# Title：

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

# Abstract：

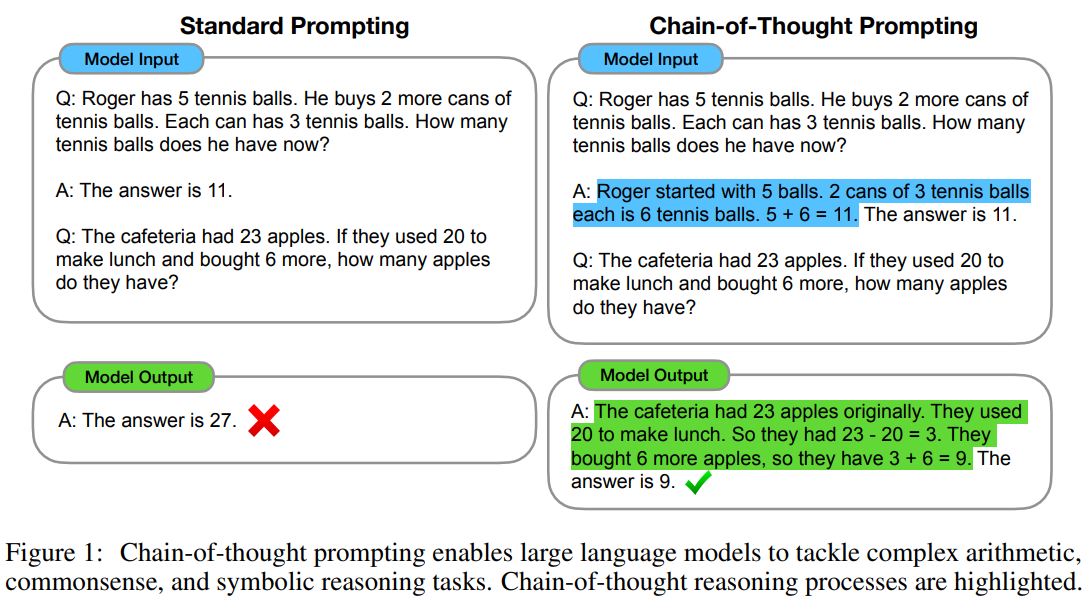
语言模型的规模越来越大，一般从三个角度衡量语言模型的规模，1是计算量（FLOPS）2是训练数据的大小（number of tokens）3是模型本身参数量的大小。但是即使是现在比较大的语言模型，现在也很难在推理上有好的表现（数学、符号、常识推理）。CoT是一系列短句子，表明了人类在推理的时候的一系列步骤。

实验结果非常好，在用到PALM模型，CoT能在GSM8K这种比较难的数学数据集上取得新的SOTA

# Introduction

语言模型在达到100B参数量的时候就能在分类任务上取得非常好的结果，作者把这种任务归结为system-1任务：能很快被人类所理解的任务（这个电影好看，这个东西好吃）。

然而还有一些system-2任务要求很慢和仔细的考虑（逻辑、数学、常识推理），这种任务语言模型很难“大力出奇迹”，随着参数量增长，表现并没有提升。针对这种问题，本文使用CoT这种方法解决推理任务。



左图是在few-shot情况下，问题答案一起丢给大语言模型，这样很容易犯错

CoT的方法是在给出解之前把推理部分加上扔给模型

# Chain-of-Thought Prompting

chain of thought：在应对推理任务时，在给出最终答案之前所产生的中间推理步骤，而他们的载体就是一系列短句子。作者把这些定义成CoT是想去模拟人类一步一步得出答案的过程。

CoT有几个有意思的步骤：  
1.CoT原则上能够让模型把一个多步的问题分解成中间步骤，是那些需要更多推理步骤的问题能够分配到更多的计算量。（LLM生成是token by token，CoT能让模型有机会去生成更多的中间推理步骤，那么自然就可以在解更难的问题的时候使用更多的计算量）

2.CoT提供了可解释性

3.CoT在原则上可以适用于任何人类能用语言所解的问题，不只是推理任务

4.可以通过few-shot learning这样的范式，引发推理步骤，最终给出答案

# Arithmetic Reasoning

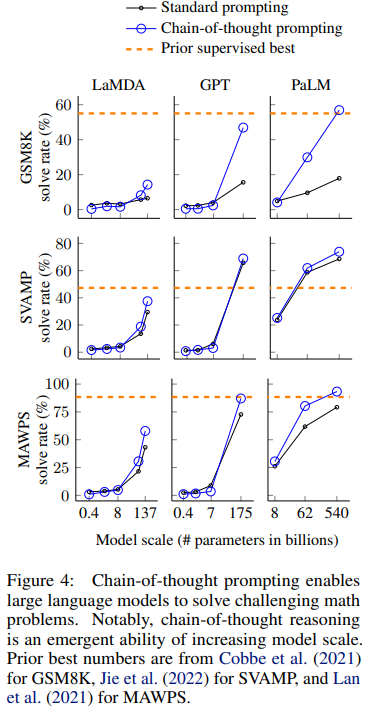
算术推理

本文的算术推理所考虑的问题范围集中在小学生的问题当中

## Chain-of-thought prompting.

人工设计了8个带有CoT设计链条的few-shot样例，在六个数据集当中同等的使用了这些样例

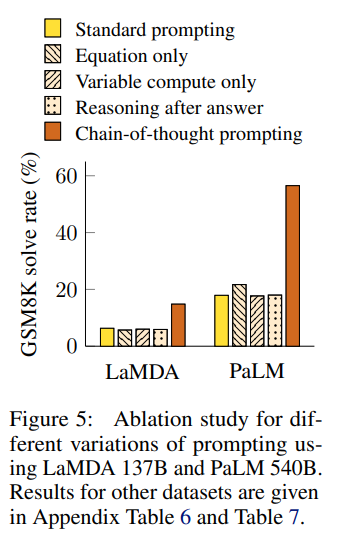
## Results



基于PaLM、GPT，LaMDA这些大语言模型来做的，除了GPT是OpenAI的，其他两个是google内部的。

模型越大，CoT的表现越好

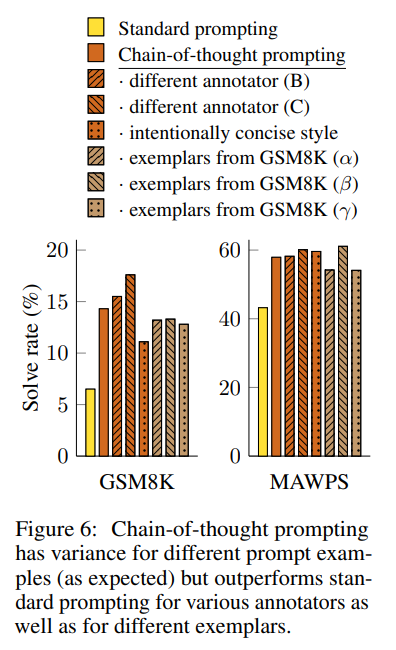
## Ablation Study



表明CoT简单，但不能再简单了

这些消融实验都表现不好

## Robustness of Chain of Thought



以为CoT对prompt敏感，作者认为有必要评测一下稳健性

实验表明带有CoT的方法比不带的要好很多

# Conclusion

我们探索了思维链提示作为一种简单且广泛适用的方法，用于增强语言模型中的推理能力。通过在算术、符号和常识推理方面的实验，我们发现思维链推理是模型规模的一个新兴特性，它允许足够大的语言模型执行推理任务，否则这些任务的扩展曲线将是平缓的。扩大语言模型可以执行的推理任务的范围将有望激发进一步研究基于语言的推理方法。