arXiv:2302.00618v1 [cs.CL] 1 Feb 2023

合成提示：为大语言模型生成思维链示例

Zhihong Shao [1 2] **Yeyun Gong** [3] **Yelong Shen** [4] **Minlie Huang** [1] **Nan Duan** [3] **Weizhu Chen** [4]

摘要

大语言模型可以通过思维链提示执行各种推理任务，这种提示引导模型通过逐步示例找到答案。然而，提示的质量取决于提供给模型的示例，而手工创建大量示例成本高昂。我们提出 SYNTHETIC PROMPTING，这是一种利用少量手工编写的示例提示模型自行生成更多示例，并选择有效示例以激发更好推理的方法。我们通过交替执行反向过程和正向过程来生成新示例。反向过程会生成一个与采样得到的推理链相匹配的问题，使问题清晰且可解。正向过程则为该问题生成更详细的推理链，从而提升示例质量。我们在数值、符号和算法推理任务上评估了该方法，结果表明，它优于现有的提示技术。

1. 引言

少样本示例，即任务输入和输出的示例，可以使大语言模型（大语言模型）无需微调即可执行各种任务（Brown 等人，2020年；Chung 等人，2022年）。大语言模型还可以通过思维链提示进一步提升性能，这种提示为任务提供中间推理步骤（Wei 等人，2022b；Kojima 等人，2022年）。然而，大语言模型的少样本表现高度依赖示例的质量，尤其是在需要复杂且多样化推理模式的推理任务中。手动创建大量多样化的示例以供示例选择，既费时又费力；而依赖数量有限的示例，则可能妨碍大语言模型对不同测试输入的泛化能力和适应能力

[1]清华大学 [2]本工作是在微软亚洲研究院实习期间完成的。 [3]微软亚洲研究院 [4]微软。联系方式：Minlie Huang <aihuang@tsinghua.edu.cn>。

能力。

本文提出一种新方法——合成提示，它利用大语言模型自身的知识和生成能力，通过自合成示例扩充有限数量的演示，然后利用扩充后的集合，促使大语言模型进行更好的推理。具体而言，给定少量种子示例（每个示例由一个问题和一系列推理步骤组成），我们提示大语言模型通过交替执行两个过程来生成更多示例：（1）反向过程，即大语言模型基于自生成推理链合成问题，从而确保问题可回答且定义明确；（2）正向过程，即大语言模型为合成问题生成推理链，使推理链更加准确，并与问题保持一致。我们重复这一过程，直到获得足够的合成示例。为了从扩充后的集合中选出最有效的演示，我们提出一种基于簇内复杂度的新选择方案，旨在通过对演示进行聚类并从每个簇中选出最复杂的演示（即推理链最长的演示），最大限度地提高其多样性和信息量。最后，我们向大语言模型提供选出的演示，提示其为测试问题生成推理链，并利用该推理链得出答案。

我们在多种推理任务上评估了我们的方法，包括数值推理、算法推理和符号推理。遵循以往的少样本设置（Wang 等人，2022b；Suzgun 等人，2022年），我们证明，我们的方法能够显著提升大语言模型的表现，较最先进的方法最多提高15.6个百分点。

我们的主要贡献如下：

- 我们提出了 SYNTHETIC PROMPTING，这是一种新颖的方法，通过向大语言模型提供提示，用自合成示例扩充有限的示例集，并利用扩充后的集合来引导大语言模型进行更好的推理。
- 我们提出了一种基于簇内复杂度的方案，用于从增强数据集中选择多样且信息丰富的演示

以进行推理。

- 我们在三个推理任务上展示了该方法的有效性，其表现显著优于以往方法。

2. 相关工作

上下文少样本学习 通过大规模无监督预训练，大语言模型（Brown 等人，2020年；Chowdhery 等人，2022年；Zhang 等人，2022a）能够通过模仿上下文示例来学习执行任务（Shin 等人，2022年）。为提升模型对提示的鲁棒性，研究者提出了指令微调（Ouyang 等人，2022年；Wei 等人，2022a；Sanh 等人，2022年；Chung 等人，2022年），通过在多种任务上训练语言模型，使其能够生成符合指令要求的理想输出。随着可控性的提升，基于上下文学习的应用蓬勃发展，包括文本生成（Yang 等人，2022年；Gao 等人，2022a）、对话生成（Thoppilan 等人，2022年）和资源构建（West 等人，2022年）。

推理提示技术思维链提示并非直接生成答案，而是提示大语言模型通过逐步推理得出答案（Wei 等人，2022b），这大幅提升了模型在众多推理任务上的表现。由少到多提示（Zhou 等人，2022年）、自我提问（Press 等人，2022年）和分解式提示（Khot 等人，2022年）等后续工作也体现了问题分解的理念，即将复杂问题拆解成一系列易于处理的子问题。这些方法都会生成自然语言推理步骤，但在计算和符号操作方面表现欠佳。程序辅助语言（PAL）提示（Gao 等人，2022b）和程序思维提示（Chen 等人，2022年）等技术提出使用结构化代码改进自然语言推理，在算术、符号和算法任务上取得了显著提升。

与提示工作流程正交的另一类工作，探索了什么样的演示才有效。评估指标包括：(1) 多样性，即选择互补的示例，使模型能够融合不同的推理方式（Li 等人，2022年；Ye 等人，2022b），或减少模型受某一种推理方式影响而产生的偏差（Zhang 等人，2022b）；(2) 推理复杂度，即选择推理复杂度最高的示例；经验研究发现，这种方法在数值推理任务上效果良好（Fu 等人，2022年）；(3) 与测试输入的相似度，即检索在结构上（Drozdov 等人，2022年）或语义上（Liu 等人，2022b）相似的示例。为确保示例兼具多样性和信息量，我们提出一种基于簇内复杂度的选择方案，从示例簇中选出最复杂的示例。所有这些选择方案都假设可以获取一组示例（无论是否带有标注）。

从大语言模型中蒸馏知识一些研究将知识从大语言模型中蒸馏为符号知识，例如结构化常识知识（West 等人，2022年）或特定任务示例（Liu 等人，2022a；Ye 等人，2022a；Huang 等人，2022年）。这些研究至少具有以下特征之一：(1) 假设能够获取训练集中的标准输入，无需生成这些输入；(2) 基于工作者与人工智能的协作来蒸馏知识；(3) 使用蒸馏得到的知识进行训练。相比之下，我们假设只能获取少量标准示例，通过向大语言模型提供提示来自动合成更多示例，并研究合成示例能否在不进一步训练的情况下，用于更好地激发模型自身的推理能力。

3. 合成提示

3.1. 概述

为了使用大语言模型执行推理任务，给定若干由问题和推理链组成的示例，通常会将它们直接拼接成提示，用于推理。本文则将这些示例视为种子示例，并提示大语言模型通过反复执行反向-正向过程，自动合成更多示例；反向过程和正向过程分别生成问题及其对应的推理链。在推理时，使用自合成示例作为提示，以更好地激发模型自身的推理能力。示例通过一种新方案进行筛选，以确保多样性和信息量。

3.2. 示例合成阶段

使用种子示例，我们通过反复执行反向-正向过程，自动合成更多示例。每个合成示例都是一个 ⟨ 问题、推理链 ⟩对。在我们的主要实验中，我们采用PAL风格推理，即推理链是代码片段，答案则通过执行代码获得。

3.2.1. 反向过程

在反向过程中，提示大语言模型先生成推理链，再生成问题。作为反向过程输出的问题，是在给定主题词、目标推理复杂度和自生成推理链的条件下合成的。图 1（左）展示了反向过程的提示示例，其中包含从种子示例和先前合成的示例中随机抽取的一些示例。示例数量等于种子示例的数量。

主题词我们假设，每个推理问题都与某个特定主题相关，且不同主题

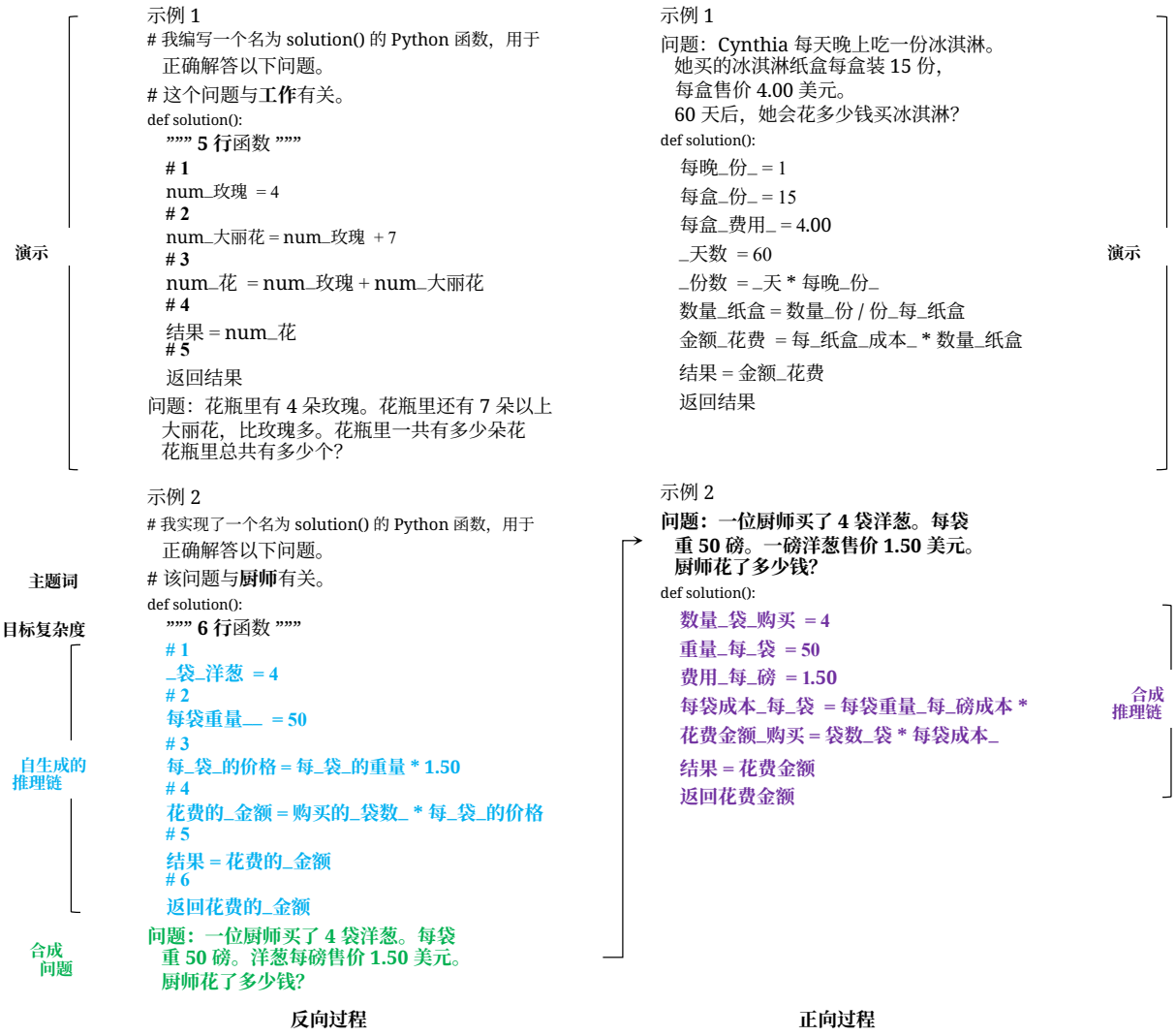

图 1：示例合成的反向过程（左）和正向过程（右）中的提示及模型补全示例。为简洁起见，我们在每个提示中只展示一个演示。自生成推理链（蓝色）、合成问题（绿色）和合成推理链（紫色）均为补全示例。在反向过程中，大语言模型根据主题词、目标推理复杂度和生成的推理链来合成问题。为了更好地控制推理复杂度，我们为推理步骤编号，例如左侧的 # 1 和 # 2。在正向过程中，大语言模型为反向过程中生成的问题合成一条更精确的推理链。反向过程中生成的问题与正向过程中生成的相应推理链共同构成一个合成示例。

可能需要不同类型的推理。例如，关于税的问题可能涉及算术运算，而关于速度的问题可能涉及单位换算。为确保合成问题的多样性，我们提示模型针对给定的主题词生成一个问题，该主题词从一组词中随机抽取。我们通过提示模型列出单词元名词来创建这个词集，并在提示中附上一些来自种子示例的随机名词。生成词集的指令是：列出 50 个名词。每个词都应仅包含一个词元。不要重复已列出的词，后面最多接 10 个来自种子示例的词。我们重复这一过程，直到获得 1,000 个不同的词，或提示重复达到 100 次。

目标复杂度 我们还希望控制合成问题的复杂度，因为更复杂的示例可能有助于模型更好地学习推理能力（Fu 等人，2022年）。我们将问题的复杂度定义为所需的数量

推理步骤，其中每一步都是由换行符分隔的一行代码。例如，图 示例 1 中的1 （左）复杂度为 5，因为它有 5 行代码。生成问题时的目标复杂度从一个范围内随机抽取，该范围从种子示例中的最低复杂度延伸到最高复杂度再加上 c。

自生成推理链 我们提示模型针对给定主题生成具有目标复杂度的推理链，然后根据该推理链生成问题。我们发现，与不经过推理链而直接生成问题相比，这种方法生成的问题更容易作答，也更明确。为了引导模型遵循目标复杂度，我们在示例中为每个推理步骤编号，例如图`#1` 和 `#2` （左）中的1 。我们会过滤掉重复的问题、至少重复一个五元组的问题，或未提及给定主题词的问题。

3.2.2. 正向过程

正向过程旨在为反向过程中合成的问题生成推理链。图1（右）展示了正向过程提示的示例，其中包含种子示例。与思维链提示不同，PAL 提示不包含最终答案，因为答案可以通过执行生成的代码获得，而非从模型输出中提取。我们观察到，正向过程中生成的推理链比反向过程中生成的推理链更相关、更准确，因为它直接以问题为条件。

我们还希望确保模型对推理链得出的答案具有足够的置信度。参照Huang 等人 (2022年)，我们用得出相同答案的抽样推理链所占比例来衡量答案的置信度。对于问题 x，我们抽样 m条推理链，并得到它们的答案 $\{a_1, a_2, ..., a_m\}$。随后，我们通过多数投票找出一致性最高的答案： $\hat{a} = \arg max_{a_i} \sum_{k=1}^{m} \mathbb{1}(a_i = a_k)$。如果超过 $m/2$条推理链得出 $\hat{a}$，我们便将其中最短的一条与合成问题关联；否则，就舍弃该问题，因为模型无法为其生成具有足够置信度的推理链。请注意，多数投票仅用于合成示例，不用于推理（第 3.3 节）。这与Wang 等人 (2022a) 不同，他们在推理时使用多数投票。

3.3. 推理阶段

在推理过程中，我们会从合成的示例中选取一部分作为模型的演示。根据Fu 等人(2022年的研究，与根据相似度选取演示相比，根据复杂度选取演示可以提升模型在推理任务上的表现。此外，根据相似度选取演示可能会引入来自演示的偏差（Zhang 等人，2022b；Lyu 等人，2022年），尤其是在演示不正确时。此外，选取彼此互补的演示，可能有助于模型融合不同类型推理中的知识（Ye 等人，2022b；Zhang 等人，2022b）。

因此，我们提出了一种基于簇内复杂度的方案，用于选择既复杂又互补的演示。具体而言，我们使用 Sentence-BERT（Reimers 与 Gurevych，2019）作为编码器，在语义嵌入空间中对合成示例进行聚类。簇的数量与推理时使用的演示数量相同。随后，我们从每个簇中选取最复杂的示例作为演示。推理过程与 PAL 提示等先前工作相同，即由模型补全给定的提示。唯一的区别在于，我们提示中的演示是根据种子示例合成的，而非固定使用种子示例。

4. 实验

4.1. 数据集

我们在七个涵盖不同推理任务的数据集上进行了实验。示例见表1。

数值推理 (1) GSM8K（Cobbe 等人，2021）是一个包含 1,319 道多样化小学数学应用题的数据集，旨在评估大语言模型进行多步数学推理的能力。(2) GSM-Hard 是 Gao 等人（Gao 等人（2022b）创建的 GSM8K 的更难版本，其做法是将问题中的数字替换为更大的数字，旨在评估大语言模型能否泛化到较大的数字。(3) SVAMP（Patel 等人，2021）是一个包含 1,000 个问题的数学应用题数据集，用于评估鲁棒性。(4) ASDiv（Miao 等人，2020年）由 2,000 道多样化的数学应用题组成。(5) SingleOp（Koncel-Kedziorski 等人，2016）由 562 道数学应用题组成。

符号推理 来自 Big-Bench Hard 的彩色物体任务（Suzgun 等人，2022年），包含 2,000 个关于给定物体的位置和颜色属性的问题。

算法推理 Repeat Copy 任务同样来自 Big-Bench Hard，包含 32 个测试示例。模型应生成一个符合给定指令要求的词语序列。

4.2. 评估设置

无论是Suzgun 等人(2022年) 还是Wang 等人(2022b），都在少样本设置下，使用基准测试评估了大语言模型在众多任务上的表现；这些设置最多只能使用 4 个标准示例。

数据集	示例
GSM8K	Patrick 有三根用过一部分的胶棒。一根还剩 1/6，第二根还剩 2/3，第三根还剩 1/2。如果胶棒原本长 12 毫米，那么剩余胶棒的总长度未使用的胶棒有多少?
彩色物体	床头柜上，你看到几件物品排成一列：一支蓝色蜡笔、一本红色笔记本、一件青绿色手链、一张洋红色纸张、一条银色狗绳和一本黑色小册子。离狗绳最远的物品是什么颜色?
Repeat Copy	把“选举进入委员会”重复三次，但每隔一个词就说一次“酷”。

表 1：示例，选自　　三个数据集。其他数值推理数据集中的问题与 GSM8K 中的问题相似。

遵循这些设置，我们默认假设每个数据集都可以使用 2 或 4 个随机示例。对于数值推理任务，我们还尝试使用 Wei 等人(2022b) 手工编写的 8 个示例；这些示例也被后续的多篇论文采用（Fu 等人，2022年；Wang 等人，2022a；Gao 等人，2022b）。我们还使用了PAL 风格的推理链，这些推理链由Gao 等人(2022b) 标注。

不使用合成的提示基线方法会使用所有给定的标准示例来构造推理提示。SYNTHETIC PROMPTING 及其变体会使用给定示例合成示例，并根据簇内复杂度选出 8 个合成示例，除非另有说明。

种子示例和合成提示见补充材料。

4.3. 基线方法

直接提示直接提示（Brown 等人，2020年）通过输入-答案对的示例提示大语言模型直接生成答案。

思维链提示思维链提示（Wei 等人，2022b）能够有效引导大语言模型进行推理，提示大语言模型先生成自然语言推理步骤，再给出答案。

PAL 提示PAL 提示（Gao 等人，2022b），作为思维链提示的一种变体，通过结构化代码提升推理能力。图1 （右侧）给出了两个示例。它不会提示大语言模型在生成内容中包含最终答案；答案通过执行代码获得。这种提示技术在许多推理任务上取得了最先进的结果。

原版 S合成提示 这是 SYNTHETIC PROMPTING 的一个变体，不同之处在于，用于问题合成的提示仅由种子示例中的问题构成。换言之，新问题通过模仿种子问题合成，不附加任何其他条件。

4.4. 实现细节

我们采用了 PAL 风格推理链，这是一种带有注释的结构化代码，其中注释是自然语言推理步骤。我们使用 InstructGPT 的 text-davinci-003 版本（Ouyang 等人，2022年）作为合成和推理的后端大语言模型。合成时使用 top-p 采样（Holtzman 等人，2020年），温度设为 0.7；推理时使用贪心解码，温度设为 0。所有数值推理数据集共用一组种子示例：当种子示例数量为 2 或 4 时，从 GSM8K 中随机抽取；当数量为 8 时，从Wei 等人 （2022b）中抽取。对于其他任务的数据集，则从各自的数据集中随机抽取种子示例。我们遵循相应的标注规范，手动为种子示例标注思维链式推理链和 PAL 风格推理链。标注内容见补充材料。对于每组种子示例，我们通过重复执行反向-正向合成 1,000 次，合成更多示例。目标复杂度范围从种子示例的最低复杂度到最高复杂度加上 c；数值推理数据集的 c 设为 4，其他数据集设为 2。在正向合成中，每个问题采样 3 条推理链。用于聚类的编码器为 all-mpnet-base-v2。

4.5. 主要结果

如表2所示，SYNTHETIC PROMPTING 的表现始终优于 PAL 提示，最高绝对提升 15.6%，这表明自合成示例可用于更好地引导大语言模型自身进行推理，效果超过仅使用种子示例。

尽管原始合成提示也使用合成示例，但并不能持续优于 PAL 提示。在 GSM8K 和 GSM-Hard 上，这些数据集包含需要复杂推导的问题，原始合成提示几乎无法优于 PAL 提示，因为它没有明确控制合成示例的推理复杂度，并且往往会合成在复杂度和信息量方面与种子示例相似的示例。值得注意的是，在 Repeat Copy 上，原始合成提示的表现显著逊于 PAL 提示

数据集	GSM8K			GSM-Hard			SVAMP			ASDiv			SingleOp			彩色物体重复复			
此前微调的最优水平	55[α]			–			57.4[β]			75.3[γ]			–			–		–	
思维链PaLM 540B	56.9			–			79.0			73.9			–			–		–	
# 黄金/种子示例	2	4	8	2	4	8	2	4	8	2	4	8	2	4	8	2	4	2	4
仅使用黄金示例的方法																			
直接法	15.4				4.3	4.4											81.7	31.3	43.8
CoT	61.8																88.8	56.3	62.5
PAL	72.5																94.4	71.9	81.3
利用合成示例的方法																			
原始 SYNTHETIC	**72.6**						___			___							93.1	53.1	81.3
SYNTHETIC	**74.7**	___			___	___									___		**97.3**	**87.5**	**84.4**

表 2：数值推理、符号推理和算法推理数据集上的准确率。此前经过微调的最先进模型包括：α：（Cobbe 等人，2021）；β：（Pi 等人，2022年）；γ：（Miao 等人，2020年）。PaLM 540B 的思维链提示准确率引自 Wei 等人（2022b）。PAL提示是此前最先进的方法，它在推理过程中使用提供的标准示例作为演示。SYNTHETIC PROMPTING 及其原始版本都将标准示例用作种子，通过提示大语言模型自动合成更多示例；在推理过程中，则从自合成示例中选择演示，以更好地激发大语言模型自身的推理能力。

使用 2 个种子示例。我们发现，选出的 2 个示例中的问题表述不当，例如：将句子“太阳很明亮”重复五次，每次强调一个不同的词。这可能是因为合成问题时没有明确考虑其推理链。第4.6.1 节展示了通过不同条件控制问题合成的益处。

我们还观察到，将种子示例数量从 2 增加到 8 并未显著提升性能，尤其是在 GSM8K 和 Repeat Copy 上。可能有以下两个原因：（1）示例合成会受到种子示例的偏向影响。种子有限时，合成示例可能不够多样，仍无法帮助处理部分测试问题。（2）尽管我们提出的示例选择方案有效（见第4.6.2节的分析），但它可能并非最优，未能充分利用合成示例。

4.6. 消融研究

我们主要在 GSM8K 和 Colored Objects 任务上进行了消融研究。

4.6.1. 问题合成所用的条件

数据集	GSM8K		彩色物体	
# 种子示例	2	4	2	4
S合成提示	**74.7**	**75.3**	**93.8**	**97.3**
不含主题词	73.9	73.2	80.2	92.9
不含目标复杂度	71.6	72.0	93.1	94.5
无推理链	72.9	73.8	86.0	92.2

表3：分析反向问题合成中使用的不同条件如何影响推理性能。

在逆向合成中，我们让大语言模型采样一个问题

方法	多样性	复杂度	正确性
SYNTHETIC 提示	0.68	8.23 / 12.5	100.0%
不含主题词	0.85	9.4 / 13.4	100.0%
不含目标复杂度	0.73	5.0 / 7.6	87.5%
无推理链	0.68	5.6 / 12.3	37.5%

表 4：使用 4 个种子示例对 GSM8K 进行分析，探讨逆向问题合成中使用的条件如何影响合成示例的质量。多样性通过合成示例之间的成对余弦相似度来衡量。复杂度衡量合成示例（左侧）和选定演示（右侧）的平均推理复杂度，两者以 / 分隔。正确性指选定的合成示例中正确示例所占的比例。

在给定主题、目标复杂度和采样得到的推理链的条件下。为了分析每个条件对问题合成的影响，我们从提示中删除了相应的行。值得注意的是，移除目标复杂度时，推理步骤的编号标记也会一并移除。如表3所示，移除任何一个条件都会导致模型在 GSM8K 和彩色物体上的性能下降。

我们进一步研究了不同条件如何影响合成示例的质量，具体考察了（1）**多样性**，以合成示例与其他示例之间的平均最大成对余弦相似度衡量；（2）**复杂度**，以推理步骤的平均数量衡量；以及（3）**正确性**，以用于推理的示例中正确示例所占的比例衡量。表4 展示了对 GSM8K 的分析。**移除主题词**会导致合成示例的多样性降低。所选示例的推理模式也受到限制；尽管所有示例都正确，但 62.5% 的问题都围绕折扣或税。**移除目标复杂度**会生成简单得多的合成示例。合成问题**而不**

以推理链为条件会对正确性产生负面影响；62.5% 的问题存在缺陷，其中 80% 无法作答，例如，

一张图像由一个 *4x4* 整数矩阵表示。矩阵的每个单元格都包含一个介于 *0* 到 *255* 之间的整数。矩阵中所有整数的平均值是多少？。

值得注意的是，尽管我们在合成问题时，也会将目标复杂度纳入提示，而不以推理链为条件，但生成的问题所需的推理步骤往往少于 SYNTHETIC 提示，这表明以编号的推理步骤为条件能更好地控制推理复杂度。

4.6.2. 示例选择方案

数据集	GSM8K		彩色物体	
# 种子示例	2	4	2	4
随机	73.4	73.1	83.0	90.0
簇质心	73.1	73.0	91.2	93.6
相似度	72.9	74.0	87.0	94.2
簇内相似度	72.9	73.2	89.3	95.4
复杂度	74.1	74.3	92.7	**97.5**
簇内复杂度	**74.7**	**75.3**	**93.8**	97.3

表5：采用不同示例选择方案时的准确率。

要充分利用合成示例，有效的选择方案至关重要。我们评估了以下6种选择方案。(1) **随机**：随机选择示例；(2) **簇质心**：选择距离每个簇质心最近的示例；(3)**相似度**：根据余弦相似度检索最相似的示例；(4) **簇内相似度**：从每个簇中选择最相似的示例；(5) **复杂度**：选择推理步骤最多的示例；(6) **簇内复杂度**：从每个簇中选择最复杂的示例。

表 5 展示了比较结果。尽管大多数选择方案的表现都优于 PAL 提示，但基于复杂度的选择方案在两项推理任务上效果最佳，而随机等其他方案则明显落后。我们提出的簇内复杂度优于复杂度方案，表明采用多样且复杂的演示具有优势。

4.6.3. 对种子示例的敏感性

为探究 SYNTHETIC PROMPTING 对种子示例的敏感程度，我们在另外两组随机种子示例上重复了实验。图 2 展示了我们的敏感性分析。SYNTHETIC PROMPTING 在不同轮次中始终优于 PAL 提示。然而，

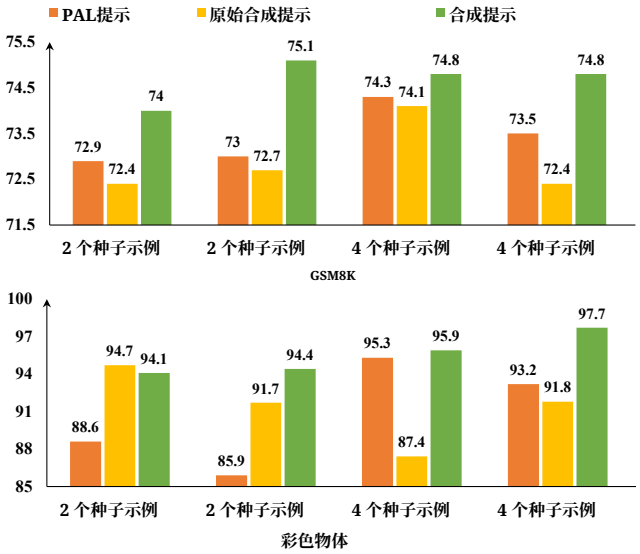

图2：GSM8K 和 Colored Objects 上的敏感性分析。我们针对每个数据集，另选了两组随机种子示例，每组分别包含 2 个和 4 个示例进行实验。

我们发现，PAL 提示效果更好的种子示例，并不一定能带来更好的 SYNTHETIC PROMPTING 效果。

4.7. 与从训练示例中选择的比较

为了衡量使用合成示例与使用从大量精心整理的示例中选出的黄金示例之间的性能差距，我们分别采用第 4.6.2节中的两种基于复杂度的选择方案（即复杂度和簇内复杂度），从 GSM8K 训练集中选出 8 个示例。由于训练示例标注了自然语言推理链（思维链式推理），我们将自然语言推理步骤的数量作为基于复杂度的选择所用的推理复杂度，并为 PAL 提示手动标注所选示例的 PAL 风格推理链。GSM8K 的训练示例种类丰富，因此复杂度和簇内复杂度两种方案都能选出多样且信息丰富的示例，在 GSM8K 测试集上达到 77.0% 的准确率，比我们 75.3% 的准确率高出 1.7 个百分点。如补充材料所示，与我们的合成示例相比，选出的黄金示例在逻辑上更复杂，推理过程也不那么直接，因此可能为大语言模型提供更多信息。

值得注意的是，使用 Wei 等人(2022b) 提供的 8 个简单示例，这些示例未经提示工程而由人工编写，准确率甚至更低，只有 71.8%。这表明示例确实很重要。在.....的情境下

种子示例
问题： Carol 的簿子里有 20 个签名，Jennifer 有 44 个。姐妹俩的暑假还剩三周，她们决定到暑假结束时，两人合计要达到 100 个签名。姐妹俩还需要收集多少个签名才能达到目标?
问题： 4 名油漆工每天工作 3/8 天，为一栋豪宅粉刷了 3 周。每位油漆工总共工作了多少小时?
使用原始合成提示生成的合成示例
问题： 有 80 人参加派对，其中 40 名男性和 40 名女性。男性中有 10 人已婚，女性中有 20 人已婚。派对上有多少名单身人士?
问题： 4 名油漆工每天工作 3/8 天，为这栋豪宅粉刷了 3 周。如果每位油漆工的时薪为 25 美元，那么这组油漆工总共赚了多少钱?
问题（无法作答）： Smith 女士的数学课上有 80 名学生参加测验，平均分为 75%。有多少名学生得分达到 95% 或以上?
来自 SYNTHETIC PROMPTING 的合成示例
给定主题词： 想法 **问题：** 如果 5 个人每人有 10 个想法，其中 5 个想法是创新的，每个需要 2 分钟来构思；另外 5 个想法是非创新的，每个需要 1 分钟来构思，那么 5 个人构思完全部 10 个想法总共需要多少分钟?
给定主题词： 办公室 **问题：** 办公室宽 20 英尺、长 30 英尺、高 10 英尺。办公室有两扇窗户，每扇宽 5 英尺、高 6 英尺，还有一扇宽 3 英尺、高 8 英尺的门。办公室墙面的总面积是多少?
给定主题词： 画廊 **问题：** 一家画廊有 10 幅绘画作品、9 件雕塑、6 张照片和 4 件混合媒介作品。绘画作品售价 200 美元，雕塑售价 500 美元，照片售价 100 美元，混合媒介作品售价 150 美元。你可以享受 15% 的折扣，并且需要支付 5% 的税。总共需要支付多少钱?

表 6：来自 GSM8K 的一组种子示例，以及分别由 SYNTHETIC PROMPTING 和其原始版本选出的合成示例。为简洁起见，我们只展示问题，因为它们的推理链都是正确的；但原始合成提示中的第三个问题无法回答。

在只能获取数量有限且可能较为简单的示例的情况下，自动合成示例以选择更有效的演示，是诱导大语言模型进行更好推理的一种有前景的方法。

4.8. 合成示例的质量分析

为了研究合成示例的质量，我们在 GSM8K 上进行了人工评估。我们分别评估了由 SYNTHETIC PROMPTING 和原始合成提示生成的 25 个随机示例。与原始合成提示相比，SYNTHETIC PROMPTING 合成的问题具有更高的复杂度（8.3 对 5.4），且错误率更低（8% 对 24% ）。

我们进一步分析所选合成示例的质量。对于 SYNTHETIC 提示，所有选出的示例都是正确的；而其原始版本中，有一个问题无法作答，另一个问题的推理则有误。

表 6 展示了两个种子示例以及一些合成示例。对于原始合成提示，前两个问题在逻辑上与种子问题相近，第三个问题则无法作答。借助合成提示，大语言模型可以生成与主题相关、需要新颖推理模式的问题，例如，第二个关于办公室的问题需要几何推理。

5. 结论

我们介绍了 SYNTHETIC PROMPTING（合成提示），这是一种利用少量示例与大语言模型进行推理的新技术。与以往工作不同，它不仅将模型用作上下文示例的使用者，还将其用作额外示例的生成器。我们展示了，通过提示大语言模型合成更多示例，与思维链提示和 PAL 提示等现有提示方法相比，可以提升模型在数值、符号和算法任务上的推理表现。

提示。

参考文献

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. 语言模型是少样本学习者。载于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., 和 Lin, H.（编），神经信息处理系统进展 *33*：*2020* 年神经信息处理系统年度会议，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020年。网址：
h
t
t
p
s://
proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstra
ct.
HTML。

Chen, W., Ma, X., Wang, X., and Cohen, W. W. 思维程序提示：将计算与数值推理任务中的推理过程分离。*CoRR*，abs/2211.12588，2022年。doi：10.48550/arXiv.2211.12588。网址 https://doi.org/10.48550/arXiv.2211.12588。

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., O mernick, M., Dai, A. M., Pillai, T. S., Pel- lat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. PaLM：借助路径扩展语言建模。*CoRR*，abs/2204.02311，2022年。doi：10.48550/arXiv.2204.02311。网址 https://doi.org/10.48550/arXiv.2204.02311。

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V. Y., Huang, Y., Dai, A. M., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., 和 Wei, J. 扩展指令微调语言模型。*CoRR*，abs/2210.11416，2022年。doi：10.48550/arXiv。2210.11416。网址：https://doi.org/10.48550/arXiv.2210.11416。Cobbe, K., Kosaraju, V., Bavarian, M., Hilton, J., Nakano, R., Hesse, C. 和 Schulman, J. 训练验证器以解决数学应用题。CoRR，abs/2110.14168，2021年。网址：https://arxiv.org/abs/2110.14168。Drozdov, A., Sch¨arli, N., Aky¨urek, E., Scales, N., Song, X., Chen, X., Bousquet, O. 和 Zhou, D. 使用大语言模型进行组合语义解析。CoRR，abs/2209.15003，2022年。doi：10.48550/arXiv. 2209.15003。网址：https://doi.org/10.48550/ arXiv.2209.15003。Fu, Y., Peng, H., Sabharwal, A., Clark, P. 和 Khot, T. 基于复杂度的多步推理提示。CoRR，abs/2210.00720，2022年。doi：10.48550/arXiv. 2210.00720。网址：https://doi.org/10.48550/ arXiv.2210.00720。Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C. 和 Guu, K. 使用语言模型研究和修订语言模型的输出，2022a。Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J. 和 Neubig, G. PAL：程序辅助语言模型。CoRR，abs/2211.10435，2022b。doi：10. 48550/arXiv.2211.10435。网址：https://doi.org/ 10.48550/arXiv.2211.10435。Holtzman, A., Buys, J., Du, L., Forbes, M. 和 Choi, Y. 神经文本退化的奇特案例。载于第八届国际学习表征会议，ICLR 2020，埃塞俄比亚亚的斯亚贝巴，2020年4月26日至30日。OpenReview.net，2020年。网址：https://openreview. net/forum?id=rygGQyrFvH。Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H. 和 Han, J. 大语言模型能够自我改进。CoRR，abs/2210.11610，2022年。doi：10.48550/arXiv. 2210.11610。网址：https://doi.org/10.48550/ arXiv.2210.11610。Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P. 和 Sabharwal, A. 分解式提示：解决复杂任务的模块化方法。CoRR，abs/2210.02406，2022年。doi：10.48550/arXiv.2210.02406。网址：https://doi.org/10.48550/arXiv.2210.02406。Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. 和 Iwa- sawa, Y. 大语言模型是零样本推理器。CoRR，abs/2205.11916，2022年。doi：10.48550/arXiv. 2205.11916。网址：https://doi.org/10.48550/ arXiv.2205.11916。

Koncel-Kedziorski, R., Roy, S., Amini, A., Kush man, N. 和 Hajishirzi, H. MAWPS：数学应用题库。载于 Knight, K., Nenkova, A. 和 Ram- bow, O.（编），*NAACL HLT 2016，2016*年北美计算语言学协会分会会议：人类语言技术，美国加利福尼亚州圣迭戈，*2016*年*6*月*12*日至*17*日，第1152–1157页。计算语言学协会，2016年。doi：10.18653/v1/n16-1136。网址：https://doi.org/10.18653/v1/n16-1136。

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., and Chen, W. 论如何提升语言模型的推理能力。2022年。doi: 10.48550/ARXIV.2206.02336。URL https://arxiv.org/abs/2206.02336。

Liu, A., Swayamdipta, S., Smith, N. A., and Choi, Y. WANLI：人工智能与工作者协作构建自然语言推理数据集。*CoRR*，abs/2201.05955，2022a。URL https://arxiv.org/abs/2201.05955。

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. 什么样的上下文示例适用于 GPT-3？载于 Agirre, E., Apidianaki, M., and Vulic, I.（编），《深度学习由内而外论文集：第三届知识抽取与集成及深度学习架构研讨会》，*DeeLIO@ACL 2022*，都柏林，爱尔兰及线上，*2022*年*5*月*27*日，第100–114页。计算语言学协会，2022b。doi: 10.18653/v1/2022.deelio-1.10。URLhttps://doi.org/10.18653/v1/2022.deelio-1.10。

Lyu, X., Min, S., Beltagy, I., Zettlemoyer, L. 和 Hajishirzi, H. Z-ICL：借助伪演示实现零样本上下文学习。*CoRR*，abs/2212.09865，2022年。doi：10. 48550/arXiv.2212.09865。URLhttps://doi.org/10.48550/arXiv.2212.09865。

Miao, S., Liang, C. 和 Su, K. 用于评估和开发英语数学文字题求解器的多样化语料库。载于 Jurafsky, D., Chai, J., Schluter, N. 和 Tetreault, J. R.（编），第*58*届计算语言学协会年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，第975–984页。计算语言学协会，2020年。doi：10.18653/v1/2020.acl-main.92。URLhttps://doi.org/10.18653/v1/2020.acl-main.92。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., 和 Lowe, R. 使用人类反馈训练语言模型，使其遵循指令。*CoRR*，abs/2203.02155，2022年。doi：10.48550/arXiv. 2203.02155。网址 https://doi.org/10.48550/arXiv.2203.02155。

Patel, A., Bhattamishra, S., 和 Goyal, N. 自然语言处理模型真的能解决简单的数学应用题吗？载于 Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-T¨ur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., 和 Zhou, Y.（编），*2021*年计算语言学协会北美分会会议论文集：人类语言技术，*NAACL-HLT 2021*，线上，*2021*年*6*月*6*日至*11*日，第2080–2094页。计算语言学协会，2021年。doi：10.18653/v1/2021.naacl-main.168。网址 https://doi.org/10.18653/v1/2021.naacl-main.168。

Pi, X., Liu, Q., Chen, B., Ziyadi, M., Lin, Z., Gao, Y., Fu, Q., Lou, J., 和 Chen, W. 像程序执行器一样进行推理。*CoRR*，abs/2201.11473，2022年。网址 https://arxiv.org/abs/2201.11473。

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., 和 Lewis, M. 衡量并缩小语言模型中的组合性差距。*CoRR*，abs/2210.03350，2022年。doi：10.48550/arXiv.2210.03350。网址 https://doi.org/10.48550/arXiv.2210.03350。

Reimers, N. 和 Gurevych, I. Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于 *2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第3982–3992页，香港，中国，2019年11月。计算语言学协会。doi：10.18653/v1/D19-1410。网址 https://aclanthology.org/D19-1410。

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., Datta, D., Chang, J., Jiang, M. T., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., F´evry, T., Fries, J. A., Teehan, R., Scao, T. L., Bi- derman, S., Gao, L., Wolf, T., and Rush, A. M. 多任务提示训练能够实现零样本任务泛化。载于第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。OpenReview.net，2022年。网址：https://openreview.net/forum?id=9Vrb9D0WI4。

Shin, S., Lee, S., Ahn, H., Kim, S., Kim, H., Kim, B., Cho, K., Lee, G., Park, W., Ha, J. 和 Sung, N. 论预训练语料库对大规模语言模型上下文学习的影响。载于 Carpuat, M., de Marn- effe, M. 和 Ru´ız, I. V. M.（编），*2022*年北美分会会议

计算语言学协会：人类语言技术，*NAACL 2022*，美国华盛顿州西雅图，*2022*年*7*月*10*日至*15*日，第 5168–5186 页。计算语言学协会，2022年。doi：10.18653/v1/ 2022.naacl-main.380。网址：
h
ttps://doi.org/10.18653/v1/2022.naacl-main.380。

Suzgun, M., Scales, N.,Sch¨arli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. 挑战 BIG-bench 任务，以及思维链能否解决这些任务。*CoRR*，abs/2210.09261，2022年。doi: 10.48550/arXiv.2210.09261。网址：https://doi.org/10.48550/arXiv.2210.09261。

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Mene- gali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhou, Y., Chang, C., Krivokon, I., Rusch, W., Pickett, M., Meier-Hellstern, K. S., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Ra- jakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fen- ton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera- Arcas, B., Cui, C., Croak, M., Chi, E. H., and Le, Q. LaMDA：面向对话应用的语言模型。*CoRR*，abs/2201.08239，2022年。网址：https://arxiv.org/abs/2201.08239。

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H. 和 Zhou, D. 自一致性提升了语言模型中的思维链推理。*CoRR*，abs/2203.11171，2022a。doi：10.48550/arXiv.2203.11171。网址：https://doi.org/10.48550/arXiv.2203.11171。

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H. G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Patel, M., Pal, K. K., Moradshahi, M., Par- mar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R.S., Karia, R., Sampat, S. K., Doshi, S., Mishra, S., Reddy, S., Patro, S., Dixit, T., Shen, X., Baral, C., Choi, Y., Smith, N. A., Hajishirzi, H. 和 Khashabi, D. Super-NaturalInstructions：通过自然语言处理任务上的声明式指令实现泛化，2022b。

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. 微调后的语言模型是零样本学习者。载于第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。OpenReview.net，2022a。网址：https://openreview.net/forum?id=gEZrGCozdqR。

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q., and Zhou, D. 思维链提示能够激发大语言模型的推理能力。*CoRR*，abs/2201.11903，2022b。网址：https://arxiv.org/abs/2201.11903。West, P., Bhagavatula, C., Hessel, J., Hwang, J. D., Jiang, L., Bras, R. L., Lu, X., Welleck, S., and Choi, Y. 符号知识蒸馏：从通用语言模型到常识模型。载于 Carpuat, M., de Marneffe, M., and Ru´ız, I. V. M.（编），*2022*年北美分会计算语言学协会会议论文集：人类语言技术，*NAACL 2022*，华盛顿州西雅图，美国，*2022*年*7*月*10*日至*15*日，第4602–4625页。计算语言学协会，2022年。doi：10.18653/v1/ 2022.naacl-main.341。网址：
h
t
tps://doi.org/10.18653/v1/2022.naacl-main.341。

Yang, K., Klein, D., Peng, N. 和 Tian, Y. DOC：通过精细的大纲控制提升长篇故事连贯性。*CoRR*，abs/2212.10077，2022年。doi：10.48550/arXiv. 2212.10077。网址 https://doi.org/10.48550/arXiv.2212.10077。

Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T. 和 Kong, L. Zerogen：通过数据集生成实现高效的零样本学习。*CoRR*，abs/2202.07922，2022a。网址 https://arxiv.org/abs/2202.07922。

Ye, X., Iyer, S., Celikyilmaz, A., Stoyanov, V., Durrett, G. 和 Pasunuru, R. 用于有效上下文学习的互补解释。*CoRR*，abs/2211.13892，2022b。doi：10.48550/arXiv.2211.13892。网址 https://doi.org/10.48550/arXiv.2211.13892。

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M. T., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D.,Koura,P.S., Sridhar, A.,Wang,T., andZettlemoyer, L. OPT：开放式预训练 Transformer 语言模型。*CoRR*，abs/2205.01068，2022a。doi：10.48550/arXiv. 2205.01068。网址 https://doi.org/10.48550/arXiv.2205.01068。

Zhang, Z., Zhang, A., Li, M., and Smola, A. 大语言模型中的自动思维链提示。*CoRR*，abs/2210.03493，2022b。doi：10.48550/arXiv. 2210.03493。网址 https://doi.org/10.48550/arXiv.2210.03493。

Zhou, D., Sch¨arli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D.,Bousquet, O., Le, Q., and Chi, E. H. 由少到多提示使大语言模型能够进行复杂推理。 *CoRR*，abs/2205.10625, 2022年。doi: 10.48550/arXiv.2205.10625。网址https://doi.org/10.48550/arXiv.2205.10625。