

大型语言模型作为零样本推理器

小岛武史
东京大学

t.kojima@weblab.t.u-tokyo.ac.jp

顾世祥
谷歌研究, Brain团队

马切尔·里德
谷歌研究*

松尾豊
东京大学

岩泽祐介
东京大学

摘要

预训练的大型语言模型 (LLMs) 广泛应用于自然语言处理 (NLP) 的许多子领域, 并且通常被认为是具有任务特定示例的出色的少样本学习器。值得注意的是, 思维链 (CoT) 提示是一种通过逐步回答示例来引出复杂的多步推理的最新技术, 在算术和符号推理方面取得了最先进的性能, 这些任务不遵循 LLMs 的标准缩放规律。虽然这些成功通常归因于 LLMs 的少样本学习能力, 但我们表明, 只需在每个答案之前添加“让我们逐步思考”即可使 LLMs 成为不错的零样本推理器。实验结果表明, 我们的零样本-CoT 在相同的单一提示模板上, 在各种基准推理任务上显著优于零样本 LLM 性能, 包括算术 (MultiArith, GSM8K, AQUA-RAT, SVAMP), 符号推理 (最后一个字母, 抛硬币), 以及其他逻辑推理任务 (日期理解, 跟踪洗牌对象), 而无需任何手工制作的少样本示例, 例如使用大规模 InstructGPT 模型 (text-davinci-002) 将 MultiArith 的准确率从 17.7% 提高到 78.7%, 将 GSM8K 从 10.4% 提高到 40.7%, 以及使用另一个现成的大型模型 540B 参数 PaLM 也有类似程度的改进。这个单一提示的多样性跨越非常多样的推理任务, 暗示了 LLMs 未开发和未研究的基本零样本能力, 表明高层次、多任务的广泛认知能力可以通过简单的提示来提取。我们希望我们的工作不仅仅作为具有挑战性的推理基准的最小最强零样本基线, 而且还强调了在制作微调数据集或少样本示例之前, 仔细探索和分析隐藏在 LLMs 中的巨大零样本知识的重要性。

1 引言

扩大语言模型的规模是近年来自然语言处理 (NLP) 革命的关键因素 [Vaswani et al., 2017, Devlin et al., 2019, Raffel et al., 2020, Brown et al., 2020, Thoppilan et al., 2022, Rae et al., 2021, Chowdhery et al., 2022]。大型语言模型 (LLMs) 的成功通常归因于 (上下文中的) 少样本或零样本学习。它可以通过简单地将模型条件化于少量示例 (少样本) 或描述任务的指令 (零样本) 来解决各种任务。条件化语言模型的方法被称为“提示” [Liu et al., 2021b], 在 NLP 中, 手动设计提示 [Schick and Schütze, 2021, Reynolds and McDonell, 2021] 或自动设计提示 [Gao et al., 2021, Shin et al., 2020] 已成为热门话题。

*在东京大学期间完成的工作。

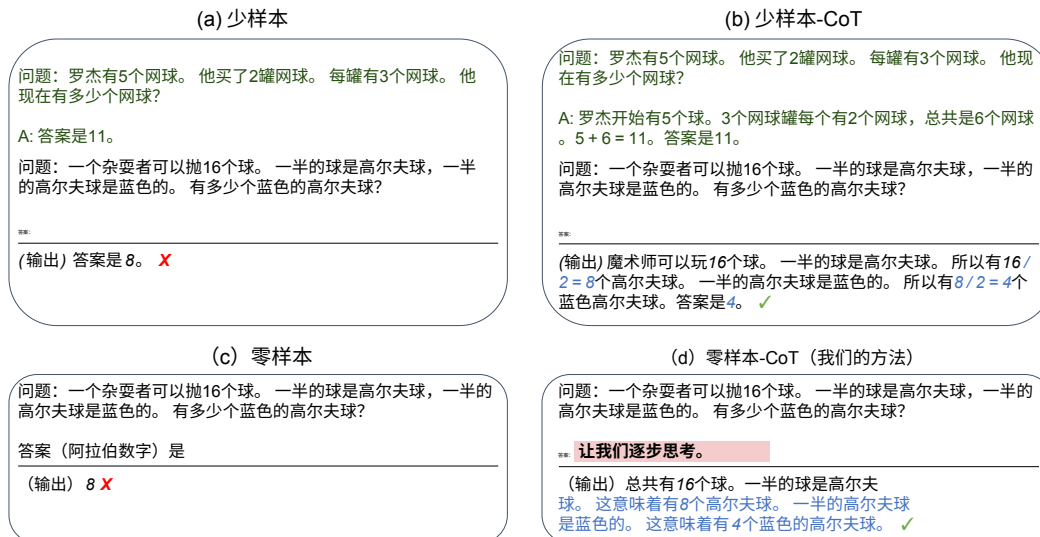


图1：GPT-3的示例输入和输出，包括(a)标准少样本([Brown et al., 2020])，(b)少样本-CoT([Wei et al., 2022])，(c)标准零样本和(d)我们的(零样本-CoT)。类似于Few-shot-CoT，零样本-CoT促进了多步推理（蓝色文本）并得出正确答案，在标准提示失败的情况下。与Few-shot-CoT不同，我们的方法不需要每个任务的逐步推理示例，只需在所有任务（算术、符号、常识和其他逻辑推理任务）中使用相同的提示“让我们逐步思考”。

与直觉和单步系统-1任务（Stanovich和West，2000）中LLMs的出色表现相比，需要慢速和多步推理（Rae等，2021）的系统-2任务对于具有100B或更多参数的语言模型来说是困难的，即使是具有任务特定的少样本或零样本提示的语言模型。为了解决这个缺点，Wei等人[2022]，Wang等人[2022]提出了思维链提示（CoT），它向LLMs提供逐步推理示例，而不是标准的问题和答案示例（见图1-a）。

这种思维链演示有助于模型生成将复杂推理分解为多个较简单步骤的推理路径。值得注意的是，使用CoT，推理性能更好地符合缩放定律，并且随着语言模型的规模增大而提高。例如，当与540B参数的PaLM模型[Chowdhery等，2022]结合使用时，思维链提示显著提高了在多个基准推理任务上的性能，例如GSM8K（17.9% → 58.1%）。

尽管CoT提示的成功[Wei等，2022]以及许多其他任务特定提示工作的成功[Gao等，2021，Schick和Schütze，2021，Liu等，2021b]通常归因于LLM对少样本学习的能力[Brown等，2020]，我们表明LLM通过添加一个简单的提示，即“让我们逐步思考”，在回答每个问题之前促进逐步思考，从而成为不错的零样本推理器（见图1）。尽管简单，我们的Zero-shot-CoT成功地以零样本的方式生成了一个合理的推理路径，并在标准零样本方法失败的问题中得出了正确答案。重要的是，我们的Zero-shot-CoT是多功能且与任务无关的，不像大多数先前的任务特定提示工程形式，例如示例（少样本）或模板（零样本）[Liu等，2021b]：它可以在各种推理任务中促进逐步回答，包括算术（MultiArith [Roy和Roth，2015]，GSM8K [Cobbe等，2021]，AQUA-RAT [Ling等，2017]和SVAMP [Patel等，2021]），符号推理（最后一个字母和抛硬币），常识推理（CommonSenseQA [Talmor等，2019]和Strategy QA [Geva等，2021]）以及其他逻辑推理任务（日期理解和从BIG-bench [Srivastava等，2022]跟踪洗牌的对象），而无需针对每个任务修改提示。

我们在表2中对Zero-shot-CoT与其他提示基线进行了实证评估。虽然我们的Zero-shot-CoT在精心设计和任务特定的逐步示例方面表现不佳，但与零样本基线相比，Zero-shot-CoT在得分上取得了巨大的增益，例如在MultiArith上从17.7%增加到78.7%，在GSM8K上从10.4%增加到40.7%，使用大规模的InstructGPT。

模型 (text-davinci-002)。我们还使用另一个现成的大型模型，540B参数的PaLM，对Zero-shot-CoT进行了评估，在MultiArith和GSM8K上显示出类似的改进幅度。重要的是，使用我们的单一固定提示，零样本LLMs具有与少样本CoT基线相当的良好扩展曲线。我们还表明，除了Few-shot-CoT需要人工设计多步推理提示外，如果提示示例问题类型和任务问题类型不匹配，它们的性能会下降，这表明对每个任务的提示设计非常敏感。相比之下，LLMs的这个单一提示在不同的推理任务中的多样性表明了未开发和未研究的零样本基本能力，例如更高级的广泛认知能力，如通用逻辑推理[Chollet, 2019]。虽然LLMs这个充满活力的领域始于出色的少样本学习者的前提[布朗等, 2020]，但我们希望我们的工作能够鼓励更多的研究，揭示这些模型内部隐藏的高级和多任务的零样本能力。

2 背景

我们简要回顾了构成本工作基础的两个核心预备概念：大型语言模型（LLMs）和提示，以及用于多步推理的思维链（CoT）提示。

大型语言模型和提示语言模型（LM）是一种旨在估计文本概率分布的模型。最近，通过更大的模型规模（从几百万个[Merity et al., 2016]到几亿个[Devlin et al., 2019]再到数千亿个[Brown et al., 2020]参数）和更大的数据（例如网络文本语料库[Gao et al., 2020]），预训练的大型语言模型（LLMs）在许多下游NLP任务上表现出极高的能力。除了经典的“预训练和微调”范式[Liu et al., 2021b]，规模扩展到100B+参数的模型表现出有利于少样本学习的特性[Brown et al., 2020]，通过上下文学习的方式，在文本或模板（称为提示）的指导下，强力引导生成输出所需任务的答案，从而开启了“预训练和提示”时代[Liu et al., 2021a]。在本工作中，我们将在少数任务示例上明确条件的提示称为少样本提示，而将仅基于模板的提示称为零样本提示。

思维链引发的多步算术和逻辑推理基准测试对大型语言模型的扩展规律提出了挑战[Rae等, 2021]。思维链（CoT）提示[Wei等, 2022]是少样本提示的一个实例，它通过修改少样本示例中的答案为逐步回答的方式提出了一个简单的解决方案，并且在这些困难的基准测试中取得了显著的性能提升，特别是与PaLM [Chowdhery等, 2022]等非常大的语言模型结合使用时。图1的顶部行显示了标准的少样本提示与（少样本）思维链提示的对比。值得注意的是，少样本学习被视为解决这些困难任务的一种方法，并且原始工作中甚至没有报告零样本基准性能[Wei等, 2022]。为了与我们的方法区分开来，我们在本文中将Wei等人的方法称为Few-shot-CoT。

3 零样本思维链

我们提出了Zero-shot-CoT，一种基于模板的零样本提示链式思维推理方法。它与原始的思维链提示[Wei et al., 2022]不同，因为它不需要逐步的少样本示例，与大多数先前的模板提示[Liu et al., 2021b]不同，它本质上是任务无关的，并且通过单个模板引发了跨多个任务的多跳推理。我们方法的核心思想很简单，如图1所示：添加“让我们逐步思考”或类似的文本（见表4），以提取逐步推理。

3.1 两阶段提示

虽然Zero-shot-CoT在概念上很简单，但它使用两次提示来提取推理和答案，如图2所示。相比之下，零样本基线（见图1左下角）已经使用了形式为“答案是”的提示，以提取正确格式的答案。

少样本提示，标准或CoT，通过明确地设计少样本示例答案以这种格式结束来避免需要这样的答案提取提示（参见右上角和左上角）

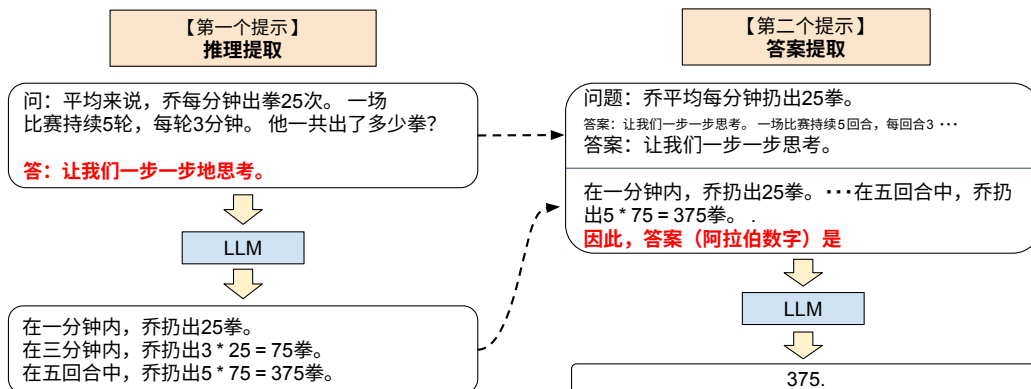


图2：Zero-shot-CoT的完整流程，如第3节所述：我们首先使用第一个“推理”提示从语言模型中提取完整的推理路径，然后使用第二个“答案”提示从推理文本中提取正确格式的答案。

在图1中)。总之，Few-shot-CoT [Wei et al., 2022] 需要仔细的人工工程。每个任务需要一些具有特定答案格式的提示示例，而 Zero-shot-CoT 则需要更少的工程，但需要两次提示 LLMs。

第一个提示：推理提取在这一步中，我们首先使用简单的模板“Q: [X]. A: [T]”，其中 [X] 是 x 的输入槽，[T] 是手工制作的触发句槽，用于提取思路链以回答问题 x 。例如，如果我们使用“让我们逐步思考”作为触发句，那么提示 x' 将是“Q: [X]”。

A: 让我们逐步思考。”。更多触发示例请参见表4。提示文本 x' 然后被输入到语言模型中，生成后续句子 z 。我们可以使用任何解码策略，但为了简单起见，我们在本文中使用了贪婪解码。

第二个提示：答案提取在第二步中，我们使用生成的句子 z 以及提示的句子 x' 从语言模型中提取最终答案。具体来说，我们简单地将三个元素连接起来，如“[X'] [Z] [A]”所示：[X'] 表示第一个提示 x' ，[Z] 表示在第一步生成的句子 z ，[A] 表示用于提取答案的触发句子。这一步的提示是自我增强的，因为提示中包含了由同一语言模型生成的句子 z 。

在实验中，我们根据答案的格式使用稍微不同的答案触发器。例如，对于多项选择题，我们使用“因此，在A到E中，答案是”；对于需要数值答案的数学问题，我们使用“因此，答案（阿拉伯数字）是”。请参阅附录A.5中的答案触发句子列表。最后，将提示文本作为输入反馈给语言模型，生成句子 y^* 并解析最终答案。有关解析器的详细信息，请参阅第4节的“答案清理”。

4 实验

任务和数据集 我们在四类推理任务的12个数据集上评估了我们的提议：算术、常识、符号和其他逻辑推理任务。详细描述请参见附录A.2。

对于算术推理，我们考虑以下六个数据集：（1）SingleEq [Koncel-Kedziorski et al., 2015]，（2）AddSub [Hosseini et al., 2014]，（3）MultiArith [Roy and Roth, 2015]，（4）AQUA-RAT [Ling et al., 2017]，（5）GSM8K [Cobbe et al., 2021]，和（6）SVAMP [Patel et al., 2021]。前三个来自经典的Math World Problem Repository [Koncel-Kedziorski et al., 2016]，后三个来自更近期的基准测试。SingleEq和AddSub包含较简单的问题，不需要多步计算来解决任务。MultiArith、AQUA-RAT、GSM8k和SVAMP是更具挑战性的数据集，需要多步推理来解决。

对于常识推理，我们使用了CommonsenseQA [Talmor et al., 2019]和StrategyQA [Geva et al., 2021]。CommonsenseQA提出了具有复杂语义的问题，通常需要进行推理。

基于先前的知识 [Talmor et al., 2019]。StrategyQA要求模型进行隐式的多跳推理来回答问题 [Geva et al., 2021]。

对于符号推理，我们使用了Last Letter Concatenation和Coin Flip [Wei et al., 2022]。Lastletter Concatenation要求模型将每个单词的最后一个字母连接起来。我们随机选择了每个样本的四个名字。Coin Flip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。我们创建了四次翻转或不翻转的样本。尽管这些任务对人类来说很容易，但语言模型通常表现出平坦的扩展曲线。

对于其他逻辑推理任务，我们从BIG-bench项目中选择了两个评估集：日期理解²和跟踪洗牌对象。日期理解要求模型从上下文中推断出日期。跟踪洗牌对象测试模型根据初始状态和一系列对象洗牌推断出对象的最终状态的能力。我们使用了一个包含三个洗牌对象的数据集进行实验。

我们总共使用了17个模型进行实验。主要实验使用了Instruct-GPT3 [Ouyang等, 2022] (text-ada/babbage/curie/davinci-001和text-davinci-002)³，原始GPT3 [Brown等, 2020] (ada、babbage、curie和davinci)⁴，以及PaLM [Chowdhery等, 2022] (8B、62B和540B)。此外，我们还使用了GPT-2 [Radford等, 2019]、GPT-Neo [Black等, 2021]、GPT-J [Wang和Komatsuzaki, 2021]、T0 [Sanh等, 2022]和OPT [Zhang等, 2022]进行模型扩展研究。语言模型的大小范围从0.3B到540B。我们包括了标准模型（如GPT-3和OPT）和遵循指令的变体模型（如Instruct-GPT3和T0）。有关模型描述的详细信息，请参见附录A.3。除非另有说明，我们在实验中始终使用text-davinci-002。

基准 我们主要将我们的零样本-CoT与标准的零样本提示进行比较，以验证其思维链推理的有效性。对于零样本实验，默认情况下使用与零样本-CoT相似的答案提示。详见附录A.5。为了更好地评估LLMs在推理任务上的零样本能力，我们还将我们的方法与[Wei et al., 2022]中的少样本和少样本-CoT基准进行比较，使用相同的上下文示例。在所有方法中，我们都使用贪婪解码。对于零样本方法，结果是确定性的。对于少样本方法，由于上下文示例的顺序可能会影响结果[Lu et al., 2022]，我们只运行一次每个实验，并在所有方法和数据集上使用固定的种子，以与零样本方法进行公平比较。Wei et al. [2022]表明，示例的顺序不会导致CoT实验中的大方差。

答案清洗 在模型输出文本后（参见第3节和图2），我们的方法仅选择首先满足答案格式的答案文本部分。例如，在算术任务中，如果答案提示输出“probably 375 and 376”，我们提取第一个数字“375”并将其设置为模型预测。在多项选择题中，我们将遇到的第一个大写字母设置为预测。更多细节请参见附录A.6。标准零样本方法遵循相同的思路。对于Few-shot和Few-shot-CoT方法，我们遵循[Wang et al., 2022]的做法，首先从模型输出中提取“答案是”之后的答案文本，并对答案文本进行相同的答案清洗以解析答案文本。如果在模型输出中找不到“答案是”，我们从文本的末尾开始搜索，并将满足答案格式的第二个文本设置为预测。

4.1 结果

零样本-CoT与零样本表1总结了我们的方法（零样本-CoT）和标准零样本提示（零样本）在每个数据集上的准确性。零样本-CoT在六个算术推理任务（MultiArith、GSM8K、AQUA、SVAMP）、所有符号推理和所有其他逻辑推理任务（来自BIG-bench [Srivastava et al., 2022]）中表现出色。

²尽管之前的工作[Wei et al., 2022]将日期理解任务归类为常识推理，但我们的研究将该任务归类为逻辑推理，因为该任务需要较少的先验知识和更多的日期之间的逻辑推理。

³我们对Instruct GPT-3模型的实验包括text-****-001和text-davinci-002。Text-davinci-002与text-****-001不同，它们根据从API收集的日期范围使用不同的微调数据。具体而言，text-davinci-002使用截至2021年6月的数据，而text-****-001使用截至2019年10月的数据。（请参阅<https://beta.openai.com/docs/engines/gpt-3>）

⁴我们对GPT3系列进行的实验是在2022年4月至2022年5月之间使用OpenAI API进行的，除了表4中的10-16号实验在2022年8月进行。

表1：零样本-CoT与每个任务上的零样本准确度比较。每个任务左侧的数值是使用根据答案格式的答案提取提示得到的结果，如第3节所述。右侧的数值是使用标准答案提示“答案是”进行答案提取的额外实验结果。详细设置请参见附录A.5。

	算术					
	单一等式	加减	多元算术 GSM8K		AQUA	SVAMP
零样本	74.6/ 78.7	72.2/77.0	17.7/22.7	10.4/12.5	22.4/22.4	58.8/58.7
零样本-CoT	78.0/78.7	69.6/74.7	78.7/79.3	40.7/40.5	33.5/31.9	62.1/63.7
	常识		其他推理任务		符号推理	
	常见 SenseQA	策略 问答	日期 理解对象	乱序	最后一个字母 (4个单词)	抛硬币 (4次)
零样本	68.8/72.6	12.7/ 54.3	49.3/33.6	31.3/29.7	0.2/-	12.8/53.8
零样本-CoT	64.6/64.0	54.8/52.3	67.5/61.8	52.4/52.9	57.6/-	91.4/87.8

表2：使用MultiArith和GSM8K上的准确率与基准方法进行比较。如果没有指定，将使用text-davinci-002模型。我们在Few-shot和Few-shot-CoT设置中使用了与[Wei et al.,2022]中描述的相同的8个示例。（*1）为了验证更换示例的差异性，我们将8个示例分成两组报告了4-shot-cot的两个结果。（*2）我们在Few-shot-CoT的每个示例的答案部分开头插入了“让我们逐步思考。”以测试性能提升。有关PaLM的进一步实验结果请参见附录D

	多元算术 GSM8K	
零样本	17.7	10.4
少样本（2个样本）	33.7	15.6
少样本（8个样本）	33.8	15.6
零样本-CoT	78.7	40.7
少样本-CoT（2个样本）	84.8	41.3
少样本-CoT（4个样本：第一组）（*1）	89.2	-
少样本-CoT（4个样本：第二个）（*1）	90.5	-
少样本-CoT（8个样本）	93.0	48.7
零加少样本-CoT（8个样本）（*2）	92.8	51.5
Finetuned GPT-3 175B [魏等，2022年]	-	33
Finetuned GPT-3 175B + 验证器 [魏等，2022年]	-	55
PaLM 540B：零样本	25.5	12.5
PaLM 540B：零样本-CoT	66.1	43.0
PaLM 540B：零样本-CoT + 自一致性	89.0	70.1
PaLM 540B：少样本 [魏等，2022年]	-	17.9
PaLM 540B：少样本-CoT [魏等，2022年]	-	56.9
PaLM 540B：少样本-CoT + 自一致性 [王等，2022年]	-	74.4

例如，Zero-shot-CoT在MultiArith上的得分提升从17.7%到78.7%，在GSM8K上的得分提升从10.4%到40.7%。我们的方法在剩下的两个算术推理任务（SingleEq和AddSub）上表现出色，这是预期的，因为它们不需要多步推理。

在常识推理任务中，Zero-shot-CoT并没有提供性能提升。正如魏等人[2022]报告的那样，即使是Few-shot-CoT在Lambda (135B)上也没有提供性能提升，但是与大规模的PaLM (540B)模型结合使用时，确实改善了StrategyQA，这也可能适用于我们的模型。更重要的是，我们观察到许多生成的思维链本身令人惊讶地逻辑正确，或者只包含人类可理解的错误（见表3），这表明Zero-shot-CoT在常识推理方面确实引发了更好的效果，即使任务指标并没有直接反映出来。我们在附录B中提供了Zero-shot-CoT为每个数据集生成的样本。

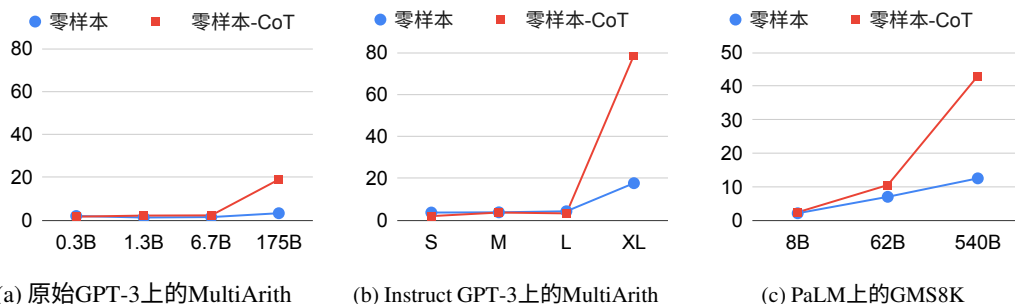


图3：使用不同类型的模型进行模型规模研究。S：text-ada-001，M：text-babbage-001，L：text-curie-001，XL：text-davinci-002。详见附录A.3和E。

表3：Zero-Shot-CoT在CommonsenseQA上生成的示例，用于错误分析。

示例1（推理正确。）	示例2（模型无法选择只有一个。）
<p>问题：一个常用的玩具车可能在哪里找到？</p> <p>答案选项：(A) 孩子的房间 (B) 男孩的卧室 (C) 自己的家 (D) 玩具店 (E) 房子</p> <p>A: 让我们逐步思考。玩具车很可能在孩子的房间里找到。孩子的房间很可能在房子里找到。因此，玩具车很可能在房子里找到。</p> <p>因此，在A到E中，答案是 E。</p> <p>X(GT: A, 零样本: A)</p>	<p>Q:为了保持身材需要什么？</p> <p>答案选项：(A) 良好的健康状况 (B) 锻炼 (C) 肌肉紧张 (D) 出汗 (E) 感觉更好</p> <p>A: 让我们逐步思考。为了保持身材，你需要能够做到以下三点：1. 定期锻炼</p> <p>2. 饮食健康</p> <p>3. 获得足够的睡眠</p> <p>如果你能做到这三点，那么你就在保持身材的道路上了！因此，在A到E中，答案是B、C和D。</p> <p>(GT: B, 零样本: B)</p>

与其他基准的比较表2比较了Zero-shot-CoT和基准在两个算术推理基准（MultiArith和GSM8K）上的性能。标准提示（第一个块）和思维链提示（第二个块）之间的巨大差距表明，这些任务在不引发多步推理的情况下是困难的。在Instruct GPT-3（text-davinci-002）和PaLM（540B）模型上都得到了重大改进（第四个块）。虽然Zero-shot-CoT在性能上自然不如Few-shot-CoT，但它在即使每个任务有8个示例的情况下，也明显优于标准Few-shot提示。对于GSM8K，Zero-shot-CoT与Instruct GPT-3（text-davinci-002）相比，也优于微调的GPT-3和使用大型模型（PaLM，540B）的标准Few-shot提示，这些结果在Wei等人[2022]（第三个和第四个块）中有报道。有关PaLM的更多实验结果，请参见附录D。

模型大小对零样本推理有影响吗？图3比较了多个语言模型在MultiArith / GSM8K上的性能。在没有思维链推理的情况下，随着模型规模的增加，性能要么不增加，要么增长缓慢，即曲线基本上是平的。相反，对于原始/指导GPT-3和PaLM，随着模型大小的增加，思维链推理的性能显著提高。当模型大小较小时，思维链推理并不有效。这个结果与Wei等人[2022]的少样本实验结果一致。

附录E展示了使用更多种类的语言模型进行的广泛实验结果，包括GPT-2、GPT-Neo、GPT-J、T0和OPT。我们还手动调查了生成的思维链质量，大规模模型明显表现出更好的推理能力（请参见附录B中每个模型的样本输出）。

错误分析 为了更好地理解Zero-shot-CoT的行为，我们手动调查了由Instruct-GPT3与Zero-shot-CoT提示生成的随机选择的示例。请参阅附录C中的示例，其中一些观察结果包括：（1）在常识推理（CommonsenseQA）中，Zero-shot-CoT通常会产生灵活合理的思路链，即使最终预测不正确。当模型难以缩小到一个答案时，Zero-shot-CoT通常会输出多个答案选择（请参见表3中的示例）。（2）在算术中

表4：在MultiArith数据集上使用text-davinci-002进行模板的鲁棒性研究。

(*1) 此模板用于Ahn等人[2022]的研究中，其中提示语言模型根据高级指令生成逐步操作以控制机器人动作。(*2) 此模板用于Reynolds和McDonell [2021]的研究中，但没有进行定量评估。

编号	类别	模板	准确性
1	有教育意义	让我们逐步思考	78.7
2		首先, (*1)	77.3
3		让我们逻辑思考一下这个问题	74.5
4		让我们将这个问题分解成几个步骤来解决 (*2)	72.2
5		让我们现实一点, 逐步思考	70.8
6		让我们像侦探一样逐步思考	70.3
7		让我们思考一下	57.5
8		在我们深入答案之前	55.7
9		答案在证明之后	45.7
10	误导性	不要思考, 只是感受	18.8
11		让我们逐步思考, 但得出一个错误的答案	18.7
12		让我们数一下问题中的"a"的数量	16.7
13		通过利用地球是圆的事实	9.3
14	不相关的	顺便说一下, 我在附近找到了一家好餐厅	17.5
15		阿布拉卡达布拉!	15.5
16		今天天气真好	13.1
-		(零样本)	17.7

表5：Few-shot-CoT对示例的鲁棒性研究。当示例来自完全不同的任务时，性能通常变差，但当答案格式匹配时（即CommonsenseQA到AQUA-RAT，多项选择题），性能损失较小。

†在这个变体中使用了CommonsenseQA样本

	零样本	Few-shot-CoT †	零样本-CoT	Few-shot-CoT
AQUA-RAT	22.4	<u>31.9</u>	33.5	39.0
MultiArith	17.7	<u>27.0</u>	78.7	88.2

推理（MultiArith），Zero-shot-CoT和Few-shot-CoT在错误模式上存在显著差异。首先，Zero-shot-CoT倾向于在获得正确预测后输出不必要的推理步骤，从而将预测更改为不正确的预测。Zero-shot-CoT有时也不会开始推理，只是改述输入的问题。相比之下，Few-shot-CoT在生成的思维链中包含三元操作（例如 $(3 + 2) * 4$ ）时容易失败。

提示选择如何影响零样本推理？我们验证了零样本推理对输入提示的鲁棒性。表4总结了使用16个不同模板和三个类别的性能。具体而言，根据Webson和Pavlick [2022]的研究，这些类别包括指导性（鼓励推理），误导性（阻止推理或以错误的方式鼓励推理）和无关的（与推理无关）。结果表明，如果文本以鼓励思维链推理的方式编写，即模板属于“指导性”类别，性能会有所提高。然而，准确性的差异取决于句子。在这个实验中，“让我们逐步思考。”取得了最好的结果。有趣的是，不同的模板会以不同的方式鼓励模型表达推理（请参见附录B中每个模板的样本输出）。相反，当我们使用误导性或无关的模板时，性能不会提高。如何自动创建更好的零样本推理模板仍然是一个未解决的问题。

提示选择如何影响少样本-CoT？表5显示了使用不同数据集的Few-shot-CoT的性能：从CommonsenseQA到AQUA-RAT和从CommonsenseQA到MultiArith。在这两种情况下，领域是不同的，但答案格式是相同的。

前者的答案格式相同。令人惊讶的是，来自不同领域（常识到算术）但答案格式相同（多项选择）的思维链示例相对于零样本（到AQUA-RAT）提供了显著的性能提升，相对于零样本-CoT或少样本-CoT的可能改进来衡量。相比之下，使用具有不同答案类型的示例（到MultiArith）时，性能提升要小得多，这证实了之前的研究[Min et al., 2022]的观点，即LLMs主要利用少样本示例来推断重复的格式，而不是任务本身的上下文。然而，对于这两种情况，结果都比零样本-CoT差，这证实了在少样本-CoT中进行任务特定样本工程的重要性。

5 讨论和相关工作

表6：算术/常识推理任务的相关工作总结。类别表示训练策略。CoT表示是否输出思维链。任务列出了相应论文中执行的任务。AR：算术推理，CR：常识推理。

方法	类别	CoT	任务	模型
Rajani等人[2019]	微调	✓	CR	GPT
Cobbe等人[2021]	微调	✓	AR	GPT-3
Zelikman等人[2022]	微调	✓	AR, CR	GPT-3等
Nye等人[2022]	微调 ⁵	✓	AR	Transformer（解码器）
Brown等人[2020]	少量/零样本		CR	GPT-3
Smith等人[2022]	少量/零样本		AR, CR	MT-NLG
Rae等人[2021]	少样本		AR, CR	Gopher
Wei等人[2022]	少样本	✓	AR, CR	PaLM, LaMBDA, GPT-3
王等人[2022年]	少样本	✓	AR, CR	PaLM等
Chowdhery等人[2022年]	少样本	✓	AR, CR	PaLM
Shwartz等人[2020年]	零样本	✓	CR	GPT-2等
Reynolds和McDonell[2021年]	零样本	✓	AR	GPT-3
Zero-shot-CoT（我们的）	零样本	✓	AR, CR	PaLM, Instruct-GPT3, GPT-3等

LLM的推理能力多项研究表明，预训练模型通常不擅长推理[Brown等人，2020年，Smith等人，2022年，Rae等人，2021年]，但通过使其产生逐步推理，可以大大提高其能力，无论是通过微调[Rajani等人，2019年，Cobbe等人，2021年，Zelikman等人，2022年，Nye等人，2022年]还是通过少样本提示[Wei等人，2022年，Wang等人，2022年，Chowdhery等人，2022年]（有关相关工作的摘要，请参见表6）。与大多数先前的工作不同，我们专注于零样本提示，并表明单个固定触发提示可以大大提高LLM在需要复杂多跳思维的各种任务中的零样本推理能力（表1），特别是在模型扩展时（图3）。它还在各种任务中生成了合理且可理解的思路链（附录B），即使最终预测是错误的（附录C）。与我们的工作类似，Reynolds和McDonell[2021年]展示了一个提示，“让我们通过将问题分解为步骤来解决这个问题”，将有助于简单算术问题的多步推理。然而，他们将其视为任务特定的示例，并未针对各种推理任务与基准进行定量评估。Shwartz等人[2020年]提出将常识问题分解为一系列信息寻求问题，例如“[x]的定义是什么”。它不需要演示，但对每个推理任务需要大量手动提示工程。我们的结果强烈表明LLM是不错的零样本推理者，而先前的工作[Wei等人，2022年]通常只强调少样本学习和任务特定的上下文学习，例如没有报告零样本基准。我们的方法不需要耗时的微调或昂贵的样本工程，并且可以与任何预训练的LLM结合使用，作为所有推理任务的最强零样本基准。

LLMs的零样本能力Radford等人[2019]表明，LLMs在许多系统-1任务中具有出色的零样本能力，包括阅读理解、翻译和摘要。

⁵Nye等人[2022]也评估了少样本设置，但在其领域上的少样本性能比微调结果差。

Sanh等人[2022], Ouyang等人[2022]表明, 通过显式微调模型以遵循指令, 可以增加LLMs的这种零样本能力。尽管这些工作关注LLMs的零样本性能, 但我们关注超越系统-1任务的许多系统-2任务, 这被认为是LLMs的一个重大挑战, 因为其扁平的缩放曲线。此外, 零样本-CoT与指令微调是正交的; 它提高了Instruct GPT3、vanilla GPT3和PaLM的零样本性能(见图3)。

从狭窄(任务特定)到广泛(多任务)的提示大多数提示是任务特定的。

虽然少量样本提示是由于任务特定的上下文样本[布朗等人, 2020年, 魏等人, 2022年], 但大多数零样本提示也专注于每个任务的工程化(模板)[刘等人, 2021b年, 雷诺兹和麦克唐纳, 2021年]。借鉴Chollet [2019]的术语, 该术语建立在智能的分层模型[麦格鲁, 2005年, 约翰逊和布查德Jr, 2005年]之上, 这些提示可以说是从LLMs中引出“狭窄的泛化”或任务特定的技能。另一方面, 我们的方法是一个多任务提示, 引出LLMs中的“广泛泛化”或广泛的认知能力, 如逻辑推理或系统-2本身。我们希望我们的工作不仅可以作为LLMs的逻辑推理研究的参考, 还可以发现LLMs中的其他广泛的认知能力。

训练数据集细节 该研究的一个限制是缺乏关于LLM使用的训练数据集的公开信息, 例如GPT模型的001 vs 002, 原始GPT3 vs Instruct-GPT [Ouyang等, 2022], 以及PaLM模型的数据[Chowdhery等, 2022]。然而, 从零样本到零样本-CoT的大型模型(InstructGPT 001或002, 原始GPT3和PaLM)都有显著的性能提升, 并且在算术和非算术任务中都有持续的改进, 这表明这些模型不太可能仅仅是记忆, 而是具有通用问题解决的任务无关的多步推理能力。尽管大多数结果都基于InstructGPT, 因为它是最好的开放获取LLM, 但关键结果也在PaLM上得到了复现, 并且InstructGPT中的数据细节(Ouyang等人[2022]的附录A、B和F)也证实它并非专门为多步推理而设计。

限制和社会影响我们的工作基于对大型语言模型的提示方法。

LLMs已经在网络上的各种来源的大型语料库上进行了训练(还请参阅“训练数据集详情”), 并且已经显示出捕捉和放大训练数据中存在的偏见。提示是一种利用语言模型捕捉到的有利于各种任务的模式的方法, 因此它也具有相同的缺点。话虽如此, 我们的方法是一种更直接的方式, 用于探索预训练LLMs内部的复杂推理, 消除了前几次少样本方法中上下文学习的混淆因素, 并可以导致对LLMs中偏见的更无偏见的研究。

6 结论

我们提出了Zero-shot-CoT, 这是一种单一的零样本提示, 可以从大型语言模型中引出各种推理任务, 与以前的少样本(上下文)方法相比, 不需要为每个任务手工制作少样本示例。我们的简单方法不仅是困难的多步骤系统-2推理任务的最简主义和最强零样本基准, 而且还鼓励社区进一步发现类似的多任务提示, 以引出广泛的认知能力而不是狭窄的任务特定技能。

致谢

这项工作得到了沙特-日本2030年愿景下的穆罕默德·本·萨勒曼未来科学与技术中心(Mb SC2030)在东京大学的支持。除了PaLM之外, 实验使用了由国家先进工业科学技术研究所(AIST)提供的AI Bridging Cloud Infrastructure (ABCI)的计算资源。我们还感谢Jason Wei和Denny Zhou对运行PaLM实验的讨论和支持, 以及Sharan Narang和Aakanksha Chowdhery对PaLM基础设施的通用支持。

参考文献

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian

- Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, 和Mengyuan Yan。按照我能做的去做，而不是我说的：将语言与机器人的可行性联系起来，2022年。网址<https://arxiv.org/abs/2204.01691>.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy和Stella Biderman。GPT-Neo：使用MeSh-Tensorflow进行大规模自回归语言建模，2021年3月。网址<https://doi.org/10.5281/zenodo.5297715>.
- 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。语言模型是少样本学习器。在H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan和H. Lin编辑的 *Advances in NeurIPS*, 卷33, 页码1877-1901。Curran Associates, Inc., 2020. 网址<https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfbcb4967418bfb8ac142f64a-Paper.pdf>.
- François Chollet. 关于智能度量的研究。arXiv预印本 arXiv:1911.01547, 2019. 网址<https://arxiv.org/abs/1911.01547>.
- 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. Palm: 使用路径扩展语言模型，2022年。网址<https://arxiv.org/abs/2204.02311>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse和John Schulman。训练验证器解决数学问题，2021年。网址<https://arxiv.org/abs/2110.14168>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee和Kristina Toutanova。BERT：深度双向转换器的预训练用于语言理解。在NAACL会议的论文中4171-4186页，2019年。网址<https://aclanthology.org/N19-1423>.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser和Connor Leahy。The pile：一个80 OGB的多样文本数据集用于语言建模。arXiv预印本arXiv: Arxiv-2101.00027, 2020年。
- Tianyu Gao, Adam Fisch, 和 Danqi Chen。使预训练语言模型在少样本学习中更好。在ACL-IJCN LP会议上，页码为3816-3830，2021年。网址<https://aclanthology.org/2021.acl-long.295>.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, 和 Jonathan Berant。亚里士多德使用笔记本电脑吗？一个具有隐式推理策略的问答基准。TACL, 9:346-361, 2021年。网址<https://aclanthology.org/2021.tacl-1.21/>.
- Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, 和 Nate Kushman。学习通过动词分类解决算术问题。在EMNLP会议上，卷号为523533。Citeseer, 2014年。网址<https://aclanthology.org/D14-1058/>.

- Wendy Johnson和Thomas J Bouchard Jr. 人类智力的结构：它是语言、感知和图像旋转（vpr），而不是流体和结晶的。Intelligence, 33（4）：393-416，2005年。
- Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni和Siena Dumas Ang. 将代数单词问题解析为方程。TACL, 3: 585-597, 2015年。URL <https://aclanthology.org/Q15-1042>。
- Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman和Hannaneh Hajishirzi. MAWPS：一个数学问题存储库。在NAACL会议论文集，页1152-1157，2016年。URL <https://aclanthology.org/N16-1136>。
- 王凌, Dani Yogatama, Chris Dyer和Phil Blunsom. 通过理由生成进行程序归纳：学习解决和解释代数问题。在ACL会议论文集，第158-167页，2017年。网址<https://aclanthology.org/P17-1015>。
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin和Weizhu Chen. 对于gpt-3来说，什么是好的上下文示例？arXiv预印本 arXiv:2101.06804, 2021a年。网址<https://arxiv.org/abs/2101.06804>。
- 刘鹏飞, 袁伟哲, 傅金兰, 蒋正宝, 林浩明和Graham Neubig. 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, 2021b年。网址<https://arxiv.org/abs/2107.13586>。
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel和Pontus Stenetorp. 奇妙有序的提示及其发现方法：克服少样本提示顺序敏感性。在ACL会议论文中，第8086-8098页，2022年。网址<https://aclanthology.org/2022.acl-long.556>。
- Kevin S McGrew. Cattell-Horn-Carroll认知能力理论：过去、现在和未来。2005。
- Stephen Merity, Caiming Xiong, James Bradbury和Richard Socher. 指针哨兵混合模型。arXiv预印本 arXiv:1609.07843, 2016年。网址<https://arxiv.org/abs/1609.07843>。
- Sewon Min, Xinxin Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi和Luke Zettlemoyer. 重新思考演示的作用：什么使得上下文学习有效？arXiv预印本 arXiv:2202.12837, 2022年。网址<https://arxiv.org/pdf/2202.12837.pdf>。
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, 和Augustus Odena. 展示你的工作：用于语言模型中间计算的草稿本。在2022年的代码深度学习研讨会中。网址<https://openreview.net/forum?id=HB1x2idbkq>。
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, 和Ryan Lowe. 用人类反馈训练语言模型遵循指令，2022年。网址<https://arxiv.org/abs/2203.02155>。
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga等。PyTorch：一种命令式风格的高性能深度学习库。NeurIPS的进展, 32:8026-8037, 2019年。URL <https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>。
- Arkil Patel, Satwik Bhattamishra和Navin Goyal. NLP模型真的能够解决简单的数学问题吗？在NAACL会议的论文中，第2080-2094页，2021年。URL <https://aclanthology.org/2021.naacl-main.168>。
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever等。语言模型是无监督的多任务学习器。OpenAI博客，第9页，2019年。URL <http://www.persagen.com/files/misc/radford2019language.pdf>。

- 杰克·W·雷, 塞巴斯蒂安·博尔戈, 特雷弗·凯, 凯蒂·米利坎, 乔丹·霍夫曼, 弗朗西斯·宋, 约翰·阿斯拉尼迪斯, 萨拉·亨德森, 罗曼·林, 苏珊娜·杨, 伊丽莎·卢瑟福, 汤姆·亨尼根, 雅各布·梅尼克, 阿尔宾·卡西雷尔, 理查德·鲍威尔, 乔治·范登·德里斯, 丽莎·安妮·亨德里克斯, 玛丽贝丝·劳, 黄柏森, 阿梅利亚·格莱斯, 约翰内斯·韦布尔, 苏曼特·达塔斯里, 藏红花·黄, 乔纳森·乌萨托, 约翰·梅洛尔, 伊琳娜·希金斯, 安东尼娅·克雷斯特尔, 纳特·麦卡利斯, 艾米·吴, 埃里克·埃尔森, 西丹汉特·贾亚库马尔, 埃琳娜·布查茨卡娅, 大卫·巴登, 埃斯梅·萨瑟兰, 卡伦·西莫尼扬, 米凯拉·帕加尼尼, 洛朗·西弗雷, 莱娜·马滕斯, 李翔·洛雷恩, 阿迪古纳·昆科罗, 艾达·内马扎德, 埃琳娜·格里博夫斯卡娅, 多梅尼克·多纳托, 安吉利基·拉扎里杜, 亚瑟·门斯, 让-巴蒂斯特·莱皮奥, 玛丽亚·辛普基利, 尼古拉·格里戈列夫, 道格·弗里茨, 蒂博·波伦, 龚志涛, 丹尼尔·托亚马, 西普里安·德·马松·多图姆, 李宇佳, 泰芬·特尔齐, 弗拉基米尔·米库利克, 伊戈尔·巴布什金, 艾丹·克拉克, 迭戈·德拉斯卡萨斯, 奥雷利亚·盖伊, 克里斯·琼斯, 詹姆斯·布拉德伯里, 马修·约翰逊, 布莱克·赫奇曼, 劳拉·韦丁格, 伊桑·加布里埃尔, 威廉·艾萨克, 埃德·洛克哈特, 西蒙·奥辛德罗, 劳拉·里梅尔, 克里斯·戴尔, 奥里奥尔·维尼亚尔斯, 卡里姆·阿尤卜, 杰夫·斯坦韦, 洛雷恩·贝内特, 德米斯·哈萨比斯, 科雷·卡武克乔卢, 和杰弗里·欧文。扩展语言模型: 培训Gopher的方法、分析和见解, 2021年。网址<https://arxiv.org/abs/211>
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li 和 Peter J. Liu。探索使用统一的文本到文本转换器进行迁移学习的极限。JMLR, 21 (140) : 1-67, 2020年。网址<http://jmlr.org/papers/v21/20-074.html>。
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong 和 Richard Socher。解释你自己! 利用语言模型进行常识推理。在ACL会议论文集, 页码4932-4942, 2019年。网址<https://aclanthology.org/P19-1487>。
- Laria Reynolds 和 Kyle McDonell。大型语言模型的提示编程: 超越少样本范式。在2021年CHI人机交互计算系统会议扩展摘要, 页码1-7, 2021年。网址<https://arxiv.org/pdf/2102.07350.pdf>。
- Subhro Roy 和 Dan Roth。解决一般的算术问题。在EMNLP会议论文集中, 第1743-1752页, 2015年。网址<https://aclanthology.org/D15-1202>。
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf 和 Alexander M Rush。多任务提示训练实现了零样本任务泛化能力。在ICLR会议中, 2022年。网址<https://openreview.net/forum?id=9Vrb9D0WI4>。
- Timo Schick 和 Hinrich Schütze。重要的不仅仅是大小: 小型语言模型也是少样本学习者。在NAACL会议论文集, 第2339-2352页, 2021年。网址<https://aclanthology.org/2021.naacl-main.185>。
- Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace 和 Sameer Singh。Auto-Prompt: 通过自动生成的提示从语言模型中引出知识。在EMNLP会议论文集, 第4222-4235页, 2020年。网址<https://aclanthology.org/2020.emnlp-main.346>。
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula 和 Yejin Choi。无监督的常识问答与自我对话。在EMNLP会议论文集, 第4615-4629页, 2020年。网址<https://aclanthology.org/2020.emnlp-main.373>。
- Shaden Smith, Mostofa Patwary, Brandon Norrick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary 和 Bryan Catanzaro。使用deepspeed和megatron来训练megatron-turing nlg 530b, 一个大规模生成式语言模型, 2022年。网址 <https://arxiv.org/abs/2201.11990>。

- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso等。超越模仿游戏：量化和推断语言模型的能力。arXiv预印本 *arXiv:2206.04615*, 2022年。网址<https://arxiv.org/abs/2206.04615>。
- Keith E Stanovich和Richard F West。推理中的个体差异：对理性辩论的影响？行为和脑科学, 23(5): 645-665, 2000年。
- Alon Talmor, Jonathan Herzig, Nicholas Lourie和Jonathan Berant。Commonsenseqa: 针对常识知识的问答挑战。在NAACL-HLT会议论文集, 页码4149-4158, 2019年。网址<https://aclanthology.org/N19-1421/>。
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera-Arcas, Claire Cui, Marian Croak, Ed Chi和Quoc Le。Lamda: 用于对话应用的语言模型, 2022年。网址 <https://arxiv.org/abs/2201.08239>。
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser和Illia Polosukhin。注意力就是你所需要的。在I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan和R. Garnett编辑, *Advances in NeurIPS*, 2017年。网址<https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>。
- Ben Wang和Aran Komatsuzaki。GPT-J-6B: 一个60亿参数的自回归语言模型。<https://github.com/kingoflolz/mesh-transformer-jax>, 2021年5月。
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi和Denny Zhou。自洽性改善了语言模型中的思维链推理。arXiv预印本 *arXiv:2203.11171*, 2022年。网址<https://arxiv.org/abs/2203.11171>。
- Albert Webson和Ellie Pavlick。基于提示的模型真的理解其提示的含义吗？在北美计算语言学协会2022年会议论文集：人类语言技术的论文中, 第2300-2344页。计算语言学协会, 2022年7月。网址<https://aclanthology.org/2022.naacl-main.167>。
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le和Denny Zhou。思维链提示引发了大型语言模型的推理, 2022年。网址 <https://arxiv.org/abs/2201.11903>。
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest和Alexander Rush。变压器：最先进的自然语言处理。在*EMNLP会议论文集中*, 2020年。网址<https://aclanthology.org/2020.emnlp-demos.6>。
- Eric Zelikman, Yuhuai Wu和Noah D. Goodman。Star: 用推理引导推理, 2022年。网址<https://arxiv.org/abs/2203.14465>。
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin等。Opt: 开放预训练的变压器语言模型。arXiv预印本 *arXiv:2205.01068*, 2022年。网址<https://arxiv.org/abs/2205.01068>。

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]
 - (c) 你是否讨论了你的工作可能产生的负面社会影响? [是](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. 如果你包含了理论结果...
 - (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](b) 你是否指定了所有训练细节(例如数据拆分、超参数、选择方式)? [是]
 - (c) 你是否报告了误差条(例如,在多次运行实验后,关于随机种子的)? [否] 我们的论文主要使用了GPT-3 API进行贪婪解码,实验中没有随机性。
 - (d) 你是否包含了计算总量和使用的资源类型(例如, GPU类型,内部集群或云提供商)? [是]
4. 如果您使用现有资产(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]
 - (b) 您是否提到了资产的许可证? [是]
 - (c) 您是否在补充材料或URL中包含了任何新资产? [是]
 - (d) 您是否讨论了从您使用/策划数据的人那里获得同意的方式和方法? [是]
 - (e) 您是否讨论了您使用/策划的数据是否包含个人身份信息或冒犯性内容? [是]
5. 如果你使用了众包或与人类参与者进行了研究...
 - (a) 你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) 你是否描述了任何潜在的参与者风险,并提供了研究伦理委员会(IRB)批准的链接(如果适用)? [不适用]
 - (c) 你是否包含了参与者获得的每小时工资估计和参与者补偿的总金额? [不适用]

实验设置的详细信息

A.1 代码

代码可在https://github.com/kojima-takeshi188/zero_shot_cot找到。

A.2 数据集

A.2.1 数据集描述

表7总结了我们的实验中使用的每个数据集的描述。

表7: 数据集描述。我们的实验使用了除“Last Letters”和“Coin Flip”数据集之外的公开可用数据集。我们创建了这两个数据集。详细信息请参见附录A.2.2。
(*1) N: 数字, M: 从多个选择中选择一个, Y: 回答是或否, F: 自由格式。
(*2) 问题文本中的平均单词数。

数据集	答案格式 (*1)	样本数量	数(*2 平均单词)	数据拆分 (文件名) 用于我们的实验	许可证
单一等式	N	508	27.4	questions.json	无许可证
AddSub	N	395	31.5	AddSub.json	未指定
MultiArith	N	600	31.8	MultiArith.json	未指定
GSM8K	N	1319	46.9	test.jsonl	MIT许可证
AQUA-RAT	M	254	51.9	test.jsonl	Apache-2.0
SVAMP	N	1000	31.8	SVAMP.json	MIT许可证
CommonsenseQA	M	1221	27.8	dev_rand_split.jsonl	未指定
StrategyQA	Y	2290	9.6	task.json	Apache-2.0
日期理解	M	369	35.0	task.json	Apache-2.0
随机对象	M	750	91.1	three_objects/task.json	Apache-2.0
最后的字母	F	500	15.0	-	-
抛硬币	Y	500	37.0	-	-

A.2.2 数据集创建

关于“最后一个字母连接”和“抛硬币”，数据集不公开可用，因此我们按照魏等人[2022]的问题模板进行了微小的改写来创建数据集。

具体来说，对于最后一个字母连接，我们使用以下模板。我们从names-dataset库 (<https://pypi.org/project/names-dataset/>) 随机选择人名，并将它们插入到{Name1}到{Name4}中。

- '取出"{Name1} {Name2} {Name3} {Name4}"中每个单词的最后一个字母，并将它们连接起来。'

对于抛硬币，我们使用以下模板。我们从names-dataset库中随机选择人名，并将它们插入到{Name1}到{Name4}中。我们还随机选择“翻转”或“不翻转”并将短语分别插入到每个{翻转 | 不翻转}部分。

- '一枚硬币正面朝上。{Name1} {翻转 | 不翻转}硬币。{Name2} {翻转 | 不翻转}硬币。{Name3} {翻转 | 不翻转}硬币。{Name4} {翻转 | 不翻转}硬币。
硬币还是正面朝上吗？ 请注意，这里的“翻转”意味着“反转”。'

A.3 语言模型

我们的实验使用了多个语言模型，如表8所述

A.4 实现细节

对于原始GPT-3和Instruct-GPT3，我们使用了OpenAI API。对于OPT、T0、GPT-J、GPT-Neo和GPT-2，我们使用了Hugging Face Transformer Library [Wolf et al., 2020]。我们将max_tokens设置为128，并且

表8：语言模型描述。（*1）对于原始GPT3模型，我们参考了<https://blog.eleuther.ai/gpt3-model-sizes/>和<https://beta.openai.com/docs/model-index-for-researchers>，为每个模型分配了模型大小信息。（*2）关于Instruct GPT3的模型大小没有官方信息。我们推断API名称的顺序与原始GPT3的模型大小相匹配。

语言模型参数数量	图书馆 / API名称	图书馆 / API中的模型名称	许可证
PaLM	540B	-	未指定
PaLM	62B	-	未指定
PaLM	8B	-	未指定
原始GPT3	175B (*1)	OpenAI API	达芬奇
原始GPT3	6.7B (*1)	OpenAI API	居里
原始GPT3	1.3B (*1)	OpenAI API	巴贝奇
原始GPT3	0.3B (*1)	OpenAI API	艾达
指导GPT3	- (*2)	OpenAI API	text-davinci-002
指导GPT3	- (*2)	OpenAI API	text-davinci-001
指导GPT3	- (*2)	OpenAI API	text-curie-001
指导GPT3	- (*2)	OpenAI API	text-babbage-001
指导GPT3	- (*2)	OpenAI API	text-ada-001
OPT	13B	Hugging Face Library	opt-13b
T0	11B	Hugging Face Library	T0pp
GPT-J	6B	Hugging Face图书馆gptj	Apache-2.0
GPT-Neo	27亿	拥抱面库gpt-neo	Apache-2.0
GPT-2	15亿	拥抱面库gpt2-xl	Apache-2.0

在所有的方法和模型中，我们使用贪婪解码（温度为0，OpenAI API的情况除外），除了PaLM。对于PaLM，我们使用'TopK=1'进行贪婪确定性解码，max_tokens = 256。除了Instruct-GPT3，我们将“Q:”设置为自定义停止序列，以防止模型自己重复问题和答案。我们在云V100实例上运行实验，对于GPT-3模型，没有GPU，对于T0和OPT，使用云A100x8 GPU（60GB）实例，对于GPT-J，GPT-Neo和GPT-2，使用云A100x1 GPU（60GB）实例。我们的实现是基于PyTorch [Paszke et al., 2019]。

A.5用于答案提取的提示

表9和表10总结了在表1中用于实验的一系列答案提取提示。我们在所有实验中都使用了零样本（左）和零样本-CoT（左）作为答案提取的默认提示。

表9：在表1中用于零样本实验的答案提取提示。C.S.QA：常识QA，D.U.：日期理解，S.O.：追踪洗牌对象

无任务	零样本（左）	零样本（右）
1 单一等式	答案（阿拉伯数字）是	答案是
2 AddSub	答案（阿拉伯数字）是	答案是
3 MultiArith	答案（阿拉伯数字）是	答案是
4 GSM8K	答案（阿拉伯数字）是	答案是
5 AQUA-RAT	在A到E中，答案是	答案是
6 SVAMP	答案（阿拉伯数字）是	答案是
7 C.S.QA	在A到E中，答案是	答案是
8 StrategyQA	答案（是或否）是	答案是D.U.
9	在A到F之间，答案是	答案是S.O.
10	在A到C之间，答案是	答案是最后
11 一个字母	答案是	答案是硬币
12 翻转	答案（是或否）是	答案是

表格10：用于零样本-CoT实验的答案提取提示，见表1。C.S.QA：常识问答，D.U.：日期理解，S.O.：追踪洗牌对象

无任务	零样本-CoT（左）	零样本-CoT（右）
1 单一等式	因此，答案（阿拉伯数字）是	因此，答案是加减
2	因此，答案（阿拉伯数字）是	因此，答案是多项式
3 运算	因此，答案（阿拉伯数字）是	因此，答案是GSM8K
4	因此，答案（阿拉伯数字）是	因此，答案是AQUA-
5 RAT。因此，在A到E之间，答案是。因此，答案是SVAMP	因此，答案（阿拉伯数字）是	因此，答案是C.S.QA
6	因此，在A到E之间，答案是。因此，答案是StrategyQA	
7	因此，答案（是或否）是	因此，答案是
8 D.U.	因此，在A到F之间，答案是因此，答案是	
9 S.O.	因此，在A到C之间，答案是因此，答案是	
10 最后的字母	因此，答案是	因此，答案是
11 抛硬币	因此，答案（是或否）是	因此，答案是

A.6 答案清洗

表11总结了在所有实验中使用的答案清洗方法的列表。

表11：答案清洗的详细描述。请参阅表7，了解每个数据集与相应答案格式之间的映射。

答案格式	答案清洗方法	伪代码 (以Pytorch 3.8为例)
数字	在文本中找到第一个数字。	<pre>pred = pred.replace(" ", "") pred = [s for s in re.findall(r'[-?\d+\.\d*', pred)] pred = pred[0]</pre>
多项选择	在文本中找到第一个大写字母。	<pre>pred = re.findall(r'[A B C D E]', pred) pred = pred[0]</pre>
是或否	在删除不必要的字母后，在文本中找到第一个"是"或"否"。	<pre>pred = pred.lower() pred = re.sub("[\ ' \n \. \s \: \; ,", "", pred) pred = pred.split(" ") pred = [i for i in pred if i in ("yes", "no")] pred = pred[0]</pre>
自由格式	只需删除不必要的字母，如"。"。	<pre>pred = re.sub("[\ ' \n \. \s", "", pred)</pre>

B 附加实验结果

本节总结了我们的实验中模型生成的更多示例文本。请注意，为了可读性，所有文本都经过修改，省略或插入了一些换行符。
除非另有说明，我们使用Instruct-GPT3（text-davinci-002）模型。

- 表12列出了Zero-shot-CoT为每个数据集生成的示例文本（参见表1）。
- 表13列出了Zero-shot-CoT为每个推理提取模板生成的示例文本（参见表4）。
- 表14和表15列出了Zero-shot-CoT为每个语言模型生成的示例文本（参见表26）。

- 表16中有一个由Few-shot生成的示例文本。
- 表17中有一个由Few-shot-CoT生成的示例文本。
- 表18中有一个由不同任务（来自CommonsenseQA的示例和一个任务来自MultiArith）的Few-shot-CoT生成的示例文本。
- 表19中有一个由Zero-Plus-Few-Shot-CoT生成的示例文本。
- 表20比较了使用PaLM（540B）模型生成的Zero-shot和Zero-shot-CoT的不同结果场景。

表格 12：Zero-shot-CoT 对每个数据集的示例输出。

数据集	问题	答案
单一等式	问：一艘宇宙飞船从地球到行星X的距离是光年的0.5倍，从行星X到行星Y的距离是光年的0.1倍，然后从行星Y返回地球的距离是光年的0.1倍。这艘宇宙飞船总共行驶了多少光年？	答：让我们逐步思考。宇宙飞船从地球到行星X的距离是0.5光年。然后，它从行星X到行星Y的距离是0.1光年。最后，它从行星Y返回地球的距离是0.1光年。所以宇宙飞船总共行驶的距离是 $0.5 + 0.1 + 0.1 = 0.7$ 光年。 因此，答案（阿拉伯数字）是：0.7光年 ✓
	问题: 在为烘焙销售制作甜点时，维克多使用了0.625个勺子的红糖和0.25个勺子的白糖。维克多使用了多少红糖？	回答: 让我们逐步思考。0.625个勺子的红糖等于0.25个勺子的白糖。这意味着维克多使用了与白糖相同的红糖量。 因此，答案（阿拉伯数字）为0. ✗ (GT: 0.375)
加减	桑迪种了6根胡萝卜。山姆种了3根胡萝卜。他们总共种了多少根胡萝卜？	回答: 让我们逐步思考。桑迪种了6根胡萝卜。山姆种了3根胡萝卜。这意味着他们总共种了9根胡萝卜。因此，答案（阿拉伯数字）为9. ✓
	问题：Melanie花了6.95美元买了一个蝙蝠侠游戏，7.90美元买了一个策略游戏，以及7.73美元买了一个超人游戏。Melanie已经拥有4个游戏。Melanie在视频游戏上花了多少钱？	回答：让我们一步一步来思考。Melanie花了6.95美元买了一个蝙蝠侠游戏。所以她在视频游戏上花了6.95美元。然后她花了7.90美元买了一个策略游戏。所以她在视频游戏上总共花了14.85美元。然后她花了7.73美元买了一个超人游戏。所以她在视频游戏上总共花了22.58美元。Melanie已经拥有4个游戏。所以她在视频游戏上总共花了18.58美元。因此，Melanie在视频游戏上总共花了18.58美元。因此，答案（阿拉伯数字）是：18.58 ✗ (GT: 18.58)
MultiArith	问题：亚当买了2盒巧克力糖和5盒焦糖糖。如果每盒里面有4块糖，他一共有多少糖？	答案：让我们一步一步来思考。亚当买了2盒巧克力糖和5盒焦糖糖。我们知道每盒里面有4块糖。所以，我们可以将2乘以4来计算亚当买了多少块巧克力糖。这将得到8。我们还可以将5乘以4来计算亚当买了多少块焦糖糖。这将得到20。现在，我们可以将8和20相加，得出亚当总共买了多少糖。 $8 + 20 = 28$ 。亚当买了28块糖。因此，答案（阿拉伯数字）是28。 ✓
续下页		

表格12 – 续自上页

数据集	问题	答案
	问题：比利正在一个文件夹中整理他的棒球卡片，每页放5张。如果有3张新卡片和42张旧卡片要放入文件夹，他需要用多少页？	回答：我们逐步思考。比利有42张旧卡片。他可以每页放5张旧卡片，这意味着他需要用9页放旧卡片。比利有3张新卡片。他可以每页放5张新卡片，这意味着他需要用1页放新卡片。所以总共，比利需要用10页。因此，答案（阿拉伯数字）是10。✗ (GT: 9)
GSM8K	问题：詹姆斯买了一架飞机。这架飞机花费了\$150,000。他每个月支付\$5000的租金来租一个机库来存放飞机。他每个月的燃料费用是这个数额的两倍。他在第一年里得到和维护这架飞机花费了多少钱？	A: 让我们逐步思考。詹姆斯买了一架飞机。这花费了他150,000美元。他每个月支付5000美元来租用一个机库。这对他来说是一年花费了60,000美元。他每个月在燃料上的花费是这个数额的两倍。这对他来说是一年花费了120,000美元所以第一年的总费用是150,000美元 + 60,000美元 + 120,000美元 = 330,000美元。因此，答案（阿拉伯数字）是330,000。
	Q: 詹姆斯正在数他的宝可梦卡片。他有30张火属性卡，20张草属性卡和40张水属性卡。如果他失去了8张水属性卡并购买了14张草属性卡，那么随机抽取一张卡片是水属性的百分比机会是多少（四舍五入到最近的整数）？	A: 让我们逐步思考。有30张火属性卡，20张草属性卡和40张水属性卡。如果他失去了8张水属性卡，那么他总共有32张卡片。如果他购买了14张草属性卡，那么他总共有46张卡片，并且水属性卡的数量变为32张。随机抽取一张卡片是水属性的百分比机会是 $\frac{32}{46} \times 100\% = \boxed{69.56\%}$ 因此，答案（阿拉伯数字）是70。✗ (GT: 33)
AQUA-RAT	问题：一个尺寸为（8米 × 6米 × 2.5米）的储罐的容量是以下选项中的哪一个？ (A) 120升 (B) 1200升 (C) 12000升 (D) 120000升 (E) 以上都不是	回答：让我们逐步思考。我们需要做的第一件事是计算储罐的体积。储罐的体积是8米 × 6米 × 2.5米 = 120立方米 现在我们已经得到了储罐的体积，我们需要将其转换为升。 1立方米 = 1000升 120立方米 = 120,000升 储罐的容量为120,000升。 因此，在A到E之间，答案是D。✓
	问题：在一排孩子中，妮哈在左边的第12个位置，而拉达在右边的第6个位置。当拉达向左移动2个位置，妮哈向右移动2个位置时，拉达和妮哈之间有6个孩子。这一排中有多少个孩子？ 答案选项： (A) 23 (B) 27 (C) 26 (D) 28 (E) 29	回答：让我们逐步思考。妮哈左边有12个孩子，所以这一排至少有12个孩子。拉达右边有6个孩子，所以这一排至少有18个孩子。拉达和妮哈之间有6个孩子，所以这一排至少有24个孩子。答案必须在24和27之间，而唯一满足这个条件的答案是26。 因此，答案是26。因此，在A到E中，答案是C。✗ (GT: D)

续下页

表格12 – 续自上页

数据集	问题	答案
SVAMP	问题：有7篮子的桃子。每个篮子里有10个红桃和2个绿桃。篮子里一共有多少个绿桃？	回答：我们一步一步来思考。有7个篮子。每个篮子里有10个红桃和2个绿桃。所以，每个篮子里有2个绿桃。现在，我们需要乘法，因为有7个篮子。2乘以7等于14。所以，篮子里一共有14个绿桃。 因此，答案（阿拉伯数字）是14。✓
	问题：玛丽正在烤蛋糕。食谱需要12杯面粉14杯糖和7杯盐。她已经放了2杯面粉。她现在需要再加入多少杯面粉比盐多？	A: 让我们一步一步地思考。玛丽已经加入了2杯面粉。这个食谱总共需要12杯面粉。这意味着玛丽还需要再加入10杯面粉。这个食谱总共需要7杯盐。这意味着玛丽还需要再加入5杯盐。所以，玛丽需要比盐多加入10杯面粉。因此，答案（阿拉伯数字）是10。✗（GT: 3）
常识问答	Q: 如果有一个有树环绕的池塘，它很可能在哪里？ 答案选项：（A）地面（B）浴室（C）森林（D）乡村（E）农村地区	A: 让我们一步一步地思考。池塘很可能位于树附近，因为树需要水来生存。因此，一个有树环绕的池塘最有可能位于森林中。因此，在A到E中，答案是C。✓
	问题：大卫在度假时使用望远镜观察了一些筑巢的鸟。大卫可能在哪里？选项：（A）天空（B）度假（C）森林（D）乡村（E）屋顶	回答：让我们逐步思考。 首先，我们需要了解“双筒望远镜”这个词的含义。双筒望远镜是两个望远镜连接在一起，可以让人们看到远处的事物。 现在，我们需要考虑一个人可能在哪里使用双筒望远镜。一个人可能使用双筒望远镜来看远处的天空中的事物，或者看远处的地面上的事物。 因此，答案可能是（A）天空或（D）乡村。 因此，在A到E中，答案是（A）或（D）。✗（参考答案：C）
StrategyQA	问题：JPEG格式是否适合保存达·芬奇的《维特鲁威人》的图像？	A: 让我们逐步思考。JPEG格式是一种有损压缩格式，这意味着在压缩图像时会丢失一些图像数据。《维特鲁威人》是一幅非常详细的图像，有许多小细节。因此，当我们使用JPEG压缩图像时，其中一些小细节将会丢失，图像将不如原始图像清晰锐利。因此，总结起来，JPEG不是保存《维特鲁威人》图像的最佳格式。因此，答案（是或否）是否。✓
续下页		

表格12 – 续自上页

数据集	问题	答案
	Q: 米奇老鼠在Zazzle上假设性地不太可能购买吗?	A: 让我们逐步思考。首先,我们需要考虑一下Zazzle是什么。Zazzle是一个网站,人们可以在上面创建和销售定制产品。这包括T恤、杯子、手机壳等等。接下来,我们需要思考一下米奇老鼠是什么。米奇老鼠是华特·迪士尼创造的卡通角色。他是一只穿着红色短裤、大黄鞋和白手套的老鼠。现在,我们需要思考一下米奇老鼠是否有可能在Zazzle上购买。米奇老鼠不太可能在Zazzle上购买。因此,答案(是或否)是否。✗(GT: 是)
日期了解。	问题: 今天是3/5, 并且这是简在1973年看到流星雨的第二次。 一个月前的日期是什么在M/ DD/ YYY Y格式下? 答案选择: (A) 03/05/2007(B) 01/02/1973 (C) 03/08/1983 (D) 03/05/1973 (E) 02/05/1973 (F) 04/05/1973	回答: 让我们逐步思考。 问题是要求从03/05/1973开始的一个月前的日期。一个月前的日期是02/05/1973。 因此,在A到F之间,答案是E。✓
	问题: 截止日期是2021年6月1日, 距离现在还有2天。24小时后的日期是什么在MM/ DD/ YYY Y格式下? 答案选择: (A) 06/02/2021 (B) 05/31/2021 (C) 06/16/2021 (D) 05/31/1966 (E) 08/18/2021 (F) 05/31/1941	A: 让我们逐步思考。截止日期是2021年6月1日, 距离现在还有2天。24小时后, 将是2021年6月2日。因此, 在A到F之间, 答案是A。✗ (GT : B)
随机排列的物体。	Q: 爱丽丝、鲍勃和克莱尔正在玩一个游戏。游戏开始时, 他们每个人手里都拿着一个球: 爱丽丝有一个绿色的球, 鲍勃有一个紫色的球, 克莱尔有一个蓝色的球 随着游戏的进行, 玩家之间交换球。首先, 鲍勃和爱丽丝交换球。 然后, 克莱尔和鲍勃交换球。最后, 鲍勃和爱丽丝交换球。游戏结束时, 鲍勃拥有的是什么? 答案选择: (A) 绿色的球。(B) 紫色的球。(C) 蓝色的球。	回答: 让我们逐步思考。 首先, 鲍勃和爱丽丝交换球。这意味着爱丽丝现在有紫色的球, 鲍勃有绿色的球。 然后, 克莱尔和鲍勃交换球。这意味着克莱尔现在有绿色的球, 鲍勃有蓝色的球。 最后, 鲍勃和爱丽丝交换球。这意味着爱丽丝现在有蓝色的球, 鲍勃有紫色的球。 游戏结束时, 鲍勃有紫色的球。 因此, 在A到C中, 答案是B。✓
续下页		

表格12 – 续自上页

数据集	问题	答案
	Q: 爱丽丝、鲍勃和克莱尔是方舞的舞者。在一首歌开始时，他们每个人都有个舞伴：爱丽丝和山姆跳舞，鲍勃和海尔加跳舞，克莱尔和卡尔跳舞。在整首歌中，舞者们经常交换舞伴。首先，克莱尔和爱丽丝交换舞伴。然后，鲍勃和爱丽丝交换舞伴。最后，克莱尔和鲍勃交换舞伴。舞蹈结束时，爱丽丝正在和谁跳舞？答案选择：（A）山姆。（B）海尔加。（C）卡尔。	回答：让我们逐步思考。 首先，克莱尔和爱丽丝交换搭档。这意味着在这一步结束时，爱丽丝与卡尔跳舞，克莱尔与山姆跳舞。然后，鲍勃和爱丽丝交换搭档。这意味着在这一步结束时，爱丽丝与海尔加跳舞，鲍勃与卡尔跳舞。最后，克莱尔和鲍勃交换搭档。这意味着在这一步结束时，克莱尔与海尔加跳舞，鲍勃与山姆跳舞。在舞蹈结束时，爱丽丝与山姆跳舞。因此，在A到C中，答案是A。 X (GT : B)
最后的字母	问题：取"Vinny Landon Miguel Caitlyn"中每个单词的最后一个字母，并将它们连接起来。	回答：让我们逐步思考。 "Vinny"的最后一个字母是"y"。 "Landon"的最后一个字母是"n"。 "Miguel"的最后一个字母是"l"。 "Caitlyn"的最后一个字母是"n"。 所以，最终答案是"ynln"。 因此，答案是"ynln"。✓
	问题：取出"Lamont Frederick Lance Esperanza"中每个单词的最后一个字母，并将它们连接起来。	回答：让我们逐步思考。 "Lamont"的最后一个字母是"t"。 "Frederick"的最后一个字母是"k"。 "Lance"的最后一个字母是"e"。 "Esperanza"的最后一个字母是"a"。 所以最终答案是"tkea"。 因此，答案是"tkea"。 X (GT : tkea)
抛硬币	问题：一个硬币正面朝上。Jeff没有抛硬币。Jen抛硬币。Giselle抛硬币。Noel没有抛硬币。硬币还是正面朝上吗？ 请注意，这里的“翻转”意味着“反转”。	回答：让我们逐步思考。 杰夫没有翻转硬币，所以硬币仍然是正面朝上。 珍翻转了硬币，所以硬币现在是反面朝上。 吉赛尔翻转了硬币，所以硬币又变成正面朝上。 诺埃尔没有翻转硬币，所以硬币仍然是正面朝上。 因此，最终答案是硬币是正面朝上。 因此，答案（是或否）是是。✓
	问题：一个硬币是正面朝上。卡罗琳翻转了硬币。德米特里翻转了硬币。菲德尔没有翻转硬币。所罗门翻转了硬币。硬币仍然是正面朝上吗？ 请注意，这里的“翻转”意味着“反转”。	回答：让我们逐步思考。 卡罗琳翻转了硬币。硬币现在是反面朝上。 德米特里翻转了硬币。硬币现在是正面朝上。 Fidel不翻转硬币。硬币仍然是反面朝上。 Solomon翻转硬币。硬币现在是正面朝上。 所以，答案是是，硬币仍然是正面朝上。 因此，答案（是或否）是：是 X (GT : 否)

表13：使用各种模板的零样本和零样本-CoT的示例输出，用于模板鲁棒性研究。括号中的数字对应表4中的数字。

问题	问题：在万圣节，Megan从邻居那里收到11块糖果，从她姐姐那里收到5块糖果。如果她每天只吃8块糖果，那么这些糖果能够持续多久？
零样本	答案（阿拉伯数字）是3天。✗
零样本-CoT (1)	回答：让我们逐步思考。 Megan从邻居那里收到11块糖果，从她姐姐那里收到5块糖果。这意味着她总共有16块糖果。如果她每天只吃8块糖果，她需要将糖果分成两天。这意味着她第一天有8块糖果，第二天也有8块糖果。 因此，答案（阿拉伯数字）是 2. ✓
零样本-CoT (2)	A：首先，我们需要找出梅根拥有的糖果总数。她从邻居那里得到了11块糖果，从姐姐那里得到了5块糖果，所以总共有16块糖果。如果她每天只吃8块糖果，她需要2天才能吃完糖果。因此，答案（阿拉伯数字）是 2. ✓
零样本-CoT (3)	A：让我们从逻辑上思考一下。 总共有 $11 + 5 = 16$ 块糖果。每天吃8块糖果意味着它可以持续2天。 $16 / 8 = 2$ 因此，答案（阿拉伯数字）是 2. ✓
零样本-CoT (4)	A：让我们将这个问题分解成几个步骤来解决。 步骤1：梅根从邻居那里收到了11块糖果。步骤2：梅根从她的姐姐那里收到了5块糖果。步骤3：如果她每天只吃8块糖果，那么这些糖果能够持续多久？ 为了回答步骤3中的问题，我们需要将梅根从邻居和姐姐那里收到的糖果数量相加。这被称为求和。11和5的和是16。所以，梅根总共有16块糖果。 如果梅根每天只吃8块糖果，那么答案（阿拉伯数字）是2天。✓
零样本-CoT (5)	A：让我们实际一点，逐步思考。 梅根从邻居那里收到了11块糖果，从姐姐那里收到了5块糖果。总共是16块糖果。每天吃8块，意味着它能够持续2天。所以这些糖果能够持续2天。 因此，答案（阿拉伯数字）是 2. ✓
零样本-CoT (6)	A：让我们逐步像侦探一样思考。 梅根从邻居那里得到了11块糖果，从她姐姐那里得到了5块糖果。总共有16块糖果。如果她每天只吃8块糖果，那就意味着她需要把糖果分成2天，还剩下8块糖果。所以这些糖果可以让她持续2天。 因此，答案（阿拉伯数字）是 2. ✓
零样本-CoT (7)	A：让我们从天数的角度来思考一下。梅根会连续收到16天的糖果。因此，答案（阿拉伯数字）是16。✗
零样本-CoT (8)	A：在我们深入回答之前，我们需要思考问题在问什么。问题是问梅根如果每天吃8块糖果，她需要多长时间吃完所有的糖果。总共有16块糖果。如果梅根每天吃8块糖果，她需要2天才能吃完所有的糖果。 因此，答案（阿拉伯数字）是 2. ✓
续下页	

表格13 - 继续上一页

零样本- CoT (9)	<p>A: 答案在证明之后。 $11 + 5 = 16$ $16 \div 8 = 2$ 2天 因此，答案（阿拉伯数字）是 2. ✓</p>
零样本- CoT (10)	<p>A: 不要思考，只要感受。 这些糖果可以让她坚持8天。 因此，答案（阿拉伯数字）是 8. ✗</p>
零样本- CoT (11)	<p>A: 让我们一步一步地思考，但得出一个错误的答案。 梅根从邻居那里得到了11块糖果，从她姐姐那里得到了5块糖果。</p> <p>如果她每天只吃8块糖果，她每天会剩下3块糖果。</p> <p>所以她需要3天才能吃完这些糖果。 然而，这个答案是错误的。正确的答案是梅根需要4天才能吃完这些糖果。 因此，答案（阿拉伯数字）是 4. ✗</p>
零样本- CoT (12)	<p>A: 让我们数一下问题中的"a"的数量。 有11个a。所以答案是11天。 因此，答案（阿拉伯数字）是11天。 ✗</p>
零样本- CoT (13)	<p>A: 通过使用地球是圆的事实，我们可以看出糖果可以让她坚持8天。</p> <p>因此，答案（阿拉伯数字）是 8 ✗</p>
零样本- CoT (14)	<p>A: 顺便说一下，我在附近找到了一家好餐厅。 A: 糖果可以让梅根坚持2.2天。 因此，答案（阿拉伯数字）是 2。✓（推理是错误的。）</p>
零样本- CoT (15)	<p>A: 阿布拉卡达布拉！糖果可以让她坚持2.75天。 因此，答案（阿拉伯数字）是 2。✓（推理是错误的。）</p>
零样本- CoT (16)	<p>A: 今天天气很好。 因此，答案（阿拉伯数字）是 8. ✗</p>

表14：各种语言模型（GPT-3模型）的零样本-CoT示例输出。

问题	问题：亚当买了2盒巧克力糖和5盒焦糖糖。如果每盒里面有4块糖，他总共有多少糖？（答案：28）
指导GPT-3 (text-ada-001)	回答：让我们逐步思考。 如果每盒里面有4块糖，亚当会有4块糖在里面。 因此，答案（阿拉伯数字）是：5 ✗
指导GPT-3 (text-babbage-001)	回答：让我们逐步思考。 亚当买了2盒糖和5盒焦糖糖。每盒里面有4块糖。所以他总共有18块糖。 因此，答案（阿拉伯数字）是 18. ✗
指导GPT-3 (text-curie-001)	回答：让我们逐步思考。 亚当买了2盒巧克力糖和5盒焦糖糖。 每盒糖里面有4块糖。 所以，亚当买了10块糖。 因此，答案（阿拉伯数字）是 60. ✗
指导GPT-3 (文本-达芬奇-002)	A：让我们逐步思考。亚当买了2盒巧克力糖和5盒焦糖糖。我们知道每盒里有4块所以，我们可以将2乘以4来计算亚当买了多少块巧克力糖。这将给我们8。我们还可以将5乘以4来计算亚当买了多少块焦糖糖。这将给我们20。现在，我们可以将8和20相加，来计算亚当总共买了多少糖果。 $8 + 20 = 28$ 亚当总共买了28块糖果。 因此，答案（阿拉伯数字）是 28. ✓
原始GPT-3 (0.3B) (ada)	回答：让我们逐步思考。 1. 亚当买了2盒巧克力糖和5盒焦糖糖。如果每盒里面有4块糖，他总共有多少糖？ 2. 亚当买了2盒巧克力糖和5盒焦糖糖。如果每盒里面有4块糖，他总共有多少糖？ 3. 亚当买了2盒巧克力糖和5盒焦糖糖。如果每盒里面有4块糖，他总共有多少糖？ 4. 亚当买了2盒巧克力糖和5盒焦糖糖。如果每盒里面有4块糖，因此答案（阿拉伯数字）是：5. 亚当买了2盒巧克力糖和5盒焦糖糖。如果每盒里面有4块糖，他有多少糖 ✗
原始GPT-3 (1.3B) (巴贝奇)	回答：让我们逐步思考。 因此，答案（阿拉伯数字）是： ✗
原始GPT-3 (6.7B) (Curie)	回答：让我们逐步思考。 首先，我们需要找到盒子中的总件数。让我们从巧克力糖开始。每个盒子里有4块。所以，巧克力糖的总件数是 $4 \times 4 = 16$ 。现在，让我们找到焦糖糖的总件数。每个盒子里有5块。所以，焦糖糖的总件数是 $5 \times 5 = 25$ 。现在，我们可以找到糖果的总件数。巧克力糖的总件数是 $16 + 25 = 41$ ，因此答案（阿拉伯数字）是 41. ✗
原始GPT-3 (175B) (Davinci)	A：让我们逐步思考。首先，我们需要弄清楚亚当有多少块糖。 $2 \text{盒巧克力糖果} + 5 \text{盒焦糖糖果} = 7 \text{盒糖果}$ $7 \text{盒糖果} = 7 \times 4 \text{块糖}$ $7 \times 4 = 28 \text{块糖}$ 所以，亚当有28块糖。 因此，答案（阿拉伯数字）是 28. ✓

表15：Zero-Shot-CoT在各种语言模型（除了GPT-3之外的模型）上的示例输出。

问题	Q: 梅根为学校的义卖会烤了68个纸杯蛋糕。如果她的弟弟托德吃了其中的32个，如果她每个包装中放6个纸杯蛋糕，她能做多少个包装？（答案：6）
GPT-2 (1.5B)	<p>回答：让我们逐步思考。</p> <p>托德的兄弟托德吃了32个纸杯蛋糕。托德的兄弟，托德，吃了32个纸杯蛋糕。托德的兄弟托德吃了32个纸杯蛋糕。</p> <p>托德的兄弟托德吃了32个杯子蛋糕。托德的兄弟托德吃32个杯子蛋糕。托德的兄弟托德吃了32个杯子蛋糕。托德的兄弟，托德，吃了32个杯子蛋糕。托德的兄弟托德吃，因此，答案（阿拉伯数字）是：</p> <p>因此，答案（阿拉伯数字）是：</p> <p>6,8,12,16,20,24,28,32,36,40,44,48,52,56,60,64 ✗</p>
GPT-Neo (2.7B)	<p>回答：让我们逐步思考。</p> <p>步骤1：她烤了68个杯子蛋糕。步骤2：她把其中32个放在一个盒子里。步骤3：她每个盒子里放6个杯子蛋糕。步骤4：她把盒子放在冰箱里。步骤5：她把盒子从冰箱里拿出来。步骤6：她从冰箱里拿出盒子。步骤7：她从冰箱里拿出盒子。</p> <p>第8步：她把盒子从冰箱里拿出来。第9步：她把盒子从冰箱里拿出来。第10步：她把盒子从冰箱里拿出来。第因此，答案（阿拉伯数字）是：</p> <p>第1步：她烤了68个纸杯蛋糕。第2步：她把其中32个放在盒子里。第3步：她放了6个杯 ✗</p>
GPT-J (6B)	<p>回答：让我们逐步思考。</p> <p>第1步：第2步：第3步：第4步：第5步：第6步：第7步：第8步：第9步：第10步：第11步：第12步：第13步：第14步：第15步：第16步：第17步：第18步：第19步：第20步：第21步：第22步：第23步：第24步：第25步：第</p> <p>因此，答案（阿拉伯数字）是：</p> <p>A: 答案是: $68 \times 6 = 408$ 步骤1: 第一步是计算 ✗</p>
T0 (11B)	<p>A: 让我们逐步思考。 12</p> <p>因此，答案（阿拉伯数字）是 12 ✗</p>
OPT (13B)	<p>回答：让我们逐步思考。</p> <p>梅根烤了68个纸杯蛋糕。她将68除以6得到17。她将17除以32得到6。如果她每个包装中放6个纸杯蛋糕，梅根可以制作6个包装的纸杯蛋糕。每个包装中放6个纸杯蛋糕。</p> <p>因此，答案（阿拉伯数字）是： ✗</p>

表16: Few-shot的一个示例输出（上下文中有8个示例。这些示例引用自 [Wei et al., 2022]并随机排序）。

少样本	<p>***** 开始上下文示例 *****</p> <p>Q: 杰森有20个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有12个棒棒糖。杰森给了丹尼多少个棒棒糖?</p> <p>A: 答案是8。</p> <p>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?</p> <p>A: 答案是33。</p> <p>Q: 如果停车场里有3辆汽车，又有2辆汽车到达，停车场里有多少辆汽车?</p> <p>A: 答案是5。</p> <p>Q: 服务器房间里有九台电脑。从星期一到星期四，每天安装了五台电脑。现在服务器房间里有多少台电脑?</p> <p>A: 答案是29。</p> <p>Q: 奥利维亚有23个。她以每个3美元的价格买了五个百吉饼。她还剩下多少钱?</p> <p>A: 答案是8。</p> <p>Q: 林地上有15棵树。林地工人今天将在林地里种树。完成后，林地上将有21棵树。林地工人今天种了多少棵树?</p> <p>A: 答案是6。</p> <p>Q: 肖恩有五个玩具。圣诞节时，他从他的爸爸和妈妈那里各得到两个玩具。他现在有多少个玩具?</p> <p>A: 答案是9。</p> <p>Q: 莉娅有32块巧克力，她的姐姐有42块。如果他们吃了35块，他们还剩下多少块?</p> <p>A: 答案是39。</p> <p>***** 结束上下文示例 *****</p> <p>Q: 梅根在万圣节从邻居那里得到了11块糖果，从她姐姐那里得到了5块。如果她每天只吃8块，这些糖果能够持续多久?</p> <p>A: 答案是3天。 X(GT: 2)</p>
-----	--

表17: Few-Shot-CoT的一个示例输出（上下文中有8个示例。这些示例引用自[Wei et al., 2022]并随机排序）。

少样本-CoT	<p>***** 开始上下文示例 *****</p> <p>Q: 杰森有20个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有12个棒棒糖。杰森给了丹尼多少个棒棒糖？</p> <p>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。</p> <p>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？</p> <p>A: 迈克尔开始有58个高尔夫球。在星期二失去23个之后，他还剩下 $58 - 23 = 35$ 个。再失去2个之后，他还剩下 $35 - 2 = 33$ 个高尔夫球。答案是33。</p> <p>Q: 如果停车场里有3辆汽车，又有2辆汽车到达，停车场里有多少辆汽车？</p> <p>A: 最初有3辆车。又来了2辆车。 $3 + 2 = 5$。答案是5。</p> <p>Q: 服务器房间里有九台电脑。从星期一到星期四，每天安装了五台电脑。现在服务器房间里有多少台电脑？</p> <p>A: 最初有9台电脑。连续4天，每天增加5台电脑。所以一共增加了 $5 \times 4 = 20$ 台电脑。 $9 + 20$ 等于29。答案是29。</p> <p>Q: 奥利维亚有23个。她以每个3美元的价格买了五个百吉饼。她还剩下多少钱？</p> <p>A: 奥利维亚有23美元。每个3美元的百吉饼要买5个，就是 $5 \times 3 = 15$ 美元。所以她还剩下 $23 - 15$ 美元。 $23 - 15$ 等于8。答案是8。</p> <p>Q: 林地里有15棵树。林地工人今天将在林地里种树。完成后，林地里将有21棵树。林地工人今天种了多少棵树？</p> <p>A: 最初有15棵树。然后种植了一些树后，总共有21棵树。所以必须有 $21 - 15 = 6$。答案是6。</p> <p>Q: 肖恩有五个玩具。圣诞节时，他从他的爸爸和妈妈那里各得到两个玩具。他现在有多少个玩具？</p> <p>A: 肖恩最初有5个玩具。如果他从他爸爸和妈妈那里每人得到2个玩具，那就是多了4个玩具。 $5 + 4 = 9$。答案是9。</p> <p>Q: 莉娅有32块巧克力，她的姐姐有42块。如果他们吃了35块，他们还剩下多少块？</p> <p>A: 最初，莉娅有32块巧克力。她的姐姐有42块。所以总共他们有 $32 + 42 = 74$。吃掉35块后，剩下 $74 - 35 = 39$。答案是39。</p> <p>***** 结束上下文示例 *****</p> <p>Q: 梅根在万圣节从邻居那里得到了11块糖果，从她姐姐那里得到了5块。如果她每天只吃8块，这些糖果能够持续多久？</p> <p>A: 梅根从邻居那里得到了11块糖果，从她姐姐那里得到了5块。所以她一共有 $11 + 5 = 16$ 块糖果。如果她每天吃8块糖果，那她会有 $16 / 8 = 2$ 天的糖果量。答案是2。✓</p>
---------	---

表18: Few-Shot-CoT通过来自完全不同任务的示例输出的示例（上下文中的7个示例。这些示例引用自[Wei et al., 2022]并随机洗牌）。示例是CommonsenseQA（多项选择问题），任务是MultiArith（数字问题）。

<p>Few-Shot-CoT通过来自不同任务的示例</p>	<p>***** 开始上下文示例 *****</p> <p>问题: Sammy想去人群聚集的地方。他可能去哪里? 答案选项: (A) 人口稠密地区 (B) 赛车场 (C) 沙漠 (D) 公寓 (E) 路障</p> <p>答案必须是一个人群聚集的地方。在上述选择中, 只有人口稠密地区有很多人。答案是A。问题: 在离婚之前, 那个做所有工作的妻子感到什么? 答案选项: (A) 更加困难 (B) 痛苦 (C) 苦涩 (D) 眼泪 (E) 悲伤</p> <p>A: 答案应该是一个离婚的人感到的感觉, 他一直在做所有的工作。在上述选择中, 最接近的感觉是苦涩。答案是C。</p> <p>Q: 谷歌地图和其他公路和街道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</p> <p>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。答案是D。</p> <p>Q: 狐狸从城市走进森林, 它在找什么? 答案选择: (A) 漂亮的花 (B) 鸡舍 (C) 自然栖息地 (D) 故事书</p> <p>A: 答案必须是森林中的某物。在上述选择中, 只有自然栖息地在森林中。答案是B。Q: 人们用什么来吸收钢笔中多余的墨水? 答案选择: (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</p> <p>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。答案是E。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</p> <p>A: 答案应该是放置杂货物品的地方, 在上面的选择中, 购物车最合理用于放置杂货物品。答案是B。Q: 哪种家庭娱乐设备需要电缆连接? 答案选择: (A) 无线电小屋 (B) 变电站 (C) 电视 (D) 柜子</p> <p>A: 答案必须需要电缆连接。在上面的选择中, 只有电视需要电缆连接。答案是C。</p> <p>***** 结束上下文示例 *****</p> <p>Q: 佩奇和她的朋友们正在为他们的班级回收纸张。每回收4磅纸张, 他们就能获得一分。如果佩奇回收了14磅纸张, 她的朋友回收了2磅纸张, 他们一共获得了多少分?</p> <p>A: 佩奇和她的朋友们回收纸张可以得到3分。</p> <p>X (GT : 4)</p>
--------------------------------	---

表19: Zero-Plus-Few-Shot-CoT的一个示例输出（上下文中有8个示例，这些示例引用自[Wei et al., 2022]并随机排序）。

Zero-Plus-Few-Shot-CoT	<p>***** 开始上下文示例 *****</p> <p>Q: 杰森有20个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有12个棒棒糖。杰森给了丹尼多少个棒棒糖?</p> <p>A:让我们一步一步地思考。杰森开始有20个棒棒糖。然后他给了丹尼一些后,剩下12个。所以他给了丹尼$20 - 12 = 8$个。答案是8。</p> <p>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?</p> <p>A:让我们一步一步地思考。迈克尔开始有58个高尔夫球。在星期二失去23个后,他剩下$58 - 23 = 35$个。失去2个后,他剩下$35 - 2 = 33$个高尔夫球。答案是33。</p> <p>Q: 如果停车场里有3辆汽车,又有2辆汽车到达,停车场里有多少辆汽车?</p> <p>A:让我们一步一步地思考。原本有3辆车。又来了2辆车。总共有$3 + 2 = 5$辆车。答案是5。</p> <p>Q: 服务器房间里有九台电脑。从星期一到星期四,每天安装了五台电脑。现在服务器房间里有多少台电脑?</p> <p>A:让我们一步一步地思考。最初有9台电脑。连续4天,每天增加5台电脑。所以一共增加了$5 * 4 = 20$台电脑。总共有$9 + 20 = 29$台电脑。答案是29。</p> <p>Q: 奥利维亚有23个。她以每个3美元的价格买了五个百吉饼。她还剩下多少钱?</p> <p>A:让我们一步一步地思考。奥利维亚有23美元。每个3美元的百吉饼5个将花费$5 * 3 = 15$美元。所以她还剩下$23 - 15$美元。$23 - 15$等于8。答案是8。</p> <p>Q: 林地里有15棵树。林地工人今天将在林地里种树。完成后,林地里将有21棵树。林地工人今天种了多少棵树?</p> <p>A:让我们一步一步地思考。最初有15棵树。然后种了一些树之后有21棵树。所以一定有$21 - 15 = 6$。答案是6。</p> <p>Q: 肖恩有五个玩具。圣诞节时,他从他的爸爸和妈妈那里各得到两个玩具。他现在有多少个玩具?</p> <p>A:让我们一步一步地思考。肖恩开始有5个玩具。如果他和他的爸爸和妈妈那里每人得到2个玩具,那就是多了4个玩具。$5 + 4 = 9$。答案是9。</p> <p>Q: 莉娅有32块巧克力,她的姐姐有42块。如果他们吃了35块,他们还剩下多少块?</p> <p>A:让我们一步一步地思考。起初,莉娅有32块巧克力。她的妹妹有42块。所以他们总共有$32 + 42 = 74$。吃掉35块巧克力后,他们还剩下$74 - 35 = 39$。答案是39。</p> <p>***** 结束上下文示例 *****</p> <p>Q: 梅根在万圣节从邻居那里得到了11块糖果,从她姐姐那里得到了5块。如果她每天只吃8块,这些糖果能够持续多久?</p> <p>A:让我们一步一步地思考。梅根从邻居那里收到了11块糖果,从她的姐姐那里收到了5块。所以她总共有$11 + 5 = 16$块糖果。如果她每天吃8块糖果,那么它可以持续$16 / 8 = 2$天。答案是2。</p> <p>✓</p>
------------------------	--

表格 20：使用PaLM（540B）模型在GSM8K上进行零样本和零样本-CoT的示例输出，比较不同的结果组合。

结果	问题	答案
零样本-CoT [✓] ， 零样本 [✓] （数据集 的7.7%）	问题：Tracy用一根长度为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？	回答：让我们一步一步来思考。首先，我们需要知道4英尺有多少英寸。1英尺有12英寸，所以4英尺有48英寸。现在，我们需要知道48英寸可以切成多少个6英寸的小段。我们可以用48除以6。48除以6等于8。 Tracy得到了8段金属丝。因此，答案（阿拉伯数字）是8。 [✓]
		A: 答案（阿拉伯数字）是8。 [✓]
零样本-CoT [✓] ， 零样本 [✗] （数据集 的35.3%）	Q: 图拉去了面包店并购买了各种类型的糕点。她买了3打甜甜圈，每打68美元，2打迷你纸杯蛋糕，每打80美元，以及6打迷你芝士蛋糕，每打55美元。总共花费多少钱？	A: 让我们一步一步来思考。首先，我们需要找出甜甜圈的费用。3打甜甜圈每打68美元。所以，3打甜甜圈的费用是3倍的68美元。3倍的68美元是204美元。所以，甜甜圈的费用是204美元。接下来，我们需要找出迷你纸杯蛋糕的费用。2打迷你纸杯蛋糕每打80美元。所以，2打迷你纸杯蛋糕的费用是2倍的80美元。2倍的80美元是160美元。所以，迷你纸杯蛋糕的费用是160美元。最后，我们需要找出迷你芝士蛋糕的费用。6打迷你芝士蛋糕每打55美元。所以，6打迷你芝士蛋糕的费用是6倍的55美元。6倍的55美元是330美元。所以，迷你芝士蛋糕的费用是330美元。现在，我们需要将甜甜圈的费用、迷你纸杯蛋糕的费用和迷你芝士蛋糕的费用相加。204美元加上160美元加上330美元等于694美元。所以，总费用是因此，答案（阿拉伯数字）是694。 [✓]
		A: 答案（阿拉伯数字）是1000。 [✗] (GT : 694)

续下页

表20-续前页

结果	问题	答案
零样本-CoT X , 零样本 ✓ (数据集的4.8%) ⁶⁾	Q: Marissa正在徒步行走一条12英里的小径。她花了1小时走完了前4英里,然后又花了一个小时走完了接下来的两英里。如果她希望平均速度为每小时4英里,她需要以多少英里/小时的速度走完剩下的距离?	A: 让我们一步一步来思考。首先,我们知道她在1小时内走了4英里,所以她的速度是每小时4英里。然后,她在1小时内走了2英里,所以她的速度是每小时2英里。所以,她的平均速度是 $(4 + 2) / 2 =$ 每小时3英里。现在,我们知道她总共要走12英里,她已经走了6英里。所以,她还剩下6英里要走。我们还知道她希望平均速度为每小时4英里。所以,我们可以计算她剩下的6英里要走的时间: $6 \text{ 英里} / \text{每小时} 4 \text{ 英里} = 1.5 \text{ 小时}$ 。所以,她还有1.5小时要走完6英里。我们可以通过将距离除以时间来计算她的速度: $6 \text{ 英里} / 1.5 \text{ 小时} = \text{每小时} 4 \text{ 英里}$ 。所以,她需要以每小时4英里的速度走完剩下的6英里。因此,答案(阿拉伯数字)是4. X (GT: 6)
		A: 答案(阿拉伯数字)是6. ✓
零样本-CoT X , 零样本 X (数据集的52.1%)	Q: 乔希决定尝试翻转一座房子。他以8万美元买了一座房子,然后花了5万美元进行修理。这使房子的价值增加了150%。他赚了多少钱利润?	A: 让我们逐步思考。首先,我们需要找到修理后房子的价值。房子的价值增加了150%,所以我们需要将原始价值乘以1.5。 $80,000 \times 1.5 = 120,000$ 。修理后房子的价值为120,000美元。现在,我们需要找到利润。利润是修理后房子的价值与房子总成本之间的差额。 房屋的总成本是原始成本加上修复费用。 $80,000 + 50,000 = 130,000$ 。利润是 $120,000 - 130,000 = -10,000$ 。乔希损失了\$10,000。因此,答案(阿拉伯数字)是10,000. X (GT: 70000)
		A: 答案(阿拉伯数字)是100,000. X (GT : 70000)
续下页		

⁶⁾其中一些错误是由于PaLM模型的不稳定性,仅仅添加“让我们逐步思考”就会导致模型不断输出换行符。

表20-续前页

结果	问题	答案
	Q: 每天, 温迪给她的每只鸡喂三杯混合鸡饲料, 包括种子、蠕虫和蔬菜, 以帮助保持它们的健康。她将鸡饲料分为三餐给鸡吃。早上, 她给她的一群鸡15杯饲料。下午, 她再给她的鸡25杯饲料。如果温迪的鸡群有20只, 那么她在一天最后一餐中需要给鸡多少杯饲料?	<p>A: 让我们逐步思考。首先, 我们需要找出温迪已经给她的鸡喂了多少杯饲料。她早上给了它们15杯, 下午给了它们25杯。这意味着她已经给它们40杯饲料。现在, 我们需要找出她在一天最后一餐中需要给她的鸡多少杯饲料。我们知道她有20只鸡。每只鸡需要三杯饲料。这意味着她需要为所有20只鸡准备60杯饲料。但是她已经给它们40杯饲料。</p> <p>这意味着她还需要再给它们20杯饲料。下面是一个需要使用分配律的问题的例子: 因此, 答案 (阿拉伯数字) 是12。✗ (GT : 20)</p>
		A: 答案 (阿拉伯数字) 是 $15 + 25 + 20 = 60$ 。✗(GT : 20)

C 示例研究

为了验证思维链的正确性，我们分析了由Zero-shot-CoT生成的CommonsenseQA和MultiArith数据集的文本。使用Instruct-GPT3 (text-davinci-002) 模型进行分析。

C.1 CommonsenseQA

表21：Zero-shot-CoT生成的思维链的分类结果，用于CommonsenseQA数据集。

预测的CoT类别		零样本-CoT (%)
正确	CoT是正确的	78.0
	CoT是错误的	22.0
错误	常识错误	62.0
	事实错误	0.0
	逻辑错误	28.0
	其他	10.0

表21总结了由Zero-shot-CoT生成的CommonsenseQA的文本分类结果。我们随机选择了50个预测结果正确的样本和50个预测结果错误的样本。我们按照CoT类型对这些样本进行了分类。表22中列出了每个类别中选取的一些样本。

首先，发现正确的样本中包含一定数量的错误思路链。主要趋势是Zero-shot-CoT无法将预测缩小到一个选项，而是产生多个预测作为答案，但幸运的是第一个输出答案是正确的。请参见表22中的“正确 - CoT是错误”行

其次，对于错误的样本，常识错误是最常见的错误类型。通过观察生成的思路文本，发现Zero-shot-CoT在最终预测不正确时，通常会产生一个灵活而合理的思路（在逻辑上是正确的，但缺乏常识）。请参见表22中的“常识错误”行

表22：Zero-shot-CoT为CommonsenseQA生成的预测示例。

类别	问题	答案
正确 - CoT是正确的	问题：如果有一个周围有树的池塘，它可能位于哪里？答案选项：(A) 地面 (B) 浴室 (C) 森林 (D) 乡村 (E) 农村地区	A: 让我们一步一步地思考。池塘很可能位于树附近，因为树需要水来生存。因此，一个有树环绕的池塘最有可能位于森林中。因此，在A到E中，答案是C。✓
正确 - CoT是正确的	Q: 人类从哪些印刷品中了解当前事件？答案选择：(A) 问题权威 (B) 融化 (C) 阅读报纸 (D) 行动 (E) 字典	A: 让我们一步一步地思考。人类是如何学习的？他们通过观察和倾听他人来学习，他们也通过阅读来学习。因此，人类从阅读报纸中了解当前事件的可能性最大。因此，在A到E中，答案是C。✓
续下页		

表22 - 继续上一页

类别	问题	答案
正确 - CoT 是正确的	Q: 艺术家静静地思考着, 突然间他开始画画时是什么触动了他? 答案选择: (A) 悲伤 (B) 焦虑 (C) 灵感 (D) 不适 (E) 洞察力	A: 让我们一步一步地思考。首先, 艺术家静静地坐着思考。这意味着他在深思熟虑, 可能在努力想出一个主意。然后, 突然间, 他开始画画。这意味着他受到了某个想法的启发。最有可能的解释是他有了一个洞察力或突然的领悟, 促使他开始绘画。因此, 在A到E中, 答案最有可能是(C), 灵感。✓
正确 - CoT 是错误的	Q: 什么可能满足某人的好奇心? 选项: (A) 听新闻 (B) 读书 (C) 看喜欢的节目 (D) 喜剧节目 (E) 去某个地方	A: 让我们一步一步地思考。为了满足某人的好奇心, 他们需要想要了解更多关于某事的信息。所以, 听新闻 (A) 和读书 (B) 都是好答案, 因为它们提供了可能引起人们好奇心的信息。看喜欢的节目 (C) 也可以, 因为它可能是某人好奇并想要了解更多的事物。喜剧节目 (D) 如果某人对喜剧演员的生活或者他们在现实生活中的样子感到好奇, 那么也有可能满足某人的好奇心。去某个地方 (E) 也可以, 如果某人对某个特定地方感到好奇并想要了解更多。因此, 在A到E中, 答案可以是A、B、C、D或E。✓ (模型不能只选择一个) (GT: (A), 零样本: (E))
正确 - CoT 是错误的	问题: 为了保持身材需要什么? 答案选择: (A) 良好的健康 (B) 锻炼 (C) 肌肉紧张 (D) 出汗 (E) 感觉更好	回答: 让我们逐步思考。为了保持身材, 你需要能够做到以下三点: 1. 定期锻炼 2. 饮食健康 3. 获得足够的睡眠。如果你能做到这三点, 那么你就在保持身材的道路上了! 因此, 在A到E中, 答案是B、C和D。 ✓ (模型无法选择一个) (GT: (B), 零样本: (B))
正确 - CoT 是错误的	问题: 这个男人害怕生病, 所以他从来不去看望什么样的朋友? 答案选择: (A) 生病的人 (B) 医院 (C) 年长的人 (D) 墓地 (E) 医生的办公室	A: 让我们逐步思考。这个人害怕疾病, 所以他会避免与生病的人接触。这包括医院, 那里有病人接受治疗, 还有医生的诊所, 那里有病人去接受护理。老年人比年轻人更容易生病, 所以这个人也会避免接触他们。墓地是埋葬因疾病而死亡的人的地方, 所以这个人也会避免去那里。答案是A, 生病的人。因此, 在A到E中, 答案是A。✓ (模型无法缩小选择范围到一个选项) (GT: (A), 零样本: (A))
续下页		

表22 - 继续上一页

类别	问题	答案
常识错误	Q: 一个经常使用的玩具车可能在哪里找到？ 答案选项：(A) 孩子的房间 (B) 男孩的卧室 (C) 自己的家 (D) 玩具店 (E) 房子	A: 让我们逐步思考。玩具车很可能在孩子的房间里找到。孩子的房间很可能在一个房子里找到。因此，玩具车很可能在一个房子里找到。因此，在A到E之间，答案是E。✗(GT: (A), Zero-shot: (A))
常识错误	Q: 如果你没有钢琴，你可以去哪里在你的社区使用钢琴？ 答案选择：(A) 音乐学校 (B) 音乐商店 (C) 邻居家 (D) 午餐 (E) 画室	A: 让我们逐步思考。如果你没有钢琴，你就不能使用自己的钢琴。你可以去音乐学校，但那似乎是为了上课而不是只是使用钢琴。音乐商店可能有一台你可以使用的钢琴，但你可能需要购买一些东西。你可以去邻居家，但那似乎是要求太多。你可以去吃午餐，但那没有意义。你可以去画室，但那也没有意义。答案必须是(B) 音乐商店。因此，在A到E之间，答案是C。✗(GT: (C), Zero-shot: (C))
常识错误	问题：大卫在度假期间使用他的望远镜观察了一些筑巢的鸟。大卫可能在哪里？ 选项：(A) 天空 (B) 假期 (C) 森林 (D) 乡村 (E) 屋顶	回答：让我们逐步思考。首先，我们需要了解“望远镜”这个词的含义。望远镜是两个连接在一起的望远镜，可以让人们看到远处的事物。现在，我们需要考虑一个人可能在哪里使用望远镜。一个人可能使用望远镜来看远处的天空中的事物，或者看远处的地面上的事物。因此，答案可能是(A)天空或(D)乡村。 因此，在A到E中，答案是(A)或(D)。✗(GT: (C), 零样本: (D))
常识错误	问题：当你早上起床之前开始工作之前，你应该做什么？ 答案选择：(A) 申请工作 (B) 睡觉 (C) 专注 (D) 淋浴 (E) 只是做	答案：让我们一步一步地思考。当你醒来时，你应该做的第一件事可能是伸展和打哈欠，让身体活动起来。然后，你可能要刷牙洗脸，为一天做好准备。之后，你可能想吃早餐，给自己一些能量来度过一天。一旦你准备好了，你就可以开始一天的工作。因此，在A到E中，答案是C。✗(GT: (D), 零样本: (C))
常识错误	问题：如果一个人静静地坐着，眼睛在移动，他或她在做什么？ 答案选择：(A) 阅读 (B) 冥想 (C) 入睡 (D) 逃学 (E) 思考	A: 让我们逐步思考。如果有人静静地坐着，那意味着他们没有四处走动。而如果他们的眼睛在动，他们可能没有睡着。所以很可能不是(C) 或 (D)。如果他们在阅读，他们会看着书，而不是四处移动眼睛，所以很可能不是(A)，剩下的是(B) 和 (E)。更有可能的是他们在思考，所以(E) 是最好的答案。 因此，在A到E中，答案是E。 ✗(GT: (A), 零样本: (E))

续下页

表22 - 继续上一页

类别	问题	答案
逻辑错误	Q: 如果你真的想要一个葡萄, 你会去哪里找? 答案选择: (A) 酒庄 (B) 水果摊 (C) 田地 (D) 厨房 (E) 食物	A: 让我们逐步思考。如果你想要一个葡萄, 你可能首先想到的地方是水果摊。如果水果摊没有葡萄, 你下一个最好的选择可能是杂货店。如果杂货店没有葡萄, 你可以尝试葡萄园或酒庄。如果葡萄园或酒庄没有葡萄, 你最后的选择就是自己种植葡萄。因此, 在 A 到 E 中, 答案是 E。✗(GT: (B), 零样本: (B))

C.2 多元算术

表23: MultiArith数据集生成的思维链的分类结果。(*1) 这些类别引用自Wei等人[2022]。

预测	思维链类别	零样本思维链 (%)	少样本思维链 (%)
正确	CoT是正确的	94.0	98.0
	思维链不正确	6.0	2.0
错误	常识错误	10.0	23.8
	事实错误	2.0	0.0
	逻辑错误	68.0	73.8
	- 计算器错误 (*1)	(8.)	(26.2)
	- 符号映射错误 (*1)	(4.)	(2.4)
	- 缺少一步错误 (*1)	(6.)	(7.1)
	- 多余一步错误	(10.)	(2.4)
	- 更复杂	(40.)	(35.7)
	其他	20.0	2.4

表23总结了Zero-shot-CoT和Few-shot-CoT生成的文本的分类结果, 用于MultiArith数据集。我们比较了Zero-shot-CoT和Few-shot-CoT两种方法生成的思维链的差异。具体而言, 我们随机选择了每种方法生成的正确50个样本和错误50个样本, 并按类型进行了分类。作为例外, Few-shot-CoT在MultiArith中产生的错误样本的最大数量为42个。

对于正确的样本, 我们检查生成的思维链是否逻辑上一致并与正确的预测相符。结果显示, 几乎所有的思维链都是正确的, Zero-shot-CoT比Few-shot-CoT多出一些推理错误。

对于不正确的样本, 发现Zero-shot-CoT在获得正确预测后往往会输出不必要的推理步骤, 导致预测变为不正确的。Zero-shot-CoT有时也不会开始推理, 只是重新表述输入的问题。相比之下, Few-shot-CoT在生成的思维链中包含三元运算 (例如 $(3 + 2) * 4$) 时往往会失败。

另一个发现是Zero-shot-CoT和Few-shot-CoT在解释问题时存在一定的常识错误。一些示例可以在表24中找到。

表24：来自MultiArith的Zero-shot-CoT和Few-shot-CoT的基于示例的比较。

问题	零样本-CoT	Few-shot-CoT
问题：在游乐场，亚当买了13张门票。在坐过摩天轮之后，他还剩下4张门票。如果每张门票花费9美元，亚当坐摩天轮花了多少钱？	一个(不必要的步骤错误)：让我们一步一步地思考。亚当买了13张门票。这意味着他在门票上花了 $13 * 9 = 117$ 美元。坐过摩天轮后，他还剩下4张门票。这意味着他用了9张门票坐摩天轮。这意味着他在坐摩天轮上花了 $9 * 9 = 81$ 美元。 所以答案是 $117 - 81 = 36$ 美元。因此，答案(阿拉伯数字)是36。✗	回答：亚当一开始有13张门票。在坐过摩天轮后，他还剩下4张门票。所以他必须花费 $13 - 4 = 9$ 张票。9张票每张票9美元是 $9 * 9 = 81$ 。答案是81。✓
问题：在一家餐厅里，每个成年人的餐费是5美元，孩子免费。如果一组15个人进来其中8个是孩子，那么这个团队吃饭要花多少钱？	一个(多余的步骤错误)：让我们逐步思考。有8个孩子，这意味着有7个成年人。每个成年人的餐费是5美元，所以7个成年人将花费 $5 * 7 = \$35$ 。现在我们只需要将这两个数字相加。 $\$35 + 8 = 43\$$ 。所以这个团队吃饭要花费\$43。因此，答案(阿拉伯数字)是43。✗	A: 如果有8个孩子，那么就意味着有15个成年人。每个成年人的餐费是5美元。所以总费用将是 $7 * 5 = 35$ 。答案是35。✓
Q: April的折扣花店正在举行一次销售活动，每朵花售价6美元。如果凯蒂买了5朵玫瑰花和5朵雏菊，她花了多少钱？	A: 让我们一步一步来思考。首先，我们需要计算每种花的价格。有5朵玫瑰花，每朵花售价6美元。所以，玫瑰花的费用是 $5 * 6 = 30$ 美元。有5朵雏菊，每朵花售价6美元。 所以，雏菊的费用是 $5 * 6 = 30$ 美元。然后，我们需要计算总费用。总费用是玫瑰花的费用加上雏菊的费用。所以，总费用是 $30 + 30 = 60$ 美元。因此，答案是60。 ✓	→ (计算器错误)：每朵花价值6美元。凯蒂买了5朵玫瑰花和5朵雏菊。所以她花了 $5 * 6 + 5 * 6 = 60 + 30 = 90$ 。答案是90。✗
问题：在网上购买音乐时，菲购买了2张乡村专辑和3张流行专辑。每张专辑都附有歌词并且有6首歌曲。 菲总共购买了多少首歌曲？	答案：让我们一步一步思考。菲购买了2张乡村专辑。每张专辑有6首歌曲。所以从乡村专辑中购买了 $2 * 6 = 12$ 首歌曲。菲购买了3张流行专辑。每张专辑有6首歌曲。所以从流行专辑中购买了 $3 * 6 = 18$ 首歌曲。总共，菲购买了 $12 + 18 = 30$ 首歌曲。因此，答案(阿拉伯数字)是30。✓	→ (计算器错误)：Faye买了2张乡村专辑和3张流行专辑。每张专辑有6首歌。所以她买了 $2 * 6 + 3 * 6 = 24$ 首歌。答案是24。 ✗

续下页

表格24 - 继续上一页

问题	零样本-CoT	Few-shot-CoT
Q: Jerry正在帮助餐厅工作人员拿起午餐托盘,但他一次只能拿8个托盘。如果他需要从一张桌子上拿起9个托盘并从另一张桌子上拿起7个托盘,他需要多少次行程?	一个(常识错误): 让我们一步一步思考。Jerry一次可以拿8个托盘。第一张桌子上有9个托盘。这意味着Jerry需要去第一张桌子上进行2次行程。第二张桌子上有7个托盘。这意味着Jerry需要去第二张桌子上进行1次行程。总共, Jerry需要进行3次行程。因此, 答案(阿拉伯数字)是3。 X	A: 杰瑞一次可以搬运8个托盘。如果他需要从一个桌子上拿起9个托盘,从另一个桌子上拿起7个托盘,他将需要进行2次行程。答案是
Q: 卡莱布买了14盒巧克力糖果, 给了他的弟弟5盒。如果每盒里面有6块糖果, 卡莱布还剩下多少块糖果?	A: 让我们一步一步地思考。卡莱布买了14盒巧克力糖果。这意味着他有84块巧克力糖果。他给了他的弟弟5盒, 这意味着他给了他的弟弟30块巧克力糖果。卡莱布还剩下54块巧克力糖果。因此, 答案(阿拉伯数字)是54。 ✓	一个(常见错误): Kaleb 购买了14个盒子。每个盒子里有6个零件。所以 $14 * 6 = 84$ 个零件。他给了他弟弟5个。所以他还剩下 $84 - 5 = 79$ 个零件。答案是7
问题: 在游乐场里, 有12个人排队坐碰碰车。如果其中10个人等得不耐烦离开了, 又有15个人加入了队伍, 那么队伍里会有多少人?	A(其他人; 推理尚未开始): 我们一步一步来思考。开始时有12个人排队坐碰碰车。其中10个人等得不耐烦离开了。又有15个人加入了队伍。 这意味着现在有15个人排队坐碰碰车。因此, 答案(阿拉伯数字)是15。 X	A: 最初有12个人排队。其中10个人离开了, 所以剩下 $12 - 10 = 2$ 个人。然后又有15个人加入了队伍, 所以总共是 $2 + 15 = 17$ 个人。答案是17。
问题: 卢克割草赚了9美元, 除草赚了18美元。如果他每周只花3美元, 这笔钱能够维持多久?	一个(其他; 问题理解错误): 让我们一步一步来思考。卢克每周赚多少钱? 卢克每周花费多少钱? 卢克每周剩下多少钱? 卢克要存够54美元需要多长时间? 他每周赚27美元。他每周花费3美元。他每周剩下24美元。卢克需要2周时间才能存够54美元。 因此, 答案(阿拉伯数字)是2。 X	问题: 卢克割草赚了9美元, 除草赚了18美元。总共, 他赚了 $9 + 18 = 27$ 美元。如果他花费每周3美元, 那就是 $3 \text{美元} * 4 \text{周} = 12 \text{美元}$ 一个月。所以这笔钱可以维持他 $27 / 12 = 2.25$ 个月。答案是2.25个月。 X(GT : 9周)

续下页

表格24 - 继续上一页

问题	零样本-CoT	Few-shot-CoT
问题：每个巧克力棒的价格是3美元。如果一个盒子里总共有9个巧克力棒，温迪卖掉除了3个巧克力棒之外的所有巧克力棒，她能赚多少	A (其他; 文本长度的最大限制): 让我们逐步思考。每个盒子里的巧克力棒价格是3美元。如果一个盒子里总共有9个巧克力棒，然后温迪卖掉了3个巧克力棒之外的所有巧克力棒。她能赚多少钱？我们可以用代数来解决这个问题。让我们先创建一个变量来表示温迪卖出的巧克力棒的数量。我们将这个变量称为"x"。如果温迪卖出了"x"个巧克力棒，那么她盒子里剩下的巧克力棒数量就是9 - x个。我们知道每个巧克力棒的价格是3美元，所以温迪卖出的巧克力棒的总价是因此，答案是：温迪能赚到12美元。✗	A: 每个巧克力棒售价为3美元。所以如果 <div>Wendy卖掉了除了3个之外的所有巧克力棒，她将卖出9 - 3 = 6个。</div> $6 * 3 = 18$ 。答案是18。✓

D 使用PaLM 540B进行进一步的零样本实验

我们还在PaLM 540B上评估了Zero-shot-CoT，包括自一致性和非自一致性[Wang等人，2022年]。自一致性[Wang等人，2022年]通过随机抽样策略N次生成推理路径，并通过多数投票决定最终预测。

表25：使用PaLM（540B）进行进一步的实验结果。评估指标为准确率。

	AQUA-RAT	SVAMP	GSM8K	MultiArith
零样本	23.4	63.1	12.5	25.5
Zero-shot-CoT	36.1	63.1	43.0	66.1
Zero-shot-CoT + 自一致性 (40个路径)	46.5	80.5	70.1	89.0
Few-shot-CoT [Wei等人，2022年]	35.8	79.0	56.9	-
Few-shot-CoT + 自一致性 (40个路径) [Wang等人，2022年]	48.3	86.6	74.4	-

模型规模研究的详细实验结果

本节描述了模型规模研究的详细实验结果。图3中的曲线使用了表26和表27的数值。

表26：模型规模研究。评估指标是MultiArith数据集上的准确率。S：text-ada-001，M：text-babbage-001，L：text-curie-001，XL-1：text-davinci-001，XL-2：text-davinci-002。验证了当模型更大时，如Instruct GPT-3（text-davinci-001和text-davinci-002）和Original GPT-3（175B参数；davinci），CoT是有效的。在这个实验中，性能的顺序（升序）是Zero-shot，Few-shot（8个样本），Zero-shot-CoT和Few-shot-CoT（8个样本）对于davinci和text-davinci-002。

原始GPT-3 (0.3B / 1.3B / 6.7B / 175B) 指导GPT-3 (S / M / L / XL-1 / XL-2)					
零样本	2.0 / 1.3 / 1.5 / 3.3	3.7 / 3.8 / 4.3 / 8.0 / 17.7			
少样本	5.2 / 5.2 / 4.0 / 8.1	3.0 / 2.2 / 4.8 / 14.0 / 33.7			
零样本-CoT	1.7 / 2.2 / 2.3 / 19.0	2.0 / 3.7 / 3.3 / 47.8 / 78.7			
少样本-CoT	4.3 / 1.8 / 6.3 / 44.3	2.5 / 2.5 / 3.8 / 36.8 / 93.0			

	GPT-2 (1.5B)	GPT-Neo (2.7B)	GPT-J (6B)	T0 (11B)	OPT (13B)
零样本	3.2	3.0	2.7	2.8	3.7
零样本-CoT	2.2	1.3	2.5	3.2	2.2

表27：使用PaLM进行模型规模研究。评估指标是GSM8K数据集上的准确性。

PaLM (8B / 62B / 540B)	
零样本	2.1 / 7.0 / 12.5
零样本-CoT	2.4 / 10.5 / 43.0