Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters

Boshi Wang[1] **Sewon Min**[2] **Xiang Deng**[1] **Jiaming Shen**[3] **You Wu**[3]
Luke Zettlemoyer[2] **Huan Sun**[1]
[1]The Ohio State University [2]University of Washington [3]Google Research
{wang.13930,deng.595,sun.397}@osu.edu
{sewon,lsz}@cs.washington.edu, {jmshen,wuyou}@google.com

Abstract

Chain-of-Thought (CoT) prompting can dramatically improve the multi-step reasoning abilities of large language models (LLMs). CoT explicitly encourages the LLM to generate intermediate rationales for solving a problem, by providing a series of reasoning steps in the demonstrations. Despite its success, there is still little understanding of what makes CoT prompting effective and which aspects of the demonstrated reasoning steps contribute to its performance. In this paper, we show that CoT reasoning is possible even with invalid demonstrations—prompting with invalid reasoning steps can achieve over 80-90% of the performance obtained using CoT under various metrics, while still generating coherent lines of reasoning during inference. Further experiments show that other aspects of the rationales, such as being relevant to the query and correctly ordering the reasoning steps, are much more important for effective CoT reasoning. Overall, these findings both deepen our understanding of CoT prompting, and open up new questions regarding LLMs' capability to learn to reason in context.[1]

1 Introduction

Large language models (LLMs) can perform new tasks during inference when prompted with a few demonstrations (Brown et al., 2020). Chain-of-Thought (CoT) prompting (Wei et al., 2022) can (Figure 1) improve the ability of sufficiently large LLMs to do complex and multi-step reasoning. In addition to (query, answer) example-pair demonstrations, CoT prompting includes a *rationale* (colored part in Figure 1) for each example, i.e., a series of reasoning steps towards the answer, which encourages the LLM to explicitly generate its intermediate reasoning process before predicting the final answer. Despite its successes, there is little understanding of what makes CoT prompting effective and which aspects of the demonstrated reasoning steps contribute to its performance. Recent findings also reveal that in-context learning could be very different from fine-tuning/training; for example, Min et al. (2022) and Webson and Pavlick (2022) show that providing random labels or misleading instructions in context only marginally harms model performance for certain tasks. Inspired by this work, we take a closer look at CoT prompting to study how and why it works.

We design a series of ablation experiments

[1]Our code and model input/output are available here.

	In-context Demonstration	Inference by LLM
Query	Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?
Standard	The answer is 39.	The answer is 18. ✗
CoT	Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of 12 + 24 = 36 pages. Now she has 120 - 36 = 84 pages left. Since she wants to read half of the remaining pages, she should read 84 / 2 = 42 pages. The answer is 42. ✓
Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.	Yesterday, Julie read 12 pages. Today, she read 12 * 2 = 24 pages. So she read a total of 12 + 24 = 36 pages. Now she needs to read 120 - 36 = 84 more pages. She wants to read half of the remaining pages tomorrow, so she needs to read 84 / 2 = 42 pages tomorrow. The answer is 42. ✓

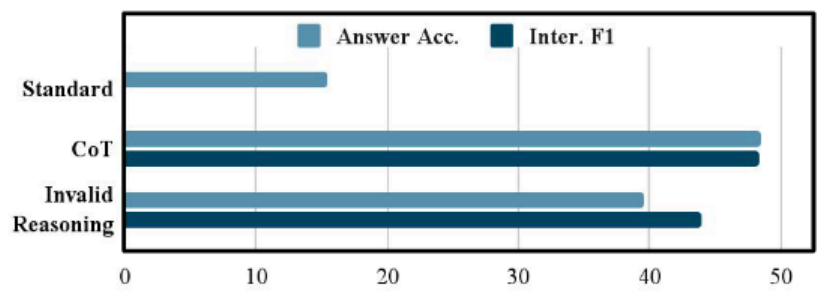

Figure 1: Results of standard prompting, Chain-of-Thought (CoT) prompting, and our ablation setting with invalid reasoning (§4). We show one demonstration example and one inference example for arithmetic reasoning, where the rationale is in color (green: valid, yellow: invalid). We find that valid reasoning for the demonstrations matters only a small portion to the performance of CoT—by providing rationales with invalid reasoning, LLMs can achieve over 80-90% of the performance of CoT under various metrics while performing logically sound and pertinent reasoning.

走向理解思维链：什么重要的实证研究

王博士　　孙焕（音）

[1] 俄亥俄州州立大学　　[2] 华盛顿大学　　[3] Google Research

{wang.13930, deng.595, sun.397}@osu.edu
{sewon, lsz}@cs.washington.edu, {jmshen,
wuyou}@google.com

摘要

思想链提示可以显著提高大型语言模型的多步推理能力。CoT 明确鼓励 LLM 通过在演示中提供一系列推理步骤来生成解决问题的中间原理。尽管它的成功，仍然有很少的理解是什么使 CoT 提示有效的，哪些方面的演示推理步骤有助于其性能。在本文中，我们表明，CoT 推理是可能的，即使与无效的演示提示与无效的推理步骤可以实现超过 80-90%的性能，使用 CoT 在各种指标下，同时仍然产生连贯的推理线在推理过程中。进一步的实验表明，原理的其他方面，例如与查询相关和正确排序推理步骤，对于有效的 CoT 推理来说更为重要。 总的来说，这些发现都加深了我们对 CoT 提示的理解，并就 LLM 在上下文中学习推理的能力提出了新的问题。

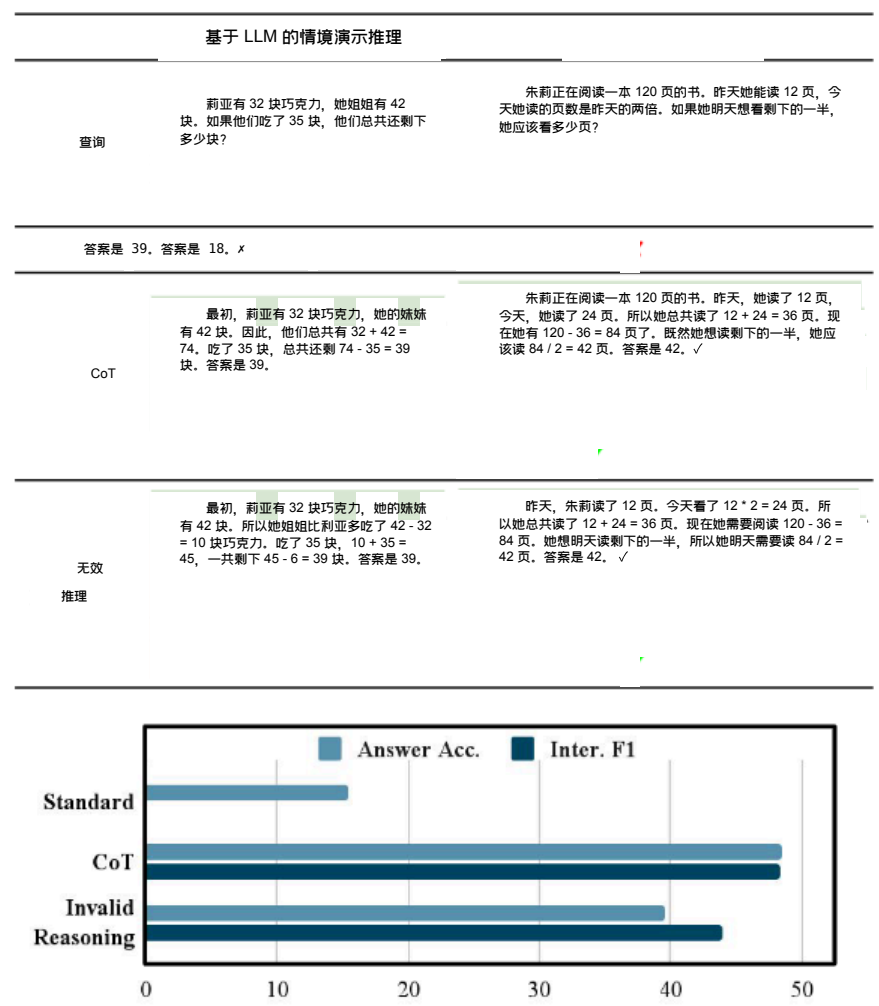

图 1：标准提示、思维链（CoT）提示和我们的消融设置与无效推理（§4）的结果。我们展示了算术推理的一个演示示例和一个推理示例，其中基本原理是彩色的（绿色：有效，黄色：无效）。我们发现，有效的推理的演示事项只有一小部分的性能 CoT-通过提供理由与无效的推理，LLM 可以实现超过 80-90%的性能 CoT 根据各种指标，同时执行逻辑健全和相关的推理。

1　介绍

大型语言模型（LLM）可以在推理过程中执行新的任务，当提示一些演示时（Brown et al.，2020 年）。思想链（CoT）提示（Wei 等人，2022）可以（图 1）提高足够大的 LLM 进行复杂和多步推理的能力。除了（查询，答案）示例对演示之外，CoT 提示还包括每个示例的基本原理（图 1 中的彩色部分），即，一系列的推理步骤的答案，这鼓励 LLM 显式地生成其中间推理过程之前预测的最终答案。尽管它取得了成功，但人们对是什么使 CoT 激励有效却知之甚少

以及所演示的推理步骤的哪些方面有助于其性能。最近的研究结果还表明，在上下文中学习可能与微调/训练有很大的不同;例如，Min 等人（2022）和 Webson 和 Pavlick（2022）表明，在上下文中提供随机标签或误导性指令只会轻微损害某些任务的模型性能。受这项工作的启发，我们仔细研究了 CoT 提示，以研究它是如何工作的以及为什么工作。

[1] 我们的代码和模型输入/输出可以在这里找到。

我们设计了一系列的烧蚀实验

where we deliberately change different aspects of the demonstrated rationales and measure how the model performance varies accordingly (§4, §5). On two representative multi-step reasoning tasks—arithmetic reasoning and multi-hop factual question answering (QA), we find that **the validity of reasoning matters only a small portion to the performance**—by providing rationales with completely invalid reasoning steps, the LLM can still achieve over 80-90% of the performance of CoT under various metrics while generating coherent lines of reasoning towards the answer (§4). Through further examinations, we identify and formulate other aspects of a CoT rationale (§5), and find that **being relevant to the query and correctly ordering the reasoning steps are the key** for the effectiveness of CoT prompting.

Overall, our findings suggest that what LLMs *learn* about how to reason under CoT prompting could be limited. Rather, they have already gained a lot of such "reasoning abilities" from pretraining, and the demonstrations may mainly specify an output space/format that regularizes the model generation to look step-by-step while being in order and relevant to the query. Our work suggests a new way of interpreting the evaluation scores in view of the prior knowledge LLMs possess, and leads to reflections on benchmarking few-shot reasoning which we discuss in §6.

2 Background & Study Formulation

Chain-of-Thought (CoT) prompting. Different from the standard way of prompting language models where a set of (query, answer) pairs are given as demonstrations (Brown et al., 2020), CoT prompting (Wei et al., 2022) additionally includes a rationale (Figure 1, colored) for each example, encouraging the model to verbalize the intermediate reasoning steps for solving the task. Such a technique has been shown to improve the performance of LLMs with sufficient scale on complex reasoning, sometimes to a large degree especially on arithmetic reasoning, multi-hop question answering, and symbolic reasoning.

Components of a CoT rationale. We identify two distinct components of a CoT rationale (examples in Table 1):

- Bridging objects: the key and necessary objects that the model needs to traverse in order to make a successful final prediction. For arithmetic reasoning, the bridging objects are defined to be the numeric part (numbers & equations) of the rationale, and for factual QA, the bridging objects are defined to be the subject & object entities.

- Language templates: the complementary parts of bridging objects, which serve as textual hints and relations/predicates that guide the model to derive the correct bridging objects along the reasoning process.

Arithmetic Reasoning	Multi-hop QA
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Q: Who is the grandchild of Dambar Shah?
A: Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	A: Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Table 1: Bridging objects and language templates of a Chain-of-Thought rationale. Here we illustrate with one in-context exemplar for each task we experiment with.

Research questions. In Chain-of-Thought prompting, correct bridging objects and language templates are provided as demonstrations to show the LLM how to reason. While CoT achieves impressive performance, we are interested in the following questions: *are ground truth bridging objects/language templates important? If not, what would be the key aspects that are needed for the LLM to reason properly?* These questions are the main focus of our study, which will be discussed in §4 and §5.

3 Experimental Setup

3.1 Datasets & In-context Exemplars

We experiment on two representative tasks involving multi-step reasoning: arithmetic reasoning & multi-hop factual question answering (QA). We select benchmarks on which CoT prompting brings significant improvements over standard prompting, as shown in previous work (Wei et al., 2022; Press et al., 2022); they are more suitable for our study, since our goal is to understand how different aspects of the Chain-of-Thought rationales contribute to the performance of CoT prompting. For arithmetic reasoning, we experiment on GSM8K (Cobbe et al., 2021), one of the most challeng-

我们故意改变所证明的原理的不同方面，并测量模型性能如何相应地变化（§4，§5）。在两个有代表性的多步推理任务-算术推理和多跳事实问题回答（QA）上，我们发现推理的有效性对性能的影响很小-通过提供完全无效的推理步骤，LLM 仍然可以在各种度量下实现 CoT 的 80-90%以上的性能，同时生成对答案的连贯推理线（§4）。通过进一步的研究，我们识别并制定了 CoT 基本原理的其他方面（§5），并发现与查询相关并正确排序推理步骤是 CoT 提示有效性的关键。

总的来说，我们的研究结果表明，LLM 学习如何在 CoT 提示下推理可能是有限的。相反，他们已经从预训练中获得了很多这样的“推理能力”，并且演示可能主要指定输出空间/格式，该输出空间/格式将模型生成规则化，以便在有序且与查询相关的同时逐步查看。我们的工作提出了一种新的方式来解释评估分数的先验知识 LLM 拥有，并导致对基准测试少数镜头推理的反思，我们在§6 中讨论。

2 背景和研究方案

思路链（CoT）提示。不同

根据提示语言模型的标准方式，其中给出一组（查询，回答）对作为示范（Brown 等人，2020）、CoT 提示（Wei 等人，2022）还包括每个示例的基本原理（图 1，彩色），鼓励模型用语言描述解决任务的中间推理步骤。这种技术已被证明可以提高 LLM 在复杂推理上的性能，有时在很大程度上，特别是在算术推理，多跳问答和符号推理上。

CoT 原理的组成部分。我们确定了 CoT 原理的两个不同组成部分（表 1 中的示例）：

- 桥接对象：模型需要遍历的关键和必要对象，以便成功进行最终预测。对于算术推理，桥接对象被定义为基本原理的数字部分（数字和等式），而对于事实 QA，桥接对象被定义为主体和客体实体。

算术推理	多跳 QA
问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？	问：谁是丹巴尔沙阿的孙子？
A：原来，莉亚有 32 块巧克力，她姐姐有 42 块。因此，他们总共有 32 + 42 = 74。吃了 35 块，总共还剩 74 - 35 = 39 块。答案是 39。	答：丹巴尔沙阿（？- 1645 年，他是 Krishna Shah 的父亲。 Rudra Shah 是 Krishna Shah 的孩子。- 1661）。所以最后的答案（孙子的名字）是：Rudra Shah。

表 1：思想链原理的桥接对象和语言模板。在这里，我们用一个上下文范例来说明我们实验的每个任务。

- 语言模板：桥接对象的补充部分，用作文本提示和关系/谓词，引导模型沿着推理过程导出正确的桥接对象。

研究问题。在思想链提示中，提供了正确的桥接对象和语言模板作为演示，向 LLM 展示如何推理。虽然 CoT 实现了令人印象深刻的性能，但我们对以下问题感兴趣：地面实况桥接对象/语言模板重要吗？如果不是，LLM 正确推理所需的关键方面是什么？这些问题是我们研究的主要焦点，我们将在第 4 和第 5 节中讨论。

3 实验装置

3.1 数据集和上下文示例

我们对两个涉及多步推理的代表性任务进行了实验：算术推理和多跳事实问题回答（QA）。我们选择了 CoT 提示比标准提示带来显著改进的基准，如以前的工作所示（Wei 等人，2022 年; Press 等人，2022 年）;它们更适合我们的研究，因为我们的目标是了解思维链原理的不同方面如何有助于 CoT 提示的表现。对于算术推理，我们在 GSM 8 K 上进行了实验（Cobbe 等人，2021 年），最受欢迎的一个-

ing mathematical reasoning benchmarks available which is also repeatedly adopted by prior work as a key benchmark for arithmetic reasoning; for multi-hop factual QA, we experiment on Bamboogle, a dataset of compositional questions constructed by Press et al. (2022). Due to budget considerations, we uniformly sample 800 out of the 1319 test examples for GSM8K for evaluation. We evaluate on all 125 test samples for Bamboogle.

We base our experiments on the original prompt exemplars, i.e., the set of (query, rationale, answer) pairs released by Wei et al. (2022) and Press et al. (2022), with slight editing to make the structure more consistent and reduce redundancy, which makes our ablations more convenient to conduct. These edits only slightly affect the performance of CoT; we show our edited demonstration examples and include more details in Appendix A.1.

3.2 Backbone Language Model

We use InstructGPT-175B[2] (Ouyang et al., 2022; Brown et al., 2020) `text-davinci-002` as our backbone LLM, which is one of the most performant and widely-used LLMs with public APIs and has demonstrated strong performance under CoT prompting (Wei et al., 2022). We report its results and analyze them in the main content. In addition, we also test on `text-davinci-003` (a very recent improved version of `text-davinci-002`), PaLM (Chowdhery et al., 2022) and Flan-PaLM (Chung et al., 2022), where the results and discussion could be found in Appendix A.3. All generations are done by greedy decoding (i.e., sampling with zero temperature) as in the original CoT work (Wei et al., 2022).

3.3 Evaluation

Prior work mainly performs evaluation using the correctness of the final answer, which could be viewed as an *extrinsic* way of assessing the predicted rationales. However, this may not align well with the actual quality of the rationale in many cases, as mentioned in Huang and Chang (2022). For example, a rationale that is correct for all but the last step (and hence derives the wrong final answer) would still be assigned a zero score, while a rationale that is wrong/incomplete but reaches the correct final answer would be assigned a full score. Therefore, in addition to extrinsic evaluation (**Answer Accuracy** for GSM8K, **Answer F1** for Bamboogle), we perform *intrinsic* evaluation where we measure the Recall/F1 (**Inter.**[3] **Recall/F1**) of the bridging objects which need to be derived by the LLM (i.e., those that do not appear in the query). For GSM8K, since annotations for ground truth reasoning steps are available, we use the derived numbers in the annotated steps as a proxy for bridging objects.[4] For Bamboogle, we manually annotate the bridging objects (intermediate entities) and measure their recall. While it is still possible for the model to reach correct bridging objects with the wrong language templates, we manually verify that this rarely happens; details are included in Appendix A.2.

4 How Much Does Valid Reasoning Matter?

Intuitively, one of the most important aspects of a Chain-of-Thought rationale would be its logically valid and sound reasoning. If we provide rationales with invalid reasoning steps in the demonstrated examples instead, we should expect the LLM to fail to reason properly and gain little or even negative improvements compared with standard prompting (where no rationale is given), since we are teaching the LLM to reason in the wrong way which could be even worse than not doing so at all. To test this intuition, we design an ablation study where we construct invalid reasoning steps for the demonstrated rationales, and measure its influence on model behavior.

4.1 Constructing Invalid Chain of Reasoning

We manually write rationales with invalid reasoning for *all* the in-context demonstration examples. Since our focus here is to investigate the importance of the validity of reasoning, we only ablate the parts in a CoT rationale which are involved with derivations that are logically sound and helpful for answering the query. More specifically, we keep the premise steps which are copies/paraphrases of facts from the query, and change the subsequent steps such that they do not logically derive the final answer. Importantly, we are *not* adopting an adversarial/counterfactual perturbation setting where

[2]We also tried the original GPT-3 175B without instruction-finetuning in our preliminary experiments, but find that CoT prompting does not yield much performance gain than standard prompting, echoing Fu et al. (2022).

[3]Abbreviation for "Intermediate".

[4]We do not use whole equations since we observe that the LLM may express the mathematical equation in different ways, e.g., "5 plus 3 is 8", "5 + 3 = 8".

现有的数学推理基准也被先前的工作反复采用作为算术推理的关键基准;对于多跳事实 QA，我们在 Bambobaly 上进行了实验，这是 Press 等人构建的组成问题数据集。出于预算考虑，我们从 1319 个 GSM 8 K 测试示例中统一抽取 800 个进行评估。我们对所有 125 个测试样本进行了评估。

我们的实验基于原始提示样本，即，Wei 等人（2022）和 Press 等人（2022）发布的（查询、理由、答案）对集合，略微编辑以使结构更加一致并减少冗余，这使得我们的消融更便于进行。这些编辑只会轻微影响 CoT 的性能;我们在附录 A.1 中展示了编辑后的演示示例并包含更多细节。

3.2 主干语言模型

我们使用 InstructGPT-175 B（Ouyang 等人，2022; Brown 等人，2020）text-davinci-002 作为我们的骨干 LLM，这是具有公共 API 的最具性能和广泛使用的 LLM 之一，并且在 CoT 提示下表现出强大的性能（Wei 等人，2022 年）。我们报告了它的结果，并在主要内容中进行了分析。 此外，我们还测试了 text-davinci-003（text-davinci-002 的最新改进版本）、PaLM（Chowdhery 等人，2022）和 Flan-PaLM（Chung 等人，2022），其结果和讨论见附录 A.3。所有的生成都是通过贪婪解码来完成的（即，在零温度下取样），如在原始 CoT 工作中那样（Wei 等人，2022 年）。

3.3 评价

先前的工作主要是使用最终答案的正确性进行评估，这可以被视为评估预测依据的外在方式。然而，在许多情况下，这可能与理论依据的实际质量不一致，如 Huang 和 Chang（2022）所述。例如，对于除了最后一步之外的所有步骤都是正确的（因此得出错误的最终答案）的基本原理仍将被分配零分，而错误/不完整但达到正确的最终答案的基本原理将被分配满分。

得分因此，除了外在的评价，

（GSM 8 K 应答准确度，

F1 for Bambosit），我们执行内在评估，其中我们测量需要由 LLM 导出的桥接对象的召回/F1（Inter.Recall/F1）（即，那些没有出现在查询中的）。对于 GSM 8 K，由于地面真值推理步骤的注释是可用的，因此我们使用注释步骤中的派生数字作为桥接对象的代理;对于 BamboCast，我们手动注释桥接对象（中间实体）并测量它们的召回率。虽然模型仍然有可能用错误的语言模板到达正确的桥接对象，但我们手动验证这种情况很少发生;细节包含在附录 A.2 中。

4 有效的推理有多重要？

直觉上，思维链理论最重要的方面之一是它的逻辑有效和合理的推理。如果我们在演示的例子中提供了无效推理步骤的基本原理，我们应该期望 LLM 无法正确推理，并且与标准提示（没有给出基本原理）相比，获得很少甚至是负面的改进，因为我们正在教 LLM 以错误的方式推理，这可能比根本不这样做更糟糕。为了测试这种直觉，我们设计了一个消融研究，我们构建了无效的推理步骤所证明的理由，并衡量其对模型行为的影响。

4.1 构造无效推理链

我们手动编写的理由与无效的推理，所有的上下文演示示例。由于我们在这里的重点是调查推理的有效性的重要性，我们只切除了 CoT 基本原理中涉及逻辑上合理的推导并有助于回答查询的部分。更具体地说，我们保留了前提步骤，这些步骤是查询中事实的副本/释义，并更改后续步骤，以便它们在逻辑上不会导出最终答案。重要的是，我们不采用对抗/反事实扰动设置，

[2] 我们还在初步实验中尝试了原始的 GPT-3 175 B，但发现 CoT 提示并没有比标准提示产生太多的性能增益，这与 Fu et al.（2022）相呼应。

[3] “中间体”的缩写。

[4] 我们不使用整个方程，因为我们观察到 LLM 可以以不同的方式表达数学方程，例如，“5 加 3 等于 8”，“5 + 3 = 8”。

minimal alterations are applied to make the reasoning invalid; instead, we apply rather drastic changes where we change both the bridging objects and language templates and hence little valid reasoning exists to help solve the query. The full prompts in our setting are included in Appendix A.4.

For example, consider an in-context demonstration (see ① in Table 4) for arithmetic reasoning. Here the query is *"Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?"*. For the 1st entailment step which should sum *"32"* and *"42"* to get the total amount *"32 + 42 = 74"* as in CoT, we instead write *"So her sister had 42 - 32 = 10 chocolates more than Leah has."* which has both the wrong bridging object and language template, and is completely unhelpful for solving the problem. The subsequent steps are written based on the previous steps, and in the end, answer the question whereas the rationale does not in any way lead to the answer logically. While the step itself still describes something that could be entailed in the example we just gave, this is not the case generally and most of the steps we write are neither helpful nor entailments from earlier steps. For example, the next step *"After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total"* makes use of unwarranted information (*"6"*) and has no valid entailment anywhere. We illustrate our construction using another example for factual QA, where the question is *"Who is the grandchild of Dambar Shah?"*. Here, we write a rationale that finds the kingdom of *"Dambar Shah"* and then a child of the person who established the kingdom, which does not lead to *"the grandchild of Dambar Shah"*.

4.2 Results & Analysis

Quantitative results. Table 2 summarizes the quantitative results for `text-davinci-002`. We include additional results and discussion for `text-davinci-003`, PaLM and Flan-PaLM in Appendix A.3. LLMs can achieve surprisingly high performance when provided with invalid reasoning steps for the demonstrations (①). In particular, under **Inter. Recall/Inter.F1**, i.e., intrinsic evaluation, which is arguably a more faithful measurement of the rationale quality (§3.3), all LLMs we tested can retain over 90% of the performance achieved under CoT prompting.

For GSM8K where there are large variations in the difficulty levels (here, we use the number of reasoning steps required to solve a problem as its difficulty level) of the problem instances, we additionally examine the model performance separately for each difficulty level. The results are shown in Figure 2. The performance drop is also uniform across samples with different levels of difficulty. At the instance level, after omitting samples where both settings get the correct/wrong answer, there is a significant portion for the remaining ones (62/196 for GSM8K, 6/20 for Bamboogle) where CoT gets the wrong answer and the invalid reasoning setting gets the correct answer. This further strengthens the finding that there is no strong connection between the reasoning validity of the demonstrations and the quality of the model predictions.

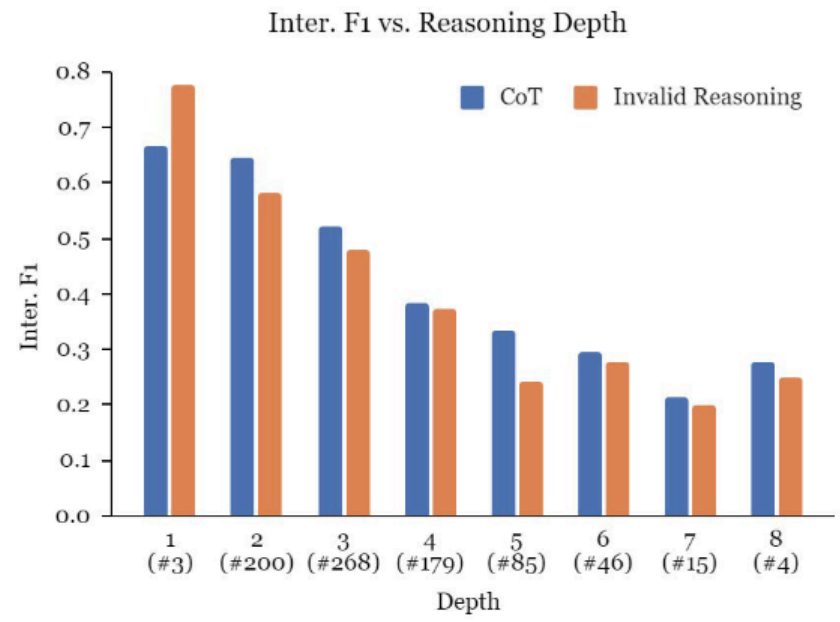

Figure 2: Model performance using CoT and demonstrations with invalid reasoning for examples with different reasoning depths on GSM8K. The number of samples for each reasoning depth is shown below (led by *"#"*). The performance drop is consistent across different levels of difficulty.

Qualitative analysis. By checking the generated rationales for the invalid reasoning setting, we find that overall they look indistinguishable from the rationales generated by CoT prompting. In almost all cases where the predicted final answer is correct, the rationales do reach the answer with valid and sound reasoning steps (as in CoT), drastically different from those in the given demonstrations; for cases where the final answer is wrong, the errors the LLM makes are also in the same types with the errors made under CoT prompting. To compare the distribution of errors between CoT and the invalid reasoning setting, we examine 20 samples from GSM8K where CoT gets the correct final answer and the invalid reasoning setting gets the wrong answer, and another 20 examples for the opposite case. We use the same error categorizations as in

应用最小的改变以使推理无效;相反，我们应用相当剧烈的改变，其中我们改变桥接对象和语言模板两者，并且因此几乎不存在有效的推理来帮助解决查询。我们设置中的完整提示包含在附录 A.4 中。

例如，考虑算术推理的上下文演示（参见表 4 中的 1）。

这里的查询是“Leah 有 32 块巧克力，她的妹妹有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？”对于第一个蕴涵步骤，应该将“32”和“42”相加以得到总量“32 + 42 = 74”，如 CoT 中所示，我们改为写“所以她的妹妹有 42 - 32 = 10 个巧克力，

莉亚有。”它既有错误的桥接对象，又有错误的语言模板，对解决问题完全没有帮助。接下来的步骤是基于前面的步骤编写的，最后，回答问题，而基本原理并不以任何方式导致逻辑上的答案。虽然步骤本身仍然描述了我们刚刚给出的例子中可能包含的东西，但通常情况下并非如此，我们编写的大多数步骤既没有帮助，也没有从前面的步骤中导出。例如，下一个步骤“在吃了 35 块之后，由于 10 + 35 = 45，他们总共剩下 45 - 6 = 39 块”使用了无根据的信息（“6”），并且在任何地方都没有有效的蕴涵。我们用另一个事实问答的例子来说明我们的结构，其中的问题是“谁是

丹巴尔·沙阿的孙子？”在这里，我们写一个

这是一个合理的解释，即发现“丹巴尔沙阿”的王国，然后是建立王国的人的孩子，这并不导致“丹巴尔沙阿的孙子”。

4.2 结果与分析

定量结果。表 2 总结了 text-davinci-002 的定量结果。我们在附录 A.3 中纳入了 text-davinci-003、PaLM 和 Flan-PaLM 的其他结果和讨论。当为演示提供无效的推理步骤时，LLM 可以实现令人惊讶的高性能（1）。特别是在国际米兰。调用/Inter.F1，即，内在评估，这可以说是一个更忠实的衡量基本原理质量（§3.3），我们测试的所有 LLM 可以保持超过 90%的性能下实现 CoT 提示。

对于 GSM 8 K，难度级别变化很大（在此，我们使用

解决问题所需的推理步骤作为其难度级别）的问题实例，我们还检查了模型的性能分别为每个难度级别。结果如图 2 所示。在不同难度的样本中，性能下降也是一致的。在实例级别，在省略两个设置都得到正确/错误答案的样本后，剩余的样本中有相当大的一部分（GSM 8 K 为 62/196，BamboCan 为 6/20）CoT 得到错误答案，无效推理设置得到正确答案。这进一步加强了这一发现，即在演示的推理有效性和模型预测的质量之间没有很强的联系。

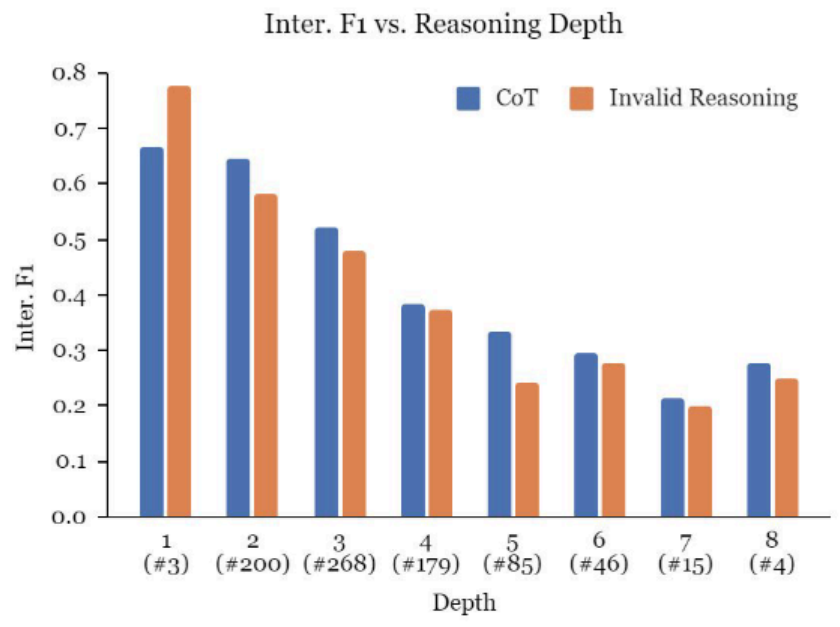

图 2：使用 CoT 的模型性能以及在 GSM 8K 上具有不同推理深度的示例的无效推理演示。每个推理深度的样本数量如下所示（以“#”开头）。性能下降在不同的难度水平上是一致的。

定性分析通过检查生成的理由无效的推理设置，我们发现，总体上看，他们看起来从 CoT 提示生成的理由没有区别。在几乎所有的情况下，预测的最终答案是正确的，基本原理确实达到了有效和合理的推理步骤（如 CoT），与给定的演示中的那些完全不同的答案;对于最终答案是错误的情况下，LLM 犯的错误也与 CoT 提示下犯的错误类型相同。为了比较 CoT 和无效推理设置之间的错误分布，我们检查了来自 GSM 8 K 的 20 个样本，其中 CoT 得到正确的最终答案，无效推理设置得到错误的答案，以及另外 20 个相反情况的例子。我们使用与

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.4	N/A	20.6
CoT (Chain-of-Thought prompting)	43.9	48.3	48.5	45.2	45.2
① Invalid Reasoning	39.8	43.9	39.5	44.4	39.4
② No *coherence* for bridging objects	35.3	39.2	35.8	40.8	37.4
③ No relevance for bridging objects	21.4	26.2	27.5	39.6	34.0
④ No *coherence* for language templates	24.1	28.3	25.8	35.2	32.1
⑤ No relevance for language templates	29.5	34.0	32.8	40.4	29.4
⑥ No *coherence*	25.2	29.4	23.1	39.6	33.8
⑦ No relevance	9.6	11.9	11.0	36.8	23.9

Table 2: Intrinsic and extrinsic evaluation results under InstructGPT (`text-davinci-002`) for all settings in our experiments. Results for `text-davinci-003`, PaLM and Flan-PaLM could be found in Appendix A.3.

Error Types	**CoT** correct & **IR** wrong	**CoT** wrong & **IR** correct
Calculation	20%	20%
One step missing	35%	25%
Semantic understanding	45%	55%

Table 3: Distribution of error types of 20 examples from GSM8K where Chain-of-Thought (**CoT**) prompting reaches the correct answer and the Invalid Reasoning setting (**IR**) reaches a wrong answer, and 20 examples for the opposite case.

Wei et al. (2022) for the qualitative analysis, and summarize the results in Table 3. The distributions of errors in both cases are highly similar.

Summary. Combining the quantitative and qualitative results, we can see that there is a low chance for any systematic difference between CoT and the invalid reasoning setting to exist. The LLM still tries and manages to generate logically sound and pertinent reasoning decently, and ablating the validity of reasoning for the demonstrations only brings a small performance degradation. This opens the question: *If valid reasoning is not required, what are the key aspects that determine the effectiveness of CoT prompting?*

5 What are the Key Aspects of Chain-of-Thoughts?

Re-examining the rationales in our ablation setting in §4, we can find that even though the reasoning is invalid, they have the following properties:

- The rationales still use information from the query; more specifically, they still start from bridging objects mentioned in the query, and the language templates are related to the query. Recall our running example for arithmetic reasoning (Table 4), even though the reasoning here is wrong, the numbers *"32"* and *"42"* are kept from the query, and the language templates are still about *"Leah"*, *"Leah's sister"* and *"Chocolates"*, and try to seek the answer to the query. Therefore, the rationale is still relevant to the query being asked.
- Each step of a rationale still follows the previous steps. Using again the same example, the bridging object (equation in this case) *"42 - 32 = 10"* in the first entailment step uses numbers from previous steps; likewise, the language template *"So her sister had _ chocolates more than Leah has"* is something that follows after the earlier steps. Hence, overall, the rationale still appears to be coherent.

We formulate two notions that capture these two aspects of a rationale in what follows.

Relevance. A component of the rationale has relevance if it is based on the corresponding component from the query. For bridging objects, this could be formally defined as using the exact same objects mentioned in the query (numbers for arithmetic reasoning and entities for factual QA); for language templates, they have relevance if they are still about the same set of entities/relations as the query, and allude to the question being asked. For example, a template about *"Patricia"* and *"hair"* would not have relevance to a query about *"Leah"* and *"Chocolates"*, and similarly, a template that attempts to find the *"brother-in-law"* of the topic entity does not have relevance to a query which seeks the *"grandchild"* (Table 4).

		GSM 8K 标准			班博什	
		中间截留召回国际米兰。F1 应答接入接口回忆答案 F1				
STD（标准提示）		N/A	N/A	15.4	N/A	20.6
CoT（思维链提示）		43.9	48.3	48.5	45.2	45.2
◯ 1	无效推理	39.8	43.9	39.5	44.4	39.4
◯ 2	桥接对象无连贯性	35.3	39.2	35.8	40.8	37.4
◯ 3	与桥接对象无关	21.4	26.2	27.5	39.6	34.0
◯ 4	语言模板没有连贯性	24.1	28.3	25.8	35.2	32.1
◯ 5	与语言模板无关	29.5	34.0	32.8	40.4	29.4
◯ 6	无连贯性	25.2	29.4	23.1	39.6	33.8
◯ 7	无关	9.6	11.9	11.0	36.8	23.9

表 2：在我们的实验中，对于所有设置，在 InstructGPT（text-davinci-002）下的内在和外在评估结果。text-davinci-003、PaLM 和 Flan-PaLM 的结果见附录 A.3。

错误类型	CoT 正确 IR 错误（& R）	CoT 错误 IR 正确（& R）
计算	20%	20%
少了一步	35%	25%
语义理解	45%	55%

表三：来自 GSM 8 K 的 20 个示例的错误类型的分布，其中思想链（CoT）提示达到正确答案，无效推理设置（IR）达到错误答案，以及 20 个相反情况的示例。

Wei 等人（2022）进行定性分析，并将结果总结在表 3 中。两种情况下的误差分布非常相似。摘要结合定量和定性的结果，我们可以看到，CoT 和无效推理设置之间存在任何系统差异的可能性很低。LLM 仍然试图并设法生成逻辑上合理的和适当的推理体面，和消融的有效性推理的示范只带来一个小的性能下降。这将打开

问题：如果不需要有效的推理，那么决定 CoT 提示有效性的关键因素是什么？

5 思维链的关键方面是什么？

重新检查第 4 节中消融设置的基本原理，我们可以发现，即使推理无效，它们具有以下特性：

- 基本原理仍然使用来自查询的信息;更具体地说，它们仍然从查询中提到的桥接对象开始

并且语言模板与查询相关。回想一下我们的算术推理运行示例（表 4），即使这里的推理是错误的，数字“32”和“42”从查询中保留下来，语言模板仍然是关于“Leah”，“Leah's sister”和“Chocolates”，并尝试寻找查询的答案。因此，理由仍然与提出的质询相关。

- 基本原理的每一步仍然遵循前面的步骤。再次使用相同的例子，第一个蕴涵步骤中的桥接对象（在这种情况下为等式）“42 - 32 = 10”使用了前面步骤中的数字;同样，语言模板“So her sister had _ chocolates more than Leah has”是前面步骤之后的内容。因此，总的来说，理由似乎仍然是一致的。

我们在下文中提出了两个概念，以捕捉这两个方面的基本原理。本案无关如果依据的组成部分基于查询中的相应组成部分，则该组成部分具有相关性。对于桥接对象，这可以正式定义为使用查询中提到的完全相同的对象（算术推理的数字和事实 QA 的实体）;对于语言模板，如果它们仍然与查询相同的实体/关系集，并且暗示正在询问的问题，则它们具有相关性。例如，关于“Patricia”和“hair”的模板将与关于“Leah”和“Chocolates”的查询不相关，并且类似地，试图找到主题实体的“brother-in-law”的模板与寻找“grandchild”的查询不相关（表 4）。

Coherence. A component of the rationale has coherence if it is in the correct order, i.e., later steps could not be pre-conditions for earlier steps and reversely, earlier steps could not be based on later steps. For example, a rationale where *"32 + 42 = 74"* appears before the introduction of *"32"* or *"42"* would not have coherence on bridging objects, and similarly for language templates.

In what follows, we design a set of ablation settings to examine the impact of these two aspects for different components of a CoT-like rationale.

5.1 Ablation Settings

In order not to introduce mixed effects which could make the results not well-controlled, we base the ablation settings on top of the CoT prompts instead of the setting in §4.

Given the two components (bridging objects and language templates) and the two aspects (relevance and coherence) of the rationale, there are naturally four ablation settings where each could examine one aspect of a certain component. We also experiment with two other settings: no relevance where neither bridging objects nor language templates have relevance, and *no coherence* which is defined analogously (⑥, ⑦ in Table 4).

Destroying relevance. We perform random substitutions to ablate the relevance of a certain component. For ablating the relevance of bridging objects, we randomly sample alternatives (numbers for GSM8K, entities for Bamboogle) for those from the query, and change the bridging objects in the subsequent steps correspondingly to maintain the coherence of the rationale. Using our running example, we randomly replace the bridging objects from the query: *"32"*→*"19"*, *"42"*→*"31"* and *"35"*→*"29"*, then change the bridging object from the first entailment step from *"32 + 42 = 74"* to *"19 + 31 = 50"*, and so on so forth. To ablate the relevance of language templates, for GSM8K, we randomly sample an annotated rationale from the training set, and use its template in place of the original template. For Bamboogle, we manually replace the template with an alternative which is irrelevant to the query.

Destroying coherence. Ablating the coherence is rather straightforward, where we randomly shuffle the components and permute their orderings.

5.2 Results & Analysis

The results could be found in Table 2, and we include additional results for `text-davinci-003`, PaLM and Flan-PaLM in Appendix A.3. We summarize the main findings in what follows.

Relevance and coherence are key for the performance of CoT prompting. It can be seen that most of the settings for this section (②-⑦) have rather large performance drops from CoT, where the low-performing ones approach or even underperform standard prompting. This suggests that overall, relevance and coherence are key for the performance of CoT.

Keeping relevance is crucial. The no relevance setting ⑦ where both components of the rationale have no relevance achieves significantly poorer performance than other ablation settings, and even underperforms standard prompting (STD) where no rationale is given on GSM8K. To see why such low performance happens, we manually examine the generated rationales under this setting for 20 examples on GSM8K. We find that the LLM is indeed generating irrelevant rationales (both bridging objects and language templates) for 15 out of 20 examples. Many of the rationales have recurring topics (e.g., "cats and dogs", "passengers and buses") which we hypothesize are frequent patterns in the portion relevant to mathematics in the pretraining corpora. Overall, this suggests that a certain level of relevance is crucial for the LLM to stick to the query being asked.

Relevance matters more than coherence for bridging objects. Providing incoherent bridging objects (②) achieves better performance than providing irrelevant bridging objects (③), especially on the more challenging GSM8K dataset (39.2 *v.s.* 26.2 **Inter. F1**). which indicates that it is important for the bridging objects to be relevant, but not as important to have them in the right order to guide the LLM along the reasoning process. We quantitatively measure the coverage of bridging objects from the query for the generated rationales, and find that the settings with no relevance for bridging objects (③, ⑦) do have significantly lower coverage (below 60%) than other settings (around 80%).

Coherence of language templates is important. Different from the coherence of bridging objects ②, the coherence of language templates ④ matters a lot to the performance of CoT prompting. By examining the predicted rationales, we find that the LLM is indeed generating rationales with incoherent language templates (14 out of 20 examples), which negatively affects reasoning.

连贯性。合理性的一个组成部分如果顺序正确，则具有连贯性，即，后来的步骤不能成为先前步骤的先决条件，而且，先前的步骤不能以后来的步骤为基础。例如，在引入“32”或“42”之前出现“32 + 42 = 74”的基本原理在桥接对象上不具有连贯性，语言模板也是如此。

在下文中，我们设计了一组消融设置，以检查这两个方面对 CoT 类原理的不同组成部分的影响。

5.1 消融设置

为了不引入可能导致结果无法得到良好控制的混合效应，我们将消融设置基于 CoT 提示而不是§4 中的设置。

考虑到基本原理的两个组成部分（桥接对象和语言模板）和两个方面（相关性和连贯性），自然有四个消融设置，每个设置可以检查某个组成部分的一个方面。我们还试验了另外两种设置：无关联性，即桥接对象和语言模板都没有关联性，以及无连贯性，这是类似定义的（表 4 中的 6，7）。

破坏相关性。我们执行随机替换来消除某个组件的相关性。为了消除桥接对象的相关性，我们随机抽取查询中的替代项（GSM 8K 的数字，Bambobox 的实体），并在随后的步骤中相应地改变桥接对象以保持原理的一致性。使用我们正在运行的示例，我们随机替换查询中的桥接对象：“32”→“19”，“42”→“31”和“35”→“29”，然后将第一个蕴涵步骤中的桥接对象从“32 + 42 = 74”更改为“19 + 31 = 50”，依此类推。为了消除语言模板的相关性，对于 GSM8K，我们从训练集中随机抽取一个注释的基本原理，并使用其模板代替原始模板。对于 BamboCast，我们手动将模板替换为与查询无关的替代模板。

破坏连贯性。增强相干性是相当简单的，我们随机地打乱组件并排列它们的顺序。

5.2 结果与分析

结果见表 2，我们还包括 text-davinci-003 的其他结果，PaLM 和 Flan-PaLM 见附录 A.3。我们将主要发现总结如下。

相关性和连贯性是 CoT 激励绩效的关键。可以看出，该部分（2 - 7）的大多数设置相对于 CoT 具有相当大的性能下降，其中低性能的设置接近甚至低于标准提示。这表明，总体而言，相关性和一致性是 CoT 绩效的关键。

保持相关性至关重要。无相关性设置 7（其中两个基本原理组件均无相关性）的性能明显低于其他消融设置，甚至低于 GSM 8K 上未给出基本原理的标准提示（STD）。为了了解为什么会发生这种低性能，我们手动检查了在此设置下生成的理由，并在 GSM8K 上进行了 20 个示例。我们发现，LLM 确实为 20 个示例中的 15 个生成了不相关的原理（桥接对象和语言模板）。许多理由都有重复出现的主题（例如，“猫和狗”，“乘客和公共汽车”），我们假设这些是预训练语料库中与数学相关的部分中的频繁模式。总的来说，这表明一定程度的相关性对于 LLM 坚持所问的查询至关重要。

对于连接对象而言，相关性比连贯性更重要。提供不相干的桥接对象（2）比提供不相关的桥接对象（3）实现了更好的性能，特别是在更具挑战性的 GSM 8K 数据集上（39.2 vs. 26.2 中间截留 F1）。这表明，重要的是桥接对象是相关的，但重要的是使它们以正确的顺序引导 LLM 沿着推理过程。我们从生成的基本原理查询中定量测量桥接对象的覆盖率，发现与桥接对象无关的设置（3，7）的覆盖率（低于 60%）明显低于其他设置（约 80%）。

语言模板的一致性很重要。

不同于衔接对象的连贯性

语言模板 4 的连贯性对 CoT 提示的效果有很大影响。通过检查预测的基本原理，我们发现 LLM 确实是用不连贯的语言模板生成基本原理（20 个例子中有 14 个），这对推理产生了负面影响。

Prompt Setting	**Example Query (Arithmetic Reasoning)** *Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?*	**Example Query (Factual QA)** *Who is the grandchild of Dambar Shah?*
STD (Standard prompting)	39	So the final answer is: Rudra Shah.
CoT (Chain-of-Thought)	Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
① Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the king of the Gorkha Kingdom. The Gorkha Kingdom was established by Prince Dravya Shah. Dravya Shah has a child named Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
② No *coherence* for bridging objects	Originally, Leah had 32 + 42 = 74 chocolates and her sister had 32. So in total they had 74 - 35 = 39. After eating 35, they had 42 pieces left in total. The answer is 39.	Krishna Shah was the father of Rudra Shah. Dambar Shah (? - 1645) was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
③ No relevance for bridging objects	Originally, Leah had 19 chocolates and her sister had 31. So in total they had 19 + 31 = 50. After eating 29, they had 50 - 29 = 21 pieces left in total. The answer is 21.	Metis Amando was the father of David Amando. Randall Amando was the child of David Amando. So the final answer (the name of the grandchild) is: Randall Amando.
④ No *coherence* for language templates	After eating 32, they had 42 pieces left in total. Originally, Leah had 32 + 42 = 74 chocolates and her sister had 35. So in total they had 74 - 35 = 39. The answer is 39.	Dambar Shah (? - 1645) was the child of Krishna Shah. Krishna Shah (? - 1661) was the father of Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
⑤ No relevance for language templates	Patricia needs to donate 32 inches, and wants her hair to be 42 inches long after the donation. Her hair is 35 inches long currently. Her hair needs to be 32 + 42 = 74 inches long when she cuts it. So she needs to grow 74 - 35 = 39 more inches. The answer is 39.	The husband of Dambar Shah (? - 1645) is Krishna Shah. Krishna Shah (? - 1661) has a brother called Rudra Shah. So the final answer (the name of the brother-in-law) is: Rudra Shah.
⑥ No *coherence*	After eating 32 + 42 = 74, they had 32 pieces left in total. Originally, Leah had 74 - 35 = 39 chocolates and her sister had 35. So in total they had 42. The answer is 39.	Krishna Shah was the child of Rudra Shah. Dambar Shah (? - 1645) was the father of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
⑦ No relevance	Patricia needs to donate 19 inches, and wants her hair to be 31 inches long after the donation. Her hair is 29 inches long currently. Her hair needs to be 19 + 31 = 50 inc long when she cuts it. So she needs to grow 50 - 29 = 21 more inches. The answer is 21.	The husband of Metis Amando is David Amando. David Amando has a brother called Randall Amando. So the final answer (the name of the brother-in-law) is: Randall Amando.

Table 4: Examples for all settings in our experiments.

6 Discussion

The results from §4 and §5 open up new questions regarding learning to reason in context for LLMs, which we discuss next.

***Do* LLMs learn to reason from CoT demonstrations?** Given the surprisingly high performance obtained by ablating the validity of reasoning for the in-context rationales, it can be concluded that what the LLM learns from the demonstrations about how to reason properly is limited—rather, the LLM has already gained a lot of such complex reasoning ability from pretraining (at least for tasks we experiment on), and the provided reasoning steps serve more as the role of an output format/space, that regularizes the LLM to generate rationales that look step-by-step while being coherent and relevant to the query. Moreover, results obtained from recent stronger models including `text-davinci-003` and Flan-PaLM (see Appendix A.3) suggest that LLMs suffer further less from the ablations when they have more prior knowledge about the task. In particular, for Flan-PaLM which is directly trained on both arithmetic reasoning and factual QA in CoT fashion and hence has immense knowledge on these tasks (Chung et al., 2022), it could be seen that none of the ablations has significant impacts on its performance. On the positive side, this indicates that LLMs can effectively utilize their prior knowledge to solve new problems. However, from another perspective, if we view the invalid reasoning setting as a *task* where the goal is to generate invalid reasoning steps for the query, then the LLM has basically failed to capture the task as it still tries to predict valid reasoning steps. This leads to the concern that LLMs may over-rely on their prior knowledge and ignore important information in the context that are presumably rare in the pretraining distribution, including those that are crucial for

提示设置	算术推理（Arithmetic Reasoning） 莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？	示例查询（事实 QA） 谁是丹巴尔·沙阿的孙子？
STD（标准提示）	39	所以最后的答案是：Rudra Shah。
思想链（Chain of Thought）	最初，莉亚有 32 块巧克力，她的妹妹有 42 块。因此，他们总共有 32 + 42 = 74。吃了 35 块，总共还剩 74 - 35 = 39 块。答案是 39。	丹巴尔沙阿（？- 1645 年，他是 Krishna Shah 的父亲。Rudra Shah 是 Krishna Shah 的孩子。1661）。所以最后的答案（孙子的名字）是：Rudra Shah。
1、无效推理	最初，莉亚有 32 块巧克力，她的妹妹有 42 块。所以她姐姐比利亚多吃了 42 - 32 = 10 块巧克力。吃了 35 块，10 + 35 = 45，一共剩下 45 - 6 = 39 块。答案是 39.	丹巴尔沙阿（？- 1645 年，他是廓尔喀王国的国王。廓尔喀王国由 Dravya Shah 王子建立。Dravya Shah 有一个名叫 Rudra Shah 的孩子。所以最后的答案（孙子的名字）是：Rudra Shah。
楔接对象无连贯性	最初，Leah 有 32 + 42 = 74 块巧克力，她的妹妹有 32 块。所以他们总共有 74 - 35 = 39。吃了 35 块，一共还剩下 42 块。答案是 39。	Krishna Shah 是 Rudra Shah 的父亲。丹巴尔沙阿（？- 1645）是 Krishna Shah（？1661）。所以最后的答案（孙子的名字）是：Rudra Shah。
与桥接对象无关	最初，莉亚有 19 块巧克力，她的妹妹有 31 块。所以他们总共有 19 + 31 = 50。吃完 29 块后，总共还剩 50 - 29 = 21 块。答案是 21。	梅提斯·阿曼多是大卫·阿曼多的父亲。兰德尔·阿曼多是大卫·阿曼多的孩子。所以最后的答案（孙子的名字）是：兰德尔·阿曼多。
语言模板没有连贯性	吃了 32 块，一共还剩下 42 块。最初，Leah 有 32 + 42 = 74 块巧克力，她的妹妹有 35 块。所以他们总共有 74 - 35 = 39。答案是 39。	丹巴尔·沙阿（？- 1645 年）是 Krishna Shah 的孩子。Krishna Shah（？- 1661 年）是鲁德拉沙的父亲。所以最后的答案（孙子的名字）是：Rudra Shah。
与语言模板无关	帕特里夏需要捐赠 32 英寸，并希望她的头发是 42 英寸长后捐赠。她的头发现在有 35 英寸长。当她剪头发时，她的头发需要 32 + 42 = 74 英寸长。所以她需要再长 74 - 35 = 39 英寸。答案是 39。	丹巴尔·沙阿的丈夫（？- 第 1645 章是我的 Krishna Shah（？- 1661 年）有一个兄弟叫鲁德拉沙。所以最后的答案（姐夫的名字）是：Rudra Shah。
6 无连贯性	吃了 32 + 42 = 74，一共还剩下 32 块。最初，Leah 有 74 - 35 = 39 块巧克力，她的妹妹有 35 块。总共有 42 个。答案是 39。	Krishna Shah 是 Rudra Shah 的儿子。丹巴尔沙阿（？- 1645）是 Krishna Shah（？1661）。所以最后的答案（孙子的名字）是：Rudra Shah。
7 无相关性	帕特里夏需要捐赠 19 英寸，并希望她的头发是 31 英寸长后，捐赠。她的头发现在有 29 英寸长。当她剪头发时，她的头发需要长 19 + 31 = 50 英寸，所以她需要再长 50 - 29 = 21 英寸。答案是 21。	梅提斯·阿曼多的丈夫是大卫·阿曼多。大卫·阿曼多有个哥哥叫兰德尔·阿曼多。所以最后的答案（姐夫的名字）是：兰德尔·阿曼多。

表 4：我们实验中所有设置的示例。

6 讨论

§4 和§5 的结果为学习 LLM 的上下文推理提出了新的问题，我们将在下面讨论。

LLM 是否从 CoT 演示中学习推理？考虑到通过消除上下文推理的有效性获得的令人惊讶的高性能，可以得出结论，LLM 从演示中学习如何正确推理是有限的-相反，LLM 已经从预训练中获得了很多这种复杂的推理能力（至少对于我们实验的任务），并且所提供的推理步骤更多地充当输出格式/空间的角色，这使 LLM 规则化，以生成看起来一步一步的基本原理，同时与查询一致和相关。此外，从最近更强的模型（包括 text-davinci-003 和 Flan-PaLM）中获得的结果（参见附录 A

3)这表明，LLM 遭受更少的消融时，他们有更多的先验知识的任务。特别地，对于 Flan-PaLM，其以 CoT 方式直接在算术推理和事实 QA 上进行训练，因此在这些任务上具有丰富的知识（Chung 等人，2022），可以看出，消融对其性能没有显著影响。从积极的方面来看，这表明 LLM 可以有效地利用他们的先验知识来解决新问题。然而，从另一个角度来看，如果我们将无效推理设置视为目标是为查询生成无效推理步骤的任务，则 LLM 基本上无法捕获任务，因为它仍然试图预测有效推理步骤。这导致人们担心 LLM 可能会过度依赖他们的先验知识，并忽略背景中的重要信息，这些信息在预训练分布中可能很少见，包括那些对学习至关重要的信息。

specifying the task semantics (Jang et al., 2023).

***Can* LLMs learn to reason in-context?** We note that what we find does not in any way diminish the *potential* of learning to reason in context for LLMs; recent work has also shown evidence that learning in context is possible and could be powerful (Garg et al., 2022; Akyürek et al., 2023). Rather, our findings show that the existing successes of CoT are not sufficient for establishing that LLMs are good *few-shot learners* of reasoning; instead, the pretraining corpora have already forged them to be good reasoners on the tasks being evaluated, and the main role that the demonstrations play is to elicit such reasoning skills.

Reflections on benchmarking few-shot reasoning. An important topic on benchmarking in the era of large pre-trained language models is to quantify the level of prior knowledge the LLM has gained about the end task being evaluated, which is crucial for assessing how well can the model truly extrapolate from pretraining and acquire new skills (Chollet, 2019). One direct way is to look into the pretraining corpora when it is accessible, e.g., Razeghi et al. (2022) investigates the correlation between the model performance and the frequency of terms from the test instances in the pretraining data. However, the pretraining corpora are not always accessible, and low-level statistics are usually not adequate when the topics of interest are abstract and high-level skills such as reasoning. Along this direction, our work could be regarded as a way to approximately quantify the prior knowledge that the LLM possesses on multi-step reasoning. Our findings indicate that evaluations on alternative benchmarks where LLMs have less prior knowledge are needed to more faithfully assess the LLMs' abilities on learning to reason from few-shot demonstrations.

7 Related Work

There have been several subsequent work of Chain-of-Thought prompting since its introduction. Wang et al. (2023) proposes to sample a diverse set of reasoning paths instead of performing greedy decoding, and marginalize over the sampled paths to select the most consistent answer. Zhang et al. (2023) proposes a method for automatically constructing the in-context exemplars for CoT. Chen et al. (2022) explores program-based CoT which can better disentangle computation from reasoning. In this paper, we are primarily focused on understanding the effectiveness of the original CoT prompting method where we use the same experimental settings (e.g., greedy decoding) and base our experiments on the same few-shot exemplars used. We believe our findings could also apply to some of the subsequent variants of CoT prompting.

A few recent work focuses on understanding/analyzing CoT prompting. Madaan and Yazdanbakhsh (2022) investigates the importance of different components of the demonstrated CoT rationales by changing them to be *counterfactual*. They only experiment with limited ways of changing the rationales to be *wrong* including using incorrect calculations (e.g., *"5 + 4 = 7"*) or entities. For most of their settings, even though the rationales are made counterfactual, they are still *correct* since the query is changed accordingly (see, e.g., Table 48 of their paper). Concurrent to our work, Ye et al. (2022) also explores how the model performance could be affected by corrupting the CoT rationales. They experiment with using incorrect calculations and *dropping* (parts of) the bridging objects/language templates, which are different from our ablation designs. Saparov and He (2023) investigates systematically evaluating CoT by creating a synthetic QA dataset based on first-order logic, which allows for parsing the generated rationales into symbolic proofs for formal analysis. Overall, to our knowledge, we are the first to show that it is possible to have CoT rationales that are wrong and drastically deviate from the gold ones while still maintaining high model performance.

In general in-context learning (ICL), Min et al. (2022) shows that for a wide range of tasks in natural language understanding with categorical label space (classification and multi-choice), ground truth input-label mappings matter very little for end-task performance, and other aspects such as the label space, overall format and the distribution of text are the key. Building on this work, Yoo et al. (2022) finds that the correct input-label correspondence could have varying impacts based on the task and experimental configurations, and Wei et al. (2023) finds that models with larger scale can override semantic priors and learn input-label mapping in context. Webson and Pavlick (2022) finds that for instruction models, the performance on natural language inference tasks has small degradations under irrelevant or misleading instructions. Xie et al. (2022) provides theoretical analysis of ICL by formulating it as Bayesian inference. Our work could be viewed as an attempt to empirically under-

包括对于指定任务语义至关重要的那些（Jang 等人，2023 年）。

LLM 可以学习在上下文中推理吗？我们注意到，我们的发现并没有以任何方式削弱学习在 LLM 的背景下进行推理的潜力;最近的工作也表明，在背景下学习是可能的，并且可能是强大的（Garg 等人，2022 年; Akyürek 等人，2023 年）。相反，我们的研究结果表明，现有的成功 CoT 是不足以建立 LLM 是很好的少数拍摄学习者的推理;相反，预训练语料库已经锻造他们是很好的推理机上的任务进行评估，和演示发挥的主要作用是引出这样的推理技能。

对基准少数镜头理性的反思-

ing.在大型预训练语言模型时代，基准测试的一个重要主题是量化 LLM 获得的关于正在评估的最终任务的先验知识水平，这对于评估模型从预训练中真正推断并获得新技能的能力至关重要（Chollet，2019）。一种直接的方法是在预训练语料库可访问时查看它，例如，Razeghi et al.（2022）研究了模型性能与预训练数据中测试实例中术语频率之间的相关性。然而，预训练语料库并不总是可访问的，并且当感兴趣的主题是抽象的和高级技能（如推理）时，低级统计数据通常是不够的。沿着这个方向，我们的工作可以被视为一种方式来近似量化的先验知识，LLM 拥有的多步推理。 我们的研究结果表明，替代基准的评估，其中 LLM 有较少的先验知识，需要更忠实地评估 LLM 的能力，从几个镜头的演示学习的原因。

7 相关工作

自引入以来，有几个后续的工作链思想提示。Wang et al.（2023）建议对不同的推理路径集进行采样，而不是执行贪婪解码，并对采样路径进行边缘化，以选择最一致的答案。Zhang et al.（2023）提出了一种自动构建 CoT 上下文范例的方法。Chen et al.（2022）探索了基于程序的 CoT，它可以更好地将计算与推理分开。本文我们主要集中于理解原始 CoT 提示方法的有效性，其中我们使用相同的实验设置（例如，贪婪解码），并将我们的实验基于所使用的相同的几个样本。我们相信我们的发现也可以应用于 CoT 提示的一些后续变体。

最近的一些工作集中在理解/分析 CoT 提示。Madaan 和 Yazdanbakhsh（2022）通过将其改变为反事实，研究了所证明的 CoT 理论的不同组成部分的重要性。他们只尝试了有限的方法来改变错误的原理，包括使用不正确的计算（例如，“5 + 4 = 7”）或实体。对于它们的大多数设置，即使基本原理是反事实的，它们仍然是正确的，因为查询相应地改变了（参见，例如，第 48 页）。与我们的工作同时，Ye et al.（2022）还探讨了模型性能如何受到破坏 CoT 理论的影响。他们尝试使用不正确的计算和丢弃（部分）桥接对象/语言模板，这与我们的消融设计不同。 Saparov 和 He（2023）通过创建基于一阶逻辑的合成 QA 数据集来系统地评估 CoT，该数据集允许将生成的原理解析为符号证明以进行形式分析。总的来说，据我们所知，我们是第一个证明 CoT 理论是错误的，并大大偏离黄金理论，同时仍然保持高模型性能的人。

在一般情况下，在上下文学习（ICL），民等。（2022）表明，对于具有分类标签空间（分类和多选）的自然语言理解中的各种任务，真实输入-标签映射对最终任务性能的影响很小，而标签空间，整体格式和文本分布等其他方面是关键。在这项工作的基础上，Yoo 等人（2022）发现，正确的输入-标签对应关系可能会根据任务和实验配置产生不同的影响，Wei 等人（2023）发现，具有较大规模的模型可以覆盖语义先验并在上下文中学习输入-标签映射。Webson 和 Pavlick（2022）发现，对于指令模型，在不相关或误导性指令下，自然语言推理任务的性能略有下降。Xie 等人（2022）通过将其公式化为贝叶斯推理提供了 ICL 的理论分析。我们的工作可以被看作是一种尝试，以经验为基础-

stand ICL in sequence generation tasks requiring multi-step reasoning.

8 Conclusion

In this paper, we aim to better understand Chain-of-Thought prompting through a series of ablation experiments that unveil the impact of different aspects of a CoT rationale. We find that 1) the validity of reasoning in the prompting examples matters only a small portion to the performance; 2) relevance to the input query and following the order along the reasoning steps are the key to the effectiveness of CoT prompting. Overall, our findings deepen the understanding of CoT prompting, and open up new questions/reflections regarding LLMs' capability of learning to reason in context.

Limitations

Experiments on other types of reasoning tasks. In addition to the two representative reasoning tasks (arithmetic reasoning and multi-hop question answering) that we experiment on, there are also other tasks where CoT prompting brings significant improvements over standard prompting shown by previous work, many of which are symbolic reasoning tasks such as Last letter concatenation, Coin flip from Wei et al. (2022) and Temporal Sequences, Tracking Shuffled Objects from BIG-Bench (Srivastava et al., 2022; Suzgun et al., 2022). However, most (if not all) tasks there are highly *template-based* and hence the reasoning steps have little variations, both within each example and across different examples. This makes it difficult for us to conduct our ablation studies on these tasks. Take the example of Last letter concatenation, a task about concatenating the last letters of a given sequence of words (e.g., *"Amy Brown"* $\rightarrow$ *"yn"*). Here, every step in the rationale except the last is in the form *"The last letter of X is Y"* where X is some word in the given sequence and Y is the last letter of X. Hence, the language templates are the same and there is no sense of order among the steps (the order is completely characterized by the given sequence instead), and our ablation settings will not apply well. Extending our ablation designs to these "reduced" cases is one of the items we want to explore in the future.

A more systematic treatment of "invalid reasoning". We manually write rationales with invalid reasoning for the experiments in §4 since automatically synthesizing such rationales turns out to be challenging, mostly due to the informal nature of the tasks we experiment on (relatedly, the original CoT rationales are also human-written). We intend to give a more systematic treatment of the invalid reasoning setting in the future, e.g., following the categorizations of informal logical fallacies (Copi et al., 2016).

Improvements on intrinsic evaluation. Our intrinsic evaluation of the generated rationales is based on the correctness of bridging objects, which, even though is a good indicator of the quality of language templates (Appendix A.2) in our experiments, may not be a good metric in general cases. It also relies on ground truth bridging objects, which are usually not available and costly to annotate. Toward this end, one direction we want to explore further is to develop ways to conduct more comprehensive and reference-free intrinsic evaluations. Recent papers such as Golovneva et al. (2023) have also done promising work along this line.

Acknowledgements

The authors would like to thank the anonymous reviewers and colleagues from the OSU NLP group for their thoughtful comments. This research was supported in part by Google Faculty Award, Google Research Scholar Award, NSF IIS 1815674, NSF CAREER 1942980, NSF OAC-2112606, and Ohio Supercomputer Center (Center, 1987). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ohio Supercomputer Center. 1987. Ohio supercomputer center.

我们的工作可以被视为一种尝试，以经验为基础，了解 ICL 在序列生成任务，需要多步推理。

8 结论

在本文中，我们的目标是更好地理解 Chain-ofThought 提示通过一系列消融实验，揭示了 CoT 理论的不同方面的影响。我们发现：1）提示示例中推理的有效性对性能的影响很小; 2）与输入查询的相关性和遵循沿着推理步骤的顺序是 CoT 提示有效性的关键。总体而言，我们的研究结果加深了对 CoT 提示的理解，并就 LLM 在上下文中学习推理的能力提出了新的问题/思考。

限制

其他类型推理任务的实验。

除了两个有代表性的推理任务外，（算术推理和多跳问题回答），也有其他任务，其中 CoT 提示带来了显着的改进，比以前的工作所示的标准提示，其中许多是符号推理任务，如最后一个字母连接，硬币翻转从魏等人。（2022）和时间序列，从 BIG-Bench 跟踪混洗对象（Srivastava 等人，2022 年; Suzgun 等人，2022 年）。然而，大多数（如果不是全部的话）任务都是高度基于模板的，因此推理步骤在每个示例内和不同示例之间几乎没有变化。这使得我们很难对这些任务进行消融研究。以最后一个字母连接为例，这是一个关于连接给定单词序列的最后一个字母的任务（例如，“艾米·布朗”→“yn”）。 在这里，除了最后一步之外，基本原理中的每一步都是“X 的最后一个字母是 Y”的形式，其中 X 是给定序列中的某个单词，Y 是 X 的最后一个字母。因此，语言模板是相同的，步骤之间没有顺序感（顺序完全由给定的序列表征），我们的消融设置将无法很好地应用。将我们的消融设计扩展到这些“减少”的病例是我们未来想要探索的项目之一。

更系统地处理“无效理由-

ing”。我们手动为§4 中的实验编写具有无效推理的基本原理，因为自动合成这些基本原理被证明是

挑战性，主要是由于我们实验的任务的非正式性质（相关地，最初的 CoT 理论也是人类编写的）。我们打算在将来对无效推理设置进行更系统的处理，例如，按照非正式逻辑谬误的分类（Copi 等人，2016 年）。

内在评价的改进。我们的-

生成的基本原理的本质评估是基于桥接对象的正确性，尽管在我们的实验中桥接对象是语言模板质量的良好指标（附录 A.2），但在一般情况下可能不是一个好的度量标准。它还依赖于地面实况桥接对象，这些对象通常不可用并且注释成本高。为此，我们希望进一步探索的一个方向是开发进行更全面和无参考的内在评估的方法。最近的论文，如 Golovneva 等人（2023）也沿着这条路线做了有希望的工作。

确认

作者要感谢匿名评论者和来自 OSU NLP 小组的同事们的深思熟虑的评论。这项研究得到了 Google Faculty Award、Google Research Scholar Award、NSF IIS 1815674、NSF CAREER 1942980、NSF OAC-2112606 和俄亥俄州超级计算机中心（Center，1987）的部分支持。本文所载的观点和结论是作者的观点和结论，不应被解释为代表美国政府的官方政策，无论是明示还是暗示。美国政府被授权为政府目的复制和分发重印本，尽管此处有任何版权声明。

引用

Ekin Akyürek，Dale Schuurmans，Jacob Andreas，Tengyu Ma，and Denny Zhou. 2023.什么学习算法是上下文学习？调查

线性模型。在第十一届国际学习代表大会上。

Tom Brown，Benjamin Mann，Nick Ryder，Melanie Subbiah，Jared D Kaplan，Prafulla 达里瓦尔，Arvind Neelakantan，Pranav Shyam，Girish Sastry，阿曼达 Askell，et al. 2020.语言模型是少镜头的

学习者神经信息处理系统的进展，33：1877 - 1901。

俄亥俄州超级计算机中心。1987.俄亥俄州超级计算机中心。

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Irving Copi, Carl Cohen, and Victor Rodych. 2016. *Introduction to logic*. Routledge.

Yao Fu, Hao Peng, and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pages 30583–30598. Curran Associates, Inc.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In

Wenhu Chen，Xueguang Ma，Xinyi Wang，and William W Cohen. 2022. 思维提示程序：将计算与推理分离，用于数值推理任务。arXiv 预印本 arXiv：2211.12588。

弗朗索瓦·肖莱 2019.关于智力的测量。

arXiv 预印本 arXiv：1911.01547。

Aakanksha Chowdhery，沙兰 Narang，Jacob Devlin，Maarten Bosma，Gaurav Mishra，Adam Roberts，Paul Barham，Hyung Won Chung，Charles 萨顿，塞巴斯蒂安 Gehrmann，et al. 2022. Palm：使用路径扩展语言建模。2204. 02311.

Hyung Won Chung，Le Hou，Shayne Longpre，Barret Zoph，Yi Tay，William Fedus，Eric Li，Xuezhi Wang，Mostafa Dehghani，Siddhartha Brahma，et al.

2022.扩展指令微调语言模型。
2210.11416.第一次世界大战

Karl Cobbe，Vineet Kosaraju，Mohammad Bavarian，Jacob Hilton，Reiichiro Nakano，Christopher Hesse 和 John Schulman. 2021.培训 Verifier to Solve Math Word Problems。arXiv preprint arXiv：2110.14168.（arXiv preprint）

欧文科皮，卡尔科恩，和维克托罗迪奇。2016.逻辑学导论。我是说，

姚复，郝鹏，图沙尔·霍特。2022. gpt 如何获得它的能力？追溯语言模型的涌现能力到它们的来源。姚复的概念。

Shivam Garg，Dimitris Tsipras，珀西 S Liang 和 Gregory Valiant。2022. transformers 在上下文中能学到什么？简单函数类的案例研究。在
神经信息处理系统的进展，
第 35 卷，第 30583-30598 页。柯兰联营公司

Olga Golovneva、莫亚彭陈、Spencer Poff、Martin 科雷多尔、Luke Zettlemoyer、Maryam FazelZarandi 和 Asli Celikyilmaz。2023. ROSCOE：一套用于逐步推理评分的指标。在
第十一届学习表征国际会议。

黄洁和 Kevin Chen-Chuan Chang。2022.大规模语言模型的推理：一个综述。
arXiv 预印本 arXiv：2212.10403。

Joel Jang，Seonghyeon Ye，and Minjoon Seo. 2023.大型语言模型真的能理解提示吗？
一个有否定提示的案例研究转让方面的
学习自然语言处理研讨会，第 52-62 页。PMLR。

阿曼·马丹和阿米尔·亚兹丹巴赫什。2022.文本和模式：有效的思路链需要两个人来跳探戈。arXiv 预印本 arXiv：2209.07686。

Sewon Min，Xinxi Lyu，Ari Holtzman，Mikel Artetxe，Mike 刘易斯，Hannaneh Hajishirzi 和 Luke Zettle-
莫耶 2022.重新思考示威的作用：
是什么让情境学习发挥作用？在进行中-
2022 年自然语言处理经验方法会议，第 11048-11064 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Long Ouyang，Jeff Wu，Xu Jiang，Diogo Almeida，卡罗尔 L 温赖特，Pamela Mishkin，Chong Zhang，Sandhini Agarwal，Katarina Slama，Alex Ray，et al.
2022. 训练语言模型遵循指令并提供人类反馈。arXiv 预印本 arXiv：2203.02155。

Ofir Press，Muru Zhang，Sewon Min，Ludwig 施密特，Noah A Smith 和 Mike 刘易斯。2022.测量和缩小语言模型中的组合性差距。2210.03350.我的天啊!

Yasaman Razeghi，Robert L Logan IV，Matt Gardner 和 Sameer Singh。2022.预训练词频对少数数字推理的影响。在
计算线协会的调查结果-
guistics：EMNLP 2022，第 840-854 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

阿布海尔·萨帕罗夫和何荷。2023.语言模型是贪婪的推理机：一个系统的形式分析
一系列的思考在第十一届国际学习代表大会上。

Aarohi Srivastava，Abhinav Rastogi，Abhishek Rao，Abu Awal Md Shoeb，Abubakar Abid，Adam 菲施，Adam R Brown，Adam Santoro，Aditya Gupta，Adrià Garriga-Alonso，et al. 2022. 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。

Mirac Suzgun，Nathan Scales，Nathanael Schärli，塞巴斯蒂安 Gehrmann，Yi Tay，Hyung Won Chung，Aakanksha Chowdhery，Quoc V Le，艾德 H Chi，Denny Zhou，et al. 2022.思考大型工作台任务以及思维链是否可以解决它们。arXiv 预印本 arXiv：2210.09261。

王学智，魏杰，Dale Schuurmans，Quoc V Le，艾德 H. Chi，沙兰纳朗，Aakanksha Chowdhery，和 Denny Zhou. 2023.自我一致性改进了语言模型中的思维推理链。在
第十一届学习表征国际会议。

艾伯特·韦森和艾莉·帕夫利克 2022.基于 XML 的模型是否真正理解其
提示？在计算语言学协会北美分会 2022 年会议论文集：人类语言技术，第 2300-2344 页，西雅图，美国。

计算语言学协会。

Jason Wei，Xuechi Wang，Dale Schuurmans，Maarten Bosma，brian ichter，Fei Xia，艾德池，Quoc V Le，和 Denny Zhou. 2022.在大型语言模型中，思维链的提示引发推理。在

Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

神经信息处理系统进展，
Volume 35，Pages 24824-24837. Curran Associates，Inc.

In Jerry Wei，Jason Wei，Yi Tay，Dustin Tran，Albert Webson，Yifeng Lu，Xinyun Chen，Hanxiao Liu，Da Huang，Denny Zhou，et al. 2023.更大的语言模型以不同的方式进行上下文学习。arXiv 预印本 arXiv：2303.03846。

Sang Michael Xie，Aditi Raghunathan，珀西梁和 Tengyu Ma。2022.情境学习的内隐 baidu 推理解释。在国际学习代表会议。

凯伊、史里瓦山伊耶尔、西利奇尔、维斯斯托雅诺夫、格雷格·杜兰特、拉玛坎特帕索鲁 2022.

有效的情境学习的补充解释。arXiv 预印本 arXiv：2211.13892。

姜敏宥俊金贤珠慧妍秀赵尚宇孪尚宇孪尚久俊 2022. Ground-truth labels matter：A deeper look into input-label demonstrations. In Pro-

2022 年自然语言处理经验方法会议，第 2422- 2437 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

张卓生、阿斯顿、李穆和亚历克斯·斯莫拉。2023.大型语言模型中的自动思维链提示。在第十一届国际学习代表大会上。

A Appendix

A.1 Chain of Thought Exemplars

We base our experiments on the original prompt exemplars released by Wei et al. (2022); Press et al. (2022) with slight editing to make the structure more consistent and reduce redundancy, which makes our ablations more convenient to conduct. The edited CoT prompts for arithmetic reasoning and multi-hop QA could be found in Table 9 and Table 10 respectively. We mainly perform the following edits: 1) shift premise steps (copy/paraphrase of facts from the query) to the beginning steps of the rationale; 2) add/expand the language templates for steps with no/over-concise language templates; 3) remove unnecessary steps/information that are unhelpful for answering the query.

Overall, these edits only slightly affect the performance of CoT. A comparison of the performance is shown in Table 5.

A.2 More Details on Intrinsic Evaluation

We use Recall/F1 of the bridging objects as the metrics for intrinsic evaluation of the generated rationales. While the metrics don't take into account the quality of the language templates, we examine the predicted rationales for 20 random examples under each setting we tested except standard prompting (which does not generate any rationale), and find that for all the examples, whenever the LLM reaches a correct bridging object, the corresponding language template within the step is also correct. This suggests that overall, the correctness of bridging objects is a very good indicator of the quality of the reasoning steps.

A.3 Additional Results & Discussion

Table 6 includes results for `text-davinci-003`, `text-davinci-002`'s very recent improved version.

Comparing with the results from `text-davinci-002` (Table 2), it could be seen that `text-davinci-003` brings large performance improvements, especially under the ablation settings. In particular, providing invalid reasoning for the rationales (①) overall only marginally harms the performance, and even outperforms CoT on GSM8K under intrinsic evaluation. This suggests that `text-davinci-003` is equipped with even stronger multi-step "reasoning" abilities on the evaluated tasks through pre-training, and learns little about how to reason from the demonstrations.

For the remaining settings where we ablate the relevance/coherence (②-⑦), the same trend can be observed on the challenging GSM8K dataset, e.g., the model still suffers a lot when providing rationales that are irrelevant or have incoherent language templates. For the relatively easier Bamboogle dataset, the high model capacity indicated by its impressive performance has basically erased significant impacts from the ablations, with the only standing observation that the model still needs the rationales to be relevant to maintain its performance.

Overall, from the performance achieved by `text-davinci-002` and `text-davinci-003`, we can observe a general trend where LLMs suffer less from the ablations when they have more prior knowledge about the task. To further explore this, we test on Flan-PaLM (Chung et al., 2022), the instruction-tuned version of PaLM (Chowdhery et al., 2022) that is directly trained on both arithmetic reasoning and factual QA in CoT fashion during instruction tuning, and hence has immense knowledge on these tasks. The results are shown in Table 7. It could be seen that none of the ablations has significant impacts on the model performance, which further strengthens this pattern. On the positive side, this indicates that LLMs can effectively utilize their prior knowledge to solve new problems; however, this also leads to the concern that LLMs may over-rely on their prior knowledge and ignore important information in the context, including those that are crucial for specifying the task semantics (Jang et al., 2023).

We also test PaLM, which is a non-instruction-finetuned LLM that exhibits strong CoT reasoning ability. The results are included in Table 8. Overall, similar observations could be found, which suggests that our findings are not exclusive to instruction-tuned models. There are some inconsistencies between the performance from PaLM and InstructGPT on Bamboogle, where the importance of coherence and relevance for bridging objects is flipped. This could be the consequence of instruction tuning, and differences in pretraining corpora and model scales.

A.4 Full List of Prompts

Full prompts for all settings in our experiments are included in Table 9-24.

A 附录

A.1 思维链范例

我们的实验基于 Wei et al.（2022）发布的原始提示样本; Press et al.（2022）进行了轻微编辑，以使结构更加一致并减少冗余，这使得我们的消融更便于进行。针对算术推理和多跳 QA 的编辑后 CoT 提示分别见表 9 和表 10。我们主要进行以下编辑：1）将前提步骤（从查询中复制/解释事实）转移到基本原理的开始步骤; 2）为没有/过于简洁的语言模板的步骤添加/扩展语言模板; 3）删除对回答查询毫无帮助的不必要步骤/信息。

总的来说，这些编辑对 CoT 的性能只有轻微的影响。性能比较如表 5 所示。

A.2 关于内在评估的更多细节

我们使用的召回/F1 的桥接对象作为度量生成的理由的内在评价。虽然这些指标没有考虑到语言模板的质量，但我们在测试的每个设置下检查了 20 个随机示例的预测基本原理，除了标准提示（不生成任何基本原理），并发现对于所有示例，只要 LLM 到达正确的桥接对象，步骤中相应的语言模板也是正确的。这表明，总体而言，桥接对象的正确性是推理步骤质量的一个非常好的指标。

A.3 其他结果和讨论

表 6 包括 text-davinci-003 的结果，text-davinci-002 的最新改进版本。

与实验结果比较，

text-davinci-002（表 2），可以看出 text-davinci-003 带来了很大的性能改进，尤其是在消融设置下。特别是，为基本原理（1）提供无效的推理总体上只会略微损害性能，甚至在内在评估下在 GSM 8 K 上优于 CoT。这表明 text-davinci-003 通过预训练在被评估的任务上具有更强的多步“推理”能力，并且从演示中几乎没有学到如何推理。

对于我们消除相关性/一致性的其余设置（2 - 7），可以在具有挑战性的 GSM 8 K 数据集上观察到相同的趋势，例如，当提供不相关或具有不连贯的语言模板的基本原理时，该模型仍然受到很大影响。对于相对简单的 Bambobaly 数据集，其令人印象深刻的性能所表明的高模型容量基本上消除了消融的显著影响，唯一的长期观察结果是模型仍然需要相关的基本原理来保持其性能。

总的来说，从

text-davinci-002 和 text-davinci-003 中，我们可以观察到一个总体趋势，即当 LLM 对任务具有更多先验知识时，LLM 遭受消融的程度更低。为了进一步探索这一点，我们在 Flan-PaLM 上进行测试（Chung 等人，2022）、PaLM 的抑制调谐版本（Chowdhery 等人，2022），其在指令调整期间以 CoT 方式直接接受算术推理和事实 QA 的训练，因此在这些任务上具有丰富的知识。其结果如表 7 所示。可以看出，消融均未对模型性能产生显著影响，这进一步强化了该模式。从积极的方面来看，这表明 LLM 可以有效地利用他们的先验知识来解决新问题;然而，这也导致了 LLM 可能过度依赖他们的先验知识并忽略上下文中的重要信息的担忧，包括那些对于指定任务语义至关重要的信息（Jang et al.，2023 年）。

我们还测试了 PaLM，这是一个非线性微调的 LLM，表现出很强的 CoT 推理能力。结果见表 8。总的来说，可以发现类似的观察结果，这表明我们的研究结果并不局限于预防调整模型。PaLM 和 InstructGPT 在 Bambobile 上的性能之间存在一些不一致之处，其中连贯性和相关性对于桥接对象的重要性被翻转。这可能是指令调整的结果，以及预训练语料库和模型量表的差异。

A.4 完整列表

我们实验中所有设置的完整提示见表 9-24。

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
Chain-of-Thought (Original)	44.5	48.7	48.1	44.8	43.1
Chain-of-Thought (After Editing)	43.9	48.3	48.5	45.2	45.2

Table 5: Performance comparison (under `text-davinci-002`) of the Chain-of-Thought exemplars before/after our editing.

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.2	N/A	25.1
CoT (Chain-of-Thought prompting)	48.4	53.1	54.5	61.6	59.5
① Invalid Reasoning	50.2	53.5	51.5	60.8	56.4
② No *coherence* for bridging objects	46.5	51.5	50.4	59.2	55.2
③ No relevance for bridging objects	32.5	38.3	47.2	60.4	56.9
④ No *coherence* for language templates	37.8	43.3	41.9	57.2	51.4
⑤ No relevance for language templates	44.6	49.9	51.8	62.4	59.3
⑥ No *coherence*	34.5	39.4	31.0	57.6	55.2
⑦ No relevance	15.5	17.8	16.2	50.0	49.0

Table 6: Intrinsic and extrinsic evaluation results under `text-davinci-003` for all settings. Discussions are included in Appendix A.3.

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	21.8	N/A	36.5
CoT (Chain-of-Thought prompting)	72.2	73.0	63.8	57.6	56.9
① Invalid Reasoning	71.8	72.6	64.4	55.6	52.8
② No *coherence* for bridging objects	72.1	72.9	65.8	51.6	49.3
③ No relevance for bridging objects	71.1	71.9	64.6	54.0	52.8
④ No *coherence* for language templates	71.6	72.2	63.9	54.0	52.0
⑤ No relevance for language templates	71.9	72.7	64.9	55.2	53.5
⑥ No *coherence*	71.7	72.5	64.2	54.4	54.0
⑦ No relevance	70.7	71.6	64.5	50.0	51.9

Table 7: Intrinsic and extrinsic evaluation results under Flan-PaLM (Chung et al., 2022), the instruction-tuned version of PaLM for all settings. Discussions are included in Appendix A.3.

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.0	N/A	31.0
CoT (Chain-of-Thought prompting)	36.6	40.6	37.0	54.0	54.8
① Invalid Reasoning	33.9	36.9	31.8	50.4	46.1
② No *coherence* for bridging objects	30.3	35.0	33.5	33.6	25.7
③ No relevance for bridging objects	15.5	20.1	21.2	47.2	47.7
④ No *coherence* for language templates	23.1	27.3	21.9	40.4	35.5
⑤ No relevance for language templates	19.5	22.9	20.4	38.4	30.6
⑥ No *coherence*	23.9	28.3	24.1	39.6	33.6
⑦ No relevance	12.1	16.4	16.4	28.4	14.3

Table 8: Intrinsic and extrinsic evaluation results under PaLM. Discussions are included in Appendix A.3.

	GSM 8K 标准			班博什	
	中间截留召回国际米兰。F1 应答接入接口回忆答案 F1				
Chain of Thought（原创）	44.5	48.7	48.1	44.8	43.1
思想链（编辑后）	43.9	48.3	48.5	45.2	45.2

表 5：我们编辑之前/之后的思想链范例的性能比较（根据 text-davinci-002）。

		GSM 8K 标准			班博什	
		中间截留召回国际米兰。F1 应答接入接口回忆答案 F1				
STD（标准提示）N/A N/A 15.2 N/A 25.1 CoT（思维链提示）48.4 53.1 54.5 61.6 59.5						
1	无效推理	50.2	53.5	51.5	60.8	56.4
2	桥接对象无连贯性	46.5	51.5	50.4	59.2	55.2
3	与桥接对象无关	32.5	38.3	47.2	60.4	56.9
4	语言模板没有连贯性	37.8	43.3	41.9	57.2	51.4
5	与语言模板无关	44.6	49.9	51.8	62.4	59.3
6	无连贯性	34.5	39.4	31.0	57.6	55.2
7	无关	15.5	17.8	16.2	50.0	49.0

表 6：所有设置下 text-davinci-003 的内在和外在评价结果。讨论情况见附录 A.3。

		GSM 8K 标准			班博什	
		中间截留召回国际米兰。F1 应答接入接口回忆答案 F1				
STD（标准提示）N/A N/A 21.8 N/A 36.5 CoT（思维链提示）72.2 73.0 63.8 57.6 56.9						
1	无效推理	71.8	72.6	64.4	55.6	52.8
2	桥接对象无连贯性	72.1	72.9	65.8	51.6	49.3
3	与桥接对象无关	71.1	71.9	64.6	54.0	52.8
4	语言模板没有连贯性	71.6	72.2	63.9	54.0	52.0
5	与语言模板无关	71.9	72.7	64.9	55.2	53.5
6	无连贯性	71.7	72.5	64.2	54.4	54.0
7	无关	70.7	71.6	64.5	50.0	51.9

表 7：Flan-PaLM 下的内在和外在评估结果（Chung 等人，2022），用于所有设置的 PaLM 的预防调整版本。讨论情况见附录 A.3。

		GSM 8K 标准			班博什	
		中间截留召回国际米兰。F1 应答接入接口回忆答案 F1				
STD（标准提示）N/A N/A 15.0 N/A 31.0 CoT（思维链提示）36.6 40.6 37.0 54.0 54.8						
1	无效推理	33.9	36.9	31.8	50.4	46.1
2	桥接对象无连贯性	30.3	35.0	33.5	33.6	25.7
3	与桥接对象无关	15.5	20.1	21.2	47.2	47.7
4	语言模板没有连贯性	23.1	27.3	21.9	40.4	35.5
5	与语言模板无关	19.5	22.9	20.4	38.4	30.6
6	无连贯性	23.9	28.3	24.1	39.6	33.6
7	无关	12.1	16.4	16.4	28.4	14.3

表 8：PaLM 下的内在和外在评价结果。讨论情况见附录 A.3。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 15 trees originally. Then there were 21 trees after the Grove workers planted some more. So there must have been 21 - 15 = 6 trees that were planted. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 3 cars. Then 2 more cars arrive. Now 3 + 2 = 5 cars are in the parking lot. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 lollipops originally. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8 lollipops. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 5 toys. He then got 2 toys each from his mom and dad. So he got 2 * 2 = 4 more toys. Now he has 5 + 4 = 9 toys. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 9 computers. For each day from monday to thursday, 5 more computers were installed. So 4 * 5 = 20 computers were added. Now 9 + 20 = 29 computers are now in the server room. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 58 golf balls. He lost 23 on Tuesday, and lost 2 more on wednesday. So he had 58 - 23 = 35 at the end of Tuesday, and 35 - 2 = 33 at the end of wednesday. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars. She bought 5 bagels for 3 dollars each. So she spent 5 * 3 = 15 dollars. Now she has 23 - 15 = 8 dollars left. The answer is 8.

Table 9: Full prompt for Chain-of-Thought prompting in our experiments (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was founded by Gianni Versace. Gianni Versace was shot and killed on July 15, 1997. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The director of the film FAQ: Frequently Asked Questions is Carlos Atanes. The director of the film The Big Money is John Paddy Carstairs. The nationality of Carlos Atanes is Spanish. The nationality of John Paddy Carstairs is British. Spanish is not equal to British. So the final answer (whether they have the same nationality) is: No.

Table 10: Full prompt for Chain-of-Thought prompting in our experiments (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？A：原来有 15 棵树。然后有 21 棵树后，格罗夫工人种植了一些。因此，必须种植 21 - 15 = 6 棵树。答案是 6。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？A：原来有 3 辆汽车。又来了两辆汽车。现在停车场有 3 + 2 = 5 辆汽车。答案是 5。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？A：原来，莉亚有 32 块巧克力，她姐姐有 42 块。因此，他们总共有 32 + 42 = 74。吃了 35 块，总共还剩 74 - 35 = 39 块。答案是 39。问：杰森有 20 个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？杰森本来有 20 个棒棒糖。然后他给了丹尼一些后有 12 个。所以他给了丹尼 20 - 12 = 8 个棒棒糖。答案是 8。问：肖恩有五个玩具。圣诞节时，他从爸爸妈妈那里各得到了两个玩具。他现在有多少玩具？答：肖恩从 5 个玩具开始。他从爸爸妈妈那里得到了两个玩具。所以他得到了 2 * 2 = 4 个玩具。

现在他有 5 + 4 = 9 个玩具。答案是 9。

问：服务器机房里有九台电脑。从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？

答：最初有 9 台电脑。从星期一到星期四，每天都多安装 5 台电脑。

所以 4 * 5 = 20 台计算机被添加。现在，9 + 20 = 29 台计算机现在在服务器室。答案是 29。问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。星期三，他又输了两个。星期三结束时他有几个高尔夫球？

A：迈克尔一开始有 58 个高尔夫球。他在周二失去了 23 分，周三又失去了 2 分。所以他在周二结束时有 58 - 23 = 35，在周三结束时有 35 - 2 = 33。答案是 33。

问：奥利维亚有 23 美元。她买了五个百吉饼，每个三美元。她还剩多少钱？A：奥利维亚有 23 美元。她买了 5 个百吉饼，每个 3 美元。所以她花了 5 * 3 = 15 美元。现在她还剩 23 - 15 = 8 美元。答案是 8。表 9：在我们的实验（算术推理）中对思想链提示的完整提示。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

答：西奥多·海克尔去世时 65 岁。哈里·沃恩·沃特金斯去世时 69 岁。所以最后的答案（这个人的名字）是：哈里沃恩沃特金斯。

问题：为什么 Versus 的创始人死了？

Versus 由 Gianni Versace 创立。Gianni Versace 于 1997 年 7 月 15 日被枪杀。所以最终的答案（死亡原因）是：枪杀。

问：谁是丹巴尔·沙阿的孙子？

答：丹巴尔沙阿（？- 1645 年，他是 Krishna Shah 的父亲。Rudra Shah 是 Krishna Shah 的孩子。

- 1661）。所以最后的答案（孙子的名字）是：Rudra Shah。问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？

答：电影的导演常见问题：常见问题是卡洛斯阿坦尼斯。电影《大款》的导演是约翰·帕迪卡斯泰尔斯。卡洛斯阿坦尼斯的国籍是西班牙。约翰·帕迪卡斯泰尔斯的国籍是英国。西班牙语不等于英国语。因此，最后的答案（他们是否有相同的国籍）是：没有。

表 10：在我们的实验中对思想链提示的完整提示（事实 QA）。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 15 trees originally. Then there were 21 trees after the Grove workers planted some more. Now 15 + 21 = 36. Since there were 6 workers in the grove, so the grove workers planted 36 / 6 = 6 trees today. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 3 cars. Then 2 more cars arrive. Now 3 * 2 = 6 cars come. So 6 - 1 = 5 cars are in the parking lot. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 lollipops originally. Then he had 12 after giving some to Denny. Now 20 + 12 = 32. Jason has 4 times what Denny has, so he gave Denny 32 / 4 = 8 lollipops. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 5 toys. He then got 2 toys each from his mom and dad. Now 5 - 2 = 3. So he has 3 * 3 = 9 toys now for Christmas. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 9 computers. For each day from monday to thursday, 5 more computers were installed. Now 9 * 5 = 45 computers. Since 4 * 4 = 16, now 45 - 16 = 29 computers are now in the server room. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 58 golf balls. He lost 23 on Tuesday, and lost 2 more on wednesday. So compared with wednesday, he lost 23 - 2 = 21 more balls on Tuesday. So he had 58 - 21 = 37 golf balls at the end of wednesday. The answer is 37.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars. She bought 5 bagels for 3 dollars each. So she earned 23 - 5 = 18 dollars. Now 18 / 3 = 6. So she has 6 + 2 = 8 dollars left. The answer is 8.

Table 11: Full prompt for "invalid reasoning" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker wrote an essay, Kierkegaard and the Philosophy of Inwardness in 1913. Harry Vaughan Watkins played his final Wales international against England in January 1906. So the final answer (the name of the person) is: Theodor Haecker.

Question: Why did the founder of Versus die?
Answer: Versus was a diffusion line of the Italian luxury fashion house Versace, which began in 2009. 2009 is the year American singer Michael Jackson died of acute propofol and benzodiazepine intoxication. So the final answer (reason of death) is: Intoxication.

Question: Who is the grandchild of Dambar Shah?
Answer: Dambar Shah (? - 1645) was the king of the Gorkha Kingdom. The Gorkha Kingdom was established by Prince Dravya Shah. Dravya Shah has a child named Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: FAQ: Frequently Asked Questions is a feature-length dystopian movie. The Big Money is a 1958 comedy film. Dystopian stories mostly take place in British. Comedy stories mostly happen in Australia. British is not equal to Australia. So the final answer (whether they have the same nationality) is: No.

Table 12: Full prompt for "invalid reasoning" setting (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？A：原来有 15 棵树。然后有 21 棵树后，格罗夫工人种植了一些。15 + 21 = 36 由于格罗夫有 6 名工人，所以格罗夫的工人今天种植了 36 / 6 = 6 棵树。答案是 6。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？A：原来有 3 辆汽车。又来了两辆汽车。现在 3 * 2 = 6 辆汽车来了。所以 6 - 1 = 5 辆汽车在停车场。答案是 5。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？A：原来，莉亚有 32 块巧克力，她姐姐有 42 块。所以她姐姐比利亚多吃了 42 - 32 = 10 块巧克力。吃了 35 块，10 + 35 = 45，一共剩下 45 - 6 = 39 块。答案是 39。问：杰森有 20 个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？杰森本来有 20 个棒棒糖。然后他给了丹尼一些后有 12 个。20 + 12 = 32。杰森的棒棒糖是丹尼的 4 倍，所以他给了丹尼 32 / 4 = 8 个棒棒糖。答案是 8。问：肖恩有五个玩具。圣诞节，他从爸爸妈妈那里得到了两个玩具。他现在有多少玩具？答：肖恩从 5 个玩具开始。他从爸爸妈妈那里得到了两个玩具。现在 5 - 2 = 3 所以他现在有 3 * 3 = 9 个玩具作为圣诞礼物。答案是 9。问：服务器机房里有九台电脑。 从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？

答：最初有 9 台电脑。从星期一到星期四，每天都多安装 5 台电脑。9 * 5 = 45 台计算机。由于 4 * 4 = 16，现在服务器机房中有 45 - 16 = 29 台计算机。答案是 29。问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。周三，他又输了两场。星期三结束时他有几个高尔夫球？A：迈克尔一开始有 58 个高尔夫球。他在周二失去了 23 分，周三又失去了 2 分。因此，与周三相比，他在周二多丢了 23 - 2 = 21 个球。所以他在周三结束时有 58 - 21 = 37 个高尔夫球。

答案是 37。

问：奥利维亚有 23 美元。她买了 5 个百吉饼，每个 3 美元。她还剩多少钱？A：奥利维亚有 23 美元。她买了 5 个百吉饼，每个 3 美元。所以她赚了 23 - 5 = 18 美元。18/3 = 6 所以她还剩 6 + 2 = 8 美元。答案是 8。表 11："无效推理"设置的全提示（算术推理）。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

答：西奥多·海克尔（Theodor Haecker）在 1913 年写了一篇文章《克尔凯郭尔与内在哲学》（Kierkegaard and the Philosophy of Inwardness）。哈里·沃恩·沃特金斯在 1906 年 1 月对阵英格兰的比赛中参加了他的最后一场威尔士国际比赛。所以最后的答案（这个人的名字）是：西奥多·海克尔。

问题：为什么 Versus 的创始人死了？

回答：Versus 是意大利奢侈时尚品牌 Versace 的一个扩散线，始于 2009 年。2009 年是美国歌手迈克尔杰克逊死于急性异丙酚和苯二氮卓类药物中毒的一年。所以最终的答案（死亡原因）是：中毒。

问：谁是丹巴尔·沙阿的孙子？

答：丹巴尔沙阿（？- 1645 年，他是廓尔喀王国的国王。廓尔喀王国由 Dravya Shah 王子建立。Dravya Shah 有一个名叫 Rudra Shah 的孩子。所以最后的答案（孙子的名字）是：Rudra Shah。

问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？

回答：常见问题：常见问题是一部长篇反乌托邦电影。《大富翁》（The Big Money）是一部 1958 年的喜剧电影。反乌托邦的故事大多发生在英国。喜剧故事大多发生在澳大利亚。英国不等于澳大利亚。因此，最后的答案（他们是否有相同的国籍）是：没有。

表 12："无效推理"设置的完整提示（事实 QA）。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 21 - 15 = 6 trees originally. Then there were 15 trees after the Grove workers planted some more. So there must have been 21 trees that were planted. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 3 + 2 = 5 cars. Then 3 more cars arrive. Now 2 cars are in the parking lot. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 32 + 42 = 74 chocolates and her sister had 32. So in total they had 74 - 35 = 39. After eating 35, they had 42 pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 - 12 = 8 lollipops originally. Then he had 20 after giving some to Denny. So he gave Denny 12 lollipops. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 4 toys. He then got 5 + 4 = 9 toys each from his mom and dad. So he got 5 more toys. Now he has 2 * 2 = 4 toys. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 5 computers. For each day from monday to thursday, 4 * 5 = 20 more computers were installed. So 9 + 20 = 29 computers were added. Now 9 computers are now in the server room. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 2 golf balls. He lost 23 on Tuesday, and lost 35 - 2 = 33 more on wednesday. So he had 58 at the end of Tuesday, and 58 - 23 = 35 at the end of wednesday. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 5 * 3 = 15 dollars. She bought 5 bagels for 23 - 15 = 8 dollars each. So she spent 3 dollars. Now she has 23 dollars left. The answer is 8.

Table 13: Full prompt for "no coherence for bridging objects" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: 65 was Harry Vaughan Watkins years old when he died. 65 was 69 years old when he died. Theodor Haecker is bigger than 69. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was shot and founded. Gianni Versace was killed on July 15, 1997 by Gianni Versace. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Krishna Shah was the father of Rudra Shah. Dambar Shah (? - 1645) was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The director of John Paddy Carstairs is John Paddy Carstairs. The director of British is Spanish. The nationality of Carlos Atanes is British. The nationality of John Paddy Carstairs is film FAQ: Frequently Asked Questions. Carlos Atanes is not equal to film The Big Money. So the final answer (whether they have the same nationality) is: No.

Table 14: Full prompt for "no coherence for bridging objects" setting (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？

答：原来有 21 - 15 = 6 棵树。然后有 15 棵树后，格罗夫工人种植了一些。

所以一定有 21 棵树被种下了。答案是 6。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？答：原来是 3 + 2 = 5 辆汽车。又来了三辆汽车。现在停车场有两辆汽车。答案是 5。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？答：最初，莉亚有 32 + 42 = 74 块巧克力，她妹妹有 32 块。所以他们总共有 74 - 35 = 39。吃了 35 块，一共还剩下 42 块。答案是 39。问：杰森有 20 个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？A：杰森本来有 20 - 12 = 8 个棒棒糖。然后他给了丹尼一些后有 20 个。所以他给了丹尼 12 根棒棒糖。答案是 8。问：肖恩有五个玩具。圣诞节，他从爸爸妈妈那里得到了两个玩具。他现在有多少玩具？A：Shawn 从 4 个玩具开始。然后他从他的妈妈和爸爸那里得到了 5 + 4 = 9 个玩具。他又买了五个玩具。

现在他有 2 * 2 = 4 个玩具。答案是 9。

问：服务器机房里有九台电脑。从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？答：最初有 5 台电脑。从星期一到星期四，每天安装 4 * 5 = 20 台计算机。9 + 20 = 29 台计算机。现在，9 台计算机在服务器室。答案是 29。问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。周三，他又输了两场。星期三结束时他有几个高尔夫球？

A：迈克尔从两个高尔夫球开始。他在周二输了 23 场，周三又输了 35 - 2 = 33 场。所以周二结束时他有 58，周三结束时 58 - 23 = 35。答案是 33。

问：奥利维亚有 23 美元。她买了五个百吉饼，每个三美元。她还剩多少钱？A：奥利维亚有 5 * 3 = 15 美元。她买了 5 个百吉饼，每个 23 - 15 = 8 美元。她花了 3 美元。现在她只剩下 23 美元了。答案是 8。

表 13：“桥接对象无一致性”设置的完整提示（算术推理）。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

回答：哈里·沃恩·沃特金斯去世时 65 岁。65 岁时，他去世了。西奥多·海克尔比 69 岁大。所以最后的答案（这个人的名字）是：哈里·沃恩·沃特金斯。

问题：为什么 Versus 的创始人死了？

答：对射成立。Gianni Versace 于 1997 年 7 月 15 日被 Gianni Versace 杀害。所以最终的答案（死亡原因）是：枪杀。

问：谁是丹巴尔·沙阿的孙子？

答：克里希纳·沙阿是鲁德拉·沙阿的父亲。丹巴尔·沙阿（？ - 1645）是 Krishna Shah（？ - 1661）。所以最后的答案（孙子的名字）是：Rudra Shah。问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？

答：约翰帕迪卡斯泰尔斯的导演是约翰帕迪卡斯泰尔斯。英国导演是西班牙人。卡洛斯阿坦尼斯的国籍是英国。John Paddy 卡斯泰尔斯的国籍是电影常见问题：常见问题。卡洛斯阿坦尼斯是不等于电影的大钱。因此，最后的答案（他们是否有相同的国籍）是：没有。

表 14：“桥接对象无一致性”设置的完整提示（事实 QA）。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 4 trees originally. Then there were 8 trees after the Grove workers planted some more. So there must have been 8 - 4 = 4 trees that were planted. The answer is 4.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 18 cars. Then 9 more cars arrive. Now 18 + 9 = 27 cars are in the parking lot. The answer is 27.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 19 chocolates and her sister had 31. So in total they had 19 + 31 = 50. After eating 29, they had 50 - 29 = 21 pieces left in total. The answer is 21.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 37 lollipops originally. Then he had 14 after giving some to Denny. So he gave Denny 37 - 14 = 23 lollipops. The answer is 23.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 8 toys. He then got 6 toys each from his mom and dad. So he got 6 * 2 = 12 more toys. Now he has 8 + 12 = 20 toys. The answer is 20.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 23 computers. For each day from monday to thursday, 10 more computers were installed. So 4 * 10 = 40 computers were added. Now 23 + 40 = 63 computers are now in the server room. The answer is 63.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 46 golf balls. He lost 27 on Tuesday, and lost 6 more on wednesday. So he had 46 - 27 = 19 at the end of Tuesday, and 19 - 6 = 13 at the end of wednesday. The answer is 13.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 48 dollars. She bought 7 bagels for 6 dollars each. So she spent 7 * 6 = 42 dollars. Now she has 48 - 42 = 6 dollars left. The answer is 6.

Table 15: Full prompt for “no relevance for bridging objects” setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Albin Barack was 49 years old when he died. Carl Clemens was 55 years old when he died. 55 is bigger than 49. So the final answer (the name of the person) is: Carl Clemens.

Question: Why did the founder of Versus die?
Answer: The gang was founded by John Vitti. John Vitti drowned and got killed on February 2009. So the final answer (reason of death) is: drowning.

Question: Who is the grandchild of Dambar Shah?
Answer: Metis Amando was the father of David Amando. Randall Amando was the child of David Amando. So the final answer (the name of the grandchild) is: Randall Amando.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The director of "The Forgortten Bride" is Paul Cuevas. The director of "Grace and the Rose" is Ronnie Dixon. The nationality of Paul Cuevas is Australia. The nationality of Ronnie Dixon is France. Australia is not equal to France. So the final answer (whether they have the same nationality) is: No.

Table 16: Full prompt for “no relevance for bridging objects” setting (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？A：原来有四棵树。然后有 8 棵树后，格罗夫工人种植了一些。因此，必须种植 8 - 4 = 4 棵树。答案是 4。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？A：原来有 18 辆汽车。然后又来了 9 辆汽车。现在停车场里有 18 + 9 = 27 辆汽车。答案是 27。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？答：最初，莉亚有 19 块巧克力，她妹妹有 31 块。所以他们总共有 19 + 31 = 50。吃完 29 块后，总共还剩 50 - 29 = 21 块。答案是 21。问：杰森有 20 根棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？杰森本来有 37 个棒棒糖。然后他给了丹尼一些后有 14 个。所以他给了丹尼 37 - 14 = 23 个棒棒糖。答案是 23。问：肖恩有五个玩具。圣诞节时，他从爸爸妈妈那里各得到了两个玩具。他现在有多少玩具？A：Shawn 从 8 个玩具开始。他从爸爸妈妈那里得到了 6 个玩具。所以他得到了 6 * 2 = 12 多个玩具。

现在他有 8 + 12 = 20 玩具。答案是 20。

问：服务器机房里有九台电脑。从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？答：最初有 23 台计算机。从星期一到星期四，每天都要多安装 10 台电脑。4 * 10 = 40 台计算机。现在服务器机房中有 23 + 40 = 63 台计算机。答案是 63。

问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。星期三，他又输了两个。星期三结束时他有几个高尔夫球？A：迈克尔开始时有 46 个高尔夫球。他在周二失去了 27 分，周三又失去了 6 分。所以他在周二结束时有 46 - 27 = 19，在周三结束时有 19 - 6 = 13。答案是 13。

问：奥利维亚有 23 美元。她买了五个百吉饼，每个三美元。她还剩多少钱？A：奥利维亚有 48 美元。她买了 7 个百吉饼，每个 6 美元。所以她花了 7 * 6 = 42 美元。现在她还剩 48 - 42 = 6 美元。答案是 6。表 15：“与桥接对象无关”设置的完整提示（算术推理）。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

回答：阿尔宾巴拉克去世时 49 岁。卡尔·克莱门斯去世时 55 岁。55 比 49 大。所以最后的答案（这个人的名字）是：卡尔·克莱门斯。

问题：为什么 Versus 的创始人死了？

答：这个团伙是由约翰·维蒂创立的。约翰·维蒂于 2009 年 2 月溺水身亡。所以最终的答案（死亡原因）是：溺水。

问：谁是丹巴尔·沙阿的孙子？

回答：梅蒂斯·阿曼多是大卫·阿曼多的父亲。兰德尔·阿曼多是大卫·阿曼多的孩子。所以最后的答案（孙子的名字）是：兰德尔·阿曼多。

问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？

答：《被遗忘的新娘》的导演是保罗·奎瓦斯。《格蕾丝与玫瑰》的导演是罗尼狄克逊。Paul Cuevas 的国籍是澳大利亚。罗尼狄克逊的国籍是法国。澳大利亚不等于法国。因此，最后的答案（他们是否有相同的国籍）是：没有。

表 16：“与桥接对象无关”设置的完整提示（事实 QA）。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Then there were 15 trees after the Grove workers planted some more. So there must have been 21 trees that were planted. There are 21 - 15 = 6 trees originally. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Then 3 more cars arrive. Now 2 cars are in the parking lot. There are originally 3 + 2 = 5 cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: After eating 32, they had 42 pieces left in total. Originally, Leah had 32 + 42 = 74 chocolates and her sister had 35. So in total they had 74 - 35 = 39. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Then he had 20 after giving some to Denny. So he gave Denny 12 lollipops. Jason had 20 - 12 = 8 lollipops originally. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Now he has 5 toys. So he got 2 more toys. Shawn started with 2 * 2 = 4 toys. He then got 5 + 4 = 9 toys each from his mom and dad. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: So 9 computers were added. Now 5 computers are now in the server room. There were originally 4 * 5 = 20 computers. For each day from monday to thursday, 9 + 20 = 29 more computers were installed. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: So he had 58 at the end of Tuesday, and 23 at the end of wednesday. He lost 2 on Tuesday, and lost 58 - 23 = 35 more on wednesday. Michael started with 35 - 2 = 33 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Now she has 23 dollars left. So she spent 5 dollars. Olivia had 3 dollars. She bought 5 * 3 = 15 bagels for 23 - 15 = 8 dollars each. The answer is 8.

Table 17: Full prompt for “no coherence for language template” setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker is bigger than 65. Harry Vaughan Watkins was 69 years old when he died. 69 was 65 years old when he died. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was killed on July 15, 1997. Gianni Versace was founded by Gianni Versace and shot. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Dambar Shah (? - 1645) was the child of Krishna Shah. Krishna Shah (? - 1661) was the father of Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The nationality of film FAQ: Frequently Asked Questions is not equal to Carlos Atanes. The nationality of film The Big Money is John Paddy Carstairs. The director of Carlos Atanes is Spanish. The director of John Paddy Carstairs is British. Spanish is British. So the final answer (whether they have the same nationality) is: No.

Table 18: Full prompt for “no coherence for language template” setting (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？A：格罗夫的工人们又种了一些树，然后有了 15 棵树。所以一定有 21 棵树被种下了。原来有 21 - 15 = 6 棵树。答案是 6。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？然后又来了三辆汽车。现在停车场有两辆汽车。原来有 3 + 2 = 5 辆汽车。答案是 5。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？A：吃了 32 块，总共还剩 42 块。最初，Leah 有 32 + 42 = 74 块巧克力，她的妹妹有 35 块。所以他们总共有 74 - 35 = 39。答案是 39。问：杰森有 20 个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？答：然后他给了丹尼一些后就有了 20 个。所以他给了丹尼 12 根棒棒糖。杰森原本有 20 - 12 = 8 个棒棒糖。答案是 8。问：肖恩有五个玩具。圣诞节时，他从爸爸妈妈那里各得到了两个玩具。他现在有多少玩具？现在他有五个玩具。他又得到了两个玩具。Shawn 从 2 * 2 = 4 玩具开始。然后他从他的妈妈和爸爸那里得到了 5 + 4 = 9 个玩具。答案是 9。问：服务器机房里有九台电脑。从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？ A：所以增加了 9 台电脑。现在，5 台计算机在服务器室。最初有 4 * 5 = 20 台计算机。从星期一到星期四，每天安装 9 + 20 = 29 台计算机。答案是 29。问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。周三，他又输了两场。星期三结束时他有几个高尔夫球？A：所以他在星期二结束时有 58 个，星期三结束时有 23 个。他在周二输了 2 场，周三又输了 58 - 23 = 35 场。迈克尔一开始是 35 - 2 = 33 高尔夫球。答案是 33。

问：奥利维亚有 23 美元。她买了 5 个百吉饼，每个 3 美元。她还剩多少钱？现在她还剩下 23 美元。她花了 5 美元。奥莉维亚有 3 美元。她买了 5 * 3 = 15 个百吉饼，每个 23 - 15 = 8 美元。答案是 8。表 17：“语言模板无连贯性”设置的全提示（算术推理）。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

答：西奥多·海克尔大于 65 岁。哈里·沃恩·沃特金斯去世时 69 岁。69 岁时，他去世了。所以最后的答案（这个人的名字）是：哈里·沃恩·沃特金斯。

问题：为什么 Versus 的创始人死了？

回答：Versus 于 1997 年 7 月 15 日被杀。Gianni Versace 由 Gianni Versace 创立并拍摄。所以最终的答案（死亡原因）是：枪杀。

问：谁是丹巴尔·沙阿的孙子？

答：丹巴尔沙阿（？- 1645 年）是 Krishna Shah 的孩子。Krishna Shah（？- 1661 年）是鲁德拉沙的父亲。所以最后的答案（孙子的名字）是：Rudra Shah。

问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？

回答：影片的国籍问题：常见问题不等于卡洛斯阿坦尼斯。电影《大款》的国籍是约翰·帕迪卡斯泰尔斯。卡洛斯阿坦尼斯的导演是西班牙人。《约翰·帕迪》的导演卡斯泰尔斯是英国人。西班牙语是英国的。因此，最后的答案（他们是否有相同的国籍）是：没有。

表 18：“语言模板无一致性”设置的完整提示（事实 QA）。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Then there were 21 - 15 = 6 trees after the Grove workers planted some more. So there must have been 15 trees that were planted. There are 21 trees originally. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Then 3 + 2 = 5 more cars arrive. Now 3 cars are in the parking lot. There are originally 2 cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: After eating 32 + 42 = 74, they had 32 pieces left in total. Originally, Leah had 74 - 35 = 39 chocolates and her sister had 35. So in total they had 42. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Then he had 20 - 12 = 8 after giving some to Denny. So he gave Denny 20 lollipops. Jason had 12 lollipops originally. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Now he has 4 toys. So he got 5 + 4 = 9 more toys. Shawn started with 5 toys. He then got 2 * 2 = 4 toys each from his mom and dad. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: So 5 computers were added. Now 4 * 5 = 20 computers are now in the server room. There were originally 9 + 20 = 29 computers. For each day from monday to thursday, 9 more computers were installed. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: So he had 2 at the end of Tuesday, and 23 at the end of wednesday. He lost 35 - 2 = 33 on Tuesday, and lost 58 more on wednesday. Michael started with 58 - 23 = 35 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Now she has 5 * 3 = 15 dollars left. So she spent 5 dollars. Olivia had 23 - 15 = 8 dollars. She bought 3 bagels for 23 dollars each. The answer is 8.

Table 19: Full prompt for "no relevance for language template" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker has 65 golf balls. Harry Vaughan Watkins has 69 golf balls. 69 balls are more than 65 balls. So the final answer (the person who has more golf balls) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: The leader of Versus was Gianni Versace. Gianni Versace shot three people and got into jail. So the final answer (reason for imprisonment) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: The husband of Dambar Shah (? - 1645) is Krishna Shah. Krishna Shah (? - 1661) has a brother called Rudra Shah. So the final answer (the name of the brother-in-law) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The author of the film FAQ: Frequently Asked Questions is Carlos Atanes. The author of film The Big Money is John Paddy Carstairs. The wife of Carlos Atanes is from Spanish. The wife of John Paddy Carstairs is from British. Spanish is warmer than British. So the final answer (the country which is warmer) is: Spanish.

Table 20: Full prompt for "no relevance for language template" setting (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？答：然后有 21 - 15 = 6 棵树后，格罗夫工人种植了一些。所以一定有 15 棵树被种下了。原来有 21 棵树。答案是 6。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？A：然后 3 + 2 = 5 辆汽车就到了。现在停车场有三辆汽车。原来有两辆汽车。答案是 5。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？答：吃了 32 + 42 = 74，总共还剩 32 块。最初，Leah 有 74 - 35 = 39 块巧克力，她的妹妹有 35 块。总共有 42 个。答案是 39。问：杰森有 20 个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？A：然后他给了丹尼一些后有 20 - 12 = 8。所以他给了丹尼 20 个棒棒糖。杰森本来有 12 个棒棒糖。答案是 8。问：肖恩有五个玩具。圣诞节，他从爸爸妈妈那里得到了两个玩具。他现在有多少玩具？A：他现在有四个玩具。所以他得到了 5 + 4 = 9 个玩具。Shawn 从 5 个玩具开始。然后他从他的妈妈和爸爸那里得到了 2 * 2 = 4 个玩具。答案是 9。问：服务器机房里有九台计算机。从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？ 答：所以增加了 5 台计算机。现在 4 * 5 = 20 台计算机现在在服务器室。最初有 9 + 20 = 29 台计算机。从星期一到星期四，每天多安装 9 台电脑。答案是 29。问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。星期三，他又输了两个。星期三结束时他有几个高尔夫球？A：所以他在星期二结束时有 2 个，星期三结束时有 23 个。他在周二以 35 - 2 = 33 输掉了比赛，周三又输了 58 场。迈克尔以 58 - 23 = 35 高尔夫球开始。答案是 33。

问：奥利维亚有 23 美元。她买了五个百吉饼，每个三美元。她还剩多少钱？现在她还剩 5 * 3 = 15 美元。她花了 5 美元。奥利维亚有 23 - 15 = 8 美元。她买了 3 个百吉饼，每个 23 美元。答案是 8。

表 19：“与语言模板无关”设置的完整提示（算术推理）。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

答：西奥多·海克尔有 65 个高尔夫球。哈里·沃恩·沃特金斯有 69 个高尔夫球。69 个球比 65 个球多。所以最后的答案（拥有更多高尔夫球的人）是：哈里·沃恩·沃特金斯。

问题：为什么 Versus 的创始人死了？

回答：Versus 的领导者是 Gianni Versace。Gianni Versace 枪杀了三个人，并入狱。所以最后的答案（监禁的原因）是：枪杀。

问：谁是丹巴尔·沙阿的孙子？

答：丹巴尔沙阿的丈夫（？ - 第 1645 章是我的 Krishna Shah（？ - 1661 年）有一个兄弟叫鲁德拉沙。所以最后的答案（姐夫的名字）是：Rudra Shah。

问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？

答：电影《常见问题：常见问题》的作者是卡洛斯阿坦尼斯。电影 The Big Money 的作者是 John Paddy 卡斯泰尔斯。卡洛斯阿坦尼斯的妻子来自西班牙。约翰·帕迪的妻子卡斯泰尔斯来自英国。西班牙语比英国语更温暖。所以最后的答案（更温暖的国家）是：西班牙。

表 20：“与语言模板无关”设置的完整提示（事实 QA）。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Then there were 21 - 15 = 6 trees after the Grove workers planted some more. So there must have been 15 trees that were planted. There are 21 trees originally. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Then 3 + 2 = 5 more cars arrive. Now 3 cars are in the parking lot. There are originally 2 cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: After eating 32 + 42 = 74, they had 32 pieces left in total. Originally, Leah had 74 - 35 = 39 chocolates and her sister had 35. So in total they had 42. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Then he had 20 - 12 = 8 after giving some to Denny. So he gave Denny 20 lollipops. Jason had 12 lollipops originally. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Now he has 4 toys. So he got 5 + 4 = 9 more toys. Shawn started with 5 toys. He then got 2 * 2 = 4 toys each from his mom and dad. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: So 5 computers were added. Now 4 * 5 = 20 computers are now in the server room. There were originally 9 + 20 = 29 computers. For each day from monday to thursday, 9 more computers were installed. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: So he had 2 at the end of Tuesday, and 23 at the end of wednesday. He lost 35 - 2 = 33 on Tuesday, and lost 58 more on wednesday. Michael started with 58 - 23 = 35 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Now she has 5 * 3 = 15 dollars left. So she spent 5 dollars. Olivia had 23 - 15 = 8 dollars. She bought 3 bagels for 23 dollars each. The answer is 8.

Table 21: Full prompt for "no coherence" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: 65 is bigger than Harry Vaughan Watkins. 65 was 69 years old when he died. Theodor Haecker was 69 years old when he died. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was shot and killed on July 15, 1997. Gianni Versace was founded by Gianni Versace. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Krishna Shah was the child of Rudra Shah. Dambar Shah (? - 1645) was the father of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The nationality of John Paddy Carstairs is not equal to John Paddy Carstairs. The nationality of British is Spanish. The director of Carlos Atanes is British. The director of John Paddy Carstairs is film FAQ: Frequently Asked Questions. Carlos Atanes is film The Big Money. So the final answer (whether they have the same nationality) is: No.

Table 22: Full prompt for "no coherence" setting (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？答：然后有 21 - 15 = 6 棵树后，格罗夫工人种植了一些。所以一定有 15 棵树被种下了。原来有 21 棵树。答案是 6。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？A：然后 3 + 2 = 5 辆汽车就到了。现在停车场有三辆汽车。原来有两辆汽车。答案是 5。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？答：吃了 32 + 42 = 74，总共还剩 32 块。最初，Leah 有 74 - 35 = 39 块巧克力，她的妹妹有 35 块。总共有 42 个。答案是 39。问：杰森有 20 个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？A：然后他给了丹尼一些后有 20 - 12 = 8。所以他给了丹尼 20 个棒棒糖。杰森本来有 12 个棒棒糖。答案是 8。问：肖恩有五个玩具。圣诞节，他从爸爸妈妈那里得到了两个玩具。他现在有多少玩具？A：他现在有四个玩具。所以他得到了 5 + 4 = 9 个玩具。Shawn 从 5 个玩具开始。然后他从他的妈妈和爸爸那里得到了 2 * 2 = 4 个玩具。答案是 9。问：服务器机房里有九台计算机。从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？ 答：所以增加了 5 台计算机。现在 4 * 5 = 20 台计算机现在在服务器室。最初有 9 + 20 = 29 台计算机。从星期一到星期四，每天多安装 9 台电脑。答案是 29。问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。星期三，他又输了两个。星期三结束时他有几个高尔夫球？A：所以他在星期二结束时有 2 个，星期三结束时有 23 个。他在周二以 35 - 2 = 33 输掉了比赛，周三又输了 58 场。迈克尔以 58 - 23 = 35 高尔夫球开始。答案是 33。

问：奥利维亚有 23 美元。她买了五个百吉饼，每个三美元。她还剩多少钱？现在她还剩 5 * 3 = 15 美元。她花了 5 美元。奥利维亚有 23 - 15 = 8 美元。她买了 3 个百吉饼，每个 23 美元。答案是 8。

表 21：“无一致性”设置的完整提示（算术推理）。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

回答：65 岁比哈里·沃恩·沃特金斯还大。65 岁时，他去世了。西奥多·海克尔去世时 69 岁。所以最后的答案（这个人的名字）是：哈里·沃恩·沃特金斯。

问题：为什么 Versus 的创始人死了？

答：1997 年 7 月 15 日，维索斯被枪杀。Gianni Versace 由 Gianni Versace 创立。所以最终的答案（死亡原因）是：枪杀。

问：谁是丹巴尔·沙阿的孙子？

答：克里希纳·沙阿是鲁德拉·沙阿的孩子。丹巴尔·沙阿（？ - 1645）是 Krishna Shah（？ - 1661）。所以最后的答案（孙子的名字）是：Rudra Shah。问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？答：约翰·帕迪卡斯泰尔斯的国籍不等于约翰·帕迪卡斯泰尔斯。的国籍

英国是西班牙。卡洛斯阿坦尼斯的导演是英国人。导演约翰帕迪卡斯泰尔斯是电影常见问题：常见问题。卡洛斯阿坦尼斯是电影的大金钱。因此，最后的答案（他们是否有相同的国籍）是：没有。

表 22：“无一致性”设置的完整提示（事实 QA）。

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Tom started with 4 apples. Then he had 8 after borrowing some from Amy. So he borrowed Amy 8 - 4 = 4. The answer is 4.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Benjamin has 18 gloves originally. Then he got 9 more gloves. So he has 18 + 9 = 27 gloves now. The answer is 27.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Patricia needs to donate 19 inches, and wants her hair to be 31 inches long after the donation. Her hair is 29 inches long currently. Her hair needs to be 19 + 31 = 50 inches long when she cuts it. So she needs to grow 50 - 29 = 21 more inches. The answer is 21.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: There were 37 trains originally. Then there were 14 after some were driven away. So there should be 37 - 14 = 23 that were driven away. The answer is 23.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: The taxi has a ride fee of 8 dollars. Michelle rode the taxi for 6 miles with 2 dollars per mile. So the taxi charge is 6 * 2 = 12. So the total amount that Michelle paid for the ride was 8 + 12 = 20. The answer is 20.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: Haley is currently 23 inches tall. She grows at the rate of 10 inches every year for 4 years. So she will have grown by 10 * 4 = 40 inches. Her height after 4 years will be 23 + 40 = 63 inches. The answer is 63.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Abigail had 46 dollars in her purse originally. She spent $27 in the store, and has 6$ left now. After going shopping, she had 46 - 27 = 19 dollars left. So she lost 19 - 6 = 13 dollars. The answer is 13.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: George earned 48 in total. He sold 7 cars for 6 dollars each. So he earned 7 * 6 = 42 dollars from them. The lego set cost was then 48 - 42 = 6. The answer is 6.

Table 23: Full prompt for "no relevance" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Albin Barack has 49 golf balls. Carl Clemens has 55 golf balls. 55 balls are more than 49 balls. So the final answer (the person who has more golf balls) is: Carl Clemens.

Question: Why did the founder of Versus die?
Answer: The leader of the gang was John Vitti. John Vitti drowned three people and got into jail. So the final answer (reason for imprisonment) is: drowning.

Question: Who is the grandchild of Dambar Shah?
Answer: The husband of Metis Amando is David Amando. David Amando has a brother called Randall Amando. So the final answer (the name of the brother-in-law) is: Randall Amando.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The author of "The Forgortten Bride" is Paul Cuevas. The author of "Grace and the Rose" is Ronnie Dixon. The wife of Paul Cuevas is from Spanish. The wife of Ronnie Dixon is from British. Spanish is warmer than British. So the final answer (the country which is warmer) is: Spanish.

Table 24: Full prompt for "no relevance" setting (factual QA).

问：格罗夫里有 15 棵树。格罗夫的工人今天将在格罗夫植树。在他们完成后，将有 21 棵树。格罗夫的工人们今天种了多少棵树？

汤姆从四个苹果开始。然后他从艾米那里借了一些后有 8 个。所以他借了艾米 8 - 4 = 4。

答案是 4。

问：如果停车场里有 3 辆汽车，又有 2 辆汽车到达，那么停车场里有多少辆汽车？本杰明原来有 18 只手套。然后他又得到了 9 只手套。所以他现在有 18 + 9 = 27 只手套。答案是 27。

问：莉亚有 32 块巧克力，她姐姐有 42 块。如果他们吃了 35 块，他们总共还剩下多少块？答：帕特里夏需要捐赠 19 英寸，并希望她的头发是 31 英寸长的捐赠后。她的头发现在有 29 英寸长。她的头发需要剪到 19 + 31 = 50 英寸长，所以她需要再长 50 29 = 21 英寸。答案是 21。问：杰森有 20 个棒棒糖。他给了丹尼一些棒棒糖。现在杰森有 12 个棒棒糖了。杰森给了丹尼多少棒棒糖？答：最初有 37 列火车。然后有 14 个，有些人被赶走了。所以应该有 37 - 14 = 23 人被赶走。答案是 23。问：肖恩有五个玩具。圣诞节时，他从爸爸妈妈那里各得到了两个玩具。他现在有多少玩具？出租车的车费是 8 美元.米歇尔坐出租车走了 6 英里，每英里 2 美元。所以出租车的费用是 6 * 2 = 12。所以米歇尔支付的总金额是 8 + 12 = 20。答案是 20。问：服务器机房里有九台计算机。从周一到周四，每天都要安装五台电脑。服务器机房中现在有多少台计算机？A：目前，哈莉身高 23 英寸。她以每年 10 英寸的速度生长了 4 年。所以她会长 10 * 4 = 40 英寸。4 年后她的身高将是 23 + 40 = 63 英寸。答案是 63。问：迈克尔有 58 个高尔夫球。周二，他丢了 23 个高尔夫球。周三，他又输了两场。星期三结束时他有几个高尔夫球？阿比盖尔的钱包里原来有 46 美元。她在店里花了 27 美元，现在还剩 6 美元。买完东西，她还剩 46 - 27 = 19 美元。所以她输了 19 - 6 = 13 美元。答案是 13。

问：奥利维亚有 23 美元。她买了五个百吉饼，每个三美元。她还剩多少钱？A：乔治总共赚了 48 英镑。他卖了 7 辆汽车，每辆 6 美元。所以他从他们那里赚了 7 * 6 = 42 美元。乐高套装的成本是 48 - 42 = 6。答案是 6。

表 23：“无相关性”设置的完整提示（算术推理）。

问题：谁活得更长，西奥多·海克尔还是哈里·沃恩·沃特金斯？

回答：阿尔宾巴拉克有 49 个高尔夫球。卡尔·克莱门斯有 55 个高尔夫球。55 个球比 49 个球多。所以最终的答案（拥有更多高尔夫球的人）是：卡尔·克莱门斯。

问题：为什么 Versus 的创始人死了？

答：这伙人的首领是约翰·维蒂。约翰·维蒂淹死了三个人，进了监狱。所以最后的答案（监禁的原因）是：溺水。

问：谁是丹巴尔·沙阿的孙子？

答：梅提斯·阿曼多的丈夫是大卫·阿曼多。大卫·阿曼多有个哥哥叫兰德尔·阿曼多。

所以最后的答案（姐夫的名字）是：兰德尔·阿曼多。

问：这两个导演的电影常见问题：常见问题和导演的电影大款来自同一个国家？

回答：《被遗忘的新娘》的作者是保罗·奎瓦斯。《格蕾丝与玫瑰》的作者是罗尼狄克逊。保罗·奎瓦斯的妻子来自西班牙。罗尼狄克逊的妻子来自英国。西班牙语比英国语更温暖。所以最后的答案（更温暖的国家）是：西班牙。

表 24：“无相关性”设置的完整提示（事实 QA）。