

# 链式思维提示在大型语言模型中激发推理能力

杰森·魏 薛志旺 戴尔·舒尔曼斯 马尔滕·博斯马  
布莱恩·伊奇特 费·夏 艾德·H·齐 郭克·V·莱 丹尼·周

谷歌研究, Brain团队  
{jasonwei,dennyzhou}@google.com

## 摘要

我们探索了如何通过生成一系列中间推理步骤来显著提高大型语言模型执行复杂推理的能力。特别是,我们展示了这种推理能力如何通过一种简单的方法——链式思维提示,在足够大的语言模型中自然地出现。在这种方法中,通过提供一些链式思维演示作为提示的示例。

对三个大型语言模型进行的实验表明,链式思维提示改善了在一系列算术、常识和符号推理任务上的性能。实证收益可能非常显著。例如,仅通过八个链式思维示例提示一个PaLM 540B,在数学问题的GSM8K基准测试中实现了最先进的准确性,甚至超过了微调的GPT-3与验证器。

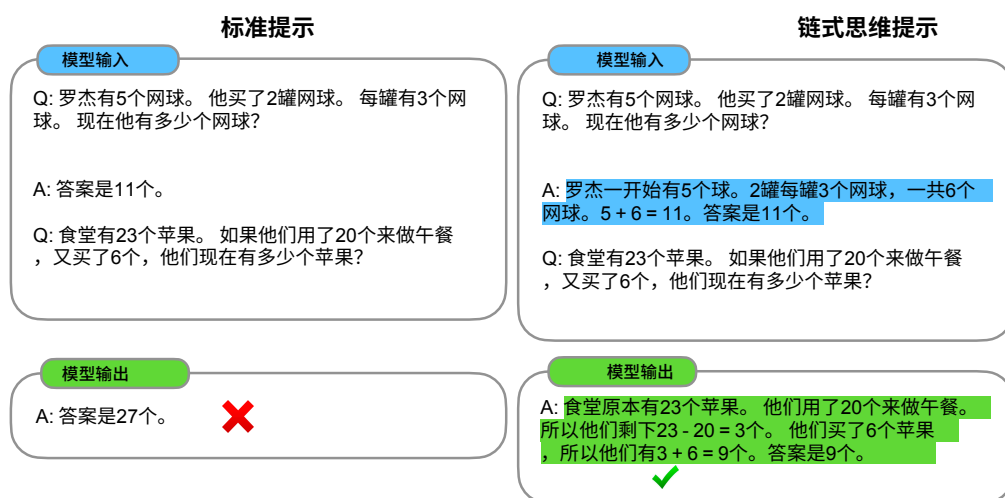


图1: 链式思维提示使大型语言模型能够处理复杂的算术、常识和符号推理任务。突出显示链式思维推理过程。

## 1 引言

最近，语言模型（Peters等，2018；Devlin等，2019；Brown等，2020，等等）已经彻底改变了自然语言处理领域。扩大语言模型的规模已经被证明具有一系列的好处，例如提高性能和样本效率（Kaplan等，2020；Brown等，2020，等等）。然而，仅仅扩大模型规模并不足以在算术、常识和符号推理等具有挑战性的任务上取得高性能（Rae等，2021）。

本研究探讨了一种简单的方法，通过两个思路来解锁大型语言模型的推理能力。首先，算术推理技术可以通过生成导致最终答案的自然语言解释来受益。之前的研究通过从头开始训练（Ling等人，2017）或微调预训练模型（Cobbe等人，2021），以及使用形式语言而不是自然语言的神经符号方法（Roy和Roth，2015；Chiang和Chen，2019；Amini等人，2019；Chen等人，2019）赋予模型生成自然语言中间步骤的能力。其次，大型语言模型提供了令人兴奋的机会

通过提示，实现上下文少样本学习的前景。也就是说，不需要为每个新任务微调单独的语言模型检查点，只需用几个输入-输出示例提示模型即可。令人惊讶的是，这在一系列简单的问答任务中取得了成功（Brown等，2020年）。

然而，上述两种想法都有关键的局限性。对于基于理性增强的训练和微调方法来说，创建一组大量高质量的理性是昂贵的，这比普通机器学习中使用的简单输入-输出对要复杂得多。对于布朗等人（2020）使用的传统少样本提示方法，在需要推理能力的任务上效果不佳，并且通常随着语言模型规模的增加改进不大（Rae等人，2021）。在本文中，我们结合了这两种想法的优势，以一种避免它们局限性的方式。具体而言，我们探索了语言模型在推理任务中执行少样本提示的能力，给定一个由三元组组成的提示：〈输入，思维链，输出〉。思维链是一系列中间的自然语言推理步骤，导致最终的输出，我们将这种方法称为链式思维提示。图1展示了一个示例提示。

我们在算术、常识和符号推理基准测试中进行了实证评估，结果显示链式思维提示优于标准提示，有时效果显著。图2展示了一个这样的结果——在数学问题的GSM8K基准测试中（Cobbe等，2021），使用PaLM 540B的链式思维提示明显优于标准提示，并取得了新的最先进性能。仅使用提示的方法很重要，因为它不需要大量的训练数据集，并且单个模型检查点可以执行许多任务而不会丧失广泛性。这项工作强调了大型语言模型如何通过少量自然语言数据的示例来学习任务（例如通过大型训练数据集自动学习输入和输出之间的模式）。

## 2 链式思维提示

考虑解决复杂推理任务（例如多步数学问题）时的自己的思考过程。通常将问题分解为中间步骤，并在给出最终答案之前解决每个步骤：“简给妈妈2朵花后，她有10...然后她给爸爸3朵花后，她将有7...所以答案是7。”本文的目标是赋予语言模型生成类似的能力——一系列连贯的中间推理步骤，导致问题的最终答案。我们将展示足够大的

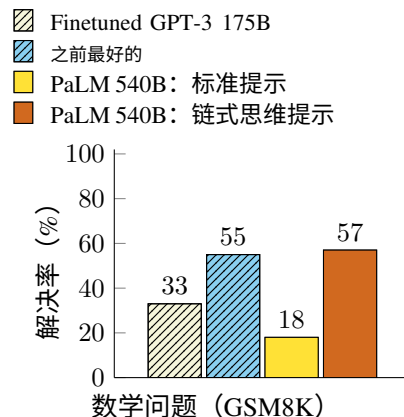


图2: PaLM 540B使用链式思维提示在数学单词问题的GSM8K基准测试中实现了最新的最佳性能。

Finetuned GPT-3和之前最好的来自Cobbe等人（2021）。

如果在示例中提供链式思维推理的演示，语言模型可以生成思维链。

图1展示了一个模型产生思维链来解决一个数学问题的例子，否则它会得到错误的答案。在这种情况下，思维链类似于一个解决方案，并且可以被解释为一个解决方案，但我们仍然选择称之为思维链，以更好地捕捉到它模拟了逐步思考的过程来得出答案的思想（而且，解决方案/解释通常在最终答案之后出现（Narang等，2020年；Wiegreffe等，2022年；Lampinen等，2022年，等等））。

链式思维提示作为促进语言模型推理的方法具有几个吸引人的特点。

1. 首先，从原则上讲，思维链允许模型将多步问题分解为中间步骤，这意味着可以为需要更多推理步骤的问题分配额外的计算。
2. 其次，一连串的思维提供了对模型行为的可解释窗口，暗示了它可能是如何得出特定答案的，并提供了调试推理路径错误的机会（尽管完全描述支持答案的模型计算仍然是一个未解之谜）。
3. 第三，链式思维推理可以用于数学问题、常识推理和符号操作等任务，并且在原则上可能适用于人类可以通过语言解决的任何任务。
4. 最后，通过将一连串思维序列的示例包含在少样本提示的示例中，可以很容易地在足够大的现成语言模型中引发链式思维推理。

在实证实验中，我们将观察链式思维提示在算术推理（第3节）、常识推理（第4节）和符号推理（第5节）中的实用性。

### 3 算术推理

我们首先考虑图1中所示的数学问题，这些问题衡量了语言模型的算术推理能力。尽管对人类来说很简单，但算术推理是一项语言模型经常难以应对的任务（Hendrycks等，2021；Patel等，2021，等等）。令人惊讶的是，当与具有540B参数的语言模型一起使用时，链式思维提示在多个任务上的表现与任务特定的微调模型相当，甚至在具有挑战性的GSM8K基准测试中取得了新的最佳结果（Cobbe等，2021）。

#### 3.1 实验设置

我们在多个基准测试中探索了链式思维提示对于各种语言模型的应用。

基准测试。我们考虑了以下五个数学问题基准测试：（1）数学问题的GSM8K基准测试（Cobbe等，2021）；（2）具有不同结构的数学问题的SVAMP数据集（Patel等，2021）；（3）多样化数学问题的ASDiv数据集（Miao等，2020）；（4）代数问题的AQuA数据集；以及（5）MAWPS基准测试（Koncel-Kedziorski等，2016）。附录表12中给出了示例问题。

标准提示。对于基准，我们考虑标准的少样本提示，由Brown等人（2020）推广，即在输出测试示例的预测之前，语言模型会给出上下文示例的输入-输出对。示例被格式化为问题和答案。模型直接给出答案，如图1（左）所示。

链式思维提示。我们提出的方法是在少样本提示中为每个示例增加一个关联答案的思维链，如图1（右）所示。由于大多数数据集只有一个评估集，我们手动组成了一组包含思维链提示的八个少样本示例，图1（右）展示了一个思维链示例，完整的示例集在附录表20中给出。（这些特定示例没有经过提示工程处理；鲁棒性在第3.4节和附录A.2中进行了研究。）为了调查这种形式的链式思维提示是否能够成功激发一系列推理能力

<p><b>数学问题 (自由回答)</b></p> <p>问题: 罗杰有5个网球。他买了2罐网球。每罐有3个网球。他现在有多少个网球?</p> <p>答案: 罗杰最初有5个网球。每罐有3个网球, 2罐共6个网球。5 + 6 = 11。答案是11。</p>	<p><b>数学问题 (多项选择)</b></p> <p>问题: 输入数字1到500需要多少次按键?</p> <p>答案选择: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788</p> <p>A: 有9个一位数从1到9。从10到99有90个两位数。从100到500有401个三位数。9 + 90(2) + 401(3) = 1392。答案是 (b)。</p>	<p><b>CSQA (常识问题)</b></p> <p>问题: 萨米想去人多的地方。他可能去哪里?</p> <p>选项: (a) 赛车场 (b) 人口密集区 (c) 沙漠 (d) 公寓 (e) 路障</p> <p>答案: 答案必须是一个人的地方。赛车场、沙漠、公寓和路障都没有很多人, 但人口密集区有。所以答案是(b)。</p>
<p><b>StrategyQA</b></p> <p>问题: 是或否: 一个梨会在水中沉没吗?</p> <p>答案: 一个梨的密度约为0.6克/立方厘米, 小于水的密度。因此, 一个梨会浮起来。所以答案是否定的。</p>	<p><b>日期理解</b></p> <p>Q: 音乐会原定于06/01/1943举行, 但因为延迟了一天而改到今天。10天前的日期是多少, 格式为MM/DD/YYYY?</p> <p>A: 06/01/1943的后一天是06/02/1943, 所以今天是06/02/1943。今天的前10天是05/23/1943。所以答案是05/23/1943。</p>	<p><b>体育理解</b></p> <p>Q: 下面的句子是否合理? "Joao Moutinho在NFC冠军赛中接到了屏幕传球。"</p> <p>A: Joao Moutinho是一名足球运动员。NFC冠军赛是美式足球的一部分, 而不是足球。所以答案是否定的。</p>
<p><b>SayCan (指导机器人)</b></p> <p>人类: 你怎么给我带来一些不是水果的东西?</p> <p>解释: 用户想要吃一些不是水果的东西。能量棒不是水果, 所以我会给用户带来一个能量棒。计划: 1. 寻找(能量棒) 2. 挑选(能量棒) 3. 寻找(用户) 4. 放置(能量棒) 5. 完成()。</p>	<p><b>最后一个字母连接</b></p> <p>Q: 取"Lady Gaga"中每个单词的最后一个字母并连接起来。</p> <p>A: "Lady"的最后一个字母是"y"。"Gaga"的最后一个字母是"a"。将它们连接起来得到"ya"。所以答案是ya。</p>	<p><b>抛硬币 (状态跟踪)</b></p> <p>问题: 一枚硬币正面朝上。Maybelle抛掷硬币。Shalonda没有翻转硬币。硬币还是正面朝上吗?</p> <p>A: 硬币是由Maybelle抛掷的。所以硬币被抛掷了1次, 这是一个奇数。硬币起初是正面朝上的, 所以经过奇数次翻转后, 它将变为反面朝上。所以答案是不。</p>

图3: 算术、常识和符号推理基准测试的(输入、思维链、输出)三元组示例。思维链被突出显示。完整的提示请参见附录G。

对于数学问题, 我们使用了这八个思维链示例的单一集合, 用于除了AQuA之外的所有基准测试, AQuA是多项选择而不是自由回答。对于AQuA, 我们使用了训练集中的四个示例和解决方案, 详见附录表21。

语言模型。我们评估了五个大型语言模型。第一个是GPT-3 (Brown等, 2020), 我们使用text-ada-001、text-babbage-001、text-curie-001和text-davinci-002, 这些模型预计对应于350M、1.3B、6.7B和175B参数的InstructGPT模型 (Ouyang等, 2022)。第二个是LaMDA (Thoppilan等, 2022), 它有422M、2B、8B、68B和137B参数的模型。第三个是PaLM, 它有8B、62B和540B参数的模型。

第四个是UL2 20B (Tay等人, 2022年), 第五个是Codex (Chen等人, 2021年, code-davinci-002在OpenAI API中)。我们通过贪婪解码从模型中进行采样 (尽管后续的研究表明, 通过对许多采样生成的的大多数最终答案进行跟进工作可以改进链式思维提示 (Wang等人, 2022a))。对于LaMDA, 我们报告了五个随机种子的平均结果, 每个种子都有一个不同的随机排序的示例。由于LaMDA实验在不同种子之间没有显示出很大的差异, 为了节省计算资源, 我们报告了所有其他模型的单个示例顺序的结果。

### 3.2 结果

链式思维提示的最强结果总结在图4中, 附录中的表2显示了每个模型集合、模型大小和基准测试的所有实验输出。有三个关键点。首先, 图4显示链式思维提示是模型规模的一种新兴能力 (Wei等人, 2022b)。也就是说, 链式思维提示对于小型模型的性能没有积极影响, 只有在使用大约1000亿参数的模型时才能获得性能提升。我们在定性上发现, 规模较小的模型产生流畅但不合逻辑的思维链, 导致性能低于标准提示。

其次，链式思维提示对于更复杂的问题有更大的性能提升。例如，对于GSM8K（基准性能最低的数据集），最大的GPT和PaLM模型的性能提升超过了两倍。另一方面，对于MAWPS中最简单的子集SingleOp，只需要一步就能解决的问题，性能改进要么是负数，要么非常小（见附录表3）。

### 第三，通过GPT-3

175B和PaLM 540B的链式思维提示与通常在标记训练数据集上微调任务特定模型的先前最先进技术相比，表现出色。图4展示了PaLM 540B如何使用链式思维提示在GSM8K、SVAMP和MAWPS上取得了新的最先进水平（尽管需要注意的是，标准提示已经超过了SVAMP的先前最佳水平）。在另外两个数据集AQuA和ASDiv上，使用链式思维提示的PaLM接近最先进水平的水平（附录表2）。

为了更好地理解为什么链式思维提示有效，我们手动检查了LaMDA 137B在GSM8K上生成的思维链的50个随机示例。在模型返回正确的最终答案的50个随机示例中，除了两个巧合地得出了正确答案之外，所有生成的思维链都是逻辑上和数学上正确的（请参见附录D.1和表8，其中列出了正确的模型生成的思维链的示例）。我们还随机检查了50个模型给出错误答案的随机示例。

这个分析的总结是，46%的思维链几乎是正确的，除了一些小错误（计算器错误，符号映射错误，ping错误，或者缺少一个推理步骤），而另外54%的思维链在语义理解或连贯性方面存在重大错误（见附录D.2）。为了对为什么扩展能够提高链式思维推理能力提供一点见解，我们对PaLM 62B模型的错误进行了类似的分析，并观察扩展到PaLM 540B模型是否能够修复这些错误。总结是，将PaLM扩展到540B可以修复62B模型中大部分缺少一个步骤和语义理解错误（见附录A.1）。

### 3.3 割除研究

使用链式思维提示所观察到的好处引发了一个自然的问题，即是否可以通过其他类型的提示来获得相同的性能改进。图5展示了一个割除研究，包括三种链式思维的变体，如下所述。

仅方程。链式思维提示可能有助于推理的一个原因是它产生了要评估的数学方程，因此我们测试了一种变体，在给出答案之前，模型只被提示输出一个数学方程。图5显示，仅方程提示对于GSM8K的帮助不大，这意味着在没有链式思维中的自然语言推理步骤的情况下，GSM8K问题的语义太具挑战性，无法直接转化为方程。然而，对于一步或两步问题的数据集，我们发现仅方程提示确实提高了性能，因为方程可以很容易地从问题中推导出来（见附录表6）。

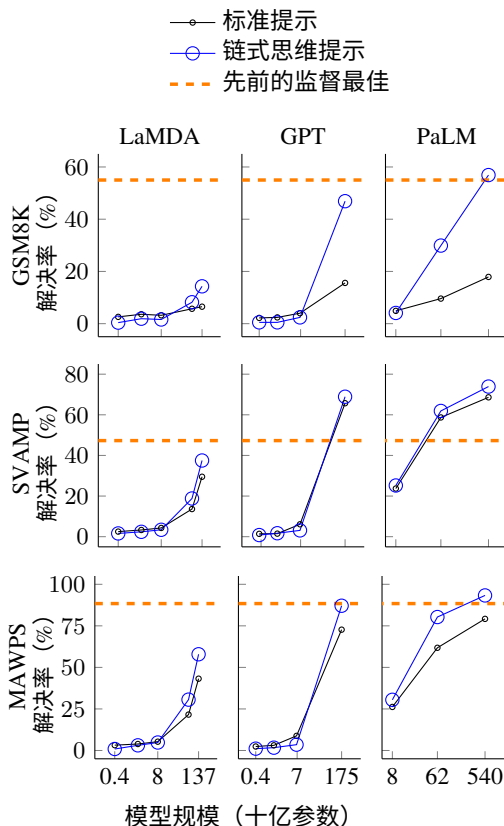


图4：链式思维提示使得大型语言模型能够解决具有挑战性的数学问题。值得注意的是，链式思维推理是随着模型规模的增加而出现的一种新能力。先前的最佳结果来自Cobbe等人（2021）的GSM8K，Jie等人（2022）的SVAMP，以及Lan等人（2021）的MAWPS。



仅计算变量。另一个直觉是，思维链允许模型在更难的问题上花费更多的计算（即中间标记）。为了将变量计算的效果与思维链推理隔离开来，我们测试了一种配置，其中模型被提示仅输出与解决问题所需的方程中的字符数相等的点序列（...）。这个变体的性能与基线相当，这表明仅仅变量计算并不是链式思维提示成功的原因，而且通过自然语言表达中间步骤似乎有用。

回答后的思维链。链式思维提示的另一个潜在好处可能仅仅是这样的提示使得模型能够更好地访问在预训练期间获得的相关知识。因此，我们测试了一种替代配置，其中思维链提示仅在回答之后给出，以确定模型是否实际上依赖于生成的思维链来给出最终答案。这个变体的性能与基线相当，这表明思维链中体现的顺序推理对于除了激活知识之外的其他原因也是有用的。

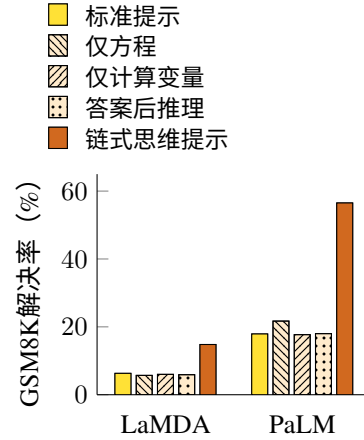


图5：使用LaMDA 137B和PaLM 540B进行不同提示变体的削减研究。其他数据集的结果在附录表6和表7中给出。

### 3.4 链式思维的稳健性

对示例的敏感性是提示方法的一个关键考虑因素——例如，改变少样本示例的排列可以导致GPT-3在SST-2上的准确性从接近随机（54.3%）到接近最先进（93.4%）（Zhao等，2021年）。在本节的最后一小节中，我们评估了对不同标注者编写的思维链的稳健性。除了上述使用由标注者A编写的思维链的结果外，本文的另外两位合著者（标注者B和C）独立为相同的少样本示例编写了思维链（附录H中显示）。标注者A还按照Cobbe等人（2021年）给出的解决方案风格，编写了另一种更简洁的思维链。<sup>1</sup>

图6展示了LaMDA 137B在GSM8K和MAWPS上的结果（其他数据集的消融结果见附录表6/表7）。尽管不同的思维链注释存在差异，这在使用基于示例的提示时是可以预期的（Le Scao和Rush, 2021; Reynolds和McDonell, 2021; Zhao等, 2021），但所有思维链提示集合都以较大的优势超越了标准基线。这个结果表明，成功使用思维链不依赖于特定的语言风格。

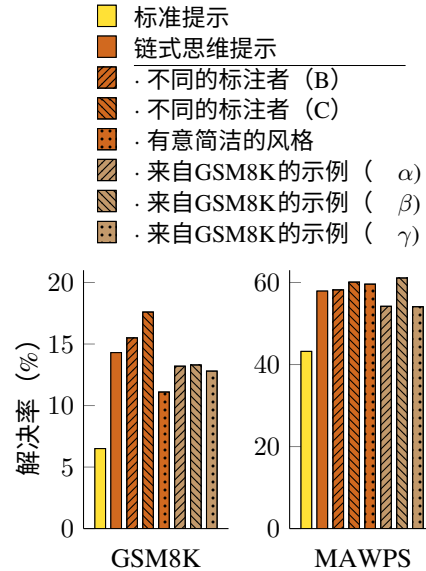


图6：链式思维提示对于不同的提示示例具有方差（如预期），但在不同的标注者和不同的示例中优于标准提示。

为了确认成功的链式思维提示对其他示例集也适用，我们还进行了实验。从GSM8K训练集中随机抽取了三组八个示例，这是一个独立的实验。

<sup>1</sup>例如，原始的思维链使用了几个短句（“最初有9台计算机。连续4天，每天增加5台计算机。所以 $5 * 4 = 20$ 台计算机被添加。 $9 + 20$ 等于29。”），简洁的思维链会这样写“ $5 * 4 = 20$ 台新计算机被添加。所以现在服务器房间里有 $9 + 20 = 29$ 台新计算机”。

源代码（此数据集中的示例已包含类似思维链的推理步骤）。<sup>2</sup>图6显示，这些提示与我们手动编写的示例相比表现出色，也明显优于标准提示。

除了对注释者、独立编写的思维链、不同的示例和各种语言模型具有鲁棒性外，我们还发现链式思维提示在算术推理中对不同示例顺序和不同示例数量具有鲁棒性（见附录A.2）。

## 4常识推理

尽管思维链特别适用于数学问题，但基于语言的思维链实际上适用于广泛的常识推理问题，这些问题涉及在一般背景知识的假设下对物理和人类交互进行推理。常识推理对于与世界互动至关重要，但目前的自然语言理解系统仍无法达到（Talmor等，2021年）。

**基准测试。**我们考虑了五个数据集，涵盖了各种常识推理类型。流行的CSQA（Talmor等，2019）提出了关于世界的常识问题，涉及复杂的语义，通常需要先前的知识。StrategyQA（Geva等，2021）要求模型推断出一种多跳策略来回答问题。我们从BIG-bench项目（BIG-bench collaboration, 2021）中选择了两个专门的评估集：日期理解，涉及从给定上下文中推断日期；体育理解，涉及确定与体育相关的句子是否合理。最后，SayCan数据集（Ahn等，2022）涉及将自然语言指令映射到离散集合中的机器人动作序列。图3显示了所有数据集的思维链注释示例。

**提示。**我们遵循与前一部分相同的实验设置。对于CSQA和StrategyQA，我们从训练集中随机选择了一些示例，并手动组成了思维链，作为少样本示例。两个BIG-bench任务没有训练集，所以我们选择了评估集中的前十个示例作为少样本示例，并报告其余评估集上的结果。对于SayCan，我们使用了Ahn等人在2022年使用的训练集中的六个示例，并手动组成了思维链。

**结果。**图7突出了PaLM的这些结果（LaMDA、GPT-3和不同模型规模的完整结果在表4中显示）。对于所有任务，增加模型大小可以提高标准提示的性能；链式思维提示进一步提高了性能，对于PaLM 540B来说改进效果最大。通过链式思维提示，PaLM 540B在策略问答（75.6% vs 69.4%）上相对于基准模型取得了强大的性能，并在体育理解上超过了一个没有辅助的体育爱好者（95.4% vs 84%）。这些结果表明，链式思维提示也可以提高在需要各种常识推理能力的任务上的性能（尽管在CSQA上的增益很小）。

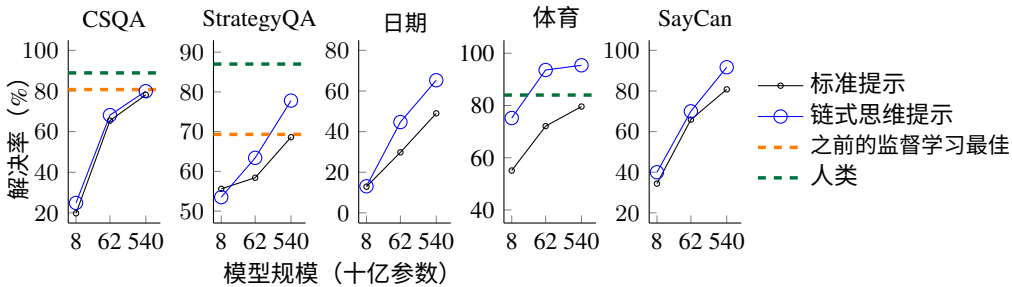


图7：链式思维提示还提高了语言模型的常识推理能力。这里展示的语言模型是PaLM。之前的最佳成绩来自CSQA（Talmor等，2019）和StrategyQA（Geva等，2021）的排行榜（仅限单模型，截至2022年5月5日）。表4展示了使用不同大小的LaMDA、GPT-3和PaLM的其他结果。

<sup>2</sup>我们随机选择例子  $\leq 60$  个标记以适应我们的输入上下文窗口，并且还将例子限制在  $\leq 2$  个步骤以便与我们编写的八个示例进行公平比较。

## 5 符号推理

我们最终的实验评估考虑了符号推理，这对人类来说很简单，但对语言模型来说可能具有挑战性。我们展示了链式思维提示不仅使语言模型能够在标准提示设置中执行具有挑战性的符号推理任务，而且还有助于在推理时处理比少样本示例中更长的输入。

任务。我们使用以下两个玩具任务。

- 最后一个字母连接。这个任务要求模型将一个名字中的单词的最后一个字母连接起来（例如，“Amy Brown” → “yn”）。<sup>3</sup>这是一个更具挑战性的版本的首字母连接，语言模型已经可以在没有链式思维的情况下执行。<sup>3</sup>我们通过随机连接来自姓名普查数据中前一千个名字和姓氏的名字来生成完整的名字（<https://namecensus.com/>）。
- 硬币翻转。这个任务要求模型回答在人们翻转或不翻转硬币后，硬币是否仍然是正面朝上（例如，“一个硬币是正面朝上的。菲比翻转了硬币。奥斯瓦尔多没有翻转硬币。硬币还是正面朝上吗？” → “不是”）。

由于这些符号推理任务的构建是明确定义的，对于每个任务，我们考虑一个领域内的测试集，其中的例子与步骤数相同

对于训练/少样本示例，以及一个域外(OOD)测试集，评估示例中的步骤比示例中的步骤多。对于最后一个字母的连接，模型只看到具有两个单词的名字示例，然后对具有3个和4个单词的名字执行最后一个字母的连接。<sup>4</sup>我们在硬币翻转任务中也是如此，对于可能的翻转次数。我们的实验设置使用与前两个部分相同的方法和模型。我们再次为每个任务的少样本示例手动组成思维链，如图3所示。

结果。这些领域内和领域外评估的结果如图8所示，PaLM的结果在附录表5中显示。<sup>5</sup>使用PaLM 540B，链式思维提示可以达到近100%的解决率（需要注意的是，标准提示已经可以解决PaLM 540B的硬币翻转问题，但对于LaMDA 137B则不行）。需要注意的是，这些领域内评估是“玩具任务”，因为少样本示例中的思维链已经提供了完美的解决结构；模型所需做的就是测试时使用新符号重复相同的步骤。然而，小型模型仍然失败了——对于这三个任务，只有拥有100B模型参数规模的模型才能在未见过的符号上执行抽象操作的能力。

至于领域外评估，标准提示对于这两个任务都失败了。通过链式思维提示，语言模型在上升曲线上实现了扩展（尽管性能低于领域内设置）。因此，链式思维提示有助于语言模型在足够规模下实现长度的泛化，超越已见思维链。

## 6 讨论

我们已经探索了链式思维提示作为一种简单的机制，用于激发大型语言模型中的多步推理行为。我们首先发现，链式思维提示在算术推理方面大幅提高了性能，改进效果比消融实验更强，并且对不同的注释者、示例和语言模型都具有鲁棒性（第3节）。接下来，

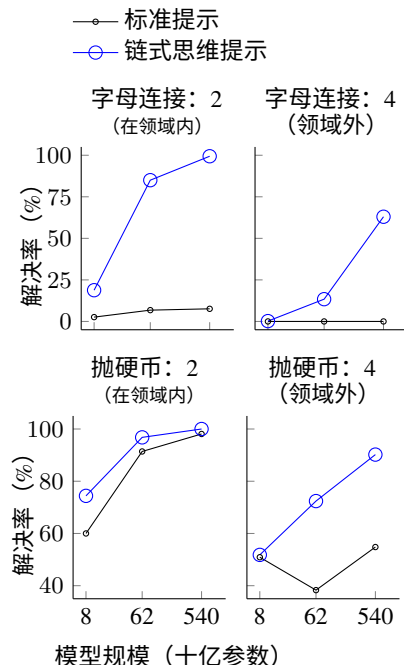


图8：使用链式思维提示有助于在两个符号推理任务中实现对更长序列的泛化。

<sup>3</sup>我们使用GPT-3达芬奇测试了10个常见的名字，除了一个之外，它都答对了。

<sup>4</sup>对于长度超过2个单词的名字，我们将多个姓和名字连接在一起。



常识推理实验证明了链式思维推理的语言性质使其具有普适性（第4节）。最后，我们展示了对于符号推理，链式思维提示有助于在更长的序列长度上进行OOD泛化（第5节）。在所有实验中，链式思维推理仅通过提示一个现成的语言模型来激发。在撰写本文的过程中，没有对语言模型进行微调。

链式思维推理的出现是模型规模的结果，这是一个主要的主题（Wei等人，2022b）。对于许多推理任务，标准提示具有平坦的缩放曲线，而链式思维提示导致缩放曲线显著增加。链式思维提示似乎扩展了大型语言模型能够成功执行的任务范围-换句话说，我们的工作强调标准提示只提供了大型语言模型能力的下限。这个观察可能引发比回答更多的问题-例如，随着模型规模的进一步增加，我们可以期望推理能力提高多少？

还有哪些提示方法可以扩展语言模型可以解决的任务范围？

至于限制，我们首先要说明的是，尽管思维链模拟了人类推理的思维过程，但这并不能回答神经网络是否真正“推理”，这是一个悬而未决的问题。其次，尽管在少样本情况下手动增加示例的思维链成本很低，但这种注释成本可能对微调来说是禁止的（尽管可以通过合成数据生成或零样本泛化来克服这个问题）。第三，没有正确推理路径的保证，这可能导致正确和错误的答案；改进语言模型的事实生成是未来工作的一个开放方向（Rashkin等人，2021；Ye和Durrett，2022；Wiegrefe等人，2022，等等）。最后，链式思维推理只在大型模型规模下出现，这使得在实际应用中成本高昂；进一步的研究可以探索如何在较小的模型中引发推理。

## 7 相关工作

这项工作受到许多研究领域的启发，我们在扩展的相关工作部分（附录C）中详细介绍。在这里，我们描述了两个方向和相关论文，可能是最相关的。

第一个相关方向是使用中间步骤来解决推理问题。Ling等人（2017年）通过一系列中间步骤使用自然语言解释来解决数学问题。他们的工作与使用形式语言进行推理的文献形成了鲜明对比（Roy等人，2015年；Chiang和Chen，2019年；Amini等人，2019年；Chen等人，2019年）。Cobbe等人（2021年）通过创建一个更大的数据集并使用它来微调预训练的语言模型，扩展了Ling等人（2017年）的工作，而不是从头开始训练模型。在程序合成领域，Nye等人（2021年）利用语言模型通过首行到行预测中间计算结果来预测Python程序的最终输出，并且展示了他们的逐步预测方法比直接预测最终输出效果更好。

自然地，这篇论文也与关于提示的大量最新工作密切相关。自从Brown等人（2020）提出了少样本提示的普及以来，几种通用方法已经改进了模型的提示能力，例如自动学习提示（Lester等人，2021）或给模型提供描述任务的指令（Wei等人，2022a；Sanh等人，2022；Ouyang等人，2022）。而这些方法改进或增强了提示的输入部分（例如，附加到输入之前的指令），我们的工作则采取了与之正交的方向，通过链式思维来增强语言模型的输出。

## 8 结论

我们探索了链式思维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。通过对算术、符号和常识推理的实验，我们发现链式思维推理是模型规模的一种新兴属性，使得足够大的语言模型能够执行在其他情况下具有平坦缩放曲线的推理任务。

扩大语言模型可以执行的推理任务范围，希望能激发对基于语言的推理方法的进一步研究。

## 致谢

我们感谢Jacob Devlin、Claire Cui、Andrew Dai和Ellie Pavlick对本文提供的反馈。我们感谢Jacob Austin、Yuhuai Wu、Henryk Michalewski、Aitor Lewkowycz、Charles Sutton和Aakanksha Chowdhery的有益讨论。我们感谢Sid Maxwell在原始手稿中提醒我们关于手动错误分析中的一个错误。

## 参考文献

Michael Ahn、Anthony Brohan、Noah Brown、Yevgen Chebotar、Omar Cortes、Byron David、Chelsa Finn、Keerthana Gopalakrishnan、Karol Hausman、Alex Herzog等。2022年。按照我能做的去做，而不是我说的去做：将语言基于机器人的可行性。arXiv预印本arXiv:2204.01691。

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi和Hannaneh Hajishirzi。2019年。MathQA：基于操作的形式主义解决可解释的数学问题。在北美计算语言学协会2019年会议论文集：人类语言技术，第1卷（长篇和短篇），明尼苏达州明尼阿波利斯。北美计算语言学协会。

Daniel Andor, Luheng He, Kenton Lee和Emily Pitler。2019年。给BERT一个计算器：通过阅读理解找到操作和参数。EMNLP。

Jacob Andreas, Dan Klein和Sergey Levine。2018年。学习与潜在语言。NAACL。

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le等。2021年。使用大型语言模型进行程序合成。arXiv预印本arXiv:2108.07732。

BIG-bench合作。2021年。超越模仿游戏：衡量和推断语言模型的能力。准备中。

Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri和Greg Durrett。2021年。灵活生成自然语言推断。EMNLP。

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和Dario Amodei。2020年。语言模型是少样本学习器。NeurIPS。

Jonathon Cai, Richard Shin和Dawn Song。2017年。通过递归使神经编程架构泛化。ICLR。

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz和Phil Blunsom。2018年。e-SNLI：自然语言推理与自然语言解释。NeurIPS。

Howard Chen, Jacqueline He, Karthik Narasimhan和Danqi Chen。2022年。合理化能改善鲁棒性吗？NAACL。

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman等。2021年。评估在代码上训练的大型语言模型。arXiv预印本arXiv:2107.03374。

Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, Dawn Song和Quoc V. Le。2019年。神经符号阅读器：用于阅读理解的分式和符号表示的可扩展集成。ICLR。

Ting-Rui Chiang和Yun-Nung Chen。2019年。用于解决和推理数学问题的语义对齐方程生成。在2019年北美计算语言学协会会议论文集：人类语言技术，卷1（长篇和短篇）中，页码2656-2668，明尼阿波利斯。计算语言学协会。

- Peter Clark, Oyvind Tafjord和Kyle Richardson. 2020年。 变压器作为语言的软推理器。 *IJCAI*。
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse和John Schulman. 2021年。 训练验证者解决数学问题。 arXiv预印本arXiv:2110.14168。
- Jacob Devlin, Ming-Wei Chang, Kenton Lee和Kristina Toutanova. 2019年。 BERT: 深度双向变压器的预训练, 用于语言理解。 *NAACL*。
- Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. 2019. 神经网络逻辑机器。 *ICLR*。
- Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. 中间注释在阅读理解中的好处。 *ACL*。
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. 亚里士多德使用笔记本电脑吗? 一个具有隐含推理策略的问答基准。 *TACL*。
- Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2022. DREAM: 揭示语言模型背后的心智模型。 *NAACL*。
- Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. 使用自然语言解释训练分类器。 *ACL*。
- 彼得·哈斯和莫希特·班萨尔。2022年。 模型何时能够从解释中学习? 一个用于理解解释数据角色的正式框架。 *ACL*。
- 丹·亨德里克斯, 科林·伯恩斯, 索拉夫·卡达瓦斯, 阿库尔·阿罗拉, 斯蒂文·巴萨特, 埃里克·唐, 唐·宋, 和雅各布·斯坦哈特。2021年。 用数学数据集衡量数学问题解决能力。 arXiv预印本 arXiv:2103.03874。
- Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni和Nate Kushman. 2014年。 学习解决带有动词分类的算术问题。 *EMNLP*。
- Zhanming Jie, Jierui Li和Wei Lu. 2022年。 学习进行演绎推理: 将数学问题解决视为复杂关系提取。 arXiv预印本arXiv:2203.10316。
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu和Dario Amodei. 2020年。 神经语言模型的缩放定律。 arXiv预印本arXiv:2001.08361。
- Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman和Hannaneh Hajishirzi. 2016年。 MAWPS: 一个数学问题存储库。 *NAACL*。
- Andrew K. Lampinen, Ishita Dasgupta, Stephanie C.Y. Chan, Kory Matthewson, Michael HenryTessler, Antonia Creswell, James L. McClelland, Jane X. Wang和Felix Hill. 2022年。 语言模型能否从上下文中学习解释? arXiv预印本arXiv:2204.02329。
- Yihuai Lan, Lei Wang, Qiyuan Zhang, Yunshi Lan, Bing Tian Dai, Yan Wang, Dongxiang Zhang, 和Ee-Peng Lim. 2021. MWPToolkit: 一个基于深度学习的数学问题求解开源框架。 arXiv预印本 arXiv:2109.00799。
- Teven Le Scao和Alexander Rush. 2021. 一个提示值多少个数据点? *NAACL*。
- Brian Lester, Rami Al-Rfou和Noah Constant. 2021. 规模的力量对于参数高效的提示调整。 *EMNLP*。
- Iddo Lev, Bill MacCartney, Christopher Manning和Roger Levy. 2004. 解决逻辑谜题: 从强大的处理到精确的语义。 第二届文本意义和解释研讨会论文集。
- Xiang Lisa Li和Percy Liang. 2021. 前缀调整: 优化生成的连续提示。 *ACL*。

郑中亮, Steven Bethard和Mihai Surdeanu. 2021年。通过内心独白解释多跳口头推理。 *NAA CL*.

Wang Ling, Dani Yogatama, Chris Dyer和Phil Blunsom. 2017年。通过理性生成进行程序归纳: 学习解决和解释代数问题。 *ACL*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi和Graham Neubig. 2021年。[预训练、提示和预测: 自然语言处理中提示方法的系统调查](#)。arXiv预印本arXiv:2107.13586.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz和Julian McAuley. 2021年。[启发常识的自然语言解释](#)。arXiv预印本arXiv:2106.13876.

Ana Marasovi'c, Iz Beltagy, Doug Downey和Matthew E Peters. 2022年。使用自然语言提示进行少样本自我合理化 *NAACL*发现。

Joshua Maynez, Shashi Narayan, Bernd Bohnet和Ryan McDonald. 2020年。关于抽象摘要的忠实性和事实性。在 *ACL*中。

Shen Yun Miao, Chao Chun Liang和Keh Yih Su. 2020年。用于评估和开发英语数学问题求解器的多样语料库。 *ACL*。

Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi和Luke Zettlemoyer. 2022年。重新思考演示的作用: 什么使得上下文学习起作用? arXiv预印本arXiv:2202.12837。

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel和Karishma Malkan. 2020年。[WT5? ! 训练文本到文本模型以解释其预测](#)。arXiv预印本arXiv:2004.14546。

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan等。2021年。展示你的工作: [使用语言模型进行中间计算的草稿本](#)。arXiv预印本arXiv:2112.00114。

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, ChongZhang, Sandhini Agarwal, Katarina Slama, Alex Ray等。2022年。训练语言模型以根据人类反馈遵循指令。arXiv预印本arXiv:2203.02155。

Arkil Patel, Satwik Bhattamishra和Navin Goyal. 2021年。自然语言处理模型真的能够解决简单的数学问题吗? *NAACL*。

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee和Luke Zettlemoyer. 2018年。深度上下文文化的词表示。 *NAACL*。

Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Yan Gao, Qiang Fu, Jian-Guang Lou和Weizhu Chen. 2022年。像程序执行者一样推理。arXiv预印本arXiv:2201.11473。

Piotr Pi'ekos, Mateusz Malinowski和Henryk Michalewski. 2021年。[通过预测推理顺序来衡量和改进BERT的数学能力](#)。 *ACL*。

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, 等。2021年。扩展语言模型: [训练Gopher的方法、分析和见解](#)。arXiv预印本arXiv:2112.11446。

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, 和 Peter J Liu. 2020年。探索使用统一的文本到文本转换器的迁移学习极限。机器学习研究杂志, 21:1-67。

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H. Hovy, 和 Yulia Tsvetkov. 2021年。SelfExplain: [神经文本分类器的自解释架构](#)。 *EMNLP*。

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong和Richard Socher. 2019年。[解释你自己! 利用语言模型进行常识推理](#)。 *ACL*。

邱然, 林彦凯, 李鹏, 周杰和刘智远。2019年。NumNet: 具有数值推理的机器阅读理解。 *EMNLP*。

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc和David Reitter。2021年。测量自然语言生成模型中的归因。arXiv预印本arXiv:2112.12870。

Gabriel Recchia。2021年。通过示范教授自回归语言模型复杂任务。arXiv预印本arXiv:2109.02102。

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch和Jason Wei。2022年。使用大型语言模型进行任意文本风格转换的配方。 *ACL*。

Laria Reynolds和Kyle McDonell。2021年。大型语言模型的提示编程: 超越少样本范式。2021年人机交互计算系统CHI会议摘要。

Subhro Roy和Dan Roth。2015年。解决一般算术问题。 *EMNLP*。

Subhro Roy, Tim Vieira和Dan Roth。2015年。关于自然语言中数量的推理。 *TACL*。

Mohammed Saeed, Naser Ahmadi, Preslav Nakov和Paolo Papotti。2021年。RuleBERT: 向预训练语言模型教授软规则。 *EMNLP*。

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja等。2022年。多任务提示训练实现零样本任务泛化。 *ICLR*。

Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu。2021。生成和排序: 一个用于数学问题的多任务框架。在计算语言学协会的发现中: *EMNLP 2021*。

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant。2019。CommonsenseQA: 一个针对常识知识的问答挑战。 *NAACL*。

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant。2020。Leap-of-thought: 教导预训练模型系统地推理隐含知识。 *NeurIPS*。

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant。2021。CommonsenseQA 2.0: 通过游戏化揭示人工智能的局限性。 *NeurIPS*数据集和基准赛道。

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby和Donald Metzler。2022年。统一语言学习范式。arXiv预印本 arXiv:2205.05131。

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du等。2022年。LaMDA: 用于对话应用的语言模型。arXiv预印本 arXiv:2201.08239。

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi和Denny Zhou。2022a年。自治性改善了语言模型中的思维链推理。arXiv预印本 arXiv:2203.11171。

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap等。2022b年。通过上下文指令对1,600多个语言任务进行基准测试。arXiv预印本 arXiv:2204.07705。

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai和Quoc V. Le。2022a年。精调语言模型是零-shot学习器。 *ICLR*。



- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler等。2022b年。大型语言模型的新兴能力。机器学习研究交易。
- Sarah Wiegrefe, Jack Hessel, Swabha Swayamdipta, Mark Riedl和Yejin Choi。2022年。重新构思人工智能协作以生成自由文本解释。 *NAACL*。
- Sarah Wiegrefe和Ana Marasović。2021年。教我解释：可解释NLP数据集综述。 *NeurIPS*。
- Sarah Wiegrefe, Ana Marasović和Noah A. Smith。2021年。测量标签与自由文本解释之间的关联。 *EMNLP*。
- Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry和Carrie J Cai。2022a年。PromptChainer：通过视觉编程链接大型语言模型提示。 *CHI*扩展摘要。
- Tongshuang Wu, Michael Terry和Carrie Jun Cai。2022b年。AI链：通过链接大型语言模型提示实现透明和可控的人工智能交互。 *CHI*。
- Yujun Yan, Kevin Swersky, Danai Koutra, Parthasarathy Ranganathan, and Milad Hashemi。2020。神经执行引擎：学习执行子程序。 *NeurIPS*。
- Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren。2021。通过组合解释改进语言模型。 *NeurIPS*。
- Xi Ye and Greg Durrett。2022。少样本情境学习中解释的不可靠性。 *arXiv预印本 arXiv:2205.03401*。
- Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu。2021。少样本领域外迁移学习的自然语言解释。 *arXiv预印本 arXiv:2112.06204*。
- Omar Zaidan, Jason Eisner, and Christine Piatko。2007。使用“注释者解释”来改进文本分类的机器学习。 *NAACL*。
- Wojciech Zaremba和Ilya Sutskever。2014年。学习执行。 *arXiv预印本 arXiv:1410.4615*。
- Eric Zelikman, Yuhuai Wu和Noah D. Goodman。2022年。STaR：用推理引导推理。 *arXiv预印本 arXiv:2203.14465*。
- Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein和Sameer Singh。2021年。使用前进行校准：改进语言模型的少样本性能。 *ICML*。
- Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong和Jian Tang。2020年。通过解释作为潜在变量的方式实现可解释的自然语言理解。 *NeurIPS*。

## 清单

1. 对于所有作者...
  - (a) 摘要和引言中提出的主要论点是否准确反映了论文的贡献和范围？ [是]
  - (b) 你是否描述了你的工作的局限性？ [是] 请参阅第6节和附录A.2。
  - (c) 你是否讨论了你的工作可能产生的负面社会影响？ [是] 我们不希望我们论文的贡献直接产生负面社会影响。然而，需要考虑的一点是，生成的思维链并不总是事实准确的，这在附录D.1中有所说明（请注意，我们不建议以事实方式或在任何现实场景中使用这样的思维链）。(d) 你是否阅读了伦理审查指南，并确保你的论文符合这些指南？ [是]
2. 如果你包含理论结果...
  - (a) 你是否陈述了所有理论结果的完整假设集？ [不适用]
  - (b) 你是否包含了所有理论结果的完整证明？ [不适用]
3. 如果你进行了实验...
  - (a) 你是否包含了重现主要实验结果所需的代码、数据和说明（可以作为补充材料或URL提供）？ [是] 我们在补充材料中包含了LaMDA和GPT-3的输入、输出和目标。虽然我们使用了专有模型，但GPT-3的结果是可以完全复现的。重现性在附录E.1中进一步讨论。
  - (b) 你是否指定了所有的训练细节（例如数据划分、超参数以及它们的选择方式）？ [是] 数据划分已经指定，超参数不适用。(c) 你是否报告了误差条（例如在多次运行实验后相对于随机种子的误差）？ [是] 使用LaMDA 137B进行多个种子的标准差，其中每个种子是不同的示例顺序，在表6和表7中给出。
  - (d) 你是否包含了使用的计算总量和资源类型（例如GPU类型、内部集群或云提供商）？ [是] 资源类型在附录E.2中描述，尽管我们没有估计计算的总量。
4. 如果您正在使用现有资产（例如代码、数据、模型）或策划/发布新资产...
  - (a) 如果您的工作使用了现有资产，您是否引用了创建者？ [是] 我们使用了两个模型，根据NeurIPS主席的建议进行了匿名处理。这些模型将在最终版本的论文中引用。
  - (b) 您是否提及了资产的许可证？ [是] 请参见附录E.3。
  - (c) 您是否在补充材料或URL中包含了任何新资产？ [是] 抛硬币和最后一个字母连接数据集是唯一的新资产，并且它们在补充材料中给出。
  - (d) 您是否讨论了从您使用/策划数据的人那里获得了同意的的问题以及如何获得的？ [N/A] 没有收集到人类数据。
  - (e) 您是否讨论了您使用/策划的数据是否包含个人可识别信息或冒犯性内容？ [N/A] 没有收集到人类数据。
5. 如果您使用众包或与人类参与者进行研究...
  - (a) 您是否包含了给参与者的完整指令文本和截图（如果适用）？ [不适用]
  - (b) 您是否描述了任何潜在的参与者风险，并提供了机构审查委员会（IRB）批准的链接（如果适用）？ [不适用]
  - (c) 您是否包括了参与者获得的预计每小时工资和总体补偿金额？ [不适用]

## 常见问题

### A.1 为什么增加模型规模可以改善链式思维提示？

成功的链式思维推理只在某些模型规模下可预测地出现的发现是有趣的。扩大语言模型的规模已被证明具有诸如提高性能和样本效率等好处（Kaplan等，2020年），但链式思维推理是一种新兴的推理方式，其成功不能仅通过推断小规模模型的性能来预测，因为对于小于100亿参数的大多数模型来说，链式思维实际上会损害性能。

为什么模型规模提高会改善链式思维提示的问题显然是多方面的，我们通过错误分析进行了初步尝试以揭示其中的见解。这个小的分析涉及手动阅读 PaLM 62B 所犯的 45 个错误，并将它们分为语义理解（20 个错误）、缺少一步（18 个错误）和其他错误（7 个错误）。“其他类别”包括幻觉、重复输出和符号映射错误。这种分类是粗略的借鉴了附录 D.2 对 LaMDA 进行的初始错误分析，其中的类别是根据需要使思维链正确的改进而构思的。

如图 9 所示，将 PaLM 扩展到 540B 参数可以修复所有三个类别中的大部分错误。图 10 给出了通过将 PaLM 扩展到 540B 修复的语义理解和缺少一步错误的示例。这个结果似乎与一个假设一致，即语言模型在模型规模的函数下获得了一系列的语义理解和逻辑推理能力（尽管需要注意的是，模型规模通常与其他因素混淆，如训练计算量）。

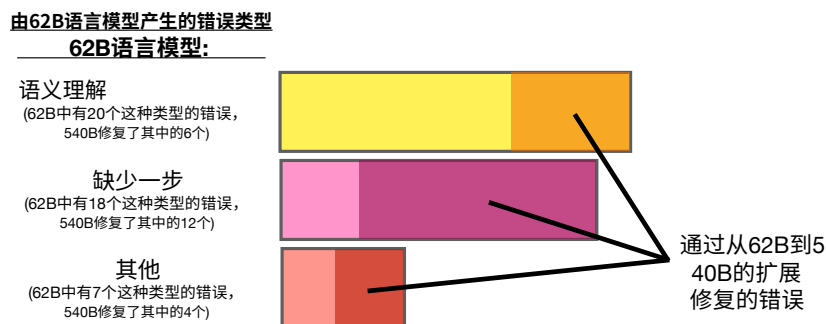


图9：PaLM 62B在45个问题中出现错误的错误分析。这些错误被归类为语义理解、缺少一步和其他。其他类别包括幻觉、重复输出和符号映射错误。将PaLM扩展到540B可以修复所有类别中的大部分错误。

关于为什么小型语言模型失败，还有三个值得注意的观点。第一个观察结果是，小型语言模型甚至在相对简单的符号映射任务上也失败了。正如在第5节中所示，即使是只需要使用与少样本示例中给定的相同思维链逻辑结构相同的符号推理任务，小型语言模型仍然失败了。第二个观察结果是，小型语言模型似乎在算术能力上本质上较弱，正如Brown等人(2020)所示，进行简单的算术运算(不需要语义理解)需要足够的模型规模。最后，我们注意到小型语言模型通常无法生成可以解析的最终答案，要么是重复，要么是逻辑上没有得出最终答案。

总结一下，链式思维推理的成功是模型规模的结果，这是一个复杂的现象，可能涉及到各种新兴的能力（语义理解、符号映射、专注于主题、算术能力、忠实度等）。未来的工作可以更全面地研究预训练数据、模型架构和优化目标的属性，以确定其对推理能力的因果影响。

<p><b>问题</b></p> <p>特蕾西用一根长4英尺的金属丝支撑花园里的番茄植物。这根金属丝被切成了6英寸长的小段。她得到了多少小段？</p>	<p><b>62B模型输出</b></p> <p>她把金属丝切成了6英寸长的小段。这意味着她得到了<math>4 * 6 = 24</math>个小段。答案是24。 (错误：语义理解错误)</p>	<p><b>540B模型输出</b></p> <p>这根金属丝长4英尺。这意味着它长<math>4 * 12 = 48</math>英寸。它被切成了6英寸长的小段。这意味着她得到了<math>48 / 6 = 8</math>个小段。答案是8。 (正确)</p>
<p><b>问题</b></p> <p>汤姆的船每小时可以行驶10英里。他从下午1点到4点航行。然后以6英里每小时的速度返回。他需要多长时间才能回来？</p>	<p><b>62B模型输出</b></p> <p>他从下午1点到4点旅行。这意味着他旅行了3个小时。他以6英里每小时的速度返回。这意味着他旅行了<math>3 * 6 = 18</math>英里。答案是18。 (错误：语义理解错误)</p>	<p><b>540B模型输出</b></p> <p>他从下午1点到4点以10英里每小时的速度旅行。这意味着他旅行了3个小时。以10英里每小时的速度旅行3个小时意味着他旅行了<math>3 * 10 = 30</math>英里。然后以6英里每小时的速度返回。这意味着他每小时旅行6英里。他必须旅行30英里，所以他需要<math>30 / 6 = 5</math>小时。答案是5。 (正确)</p>
<p><b>问题</b></p> <p>史蒂芬在线订购了杂货。他最终的账单金额为\$40.00。因为这是通过一个送货商，他们在最终总额上额外增加了25%的费用，并向他收取了\$3.00的送货费。史蒂芬还加了\$4.00的小费。在额外费用之后，史蒂芬的杂货最终价格是多少？</p>	<p><b>62B模型输出</b></p> <p>史蒂芬的杂货最终价格是\$40.00。送货商在最终总额上增加了25%，所以最终总额是\$40.00的125%。这意味着最终总额是<math>125% * \\$40.00 = \\$50.00</math>。送货商还收取了\$3.00的送货费。所以史蒂芬的杂货最终价格是<math>\\$50.00 + \\$3.00 = \\$53.00</math>。答案是\$53.00。 (错误：缺少一步计算)</p>	<p><b>540B模型输出</b></p> <p>史蒂芬的账单是\$40.00。送货商在账单上加了25%。这意味着账单增加了\$40.00的25%，即<math>0.25 * \\$40.00 = \\$10.00</math>。送货费是\$3.00。小费是\$4.00。所以总账单是<math>\\$40.00 + \\$10.00 + \\$3.00 + \\$4.00 = \\$57.00</math>。答案是57。 (正确)</p>
<p><b>问题</b></p> <p>有四所学校参加篮球比赛。每所学校都派了一支女子篮球队和一支男子篮球队，每支队伍有5名球员。每所学校还派了一名教练给每支队伍。总共，所有学校一共派了多少人？</p>	<p><b>62B模型输出</b></p> <p>有4所学校。每所学校都派了一支女子篮球队和一支男子篮球队。每支队伍都有5名球员。这意味着一共有<math>4 * 2 * 5 = 40</math>名球员。  每个团队都有一名教练。这意味着有<math>40 + 4 = 44</math>个教练。答案是44。 (错误：缺少一步错误)</p>	<p><b>540B模型输出</b></p> <p>每所学校都派出了2个队，每个队有5名球员。这意味着每所学校派出了<math>2 * 5 = 10</math>名球员。每所学校还派出了2名教练。这意味着每所学校派出了<math>10 + 2 = 12</math>人。一共有4所学校，所以所有学校一共派出了<math>4 * 12 = 48</math>人。答案是48。 (正确)</p>

图10：通过将PaLM从62B扩展到540B来修复语义理解和缺少一步错误的示例。

## A.2 提示工程的作用是什么？

提示的一个关键考虑因素是对确切提示的敏感性。大量的研究表明，提示以意想不到的方式影响语言模型（Min等，2022年）。我们创建链式思维注释的一般方法是从训练集中选择八个示例，并将推理过程分解为多个步骤，直到最终答案。

图3提供了链式思维注释的示例，并在附录G中提供了完整的提示。为了分析链式思维对提示工程的敏感性，我们进行了针对各种因素的鲁棒性实验。

- 不同的注释者。我们首先分析对三个不同注释者的鲁棒性（第3.4节和图6）。尽管性能存在显著差异（我们将在后面讨论），但在算术、常识和符号推理的八个数据集上，链式思维相对于基准模型表现出了明显的优势（表6和表7）。与Cobbe等人（2021）的注释过程类似，注释者没有收到关于具体说明的指示。

如何编写链式思维注释，除了简单地写出导致最终答案的逐步推理过程之外。因此，注释是按照每个注释者自己的语言“链式思维”写作风格编写的。

- 没有机器学习背景的注释者。GSM8K数据集（Cobbe等，2021年）方便地提供了一个由众包工人撰写的推理链训练集，这使我们能够研究来自独立来源且没有机器学习背景的推理链是否仍然有效。因此，我们随机抽取了三组具有GSM8K推理链的示例集。这些推理链注释在所有四个算术数据集上也大大优于基准线（表6），表明链式思维并不依赖于特定的注释者集。
- 不同的示例。上述不同的GSM8K示例实验（表6）还表明链式思维提示适用于不同的示例集。值得注意的是，我们在所有四个算术数据集上测试每组示例（而不是从训练集中选择示例），这表明示例不一定需要来自与测试示例相同的数据集分布。
- 不同的示例顺序。先前的研究表明，在某些情况下（例如，分类任务），提示的顺序甚至会影响到GPT-3在SST-2上的准确性，少样本示例的排列组合可以使准确率从接近随机（54.3%）到接近最佳结果（93.4%）（Zhao等，2021年）。我们在表6和表7中展示了不同示例的性能标准差。几乎在所有情况下，与提示顺序相关的标准差都相对较小。唯一的例外是硬币抛掷任务，其中示例顺序具有较高的标准差，可能是因为同一类别的示例连续出现会对模型输出产生偏差（Zhao等，2021年）。
- 不同数量的示例。我们还发现，在少样本示例数量不同的情况下，链式思维提示仍然具有优势。这在图11中展示了五个数据集（我们没有足够的计算资源来运行所有数据集）。我们还在初步实验中发现，进一步增加标准提示中的示例数量并没有带来显著的改进（例如，从8个增加到16个示例并没有使标准提示的性能足够提升以赶上链式思维提示）。
- 不同的语言模型。另一个有趣的问题是，对于一个模型有效的提示是否对其他大型语言模型也有效。我们发现，使用相同的提示，链式思维提示可以提高所有三个模型（LaMDA、GPT-3和PaLM）在除了GPT-3的CSQA和StrategyQA之外的所有数据集上的性能（表1、表4、表5）。链式思维的收益并没有完全在不同的模型之间转移，这是一个限制；进一步的工作可以研究为什么不同的预训练数据集和模型架构会影响链式思维提示的性能提升。

然而，提示工程仍然很重要。虽然对于算术推理，结果相对稳定，但我们要明确提示工程仍然很重要，并且在许多情况下可以显著提高性能。尽管大多数链式思维注释在许多情况下优于标准提示，但存在很大的变异性。例如，对于抛硬币任务，性能从注释者A的99.6%变化到注释者C的71.4%，尽管两者都高于标准提示的50.0%（见表7）。甚至有些任务需要提示工程才能获得良好的性能。在初步实验中，我们尝试使用链式思维使语言模型能够颠倒一个包含5个项目的列表的顺序。虽然两位合著者尽力而为，但无法编写出解决该任务的链式思维提示，但第三位合著者成功编写了一个完美解决该任务的链式思维提示。

如何以稳健的方式生成链式思维注释可能是未来工作的一个有趣方向。例如，一个想法是使用大型语言模型通过提示自动生成思维链（并可能在验证集上进行优化）。

### A.3链式思维提示会提高我感兴趣的任务的性能吗？

虽然链式思维提示原则上适用于任何文本到文本的任务，但对于某些任务来说，它比其他任务更有帮助。根据本文中的实验，我们的直觉是，当满足以下三个条件时，链式思维提示效果最好：（1）任务具有挑战性并需要



多步推理，（2）使用大型语言模型，以及（3）缩放曲线相对平缓。相反，当这些条件中的一个或多个不满足时，好处较小。这些直觉可能得到了算术推理结果的支持。链式思维提示的性能提升对于PaLM 540B在GSM8K上最大（具有挑战性的多步问题，平坦的缩放曲线），符合这些条件。对于只需要一两步的MAWPS子集（Single OP, SingleEq和AddSub），性能提升很小，对于这些子集，PaLM 540B的性能已经达到90%或更高（而且通常情况下，当性能已经强大时，改进的空间较小）。

尽管在本文中我们专注于多步推理任务（算术、常识和符号），但链式思维提示原则上可以应用于任何需要人类使用“思维链”解决的任务。我们将链式思维提示在这些不同任务（例如机器翻译等）上的实证评估留给未来的工作。

#### A.4 为什么仅仅使用方程作为提示对于某些算术推理数据集来说是不够的？

仅仅使用方程作为中间步骤的提示在许多数据集上确实有帮助，特别是当数据集只需要几个推理步骤时（SVAMP、ASDiv、MAWPS）。然而，对于GSM8K数据集，仅仅使用方程并没有显著提高性能。基于定性分析，我们认为这些问题对于模型来说在语义上过于具有挑战性，无法直接将其转化为数学方程。考虑一下来自LaMDA 137B的例子：

问题：迈克打乒乓球40分钟。在前20分钟，他得了4分。在后20分钟，他得了25%更多的分数。他总共得了多少分？

仅方程 (错误答案):  $(4 + 20 * 0.25) = 6$ . 答案是6.

思维链 (正确): 迈克打乒乓球40分钟。在前20分钟，他得了4分。在第二个20分钟，他得了25%更多的分数。所以他在第二个20分钟得了25%更多的分数。 $4 \times 1.25 = 5$ . 所以他在第二个20分钟得了5分。所以他总共得了9分。答案是9.

对于模型来说，直接将所有语义转化为一个方程是困难的，但是通过自然语言中的中间步骤，思维链可以更好地推理问题的每个部分。

## B 所有实验结果

本节包含了各种模型和模型大小的实验结果表格，涵盖了所有基准测试，比较了标准提示和思维链提示的效果。

对于算术推理基准，一些思维链（以及生成的方程式）是正确的，只是模型在执行算术运算时出现错误。Cobbe等人（2021）也做出了类似的观察。因此，我们可以进一步将Python程序作为外部计算器（使用Python的eval函数）添加到生成的思维链中的所有方程式中。

当思维链中有多个方程式时，我们通过字符串匹配将外部计算器的结果从一个方程式传递到后续的方程式中。如表1所示，我们可以看到在大多数任务上，添加计算器显著提升了链式思维提示的性能。

表1：链式思维提示在五个算术推理基准上优于标准提示的各种大型语言模型。所有指标均为准确率（%）。Ext. calc.：事后外部计算器仅用于算术计算。先前最佳结果来自以下几个研究。a：Cobbe等人（2021）。b & e：Pi等人（2022），c：Lan等人（2021），d：Pi et al.等人（2021）。

	提示	GSM8K	SVAMP	ASDiv	AQuA	MAWPS
之前最好的	N/A (微调)	55 <sup>a</sup>	57.4 <sup>b</sup>	75.3 <sup>c</sup>	37.9 <sup>d</sup>	88.4 <sup>e</sup>
UL2 20B	标准	4.1	10.1	16.0	20.5	16.6
	思维链	4.4 (+0.3)	12.5 (+2.4)	16.9 (+0.9)	23.6 (+3.1)	19.1 (+2.5)
	+ 扩展计算	6.9	28.3	34.3	23.6	42.7
LaMDA 137B	标准	6.5	29.5	40.1	25.5	43.2
	思维链	14.3 (+7.8)	37.5 (+8.0)	46.6 (+6.5)	20.6 (-4.9)	57.9 (+14.7)
	+ 扩展计算	17.8	42.1	53.4	20.6	69.3
GPT-3 175B (text-davinci-002)	标准	15.6	65.7	70.3	24.8	72.7
	思维链	46.9 (+31.3)	68.9 (+3.2)	71.3 (+1.0)	35.8 (+11.0)	87.1 (+14.4)
	+ 扩展计算	49.6	70.3	71.1	35.8	87.5
Codex (code-davinci-002)	标准	19.7	69.9	74.0	29.5	78.7
	思维链	63.1 (+43.4)	76.4 (+6.5)	80.4 (+6.4)	45.3 (+15.8)	92.6 (+13.9)
	+ 扩展计算	65.4	77.0	80.0	45.3	93.3
PaLM 540B	标准	17.9	69.4	72.1	25.2	79.2
	思维链	56.9 (+39.0)	79.0 (+9.6)	73.9 (+1.8)	35.8 (+10.6)	93.3 (+14.2)
	+ 扩展计算	58.6	79.8	72.6	35.8	93.5

表2：标准提示与思维链提示在五个算术推理基准上的对比。请注意，思维链提示是模型规模的一种新兴能力——只有在使用足够规模的模型时才会对性能产生积极影响。

模型		GSM8K		SVAMP		ASDiv		AQuA		MAWPS	
		标准	CoT	标准	CoT	标准	CoT	标准	CoT	标准	CoT
UL2	20B	4.1	<b>4.4</b>	10.1	<b>12.5</b>	16.0	<b>16.9</b>	20.5	<b>23.6</b>	16.6	<b>19.1</b>
LaMDA	420M	2.6	0.4	2.5	1.6	3.2	0.8	23.5	8.3	3.2	0.9
	2B	3.6	1.9	3.3	2.4	4.1	3.8	22.9	17.7	3.9	3.1
	8B	3.2	1.6	4.3	3.4	5.9	5.0	22.8	18.6	5.3	4.8
	68B	5.7	<b>8.2</b>	13.6	<b>18.8</b>	21.8	<b>23.1</b>	22.3	20.2	21.6	<b>30.6</b>
	137B	6.5	<b>14.3</b>	29.5	<b>37.5</b>	40.1	<b>46.6</b>	25.5	20.6	43.2	<b>57.9</b>
GPT	350M	2.2	0.5	1.4	0.8	2.1	0.8	18.1	8.7	2.4	1.1
	1.3B	2.4	0.5	1.5	1.7	2.6	1.4	12.6	4.3	3.1	1.7
	6.7B	4.0	2.4	6.1	3.1	8.6	3.6	15.4	13.4	8.8	3.5
	175B	15.6	<b>46.9</b>	65.7	<b>68.9</b>	70.3	<b>71.3</b>	24.8	<b>35.8</b>	72.7	<b>87.1</b>
Codex	-	19.7	<b>63.1</b>	69.9	<b>76.4</b>	74.0	<b>80.4</b>	29.5	<b>45.3</b>	78.7	<b>92.6</b>
PaLM	8B	4.9	4.1	15.1	<b>16.8</b>	23.7	<b>25.2</b>	19.3	<b>21.7</b>	26.2	<b>30.5</b>
	62B	9.6	<b>29.9</b>	48.2	46.7	58.7	<b>61.9</b>	25.6	22.4	61.8	<b>80.3</b>
	540B	17.9	<b>56.9</b>	69.4	<b>79.0</b>	72.1	<b>73.9</b>	25.2	<b>35.8</b>	79.2	<b>93.3</b>

表3：标准提示与链式思维提示在MAWPS的四个子集上的比较。划分MAWPS基准的目的是展示在简单的一步或两步问题上，大型语言模型已经取得了很高的性能（例如，SingleOp, SingleEq和AddSub），因此性能提升是最小的。SingleOp

模型				SingleEq		AddSub		MultiArith	
				标准	CoT	标准	CoT	标准	CoT
UL2	20B	24.9	27.2	18.0	20.2	18.5	18.2	5.0	10.7
LaMDA	420M	2.8	1.0	2.4	0.4	1.9	0.7	5.8	1.5
	2B	4.6	4.1	2.4	3.3	2.7	3.2	5.8	1.8
	8B	8.0	7.0	4.5	4.4	3.4	5.2	5.2	2.4
	68B	36.5	40.8	23.9	26.0	17.3	23.2	8.7	32.4
	137B	73.2	76.2	48.8	58.7	43.0	51.9	7.6	44.9
GPT	350M	3.2	1.8	2.0	0.2	2.0	1.5	2.3	0.8
	1.3B	5.3	3.0	2.4	1.6	2.3	1.5	2.2	0.5
	6.7B	13.5	3.9	8.7	4.9	8.6	2.5	4.5	2.8
	175B	90.9	88.8	82.7	86.6	83.3	81.3	33.8	91.7
Codex	-	93.1	91.8	86.8	93.1	90.9	89.1	44.0	96.2
PaLM	8B	41.8	46.6	29.5	28.2	29.4	31.4	4.2	15.8
	62B	87.9	85.6	77.2	83.5	74.7	78.2	7.3	73.7
	540B	94.1	94.1	86.5	92.3	93.9	91.9	42.2	94.7

表4：标准提示与链式思维提示在五个常识推理基准上的对比。链式思维提示是模型规模的一种新兴能力——只有在与足够大规模的模型一起使用时才会对性能产生积极影响。

模型		CSQA		StrategyQA		日期		体育		SayCan	
		标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT				
UL2	20B	34.2	<b>51.4</b>	59.0	53.3	13.5	<b>14.0</b>	57.9	<b>65.3</b>	20.0	<b>41.7</b>
LaMDA	420M	20.1	19.2	46.4	24.9	1.9	1.6	50.0	49.7	7.5	7.5
	2B	20.2	19.6	52.6	45.2	8.0	6.8	49.3	57.5	8.3	8.3
	8B	19.0	20.3	54.1	46.8	9.5	5.4	50.0	52.1	28.3	33.3
	68B	37.0	<b>44.1</b>	59.6	<b>62.2</b>	15.5	<b>18.6</b>	55.2	<b>77.5</b>	35.0	<b>42.5</b>
	137B	53.6	<b>57.9</b>	62.4	<b>65.4</b>	21.5	<b>26.8</b>	59.5	<b>85.8</b>	43.3	<b>46.6</b>
GPT	350M	14.7	15.2	20.6	0.9	4.3	0.9	33.8	41.6	12.5	0.8
	1.3B	12.0	19.2	45.8	35.7	4.0	1.4	0.0	26.9	20.8	9.2
	6.7B	19.0	<b>24.0</b>	53.6	50.0	8.9	4.9	0.0	4.4	17.5	<b>35.0</b>
	175B	79.5	73.5	65.9	65.4	43.8	<b>52.1</b>	69.6	<b>82.4</b>	81.7	<b>87.5</b>
Codex	-	82.3	77.9	67.1	<b>73.2</b>	49.0	<b>64.8</b>	71.7	<b>98.5</b>	85.8	<b>88.3</b>
PaLM	8B	19.8	<b>24.9</b>	55.6	53.5	12.9	13.1	55.1	<b>75.2</b>	34.2	<b>40.0</b>
	62B	65.4	<b>68.1</b>	58.4	<b>63.4</b>	29.8	<b>44.7</b>	72.1	<b>93.6</b>	65.8	<b>70.0</b>
	540B	78.1	<b>79.9</b>	68.6	<b>77.8</b>	49.0	<b>65.3</b>	80.5	<b>95.4</b>	80.8	<b>91.7</b>

表5：标准提示与链式思维提示使得在两个符号操作任务上能够推广到更长的推理示例。

模型		最后一个字母连接						抛硬币（状态跟踪）					
		2		OOD: 3		OOD: 4		2		OOD: 3		OOD: 4	
		标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT	标准CoT
UL2	20B	0.6	<b>18.8</b>	0.0	0.2	0.0	0.0	70.4	67.1	51.6	52.2	48.7	50.4
LaMDA	420M	0.3	<b>1.6</b>	0.0	0.0	0.0	0.0	52.9	49.6	50.0	50.5	49.5	49.1
	2B	2.3	<b>6.0</b>	0.0	0.0	0.0	0.0	54.9	<b>55.3</b>	47.4	48.7	49.8	50.2
	8B	1.5	<b>11.5</b>	0.0	0.0	0.0	0.0	52.9	<b>55.5</b>	48.2	49.6	51.2	50.6
	68B	4.4	<b>52.0</b>	0.0	<b>0.8</b>	0.0	<b>2.5</b>	56.2	<b>83.2</b>	50.4	<b>69.1</b>	50.9	<b>59.6</b>
	137B	5.8	<b>77.5</b>	0.0	<b>34.4</b>	0.0	<b>13.5</b>	49.0	<b>99.6</b>	50.7	<b>91.0</b>	49.1	<b>74.5</b>
PaLM	8B	2.6	<b>18.8</b>	0.0	0.0	0.0	<b>0.2</b>	60.0	<b>74.4</b>	47.3	<b>57.1</b>	50.9	<b>51.8</b>
	62B	6.8	<b>85.0</b>	0.0	<b>59.6</b>	0.0	<b>13.4</b>	91.4	<b>96.8</b>	43.9	<b>91.0</b>	38.3	<b>72.4</b>
	540B	7.6	<b>99.4</b>	0.2	<b>94.8</b>	0.0	<b>63.0</b>	98.1	<b>100.0</b>	49.3	<b>98.6</b>	54.8	<b>90.2</b>

表6: 算术推理数据集的消融和鲁棒性结果。链式思维通常以较大的优势胜过消融。“仅方程式”在标准提示和链式思维提示之间表现, 它通过方程式允许中间推理步骤, 但不利用自然语言。使用不同注释者编写的提示或使用其他示例时, 链式思维提示存在差异(如预期), 但仍然大大优于标准提示。所示标准差是针对少量示例的不同顺序, 使用五个不同的随机种子。这里显示的结果是针对LaMDA 137B的, 因为对GPT-3和PaLM的额外查询都是有限且昂贵的。

	GSM8K	SVAMP	ASDiv	MAWPS
标准提示	6.5 $\pm$ 0.4	29.5 $\pm$ 0.6	40.1 $\pm$ 0.6	43.2 $\pm$ 0.9
链式思维提示	14.3 $\pm$ 0.4	36.7 $\pm$ 0.4	46.6 $\pm$ 0.7	57.9 $\pm$ 1.5
<b>消融</b>				
· 仅方程式	5.4 $\pm$ 0.2	35.1 $\pm$ 0.4	45.9 $\pm$ 0.6	50.1 $\pm$ 1.0
· 仅变量计算	6.4 $\pm$ 0.3	28.0 $\pm$ 0.6	39.4 $\pm$ 0.4	41.3 $\pm$ 1.1
· 回答后推理	6.1 $\pm$ 0.4	30.7 $\pm$ 0.9	38.6 $\pm$ 0.6	43.6 $\pm$ 1.0
<b>鲁棒性</b>				
· 不同的标注者 (B)	15.5 $\pm$ 0.6	35.2 $\pm$ 0.4	46.5 $\pm$ 0.4	58.2 $\pm$ 1.0
· 不同的标注者 (C)	17.6 $\pm$ 1.0	37.5 $\pm$ 2.0	48.7 $\pm$ 0.7	60.1 $\pm$ 2.0
· 有意简洁的风格	11.1 $\pm$ 0.3	38.7 $\pm$ 0.8	48.0 $\pm$ 0.3	59.6 $\pm$ 0.7
· 来自GSM8K的示例 ( $\alpha$ )	12.6 $\pm$ 0.6	32.8 $\pm$ 1.1	44.1 $\pm$ 0.9	53.9 $\pm$ 1.1
· 来自GSM8K ( $\beta$ ) 12.7的示例	$\pm$ 0.5	34.8 $\pm$ 1.1	46.9 $\pm$ 0.6	60.9 $\pm$ 0.8
· 来自GSM8K ( $\gamma$ ) 12.6的示例	$\pm$ 0.7	35.6 $\pm$ 0.5	44.4 $\pm$ 2.6	54.2 $\pm$ 4.7

表7: 常识和符号推理中四个数据集的消融和鲁棒性结果。链式思维通常比消融方法表现更好。链式思维提示在使用不同注释者编写的提示或使用其他示例时具有方差(如预期), 但仍然大大优于标准提示。标准差显示了使用不同的少样本提示示例顺序时的情况, 使用了五个不同的随机种子。这里的结果是针对LaMDA 137B显示的, 因为对于GPT-3和PaLM来说, 额外的查询都是有限且昂贵的。唯一的例外是我们在这里使用PaLM运行SayCan, 因为SayCan评估集只有120个示例, 因此多次运行的成本较低。

	常识			符号	
	日期	体育	SayCan	连接	硬币
标准提示	21.5 $\pm$ 0.6	59.5 $\pm$ 3.0	80.8 $\pm$ 1.8	5.8 $\pm$ 0.6	49.0 $\pm$ 2.1
链式思维提示 26.8	$\pm$ 2.1	85.8 $\pm$ 1.8	91.7 $\pm$ 1.4	77.5 $\pm$ 3.8	99.6 $\pm$ 0.3
<b>消融</b>					
· 仅计算变量	21.3 $\pm$ 0.7	61.6 $\pm$ 2.2	74.2 $\pm$ 2.3	7.2 $\pm$ 1.6	50.7 $\pm$ 0.7
· 回答后推理	20.9 $\pm$ 1.0	63.0 $\pm$ 2.0	83.3 $\pm$ 0.6	0.0 $\pm$ 0.0	50.2 $\pm$ 0.5
<b>鲁棒性</b>					
· 不同的注释者 (B)	27.4 $\pm$ 1.7	75.4 $\pm$ 2.7	88.3 $\pm$ 1.4	76.0 $\pm$ 1.9	77.5 $\pm$ 7.9
· 不同的注释者 (C)	25.5 $\pm$ 2.5	81.1 $\pm$ 3.6	85.0 $\pm$ 1.8	68.1 $\pm$ 2.2	71.4 $\pm$ 11.1



## C 扩展相关工作

链式思维提示是一种通用方法，受到几个先前方向的启发：提示式编程，自然语言解释，程序合成/执行，数值和逻辑推理，以及中间语言步骤。

### C.1 提示

大规模语言模型的最近成功引起了对通过提示提高其执行任务能力的兴趣（Brown等人（2020），参见Liu等人（2021）的调查）。本文属于通用提示方法的范畴，通过优化输入提示，使单个大型语言模型能够更好地执行各种任务（Li和Liang，2021；Lester等人，2021；Reif等人，2022，等等）。

最近的一项工作旨在通过提供描述任务的指令（Raffel等人，2020年；Wei等人，2022年a；Ouyang等人，2022年；Sanh等人，2022年；Wang等人，2022年b）来提高语言模型执行任务的能力。这一工作与之相关，因为它还通过元数据增强了输入-输出对。但是，指令是对任务输入的增强（指令通常添加到输入之前），而链式思维提示是对语言模型输出的增强。另一个相关的方向是顺序组合语言模型的输出；人机交互（HCI）工作（Wu等人，2022年a，b）表明，组合语言模型的顺序生成可以提高20人用户研究中的任务结果。

### C.2 自然语言解释

另一个密切相关的方向是使用自然语言解释（NLE），通常旨在提高模型的可解释性（Zhou等人，2020年；Wiegreffe和Marasović，2021年，等）。这一工作通常侧重于自然语言推理（Camburu等人，2018年；Yordanov等人，2021年；Bostrom等人，2021年），并且生成解释要么与最终预测同时进行，要么在最终预测之后进行（Narang等人，2020年；Majumder等人，2021年；Wiegreffe等人，2021年，2022年）。相比之下，本文考虑的思维链处理发生在最终答案之前。虽然NLE主要旨在提高神经网络的可解释性（Rajagopal等人，2021年），但链式思维提示的目标是使模型能够将多跳推理任务分解为多个步骤-可解释性只是一个副作用。Marasović等人（2022年）表明，基于提示的微调与NLE改善了NLI和分类性能，尽管他们主要关注评估解释的合理性。相比之下，我们的工作侧重于需要多跳推理的一系列算术、常识和符号任务。

### C.3 程序合成和执行

在程序合成和执行中，使用中间推理步骤已经有很长的历史（Zaremba和Sutskever，2014年，等等）。最近在这个方向上的工作包括了许多架构创新（Cai等，2017年；Dong等，2019年；Yan等，2020年），以及大型语言模型的使用（Chen等，2021年；Austin等，2021年）。与我们最接近的程序执行工作可能是Nye等人的研究（2021年），他们表明大型语言模型可以执行多达10位数的加法，求解多项式，并执行Python程序。生成一个程序然后执行它可以被视为一种推理类型，我们的工作将这种领域特定的原语推广到自然语言，这是开放领域的，并且与任何文本到文本的自然语言处理任务相关。

### C.4 数值和逻辑推理

数值和逻辑推理是机器学习和自然语言处理中长期研究的任务之一（Lev等，2004年，等等）。最近的研究还试图以各种方式将数值推理能力注入语言模型中，例如通过将BERT与一组预定义的可执行操作进行增强（Andor等，2019年），包括使用图神经网络（Ran等，2019年）和使用专门的训练程序（Pilekos等，2021年）。另一方面的研究旨在使语言模型能够进行逻辑或形式推理，通常通过使用自然语言形式规则中的规则来进行语言化（Clark等，2020年；Saeed等，2021年；Liang等，2021年）。

在这方面，最相关的工作可能是Recchia（2021年），该工作表明微调使得长手模块操作成为可能，而这在以前对表演者来说是困难的。虽然在这个方向上的工作通常是任务特定的并使用微调，但我们展示了链式思维提示在广泛的任务中都可以工作，而无需任何微调。

## C.5 中间语言步骤

广泛的先前工作表明，通过训练或微调赋予神经网络产生中间步骤的能力在各种情况下都有益处。例如，已经证明自然语言的中间步骤可以提高性能(Zaidan et al., 2007; Yao et al., 2021; Hase and Bansal, 2022; Gu et al., 2022)，提高鲁棒性。

(Chen et al., 2022)加快训练速度(Hancock et al., 2018)，减轻偏见(Dua et al., 2020)，甚至在图像和强化学习环境有所帮助(Andreas et al., 2018)。为了赋予模型产生中间步骤的能力，先前的工作通常在手动注释的训练数据集上微调模型(Camburu et al., 2018; Rajani et al., 2019, *inter alia*)，或者生成合成数据集(Talmor et al., 2020; Zelikman et al., 2022)。与这些训练或微调方法相比，我们的工作表明，只需通过提示就可以在足够规模的现成语言模型中激发各种自然语言推理能力。这种提示设置很重要，因为它允许进行中间步骤推理而无需大量标记注释，并且一个模型可以执行各种推理任务而无需任何梯度更新。

## D 附录：额外分析

### D.1 正确的思维链分析

正如主文中所提到的，我们分析了来自LaMDA 137B的50个思维链，这些思维链在GSM8K数据集中导致了正确的答案。在这50个思维链中，只有一个通过错误的推理得出了正确的答案（在表9中显示：“偶然正确”）。其他49个思维链具有正确的逻辑和数学，示例显示在表8中。其中五个在保持连贯和可理解的逻辑的同时存在轻微的缺陷：

- 一个存在不完整的陈述（在表9中显示：“正确但不完整的陈述”）。
- 一个提出了无关但正确的陈述（在表9中显示：“正确但无关的陈述”）。
- 两个在方程中省略了具体步骤，但仍然可以理解（在表9中显示：“正确但不完美的方程”）。
- 一个在数学上是正确的，但颠倒了语义（在表9中显示：“正确但颠倒的语义”）

表8：模型为GSM8K数据集生成的正确思维链示例。

---

问题：詹姆斯决定每周跑3次冲刺，每次跑3次。他每次冲刺跑60米。他每周总共跑多少米？

模型答案（正确）：詹姆斯决定每周跑3次冲刺，每次跑3次。他每次冲刺跑60米。所以他每周跑60米 x 3次冲刺 x 3次。那就是60米 x 9。答案是540。✓

---

问题：布兰登的iPhone的年龄是本iPhone的四倍。本的iPhone的年龄是苏茜的iPhone的两倍。如果苏茜的iPhone已经1岁了，布兰登的iPhone多大？

模型答案（正确）：布兰登的iPhone的年龄是本iPhone的4倍。本的iPhone的年龄是苏茜的iPhone的2倍。所以布兰登的iPhone是4 x 2 = 8倍于苏茜的iPhone。苏茜的iPhone已经1岁。所以布兰登的iPhone是8 x 1 = 8岁。答案是8。✓

---

问题:珍妮有30个棒棒糖。珍妮吃掉了其中的2个棒棒糖。剩下的棒棒糖，珍妮想要把2个放在一个袋子里。珍妮能装多少个袋子？

模型答案（正确）:珍妮一开始有30个棒棒糖。她吃掉了其中的2个。所以她剩下28个棒棒糖。她想要把2个棒棒糖放在一个袋子里。所以她能装28 / 2 = 14个袋子。答案是14。✓

---

尽管我们发现链式思维在数学问题中大多数情况下是正确的，但这很可能是通过巧合得出正确的最终答案对于自由回答问题来说是困难的。对于多项选择或二元分类问题，情况要好得多

模型可能通过错误的推理路径得出正确答案（例如，我们评估的所有常识推理数据集）。这是一个限制，未来的工作应该对这种思维链的事实性进行分析。

表9：在模型正确的50个示例中，有7个显著的案例，模型生成的思维链是不完美的。我们发现其他43个与人类可能生成的基本事实相等。

---

问题：有36只企鹅在雪地上晒太阳。其中三分之一的企鹅跳进海里游泳。

另外三分之一的企鹅进入洞穴吃晚餐。还有多少只企鹅在继续晒太阳？

模型答案（仅凭运气正确）：有36只企鹅。其中三分之一的企鹅跳进海里游泳。所以是12只企鹅。

另外三分之一的企鹅进入洞穴吃晚餐。所以是12只企鹅。

答案是12。

---

问题：约翰有2所房子，每所房子有3间卧室。每间卧室有2扇窗户。每所房子还有4扇不与卧室相连的窗户。两所房子之间总共有多少窗户？

模型答案（正确但不明确的陈述）：每所房子有3间卧室。每间卧室有2扇窗户。所以每所房子有 $2 \times 3 = 6$ 扇窗户。每所房子还有额外的4扇不与卧室相连的窗户。所以每所房子有 $4 + 4 = 8$ 扇窗户。所以每所房子有 $6 \times 2 = 12$ 扇窗户。所以两所房子总共有 $12 + 8 = 20$ 扇窗户。答案是20。

---

问题：如果安妮9岁，她的兄弟是她年龄的两倍，那么3年后她的兄弟多大？

MODEL ANSWER（正确但无关的陈述）：安妮9岁。她的兄弟是她的两倍年龄。所以她的兄弟18岁。3年后，她将12岁。所以她的兄弟将会 $18 + 3 = 21$ 岁。答案是21。

---

QUESTION：一个当地的加油站正在以每加仑3.00美元的价格出售汽油。一家应用公司在这个加油站加油时提供每加仑0.20美元的返现。如果有人购买10加仑的汽油，返现后他们的汽油价格是多少？

MODEL ANSWER（正确但不完美的方程）：汽油原价是每加仑3美元。如果你购买10加仑，那就是 $3 \times 10 = 30$ 美元。如果你每加仑减少20美分，那就是 $20 \times 10 = 2$ 美元。所以总共是 $30 - 2 = 28$ 美元。答案是28美元。

---

问题：尼莎雇佣了60名季节性工人在她的百货公司圣诞村扮演小精灵。三分之一的小精灵在孩子们吐在他们身上后辞职，然后剩下的小精灵中有10个在孩子们踢他们的小腿后辞职。还剩下多少小精灵？

模型答案（正确但不完美的方程式）：尼莎雇佣了60名季节性工人。其中三分之一辞职。所以 $60 - 1/3 = 40$ 个小精灵还剩下。然后有10个小精灵辞职。所以 $40 - 10 = 30$ 个小精灵还剩下。答案是30个小精灵。

---

问题：琼斯奶奶为消防员午餐会烤了5个苹果派。她把每个派切成8块，并把这五个派放在自助餐桌上供客人自己取用。晚上结束时，客人们拿走并吃掉了他们的派块后，还剩下14块派。客人们拿走了多少块派？

MODEL ANSWER（正确但语义颠倒）：5个馅饼被烤制并切成8块。

然后将这5个馅饼分给了客人。 $8 \times 5 = 40$ 块馅饼。客人吃了14块馅饼。 $40 - 14 = 26$

剩下的馅饼。答案是26。✓

---

## D.2 错误的思维链分析

我们还手动分析了模型在GSM8K上的50个随机抽样输出，这些输出在LaMDA 137B上是错误的。思维链的错误方式有很多，使得错误分类的设计变得复杂。我们决定将错误分类为需要进行哪些更改，以使思维链变得正确，目的是阐明如何改进模型。

我们发现许多思维链可以通过以下三类修改之一来变得正确。

- **仅计算器错误。**我们发现，除了计算器错误外，8%的思维链完全正确，换句话说，将外部计算器应用于方程，就像Cobbe等人（2021年）所做的那样，会使思维链变得正确。这种类型的错误示例在表10中显示：“仅计算器错误”。事实上，当我们添加一个Python程序作为外部计算器时，链式思维提示在LaMDA 137B GSM8K上的解决率从14.3%提高到17.3%，如表2所示。此外，34%的示例中除了其他类型的错误外还包含计算器错误。然而，我们在计算器错误之外独立进行其余的错误分类。
- **符号映射错误。**接下来，我们发现16%的思维链除了我们称之为符号映射错误的错误外是正确的。我们将符号映射错误定义为思维链除了数字符号之外是正确的，通过仅修改方程而不是单词，可以使其完全正确。正如有人可能会争辩说他们可以简单地将正确的最终方程放入任何思维链中，我们将此类别限制为思维链可以修改为完全正确的推理过程（不仅仅是最终答案）。这个错误类别的示例在表10中显示：“符号映射错误”。
- **缺少一步错误。**我们下一个错误类别是思维链，除了缺少一步之外是正确的。换句话说，这些思维链可以通过添加一个被模型忽略的额外推理步骤来重新写成正确的。这个错误类别的一个例子在表10中显示：“缺少一步错误”。我们发现22%的错误属于这个类别。

表10：错误思维链的示例，按附录D.2中描述的方式分类。

---

问题：丹种了3棵玫瑰花丛。每棵玫瑰花丛有25朵玫瑰。每朵玫瑰有8个刺。总共有多少刺？

模型答案（不正确；仅计算器错误）：丹种了3棵玫瑰花丛。每棵玫瑰花丛有25朵玫瑰。每朵玫瑰有8个刺。所以 $3 \times 25 \times 8 = 300$ 。答案是300。✗

**ERROR CATEGORY**的解释：通过将方程式输入外部计算器而不是要求模型进行计算，可以使思维链变得正确。

---

问题：吉尔每小时教课的报酬是20美元，担任啦啦队教练的报酬是30美元。如果她一年工作50周，每周作为教师工作35小时，作为教练工作15小时，她的年薪是多少？

模型答案（错误；符号映射错误）：吉尔每小时教课的报酬是20美元，担任啦啦队教练的报酬是30美元。如果她一年工作50周，每周作为教师工作35小时，作为教练工作15小时，那么她作为教师工作 $50 \times 35 = 1750$ 小时，作为教练工作 $15 \times 30 = 450$ 小时。所以她一共工作了 $1750 + 450 = 2200$ 小时。她作为教师工作了1750小时，每小时获得20美元，作为教练工作了450小时，每小时获得30美元。所以她的年薪是 $20 \times 1750 + 30 \times 450 = 36,500$ 美元。答案是36,500美元。✗

**ERROR CATEGORY**的解释：模型错误地映射了“ $15 \times 30$ ”而不是“ $15 \times 50$ ”，通过修正这个错误（使用外部计算器）可以使思维链变得正确。

---

问题：Kelian有两个烹饪菜肴的食谱，一个有20个步骤，第二个有第一个的两倍多的步骤。Kelian需要阅读多少个步骤来准备这两道菜？

模型答案（错误；缺少一步错误）：Kelian有两个食谱。一个有20个步骤。另一个有第一个的两倍多的步骤。所以第二个有40个步骤。所以Kelian需要阅读40个步骤来准备这两道菜。答案是40。✗

错误类别解释：只需添加一个额外的步骤，模型答案就可以变得正确（在这种情况下，添加20和40）。

---

我们发现剩下的思维链（50个中的27个; 54%）需要进行大量编辑才能变成正确的思维链。几乎所有情况都涉及语义理解方面的错误（见表11：“语义理解错误”），其中27个中有8个还存在不连贯的思维链，即生成的思维链中的某些陈述与之前的陈述不一致或违反基本世界知识（见表11：“不连贯的思维链错误”）。

表11：错误的思维链示例，按附录D.2中描述的方式进行分类。这些思维链有多个无法通过小修复来解决的错误。

---

问题：格雷琴有110个硬币。金币比银币多30个。格雷琴有多少个金币？

模型答案（错误；语义理解错误）：格雷琴有110个硬币。金币比银币多30个。所以有  $110 - 30 = 80$  个银币。所以有80个银币和  $110 - 80 = 30$  个金币。答案是30。✗

错误类别的解释：模型不理解110个硬币是总数，它必须解方程来得到金币和银币的数量。

---

问题：在一个舞蹈班有20名学生，其中20%的学生报名学习现代舞，剩下的学生中有25%报名学习爵士舞，其余的学生报名学习街舞。有多少百分比的学生报名学习街舞？

**MODEL ANSWER**（错误的；不连贯的思维链错误）：学生中的20%参加了现代舞。剩下的学生中有25%参加了爵士舞。其余的学生参加了街舞。因此，整个学生中参加街舞的百分比是参加街舞的学生的百分比减去参加现代舞的学生的百分比，再减去参加爵士舞的学生的百分比。因此，参加街舞的整个学生的百分比是  $(25 + 20) - (25 + 20) = 100\%$ 。答案是100%。✗

**EXPLANATION FOR ERROR CATEGORY**：这种思维链条不连贯，因为整个学生中报名跳街舞的百分比不能是报名跳街舞的学生百分比减去另一个术语。

---

总的来说，大型语言模型生成的推理过程既不一定连贯，也不一定事实正确，正如最近评估语言模型生成和解释的工作所强调的那样（Maynez等，2020年；Rashkin等，2021年；Ye和Durrett，2022年；Marasović等，2022年；Wiegrefe等，2022年）。错误的推理过程既可能导致错误的最终答案，也可能导致意外的正确最终答案（对于二元分类等任务而言，意外的正确最终答案更有可能）。改进语言模型生成的事实性，使其与上下文和世界知识相符，是语言模型研究中一个重要的方向和潜在的改进多步推理能力的方法。改进解码质量的一种潜在方法可能涉及生成多个推理路径并用验证器对每个路径进行评分，但这需要训练验证器（Cobbe等，2021年；Shen等，2021年；Thoppilan等，2022年）。

### D.3 附加鲁棒性分析

由于主要论文中的实验使用了固定数量的少样本示例（8个；受1024个标记的输入长度限制），我们验证了链式思维提示对不同数量的少样本示例的鲁棒性。我们对LaMDA 137B进行了实验，比较了链式思维提示和标准提示在五个数据集上的效果，其中标准提示的性能曲线基本平缓（最大模型未能达到高性能）。如图11所示，链式思维提示相对于标准提示的改进在不同数量的少样本示例的提示中仍然具有鲁棒性。



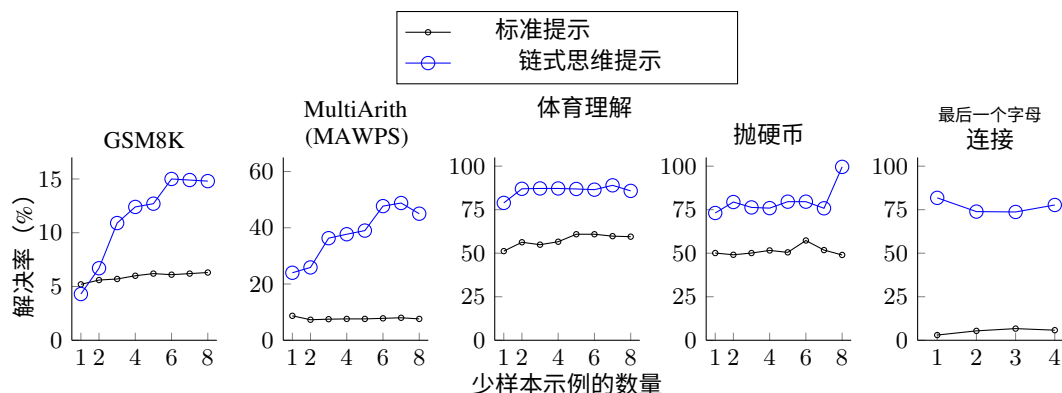


图11：链式思维提示相对于标准提示的改进在不同数量的少样本示例的提示中似乎具有鲁棒性。

表12：本文中使用的数学问题基准摘要及示例。 $N$ ：评估示例的数量。

数据集	$N$ 示例问题
GSM8K	1,319 乔希决定尝试翻转一座房子。他以8万美元购买了一座房子，然后投入5万美元进行修缮。这使房子的价值增加了150%。他赚了多少钱？
SVAMP	1,000 每个DVD包装售价76美元。如果每个包装有25美元的折扣。购买每个包装需要支付多少钱？
ASDiv	2,096 艾伦比马林多六个球。马林有九个球。艾伦有多少个球？
AQuA	254 一辆汽车以匀速直线行驶，朝着一个垂直塔的底部。从汽车上观察到塔顶，在这个过程中，仰角从45度变化到60度需要10分钟。 这辆车还需要多长时间才能到达塔底？ 答案 选项：(a) $5\sqrt{3} + 1$ (b) $6\sqrt{3} + \sqrt{2}$ (c) $7\sqrt{3} - 1$ (d) $8\sqrt{3} - 2$ (e) 以上都不是
MAWPS：单操作	562 如果一个盒子里有7个瓶盖，琳达再放进去7个瓶盖，盒子里一共有多少个瓶盖？
MAWPS：单方程	508 本尼买了一瓶2美元的软饮料和5个糖果棒。他总共花了27美元。每个糖果棒的价格是多少？
MAWPS：加减	395 花瓶里有6朵玫瑰花。玛丽从花园里剪了一些玫瑰花。现在花瓶里有16朵玫瑰花。她剪了多少朵玫瑰花？
MAWPS：多算术	600 学校食堂订购了42个红苹果和7个绿苹果供学生午餐。但是，如果只有9个学生想要水果，食堂最后多了多少个？

## 附加细节

### 版本控制

**V5 → V6.** 修正图3中的小错误。

**V4 → V5.** 添加Codex和UL2的结果。对论文的写作和风格进行了一些小改动。

**V3 → V4.** 修正了图3中的拼写错误，并添加了一些引用。

**V2 → V3.** 添加了GPT-3的结果。为数学添加了SVAMP和AQuA评估数据集。为常识添加了SayCan评估。添加了扩展相关工作部分（附录C）。为常识和符号推理添加了消融实验（表7）。添加了常见问题解答部分（附录A）。在附录B中添加了原始结果。

**V1 → V2.** 在LaMDA中添加了PaLM的结果（V1只有LaMDA）。

### E.1 可复现性声明

由于我们的结果使用了两组不公开的大型语言模型，我们采取以下措施来促进可复现性。首先，在附录G的表20-表27中提供了所有任务的确切输入提示（并强调我们没有进行任何微调，只对现成的语言模型应用提示）。其次，我们使用公开可用的GPT-3 API进行了四个模型规模的实验（text-ada-001, text-babbage-001, text-curie-001, text-davinci-002）。最后，我们将LaMDA 137B的确切输入、目标和预测结果作为一个zip文件提供在补充材料中。

### E.2 计算资源

对于我们评估的所有三个语言模型，我们只进行了基于提示的推理。本论文没有进行微调。对于LaMDA 137B的推理，我们使用TPU v3（8x8配置，64个芯片/128个核心），对于PaLM 540B的推理，我们使用TPU v4（4x4x12配置，192个芯片/384个核心）。GPT-3实验是使用公共API进行的。<sup>5</sup>

### E.3 数据集详细信息和许可证

我们列出了本论文中使用的所有算术和常识数据集的详细信息和许可证。符号推理数据集是根据第4节中的描述进行合成的。

#### 算术推理

- 数学问题存储库（Koncel-Kedziorski等，2016年）：AddSub（Hosseini等，2014年）：<https://www.cs.washington.edu/nlp/arithmetric>；MultiArith（Roy和Roth，2015年），许可证：CC BY 4.0。
- ASDiv（Miao等，2020）：<https://github.com/chaochun/nlu-asdiv-dataset>。
- AQuA（Ling等，2017）：<https://github.com/deepmind/AQuA>，许可证：<https://github.com/deepmind/AQuA/blob/master/LICENSE>。
- GSM8K（Cobbe等，2021）：<https://github.com/openai/grade-school-math>，MIT许可证：<https://github.com/openai/grade-school-math/blob/master/LICENSE>。
- SVAMP（Patel等，2021）：<https://github.com/arkilpatel/SVAMP>，MIT许可证：<https://github.com/arkilpatel/SVAMP/blob/main/LICENSE>。

#### 常识推理

- CSQA（Talmor等，2019）：<https://www.tau-nlp.org/commonsenseqa>, <https://github.com/jonathanherzig/commonsenseqa>。

<sup>5</sup><https://beta.openai.com/docs/api-reference/making-requests>

- StrategyQA (Geva等, 2021年) : 我们使用开放域设置 (仅问题集)  
来自BIG-bench合作 (2021年) : [https://github.com/google/BIG-bench/tree/main/bigbench/benchmark\\_tasks/strategyqa](https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/strategyqa)。原始数据集来自  
<https://github.com/eladsegal/strategyqa>, MIT许可证: <https://github.com/eladsegal/strategyqa/blob/main/LICENSE>。
- BIG-Bench的日期理解和体育理解 (BIG-bench合作,  
2021年) : Apache许可证v.2: <https://github.com/google/BIG-bench/blob/main/LICENSE>。
- SayCan (Ahn等, 2022年) : SayCan数据集可以在<https://say-can.github.io/>下以CC BY 4.0许可证访问。

## 附录F：输入/输出示例

表格13：LaMDA 137B在字母连接任务中产生的正确和错误思维链示例。

问题：取“Waldo Schmidt”中单词的最后一个字母并将它们连接起来。
模型回答（正确）：“Waldo”的最后一个字母是“o”。“Schmidt”的最后一个字母是“t”。将它们连接起来得到“ot”。所以答案是ot。✓
问题：取“Daniel Friedman”中单词的最后一个字母并将它们连接起来。
模型回答（错误）：“Daniel”的最后一个字母是“l”。“Friedman”的最后一个字母是“m”。将它们连接起来得到“lm”。所以答案是lm。✗

表格14：LaMDA 137B在抛硬币任务中产生的正确和错误思维链示例。

问题：一枚硬币正面朝上。安德里翻转了硬币。奥德里没有翻转硬币。硬币还是正面朝上吗？
模型答案（正确）：硬币被安德里翻转了。所以硬币被翻转了1次，是一个奇数次。硬币开始时是正面朝上的，所以经过奇数次翻转后，它会变成反面朝上。所以答案是否定的。✓
问题：一枚硬币正面朝上。克里斯蒂安没有翻转硬币。达拉斯没有翻转硬币。硬币还是正面朝上吗？
模型答案（错误）：硬币被克里斯蒂安翻转了。所以硬币被翻转了1次，是一个奇数次。硬币开始时是正面朝上的，所以经过奇数次翻转后，它会变成反面朝上。所以答案是否定的。✗

表15：LaMDA 137B在CSQA上产生的正确和错误思维链的示例。

问题：一个人开始工作时，他们在建造什么？
答案选择：(a) 时间 (b) 完成 (c) 工作 (d) 动力 (e) 疲劳
模型答案（正确）：答案必须是建造的东西。在上述选择中，只有动力是建造的。所以答案是 (d)。✓
问题：你可能在哪里找到汉堡包？
答案选择：(a) 快餐店 (b) 比萨 (c) 碎牛肉 (d) 嘴巴 (e) 牛尸体
模型答案（正确）：答案必须是能找到汉堡包的地方。在上述选择中，只有快餐店供应汉堡包。所以答案是 (a)。✓
问题：除了水和营养，你的狗还需要什么？
答案选择：(a) 骨头 (b) 魅力 (c) 抚摸 (d) 大量关注 (e) 散步
模型答案（错误）：答案必须是狗需要的东西。在上述选择中，只有骨头是狗需要的东西。所以答案是(a)。✗(正确答案是(d)。)
问题：当你在接待区等候时，你还在等待什么？
答案选择：(a) 汽车旅馆 (b) 椅子 (c) 医院 (d) 人们 (e) 酒店
模型答案（错误）：答案必须是在接待区与你一起等待的东西。在上述选择中，只有人们在接待区与你一起等待。所以答案是(e)。✗(正确答案是(d)。)

表16: LaMDA 137B在StrategyQA上产生的正确和错误思维链示例。

问题:《驱魔人》会刺激边缘系统吗?
模型回答(正确):《驱魔人》是一部恐怖电影。恐怖电影很吓人。边缘系统参与恐惧。因此,《驱魔人》会刺激边缘系统。所以答案是是。✓
问题:杰克逊·波洛克受到莱昂纳多·达·芬奇的训练吗?
模型回答(正确):莱昂纳多·达·芬奇生活在15世纪。杰克逊·波洛克生活在20世纪。因此,杰克逊·波洛克不可能接受莱昂纳多·达·芬奇的训练。所以答案是否定的。✓
问题:哈利·波特可以在韩亚航空公司预订航班吗?
模型回答(错误):哈利·波特是一个虚构的角色。因此,哈利·波特可以任何事情。所以答案是是的。✗(正确答案是“不是”)
问题:一个诡辩家会使用重剑吗?
模型回答(错误):诡辩家是一位擅长说服艺术的人。重剑是一种剑的类型。因此,诡辩家可以使用重剑。所以答案是是的。✗(正确答案是“不是”)

表17: LaMDA 137B在日期理解上产生的正确和错误思维链的示例。

问题:对于简来说,1992年5月6日就像昨天,但实际上已经过去了十年。一个月前的日期是多少(MM/DD/YYYY)?
模型回答(正确):1992年5月6日是十年前,所以今天是2002年5月6日。所以一个月前是4月6日。所以答案是04/06/2002。✓
问题:这是1899年的最后一天。24小时后的日期是多少(以MM/DD/YYYY格式)?
模型答案(正确):今天是12/31/1899。24小时后将是01/01/1900。所以答案是01/01/1900。✓
问题:简安排了明天(周二,7/9/1972)与5个人的3个约会。今天往前推一周的日期是多少(以MM/DD/YYYY格式)?
模型答案(错误):明天是7/9/1972。今天往前推一周是7/2/1972。所以答案是7/2/1972。✗(正确答案是07/01/1972。)
问题:今天是2020年的回文日期,因为日期的MMDDYYYY格式正反读都一样。明天的日期是多少(以MM/DD/YYYY格式)?
MODEL ANSWER (错误):今天是2020年02月29日。明天将是2020年03月01日。所以答案是03/01/2020。✗(正确答案是“02/03/2020”。)

表18: LaMDA 137B在体育理解方面产生的正确和错误思维链示例。

问题:以下句子是否合理?“Joao Moutinho在三垒出局。”
MODEL ANSWER (正确):Joao Moutinho是一名足球运动员。在三垒出局是棒球的一部分,而不是足球。所以答案是否定的。✓
问题:以下句子是否合理?“Malcolm Brogdon在NBA总决赛中进行了欧洲步进攻。”
MODEL ANSWER (正确):Malcolm Brogdon是一名篮球运动员。欧洲步进攻是篮球的一部分。所以答案是肯定的。✓
问题:以下句子是否合理?“德里克·怀特反手射门。”
模型答案(错误):德里克·怀特是一名篮球运动员。反手射门是篮球的一部分。✗(正确答案是“不”,反手射门是曲棍球的一部分。)
问题:以下句子是否合理?“纳泽姆·卡德里在家里出局。”
模型答案(错误):纳泽姆·卡德里是一名曲棍球运动员。在家里出局是曲棍球的一部分。所以答案是是。✗(正确答案是“不”,在家里出局是棒球的一部分。)

表格19：PaLM 540B在SayCan机器人规划任务中产生的正确和错误思维链示例。

问题:我把可乐洒在桌子上，你能把它扔掉然后给我拿点东西帮忙清理吗？
<b>MODEL ANSWER (正确):</b> 解释：用户把可乐洒在桌子上。我会扔掉可乐，然后给用户拿一个海绵。计划：找到(可乐)，拿起(可乐)，找到(垃圾)，放下(可乐)，找到(海绵)，拿起(海绵)，找到(桌子)，放下(海绵)。✓
问题：你能帮我把这个苹果堆肥吗？
<b>MODEL ANSWER (正确):</b> 解释：用户希望我把苹果堆肥。我会把苹果移到堆肥桶里。计划：找到(苹果)，拿起(苹果)，找到(垃圾)，放下(苹果)。✓
问题：你能给我拿一杯不是可乐的饮料吗？
<b>MODEL ANSWER(错误):</b> 解释：用户要求我拿一杯不是可乐的饮料，我会拿一瓶雪碧。计划：找到(雪碧)，拿起(雪碧)，找到(用户)，放下(雪碧)。✗(可行的答案包括找到柠檬苏打水、百事可乐、葡萄柚苏打水或水，因为根据表28中的定义，场景中没有雪碧。)
问题:你能给我带点咸的东西吗？
模型回答(错误):解释：用户想要咸的东西。在物品列表中有几个选项，所以我会给用户带一杯柠檬苏打。计划：找到(柠檬苏打)，拿起(柠檬苏打)，找到(用户)，给(柠檬苏打)。✗(可行的答案包括找到薯片。模型错误地将柠檬苏打误认为是咸的。)



## 附录G：完整提示

表格20：数学问题完整链式思维提示的少样本示例。这组示例用于除AQuA以外的所有数学问题数据集。

---

### 数学问题提示

问题：林中有15棵树。林地工人今天将在林地种树。完成后，将有21棵树。林地工人今天种了多少棵树？

A:最初有15棵树。然后种了一些树后，有21棵树。所以必须有 $21 - 15 = 6$ 。答案是6。

Q:停车场里有3辆车，又有2辆车到达，停车场里有多少辆车？

A:最初有3辆车。又有2辆车到达。 $3 + 2 = 5$ 。答案是5。

Q:Leah有32块巧克力，她的妹妹有42块。如果他们吃了35块，他们总共还剩下多少块？

A:最初，Leah有32块巧克力。她的妹妹有42块。所以他们总共有 $32 + 42 = 74$ 块。吃了35块后，他们还剩下 $74 - 35 = 39$ 块。答案是39。

Q:Jason有20个棒棒糖。他给了Denny一些棒棒糖。现在Jason还剩下12个棒棒糖。Jason给了Denny多少个棒棒糖？

A:杰森开始有20个棒棒糖。然后他给了丹尼一些，所以他给了丹尼 $20 - 12 = 8$ 个。答案是8。

Q:肖恩有五个玩具。圣诞节时，他从他爸爸和妈妈那里各得到两个玩具。他现在有多少玩具？

A:肖恩开始有5个玩具。如果他从他爸爸和妈妈那里各得到2个玩具，那就是多了4个玩具。 $5 + 4 = 9$ 。答案是9。

Q:服务器房间里有九台电脑。从星期一到星期四，每天增加五台电脑。现在服务器房间里有多少台电脑？

A:最初有9台电脑。连续四天，每天增加5台电脑。所以一共增加了 $5 * 4 = 20$ 台电脑。 $9 + 20$ 等于29。答案是29。

Q:迈克尔有58个高尔夫球。星期二，他丢失了23个高尔夫球。星期三，他又丢失了2个。星期三结束时他还剩下多少个高尔夫球？

A:迈克尔开始有58个高尔夫球。星期二丢失了23个后，他剩下 $58 - 23 = 35$ 个。再丢失2个后，他剩下 $35 - 2 = 33$ 个高尔夫球。答案是33。

Q:奥利维亚有23美元。她买了五个3美元的百吉饼。她还剩下多少钱？

A:奥利维亚有23美元。五个百吉饼每个3美元，总共是 $5 * 3 = 15$ 美元。所以她还剩下 $23 - 15$ 美元。 $23 - 15$ 等于8。答案是8。

---

表21：AQuA完整链式思维提示的少样本示例。

---

为代数问题提供提示
问题：约翰发现15个数字的平均值是40。如果每个数字加上10，那么数字的平均值是多少？
答案选项：(a) 50 (b) 45 (c) 65 (d) 78 (e) 64
答案：如果每个数字加上10，那么数字的平均值也会增加10。所以新的平均值将是50。答案是(a)。
问题：如果 $a/b = 3/4$ ，且 $8a + 5b = 22$ ，则求 a 的值。
答案选项：(a) $1/2$ (b) $3/2$ (c) $5/2$ (d) $4/2$ (e) $7/2$
答案：如果 $a/b = 3/4$ ，则 $b = 4a/3$ 。所以 $8a + 5(4a/3) = 22$ 。这简化为 $8a + 20a/3 = 22$ ，意味着 $44a/3 = 22$ 。所以 a 等于 $3/2$ 。答案是(b)。
问题：一个人以20公里/小时的速度旅行，用2.5小时到达目的地，那么求距离是多少？
选项：(a) 53公里 (b) 55公里 (c) 52公里 (d) 60公里 (e) 50公里
答案：这个人旅行的距离将会是20公里/小时 * 2.5小时 = 50公里。答案是(e)。
问题：打印从1到500需要多少次按键？
选项：(a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788
答案：从1到9有9个一位数。从10到99有90个两位数。从100到500有401个三位数。 $9 + 90(2) + 401(3) = 1392$ 。答案是(b)。

---

表22：全链式思维提示的少样本示例，用于最后一个字母连接任务。

---

<b>P</b> 最后一个字母连接的提示
问题：取出“Elon Musk”中单词的最后一个字母并将它们连接起来。
回答：“Elon”的最后一个字母是“n”。“Musk”的最后一个字母是“k”。将它们连接起来得到“nk”。答案是nk。
问题：取出“Larry Page”中单词的最后一个字母并将它们连接起来。
回答：“Larry”的最后一个字母是“y”。“Page”的最后一个字母是“e”。将它们连接起来得到“ye”。答案是ye。
问题：取出“Sergey Brin”中单词的最后一个字母并将它们连接起来。
回答：“Sergey”的最后一个字母是“y”。“Brin”的最后一个字母是“n”。将它们连接起来得到“yn”。答案是yn。
<b>Q</b> :取“比尔·盖茨”中每个单词的最后一个字母并连接起来。
<b>A</b> :“比尔”的最后一个字母是“l”。“盖茨”的最后一个字母是“s”。将它们连接起来得到“ls”。答案是ls。

---

表23：硬币翻转任务的全链式思维提示的少样本示例。

---

**P提示为 C硬币 F翻转**

**Q:**一枚硬币正面朝上。卡翻转了硬币。雪莉翻转了硬币。硬币还是正面朝上吗？

**A:**硬币被卡和雪莉翻转了。所以硬币被翻转了2次，这是一个偶数。硬币一开始是正面朝上的，所以经过偶数次翻转后，它仍然是正面朝上的。所以答案是是。

**Q:**一枚硬币正面朝上。杰米翻转了硬币。特蕾莎翻转了硬币。硬币还是正面朝上吗？

**A:**硬币由Jamey和Teressa抛掷。所以硬币被抛掷了2次，这是一个偶数。硬币开始时是正面朝上的，所以经过偶数次抛掷后，它仍然是正面朝上的。所以答案是是。**Q:**一个硬币是正面朝上的。May

belle抛掷了硬币。Shalonda不抛掷硬币。硬币还是正面朝上吗？

**A:**硬币由Maybelle抛掷。所以硬币被抛掷了1次，这是一个奇数。硬币开始时是正面朝上的，所以经过奇数次抛掷后，它会变成反面朝上。所以答案是不。

**Q:**一个硬币是正面朝上的。Millicent不抛掷硬币。Conception抛掷了硬币。硬币还是正面朝上吗？

**A:**这个硬币是由康塞普修翻转的。所以这个硬币被翻转了1次，是一个奇数。这个硬币开始是正面朝上的，所以经过奇数次翻转后，它将变成反面朝上。所以答案是否定的。

**Q:**一个硬币是正面朝上的。萨尔翻转了这个硬币。雷蒙德没有翻转这个硬币。这个硬币还是正面朝上吗？

**A:**这个硬币是由萨尔翻转的。所以这个硬币被翻转了1次，是一个奇数。这个硬币开始是正面朝上的，所以经过奇数次翻转后，它将变成反面朝上。所以答案是否定的。

**Q:**一个硬币是正面朝上的。康塞普修翻转了这个硬币。克里斯蒂安没有翻转这个硬币。这个硬币还是正面朝上吗？

**A:**这个硬币是由康塞普修翻转的。所以这个硬币被翻转了1次，是一个奇数。这个硬币开始是正面朝上的，所以经过奇数次翻转后，它将变成反面朝上。所以答案是否定的。

**Q:**一个硬币是正面朝上的。英嘉没有翻转这个硬币。埃莱诺没有翻转这个硬币。这个硬币还是正面朝上吗？

**A:**没有人翻转了硬币。所以硬币被翻转了0次。硬币一开始是正面朝上的，且没有被翻转，所以它仍然是正面朝上的。所以答案是是。

**Q:**一个硬币是正面朝上的。瑞安翻转了硬币。肖恩达翻转了硬币。硬币还是正面朝上吗？

**A:**硬币被瑞安和肖恩达翻转了。所以硬币被翻转了2次，这是一个偶数次数。硬币一开始是正面朝上的，所以经过偶数次翻转后，它仍然是正面朝上的。所以答案是是。

---

表格24：CSQA完整链式思维提示的少样本示例。表格中省略了答案选项之间的换行符以节省空间。

---

**P CSQA的提示**

**Q:**人们用什么来吸收钢笔中多余的墨水？ 答案选项：（a）衬衫口袋（b）书法家的手（c）墨水瓶（d）办公桌抽屉（e）吸墨纸

**A:**答案必须是能吸收墨水的物品。在上面的选择中，只有吸墨纸可以用来吸收墨水。所以答案是(e)。

**Q:**什么家庭娱乐设备需要电缆？

答案选择：（a）无线电台（b）变电站（c）电视（d）橱柜

**A:**答案必须需要电缆。在上面的选择中，只有电视需要电缆。所以答案是(c)。

**Q:**狐狸从城市走进森林，它在找什么？ 答案选择：（a）漂亮的花朵（b）

鸡舍（c）自然栖息地（d）故事书

**A:**答案必须是森林中的某物。在上面的选择中，只有鸡舍在森林中。所以答案是(b)。

**Q:**萨米想去人群聚集的地方。他可能会去哪里？ 答案选择：（a）人口稠密地区

（b）赛车场（c）沙漠（d）公寓（e）路障

**A:**答案必须是一个很多人很多的地方。在上述选择中，只有人口稠密的地区有很多人。所以答案是（a）。

**Q:**在结账前，你把葡萄放在哪里？ 答案选择：（a）嘴巴（b）购物车（c）超市（d）水果篮（e）水果市场

**A:**答案应该是在结账前放置杂货的地方。在上述选择中，购物车最合理，可以放置杂货。所以答案是（b）。

**Q:**谷歌地图和其他高速公路和街道GPS服务取代了什么？ 答案选择：（a）美国（b）墨西哥（c）乡村（d）地图册

**A:**答案必须是曾经做过谷歌地图和GPS服务所做的事情，即提供方向。在上述选择中，只有地图册用于提供方向。所以答案是（d）。

**Q:**在离婚之前，那个一直在做所有工作的妻子是什么感觉？ 答案选择：（a）更加困难（b）

痛苦（c）苦涩（d）泪水（e）伤心

**A:**答案应该是一个离婚的人，一直在做所有工作的感觉。在上面的选择中，最接近的感觉是苦涩。所以答案是(c)。

---

表25：StrategyQA的全链式思维提示的少样本示例。

<b>P策略QA的提示</b>
<b>Q:</b> 仓鼠为其他动物提供食物吗？ <b>A:</b> 仓鼠是猎物动物。猎物是捕食者的食物。因此，仓鼠为一些动物提供食物。所以答案是是。 <b>Q:</b> 布鲁克·希尔兹能在宾夕法尼亚大学成功吗？ <b>A:</b> 布鲁克·希尔兹去了普林斯顿大学。普林斯顿大学在学术上与宾夕法尼亚大学差不多严格。因此，布鲁克·希尔兹也可以在宾夕法尼亚大学成功。所以答案是是。 <b>Q:</b> 是或否：氢的原子序数的平方是否超过辣妹组合的人数？ <b>A:</b> 氢的原子序数是1。1的平方是1。辣妹组合有5个成员。因此，氢的原子序数的平方小于5。所以答案是否定的。 <b>Q:</b> 在一些大学毕业典礼上看到霜冻是常见的吗？ <b>A:</b> 大学毕业典礼可以在12月、5月和6月举行。12月是冬季，所以可能会有霜冻。因此，在一些毕业典礼上可能会有霜冻。所以答案是肯定的。 <b>Q:</b> 是或否：在越南战争（1945-46）期间，一只羊驼能生两次吗？ <b>A:</b> 越南战争持续了6个月。羊驼的妊娠期为11个月，超过了6个月。因此，在越南战争期间，一只羊驼不能生两次。所以答案是否定的。 <b>Q:</b> 是或否：一个梨会在水中沉没吗？ <b>A:</b> 一个梨的密度约为 0.6克/立方厘米 <sup>3</sup> ，低于水的密度。比水密度小的物体会浮起来。因此，一个梨会浮起来。所以答案是否定的。

表26：用于完整思维链提示的少样本示例，用于日期理解。

<b>P日期理解的提示</b>
<b>Q:</b> 2015年还有36小时。从今天起一周后的日期是多少（MM/DD/YYYY格式）？ <b>A:</b> 如果2015年还有36小时，那么它还有2天。01/01/2015的前2天是12/30/2014，所以今天是12/30/2014。所以从今天起一周后的日期是01/05/2015。所以答案是01/05/2015。 <b>Q:</b> 2019年的第一天是星期二，今天是2019年的第一个星期一。今天的日期是多少MM/DD/YYYY？ <b>A:</b> 如果2019年的第一天是星期二，那么01/01/2019是星期二。今天是第一个星期一，再过六天。所以今天是01/07/2019。所以答案是01/07/2019。 <b>Q:</b> 音乐会原定于06/01/1943举行，但因为延迟了一天而改到了今天。10天前的日期是多少MM/DD/YYYY？ <b>A:</b> 06/01/1943的后一天是06/02/1943，所以今天是06/02/1943。今天的前10天是05/23/1943。所以答案是05/23/1943。 <b>Q:</b> 今天是1969年4月19日。24小时后的日期是多少（以MM/DD/YYYY格式）？ <b>A:</b> 今天是1969年04月19日。24小时后是今天的第二天，也就是1969年04月20日。所以答案是04/20/1969。 <b>Q:</b> 简以为今天是2002年3月11日，但实际上今天是3月12日，也就是晚了1天。24小时后的日期是多少（以MM/DD/YYYY格式）？ <b>A:</b> 今天是2002年03月12日。所以24小时后的日期将是03/13/2002。所以答案是03/13/2002。 <b>Q:</b> 简在2001年的二月最后一天出生。今天是她16岁生日。昨天的日期是多少（以MM/DD/YYYY格式）？ <b>A:</b> 二月的最后一天是28日，所以简的生日是02/28/2001。今天是她16岁生日，所以今天是02/28/2017。所以昨天是02/27/2017。所以答案是02/27/2017。

表格27：用于体育理解的全链式思维提示的少样本示例。

---

**P**用于体育理解的提示

**Q:**以下句子是否合理？“Kyle Palmieri被判犯规。”

**A:**Kyle Palmieri是一名曲棍球运动员。被判犯规是曲棍球的一部分。所以答案是是。**Q:**以下句子是否合理？“Joao Moutinho在NFC冠军赛中接到了屏幕传球。”**A:**Joao Moutinho是一名足球运动员。NFC冠军赛是美式足球的一部分，而不是足球。所以答案是否定的。

**Q:**以下句子是否合理？“卡森·温兹进行了挡拆和滚筐。”

**A:**卡森·温兹是一名美式足球运动员。挡拆和滚筐是篮球的一部分，而不是足球。所以答案是不合理的。

**Q:**以下句子是否合理？“约纳斯·瓦兰丘纳斯在蜂鸣器响起前投篮得分。”

**A:**乔纳斯·瓦兰丘纳斯是一名篮球运动员。在篮球比赛中，击败蜂鸣器是其中的一部分。所以答案是肯定的。

**Q:**以下句子是否合理？“贾梅尔·默里罚球全中。”

**A:**贾马尔·默里是一名篮球运动员。罚球全中是篮球比赛的一部分。所以答案是肯定的。

**Q:**以下句子是否合理？“山姆·达诺尔德传球。”

**A:**山姆·达诺尔德是一名美式足球运动员。传球是冰球的一部分，而不是美式足球。所以答案是否定的。

**Q:**以下句子是否合理？“德雷蒙德·格林投掷Touchdown。”

**A:**德雷蒙德·格林是一名篮球运动员。投掷Touchdown是橄榄球的一部分，而不是篮球。所以答案是否定的。

**Q:**以下句子是否合理？“马尔科姆·布罗格登将球投入篮筐。”

**A:**马尔科姆·布罗格登是一名篮球运动员。将球投入篮筐是篮球比赛的一部分。所以答案是肯定的。

---



表格 28: SayCan 机器人规划任务的全链式思维提示的少样本示例。

---

**P为 SAYCAN ROBOT PLANNING 提示**

位置 = [柜台, 桌子, 用户, 垃圾桶, 碗]。

物品 = [7up, 苹果, 烧烤薯片, 茶, 多谷薯片, 可乐, 青柠苏打水, 辣椒薯片, 米饼, 橙子, 葡萄柚苏打水, 百事可乐, 红牛, 能量棒, 海绵, 水]。

机器人可以用pick(object)拿起物品, 用put(object)放下物品, 还可以用find()找到物品或位置。机器人只能理解明确列出的位置和物品。

人类: 我饿了, 你能给我带些薯片来吗。

解释: 用户饿了, 并要求带些薯片。有几种薯片可供选择, 我会给用户带烧烤薯片。

计划: 1. 找到(烧烤薯片), 2. 拿起(烧烤薯片), 3. 找到(用户), 4. 放下(烧烤薯片), 5. 完成()。

人类: 你会如何将葡萄柚饮料从桌子上移到柜台上?

解释: 用户要求我将葡萄柚饮料移到柜台上。

计划: 1. 找到(葡萄柚汽水), 2. 拿起(葡萄柚汽水), 3. 找到(柜台), 4. 放下(葡萄柚汽水), 5. 完成()。

人类: 你会如何给我带一些零食?

解释: 用户要求零食, 我会选择两样东西带来。我会带来辣椒薯片和一个苹果。

计划: 1. 找到(辣椒薯片), 2. 拿起(辣椒薯片), 3. 找到(用户), 4. 放下(辣椒薯片), 5. 找到(苹果), 6. 拿起(苹果), 7. 找到(用户), 8. 放下(苹果), 9. 完成()。

人类: 你怎样给我带来一些不是水果的食物?

解释: 用户要求一种不是水果的食物, 我会给他们带来一块能量棒。

计划: 1. 寻找(能量棒), 2. 拿起(能量棒), 3. 寻找(用户), 4. 放下(能量棒), 5. 完成()。

人类: 你怎样把米饼放进碗里, 然后把茶移到桌子上?

解释: 用户要求我完成两个任务, 我会先完成一个, 然后再完成另一个。

计划: 1. 寻找(米饼), 2. 拿起(米饼), 3. 寻找(碗), 4. 放下(米饼), 5. 寻找(茶), 6. 拿起(茶), 7. 寻找(桌子), 8. 放下(茶), 9. 完成()。

人类: 你会怎样扔掉一罐红牛?

解释: 用户要求我扔掉红牛, 我会把它放进垃圾桶。

计划: 1. 寻找(红牛), 2. 拿起(红牛), 3. 寻找(垃圾桶), 4. 放入(红牛), 5. 完成()。

人类: 给我拿一杯饮料。

解释: 用户要求一杯饮料, 有很多选择。我会给他们拿一杯水。

计划: 1. 寻找(水), 2. 拿起(水), 3. 寻找(用户), 4. 放入(水), 5. 完成()。

---

## 附录H：数学问题的备用注释者

表29：数学问题完整链式思维提示的少样本示例。这些示例与表20中的示例相同，只是思维链由不同的注释者（“注释者B”而不是“注释者A”）编写。注释者是共同作者，并熟悉链式思维提示的目标。

---

### 数学问题提示

问题：林中有15棵树。林地工人今天将在林地种树。完成后，将有21棵树。林地工人今天种了多少棵树？

A:现在在21棵树，一开始有15棵树，所以工人种了 $21 - 15 = 6$ 棵树。答案是6。

Q:停车场里有3辆车，又有2辆车到达，停车场里有多少辆车？

A:一开始有3辆车，又来了2辆，所以现在应该有 $3 + 2 = 5$ 辆车。答案是5。

Q:Leah有32块巧克力，她的妹妹有42块。如果他们吃了35块，他们总共还剩下多少块？

A:莉娅有32块巧克力，她姐姐有42块，总共他们有 $32 + 42 = 74$ 块巧克力。吃掉35块巧克力后，现在还剩下 $74 - 35 = 39$ 块巧克力。答案是39。

Q:Jason有20个棒棒糖。他给了Denny一些棒棒糖。现在Jason还剩下12个棒棒糖。Jason给了Denny多少个棒棒糖？

A:杰森一开始有20个棒棒糖，但现在只剩下12个，所以他给了丹尼8个棒棒糖。答案是8。

Q:肖恩有五个玩具。圣诞节时，他从他爸爸和妈妈那里各得到两个玩具。他现在有多少玩具？

A:肖恩从他爸爸和妈妈那里每人得到2个玩具，所以他得到了 $2 * 2 = 4$ 个，现在他会有 $5 + 4 = 9$ 个玩具。答案是9。

Q:服务器房间里有九台电脑。从星期一到星期四，每天增加五台电脑。现在服务器房间里有多少台电脑？

A:从星期一到星期四安装了5台电脑，所以总共安装了 $5 * 4 = 20$ 台电脑。一开始有9台电脑，所以现在有 $20 + 9 = 29$ 台电脑。答案是29。

Q:迈克尔有58个高尔夫球。星期二，他丢失了23个高尔夫球。星期三，他又丢失了2个。星期三结束时他还剩下多少个高尔夫球？

A:迈克尔一开始有58个高尔夫球，丢失了23个，所以他还剩下 $58 - 23 = 35$ 个。再丢失2个后，他现在还剩下 $35 - 2 = 33$ 个球。答案是33。

Q:奥利维亚有23美元。她买了五个3美元的百吉饼。她还剩下多少钱？

A:5个百吉饼每个3美元应该花费 $5 * 3 = 15$ 美元。奥利维亚一开始有23美元，所以现在她还剩下 $23 - 15 = 8$ 美元。答案是8。

---

表格30：数学问题的完整思维链示例。这些示例与表格20中的示例相同，只是思维链由不同的注释者（“注释者C”而不是“注释者A”）编写。

---

数学问题提示

问题：林中有15棵树。林地工人今天将在林地种树。完成后，将有21棵树。林地工人今天种了多少棵树？

A:我们开始有15棵树。后来我们有21棵树。差异必须是他们种植的树的数量。所以，他们必须种植了 $21 - 15 = 6$ 棵树。答案是6。

Q:停车场里有3辆车，又有2辆车到达，停车场里有多少辆车？

A:停车场已经有3辆车了。又来了2辆。现在有 $3 + 2 = 5$ 辆车。答案是5。

Q:Leah有32块巧克力，她的妹妹有42块。如果他们吃了35块，他们总共还剩下多少块？

A:莉娅有32块巧克力，莉娅的姐姐有42块。这意味着最初有 $32 + 42 = 74$ 块巧克力。已经吃掉了35块。所以总共他们还有 $74 - 35 = 39$ 块巧克力。答案是39。

Q:Jason有20个棒棒糖。他给了Denny一些棒棒糖。现在Jason还剩下12个棒棒糖。Jason给了Denny多少个棒棒糖？

A:杰森有20个棒棒糖。因为他现在只剩下12个，所以他一定把剩下的给了丹尼。他给丹尼的棒棒糖数量一定是 $20 - 12 = 8$ 个棒棒糖。答案是8。

Q:肖恩有五个玩具。圣诞节时，他从他爸爸和妈妈那里各得到两个玩具。他现在有多少玩具？

A:他有5个玩具。他从妈妈那里得到2个，所以之后他有 $5 + 2 = 7$ 个玩具。然后他从爸爸那里再得到2个，所以总共他有 $7 + 2 = 9$ 个玩具。答案是9。

Q:服务器房间里有九台电脑。从星期一到星期四，每天增加五台电脑。现在服务器房间里有多少台电脑？

A:从星期一到星期四有4天。每天增加5台电脑。这意味着总共增加了 $4 * 5 = 20$ 台电脑。一开始有9台电脑，所以现在有 $9 + 20 = 29$ 台电脑。答案是29。

Q:迈克尔有58个高尔夫球。星期二，他丢失了23个高尔夫球。星期三，他又丢失了2个。星期三结束时他还剩下多少个高尔夫球？

A:迈克尔最初有58个球。星期二他丢失了23个，所以之后他有 $58 - 23 = 35$ 个球。星期三他又丢失了2个，所以现在他有 $35 - 2 = 33$ 个球。答案是33。

Q:奥利维亚有23美元。她买了五个3美元的百吉饼。她还剩下多少钱？

A:她买了5个百吉饼，每个3美元。这意味着她在百吉饼上花了 $5 * 3 = 15$ 美元。一开始她有23美元，所以现在她有 $23 - 15 = 8$ 美元。答案是8。

---