visual commonsense reasoning, 2018. arXiv:1811.10830

[10] 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.