

MultiTool-CoT: GPT-3 Can Use Multiple External Tools with Chain of Thought Prompting

Tatsuro Inaba¹ Hirokazu Kiyomaru¹ Fei Cheng¹ Sadao Kurohashi^{1,2}

¹Kyoto University, Japan

²National Institute of Informatics, Japan

{inaba, kiyomaru, feicheng, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Large language models (LLMs) have achieved impressive performance on various reasoning tasks. To further improve the performance, we propose MultiTool-CoT, a novel framework that leverages chain-of-thought (CoT) prompting to incorporate multiple external tools, such as a calculator and a knowledge retriever, during the reasoning process. We apply MultiTool-CoT to the Task 2 dataset of NumGLUE, which requires both numerical reasoning and domain-specific knowledge. The experiments show that our method significantly outperforms strong baselines and achieves state-of-the-art performance.¹

1 Introduction

Reasoning refers to the logical process of inferring unknown facts from known facts. Solving reasoning problems requires language understanding, real-world knowledge, arithmetic calculation, and symbolic processing. Improving the reasoning capability of artificial intelligence has been a long-standing challenge and remains an active research topic to this day (Gordon et al., 2012; Sap et al., 2020).

Recently, large language models (LLMs) have achieved amazing performance on various reasoning tasks (Brown et al., 2020; Lewkowycz et al., 2022; Zhang et al., 2022; Chowdhery et al., 2022). However, the amount of real-world knowledge learned by LLMs is still constrained by the size of model parameters and the training data. This problem could be more severe in the case of sparse domain-specific knowledge. Furthermore, LLMs are based on the computation among continuous token representations, which cannot ensure accurate arithmetic calculations.

To solve these problems, previous studies propose to complement the capabilities of LLMs with

¹Our code is publicly available at <https://github.com/InabaTatsuro/MultiTool-CoT>.

an external tool, such as a web browser or a calculator (Nakano et al., 2021; Cobbe et al., 2021; Yao et al., 2022). This is performed by invoking an external tool during reasoning with LLMs and injecting the results into the reasoning process. However, previous studies have focused on using a single external tool to solve a single problem with LLMs and have not addressed different problems together.

This paper proposes MultiTool-CoT, an interactive framework that allows LLMs to use multiple external tools during reasoning. Figure 1 provides an overview. In MultiTool-CoT, LLMs solve reasoning problems by generating reasoning processes including tool triggers to invoke external tools. We let LLMs learn to invoke multiple external tools at proper reasoning steps by chain-of-thought (CoT) prompting based on few-shot learning (Wei et al., 2022).

As a proof of concept, we apply MultiTool-CoT to the Task 2 dataset of NumGLUE (Mishra et al., 2022), which requires both numerical reasoning and domain-specific knowledge. Experiments show that MultiTool-CoT significantly outperforms strong baselines and achieves state-of-the-art performance.

2 Related Work

Large language models (LLMs) can perform various tasks by *prompting* (Liu et al., 2022). As for reasoning tasks, chain-of-thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022) is known for its effectiveness, which elicits the answer with intermediate reasoning steps from LLMs.

There is a growing body of work on using an external tool to improve reasoning with LLMs. Cobbe et al. (2021) use a calculator to process mathematical formulas that appear in reasoning processes by fine-tuning LLMs to generate mathematical formulas with a tool trigger to call the calculator. Nakano et al. (2021) allow LLMs to use a

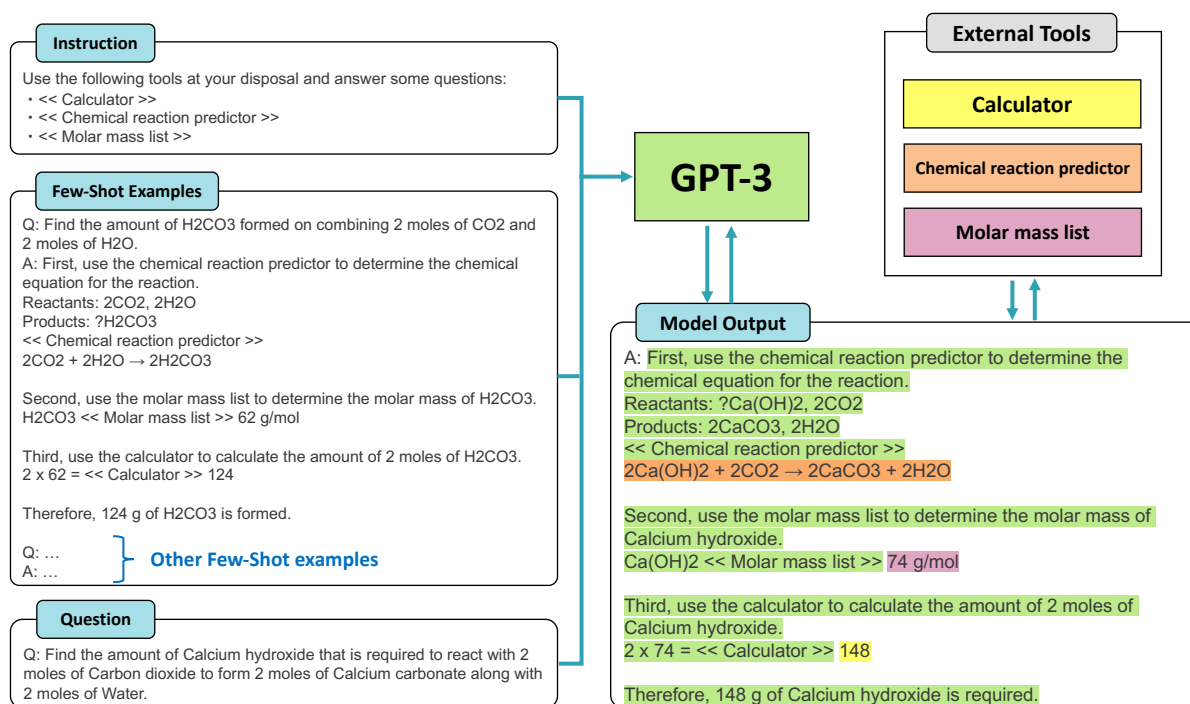


Figure 1: Overview of the MultiTool-CoT. The output of GPT-3, the calculator, the chemical reaction predictor, and the molar mass list are highlighted in green, yellow, orange, and purple, respectively.

web browser by fine-tuning LLMs to generate action codes to operate the browser. Previous studies focus on a single problem of LLMs, namely, error-prone arithmetic calculation or incomplete real-world knowledge, and address it by fine-tuning LLMs so that they can call a single external tool. In contrast, this study addresses multiple problems together by allowing LLMs to use multiple external tools. Besides, this study presents a few-shot learning-based framework (Brown et al., 2020) for doing this, which does not require fine-tuning.

A very recent study (Yao et al., 2022) proposes a few-shot learning-based method for invoking a Wikipedia API to perform knowledge-intensive reasoning tasks. However, this study has not investigated the effectiveness of using multiple external tools. A Python library named LangChain² implements a framework for allowing LLMs to use multiple external tools based on Yao et al. (2022), which is similar to ours. However, its effectiveness has not been investigated in any benchmark datasets as of this submission.

3 Method

We propose MultiTool-CoT, an interactive framework that allows LLMs to use multiple external

tools during reasoning. Figure 1 illustrates an overview.

MultiTool-CoT leverages chain-of-thought (CoT) prompting based on few-shot learning (Wei et al., 2022). Our prompt consists of an instruction specifying the available external tools, few-shot examples demonstrating several question-answer pairs with reasoning processes, and a question to be solved. We manually annotate the reasoning processes shown as few-shot examples with tool triggers marked with corresponding input data, adhering to a specific format. In this study, we let the string <<External tool name>> be a tool trigger. For example, if we use a calculator as an external tool, we annotate the reasoning processes with the tool trigger <<Calculator>> after input formulas like 2×62 .

When reasoning, GPT-3 follows the prompt and generates a reasoning process including tool triggers. If a tool trigger is generated, we stop text generation. We then extract the name of the external tool and the input for the tool from the reasoning process, execute the tool with the input, and append the result to the end of the reasoning process. After that, we restart text generation.

If we cannot execute an external tool for some reason (e.g., invalid tool input is generated), we fall back on GPT-3 and let it generate the output

²<https://langchain.readthedocs.io/en/latest>

of the tool.

We observe that the final answer value is nearly always contained in the last sentence of the reasoning process. Therefore, we apply an additional GPT-3 few-shot learning process for mapping the last sentence to the answer value by prompting several sentence-answer pairs.

4 Experiment

As a proof of concept, we applied MultiTool-CoT to solve a knowledge-based numerical reasoning task.

4.1 Dataset

We used the Task 2 dataset of NumGLUE (Mishra et al., 2022), which requires both numerical reasoning and domain-specific knowledge, mainly related to chemistry. Example (1) shows a question in the dataset.

- (1) Find the amount of Calcium hydroxide required to react with 2 moles of Carbon dioxide to form 2 moles of Calcium carbonate along with 2 moles of Water.

All the answers are given as numbers. We used 325 questions in the test split for evaluation. We evaluated the accuracy.

4.2 External Tools

We implemented the following external tools and used them in the proposed framework.

- **Calculator (CAL):** The calculator is given a mathematical formula and outputs the calculation result. The calculator is implemented using Python’s `eval` function³. Operators in mathematical formulas are replaced according to Python’s syntax. We prompt GPT-3 to output the tool trigger, `<<Calculator>>`, with a mathematical formula on the same line.
- **Chemical reaction predictor (CRP):** The chemical reaction predictor is given the chemical formula of reactants and products and outputs the chemical reaction equation by adjusting the coefficients so that the reactants and products have the same number of each atom. We prompt GPT-3 to output the tool trigger, `<<Chemical reaction`

³<https://docs.python.org/3/library/functions.html#eval>

Method	
Zero-Shot [†]	1
Zero-Shot+CoT	32.62
Few-Shot [†]	42
Few-Shot+CoT	57.85
MultiTool-CoT (CAL only)	62.77
MultiTool-CoT (CRP only)	64.31
MultiTool-CoT (MML only)	69.23
MultiTool-CoT (Ours)	85.85

Table 1: Performance in the Task 2 dataset of NumGLUE. The best result is shown in **bold**. (†) is cited from Mishra et al. (2022).

`predictor>>`, with the reactants and products on the previous two lines.

- **Molar mass list (MML):** The molar mass list is given a chemical formula and outputs its molar mass. The molar mass of the chemical formula is calculated from the atoms and their number in the formula. The molar mass of the atoms is obtained from the knowledge base listing the weight of all atoms. We prompt GPT-3 to output the tool trigger, `<<Molar mass list>>`, with a chemical formula on the same line.

4.3 Methods for Comparison

We used GPT-3 (text-davinci-003; 175B parameters) via OpenAI API⁴ and compared the following methods.

Zero-Shot We fed only the question into GPT-3 and considered the generated text as the answer.

Zero-Shot+CoT (Kojima et al., 2022) We fed the question with the sentence “Let’s think step by step.” into GPT-3 and obtained the answer with the intermediate reasoning steps. We then added the sentence fragment “Therefore, the answer (Arabic numerals) is ” after the generated text and fed it into GPT-3 to get the final answer.

Few-Shot We fed the question with few-shot examples of question-answer pairs into GPT-3 and obtained the generated text as the answer.

Few-Shot+CoT We performed the proposed method without invoking any external tools. If the tool triggers were generated, we used GPT-3 to output the result.

⁴<https://openai.com/api/>

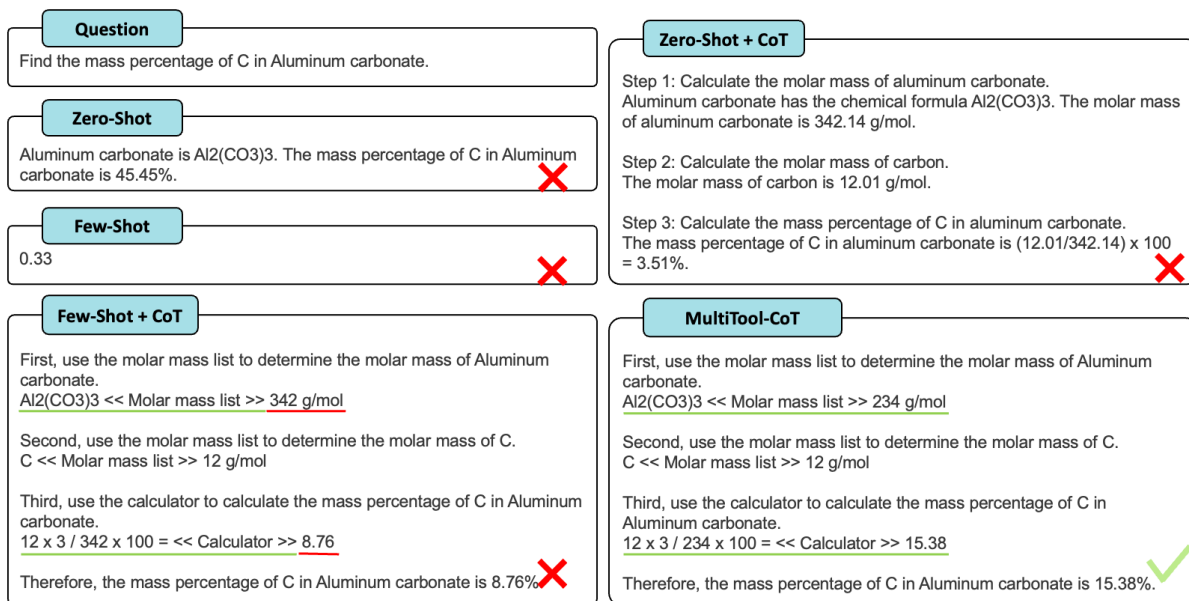


Figure 2: An improved example. The green lines indicate correct reasoning processes. The red lines indicate errors related to knowledge or arithmetic calculation.

MultiTool-CoT ($\{\text{CAL}|\text{CRP}|\text{MML}\}$ only) We performed the proposed method with one of the external tools introduced in Section 4.2. As for the other external tools, we let GPT-3 generate the result.

MultiTool-CoT (Ours) We performed the proposed method with all the external tools introduced in Section 4.2.

In few-shot settings, we used 20 questions in the training split as few-shot examples. The questions were manually selected to avoid bias in the number of external tool calls. In order to annotate the questions with reasoning processes with tool triggers, we followed a two-step process. First, we employed GPT-3 to generate the reasoning processes for solving these questions using zero-shot chain-of-thought prompting (Kojima et al., 2022), aiming to obtain reasoning processes that GPT-3 can easily follow. Then, we manually annotated the reasoning processes with tool triggers and the input and output for the corresponding external tools.

We set the temperature parameter of GPT-3 as 0 to generate constant predictions. Therefore, we report the results of single runs of the methods.

4.4 Results

Table 1 shows the results. The proposed method achieved an accuracy of 85.85, a state-of-the-art performance. We observed a significant perfor-

mance improvement compared to methods that did not use external tools and methods that used only one external tool. Note that the performance improvement from using multiple external tools is larger than the sum of the performance improvements from using each tool individually. This is because GPT-3 can fail to provide accurate answers due to a combination of different types of errors, such as incorrect arithmetic calculation and knowledge. The use of multiple external tools addressed such cases effectively, thereby improving the overall accuracy.

4.5 Case Study

Figure 2 shows an improved example. Zero-Shot and Few-Shot generated wrong answers. Zero-Shot+CoT and Few-Shot+CoT performed reasoning based on the incorrect molar mass of $\text{Al}_2(\text{CO}_3)_3$, resulting in incorrect answers. Besides, Few-Shot+CoT failed to calculate $12 \times 3 / 342 \times 100$. Our method, MultiTool-CoT, was able to answer correctly based on correct knowledge and calculation, relying on external tools. More examples are presented in Figure 3 and Figure 4 in Appendix.

Despite the excellent results, there were 46 instances in which the proposed method failed to deliver accurate answers. Upon manual investigation of all the errors, we identified that the majority of them were caused by incorrect reason-

ing processes (39%) and invalid tool inputs (35%). The remaining errors were categorized into incorrect gold answers (15%) and variations in answer formats (11%). Examples can be found in Appendix B. These errors are beyond the scope of what external tools can assist with.

5 Conclusion

We proposed MultiTool-CoT, a framework that allows LLMs to use multiple external tools, such as a knowledge retriever and a calculator, during reasoning. We applied MultiTool-CoT to a numerical reasoning task that requires knowledge of chemistry and confirmed its effectiveness. The proposed framework is general and can be applied to various tasks by changing and extending external tools. We plan to verify the effectiveness of the proposed method in other tasks in the future.

Limitations

The major limitation of the present study is that the effectiveness of the proposed method has been confirmed only for a single task. This is because most existing reasoning tasks are relatively simple that they can be solved by a single external tool at most. For example, most existing numerical reasoning tasks provide self-contained questions; that is, all the required knowledge is included in the questions. In such tasks, a calculator is all that is needed as an external tool. However, it would be rare for a single external tool to be sufficient in real-world applications such as medical text analysis. It is crucial for future work to validate the effectiveness in such realistic scenarios that necessitate the use of multiple external tools.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts,

Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [PaLM: Scaling Language Modeling with Pathways](#). *arXiv preprint arXiv:2204.02311*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. [SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of common-sense causal reasoning](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving quantitative reasoning problems with language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857. Curran Associates, Inc.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *ACM Computing Surveys*, 55(9).

- Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. [NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. [Webgpt: Browser-assisted question-answering with human feedback](#). *arXiv preprint arXiv:2112.09332*.
- Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. [Commonsense reasoning for natural language processing](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. [React: Synergizing reasoning and acting in language models](#). *arXiv preprint arXiv:2210.03629*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [OPT: Open Pre-trained Transformer Language Models](#). *arXiv preprint arXiv:2205.01068*.

	Few-Shot Examples	Acc.
CoT	5	55.38
CoT	10	56.31
CoT	20	57.85
MultiTool-CoT	5	83.69
MultiTool-CoT	10	84.00
MultiTool-CoT	20	85.85

Table 2: Performance for the different number of few-shot examples in the Task 2 dataset of NumGLUE. The best result is shown in **bold**.

A Effect of the Number of Few-shot Examples on Performance

We investigated the effect of the number of few-shot examples on performance. Table 2 shows the results. Reducing the number of few-shot examples decreased accuracy, regardless of whether external tools were used. Surprisingly, however, the drop in performance was not drastic, suggesting the strong generalization ability of GPT-3. Note that it is hopeless to further improve the performance by simply increasing the number of few-shot examples because the total number of tokens in the 20 few-shot examples is nearly 3,000 while the number of tokens that GPT-3 can process is 4,000.

B Analysis of Error Types

We manually investigated all 46 errors as described in Section 4.5. There were four types of errors: incorrect reasoning processes (39%), invalid tool inputs (35%), incorrect gold answers (15%), and variations in answer formats (11%).

Incorrect Reasoning Processes Figure 5 shows an error due to an incorrect reasoning process. GPT-3 generated an incorrect mathematical formula (underlined in red), which was expected to be $3 \times 16/160 \times 100$. Consequently, even though the calculation was performed correctly using the calculator, the final answer turned out to be incorrect.

Invalid Tool Inputs Figure 6 shows an error caused by an invalid tool input. GPT-3 generated an invalid product, CH₂Cl₂ (underlined in red), which was expected to be CCl₄. Thus, the chemical reaction predictor encountered a run-time error, resulting in an incorrect final answer.

Incorrect Gold Answers Figure 7 shows an error resulting from an incorrect gold answer. The answer predicted by the proposed method was “85 g/mol,” whereas the gold answer was “90 g/mol.”

Variations in Answer Formats Figure 8 shows an error attributed to a variation in the answer format. The answer predicted by the proposed method was “1 mole,” while the gold answer was “18 g”. Since 1 mole of water is 18g, they both represent the same quantity. However, due to the difference in the answer formats, it is considered an error.

<p>Question</p> <p>How many moles of C₂H₆ are required to react with 14 moles of Cl₂ to form 4 moles of CCl₄ along with 12 moles of HCl</p>	<p>Zero-Shot + CoT</p> <p>First, we need to calculate the number of moles of C₂H₆ required to produce 4 moles of CCl₄. This can be done using the following equation: $C_2H_6 + 3Cl_2 \rightarrow 2CCl_4 + 6HCl$ Therefore, we need 2 moles of C₂H₆ to produce 4 moles of CCl₄.</p>
<p>Zero-Shot</p> <p>8 moles of C₂H₆ are required to react with 14 moles of Cl₂ to form 4 moles of CCl₄ along with 12 moles of HCl. ✘</p>	<p>Next, we need to calculate the number of moles of C₂H₆ required to consume 14 moles of Cl₂. This can be done using the following equation: $C_2H_6 + 3Cl_2 \rightarrow 2CCl_4 + 6HCl$ Therefore, we need 6 moles of C₂H₆ to consume 14 moles of Cl₂.</p>
<p>Few-Shot</p> <p>8 ✘</p>	<p>Finally, we need to add the two values together to get the total number of moles of C₂H₆ required.</p>
<p>Few-Shot + CoT</p> <p>First, use the chemical reaction predictor to determine the chemical equation for the reaction. Reactants: ?C₂H₆, 14Cl₂ Products: 4CCl₄, 12HCl << Chemical reaction predictor >> <u>$7C_2H_6 + 14Cl_2 \rightarrow 8CCl_4 + 12HCl$</u> Therefore, 7 moles of C₂H₆ are required. ✘</p>	<p>MultiTool-CoT</p> <p>First, use the chemical reaction predictor to determine the chemical equation for the reaction. Reactants: ?C₂H₆, 14Cl₂ Products: 4CCl₄, 12HCl << Chemical reaction predictor >> <u>$2C_2H_6 + 14Cl_2 \rightarrow 4CCl_4 + 12HCl$</u> Therefore, 2 moles of C₂H₆ are required. ✔</p>

Figure 3: An improved example. The red line indicates an error in chemical reaction understanding. The green line indicates the correct reasoning process by using the chemical reaction predictor as an external tool.

<p>Question</p> <p>What is the molecular weight of a compound having 2 Aluminium, 3 Carbon, and 9 Oxygen</p>	<p>Zero-Shot + CoT</p> <p>The molecular weight of a compound is the sum of the atomic weights of all the atoms in the compound.</p>
<p>Zero-Shot</p> <p>The molecular weight of a compound with 2 Aluminium, 3 Carbon, and 9 Oxygen atoms is 270.18 g/mol. This is calculated by adding the atomic weights of each element: <u>$(2 \times 26.98) + (3 \times 12.01) + (9 \times 16.00) = 270.18 \text{ g/mol}$</u>. ✘</p>	<p>The atomic weight of Aluminium is 26.981538, Carbon is 12.0107, and Oxygen is 15.9994.</p>
<p>Few-Shot</p> <p>342 ✘</p>	<p>Therefore, the molecular weight of the compound is <u>$(2 \times 26.981538) + (3 \times 12.0107) + (9 \times 15.9994) = 262.818$</u>. ✘</p>
<p>Few-Shot + CoT</p> <p>First, use the molar mass list to determine the molar mass of Aluminium, Carbon, and Oxygen. Al << Molar mass list >> 27 g/mol C << Molar mass list >> 12 g/mol O << Molar mass list >> 16 g/mol</p> <p>Second, use the calculator to calculate the molecular weight of the compound. <u>$2 \times 27 + 3 \times 12 + 9 \times 16 = \text{<< Calculator >> } 198$</u></p> <p>Therefore, the molecular weight of the compound is 198 g/mol. ✘</p>	<p>MultiTool-CoT</p> <p>First, use the molar mass list to determine the molar mass of Aluminium, Carbon, and Oxygen. Al << Molar mass list >> 27 g/mol C << Molar mass list >> 12 g/mol O << Molar mass list >> 16 g/mol</p> <p>Second, use the calculator to calculate the molecular weight of the compound. <u>$2 \times 27 + 3 \times 12 + 9 \times 16 = \text{<< Calculator >> } 234$</u></p> <p>Therefore, the molecular weight of the compound is 234 g/mol. ✔</p>

Figure 4: An improved example. The red line indicates errors in arithmetic calculation. The green line indicates the correct reasoning process by using the calculator as an external tool.

Question

Find the mass percentage of O in Fe₂O₃

MultiTool-CoT

First, use the molar mass list to determine the molar mass of Fe₂O₃.
 Fe₂O₃ << Molar mass list >> 160 g/mol

Second, use the molar mass list to determine the molar mass of O.
 O << Molar mass list >> 16 g/mol

Third, use the calculator to calculate the mass percentage of O in Fe₂O₃.
 $2 \times 16 / 160 \times 100 =$ << Calculator >> 20

Therefore, the mass percentage of O in Fe₂O₃ is 20%. ❌

Figure 5: An example of incorrect reasoning processes.

Question

Find the number of moles of Hydrochloric acid formed on combining 1 mole of Methane and 4 moles of Chlorine

MultiTool-CoT

First, use the chemical reaction predictor to determine the chemical equation for the reaction.
 Reactants: CH₄, 4Cl₂
 Products: ?HCl, CH₂Cl₂
 << Chemical reaction predictor >>
 CH₄ + 2Cl₂ → HCl + CH₂Cl₂

Therefore, 2 moles of Hydrochloric acid are formed. ❌

Figure 6: An example of the invalid tool inputs.

Question

What is the molecular weight of 5 moles of Ammonia

MultiTool-CoT

First, use the molar mass list to determine the molar mass of Ammonia.
 NH₃ << Molar mass list >> 17 g/mol

Second, use the calculator to calculate the molecular weight of 5 moles of Ammonia.
 5 x 17 = << Calculator >> 85

Therefore, the molecular weight of 5 moles of Ammonia is 85 g/mol. ❌

Figure 7: An example of incorrect gold answers.

Question

Find the amount of H₂O formed on combining 2 moles of NaOH and 1 mole of Cl₂

MultiTool-CoT

First, use the chemical reaction predictor to determine the chemical equation for the reaction.
 Reactants: 2NaOH, Cl₂
 Products: ?H₂O, NaCl, NaClO
 << Chemical reaction predictor >>
 2NaOH + Cl₂ → H₂O + NaCl + NaClO

Therefore, 1 mole of H₂O is formed. ❌

Figure 8: An example of variations in answer formats

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
the 'Limitations' section
- A2. Did you discuss any potential risks of your work?
This study focuses on improving the reasoning performance of language models. We cannot think of particular concerns.
- A3. Do the abstract and introduction summarize the paper's main claims?
0,1
- A4. Have you used AI writing assistants when working on this paper?
We use Grammarly for grammar checking for a part of the sections of the paper.

B Did you use or create scientific artifacts?

4,5

- B1. Did you cite the creators of artifacts you used?
4,5
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5

C Did you run computational experiments?

5

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

5

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

5

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

5

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.