自动化思维链提示
在大型语言模型中

Zhuosheng Zhang[†*], **Aston Zhang**[‡], **Mu Li**[‡], **Alex Smola**[‡]
[†]上海交通大学, [‡]亚马逊网络服务

摘要

大型语言模型（LLMs）可以通过生成中间推理步骤来执行复杂的推理。 为提示演示提供这些步骤被称为思维链（CoT）提示。
CoT提示有两个主要范例。 一个利用简单的提示语“让我们逐步思考”来促进逐步思考，然后回答问题。 另一个使用一些逐个手动演示，每个演示由一个问题和一个导致答案的推理链组成。 第二个范例的卓越性能依赖于逐个手工制作特定任务的演示。我们展示了通过利用带有“让我们逐步思考”提示的LLMs来生成逐个演示的推理链，即不仅逐步思考，而且逐个思考。 然而，这些生成的链条经常出现错误。 为了减轻这些错误的影响，我们发现多样性对于自动构建演示非常重要。 我们提出了一种自动CoT提示方法：Auto-CoT。它通过多样性抽样问题并生成推理链来构建演示。

在使用GPT-3进行十个公共基准推理任务时，Auto-CoT始终能够匹配或超过需要手动设计演示的CoT范式的性能。 代码可在https://github.com/amazon-research/auto-cot找到

1 引言

大型语言模型（LLMs）[Brown et al., 2020, Thoppilan et al., 2022, Rae et al., 2021, Chowdhery et al., 2022]通过将多步骤问题分解为中间步骤，然后生成答案，在复杂推理任务上表现出色。 这种推理过程是通过一种非常新的技术引发的：思维链（CoT）提示[Wei et al., 2022a]。

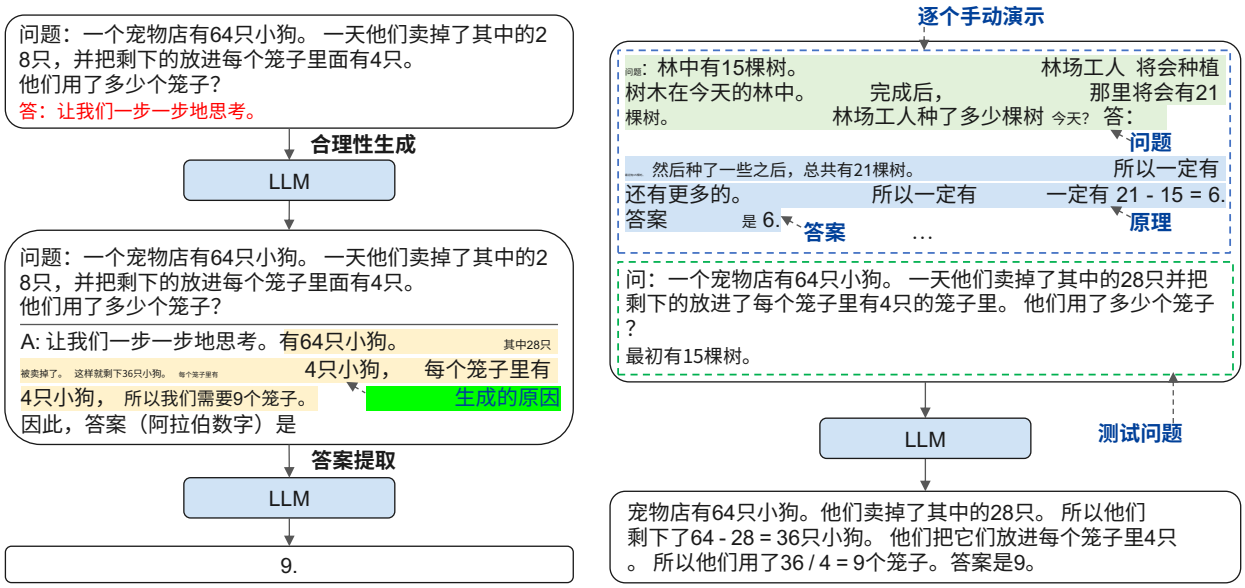
CoT提示可以分为两种主要范式。 其中一种在测试问题后添加一个类似“让我们逐步思考”的提示，以促进LLMs中的推理链[Kojima et al., 2022]。 由于这种提示范式与任务无关且不需要输入输出演示，因此被称为零射击CoT（图1左侧）。 通过零射击CoT，LLMs已经显示出良好的零射击推理能力。 另一种范式是逐个使用手动推理演示的少射击提示[Wei et al., 2022a]。 每个演示都有一个问题和一个推理链。 一个推理链由一个理由（一系列中间推理步骤）和一个预期答案组成。

由于所有演示都是手动设计的，这种范例被称为手动 **-CoT**（图1右侧）。

在实践中，手动-CoT比零射-CoT [Wei et al., 2022a, Kojima et al.,2022]获得了更强的性能。 然而，这种卓越的性能依赖于有效演示的手工草图。 具体而言，手工草图涉及到设计问题和推理链的非平凡努力，用于演示。
此外，设计任务特定演示的人力工作量更大：不同的任务，如算术
[Roy and Roth, 2015]和常识推理[Talmor et al., 2019]，需要不同的演示方式。

为了消除这种手动设计，我们提倡另一种自动-**C**oT范例，用于自动构建带有问题和推理链的演示。 具体而言，自动-**C**oT利用LLMs和“让我们逐步思考”提示来逐个生成演示的推理链，即不仅逐步思考，还逐个思考。

[*]在亚马逊网络服务实习期间完成的工作。联系人：Zhuosheng Zhang <zhangzs@sjtu.edu.cn> 和 Aston Zhang <astonz@amazon.com>



（a）零射击-CoT　　（b）手动-CoT

图1：零射击-CoT [Kojima等，2022]（使用“让我们一步一步地思考”提示）和手动-CoT [Wei等，2022a]（使用手动设计的演示一个接一个）与LLM的示例输入和输出。

然而，我们发现这个挑战不能通过简单的解决方案有效地解决。 例如，给定一个数据集的测试问题，检索语义相似的问题并调用零射击-CoT生成推理链将失败。 虽然LLMs是不错的零射击推理器，但它们并不完美：零射击-CoT仍然可能在推理链中犯错误。

为了减轻零射击-CoT的推理链错误的影响，我们的分析表明，多样性的演示问题是关键。基于这一洞察，我们提出了一种自动构建演示的Auto-CoT方法。 Auto-CoT由两个主要步骤组成。 首先，将给定数据集的问题分成几个簇。
其次，从每个簇中选择一个代表性问题，并使用零射击-CoT和简单的启发式方法生成其推理链。

我们在包括：（i）算术推理（MultiArith [Roy and Roth，2015]，GSM8K [Cobbe et al.，2021]，AQUA-RAT [Ling et al.，2017]，SVAMP [Patel et al.，2021]）；（ii）常识推理（CSQA [Talmor et al.，2019]，StrategyQA [Geva et al.，2021]）；（iii）符号推理（最后一个字母连接，抛硬币）[Wei et al.，2022a]的十个基准推理任务上评估了Auto-CoT。 实验结果表明，使用GPT-3，Auto-CoT始终与需要手动设计的Manual-CoT的性能相匹配或超过。 这表明LLMs可以通过自动构建演示来进行CoT推理。

2 相关工作

本节回顾了两个研究方向，构成了本工作的基础：用于多步推理的思维链（CoT）提示和用于从示范中引导LLM学习的上下文学习。

2.1 思维链提示

CoT提示是一种无梯度技术，用于引导LLM生成导致最终答案的中间推理步骤。魏等人[2022a]在语言模型中正式研究了CoT提示的主题。 这种技术引导LLM生成一系列连贯的中间推理步骤，导致问题的最终答案。 研究表明，LLM可以通过零-shot提示（Zero-Shot-CoT）[Kojima等，2022]或手动编写的少量示范（Manual-CoT）[魏等，2022a]进行CoT推理。

零-shot提示。Kojima等人[2022]表明，LLM是不错的零-shot推理者，其生成的解释已经反映了CoT推理。这一发现启发了我们利用自动生成的解释进行示范。 最近的一项研究[Zelikman等，2022]显示，LLM生成解释是可行的。在

他们的工作是促使LLM生成理由，并选择导致正确答案的理由。
选择需要一个带有注释答案的问题训练数据集。 相比之下，我们的工作考虑了一个更具挑战性的情景，只提供一组测试问题（没有训练数据集），这是根据Wei等人[2022a]和Kojima等人[2022]的CoT提示研究。

手动**CoT**。 通过引发**CoT**推理能力和有效的手动演示，手动**CoT**实现了更强的性能。. 推理过程的演示是手动设计的。 然而，设计问题和推理链的人力工作是非常重要的。最近的研究主要集中在手工制作更复杂的演示或利用类似集成的方法，而不是解决这个限制。 一个趋势是问题分解。 在最少到最多提示[Zh**ou**等人，2022]中，复杂问题被简化为子问题，然后按顺序解决子问题。 另一个趋势是对测试问题的多个推理路径进行投票。W**an**g等人[2022**a**]引入了一种自一致解码策略，对LL**M**的多个输出进行采样，然后对最终答案进行多数投票。W**an**g等人[2022b]和Li等人[2022]在输入空间中引入随机性，以产生更多样化的投票输出。 他们使用手动设计的演示作为种子集，并生成额外的理由：从种子集中留下一个问题，然后使用剩余的演示来为该问题生成理由。与依赖手动设计演示的前述研究线不同，我们的工作旨在通过具有竞争性能来消除手动设计。

2.2 上下文学习

CoT提示与上下文学习（ICL）密切相关[Ra dford等，2019年，Brown等，2020年]。 ICL通过将一些提示的示例作为输入的一部分，使LLM能够执行目标任务。 在没有梯度更新的情况下，ICL允许单个模型在全球范围内执行各种任务。 有多种研究线路可以改善ICL的性能：（i）检索与测试实例相关的演示，其中流行的做法是动态检索给定测试输入的相关训练示例[Rubin等，2022年，Su等，2022年]；（ii）增加细粒度信息，例如合并任务指令[Mishra等，2022年，Wei等，2022b年，Sanh等，2022年]；（iii）操作LLM的输出概率，而不是直接计算目标标签的似然性[Holtzman等，2021年，Zhao等，2021年，Min等，2022a年]。

尽管ICL取得了成功，研究[Liu et al., 2022a, Lu et al., 2022]表明，ICL的强度可能会因为上下文演示的选择而有很大差异。 具体而言，提示的格式，如措辞或演示的顺序，可能会导致性能波动[Webson and Pavlick, 2022, Zhao et al., 2021]。 最近的一项研究[Min et al., 2022b]甚至质疑了地面真实输入-输出映射的必要性：在示例中使用错误的标签只会轻微降低性能。 然而，对ICL的现有分析主要基于标准分类和多选数据集，这些数据集只有简单的<input→output>映射。 我们发现这些发现可能不适用于具有更复杂的<input→rationale→output>映射的CoT提示场景。 例如，无论是<input→rationale>映射还是<rationale→output>映射中的错误都会导致性能急剧下降（附录A.1）。

3个自动化思维链的挑战

正如刚才讨论的，ICL的性能取决于手工制作的演示。 正如在Manual-CoT [Weiet al., 2022a]中报道的，使用不同注释者编写的演示在符号推理任务中会导致高达28.2%的准确性差异，而改变演示的顺序在大多数任务中只会导致不到2%的变化。 这表明自动化思维链的关键挑战在于自动构建具有良好问题和推理链的演示。

回想一下，Manual-CoT手工制作了几个（例如8个）问题的演示。 由于基于相似性检索的方法被广泛应用于提示LLMs [Rubin et al., 2022, Su et al., 2022]，一个有希望的候选解决方案是使用基于相似性检索的方法来采样演示问题。我们遵循CoT研究中更具挑战性的假设 [Wei et al., 2022a, Kojima et al., 2022]，即只提供一组测试问题（没有训练数据集）。
根据Liu等人[2022a]的研究，我们使用Sentence-BERT [Reimers和Gurevych，2019]对问题进行编码。 对于测试数据集中的每个问题 $q$test，我们会抽样演示问题 $q_i$demo。 $(i = 1, \ldots, k)$ 从其他问题中分开。
我们设计了一种检索**-Q-CoT**方法，根据余弦相似度检索前-k个相似的问题（例如， $k = 8$）。
为了与这种基于相似度的方法进行比较，我们还测试了一种相对更多样化的方法：随机-**Q**-**C**oT，
它为每个测试问题随机抽样 k个其他测试问题。

检索-Q-CoT和随机-Q-CoT都调用了Zero-Shot-CoT [Kojima et al., 2022]来生成推理链。
c_i^{demo} 对于每个抽样的问题 q_i^{demo}，我们使用LLMs作为不错的零-shot推理器[Kojima et al., 2022]来生成理由和答案。 除非另有说明，我们使用具有175B参数的GPT-3 [Brown et al., 2020]（text-davinci-002）作为LLM。

从高层次上看，Retrieval-Q-CoT和Random-Q-CoT都采用了 q_i^{demo}, c_i^{demo}的连接 将($i = 1, \ldots, k$) 和 q^{test}作为输入，预测 q^{test}的推理链，最终包含答案（如图1右侧所示）。

令人惊讶的是，Retrieval-Q-CoT在算术数据集MultiArith [Roy and Roth, 2015]上表现不佳（表1）。 请注意，检索方法最初是在带有注释标签的任务中提出的[Rubin et al., 2022, Su et al., 2022]，然而，调用Zero-Shot-CoT并不能保证完全正确的推理链。因此，我们假设Retrieval-Q-CoT的性能较差是由于Zero-Shot-CoT生成的错误推理链。为了验证这个假设，我们在另外两个具有带有推理链注释的训练集的数据集GSM8K [Cobbe et al., 2021]和AQuA[Ling et al., 2017]上进行了Retrieval-Q-CoT的实验。 结果如表1所示，带有 †的标记。 在带有推理链注释的设置下，Retrieval-Q-CoT甚至超过了Manual-CoT。这个结果表明，在人类注释可用时，Retrieval-Q-CoT是有效的。

表1：不同采样方法的准确率（%）。 符号 †表示使用带有推理链注释的训练集。

方法	MultiArith	GSM8K	AQuA
Zero-Shot-CoT	78.7	40.7	33.5
Manual-CoT	**91.7**	46.9	35.8†
Random-Q-CoT	86.2	47.6†	36.2†
Retrieval-Q-CoT	82.8	**48.0**†	**39.7**†

尽管人工注释很有用，但这种手动工作并不简单。然而，通过零射击-CoT自动生成推理链在未解决问题采样的挑战时表现不佳。 为了设计更有效的自动-CoT，我们需要更好地了解它的挑战。

3.1 检索-Q-CoT由于相似性误导而失败

由于检索-Q-CoT使用了一些类似于手动-CoT的提示演示，因此预计检索-Q-CoT也能有竞争力。 然而，检索-Q-CoT中的推理链（包括理由和答案）是由零射击-CoT生成的：它们可能会出现导致错误答案的错误。 让我们简单地将带有错误答案的演示称为错误演示。 直观地说，当检索到与测试问题类似的问题时，由于零射击-CoT引起的错误演示可能会导致相同的LLM以类似的方式推理出错误答案（例如，复制错误）。 我们将这种现象称为相似性误导。 我们将调查相似性误导是否导致检索-Q-CoT的性能较差。

首先，我们对MultiArith数据集中的所有600个问题应用Zero-Shot-CoT。 其中，我们收集了那些Zero-Shot-CoT生成错误答案的128个问题（表示为$\mathcal{Q}$）。 正如我们所提到的，通过额外的演示，检索-$\mathcal{Q}$-CoT和随机-$\mathcal{Q}$-CoT预计能够比Zero-Shot-CoT表现更有竞争力。在Zero-Shot-CoT失败的问题中，我们将那些检索-$\mathcal{Q}$-CoT或随机-$\mathcal{Q}$-CoT仍然失败的问题称为未解决的问题。我们将未解决的问题数量除以128（$\mathcal{Q}$中的问题数量）来计算未解决率。 更高的未解决率意味着该方法更有可能像Zero-Shot-CoT一样仍然犯错误。图2显示了检索-$\mathcal{Q}$-CoT的未解决率为46.9%。

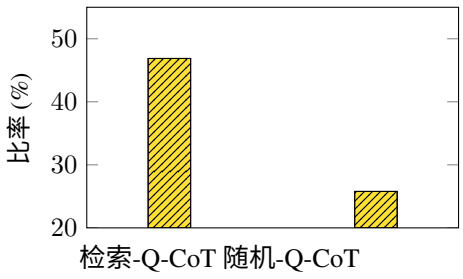


图2：未解决率。

远高于随机抽样的CoT (25.8%)。这表明，在对测试问题进行类似抽样的情况下，检索抽样的CoT受到了相似性的误导的负面影响。

为了证明检索抽样的CoT中存在未解决问题的相似性，我们在表2中提供了一个案例研究。 在左侧部分，检索到的示范问题与测试问题相似，并询问“他要花多长时间煮剩下的部分？”Zero-Shot-CoT生成的推理链回答了“总共”的问题，而不是“剩下的”问题。
在示范之后，检索抽样的CoT也因误解“剩下的”含义而失败。 相比之下，随机抽样的CoT更好地理解了“剩下的”问题，在示范中没有犯类似的错误，这要归功于相对更多样化（随机）的示范。

3.2 错误经常聚集在同一簇中

受到表[2]中的观察启发，我们使用k-means将所有600个测试问题分成k=8个簇，每个簇包含相似的问题。利用这些簇和Zero-Shot-CoT生成的推理链

[2]我们使用Sentence-BERT [Reimers and Gurevych, 2019]来编码问题，并应用 k-means进行聚类。

表2：构建演示和预测的检索-Q-CoT和随机-Q-CoT示例。 上部分包含自动构建的演示，下部分包含测试问题以及预测的推理链。

Retrieval-Q-CoT	**Random-Q-CoT**
问题：一位厨师需要煮9个土豆。 他已经煮了7个。 如果每个土豆需要3分钟煮熟，他还需要多长时间煮熟剩下的土豆? 回答：让我们一步一步来思考。 厨师已经煮了7个土豆。 这意味着他煮这7个土豆花了7 * 3分钟。 这意味着他还需要3分钟来煮熟剩下的2个土豆。 这意味着他总共需要7 * 3 + 2* 3 = 21分钟来煮熟所有9个土豆。答案是21。 ✗ **Q:**一位厨师需要煮13个土豆。 他已经煮了5个。 如果每个土豆需要煮6分钟，他还需要多长时间来煮剩下的土豆? **A:**让我们一步一步地思考。 厨师已经煮了5个土豆。 这意味着到目前为止他已经花了30分钟。总共有13个土豆。已经煮了5个，所以还剩下8个要煮。 每个土豆需要煮6分钟。 这意味着厨师需要48分钟来煮剩下的8个土豆。 总共，厨师需要78分钟来煮熟所有13个土豆。答案是78。 ✗	问题：温迪在Facebook上上传了79张照片。她把44张照片放在一个相册里，把剩下的放在5个不同的相册里。 每个相册里有多少张照片? **A:**让我们一步一步地思考。 Wendy上传了79张照片到Facebook。她把44张照片放在一个相册里。 这意味着她把剩下的照片放在了5个不同的相册里。 这意味着每个相册里会有(79-44)/5 =15张照片。答案是15。 ✓ **Q:**在一家餐厅，每份成人餐费用5美元，儿童免费。 如果一组15个人进来，其中8个是儿童，这组人的用餐费用将是多少? **A:**让我们一步一步地思考。 有8个孩子，那就意味着有7个成年人。 每个成年人的餐费是5美元，那就意味着7个成年人的费用是5美元*7=$35。 现在我们只需要把这两个数字相加。$35+8=43$。 所以这个团体的用餐费用是$43。答案是43。 ✗
Q:一位厨师需要煮15个土豆。 他已经煮了8个。 如果每个土豆需要9分钟煮熟，他还需要多长时间才能煮熟剩下的土豆? **A:**让我们一步一步地思考。 这位厨师已经煮了8个土豆。 这意味着他煮这8个土豆花了8 * 9分钟。 这意味着他需要花9分钟来煮剩下的7个土豆。 这意味着他总共需要花费8 * 9 + 7 * 9 = 135分钟来煮熟所有的15个土豆。答案是135。 ✗	**Q:**一位厨师需要煮15个土豆。 他已经煮了8个。 如果每个土豆需要9分钟煮熟，他还需要多长时间才能煮熟剩下的土豆? **A:**让我们一步一步地思考。 厨师需要煮15个土豆。 他已经煮了8个。 这意味着他还需要煮7个土豆。 每个土豆需要9分钟来煮熟。 这意味着他需要花费9 * 7 = 63分钟来煮熟剩下的土豆。答案是63。 ✓

（在第3.1节中），现在我们想知道是否存在某些簇包含Zero-Shot-CoT经常失败的问题。 因此，我们计算每个簇的错误率（错误的Zero-Shot-CoT答案/总问题数）。

如图3所示，存在一个频繁出现零射击-CoT错误（52.3%）的聚类（聚类2）。这种现象可能是普遍的，因为零射击-CoT可能缺乏解决目标任务中一些常见问题的一些技能。[3]为了方便描述，让我们将具有最高错误率的聚类称为频繁错误聚类（例如，图3中的聚类2）。 因此，以零射击方式生成的推理链的不完善性存在使用基于相似性的方法在频繁错误聚类中检索多个相似问题的风险。 对于频繁错误聚类中的测试问题，检索-Q-CoT更容易构建具有多个相似错误的演示。 因此，检索-Q-CoT经常像零射击-CoT一样犯相似的错误，这在图2中通过其更高的未解决率得到了重申。

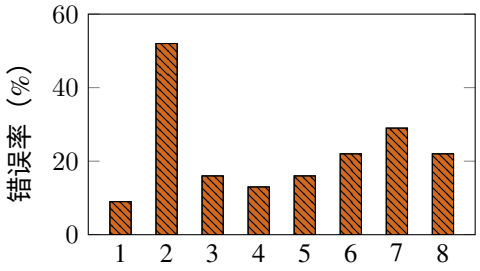


图3：相似问题的簇。

3.3 多样性可能减轻相似性误导

到目前为止的分析令人信服地表明，LLMs仍然不是完美的零射击推理器；因此，我们的目标是减轻其零射击-CoT错误的影响，特别是在Auto-CoT的设计中减轻相似性误导。正如我们稍后将展示的（第5.5节），展示少量错误（例如8个中的1或2个错误演示）不会对测试问题的整体推理性能造成伤害。 假设所有错误演示的问题都属于同一个频繁错误簇；那么从每个不同簇中抽取一个问题将导致高于$7/8 = 87.5\%$的概率构建出所有8个正确演示。 由于不同的簇反映了问题的不同语义，这种基于聚类的抽样方法可以被视为基于多样性的，与基于相似性的检索-Q-CoT形成鲜明对比。一方面，具有多样性的抽样问题可能会减轻

[3]当改变聚类数或使用其他数据集时，我们观察到类似的现象（附录A.2）。

通过相似性误导的影响（第3.1节）。另一方面，如果我们将每个演示视为一种技能，不同的演示似乎涵盖了更多解决目标问题的替代技能：即使演示中仍然存在一小部分错误（例如，$1/8$），性能也不会受到负面影响（将在图6中展示）。

然而，基于聚类的采样方法仍可能构建一小部分错误的演示，例如来自频繁错误聚类中的问题。 正如我们将在后面展示的那样，其中一些错误的演示可以通过启发式方法消除。 例如，错误的演示通常伴随着较长的问题和较长的解释。
使用简单且通用的启发式方法，例如仅考虑较短的问题和较短的解释，进一步有助于减轻不完美的零射击思维链能力的影响（附录C.2）。

4 Auto-CoT：自动化思维链提示

基于第3节中的观察和考虑，我们提出了一种自动化思维链提示（**Auto-CoT**）方法，用于自动构建带有问题和推理链的演示。 **Auto-CoT**包括两个主要阶段：（i）问题聚类：将给定数据集中的问题划分为几个聚类；（ii）演示抽样：从每个聚类中选择一个代表性问题，并使用简单的启发式方法生成其推理链。 整个过程如图4所示。

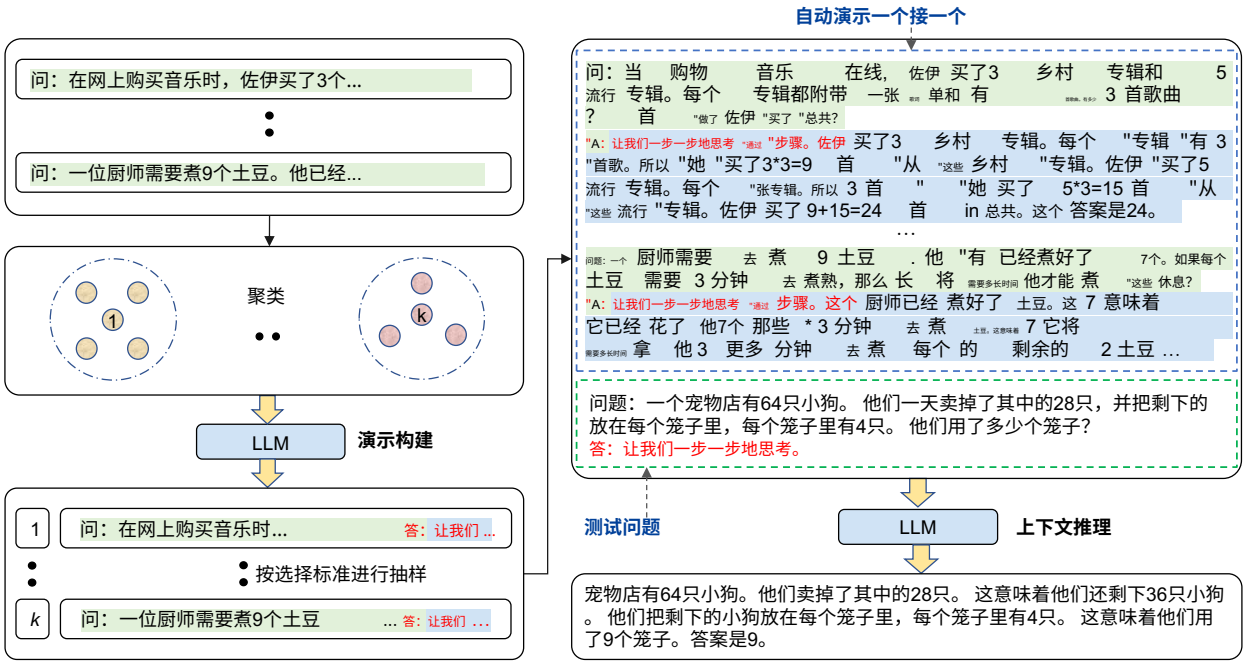


图4：Auto-CoT方法概述。 与图1中的Manual-CoT不同，演示（右侧）是使用带有“让我们逐步思考”提示的LLM自动构建的一个接一个（总共：k个）。

4.1 问题聚类

由于基于多样性的聚类可以减轻相似性导致的误导（第3.3节），我们对给定的问题集合$\mathcal{Q}$进行聚类分析。 我们首先通过Sentence-BERT [Re$i$mers和Gurevych，$_2$0$_1$9]为$\mathcal{Q}$中的每个问题计算一个向量表示。 上下文化的向量被平均以形成固定大小的问题表示。 然后，问题表示通过k-means聚类算法进行处理，以产生k个问题簇。 对于每个簇i中的问题，按照到簇中心的距离的升序将它们排序到列表$\mathbf{q}^{(i)}$=[$\mathbf{q}_{1(}{}^{i)}$, $\mathbf{q}_{2(}{}^{i)}$, ...]中。 这个问题聚类阶段在算法$_1$中总结。

4.2演示抽样

在第二阶段，我们需要为那些抽样的问题生成推理链，然后抽样满足我们选择标准的演示
。

更具体地说，我们为每个聚类^{i}（$i=1$，$\ldots, k$）构建一个演示 d（i）（问题、理由和答案的连接）。 对于聚类 i，我们在排序列表 $\mathbf{q}$（^{i}）$=[q_1$（^{i}），q_2（^{i}），$\ldots]$（通过算法$_1$获得）中迭代，直到满足我们的选择标准。 换句话说，离聚类 i中心更近的问题被认为是较早的。 假设正在考虑第 j个最接近的问题 q_j（^{i}）。 提示输入的格式为：[问题： q（^{i}）

$_j$. A: [P]]，其中 [P]是一个单一提示“让我们一步一步思考”。这个形成的输入被馈送到一个使用Zero-Shot-CoT [Ko$_{ji}$ma et al., 2022] 的LLM中，以输出由理由 $r_j^{(i)}$和提取的答案 $a_j^{(i)}$组成的推理链。然后，通过连接问题、理由和答案来构建一个候选演示 $d_j^{(i)}$，形式为 [Q: $q_j^{(i)}$, A: $r_j^{(i)} \circ a_j^{(i)}$]。

与Wei等人[2022a]中手工制作演示的标准类似，我们的选择标准遵循简单的启发式方法，以鼓励选择更简单的问题和理由：如果所选演示 $d^{(i)}$具有不超过 60个标记的问题 $q_j^{(i)}$和不超过 5个推理步骤的理由 $r_j^{(i)}$，则将其设置为所选演示 $\hat{d}_j^{(i)}$。[4]

算法 1聚类

要求：一组问题 $\mathcal{Q}$ 和演示次数 k

确保：排序后的问题 $\mathbf{q}^{(i)}=[q_1{}^{(i)}, q_2{}^{(i)}, \ldots]$对于每个聚类 i $(i=1, \ldots, k)$

过程 CLUSTER ($\mathcal{Q}$, k)
  对于每个问题 q在 $\mathcal{Q}$中执行
    通过 Sentence-BERT 对 q进行编码
  将所有编码的问题表示聚类成 k个聚类
  对于每个聚类 $i=1, \ldots, k$执行
    按照到聚类中心的距离升序排序问题 $\mathbf{q}^{(i)}=[q_1{}^{(i)}, q_2{}^{(i)}, \ldots]$
  返回 $\mathbf{q}^{(i)}$ $(i=1, \ldots, k)$

算法2构建

要求：排序的问题 $\mathbf{q}^{(i)}=[q_1{}^{(i)}, q_2{}^{(i)}, \ldots]$对于每个簇 i $(i=1, \ldots, k)$，空的演示列表 $\mathbf{d}$

确保：演示列表 $\mathbf{d}=[d^{(1)}, \ldots, d^{(k)}]$

过程 构建 ($\mathbf{q}^{(i)}, \ldots, \mathbf{q}^{(k)}$)
  对于每个簇 $i=1, \ldots, k$执行
    对于每个问题 $q_j^{(i)}$在 $\mathbf{q}^{(i)}$中执行
      使用零射击-CoT为 $q_j^{(i)}$生成理由 $r_j{}^{(i)}$和答案 $\hat{a}_j^{(i)}$
      如果$\hat{p}^{(i)}, r$如果$_j$满足选择条件则
        添加 $d^{(i)}=$ [Q: $q_j^{(i)}$, A: $r_j^{(i)} \circ$将$^{(i)}$]添加到 $\mathbf{d}$
        中断
  返回 $\mathbf{d}$

如算法2所总结的，在对所有 k聚类进行演示采样后，将构建 k个演示 $[d^{(1)}, \ldots, d^{(k)}]$. 构建的演示用于增强上下文学习的测试问题 q^{test}. 具体来说，输入是所有演示 $[d^{(1)}, \ldots, d^{(k)}]$和[Q: q^{test}. A: [P]]的连接. 将此输入提供给LLMs以获得带有答案的推理链 q^{test}（图4右侧）。

5个实验

我们简要描述了实验设置并呈现了主要实验结果。 更多实验细节和结果可以在附录中找到。

5.1 实验设置

任务和数据集。我们的方法在三类推理任务的十个基准数据集上进行评估：（i）算术推理（Mul**t**iAri**t**h [Roy **and** Ro**t**h, 2015]，GSM8K [Cobb**e** **et al.**, 2021]，A**dd**Sub [Hoss**ei**n**i** **et al.**,2014]，AQUA-RA**T** [Li**n**g **et al.**, 2017]，Si**n**gle**E**q [Ko**nce**l-**Ked**zior**ski** **et al.**, 2015]，SVAMP [**Pate**l **et al.**, 2021]）；（ii）常识推理（CSQA [**Ta**lmor **et al.**, 2019]，S**trate**gyQA [Ge**va** **et al.**, 2021]）；（iii）符号推理（最后一个字母连接，硬币翻转）[We**i** **et al.**, 2022**a**]。

实施。 我们使用公共的GPT-3 [Bro**wn** **et al.**, 2020]，版本为**te**x**t**-da**vinci**-002，参数为175B，用于LLM [Ouy**an**g **et al.**, 2022]，除非另有说明。**.** 我们选择这个LLM，因为它在公共LLM中具有最强的**C**o**T**推理性能，如K**oji**ma **et al.** [2022]和W**ei** **et al.** [2022**a**]所报道。 我们还评估C**o**d**e**x模型[Ch**en** **et al.**, 2021]（c**o**d**e**-da**vinci**-002）作为LLM。根据W**ei** **et al.** [2022**a**]的说法，除了AQuA和L**ette**r（4）以外，演示数量k为8，CSQA为7，S**trate**gyQA为6。

[4]因为Zero-Shot-CoT通常使用“\n”来分隔推理步骤，所以可以通过计算生成的解释中的“\n”标记数量来轻松实现该规则。

表3：来自三类推理任务的十个数据集的准确率。

模型	算术						常识		符号	
	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	CSQA	Strategy	Letter	Coin
零射击	22.7	12.5	77.0	22.4	78.7	58.8	72.6	54.3	0.2	53.8
Zero-Shot-CoT	78.7	40.7	74.7	33.5	78.7	63.7	64.6	54.8	57.6	91.4
少射击	33.8	15.6	83.3	24.8	82.7	65.7	**79.5**	**65.9**	0.2	57.2
Manual-CoT	91.7	46.9	81.3	35.8	86.6	68.9	73.5	65.4	59.0	97.2
自动-CoT	**92.0**	**47.9**	**84.8**	**36.5**	**87.0**	**69.5**	74.4	65.4	**59.7**	**99.9**

基线。 我们将我们的方法与四种基线方法进行比较：Zero-Shot [Kojima et al., 2022]，Zero-Shot-CoT[Kojima et al., 2022]，Few-Shot [Wei et al., 2022a]和Manual-CoT [Wei et al., 2022a]。. Zero-Shot-CoT和Manual-CoT在图1中有所说明。 Zero-Shot基线将测试问题与提示"答案是"连接起来作为LLM输入。 Few-Shot基线与Manual-CoT具有相同的LLM输入，除了从所有演示中删除了理由。

5.2 自动化思维链在十个数据集上的竞争性表现

表3比较了来自三个推理任务类别的十个数据集上的准确性。 零射和零-射-CoT的结果来自Kojima等人[2022]，少射和手动-CoT的结果来自Wei等人[2022a]，自动-CoT的结果是三次随机运行的平均值。 总体而言，自动-CoT始终能够匹配或超过需要手动设计演示的CoT范例的性能。 由于手动设计的成本，手动-CoT可能会为多个数据集设计相同的演示（例如，5/6的）。

表4：使用Codex LLM的准确性。

方法	MultiArith	GSM8K	加减
Zero-Shot-CoT	64.8	31.8	65.6
手动-CoT	**96.8**	59.4	84.6
自动-CoT	93.2	**62.8**	**91.9**

算术数据集）。相比之下，Auto-CoT更加灵活和任务自适应：每个数据集都有自己的自动构建的演示。

5.3问题聚类的可视化

图5展示了十个数据集中问题聚类的可视化结果（使用PCA投影）。 这个示例表明存在着通用模式，不同的模式可以由不同聚类中的问题来描述。 我们在附录D中提供了Auto-CoT的构建演示。

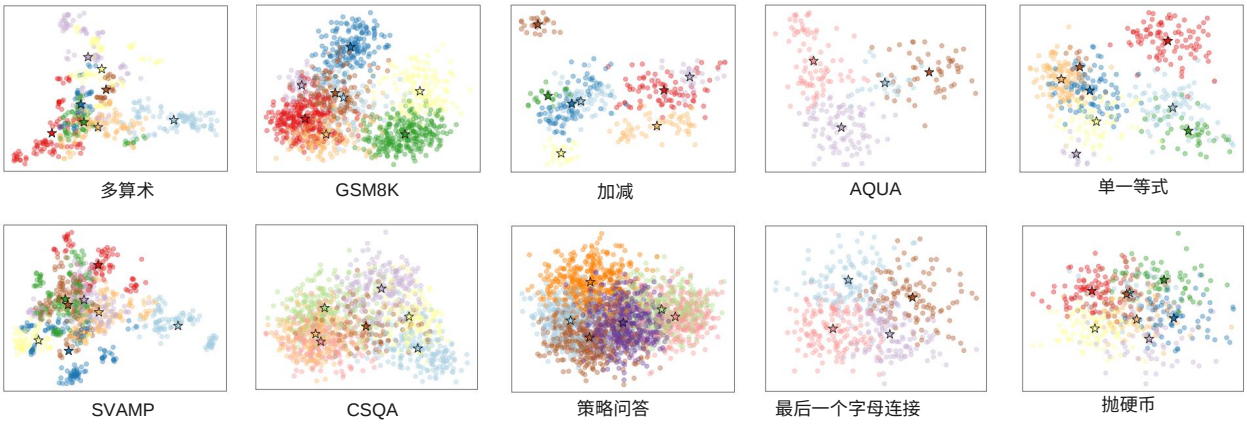


图5：十个推理任务数据集上的问题聚类。 星号表示聚类中心。

5.4使用Codex LLM的总体有效性评估

为了评估Auto-CoT在不同LLM下的总体有效性，我们将LLM更改为Codex模型[Chen et al., 2021]。 如表4所示，与使用GPT-3（text-davinci-002）LLM的表3相比，Codex LLM在Manual-CoT中导致了性能的提升。然而，使用Codex LLM，Auto-CoT的整体性能仍然与Manual-CoT相媲美，为Auto-CoT的有效性提供了额外的实证证据。

5.5 错误演示的影响

回顾我们在第3.3节中的讨论，可能存在错误的演示（答案错误）。 为了查看多样性是否可以减轻这种影响，我们设计了一个基于簇内抽样的基准线，通过从包含测试问题的同一簇中随机抽样问题来构建演示。 图6比较了在MultiArith上使用不同数量的错误演示时的准确性。与簇内抽样相比，Auto-CoT（使用基于多样性的聚类）受到错误演示的影响较小：即使提供了50%的错误演示，其性能仍然不会显著下降。

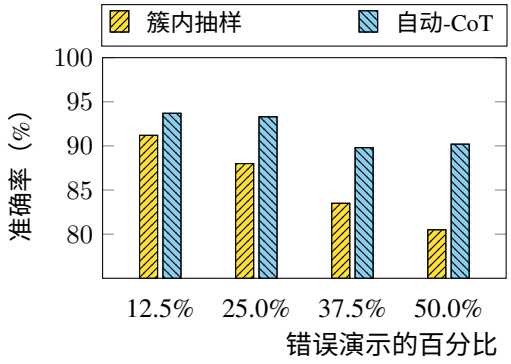


图6：错误演示的影响。

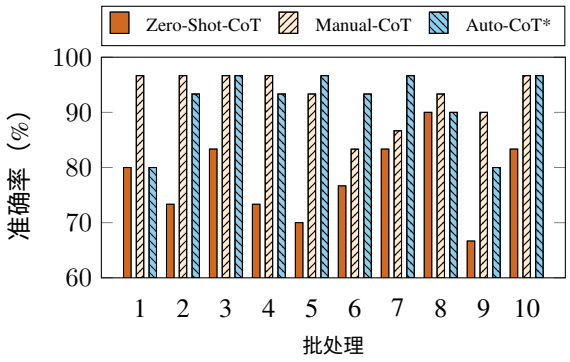


图7：流式设置的引导。

5.6 更具挑战性的流式设置

CoT研究通常假设提供了一个包含测试问题的完整数据集[Wei et al., 2022a, Kojima et al., 2022]。
根据给定的数据集，Auto-CoT会抽样问题来构建演示。 尽管如此，现在我们考虑一个更具挑战性的流式设置，其中一小批测试问题（比如m个问题）像数据流一样一次到达。

为了应对这一挑战，我们将Auto-CoT扩展为引导版本Auto-CoT*：（i）初始化一个空集合$\mathcal{M}_0$；（ii）当批次1的问题$q1$（1），...到达时，对于每个问题$q1$（1），调用零射击-CoT（由于小$_m$而无需聚类）。，$q1$（1）到q_m（1）到达时，对于每个问题$q1$（1），调用零射击-CoT（由于小$_m$而无需聚类）$q_i^{(1)}$ 获取其推理链 c（$_1$）$_i$。 添加问题链对（q_1 （$_1$），c（$_1$）$_1$），...，（q_m（$_1$），c（$_1$）$_m$）到 $\mathcal{M}_0$并调用新集合$\mathcal{M}$；（$_{iii}$）当批量 b（$b>1$）的问题 q_1 （^{b}），...，q$_m$（^{b}）到达时，使用现有问题和推理链在 $\mathcal{M}$ $_{b-}$ 1中构建演示（如Auto-CoT），并将演示用于每个 $_i$ q（^{b}）的上下文推理。 添加问题链对（q_1 （^{b}），c（$_1$ ^{b}））， ...， （q_m（^{b}）$,c$（$_{nb}$））到 $\mathcal{M}_{b-1}$并调用新集合 $\mathcal{M}_b$。

图7比较了在这种流式设置中每个批次（$m=30$）上MultiArith的准确性（扩展版本：附录中的图11）。如预期的那样，对于批次 1，Auto-CoT*和Zero-Shot-CoT获得相等的准确性。 从批次 2开始，Auto-CoT*与Manual-CoT表现相当。这个结果表明我们的方法在更具挑战性的流式设置中仍然有效。

6 结论

LLMs展示了具有CoT提示的推理能力。 Manual-CoT的卓越性能依赖于示范的手工设计。 为了消除这种手动设计，我们提出了Auto-CoT来自动构建示范。 它通过多样性抽样问题并生成推理链来构建示范。

对十个公共基准推理数据集的实验结果表明，使用GPT-3，Auto-CoT始终与需要手动设计示范的CoT范式的性能相匹配或超过。

参考文献

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan , Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom He nighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark C hen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, A lec Radford, Ilya Sutskever, and Dario Amodei. 语言模型是少样本学习者。 在Hugo Larochelle、Marc'Aurelio Ranzato、Raia Hadsell、Maria-Florina Balcan和Hsuan-Tien Lin编辑的《神经信息处理系统*33*：神经信息处理系统*2020*年年会》中，*2020*年*12*月*6*日至*12*日，虚拟，*2020*年。 网址https://proceedings.neurips.c c/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html。

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 Aguera-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。

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·阿斯拉尼德斯，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，理查德·鲍威尔，乔治·范登·德里斯切，丽莎·安妮·亨德里克斯，玛丽贝丝·劳，黄柏森，黄柏森，乔纳森·乌萨托，约翰·梅洛尔，伊琳娜·希金斯，安东尼娅·克雷斯韦尔，纳特·麦卡利斯，艾米·吴，埃里希·埃尔森，西汤·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·西蒙扬，米凯拉·帕加尼尼，洛朗·西弗雷，莱娜·马滕斯，李翔·洛林，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，安吉丽基·拉扎里杜，亚瑟·门斯，让-巴蒂斯特·莱皮奥，玛丽亚·辛普鲁凯利，尼古拉·格里戈列夫，道格·弗里茨，蒂博·索蒂奥，替博·波伦，龚志涛，丹尼尔·豆山，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，伊森·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·斯坦威，洛雷恩·贝内特，德米斯·哈萨比斯，科雷·卡武库奥卢，和杰弗里·欧文。扩展语言模型：来自训练地鼠的方法、分析和见解，2021年。 网址https://arxiv.org/abs/2

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 Sankaranarayana 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和Noah Fiedel。Palm：通过路径扩展语言模型，2022年。 网址https://arxiv.org/abs/2204.02311.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 链式思维提示在大型语言模型中引发推理。 在第三十六届神经信息处理会议（*N*eur*IPS 2022*）, *2022*a。 网址https://arxiv.org/abs/2201.11903。

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 大型语言模型是零-shot 推理器。 在第三十六届神经信息处理系统会议（*NeurIPS 2022*）, 2022。 网址https://arxiv.org/abs/2205.11916。

Subhro Roy 和 Dan Roth. 解决一般算术问题。 在*2015* 年自然语言处理实证方法会议论文集, 第*1*743-*1752*页，葡萄牙里斯本，*2015*年。 计算语言学协会。doi: *10.18653*/v*1*/D*15-1202*. 网址https://aclanthology.org/D15-1202。Alon Talmor, Jonathan Herzig, *N*icholas *L*ourie, and Jonathan Berant. *C*ommonsenseQA: 一个针对常识知识的问答挑战。 在*2019*年北美会议论文集中

计算语言学协会章节：人类语言技术，第*1*卷（长篇和
短篇），第4149-4158页，明尼阿波利斯，明尼苏达州，2019年。 计算语言学协会。doi:
10.18653/v1/N19-1421。URLhttps://aclanthology.org/N19-1421。

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
文字问题，2021年。 URLhttps://arxiv.org/abs/2110.14168。

王玲，Dani Yogatama，Chris Dyer和Phil Blunsom。通过理性生成进行程序归纳：学习解决和解释代数文字
问题。 在计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第*15*8-*1*67页，加拿大温哥华，20*17*
年。 计算语言学协会。doi：*10.18653*/v*1*/P*17-1015*。URLhttps://aclanthology.org/P17-1015。*A*rkil P
atel，Satwik Bhattamishra和Navin Goyal。NLP模型真的能解决简单的数学文字问题吗？ 在计算语言
学协会北美分会202*1*年会议论文集中：

人类语言技术，第2080-2094页，在线，2021年。 计算语言学协会。d*oi*:10.18653/v1/2021.naacl-main.168. U
RLhttps://aclanthology.org/2021.naacl-main.168.M*or* G*eva*, D*aniel* K*hasha*bi, E*la*d S*egal*, *Tusha*

r K*hot*, D*an* R*oth*, *and* J*onathan* B*erant*. Ar*istotle*使用笔记本电脑吗？ 一个具有隐含推理策略的问答基准。计算
语言学协会交易，第9卷：346-361页，2021年。 d*oi*: 10.1162/t*acl_a*_00370. UR*L*https://doi.org/10.116
2/tacl_a_00370.Er*ic* Z*elikman*, Y*uhuai* W*u*, *and* N*oah* D G*oodman*. St*ar*：用推理引导推理。*arXiv*预印本*a*

rXiv:2203.14465，2022年。 UR*L*https://arxiv.org/abs/2203.14465.

Denny Zhou、Nathanael Schärli、Le Hou、Jason Wei、Nathan Scales、Xuezhi Wang、Dale Schuurmans、Olivier Bousquet、
Quoc Le和Ed Chi。最少到最多提示使大型语言模型能够进行复杂推理。ar*Xiv*预印本
arXiv:2205.10625，2022年。 网址https://arxiv.org/abs/2205.10625。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi和Denny Zhou。自洽性改善了语言模型的思维
链推理。arXiv预印本arXiv*:2203.11171*，*2022*a年。 网址https://arxiv.org/abs/2203.11171。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi和Denny Zhou。基于理性的集成在语言模型中
的应用。arXiv预印本arXiv*:2207.00747*，*2022*b年。 网址https://arxiv.org/abs/2207.00747。Yif

ei Li、Zeqi Lin、Shizhuo Zhang、Qiang Fu、Bei Chen、Jian-Guang Lou和Weizhu Chen。关于提升语言模型推
理能力的进展。arXiv预印本arXiv*:2206.02336*，*2022*年。 网址https://arxiv.org/abs/2206.02336。

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever等。语言模型是无监督的多任
务学习者。*O*penAI博客，第9页，2019年。

Ohad Rubin，Jonathan Herzig和Jonathan Berant。学习检索上下文学习的提示。 在*2022*年北美计算语言学协
会会议论文集：人类语言技术，第*2*655-*2*671页，*2022*年。 doi：1*0*.18653/v1/*2022*.naacl-main.191。URLht
tps://aclanthology.org/2022.naacl-main.191。

Hongjin Su，Jungo Kasai，Chen Henry Wu，Weijia Shi，Tianlu Wang，Jiayi Xin，Rui Zhang，Mari Ostendorf
，Luke Zettlemoyer，Noah A Smith等。选择性注释使语言模型成为更好的少样本学习者。 arXiv预印本arX
iv*:2209.01975*，*2022*年。. URLhttps://arxiv.org/abs/2209.01975。

Swaroop Mishra，Daniel Khashabi，Chitta Baral和Hannaneh Hajishirzi。通过自然语言众包指令进行跨任务泛
化。 在计算语言学协会第*60*届年会论文集（第*1*卷：长文）中，第3470-3487页，*2022*年。 doi：*10.18653*/
v*1*/*2022*.acl-long.244。网址 https://aclanthology.org/2022.acl-long.244。

Jason Wei，Maarten Bosma，Vincent Zhao，Kelvin Guu，Adams Wei Yu，Brian Lester，Nan Du，Andrew M.
Dai和Quoc V Le。微调语言模型是零-shot学习器。 在国际学习表示会议上，2022b年。 网址https://ope
nreview.net/forum?id=gEZrGCozdqR。

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 S
harma，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，J
ason Alan Fries，Ryan Teehan，Teven Le Scao，Stella Biderman，Leo Gao，Thomas Wolf和Alexander M Rus
h。多任务提示训练实现零-shot任务泛化。 在国际学习表示会议上，2022年。 网址https://openreview.
net/forum?id=9Vrb9D0WI4。

Ari Holtzman，Peter West，Vered Shwartz，Yejin Choi和Luke Zettlemoyer。表面形式竞争：为什么最高概率答案并不总是正确的。 在*2021*年经验方法会议论文集中，第*7038-7051*页，*2021*年。 doi：*10.18*653/v*1/2021*.emnlp-main.564。URL https://aclanthology.org/2021.emnlp-main.564。

Zihao Zhao，Eric Wallace，Shi Feng，Dan Klein和Sameer Singh。使用前进行校准：提高语言模型的少样本性能。 在国际机器学习会议上，第12697-12706页，2021年。
URLhttp://proceedings.mlr.press/v139/zhao21c/zhao21c.pdf。

Sewon Min，Mike Lewis，Hannaneh Hajishirzi和Luke Zettlemoyer。噪声信道语言模型提示用于少样本文本分类。 在第*60*届计算语言学年会论文集（第*1*卷：长论文），第*5316-5330*页，*2022*年。 doi：*10.18653*/v*1/2022*.acl-long.365。URLhttps://aclanthology.org/2022.acl-long.365。

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 什么使得gpt-3的上下文示例好？ 在Deep Learning Inside *O*ut *(*DeeLI*O 2022)*会议论文集中：第3届深度学习架构知识提取与集成研讨会, 第*100-114*页, *2022*a. doi*:*10.18653/v1/2022.deelio-1.10. URLhttps://aclanthology.org/2022.deelio-1.10.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 奇妙有序的提示以及如何找到它们：克服少样本提示顺序敏感性。 在第*60*届计算语言学协会年会 *(*第*1*卷：长论文*)*, 第*8086-8098*页, *2022*. doi: 10.18653/v1/2022.acl-long.556. URLhttps://aclanthology.org/2022.acl-long.556.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 少样本参数高效微调比上下文学习更好更便宜。 ar*Xiv*预印本 ar*Xiv:2205.05638*, 2022b.. 网址https://arxiv.org/abs/2205.05638.

Albert Webson 和 Ellie Pavlick. 基于提示的模型真的理解提示的含义吗？ 在*2022*年北美计算语言学协会会议论文集*:*人类语言技术, 第*2300-2344*页, 美国西雅图, *2022*年。 计算语言学协会. doi*:* 10.18653/v1/*2022*.naacl-main.167. 网址https://aclanthology.org/2022.naacl-main.167.

Sewon Min，Xinxi Lyu，Ari Holtzman，Mikel Artetxe，Mike Lewis，Hannaneh Hajishirzi和Luke Zettlemoyer。重新思考示范的作用：什么使得上下文学习有效？ar*Xiv*预印本ar*Xiv*:*2202.12837*，2022b。 网址https://arxiv.org/abs/2202.12837。

Nils Reimers和Iryna Gurevych。句子BERT：使用Siamese BERT网络的句子嵌入。在*2019*年经验方法自然语言处理会议和第*9*届国际联合自然语言处理会议（*EMNLP-IJCNLP*）论文集，页码3982-3992，中国香港，2019年。计算语言学协会。doi：10.18653/v1/D19-1410。网址https://aclanthology.org/D19-1410。

Mohammad Javad Hosseini，Hannaneh Hajishirzi，Oren Etzioni和Nate Kushman。学习通过动词分类解决算术单词问题。 在*2014*年经验方法会议论文集（EMNL*P*）中，第*523-533*页，卡塔尔多哈，*2014*年。 计算语言学协会。doi:*10.3115*/v*1*/D*14-1058*。URLhttps://aclanthology.org/D14-1058。

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式。计算语言学协会交易，第3卷：585-597页，2015年。 doi: 10.1162/tacl_a_00160。URLhttps://aclanthology.org/Q15-1042。

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, SandhiniAgarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens,Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 使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.

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

Rik Koncel-Kedziorski，Subhro Roy，Aida Amini，Nate Kushman和Hannaneh Hajishirzi。MAWPS：一个数学问题存储库。 在*2016*年北美计算语言学协会会议论文集：人类语言技术中，页码*1152-1157*，加利福尼亚圣地亚哥，*2016*年。计算语言学协会。doi：10.18653/v1/N16-1136。URLhttps://aclanthology.org/N16-1136。

Auto-CoT挑战的扩展分析

A.1演示元素的影响。

演示是由<问题，原理，答案>组成的三元组，如图1所示。 我们随机排列演示组件之一，以查看性能如何变化。 根据表5中的结果，随机排列问题导致的性能下降最小（$91.7\% \rightarrow 73.8\%$）。 性能较好的可能原因是模型可能捕捉到了原理-答案映射模式。 这种模式预计反映了中间推理步骤如何导致最终答案。 这一发现表明，通过Zero-Shot-CoT可以利用生成的原理，因为这些原理通常在逻辑上是正确的，即使它们可能导致错误的答案[Kojima et al.，2022]。

表5：演示中不同组成部分的影响。

方法	准确性
Manual-CoT	91.7
洗牌问题	73.8
洗牌原因	43.8
洗牌答案	17.0

相比之下，洗牌原因或答案都会显著降低准确性（$91.7\% \rightarrow 43.8\%/17.0\%$）。 观察表明，原因-答案的一致性至关重要。这种不匹配实际上在Zero-Shot-CoT中发生。表6中展示了一个例子。 使用这样的演示可能会教会模型一种幻觉-在没有依据的情况下预测答案。

表6：演示中原因-答案不一致导致的错误。 第一个块显示了第二个块中测试问题所使用的演示。

问题：对于学校的烘焙义卖，温迪做了点心。 她烤了4个纸杯蛋糕和29个饼干。 销售后，她还剩下24个带回家。 她卖了多少个点心？ **A:**让我们一步一步地思考。 Wendy做了4个杯子蛋糕和29个饼干。她卖了一些糕点。 她带回家了24个。 她卖了多少糕点？我们可以建立一个方程来解决这个问题。 Wendy做了4个杯子蛋糕和29个饼干。她卖了一些糕点。 她带回家了24个。 她卖了多少糕点？我们可以建立一个方程来解决这个问题。4个杯子蛋糕+29个饼干-卖掉的糕点=24个糕点。答案是29。 ✗
Q:一位厨师需要煮13个土豆。 他已经煮了5个。 如果每个土豆需要6分钟煮熟，他还需要多长时间来煮剩下的土豆？ **A:**让我们一步一步地思考。 一位厨师需要煮15个土豆。 他已经煮了8个。 如果每个土豆需要9分钟煮熟，他还需要多长时间来煮剩下的土豆？我们可以建立一个方程来解决这个问题。 一位厨师需要煮15个土豆。 他已经煮了8个。 如果每个土豆需要9分钟煮熟，他煮完剩下的需要多长时间？我们可以建立一个方程来解决这个问题。8个土豆 + (15-8)个土豆 = 15个土豆。 答案是72分钟。 ✗

A.2 频繁错误聚类的观察

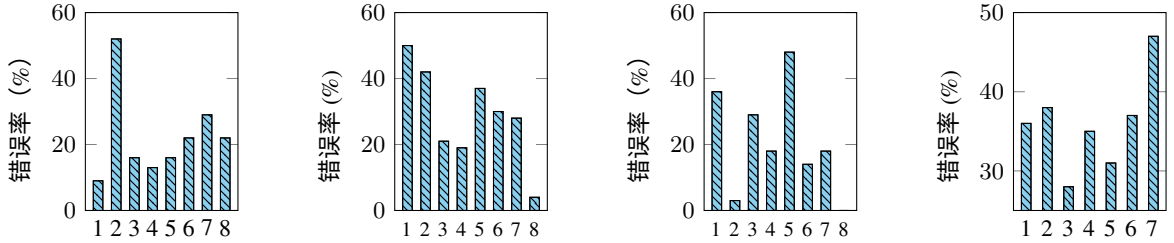
为了验证Zero-Shot-CoT是否在类似问题上失败，我们将问题聚类成几个类别，并计算每个类别中问题的错误率。 如图8所示，这些错误往往在不同数据集中聚集在一个或多个类别中。 当聚类数量发生变化时，我们观察到类似的现象，如图9所示。 这种现象已经被证明是普遍的，Zero-Shot-CoT可能缺乏解决目标任务中一些常见问题的能力。 我们将错误率最高的类别称为频繁错误聚类。

因此，生成的推理链的不完美性质存在将一组类似问题检索到频繁错误聚类中的风险。

B实验细节

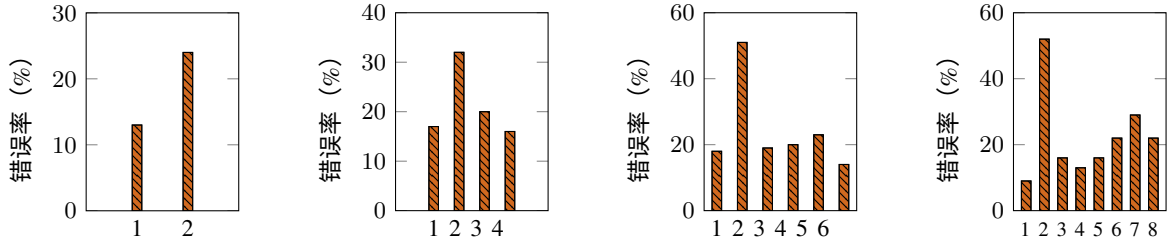
B.1任务和数据集

我们的方法在涵盖算术推理、常识推理和符号推理任务的十个基准数据集上进行评估。 数据集的统计数据如表7所示。



(a) MultiArith (Δ=43) (b) AddSub (Δ=46) (c) SingleEq (Δ=48) (d) CSQA (Δ=19)

图8：不同数据集中的问题聚类。（Δ通过最大值和最小值的差计算。



(a) 聚类数目 = 2 (b) 聚类数目 = 4 (c) 聚类数目 = 6 (d) 聚类数目 = 8

图9：不同聚类数目下的问题聚类。

算术推理。对于算术推理，我们考虑以下六个数据集：(**i**) Mul**tiArith** [**Ro**y and**Roth**, 2015]，(**ii**) GSM8K [**C**ob**e et al.**, 2021]，(**iii**) **A**ddSub [**Hosseini et al.**, 2014]，(**iv**) **AQUA** [**Ling et al.**, 2017]，(v)**Single**Eq [**Koncel-Kedzio**rs**ki et al.**, 2015]，以及 (v**i**) SV**A**MP [**Patel et al.**, 2021]。 前三个数据集来自经典的**Math Wor**ld **Prob**l**em Repos**i**tor**y [**Koncel-Kedziorski et al.**, 2016]，后三个数据集来自更近期的基准测试。

常识推理。对于常识推理，我们使用（**i**）**Commonsense**QA（**C**SQA）[**Talmor et al.**, 2019]和（**ii**）**Strategy**QA[**Geva et al.**, 2021]。 **Commonsense**QA提出了具有复杂语义的问题，通常需要基于先前知识的推理[**Talmor et al.**, 2019]。 **Strategy**QA要求模型进行隐含的多跳推理来回答问题[**Geva et al.**, 2021]。

符号推理。对于符号推理，我们使用（**i**）**Last Letter Concatenation**[**Wei et al.**, 2022**a**]和（**ii**）**Coin Flip**任务[**Wei et al.**, 2022**a**]。 **Last Letter Concatenation**要求模型将每个单词的最后一个字母连接起来。
Coin Flip的目标是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的问题。

表格 7：数据集描述。

数据集	样本数量	平均词数	答案格式	许可证
多算术	600	31.8	数量	未指定
加减	395	31.5	数量	未指定
GSM8K	1319	46.9	数量	MIT 许可证
AQUA	254	51.9	多项选择	Apache-2.0
单等式	508	27.4	数量	无许可证
SVAMP	1000	31.8	数量	MIT 许可证
CSQA	1221	27.8	多项选择	未指定
StrategyQA	2290	9.6	是或否	Apache-2.0
最后一个字母	500	15.0	字符串	未指定
抛硬币	500	37.0	是或否	未指定

B.2 实现细节

我们使用具有175B参数的GPT-3 [Brown et al., 2020] 的 text-davinci-002 版本作为 LLM [Ouyang et al., 2022]，除非另有说明。 我们选择这个模型，因为它是公开的，并且被广泛用于评估 CoT 的能力

LLM 推理 [Wei et al., 2022a, Kojima et al., 2022]。 通过 OpenAI API 访问模型。 [5]使用贪婪解码生成输出。 我们设置 max_to$k$ens = 256 和 temperature = 0。根据 Wei 等人 [2022a] 的说法，在大多数任务中，用于上下文学习的演示数量 k 为 8，AQuA 和 Last Letter Concatenation 为 4，CSQA 为 7，StrategyQA 为 6。

C分析

C.1 排序问题的比较标准

我们比较每个聚类中排序问题的不同方式，包括：（i）到聚类中心的最小距离（在Auto-CoT中采用的In-Cluster Min Dist），（ii）到聚类中心的最大距离（In-Cluster Max Dist），以及（iii）在聚类中进行随机抽样（In-Cluster Random）。为了减轻错误演示的影响，我们仅对具有正确答案的演示进行抽样分析。

表8：演示抽样的影响。

	方法	多算术
	自动-CoT	93.7
↑	In-Cluster Min Dist	93.7
	In-Cluster Random	89.2
	In-Cluster Max Dist	88.7

从表8中比较结果可以看出，如果演示与聚类中心越接近，通常效果更好。

C.2 简单启发式方法的有效性

在第4节中，我们应用简单的启发式方法来鼓励模型使用简单而准确的演示。 与魏等人[2022a]中手工制作演示的标准类似，我们的选择标准遵循简单的启发式方法，以鼓励抽样更简单的问题和解释：如果选定的演示 d（i）满足以下条件，则将其设置为候选演示：问题 q_j（i）不超过 60个标记，解释 r_j（i）不超过 5个推理步骤。[6]对于算术推理任务，除了AQuA（因为它是一个多项选择问题），我们要求答案 a（i）不为空并且出现在解释 r_j（i）中[7]以减轻解释-答案不匹配的风险（因为我们发现这样的错误是有害的，在附录A.1中）。如果问题、解释和答案满足上述条件，则构造一个候选演示 d_j（i）为第 i个聚类的问题、解释和答案的连接：[Q: q_j（i）, A: r_j（i）$\circ a$（i）]。

表格 9：演示构建的三次运行中的平均错误。

	MultiArith	AddSub	GSM8K	AQuA	SingleEq	SVAMP	CSQA	Strategy	Letter	Coin
演示数量	8	8	8	4	8	8	7	6	4	8
简单的启发式方法	0.3	1.7	1.7	1	1	0.7	2.7	2.3	0	0
无启发式方法	1.3	5	3	2.7	2	3.3	3.3	2.3	3	1

我们在使用简单的启发式方法之前和之后运行演示构建过程三次，以量化其效果。
表格 9 显示了比较结果。 简单的启发式方法减少了构建演示时的平均错误理由数量。 图 10 进一步描述了使用和不使用简单的启发式方法的错误率。 错误率通过平均错误理由数量除以演示数量来计算。 我们发现在大多数任务中，我们的方法可以将错误率保持在20%以下（$7/10$）。

[5]我们的实验在2022年7月至2022年9月期间使用OpenAI API运行。

[6]因为Zero-Shot-CoT通常使用“\n”来分隔推理步骤，所以可以通过计算生成的理由中的“\n”标记数量来轻松实现该规则。

[7]在算术推理任务中，解释通常在最后几个标记中推断出答案，如附录D中所示的演示示例。

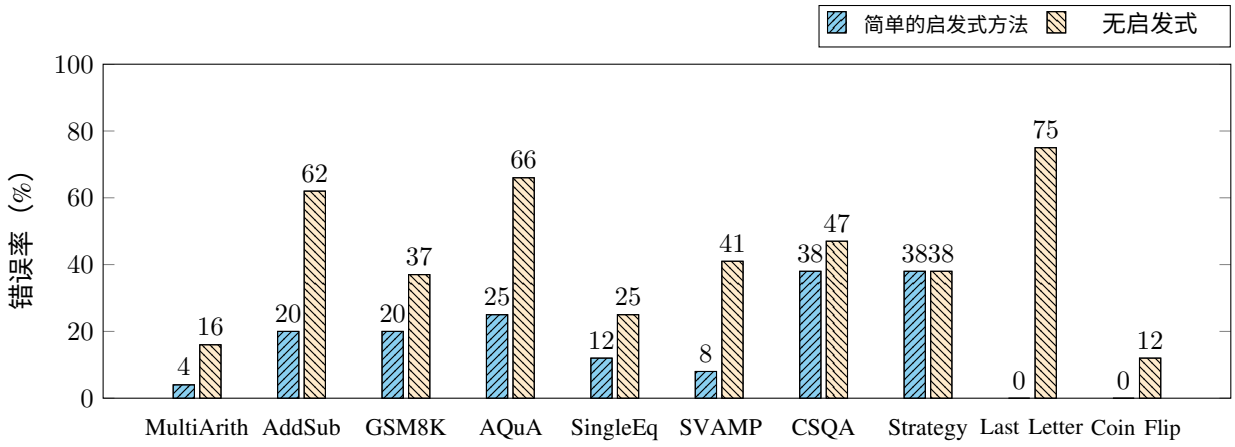


图10：抽样演示的平均错误率（%）。

C.3 扩展：更具挑战性的流设置

在第5.6节中，我们讨论了在更具挑战性的流设置中应用Auto-CoT的应用，其中一次到达一小批测试问题（例如m个问题），就像在数据流中一样。

由于页面空间限制，我们只在第5.6节中展示了前10批（总共300个测试问题）的结果。在图11中，我们展示了MultiArith中所有600个测试问题的每批准确性。如预期的那样，对于批次1，Auto-CoT*和Zero-Shot-CoT获得相等的准确性。 从第2批开始，Auto-CoT*很快与Manual-CoT表现相当。这个结果表明我们的方法在更具挑战性的流设置中仍然有效。

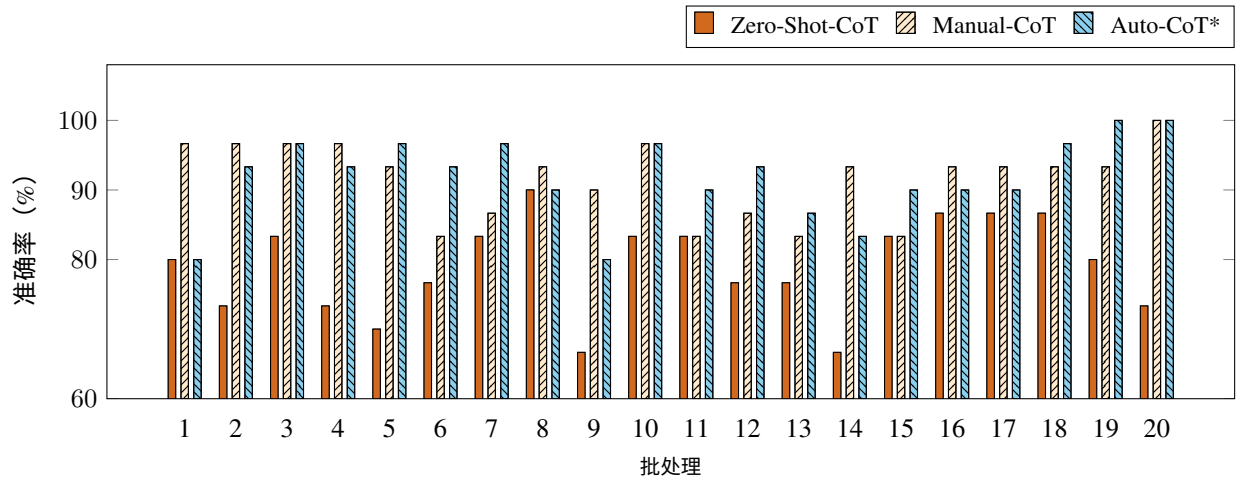


图11：流媒体设置的引导

附录D：自动构建的演示

表10：MultiArith的自动构建演示

问：在万圣节，凯蒂和她的姐姐把他们收到的糖果结合在一起。 凯蒂有8块糖果，而她的姐姐有23块。 如果他们第一天吃了8块，他们还剩下多少块？
答：让我们一步一步地思考。 凯蒂和她的姐姐总共有8 + 23 = 31块糖果。 如果他们第一天吃了8块，他们还剩下31 - 8 = 23块。答案是23。
问：一家宠物店有78只小狗。 他们一天卖掉了其中的30只，并把剩下的放进了每个笼子里有8只的笼子里。 他们用了多少个笼子？
答：让我们一步一步地思考。有78只小狗。卖掉了30只，那就意味着还剩下48只。48除以8等于6，这意味着有6个笼子，每个笼子里有8只小狗。答案是6。
Q:一个服务员有14个顾客要等候。如果有3个顾客离开，然后他又接待了39个顾客，他一共会有多少顾客？
A:让我们一步一步来思考。 这个服务员原本有14个顾客要等候。如果有3个顾客离开，那意味着他还剩下11个顾客。 如果他又接待了39个顾客，那意味着他一共有50个顾客。答案是50。
Q:比安卡正在整理书架，确保每个书架上都有8本书。如果她有5个神秘书架和4个图画书架，她一共有多少本书？
A:让我们一步一步来思考。 有5个神秘书架。每个书架上有8本书。 所以一共有40本神秘书。 有4个图画书架。每个书架上有8本书。 所以一共有32本图画书。40 + 32 = 72本书。答案是72。
问题：温迪在Facebook上上传了45张照片。她把27张照片放在一个相册里，把剩下的放在9个不同的相册里。 每个相册里有多少张照片？
回答：让我们一步一步来思考。 首先，我们知道温迪总共上传了45张照片。其次，我们知道温迪把27张照片放在一个相册里。 这意味着温迪把剩下的18张照片放在了9个不同的相册里。 这意味着每个相册里会有2张照片。答案是2。
问题：一个知识竞赛团队总共有7名成员，但在比赛中有2名成员没有出现。如果每个出现的成员得分4分，总共得分多少？
回答：让我们一步一步来思考。 团队总共有7名成员，但有2名成员没有出现。这意味着有5名成员出现了。每个出现的成员得分4分。 所以如果5名成员每人得分4分，那么总得分就是5*4=20。答案是20。
问：迈克在夏天割草赚了69美元。 如果他花了24美元买新的割草机刀片，他剩下的钱可以买多少个5美元的游戏？
答：让我们一步一步地思考。 迈克从割草中赚了69美元。 他花了24美元买新的割草机刀片。 这意味着他还剩下45美元。 每个游戏的价格是5美元，所以他可以买9个游戏。答案是9。
问：学校餐厅为学生午餐订购了8个红苹果和43个绿苹果。 但是，如果只有42个学生想要水果，餐厅最后会有多少个多余的苹果？
答：让我们一步一步地思考。 学校餐厅为学生午餐订购了8个红苹果和43个绿苹果。 这意味着他们总共订购了51个苹果。 但是，如果只有42个学生想要水果，这意味着学校餐厅会有9个多余的苹果。答案是9。

表11：用于AddSub的自动构建演示

问题：莎莉从柠檬树上摘了7个柠檬，玛丽摘了9个柠檬。 总共摘了多少个柠檬？

回答：让我们一步一步来思考。 莎莉摘了7个柠檬，玛丽摘了9个柠檬。 所以我们可以说莎莉摘了7个柠檬，玛丽摘了9个柠檬。 所以我们可以说7 + 9 = 16个柠檬总共被摘了。答案是16。

问题：戈登为班级聚会买了3.42磅的水果。 班级吃掉了2.2磅的水果。还剩下多少水果？

回答：让我们一步一步来思考。 戈登买了3.42磅的水果。 班级吃掉了2.2磅的水果。 这意味着还剩下3.42 - 2.2 = 1.22磅的水果。答案是1.22。

问题：玛丽在她的银行里有21个一角硬币和38个一分硬币。 她爸爸从玛丽那里借了18个一分硬币。 现在她有多少个一分硬币？

回答：让我们一步一步来思考。 玛丽在她的银行里有21个一角硬币和38个一分硬币。 她爸爸从玛丽那里借了18个一分硬币。所以现在，玛丽的银行里有21个一角硬币和20个一分硬币。答案是20。

问题：琼在海滩上找到了72个海贝和12个海星。 她给了艾丽莎一些她的海贝。她有28个海贝。 她给了艾丽莎多少个海贝？

回答：让我们一步一步来思考。 琼找到了72个海贝和12个海星。 她给了艾丽莎一些她的海贝。 所以我们需要找出她给了艾丽莎多少个海贝。琼有28个海贝。 所以她给了艾丽莎44个海贝。
答案是44。

问题：去年，一个国家有90171人出生，16320人移民到该国。 去年有多少新人开始在这个国家居住？

回答：让我们一步一步来思考。 首先，该国有90171人出生。 然后，16320人移民到该国。 这意味着去年有90171 + 16320 = 106491新人居住在该国。答案是106491。

问题：一家建筑公司正在重新铺设一条损坏的道路。 到目前为止，他们已经重新铺设了4938英寸的道路。 今天，他们重新铺设了805英寸的道路。 在今天之前，他们已经重新铺设了多少英寸的道路？

回答：让我们一步一步来思考。 该公司已经重新铺设了4938英寸的道路。 今天，他们重新铺设了805英寸的道路。 所以在今天之前，他们已经重新铺设了4938 - 805 = 4133英寸的道路。答案是4133。

问：抽屉里有5把剪刀和3支铅笔。 杰森把4把剪刀放在抽屉里。 现在总共有多少把剪刀？

答：让我们一步一步地思考。 抽屉里有5把剪刀和3支铅笔。 杰森把4把剪刀放在抽屉里。
现在，抽屉里有9把剪刀。答案是9。

问：公园里目前有37棵矮灌木和30棵高树。 公园工人今天将种植20棵矮灌木。
工人完成后，公园将有多少棵矮灌木？

答：让我们一步一步地思考。 公园里目前有37棵矮灌木和30棵高树。 公园工人今天将种植20棵矮灌木。 这意味着工人完成后，公园将有总共57棵矮灌木。答案是57。

表格 12：GSM8K 的自动构建演示。

Q:两列火车同时离开圣拉斐尔。 它们开始向西行驶，都行驶了80英里。 第二天，它们向北行驶，共行驶了150英里。 两天内每列火车行驶的距离是多少？

A:我们一步一步来思考。 第一天，两列火车都行驶了80英里。 所以第一列火车行驶了80英里，第二列火车也行驶了80英里。 第二天，两列火车都行驶了150英里。 所以第一列火车行驶了150英里，第二列火车也行驶了150英里。 总共，第一列火车行驶了230英里，第二列火车也行驶了230英里。答案是230。

Q:一个水箱在星期一的深度为17英尺。星期二，水箱中的水增加了7英尺。 星期三，水的深度是星期二的三分之二。水箱在星期三的水深是多少？

A:让我们一步一步地思考。 星期一，深度为17英尺。 星期二，深度比星期一多7英尺，所以深度为24英尺。 星期三，深度是星期二的三分之二，所以深度为16英尺。答案是16。

Q:在杂货店，斯隆数了100位顾客进入店内。 第二天，她比第一天多数了50位顾客。如果到第三天的总顾客数为500，那么她在第三天数了多少位顾客？

A:让我们一步一步地思考。 第一天，斯隆数了100位顾客。 第二天，她比第一天多数了50位顾客。所以，她在第二天数了150位顾客。第三天，她数了500位顾客。 因此，第三天，斯隆数了500位顾客。答案是500。

Q:劳雷尔的朋友给了她24件她的孩子不再需要的婴儿服装。 在她的婴儿派对上，劳雷尔收到了两倍于新婴儿服装的数量。 然后，劳雷尔的妈妈又送给她另外15件婴儿服装。 她为她的孩子有多少套衣服？

A:让我们一步一步地思考。 劳雷尔的朋友给了她24件婴儿服装。 所以她有24件婴儿服装。 在她的婴儿派对上，劳雷尔收到了两倍于新婴儿服装的数量。 所以她有48件婴儿服装。 然后，劳雷尔的妈妈又送给她另外15件婴儿服装。 所以她有63件婴儿服装。答案是63。

Q:帕姆和弗雷德去了一个嘉年华会。帕姆坐过了过山车2次，而弗雷德坐过了4次。 之后，他们决定各自再坐雪橇2次。 如果每次乘坐需要6张票，他们那天用了多少张票？

A:让我们一步一步地思考。 帕姆坐过了过山车2次。 所以，那就是2 * 6 = 12张票。 弗雷德坐过了过山车4次。 所以，那就是4 * 6 = 24张票。 帕姆滑过了雪橇2次。 所以，那就是2 * 6 = 12张票。 弗雷德滑过了雪橇2次。 所以，那就是2 * 6 = 12张票。 总共，他们使用了12 + 24 + 12 + 12 = 60张票。答案是60。

Q:八年级班举办了一场烘焙比赛。 凯尔西做的饼干是乔希的两倍。乔希做的饼干是苏珊娜做的四分之一。 如果苏珊娜做了36块饼干，凯尔西做了多少块？

A:让我们一步一步地思考。 乔希做的饼干是苏珊娜做的四分之一。 所以如果苏珊娜做了36块饼干，乔希做了9块饼干。 凯尔西做的饼干是乔希的两倍。所以如果乔希做了9块饼干，凯尔西做了18块饼干。答案是18。

Q:简过去8周每周都存了$4。 如果她继续每周存同样的金额，还需要多少周才能存到总共$60？

A:我们一步一步来思考。 简过去8周每周都存了$4。 这意味着她到目前为止已经存了$32。$60 - $32 = $28 $28 / $4 = 7 简还需要再存7周才能存到总共$60。
答案是7。

Q:米娅和艾玛目前年龄相差16岁。如果米娅比艾玛年轻，那么米娅现在40岁，他们的年龄平均值是多少？

A:我们一步一步来思考。 米娅目前16岁。米娅比艾玛年轻。米娅现在40岁。因此，艾玛必须是40 + 16 = 56岁。他们的年龄平均值是(40 + 56)/2 = 48岁。答案是48。

表格 13：为AQuA自动构建的演示。

Q：朱莉的院子是长方形的。 院子的一边宽100英尺。 院子的总面积是3,000平方英尺。 院子的另一边长度是多少？ 答案选择：（A）30英尺（B）20英尺（C）10英尺（D）50英尺（E）60英尺
A：让我们一步一步地思考。 我们知道院子是长方形的，所以它有四条边。 我们还知道院子的一边宽100英尺。 院子的总面积是3,000平方英尺。 我们可以建立方程：3,000 = 100x x = 30因此，院子的另一边长度是30英尺。答案是A。
Q：在一个装有红色和绿色糖果的袋子里，红色糖果与绿色糖果的比例是3:4。 如果袋子里有120颗绿色糖果，那么有多少颗红色糖果？ 答案选择：（A）90（B）80（C）95（D）100（E）85A：让我们
一步一步地思考。 每4颗绿色糖果中有3颗红色糖果。 这意味着每12颗糖果中有9颗红色糖果。 这意味着每120颗绿色糖果中有90颗红色糖果。答案是（A）。答案是A。
Q:一辆汽车以60公里/小时的速度完成一次旅程，用时20小时。 如果要在10小时内完成相同的距离，汽车需要增加多少速度？ 选项：（A）80公里/小时（B）50公里/小时（C）120公里/小时（D）70公里/小时（E）80公里/小时
A:让我们一步一步地思考。 汽车目前以60公里/小时的速度行驶。 为了在10小时内行驶相同的距离，汽车需要以6倍的速度行驶，即360公里/小时。 因此，汽车需要增加300公里/小时的速度。答案是C。答案是C。
Q:马克告诉约翰：“如果你给我一半的钱，我会有75卢比。约翰说：“如果你给我三分之一的钱，我会有75卢比。”约翰有多少钱？ 选项：（A）22（B）60（C）28（D）26（E）18
A:让我们一步一步地思考。 如果马克有75卢比，那么约翰也有75卢比。 如果马克把他一半的钱给约翰，那么马克会有37.50卢比，约翰会有112.50卢比。然而，这不是其中一个答案选择。 如果马克把他三分之一的钱给约翰，那么马克会有50卢比，约翰会有100卢比。这是其中一个答案选择，所以约翰一开始就有100卢比。答案是D。

表14：SingleEq的自动构建演示。

Q:马克的父亲给了他85美元。 马克买了10本书，每本书5美元。 马克还剩下多少钱?

A:让我们一步一步地思考。 马克的父亲给了他85美元。 马克买了10本书，每本书5美元。 这意味着马克花了50美元买书。 所以马克还剩下85美元 - 50美元 = 35美元。答案是35。

问题：蒂姆有44本书。山姆有52本书。 他们一共有多少本书?

回答：让我们一步一步来思考。 步骤1：蒂姆有多少本书? 步骤2：山姆有多少本书?
步骤3：他们一共有多少本书？蒂姆有44本书。山姆有52本书。 他们一共有96本书。答案是96。

问题：公园里目前有22棵胡桃树。 公园工人今天将种更多的胡桃树。 当工人们完成后，公园里将有55棵胡桃树。 工人们今天种了多少棵胡桃树?

回答：让我们一步一步来思考。 公园里目前有22棵胡桃树。 工人们今天将种更多的胡桃树。 当工人们完成后，公园里将有55棵胡桃树。 工人们今天种了33棵胡桃树。答案是33。

Q:Karen在一批混合坚果中添加了0.25杯核桃。后来，她又添加了0.25杯杏仁。 Karen总共放了多少杯坚果在混合坚果中?

A:让我们一步一步地思考。 Karen在一批混合坚果中添加了0.25杯核桃。这是第一步，到目前为止，她已经添加了0.25杯核桃。 后来，她添加了0.25杯杏仁。 这是第二步，到目前为止，她已经添加了0.5杯坚果（0.25杯核桃+0.25杯杏仁）。 因此，总共，Karen在混合坚果中添加了0.5杯坚果。答案是0.5。

Q:莎拉、基思、本尼和艾莉莎每人有96张棒球卡。 他们总共有多少打棒球卡?

A:让我们一步一步地思考。 一打有12打，所以我们需要将96除以12。96/12=8。因此，他们总共有8打棒球卡。答案是8。

Q:克雷格从学校走了0.2英里到达大卫的家，然后从大卫的家走了0.7英里到达自己的家。
克雷格总共走了多少英里?

A:让我们一步一步地思考。 克雷格从学校走了0.2英里到达大卫的家。 然后，克雷格从大卫的家走了0.7英里到达自己的家。 总共，克雷格走了0.2 + 0.7 = 0.9英里。答案是0.9。

Q:汤姆在海滩上找到了5个海贝。 他给了杰西卡2个海贝。 他现在有多少个海贝?

A:让我们一步一步地思考。汤姆找到了5个海贝。 汤姆给了杰西卡2个海贝。 汤姆现在有3个海贝。答案是3。

Q:布伦达的妈妈为5个人做了饼干。 如果她准备了35块饼干，并且每个人都有相同数量的饼干，那么每个人有多少块饼干?

A:让我们一步一步地思考。 布伦达的妈妈做了35块饼干。 我们知道她为5个人做了这些饼干。 那意味着她为每个人做了7块饼干。答案是7。

表15：自动构建的SVAMP演示。

Q:马可和他爸爸去采草莓。 马可爸爸的草莓重11磅。 如果他们的草莓一起重30磅。 马可的草莓重多少磅?

A:马可爸爸的草莓重11磅。 如果他们的草莓一起重30磅，那么马可的草莓重19磅。答案是19。

问题：篮子里有19个红桃，11个黄桃和12个绿桃。 篮子里有多少个红桃比黄桃多?

回答：篮子里有19个红桃和11个黄桃。19-11=8 红桃比黄桃多8个。
答案是8。

问题：面包师傅做了144个蛋糕。他卖掉了其中的71个。 然后他又做了111个蛋糕。 面包师傅做了比卖掉的蛋糕多多少个?

回答：首先，他做了144个蛋糕，卖掉了其中的71个。 所以，他还剩下73个蛋糕。 然后，他又做了111个蛋糕。 所以，他总共有73 + 111 = 184个蛋糕。现在，我们需要找出他做了比卖掉的蛋糕多多少个。总共有184个蛋糕。他卖掉了其中的71个。 所以，他做了184-71 = 113个比卖掉的蛋糕多。答案是113。

问题：布莱恩也看了看他的书。 如果他总共有42本书，每个书架上有2本书，他有多少个书架?

答案：我们知道布莱恩总共有42本书。我们知道每个书架上有2本书。 从这个我们可以推断出布莱恩有21个书架。答案是21。

问题：在一所学校里有902个女孩和811个男孩。又有44个女孩加入了学校。 现在学校里有多少个女孩?

答案：学校里有902个女孩和811个男孩。又有44个女孩加入了学校。 这意味着现在学校里有(902 + 44) = 946个女孩。答案是946。

问题：他还有26个盐水动物的水族箱。 这个水族箱总共有52只盐水动物，每个水族箱里的动物数量相同。 每个水族箱里有多少只盐水动物?

A:我们知道总共有26个水族馆和52只动物。这意味着如果我们将52除以26，我们就会知道每个水族馆里有多少动物。52除以26等于2。 所以，每个水族馆有2只海水动物。答案是2。

Q:一个邮递员必须给每个街区送25封垃圾邮件。 如果他给每个房子送5封邮件，一个街区有多少房子?

A:邮递员必须给每个街区送25封垃圾邮件。 这意味着如果邮递员给每个房子送5封垃圾邮件，他在访问5个房子后就会送出25封垃圾邮件。 因此，如果邮递员必须给每个街区送25封垃圾邮件，他必须在一个街区访问5个房子。答案是5。

问题：每天Ryan花3个小时学习英语，还花一些时间学习中文。 如果他每天总共花4个小时学习英语和中文，他花多少时间学习中文?

回答：首先，我们知道Ryan每天花3个小时学习英语。其次，我们知道Ryan每天总共花4个小时学习英语和中文。由此可推断出Ryan每天花1个小时学习中文。答案是1。

表16：CSQA的自动构建演示。

问题：你在哪里可以找到狗的房子？ 答案选择：（A）在外面找到（B）忠诚的（C）可怕的（D）冷的（E）四条腿

回答：让我们一步一步地思考。 狗需要一个房子来保护它免受天气和捕食者的影响。 房子应该足够大，让狗可以活动，但不要太大，否则狗会感到寒冷。 房子应该有一扇门，以防寒冷，有一层屋顶，以防雨水。 房子应该由坚固的材料制成，不会在风中倒塌。 找到狗屋最好的地方是宠物店或家居改进店。 你也可以在网上找到狗屋。答案是**A**。

问：你在哪里可以找到售票亭并观看音乐会？ 选项：（A）职员（B）室内（C）电影院（D）场馆（E）礼堂

答：让我们一步一步地思考。 你可以去售票亭购买音乐会的门票。 音乐会将在一个场馆举行，可以是室内或室外。 最有可能的是音乐会将在一个大礼堂或电影院室内举行。答案是D。

问：人们在工作中的目标是什么？ 选项：（A）完成工作（B）互相学习（C）杀死动物（D）戴帽子（E）互相交谈

答：让我们一步一步地思考。 人们通常去工作是为了赚钱，他们可以用来购买他们想要或需要的东西。 为了做到这一点，他们必须完成他们的工作。这通常意味着与他人合作以高效地完成工作。 有时候，人们从他们的同事那里学到新的东西。 例如，他们可能学到一项新技能或如何更好地完成工作。 最后，人们通常在休息时间或下班后与他们的同事交谈。答案是**A**。

问：你在哪里存放一个大容器？ 答案选择：（A）超市（B）工厂（C）果汁（D）旅馆（E）柜子

答：让我们一步一步地思考。 我们需要存放一个大容器。 它无法放入柜子中，所以这不是答案。 它也无法放入果汁容器中，所以这不是答案。 它可以放入工厂，但那不一定是存放它的地方。它也可以放入超市，但那不一定是存放它的地方。最有可能的答案是旅馆。答案是D。

问：如果有一个周围有树的池塘，它可能位于哪里？ 答案选择：（A）地面（B）浴室（C）森林（D）乡村（E）农村地区

答：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，一个周围有树的池塘最有可能位于森林中。答案是C。

Q:詹姆斯正在见一个朋友。 他们计划过一个慢节奏的一天。他们不想做太多事情。他们只想要什么？ 答案选择：（A）放松（B）恐慌（C）警报（D）喜悦（E）欢呼

A:让我们一步一步地思考。 詹姆斯正在见一个朋友。 他们计划过一个慢节奏的一天。他们不想做太多事情。他们只想要____。 唯一合乎逻辑的答案选择是放松。答案是A。

Q:一个人为了保持健康应该做什么？ 答案选择：（A）发烧（B）每天吃东西（C）兴奋（D）头痛（E）表达

A:让我们一步一步地思考。 为了保持健康，一个人应该做一些像吃健康食物、锻炼和充足休息的事情。 所以最符合的答案选择是（B）每天吃东西，（C）兴奋和（D）头痛。
答案是B。

表17：用于StrategyQA的自动构建演示。

Q: 素食主义者能在Chick-fil-A吃东西吗？

A: 让我们一步一步地思考。 Chick-fil-A是一家专门提供鸡肉三明治的连锁餐厅。 所以从一开始，我们可以说素食主义者可能找不到他们能吃的三明治。然而，Chick-fil-A提供沙拉和配菜，素食主义者可以吃。答案是肯定的。

Q: Justin Timberlake能在六旗乘坐Shipwreck Falls吗？

A: 让我们一步一步地思考。 首先，让我们考虑Shipwreck Falls的身高要求，即48英寸。
接下来，我们将查找Justin Timberlake的身高。 根据各种来源，Timberlake的身高在5'8"到5'11"之间。
最后，我们将比较Timberlake的身高与Shipwreck Falls的身高要求。由于Timberlake比48英寸高，他可以乘坐Shipwreck Falls。答案是肯定的。

问：蛇会害怕蜜獾吗？

答：让我们一步一步地思考。 蛇的主要攻击方式是缠绕动物并勒死它，直到它窒息。 对于蜜獾来说，这将是困难的，因为它们有厚厚的皮肤并且非常灵活。 此外，蜜獾以非常具有攻击性而闻名，它们有锋利的爪子和牙齿，可以对蛇造成很大的伤害。 因此，可以合理地说蛇会害怕蜜獾。答案是肯定的。

问：伊丽莎白女王会被埋葬在万神殿吗？

答：让我们一步一步地思考。 首先，万神殿是一座教堂，所以她有可能被埋葬在那里。 其次，伊丽莎白二世仍然健在，所以她还没有被埋葬。第三，即使她被埋葬在万神殿，我们事先可能不会知道，所以很难确定。答案是否定的。

问：在温布利体育场能看到月亮吗？

答：让我们一步一步地思考。 温布利体育场是一个室内场馆，所以只有在屋顶有个洞的情况下才能看到月亮。 即使如此，月亮也只能在夜晚才能看到。 所以答案很可能是否定的。答案是否定的。

问：美国第40任总统是否向他的朋友转发lolcats？

答：让我们一步一步地思考。 美国第40任总统是罗纳德·里根。没有证据表明罗纳德·里根曾向他的朋友转发lolcats。答案是否定的。

表18：最后一个字母连接的自动构建演示。

问：取"Lesley Luna Nadia Adriana"中每个单词的最后一个字母并连接起来。

答：让我们一步一步地思考。 "Lesley"的最后一个字母是"y"。 "Luna"的最后一个字母是"a"。 "Nadia"的最后一个字母是"a"。 "**A**driana"的最后一个字母是"a"。 所以最终答案是"yaaa"。答案是yaaa。

Q: 取"Alfonso Collin Amado Dick"中每个单词的最后一个字母并连接起来。

A: 让我们逐步思考。 "Alfonso"的最后一个字母是"o"。 "Collin"的最后一个字母是"n"。 "Amado"的最后一个字母是"o"。 "Dick"的最后一个字母是"k"。 所以最终答案是"onok"。答案是onok。

Q: 取"Cristina Saad Bridget Katie"中每个单词的最后一个字母并连接起来。

A: 让我们逐步思考。 "Cristina"的最后一个字母是"a"。 "Saad"的最后一个字母是"d"。 "Bridget"的最后一个字母是"t"。 "Katie"的最后一个字母是"e"。 所以最终答案是"adte"。答案是adte。

Q: 取"Idalia Arnoldo Marla Duane"中每个单词的最后一个字母并连接起来。

A: 让我们一步一步地思考。 "Idalia"的最后一个字母是"a"。 "Arnoldo"的最后一个字母是"o"。 "Marla"的最后一个字母是"a"。 "Duane"的最后一个字母是"e"。 所以最终答案是"aoae"。答案是aoae。

表19：用于硬币翻转的自动构建演示。

Q:一个硬币是正面朝上的。Shari翻转了硬币。Bella翻转了硬币。 Liza没有翻转硬币。 Maira没有翻转硬币。 硬币仍然是正面朝上吗？ 注意这里的"翻转"意味着"反转"。

A:让我们一步一步地思考。Shari翻转了硬币。 硬币现在是反面朝上的。Bella翻转了硬币。 硬币现在是正面朝上的。Liza没有翻转硬币。 硬币仍然是正面朝上的。Maira没有翻转硬币。 硬币仍然是正面朝上的。答案是是。

问：硬币正面朝上。乔伊斯翻转了硬币。勒罗伊翻转了硬币。 艾丽莎没有翻转硬币。 玛吉没有翻转硬币。 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：让我们一步一步地思考。 乔伊斯翻转了硬币，所以现在是反面朝上。勒罗伊翻转了硬币，所以现在又是正面朝上。 艾丽莎没有翻转硬币，所以它仍然是反面朝上。玛吉没有翻转硬币，所以它仍然是正面朝上。所以最终答案是是的，硬币仍然是正面朝上。答案是是的。

问：硬币正面朝上。乌巴尔多没有翻转硬币。卡特琳娜翻转了硬币。 弗朗西斯没有翻转硬币。林恩翻转了硬币。 硬币仍然是正面朝上吗？ 请注意，这里的“翻转”意味着“反转”。

A:让我们一步一步地思考。 乌巴尔多没有翻转硬币，所以硬币仍然是正面朝上。卡特里娜翻转了硬币，所以硬币现在是反面朝上。弗朗西斯没有翻转硬币，所以硬币仍然是反面朝上。琳恩翻转了硬币，所以硬币现在又是正面朝上。答案是是。

Q:一枚硬币是正面朝上。卡梅洛翻转了硬币。托尼翻转了硬币。苏珊翻转了硬币。 莎拉没有翻转硬币。 硬币还是正面朝上吗？ 请注意，这里的“翻转”意味着“反转”。

A:让我们一步一步地思考。卡梅洛翻转了硬币。 硬币现在是反面朝上。托尼翻转了硬币。 硬币现在是正面朝上。苏珊翻转了硬币。 硬币现在是反面朝上。莎拉没有翻转硬币。 硬币仍然是反面朝上。答案是否定的。

问：硬币正面朝上。保罗翻转硬币。塔蒂亚娜翻转硬币。 伯尼斯不翻转硬币。劳尔翻转硬币。 硬币仍然正面朝上吗？ 请注意，这里的“翻转”意味着“反转”。

答：让我们一步一步地思考。保罗翻转硬币。 这意味着硬币现在是反面朝上的。塔蒂亚娜翻转硬币。 硬币现在又是正面朝上了。 伯尼斯不翻转硬币。 硬币仍然正面朝上。劳尔翻转硬币。
硬币现在又是反面朝上了。答案是否定的。

问：硬币正面朝上。克劳迪娅翻转硬币。 科尔不翻转硬币。 马修不翻转硬币。 胡安
巴勃罗不翻转硬币。 硬币仍然正面朝上吗？ 请注意，这里的“翻转”意味着“反转”。

A:让我们逐步思考。 Claudia翻转了硬币，所以现在是反面朝上。Cole没有翻转硬币，所以它仍然是反面朝上。Matthew没有翻转硬币，所以它仍然是反面朝上。Juan Pablo没有翻转硬币，所以它仍然是反面朝上。所以答案是否定的，硬币不是正面朝上。答案是否定的。

Q:硬币是正面朝上。Aj没有翻转硬币。Jd翻转了硬币。 Maddie没有翻转硬币。 Francisca没有翻转硬币。 硬币还是正面朝上吗？ 请注意，这里的"翻转"意味着"反转"。

A:让我们逐步思考。 Aj没有翻转硬币，所以硬币仍然是正面朝上。Jd翻转了硬币，所以硬币现在是反面朝上。Maddie没有翻转硬币，所以硬币仍然是反面朝上。Francisca没有翻转硬币，所以硬币仍然是反面朝上。所以，最终答案是硬币是反面朝上。答案是否定的。

问：硬币正面朝上。阿尔伯特没有翻转硬币。 费利西亚没有翻转硬币。玛戈翻转了硬币。 帕蒂没有翻转硬币。 硬币仍然是正面朝上吗？ 请注意，这里的"翻转"意味着"反转"。

答：让我们一步一步地思考。 阿尔伯特没有翻转硬币，所以硬币仍然是正面朝上。费利西亚没有翻转硬币，所以硬币仍然是正面朝上。玛戈翻转了硬币，所以硬币现在是反面朝上。帕蒂没有翻转硬币，所以硬币仍然是反面朝上。答案是否定的。