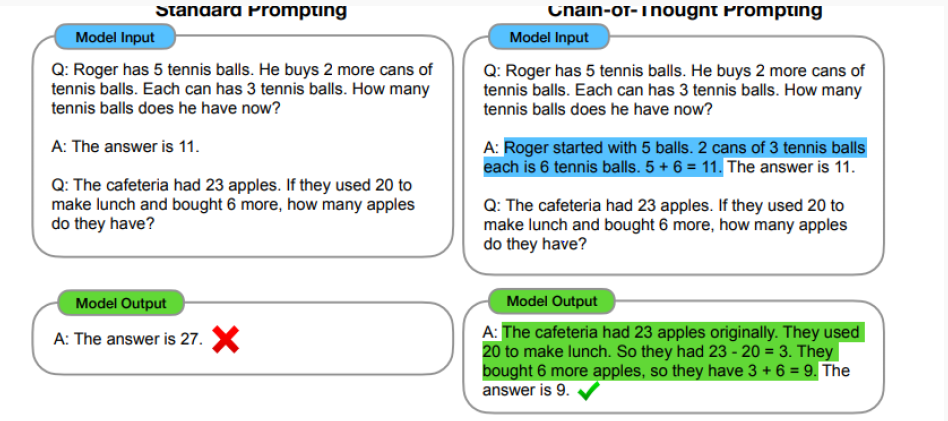
**A SUMMARY OF LEAST-TO-MOST PROMPTING ENABLES COMPLEX REASONING IN LARGE LANGUAGE MODELS**

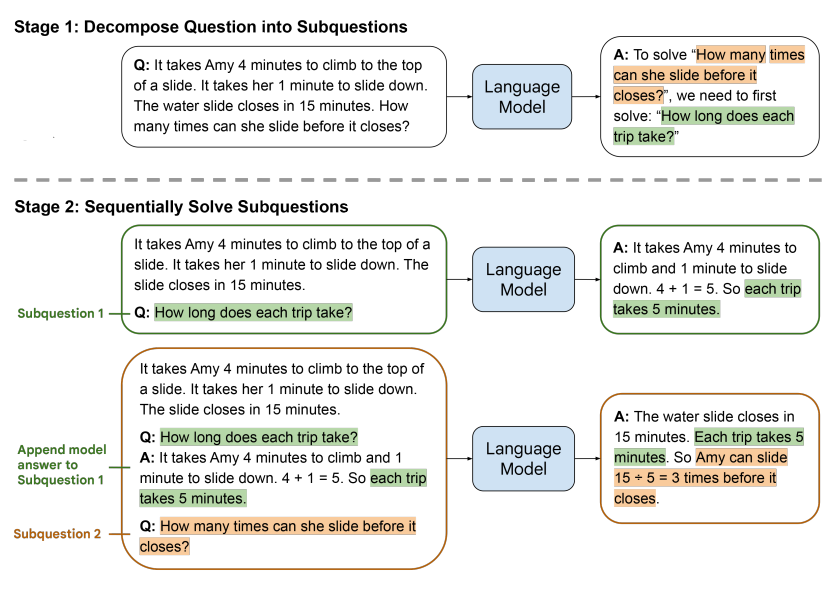
**1. INTRODUCTION**

This research paper discusses a novel approach called "least-to-most prompting," which aims to enhance the problem-solving capabilities of large language models, such as GPT-3, particularly in complex reasoning tasks. While the **"chain-of-thought" prompting approach has shown impressive results in various natural language reasoning tasks, it struggles when confronted with problems that are more challenging than the examples given in the prompts.**

****

**To address this limitation, the authors propose the "least-to-most prompting" strategy. The central idea behind this strategy is to break down a complex problem into a series of simpler subproblems and then solve them sequentially**. Importantly, solving each subproblem is made easier by the answers obtained from previously solved subproblems. The experimental results on tasks related to symbolic manipulation, compositional generalization, and mathematical reasoning indicate that least-to-most prompting allows models to generalize to more difficult problems than the prompts.

**One noteworthy finding is that when applied to the GPT-3 model, this approach significantly improves its ability to solve the compositional generalization benchmark SCAN, achieving an accuracy of at least 99% with just 14 exemplars,** compared to only 16% accuracy using the chain-of-thought prompting method. This is remarkable because existing neural-symbolic models specializing in solving SCAN are trained on a vast dataset with over 15,000 examples.



The above figure depicts the two-step process of "least-to-most prompting" in solving a math word problem. In the first stage, the problem is broken down into subproblems, and in the second stage, the subproblems are solved sequentially. Each subproblem's answer is built upon the solution to the previous subproblem. Demonstration examples for each stage are not included in this illustration.

**2. LEAST-TO-MOST PROMPTING**

"Least-to-most prompting" is a method used to teach language models, like GPT-3, how to solve complex problems by breaking them down into simpler subproblems. It involves two stages:

**1. Decomposition:** In this stage, the model is provided with a prompt that includes examples illustrating the decomposition of a problem, followed by the specific question that needs to be decomposed.

**2. Subproblem solving:** In this stage, the prompt consists of three components: (1) examples demonstrating how to solve subproblems, (2) a list of previously answered subquestions and generated solutions, and (3) the next question to be answered.

**For example, if the original problem is to be decomposed, the model first learns how to break it down into subproblems. The prompt guides the model in understanding that the original problem can be solved by addressing an intermediate problem, such as "How long does each trip take?"**

After the decomposition phase, the model is tasked with solving the subproblems individually. The solving process starts with a prompt that provides examples of how to solve problems, followed by the first subproblem (e.g., "How long does each trip take?"). The model generates an answer ("... each trip takes 5 minutes"), and this answer is appended to the previous prompt, followed by the next subproblem, which could be the original problem itself. The model continues this process until it arrives at the final answer.

Least-to-most prompting can be combined with other prompting techniques, such as "chain-of-thought" and "self-consistency," but it is not required. Depending on the task, the two stages of least-to-most prompting can also be merged into a single-pass prompt for more efficient problem-solving.

**3. RESULTS**

In this section, we present the results of applying least-to-most prompting to symbolic manipulation, compositional generalization, and math reasoning tasks. We compare these results with chain-of-thought prompting.

**Symbolic Manipulation (Last-Letter-Concatenation Task):**

For the last-letter-concatenation task, we find that chain-of-thought prompting performs well when the testing lists match the prompt exemplars in length but poorly when the testing lists are longer. In contrast, **least-to-most prompting significantly outperforms chain-of-thought prompting on length generalization.**

- Least-to-Most Prompt Context for Decomposition (Table 1): This prompt effectively decomposes long lists into sequential sublists with 100% accuracy.

- Least-to-Most Prompt Context for Solution (Table 2): These exemplars demonstrate both a base case and a recursive step in solving the task.

**Compositional Generalization (SCAN):** [**[2006.10627] Compositional Generalization by Learning Analytical Expressions (arxiv.org)**](https://arxiv.org/abs/2006.10627)

In the case of SCAN, a benchmark for compositional generalization, we observe that least-to-most prompting with large language models can solve SCAN using only a few demonstration examples without requiring training or fine-tuning.

**Math Reasoning (GSM8K and DROP):**

We apply least-to-most prompting to solve math word problems in GSM8K and DROP. GSM8K presents problems of varying complexity, measured by the number of solving steps required. Least-to-most prompting slightly improves over chain-of-thought prompting, with significant improvement observed for problems requiring at least 5 steps to solve.

In summary, least-to-most prompting shows promising results in various tasks, including improved generalization on symbolic manipulation tasks, effective solving of compositional generalization benchmarks, and a slight improvement over chain-of-thought prompting in math reasoning tasks. These results are achieved without the need for extensive training or fine-tuning.

**4. RELATED WORK**

**1. Compositional Generalization:** The paper discusses the SCAN benchmark and prior work that focused on neural-symbolic architectures and grammar induction techniques to tackle it. These approaches required complex training and grammar inference. The paper notes that least-to-most prompting achieves high accuracy on SCAN without specialized model architectures or extensive training.

**2. Easy-to-Hard Generalization:** The paper mentions tasks where test cases are more challenging than training examples. It cites studies like Neural Logic Machines (NLMs) and recurrent networks. In contrast, least-to-most prompting addresses this by breaking complex problems into simpler subproblems.

**3. Task Decomposition:** The section explores methods for breaking down tasks into subquestions. It mentions approaches like modeling prompts as virtual tokens. However, least-to-most prompting sets itself apart by not needing training or fine-tuning and by requiring sequential resolution of dependent subquestions.

**4. Question Translation and Chaining:** The paper briefly mentions methods for translating questions into SQL queries and chaining steps in large language models. Least-to-most prompting differs from these methods by combining problem decomposition and subproblem solving.

**5. LIMITATIONS**

**1. Domain-Specific Prompt Design:** Decomposition prompts may not be universally effective across different problem domains. For instance, a prompt designed for math word problems may not work well for common sense reasoning problems. To achieve optimal results, new prompts tailored to specific problem types are often necessary.

**2. Challenges in Same-Domain Generalization:** Even within the same problem domain, generalizing decomposition can be problematic. While some benchmarks may benefit from correct decomposition prompts, this is not always straightforward. Challenges can arise in breaking down complex problems into simpler subproblems, affecting generalization within the domain.

**3. Variability in Decomposition Difficulty:** The effectiveness of the least-to-most prompting approach can vary based on the complexity of the decomposition process. Some tasks, such as the "last-letter-concatenation" task, are relatively easy to decompose, leading to exceptional results. However, the level of difficulty in decomposition may not be consistent across all problem types.

**6. CONCLUSION**

In summary, the research introduced "least-to-most prompting," a method that enables language models to solve more challenging problems than those in the prompts. Empirical results across various tasks showed that least-to-most prompting outperformed traditional prompts and chain-of-thought prompting.

The paper also suggests that traditional prompting may not be the ideal way to teach reasoning skills to language models, and the future may involve transitioning to fully bidirectional conversations, allowing for immediate feedback. The least-to-most prompting approach represents a step in this direction, enabling more effective interactions with language models.