arXiv:2305.05181v2 [cs.CL] 9 Oct 2023

MoT：思维记忆助力 ChatGPT 自我改进

Xiaonan Li,XipengQiu复旦大学计算机科学学院、复旦大学上海市智能信息处理重点实验室 {lixn20, xpqiu}@fudan.edu.cn

摘要

大语言模型（LLM）在各种任务中展现出了令人瞩目的能力。然而，要从根本上提升其能力，依赖于高质量数据集或计算成本高昂的微调。相比之下，人类无需外部资源，仅凭自主思考和记忆便能轻松提升自己。本文提出一种框架 **MoT**，通过 **M**emory-**o**f-**T**hought（思维记忆）让大语言模型实现自我改进，无需标注数据集，也无需更新参数。具体而言，MoT 分为两个阶段：1. 在测试阶段之前，大语言模型在无标签数据集上进行预先思考，并将高置信度思维保存为外部记忆；2. 在测试阶段，给定一个测试问题，大语言模型会回忆相关记忆，以辅助自身推理并作答。实验结果表明，MoT 能显著提升 ChatGPT 在算术推理、常识推理、事实推理和自然语言推理方面的能力。进一步分析表明，每个组成部分都对性能提升至关重要，而且 MoT 能在各种思维链方法和大语言模型上带来稳定的提升。

1 引言

大语言模型（LLM）在广泛的自然语言处理（NLP）任务中展现出了令人惊讶的能力（Chen 等，2023；Zhang 等，2022a；Chowdhery 等，2022；Tay 等，2022；OpenAI，2023；Hoffmann 等，2022；Touvron 等，2023；Mialon 等，2023；Zhao 等，2023；Qiu 等，2020年）。值得注意的是，随着大语言模型的规模扩大到数千亿参数，它们涌现出了新的能力，例如上下文学习（Chen 等，2023；Dong 等，2022）、简单的数字运算和事实性知识查询（Wei 等，2022b）。尤其是，大语言模型的一般推理能力令自然语言处理领域印象深刻，相关技术也取得了一系列新的最先进成果（Wei 等，2022c；Kojima 等人，2022；Lampinen 等，2022；



(a) 大语言模型微调



(b) 预先思考与回忆

图 1：微调与 MoT 的比较：使用带标签数据集对大语言模型进行微调成本高昂，且需要强大的计算资源；而 MoT 无需参数更新和标注数据集，即可通过预先思考与回忆使大语言模型实现自我改进。

Wang 等人，2022b；Huang 和 Chang，2022年）。具体而言，Wei 等人(2022c) 和 Kojima 等人(2022年) 分别提出了少样本思维链和零样本思维链，通过少样本示例和简单但有效的“让我们一步一步地思考”提示来激发大语言模型的推理能力。在此基础上，Wang 等人(2022b)；Press 等人(2022年)；Zhou 等人(2022年)；Wang 等人(2023年)；Weng 等人(2022年) 进一步提出了自一致性、自问、由简到繁、规划并求解等方法，以便在各种专业场景中实现更复杂的推理。

尽管在大型语料库上预训练的大语言模型能力令人瞩目，但要从根本上提升大语言模型的性能，使其超越少样本／零样本基线方法，很大程度上仍依赖于高质量标注数据集，或对大语言模型进行成本高昂的微调。总的来说，这些方法可分为三类：1. 标注数据集 +微调： Wei 等人(2022a) 和 Sanh 等人(2022年) 分别提出 FLAN 和 T0，通过数十个精心整理的自然语言处理基准数据集来增强大语言模型的零样本能力。基于 FLAN，Chung 等人 (2022年) 从模型规模和任务数量两方面扩大训练规模，并证明，加入带有推理依据的思维链示例能够提升大语言模型的推理能力。InstructGPT（Ouyang 等人，2022年）通过使用多样化的众包指令-答案数据进行微调，提升 GPT-3 的指令遵循能力

对。2. 检索标注数据：Liu 等人（2022年）、Su 等人（2022a）和 Agrawal 等人（2022年）使用 SentenceBERT（Reimers 和 Gurevych，2019年）或 BM25（Robertson 和 Zaragoza，2009年）从标注数据集中检索相关示例，以提升大语言模型的上下文学习能力。Rubin 等人（2022年）和 Shi 等人（2022a）利用标注数据集，通过语言模型反馈训练检索器，从而为测试示例检索有用的示例。3. 使用大语言模型生成的数据进行微调：Zelikman 等人（2022年）让语言模型为标注数据集生成推理依据，并通过自我训练增强其推理能力。Magister 等人（2022年）、Ho 等人（2022年）和 Fu 等人（2023年）使用大型语言模型生成的推理路径，提升小型语言模型的推理能力。最近，Huang 等人（2022年）展示了在 PaLM（Chowdhery 等人，2022年）上进行自我训练的有效性。

由于标注高质量数据（尤其是思维链数据中的推理依据）成本高昂，微调大语言模型需要极其强大的计算资源，且计算成本很高。上述依赖微调的方法还面临两项挑战：1. 由于最强大的大语言模型（例如 GPT-4（OpenAI，2023年）和 PaLM（Chowdhery 等人，2022年；Anil 等人，2023年））仅能通过推理 API 公开使用，因此对于大多数研究人员而言，无法通过这些方法改进它们。2. 为增强特定能力而微调大语言模型成本高昂，也不利于环境保护。由于大语言模型的参数规模庞大，微调会带来高昂的模型存储和部署成本。进一步的研究表明，使用专业数据微调大语言模型可能会显著削弱其通用能力（Fu 等人，2023b）。

尽管人们投入了大量精力来收集高质量标注数据集并对大语言模型进行微调，但这既成本高昂，也可能削弱其通用能力；相比之下，人类可以通过元认知过程（Dunlosky 和 Metcalfe，2008年）和记忆机制（Tulving，2002）提升自身的推理能力，同时保留通用能力。例如，记忆有助于人类提升决策、推理、判断等方面的能力（Tulving，2002）。受此启发，我们提出了**MoT**，如图 1所示。这是一个“预先思考—再回忆”框架，旨在让大语言模型通过**思**维**之**记**忆**，在无需监督数据和参数更新的情况下实现自我改进。在预思考阶段，大语言模型会在无标签数据集上进行思考，并将思维保存为外部记忆。在测试阶段，大语言模型会召回相关记忆，以帮助推理并回答给定的测试问题。由于我们关注的是整体框架，并旨在展示其通用性和可扩展性，因此采用简单的组件来实现这两个阶段。具体而言，我们在预思考阶段使用简单的少样本思维链（Wei 等人，2022c），并结合多路径解码策略（Wang 等人，2022b），同时提出答案熵来过滤不确定的思维。对于记忆召回，我们提出了 LLM 检索，它让大语言模型自行检索相关记忆，以帮助回答测试问题。与 SBERT（Reimers 和 Gurevych，2019年）等典型的语义检索器相比，LLM检索比语义嵌入更能捕捉复杂逻辑和推理之间的深层联系。

我们的贡献总结如下：1. 据我们所知，所提出的框架首次使大语言模型能够基于记忆机制提升自身的推理能力，且无需进行参数更新或使用标注数据集。2. 我们在大量数据集上开展了全面实验，结果表明，无需参数更新或使用标注数据集，MoT 就能帮助 ChatGPT 提升算术推理、常识推理、事实推理和自然语言推理能力。进一步分析表明，各个组件都对性能提升至关重要，而且 MoT 能够在各种思维链方法和大语言模型上带来稳定的改进。3. 我们将公开代码和生成的思维链推理路径，以促进未来研究。[1]。本文通过简单组件实例化了所提出的框架，并证明了其有效性。我们希望 MoT 能启发研究人员思考如何借助记忆机制使大语言模型实现自我改进这一重要设计问题，并为进一步提升铺平道路。

2 背景：思维链

大语言模型在各种任务中展现出了令人瞩目的推理能力。思维链（CoT）提示（Wei 等人，2022c；Kojima 等人，2022年）是引导大语言模型进行推理、即生成一系列中间推理步骤以得出最终答案的主流方法。如图2所示，少样本思维链（Wei 等人，2022c；Lampinen 等人，2022）提供若干包含推理依据的示例，即问题／推理依据／答案对，并提示大语言模型

[1] https://github.com/LeeSureman/MoT

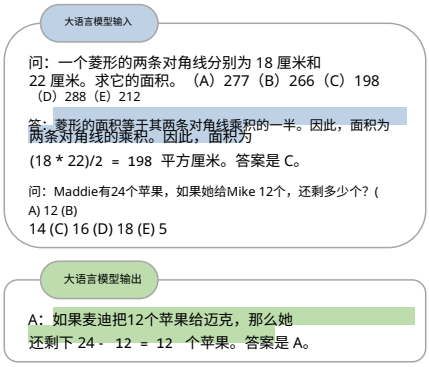


图 2：少样本思维链示意图。

生成能够导向最终答案的推理依据。零样本思维链（Kojima 等人，2022）在测试问题后添加提示“让我们一步一步地思考”，以引导大语言模型进行推理。具体而言，少样本思维链得出的答案如下：

$$s = \text{LLM}(d_1, d_2, \cdots, q_{test}) \quad (1)$$

$$a = \text{Parse-Answer}(s), \quad (2)$$

其中 $d_i = [x_i, r_i, a_i]$ 是第 i 个示例，由输入、推理依据和答案组成。少样本思维链首先根据给定的少样本思维链示例，从大语言模型解码得到 s，然后解析 s 以获取最终答案。由于示例通常采用如下格式：“[输入] [推理依据] 答案是 [答案]”，因此可以通过触发词“答案是”轻松从 中解析出答案（Wei 等人，2022c）。类似地，零样本思维链使用答案触发词，例如“因此，答案是”，从大语言模型生成的零样本推理路径中提取最终答案（Kojima 等人，2022年）。

3 方法

我们在图3中展示了框架的概览。本文主要关注如何在典型的少样本思维链场景中让大语言模型自我改进。在这一场景中，我们拥有一个冻结的大语言模型，以及一个包含少量思维链示例的无标签数据集（Wei 等人，2022c；Huang 等人，2022年）。我们还在第4.3节进一步展示了 MoT 在零样本场景中的有效性。我们的框架分为两个阶段：**1. 预先思考**在测试阶段之前，大语言模型会对无标签数据集进行思考，并将高置信度的推理路径保存为记忆。**2. 回忆** 在测试阶段，对于给定的测试问题，我们提出 LLM 检索，让大语言模型检索相关记忆，以帮助自身进行推理并作答。我们的方法不依赖高质量的带标签数据集，也不需要对大语言模型进行成本高昂的微调；即使大语言模型处于冻结状态或只能通过推理 API 使用，也依然可行。由于我们让大语言模型对无标签数据集进行思考，将其自行生成的思维保存为外部记忆，并检索相关记忆来辅助推理，因此我们将这种借助思维记忆使大语言模型自我改进的方法称为 MoT。下面将介绍这两个阶段。

3.1 预先思考

3.1.1 让大语言模型在测试阶段前思考

在这一阶段，我们让大语言模型对无标签数据集进行思考，并将生成的问题/推理依据/答案对保存为外部记忆。由于我们关注的是整体框架，并旨在展示其通用性和可扩展性，本文将“思考”机制具体实现为简单的少样本思维链（Wei 等人，2022c），并采用多路径解码策略（Wang 等人，2022b）。具体而言，对于无标签数据集中的每个样例 x from the unlabeled dataset X，我们让大语言模型采样 n 条推理路径和答案，温度设为 $T > 0$，记作$[r_1, r_2 \cdots, r_n]$ 和 $[a_1, a_2 \cdots, a_n]$。然后，我们使用多数投票选出最一致的答案，$\tilde{a} = \text{argmax}_{a_i} \sum_{j=1}^{n} \mathbb{1}(a_i = a_j)$，并将导出 $\tilde{a}$的推理路径保留为记忆。由于我们只考虑导向最一致答案的思维，保留的思维可能更加准确（Wang 等人，2022b），也能更好地帮助测试阶段。为简化流程并节省记忆空间，我们为每个无标签样例随机选择一条对应最终答案的推理路径，并将为同一问题保存多个思维留待未来研究。

3.1.2 高置信度思维过滤

由于最一致的答案并不一定会导向正确答案，而错误的示范可能导致表现变差（Yoo 等人，2022年；Lyu 等人，2022年），我们进一步提出根据不确定性对思维进行过滤。受Liu 等人(2020年)；Xin 等人(2020年)的启发，我们提出答案熵 $u(\cdot)$，以过滤掉高不确定性的思维：

$$A^* = \text{unique}(\{a_i\}_{i=1}^{n}) \quad (3)$$

$$p(a_i^*) = \sum_{j=1}^{n} \mathbb{1}(a_i^* = a_j)/n \quad (4)$$

$$u(a_i^*) = -\sum_{i=1}^{|A^*|} p(a_i^*) \log p(a_i^*) \quad (5)$$

其中 $A^* = \{a_1^*, a_2^* \cdots\}$ 是答案集合。$u(\cdot)$ 表示答案的不确定性，$u(\cdot)$ 越高，大语言模型就越不确定。我们过滤掉不确定性为

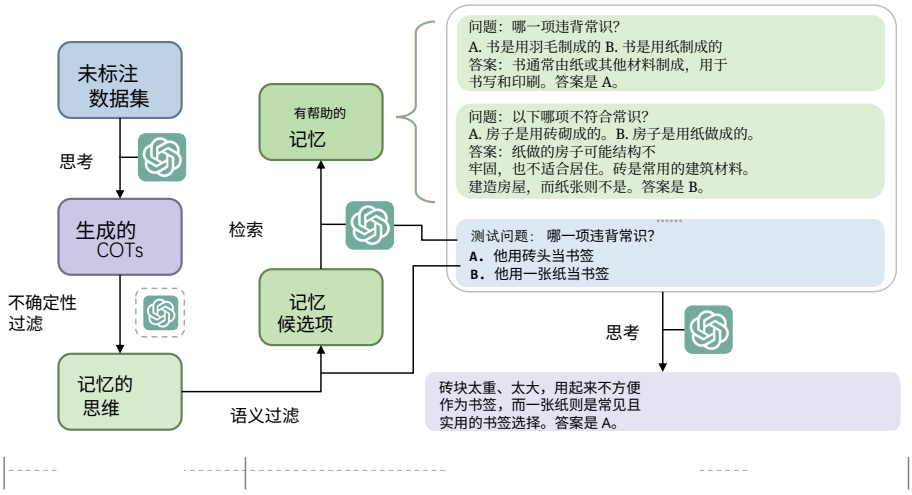


图 3：MoT 概览。

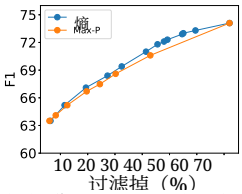


图 4：在 DROP 数据集上进行多路径解码后，准确率与过滤剔除比例之间的关系。在相同的过滤剔除比例下，按答案熵进行过滤的 F1 分数略高于按 Huang 等人 (2022年) 提出的最大概率（Max-P）进行过滤的结果。

高于 τ 且 τ 是预定义的阈值。在探索性实验（图 4）中，我们发现不确定性较低的思维更有可能是正确的：过滤越严格，剩余思维的准确率就越高。因此，答案熵可以过滤掉噪声思维，从而为回忆提供更准确的思维。与 Huang 等人（2022年）的过滤方法相比，该方法使用一致路径的数量（最大概率，简称 Max-P）作为指标，在过滤比例相同的情况下，答案熵带来的准确率略高。

经过过滤后，我们获得了思维记忆池，$M = \{m_i\}_{i=1}^{|M|}$，其中 m_i 是相应输入、推理路径和答案的拼接结果（见图3）。M 由大语言模型针对各种问题生成的高质量思维组成，因此包含了大语言模型回答测试问题所需的关键且有价值的信息。为保持后续内容的一致性，我们将 m_i称为“记忆”或“思维”。

3.2 回忆

在测试阶段，从记忆池 M 中检索相关记忆，以帮助大语言模型回答给定的测试问题 q_{test}。尽管语义嵌入器，例如 SBERT（Reimers 和 Gurevych, 2019年）能够检索语义相关的示例，以用于上下文学习（Liu 等人，2022年），但对于推理任务，它们难以充分捕捉 q_{test}与有用记忆之间深层的逻辑联系，因为单个向量无法直接体现复杂的逻辑和推理路径。鉴于大语言模型（例如 ChatGPT）已展现出令人印象深刻的强大通用自然语言理解能力，以及一定程度的自我意识（Kadavath 等人，2022年），我们提出 LLM检索，让大语言模型为自身检索有用记忆。

由于大语言模型（LLM）存在最大长度限制，因此让大语言模型直接从整个记忆池中进行选择是不可行的。受到人类记忆回忆过程的启发——我们通常会先无意识地筛选出相关记忆，然后再有意识地对其进行评估（Schacter 和 Addis，2007年；Franklin 等人，2005年），我们将 LLM检索分为两个阶段：1. 过滤掉语义上不相关的记忆，获取记忆候选项；2. 让大语言模型从记忆候选项中进行选择。

鉴于示例的多样性已被证明对大语言模型至关重要(Yeet al., 2023年; Levyet al., 2022年; Liand Qiu, 2023年)，我们遵循Li et al.(2022年)采用基于多样性聚类的方法进行记忆检索，即将整个记忆池划分为 l个簇，$\{M_{(1)}\, M^{(2)} \cdots, M^{(l)}\}$，并分别从每个簇中检索一条记忆。具体而言，对于每个簇 $M^{(i)}$，我们首先使用现成的语义嵌入器（例如 SBERT（Reimers andGurevych, 2019年））过滤掉语义上不相关的记忆，并按如下方式获取记忆候选项：

$$M_c^{(i)} = \text{top-k}_{m \in M^{(i)}}(\text{sim}(q_{test}, m)), \quad (6)$$

其中，$\text{sim}(\cdot,\cdot)$ 是语义嵌入的余弦相似度。$M_c^{(i)}$是第 i 个簇的候选项，包含 k 条记忆。

然后，我们进一步让大语言模型从每个记忆簇中选择最有用的记忆，具体如下：

$$m^{(i)} = \text{LLM}(q_{test}, M_c^{(i)}, P_{retrieval}), \quad (7)$$

其中，$P_{retrieval}$ 是用于让大语言模型检索有用记忆的提示。我们通过专门设计的模板，将测试问题 q_{test}、记忆候选项 $M_c^{(i)}$ 和 $P_{retrieval}$ 拼接起来。生成的大语言模型输入如下："参考：[$M_{(i)c,\ 1}$，$M_{c,\ 2}^{(i)}\cdots$，$M_{(i)c,\ k}$] 目标问题：[q_{test}]哪一个参考内容对你回答目标问题最有帮助？"。

通过这种方式，我们可以利用大语言模型强大的自然语言理解能力，为其自身选择最有用的 $M_c^{(i)}$ 记忆。由于这些检索到的记忆来自不同的记忆簇，因此它们不仅能够对 q_{test} 提供帮助，还具有全面性，从而促进大语言模型回答测试问题。同时，语义过滤可以预先筛除语义上不相关的记忆，因而显著减少大语言模型调用次数。

在探索性实验中，我们发现，仅向 LLM 检索提供记忆候选项的问题几乎不会影响检索结果。因此，对于每个 $M_c^{(i)}$，我们仅向大语言模型提供其问题供其选择，这可以显著降低大语言模型的推理成本。为使大语言模型更好地理解检索有用记忆的目标，并使其输出易于按照预定义格式解析，我们在输入中附加了额外指令，例如"你必须以如下格式结束："最有帮助的问题是 *question [idx]*。""。我们在附录A中展示了 LLM 检索的完整输入。

3.3 推理

给定一个测试问题 q_{test}，大语言模型可以基于检索到的记忆进行思考并输出答案，$m^{(1)}$，$m^{(2)}\cdots m^{(k)}$。具体而言，我们让大语言模型以少样本思维链的方式进行推理：

$$s = \text{LLM}(m^{(1)}, m^{(2)}, \cdots, m^{(k)}, q_{test}) \quad (8)$$

$$a = \text{Parse-Answer}(s) \quad (9)$$

在本文中，我们聚焦于整体框架，并使用简单组件对其进行实例化，即少样本思维链、简单的不确定性筛选和大语言模型检索。我们进一步在第4.3节分析 MoT 与不同思维链方法的正交性。我们将探索更多实现留待后续，例如，让大语言模型本身过滤掉不确定的思维（Weng 等人，2022年；Long，2023年），作为未来的工作。

4 实验

4.1 实验设置

数据集 我们在十个数据集上开展实验，涵盖四类任务：**算术推理：** AQuA（Ling 等人，2017）和 DROP（Dua 等人，2019年）；**自然语言推理：** 对抗性自然语言推理子集（Nie 等人，2020年），包括 ANLI-A1、ANLI-A2 和 ANLI-A3，涵盖不同难度级别；**常识推理：** OBQA（Mihaylov 等人，2018年）和 ComV（常识验证）（Wang 等人，2019年）；**事实推理：** BoolQ（Clark 等人，2019年）、FactCK（事实核查器）和 WikiQA（Srivastava 等人，2022年）。数据集概况、统计信息、划分方式和评估指标见附录 B。

方法比较 由于我们关注 MoT 能否帮助大语言模型自我改进，因此在同一个大语言模型 ChatGPT（GPT-3.5-Turbo-0301）上，将 MoT 与基线方法进行比较，包括零样本/少样本思维链和零样本/少样本直接提示。为分析推理依据和思考对 MoT 的影响，我们还将 MoT 与其两个变体进行比较：1）MoT（无推理依据），即移除检索记忆中的推理依据，从而让大语言模型直接输出答案，可视为 MoT 的少样本直接提示版本；2）MoT（无思考），即保留检索记忆中的推理依据，但要求大语言模型不使用思维链，直接回答问题。具体而言，我们将"答案是"添加为大语言模型的输出前缀，以提示其直接输出答案。通过这两个变体，我们可以分别分析推理依据和思考对 MoT 的影响。此外，我们还在标注数据集上开展 MoT 实验，即 MoT（使用黄金标准），以探究其潜在的改进空间；在此设置下，我们使用黄金标签过滤掉错误记忆。因此，MoT 不会因错误答案而退化。

实现细节 除非另有说明，我们使用公开的 Ope- nAI 语言模型"gpt-3.5-turbo-0301"，而在"text-davinci-002/003"上的实验（附录C）显示了相同的趋势。在回忆时，我们使用 SBERT（"all- mpnet-base-v2"）（Reimers 和 Gurevych，2019年）进行语义过滤。我们将簇的数量 l（即示例数量）和每个簇的记忆候选项数量 k 分别设为 4 和 10。我们进一步分析了示例数量，

方法	算术推理		ANLI			CS 推理		事实推理			AVG
	AQuA	DROP	-A1	-A2	-A3	OBQA	ComV	BoolQ	FactCK	WikiQA	
零样本	27.7	24.7	54.4	48.0	51.7	79.4	90.5	63.4	75.6	52.6	56.8
少样本	28.9	46.3	55.0	48.5	51.1	82.0	90.8	64.4	77.0	32.5	57.6
MoT（无推理过程）	27.0	59.4	56.2	50.3	52.6	**84.2**	91.0	70.1	82.1	53.9	62.7
MoT（无思考）	24.4	59.4	55.6	50.2	52.6	81.3	90.5	71.6	82.2	64.3	63.1
零样本思维链	51.7	62.2	61.9	51.6	48.5	69.2	87.1	53.0	66.0	49.9	60.1
少样本思维链	49.7	57.6	59.7	48.1	52.3	80.0	94.5	67.7	80.6	65.2	65.5
MoT	**54.1**	**65.7**	**64.6**	**52.8**	**55.2**	82.3	**95.5**	**71.5**	**82.2**	**68.0**	**69.2**
MoT（含黄金标准）	56.5	71.0	65.7	55.6	55.4	82.8	94.6	74.2	86.6	70.6	71.3

表1：ChatGPT（GPT-3.5-Turbo-0301）上的性能对比。

详见附录 D。在测试阶段，为了保证结果稳定，除非另有说明，我们使用贪心解码生成输出。完整的实现细节和少样本示例列于附录E。

4.2 主要结果

我们在表1中展示了实验结果。可以看出，MoT 在大多数数据集上的表现都显著优于基线方法，这表明 MoT 在一系列自然语言处理任务中具备最出色的综合推理能力。具体而言，MoT 分别比少样本思维链和零样本思维链高出 3.7 分和 9.1 分，这直接证明了 MoT 能够通过思维记忆使大语言模型实现自我提升，而无需标注数据集和参数更新。值得注意的是，零样本思维链在 ChatGPT 上表现出色，并且在多个数据集上优于少样本思维链，例如 AQuA、DROP、ANLI-1 和 ANLI-2。这表明，对于具备强大零样本推理能力的大语言模型而言，无关的思维链示例可能并非必需。与此同时，MoT 在所有数据集上的表现始终优于零样本思维链，这表明检索记忆确实能够发挥积极作用。

至于 MoT 的两种变体，它们的整体表现也优于零样本和少样本方法。与此同时，尽管它们直接输出答案，但在若干数据集上（如 OBQA、BoolQ 和 FactCK）的表现优于零样本思维链和少样本思维链。这类似于人类的一种常见现象：回忆相关记忆时，我们无需有意识地进行推理，仅凭直觉也能表现出色（Dijksterhuis 和 Nordgren，2006年；Todd 等人，1999）。此外，尽管 MoT（无思考）提供了推理依据，而 MoT（无推理依据）则没有，但二者的表现总体相近，这表明显式推理对于大语言模型充分利用检索记忆是必要的。简而言之，两者都 relevant 记忆和显式推理对于 MoT 在大量数据集上持续取得提升至关重要。

方法	OBQA	BoolQ	WikiQA
解码路径=*8*			
零样本思维链$_{T=0.7}$	80.4	55.9	59.2
零样本思维链$_{T=1}$	82.4	54.1	59.2
零样本思维链$_{T=1.2}$	81.0	49.1	59.4
少样本思维链$_{T=0.7}$	82.0	69.2	69.6
少样本思维链$_{T=1}$	83.6	68.2	70.8
少样本思维链$_{T=1.2}$	82.2	70.2	70.8
MoT$_{T=0.7}$	**85.0**	73.2	**73.2**
MoT$_{T=1}$	84.4	**73.5**	72.1
MoT$_{T=1.2}$	85.0	72.8	72.1
解码路径=*16*			
零样本思维链$_{T=0.7}$	81.6	58.0	63.8
零样本思维链$_{T=1}$	83.2	55.8	65.7
零样本思维链$_{T=1.2}$	83.4	53.8	64.3
少样本思维链$_{T=0.7}$	83.2	69.6	71.5
少样本思维链$_{T=1}$	83.6	69.2	71.7
少样本思维链$_{T=1.2}$	83.6	69.7	70.9
MoT$_{T=0.7}$	84.4	**73.7**	73.7
MoT$_{T=1}$	**85.0**	72.7	**74.2**
MoT$_{T=1.2}$	84.8	73.3	74.1

表 2：不同超参数下自一致性策略的性能比较。

此外，MoT（使用黄金标准）比 MoT 表现更好，这表明 MoT 应用更先进的思维链方法时具有提升潜力（Wang 等人，2023年；Zhou 等人，2022年；Zheng 等人，2023年；Long，2023年），以及验证方法（Weng 等人，2022年；Manakul 等人，2023年）。

4.3 分析

多次解码性能 在本节中，我们在自一致性策略下评估 MoT（Wang 等人，2022b），该策略会进行多次解码，并通过多数投票得出最终答案。我们在 OBQA、BoolQ 和 WikiQA 上，针对不同的采样次数和温度，将 MoT 与基线方法进行比较，结果见表 2。我们发现，MoT 在不同的解码温度和采样次数下，始终优于零样本思维链和少样本思维链，

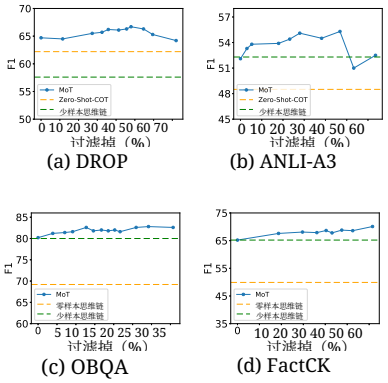
	DROP	ANLI-A3	BoolQ	WikiQA
少样本思维链	57.6	52.3	67.7	65.2
MiniLM	63.0	53.7	70.2	67.0
Instructor-base	64.2	53.2	70.2	66.9
随机	57.5	52.8	69.7	66.3
+ MPNet	64.7	53.3	70.4	67.1
+ LLM检索	**65.7**	**55.2**	**71.5**	**68.0**

表 3：检索方法比较。

这表明 MoT 具有普适性和稳定性。我们注意到，随着采样次数增加，性能提升略有减弱。这与人类身上的一种现象相似：我们对一个问题思考得越仔细，先前的准备所起的作用就越小。

LLM检索的效果 为评估 LLM检索对 MoT 的效果，我们在 DROP、ANLI-A3、BoolQ 和 WikiQA 上采用不同的检索方法进行了实验，结果如表 3所示。除 MoT 中使用的 SBERT（“all-mpnet-base-v2”，简称 MPNet）（Reimers 和 Gurevych，2019年）外，我们还比较了另外两种语义嵌入器：SBERT（“all-MiniLM-L6-v2”，简称 MiniLM）（Reimers 和 Gurevych，2019年）以及 Instructor-base（Su 等人，2022b），后者通过 330种多样化任务进行训练，支持多种应用场景。我们发现，仅使用 MPNet 进行记忆检索，相较于少样本思维链也能带来显著提升，这表明 MoT 在大语言模型 API 预算有限时仍然实用。使用大语言模型检索记忆后，性能进一步提升，这直接证明了 LLM检索的有效性。此外，我们发现 LLM检索优于所有参与比较的语义嵌入器，这表明大语言模型比语义嵌入更能捕捉复杂的推理逻辑。

过滤的影响 为了评估记忆过滤对 MoT 的影响，我们绘制了在 DROP、ANLI-A3、OBQA 和 WikiQA 上采用不同过滤阈值时的性能曲线。具体而言，我们均匀地调整答案熵的过滤阈值，并观察相应的性能。结果如图5所示。我们发现，不进行过滤的 MoT 性能显著下降，在某些数据集（如 OBQA 和 FactCK）上的表现略逊于少样本思维链，这表明错误记忆会损害大语言模型的推理，因此我们的过滤策略是

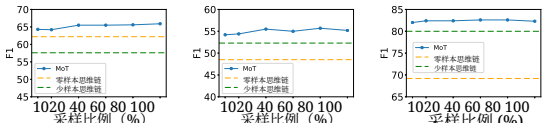


(a) DROP (b) ANLI-A3 (c) OBQA (d) FactCK

图5：过滤的效果。横轴最左端和最右端分别对应不进行过滤和最严格的过滤，即只保留所有路径都.....的记忆

可以保留得出相同答案的内容。

必要的。同时，大多数过滤阈值都能持续优于基线方法，这表明 MoT 带来的改进总体上不受过滤阈值这一超参数的影响。



(a) DROP (b) ANLI-A3 (c) OBQA

图6：有限内存规模性能。

有限内存规模性能 在现实场景中，无标签示例的数量或可用外部内存空间的大小可能受到限制，而这些因素都可能导致内存大小受限。在本节中，我们评估 MoT 在不同内存大小下的表现。具体而言，我们对按不同比例随机抽取的子集进行实验，并在图 6 中绘制相应的性能曲线。我们观察到，MoT 能够持续带来性能提升。即使仅使用原始记忆池 10% 的内容，MoT 仍然能够超越零样本思维链和少样本思维链。这些结果表明，当无标签示例或可用外部内存空间有限时，MoT 仍具有良好的可用性。

不同思维链方法间的迁移能力本节评估 MoT 在另外两种思维链方法上的表现：零样本思维链（Kojima 等人，2022年）和规划并求解（Wang 等人，2023年）。与使用“让我们一步一步地思考”来引导大语言模型进行推理的零样本思维链相比，规划并求解使用专门设计的提示，让大语言模型先制定计划，将整个任务划分为多个子任务，然后

	DROP	ANLI-A1	ANLI-A3	OBQA
少样本思维链	57.6	59.7	52.3	80.0
+MoT	65.7	65.6	55.2	82.3
零样本思维链	62.2	61.9	48.5	69.2
+MoT	66.6	65.9	54.0	81.5
规划并求解	60.5	62.6	52.1	71.4
+MoT	67.6	66.3	56.9	85.6

表4：各种思维链方法的比较。

根据计划逐一求解这些子任务，因此能够完成更复杂的推理（Wang 等人，2023年）。对于这两种思维链方法，我们分别在思考前阶段使用它们生成思维记忆池。在测试阶段，我们从相应的记忆池中检索思维，将其与测试问题拼接，然后使用相应的提示（例如，零样本思维链使用“让我们一步一步地思考”），引导大语言模型进行推理。DROP、ANLI-A1、ANLI-A3 和 OBQA 上的结果如表 4 所示。我们观察到，MoT 始终带来性能提升，这表明它在各种思维链方法中具有稳定性和通用性。此外，由于这两种思维链方法不依赖人工思维链示例，这些结果也证明了在没有人工思维链示例时 MoT 仍然有效。同时，使用更先进的思维链方法——规划并求解时，MoT 的性能进一步提升，这表明当来来出现更强大的思维链方法时，MoT 具有应用潜力。

5 相关工作

利用大语言模型生成的数据增强模型 本节介绍此前利用大语言模型生成的数据来增强模型的方法。Ye 等人（2022a）；Gao 等人（2022）；Ye 等人（2022b）提出了 ZeroGen、ProGen 和 ZeroGen$^+$，利用大语言模型生成数据集，以增强小型模型，例如长短期记忆网络。Fu 等人（2023年）；Magister 等人（2022年）；Ho 等人（2022年）利用大语言模型生成推理路径，并教会小型语言模型进行推理。Wang 等人（2022c）和 Honovich 等人（2022年）利用大语言模型生成指令数据，以提升大语言模型遵循指令的能力。Schick 等人（2023年）提出 ToolFormer，通过自生成数据学习如何使用各种工具。Zelikman 等人（2022年）和 Huang 等人（2022）分别利用大语言模型生成推理路径，并通过带标签和无标签数据集来改进自身，带标签和无标签数据集。不同于这些依赖高成本微调的方法，MoT 可通过思维记忆使大语言模型自我改进，无需参数更新，并且兼容可通过 API 访问的大语言模型。近期，Zhang 等人（2022b）；Shao 等人（2023年）利用大语言模型自动生成思维链示例。Li 等人（2022年）利用大语言模型生成知识库，并提升其开放域问答能力。这些方法可视为 MoT 的特例，采用了任务级记忆选择或任务专用记忆构建。

大语言模型的示例检索本节介绍以往用于上下文学习的示例检索方法。这些方法主要从标注数据集中检索相关的输入/输出对，供语言模型预测测试样例。Liu 等人(2022年提出利用稠密语义嵌入器检索相关样例，以改进上下文学习。Agrawal 等人(2022年) 利用 BM25 为机器翻译的上下文学习检索样例。Das 等人(2021年) 和 Hu 等人(2022年) 分别针对知识型问答和对话状态跟踪的上下文学习，设计专门的目标相似度来训练示例检索器。Rubin 等人(2022年)；Shi 等人(2022b) 利用语言模型的反馈，为语义解析训练示例检索器。Lyu 等人(2022年) 检索带有随机标签的相关样例，并提出启发式方法以减少错误标签带来的负面影响。最近，Li 等人(2023年) 提出了 UDR，这是一种面向各种自然语言处理任务的统一示例检索器，通过统一的语言模型反馈在约 40 个标注数据集上进行训练。尽管这些方法大多依赖高质量标注数据集，且只探索不带推理依据的上下文学习，但 MoT 无需标注数据集和参数更新即可让大语言模型自我改进。据我们所知，我们率先探索了具有挑战性且复杂的推理场景中的示例检索，并展示了 MoT 的有效性。

6 结论

在本文中，我们提出了 MoT，这是一种通过思维记忆让大语言模型实现自我改进的框架，无需标注数据集和参数更新。实验结果表明，MoT 能够显著提升 ChatGPT 在算术推理、常识推理、事实性推理和自然语言推理方面的能力。进一步的分析表明，

每个组件都对性能提升起着关键作用，并且 MoT 能够在各种思维链方法和大语言模型上带来一致的改进。

局限性

MoT 主要有以下局限性：

- 尽管我们提出了答案熵来过滤掉不确定的思维，但剩余的思维仍可能包含一些错误。未来我们将探索更多错误思维过滤方法（Lin 等人，2023年）。

- 本文采用了一种简单策略来利用相关记忆，即将其与测试输入 q_{test} 拼接，从而帮助大语言模型生成答案 q_{test}。我们将探索更多利用检索记忆的策略，例如，检索记忆以验证当前的推理路径 q_{test}。

- 一方面，本文迈出了基于记忆机制让大语言模型自我改进的第一步。所开展的实验仍处于安全设置中，即使用特定的无标签数据集，且大语言模型无法访问互联网或控制外部工具。因此，我们认为我们的方法和实验仍足够安全，不会对社会造成严重影响或带来不可逆转的后果。另一方面，大语言模型已表现出各种偏见（Bender 等人，2021年）。由于我们让大语言模型生成思维/记忆来帮助自身，大语言模型可能会受到生成内容中偏见的影响。我们认为，大语言模型去偏是一个重要的未来研究课题。

参考文献

Sweta Agrawal、Chunting Zhou、Mike Lewis、Luke Zettlemoyer 和 Marjan Ghazvininejad。2022年。用于机器翻译的上下文示例选择。*CoRR*，abs/2212.02437。

Rohan Anil、Andrew M. Dai、Orhan Firat、Melvin Johnson、Dmitry Lepikhin、Alexandre Passos、Siamak Shakeri、Emanuel Taropa、Paige Bailey、Zhifeng Chen、Eric Chu、Jonathan H. Clark、Laurent El Shafey、YanpingHuang、KathyMeier-Hellstern、Gau- rav Mishra、Erica Moreira、Mark Omernick、Kevin Robinson、Sebastian Ruder、Yi Tay、Kefan Xiao、Yuanzhong Xu、Yujing Zhang、Gustavo Hernández Ábrego、Junwhan Ahn、Jacob Austin、Paul Barham、Jan A. Botha、James Bradbury、Siddhartha Brahma、Kevin Brooks、Michele Catasta、Yong Cheng、Colin Cherry、Christopher A. Choquette-Choo、Aakanksha Chowdhery、Clément Crepy、Shachi Dave、Mostafa Dehghani、Sunipa Dev、Jacob Devlin、Mark Díaz、Nan Du、Ethan Dyer、Vladimir Feinberg、Fangxiaoyu Feng、Vlad Fienber、Markus Freitag、Xavier Garcia、SebastianGehrmann、LucasGonzalez、等人。2023年。PaLM 2 技术报告。*CoRR*，abs/2305.10403。

Emily M. Bender、Timnit Gebru、Angelina McMillan- Major 和 Shmargaret Shmitchell。2021年。随机鹦鹉的危险：语言模型会不会太大？载于 *FAccT '21*：*2021*年 *ACM* 公平性、问责制与透明度会议，线上活动 / 多伦多，加拿大，*2021*年*3*月*3*日至*10*日，第 610–623 页。ACM。

ZekaiChen、MariannMicsinai Balan 和 KevinBrown。2023年。语言模型是用于预后预测的少样本学习器。*CoRR*，abs/2302.12692。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vin- odkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael I sard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Henryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、Daphne Ippolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agrawal、Mark Omernick、An- drew M. Dai、Thanumalayan Sankaranarayana Pil- lai、Marie Pellat、Aitor Lewkowycz、Erica Moreira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、DouglasEck、JeffDean、Slav Petrov 和 Noah Fiedel。2022年。PaLM：借助路径扩展语言建模规模。*CoRR*，abs/2204.02311。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Eric Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma、Albert Webson、Shixiang Shane Gu、Zhuyun Dai、Mirac Suz- gun、Xinyun Chen、Aakanksha Chowdhery、Sharan Narang、Gaurav Mishra、Adams Yu、Vincent Y. Zhao、Yanping Huang、Andrew M. Dai、Hongkun Yu、Slav Petrov、Ed H. Chi、Jeff Dean、Jacob Devlin、Adam Roberts、Denny Zhou、Quoc V. Le 和 Jason Wei。2022年。扩大指令微调语言模型的规模。*CoRR*，abs/2210.11416。

Christopher Clark、Kenton Lee、Ming-Wei Chang、Tom Kwiatkowski、Michael Collins 和 Kristina Toutanova。2019年。BoolQ：探索自然语言是非问题出人意料的难度。载于《*2019*年北美计算语言学协会分会会议论文集：人类语言技术》，第*1*卷（长篇与短篇论文），第2924–2936页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Rajarshi Das、Manzil Zaheer、Dung Thai、Ameya Godbole、Ethan Perez、Jay Yoon Lee、Lizhen Tan、Lazaros Polymenakos 和 Andrew McCallum。2021年。基于案例的推理：针对知识库的自然语言查询。载于《*2021*年自然语言处理经验方法会议论文集》，*EMNLP 2021*，线上活动／蓬塔卡纳，多米尼加共和国，*2021*年*11*月7日至*11*日，第9594–9611页。计算语言学协会。

A. J. Dijksterhuis 和 Loran F Nordgren。2006年。无意识思维理论。心理科学展望，1：109–95。

Qingxiu Dong、Lei Li、Damai Dai、Ce Zheng、Zhiyong Wu、Baobao Chang、Xu Sun、Jingjing Xu、LeiLi 和 Zhifang Sui。2022年。上下文学习综述。

Dheeru Dua、Yizhong Wang、Pradeep Dasigi、Gabriel Stanovsky、Sameer Singh 和 Matt Gardner。2019年。DROP：一种要求对段落进行离散推理的阅读理解基准。载于*2019*年北美计算语言学协会分会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文），第2368–2378页，明尼阿波利斯，明尼苏达州。计算语言学协会。

J. Dunlosky 和 J. Metcalfe。2008年。元认知。SAGE 出版社。

Stan Franklin、Bernard Baars、Uma Ramamurthy 和 Matthew Ventura。2005年。意识在记忆中的作用。《大脑、心智与媒体》，1。Yao Fu、Hao Peng、Litu Ou、Ashish Sabharwal 和 Tushar Khot。2023年。面向多步推理的较小语言模型专业化。*CoRR*，abs/2301.12726。Yao Fu、Hao Peng、Litu Ou、Ashish Sabharwal 和 Tushar Khot。2023年。面向多步推理的较小语言模型专业化。*CoRR*，abs/2301.12726。Jiahui Gao、Renjie Pi、Yong Lin、Hang Xu、Jiacheng Ye、Zhiyong Wu、Xiaodan Liang、Zhenguo Li 和 Lingpeng Kong。2022年。ZeroGen$^+$：高效零样本学习中的自我引导高质量数据生成。*CoRR*，abs/2205.12679。Namgyu Ho、Laura Schmid 和 Se-Young Yun。2022年。大语言模型是推理教师。*CoRR*，abs/2212.10071。Jordan Hoffmann、Sebastian Borgeaud、Arthur Mensch、Elena Buchatskaya、Trevor Cai、Eliza Rutherford、Diego de Las Casas、Lisa Anne Hendricks、Johannes Welbl、Aidan Clark、Tom Hennigan、Eric Noland、Katie Millican、George van den Driessche、Bogdan Damoc、Aurelia Guy、Simon Osindero、Karen Simonyan、Erich Elsen、Jack W. Rae、Oriol Vinyals 和 Laurent Sifre。2022年。训练计算最优的大语言模型。*CoRR*，abs/2203.15556。

Or Honovich、Thomas Scialom、Omer Levy 和 Timo Schick。2022年。非自然指令：几乎无需人工劳动即可微调语言模型。*CoRR*，abs/2212.09689。

Yushi Hu、Chia-Hsuan Lee、Tianbao Xie、Tao Yu、Noah A. Smith 和 Mari Ostendorf。2022年。用于少样本对话状态跟踪的上下文学习。*CoRR*，abs/2203.08568。

Jiaxin Huang、Shixiang Shane Gu、Le Hou、Yuexin Wu、Xuezhi Wang、Hongkun Yu 和 Jiawei Han。2022年。大语言模型能够自我改进。*CoRR*，abs/2210.11610。

Jie Huang 和 Kevin Chen-Chuan Chang。2022年。迈向大语言模型推理：一篇综述。*CoRR*，abs/2212.10403。

Saurav Kadavath、Tom Conerly、Amanda Askell、Tom Henighan、Dawn Drain、Ethan Perez、Nicholas Schiefer、Zac Hatfield-Dodds、Nova DasSarma、Eli Tran-Johnson、Scott Johnston、Sheer ElShowk、Andy Jones、Nelson Elhage、Tristan Hume、Anna Chen、Yuntao Bai、Sam Bowman、Stanislav Fort、Deep Ganguli、Danny Hernandez、Josh Jacobson、Jack- son Kernion、Shauna Kravec、Liane Lovitt、Ka- mal Ndousse、Catherine Olsson、Sam Ringer、Dario Amodei、Tom Brown、Jack Clark、Nicholas Joseph、Ben Mann、Sam McCandlish、Chris Olah 和 Jared Kaplan。2022年。语言模型（大多）知道自己知道什么。*CoRR*，abs/2207.05221。

Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yu- taka Matsuo 和 Yusuke Iwasawa。2022年。大语言模型是零样本推理器。*CoRR*，abs/2205.11916。Andrew K. Lampinen、Ishita Dasgupta、Stephanie C. Y. Chan、Kory W. Mathewson、Mh Tessler、Antonia Creswell、James L. McClelland、Jane Wang 和 Fe- lix Hill。2022年。语言模型能否从上下文中的解释中学习？载于 计算语言学协会论文集：*EMNLP 2022*，阿布扎比，阿拉伯联合酋长国，*2022*年*12*月7日至*11*日，第537–563页。计算语言学协会。Itay Levy、Ben Bogin 和 Jonathan Berant。2022年。多样化示例可提升上下文学习中的组合泛化能力。*CoRR*，abs/2212.06800。Junlong Li、Zhuosheng Z hang 和 Hai Zhao。2022年。用于开放领域问答的自提示大语言模型。*CoRR*，abs/2212.08635。Xiaonan Li、Kai Lv、Hang Yan、Tianyang Lin、Wei Zhu、Yuan Ni、Guotong Xie、Xiaoling Wang 和 Xipeng Qiu。2023年。用于上下文学习的统一示例检索器。*CoRR*，abs/2305.04320。Xiaonan Li 和 Xipeng Qiu。2023年。为上下文学习寻找支持性示例。*CoRR*，abs/2302.13539。

Zhen Lin、Shubhendu Trivedi 和 Jimeng Sun。2023年。带置信度生成：黑盒大语言模型的不确定性量化。*CoRR*，abs/2305.19187。

Wang Ling、Dani Yogatama、Chris Dyer 和 Phil Blunsom。2017。通过生成推理过程进行程序归纳：学习解决并解释代数文字题。载于第*55*届计算语言学协会年会论文集，*ACL 2017*，温哥华，加拿大，*7*月*30*日—*8*月*4*日，第*1*卷：长篇论文，第158–167页。计算语言学协会。

Jiachang Liu、Dinghan Shen、Yizhe Zhang、Bill Dolan、Lawrence Carin 和 Weizhu Chen。2022年。GPT-3 的优质上下文示例有何特点？载于《*Deep Learning Inside Out*：深度学习架构知识提取与整合第三届研讨会论文集》，*DeeLIO@ACL 2022*，都柏林，爱尔兰及线上，*2022*年*5*月*27*日，第100–114页。计算语言学协会。

Weijie Liu、Peng Zhou、Zhiruo Wang、Zhe Zhao、Haotang Deng 和 Qi Ju。2020年。FastBERT：一种具有自适应推理时间的自蒸馏 BERT。载于第*58*届计算语言学协会年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，第6035–6044页。计算语言学协会。

Jieyi Long。2023年。大型语言模型引导的思维树。*CoRR*，abs/2305.08291。

Xinxi Lyu、Sewon Min、Iz Beltagy、Luke Zettlemoyer 和 Hannaneh Hajishirzi。2022年。Z-ICL：利用伪示范进行零样本上下文学习。*CoRR*，abs/2212.09865。

Lucie Charlotte Magister、Jonathan Mallinson、Jakub Adámek、Eric Malmi 和 Aliaksei Severyn。2022年。教小型语言模型进行推理。*CoRR*，abs/2212.08410。

Potsawee Manakul、Adian Liusie 和 Mark J. F. Gales。2023年。SelfCheckGPT：面向生成式大语言模型的零资源黑盒幻觉检测。*CoRR*，abs/2303.08896。

Grégoire Mialon、Roberto Dessì、Maria Lomeli、Christoforos Nalmpantis、Ramakanth Pasunuru、Roberta Raileanu、Baptiste Rozière、Timo Schick、Jane Dwivedi-Yu、Asli Celikyilmaz、Edouard Grave、Yann LeCun 和 Thomas Scialom。2023年。增强型语言模型：一项综述。*CoRR*，abs/2302.07842。

Todor Mihaylov、Peter Clark、Tushar Khot 和 Ashish Sabharwal。2018年。一套盔甲能导电吗？用于开放式图书问题回答的新数据集。载于*2018*年自然语言处理实证方法会议论文集，第 2381–2391 页，比利时布鲁塞尔。计算语言学协会。

Yixin Nie、Adina Williams、Emily Dinan、Mohit Bansal、Jason Weston 和 Douwe Kiela。2020年。对抗性自然语言推理：自然语言理解的新基准。载于计算语言学协会第*58*届年会论文集，第4885–4901页，线上。计算语言学协会。

OpenAI。2023年。GPT-4 技术报告。*CoRR*，abs/2303.08774。

Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul F. Christiano、Jan Leike 和 Ryan Lowe。2022年。通过人类反馈训练语言模型以遵循指令。*CoRR*，abs/2203.02155。

Ofir Press、Muru Zhang、Sewon Min、Ludwig Schmidt、Noah A. Smith 和 Mike Lewis。2022年。衡量并缩小语言模型中的组合性差距。*CoRR*，abs/2210.03350。

Xipeng Qiu、Tianxiang Sun、Yige Xu、Yunfan Shao、Ning Dai 和 Xuanjing Huang。2020年。自然语言处理预训练模型：综述。*CoRR*，abs/2003.08271。

Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。收录于*2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP 2019*）论文集，*2019*年*11*月*3*日至*7*日，中国香港，第3980–3990页。计算语言学协会。

Stephen E. Robertson 和 Hugo Zaragoza。2009年。概率相关性框架：BM25 及其后续发展。信息检索基金会趋势，3(4):333–389。

Ohad Rubin、Jonathan Herzig 和 Jonathan Berant。2022年。学习检索上下文学习提示。载于 *2022*年计算语言学协会北美分会年会论文集：人类语言技术，*NAACL 2022*，美国华盛顿州西雅图，*2022*年*7*月*10*日至*15*日，第2655–2671页。计算语言学协会。

Victor Sanh、Albert Webson、Colin Raffel、Stephen H. Bach、Lintang Sutawika、Zaid Alyafeai、Antoine Chaffin、Arnaud Stiegler、Arun Raja、Manan Dey、M Saiful Bari、Canwen Xu、Urmish Thakker、Shanya Sharma Sharma、Eliza Szczechla、Taewoon Kim、Gunjan Chhablani、Nihal V. Nayak、Debajyoti Datta、Jonathan Chang、Mike Tian-Jian Jiang、Han Wang、MatteoManica、ShengShen、ZhengXin Yong、Harshit Pandey、Rachel Bawden、Thomas Wang、Tr- ishala Neeraj、Jos Rozen、Abheesht Sharma、An- drea Santilli、Thibault Févry、Jason Alan Fries、Ryan Teehan、Teven Le Scao、Stella Biderman、Leo Gao、

Thomas Wolf 和 Alexander M. Rush。2022年。多任务提示训练实现零样本任务泛化。载于第十届国际学习表征会议，*ICLR 2022*，线上活动，*2022*年*4*月*25*日至*29*日。OpenReview.net。

Daniel L. Schacter 和 Donna Rose Addis。2007。《建构性记忆的认知神经科学：回忆过去与想象未来》。载于 Jon Driver、Patrick Haggard 和 Tim Shallice 编，《style id='1'>人类大脑中的心理过程》。牛津大学出版社。

Timo Schick、Jane Dwivedi-Yu、Roberto Dessì、Roberta Raileanu、Maria Lomeli、Luke Zettlemoyer、Nicola Cancedda 和 Thomas Scialom。2023年。Toolformer：语言模型可以自学使用工具。*CoRR*，abs/2302.04761。

Zhihong Shao、Yeyun Gong、Yelong Shen、Min- lie Huang、Nan Duan 和 Weizhu Chen。2023年。合成提示：为大语言模型生成思维链示例。*CoRR*，abs/2302.00618。

Peng Shi、Rui Zhang、He Bai 和 Jimmy Lin。2022a。XRICL：用于跨语言文本到 SQL 语义解析的跨语言检索增强上下文学习。*CoRR*，abs/2210.13693。

Peng Shi、Rui Zhang、He Bai 和 Jimmy Lin。2022b。XRICL：用于跨语言文本到 SQL 语义解析的跨语言检索增强型上下文学习。*CoRR*，abs/2210.13693。

Aarohi Srivastava、Abhinav Rastogi、Abhishek Rao、Abu Awal Md Shoeb、Abubakar Abid、Adam Fisch、Adam R. Brown、Adam Santoro、Aditya Gupta、Adrià Garriga-Alonso、Agnieszka Kluska、Aitor Lewkowycz、Akshat Agarwal、Alethea Power、Alex Ray、Alex Warstadt、Alexander W. Kocurek、Ali Safaya、Ali Tazarv、Alice Xiang、Alicia Par- rish、Allen Nie、Aman Hussain、Amanda Askell、Amanda Dsouza、Ameet Rahane、Anantharaman S. Iyer、Anders Andreassen、Andrea Santilli、Andreas Stuhlmüller、Andrew M. Dai、Andrew La、Andrew K. Lampinen、Andy Zou、Angela Jiang、Angelica Chen、Anh Vuong、Animesh Gupta、Anna Gottardi、Anto- nio Norelli、Anu Venkatesh、Arash Gholamidavoodi、Arfa Tabassum、Arul Menezes、Arun Kirubarajan、Asher Mullokandov、Ashish Sabharwal、Austin Her- rick、Avia Efrat、Aykut Erdem、Ayla Karakas 等。2022年。超越模仿游戏：量化并外推语言模型的能力。*CoRR*，abs/2206.04615。

HongjinSu、Jungo Kasai、Chen Henry Wu、WeijiaShi、Tianlu Wang、Jiayi Xin、Rui Zhang、Mari Osten- dorf、Luke Zettlemoyer、Noah A. Smith 和 Tao Yu。2022a。选择性标注让语言模型成为更好的少样本学习器。*CoRR*，abs/2209.01975。

Hongjin Su、Weijia Shi、Jungo Kasai、Yizhong Wang、Yushi Hu、Mari Ostendorf、Wen-tau Yih、Noah A. Smith、LukeZettlemoyer 和 Tao Yu。2022b。一篇嵌入器，任意任务：经过指令微调的文本嵌入。*CoRR*，abs/2212.09741。

Yi Tay、Mostafa Dehghani、Vinh Q. Tran、Xavier Garcia、Dara Bahri、Tal Schuster、Huaixiu StevenZheng、Neil Houlsby 和 Donald Metzler。2022年。统一语言学习范式。*CoRR*，abs/2205.05131。

Peter Todd、Jean Ortega、Jennifer Davis、Gerd Gigeren- zer、Daniel Goldstein、Adam Goodie、Ralph Hertwig、Ulrich Hoffrage、Kathryn Laskey、Laura Martignon 和 Geoffrey Miller。1999。《让我们变聪明的简单启发式》。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、BaptisteRozière、Naman Goyal、Eric Hambro、Faisal Azhar、Aurélien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lample。2023年。Llama：开放且高效的基础语言模型。*CoRR*，abs/2302.13971。

EndelTulving。2002。情景记忆：从心智到大脑。《心理学年度评论》，53：1–25。

Cunxiang Wang、Shuailong Liang、Yue Zhang、Xiaonan Li 和 Tian Gao。2019年。这有意义吗？为什么？关于意义建构与解释的初步研究。载于第*57*届计算语言学协会年会论文集，第4020–4026页，意大利佛罗伦萨。计算语言学协会。

Lei Wang、Wanyu Xu、Yihuai Lan、Zhiqiang Hu、Yunshi Lan、Roy Ka-WeiLee 和 Ee-Peng Lim。2023年。计划与求解提示：通过大语言模型改进零样本思维链推理。*CoRR*，abs/2305.04091。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc V. Le、Ed H. Chi 和 Denny Zhou。2022a。语言模型中的理由增强集成。*CoRR*，abs/2207.00747。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc V. Le、Ed H. Chi 和 Denny Zhou。2022b。自洽性提升语言模型中的思维链推理。*CoRR*，abs/2203.11171。

Yizhong Wang、Yeganeh Kordi、Swaroop Mishra、Al- isa Liu、Noah A. Smith、Daniel Khashabi 和 Han- naneh Hajishirzi。2022c。自我指令：利用自动生成的指令对齐语言模型。*CoRR*，abs/2212.10560。

Jason Wei、Maarten Bosma、Vincent Y. Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、An- drew M. Dai 和 Quoc V. Le。2022a。微调语言模型是零样本学习者。载于第十届国际学习表征会议，*ICLR 2022*，线上活动，*2022*年*4*月*25*日至*29*日。OpenReview.net。

Jason Wei、Yi Tay、Rishi Bommasani、Colin Raffel、Barret Zoph、Sebastian Borgeaud、Dani Yogatama、Maarten Bosma、DennyZhou、Donald Metzler、Ed H.

Chi、Tatsunori Hashimoto、Oriol Vinyals、Percy Liang、Jeff Dean 和 William Fedus。2022b。大语言模型的涌现能力。*CoRR*，abs/2206.07682。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、brian ichter、Fei Xia、Ed H. Chi、Quoc V Le 和 Denny Zhou。2022年c。思维链提示引导大语言模型进行推理。载于神经信息处理系统进展。

Yixuan Weng、Minjun Zhu、Shizhu He、Kang Liu 和 JunZhao。2022年。大语言模型是具备自我验证能力的推理者。*CoRR*，abs/2212.09561。

Ji Xin、Raphael Tang、Jaejun Lee、Yaoliang Yu 和 Jimmy Lin。2020年。DeeBERT：用于加速BERT推理的动态提前退出。载于 第*58*届计算语言学协会年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，第2246–2251页。计算语言学协会。

Jiacheng Ye、Jiahui Gao、Qintong Li、Hang Xu、Jiang- tao Feng、Zhiyong Wu、Tao Yu 和 Lingpeng Kong。2022a。ZeroGen：通过数据集生成实现高效的零样本学习。载于 *2022*年自然语言处理实证方法会议论文集（*EMNLP 2022*），阿布扎比，阿拉伯联合酋长国，*2022*年*12*月*7*日至*11*日，第 11653–11669 页。计算语言学协会。

Jiacheng Ye、Jiahui Gao、Zhiyong Wu、Jiangtao Feng、Tao Yu 和 Lingpeng Kong。2022b。Progen：通过上下文内反馈实现渐进式零样本数据集生成。载于计算语言学协会论文集：*EMNLP 2022* 研究成果，阿布扎比，阿拉伯联合酋长国，*2022*年*12*月*7*日至*11*日，第3671–3683页。计算语言学协会。

Jiacheng Ye、Zhiyong Wu、Jiangtao Feng、Tao Yu 和 Lingpeng Kong。2023年。用于上下文学习的组合式示例。*CoRR*，abs/2302.05698。

Kang MinYoo、Junyeob Kim、Hyuhng Joon Kim、Hyun- soo Cho、Hwiyeol Jo、Sang-Woo Lee、Sang-goo Lee 和 Taeuk Kim。2022年。真实标签至关重要：深入探究输入—标签示例。载于*2022*年自然语言处理实证方法会议论文集，*EMNLP 2022*，阿布扎比，阿拉伯联合酋长国，*2022*年*12*月*7*日至*11*日，第 2422–2437 页。计算语言学协会。

Eric Zelikman、Yuhuai Wu 和 Noah D. Goodman。2022年。Star：通过推理来引导推理。*CoRR*，abs/2203.14465。

Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona T. Diab、Xian Li、Xi Victoria Lin、Todor Mihaylov、Myle Ott、Sam Shleifer、Kurt Shuster、Daniel Simig、P unit Singh Koura、Anjali Sridhar、Tianlu Wang 和 Luke Zettlemoyer。2022a。OPT：开放式预训练 Transformer 语言模型。*CoRR*，abs/2205.01068。

Zhuosheng Zhang、Aston Zhang、Mu Li 和 Alex Smola。2022b。大语言模型中的自动思维链提示。*CoRR*，abs/2210.03493。

Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Be- ichen Zhang、Junjie Zhang、Zican Dong、Yifan Du、Chen Yang、Yushuo Chen、Zhipeng Chen、Jinhao Jiang、RuiyangRen、YifanLi、XinyuTang、Zikang Liu、Peiyu Liu、Jian-Yun Nie 和 Ji-Rong Wen。2023年。大语言模型综述。*CoRR*，abs/2303.18223。

Chuanyang Zheng、Zhengying Liu、Enze Xie、Zhenguo Li 和 Yu Li。2023年。渐进式提示改进大语言模型中的推理。*CoRR*，abs/2304.09797。

Denny Zhou、Nathanael Schärli、Le Hou、Jason Wei、Nathan Scales、Xuezhi Wang、Dale Schuurmans、Olivier Bousquet、Quoc Le 和 Ed H. Chi。2022年。由少到多提示使大语言模型能够进行复杂推理。*CoRR*，abs/2205.10625。

A LLM检索示例

我们在表6中展示 LLM 检索示例。

B 数据集详情

概述 我们在十个数据集上开展实验，涵盖四类任务：

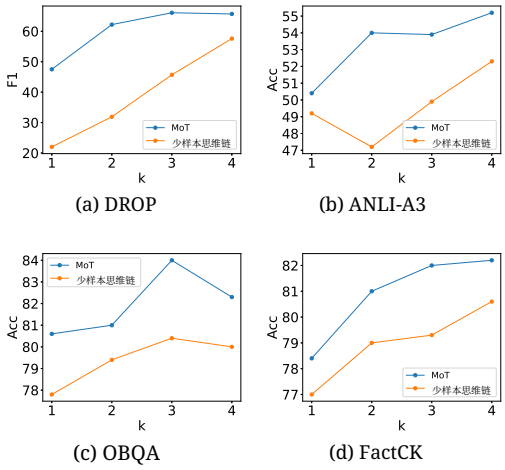
- **算术推理：** AQuA（Ling 等人，2017）：一个涵盖不同主题和难度级别的算术选择题数据集；以及 DROP（Dua 等人，2019年）：一个需要进行离散推理的阅读理解数据集；

- **自然语言推理：** 对抗性自然语言推理子集（Nie 等人，2020年），包括 ANLI-A1、ANLI-A2 和 ANLI-A3，分别涵盖不同的难度级别；

- **常识推理：** OBQA（OpenBookQA）（Mihaylov 等人，2018年）：与常识相关的问题，需要运用事实并将其应用于新情境；以及 ComV（常识验证）（Wang 等人，2019年）：一个要求从两句措辞相似的句子中识别出不合逻辑句子的数据集；

- **事实推理：** BoolQ（Clark 等人，2019年），FactCK（事实核查器）（Srivastava 等人，2022年）：一个用于测试评估事实性主张真实性能力的数据集，涵盖维基百科、COVID-19 和政治。WikiQA（Srivastava 等人，2022年）：基于随机抽取的 Wikidata 事实三元组进行的问题解答。

划分、评估指标与统计数据对于 AQuA、DROP、ANLI-A1、ANLI-A2、ANLI-A3、ComV 和 OBQA，我们使用其官方测试集进行评估。对于 BoolQ，我们遵循Wang 等人(2022a)使用验证集进行评估，因为其测试集尚未公开。对于 FactCK 和 WikiQA，我们手动将数据集划分为训练集和测试集，并将训练集中的问题作为无标签数据集，因为目前尚未发布这两个数据集的划分版本。受预算限制，对于 DROP 数据集，我们只使用其一半的无标签数据集（训练集中的问题）供大语言模型进行预先思考。对于分类或多项选择数据集，我们使用准确率作为评估指标。对于抽象式问答数据集包括 DROP 和 WikiQA，我们使用 F1 分数作为评估指标。对于 DROP，由于其一个测试样本有多个标注答案，我们遵循其原论文（Dua 等人，2019年），对所有标注答案取最大值。受预算限制，对于规模超过 1000 的评估数据集，我们随机抽取 1000 个样本进行评估。我们在表中列出了数据集的总体统计数据、过滤后记忆的规模以及评估指标7.



(a) DROP (b) ANLI-A3

(c) OBQA (d) FactCK

图7：示例数量的影响。

C 不同大语言模型上的性能

我们在 Text-Davinci-002 和 Text-Davinci-003 上进行了实验（Chen 等人，2023年；Ouyang 等人，2022年），以评估 MoT 在不同大语言模型上的通用性。结果见表5。我们观察到，MoT 在这两种大语言模型上始终优于基线方法，这表明 MoT 的效果并不依赖于某一种特定的大语言模型；未来随着更强大的大语言模型问世，MoT 还能带来进一步的提升。

D 示范数量的影响

我们比较了不同数量示例下 MoT 和少样本思维链的表现，结果如图 7所示。可以看出，在不同数量的示例下，MoT 的表现始终优于少样本思维链，这体现了 MoT 的稳定性。此外，结果表明，检索记忆中的示例比少样本思维链中的人工示例更有帮助、信息更丰富：具体而言，在 OBQA 和 DROP 上，MoT 仅用 1 或 2 个示例便能胜过使用 4 个示例的少样本思维链。

方法	ANLI-A3	OBQA	BoolQ	FactCK
Text-Davinci-002				
零样本思维链	34.2	47.2	40.8	52.2
少样本思维链	46.9	75.4	58.8	83.6
MoT	49.1	80.0	63.9	86.2
Text-Davinci-003				
零样本思维链	46.9	64.2	62.8	47.2
少样本思维链	45.3	81.6	67.7	84.0
MoT	49.3	84.6	71.0	87.0

表5：不同大语言模型上的性能比较。

E 实现细节

除非另有说明，我们使用 OpenAI 的公开语言模型“gpt- 3.5-turbo-0301”，而在“text-davinci-002/003”上进行的实验（附录C）也显示出一致的改进。由于大语言模型 API 预算有限，我们在预思考阶段启发式地设置超参数，包括生成温度 T、解码推理路径的数量 n。对于过滤阈值 τ 和记忆簇的数量 l，我们在 OBQA 上进行了探索性实验，发现 $\tau = \{0.2, 0.3, 0.4\}$ 和 $l = \{3, 4, 5\}$ 能够取得相似的性能。因此，我们分别设置 $\tau = 0.3$ 和$l = 4$。对于预先思考，我们使用温度 $T = 1.2$ 来鼓励生成更多样化的推理路径，并使用 $n = 16$次推理路径采样，除非另有说明。对于记忆召回，我们使用 SBERT（“all-mpnet-base- v2”）（Reimers 和 Gurevych，2019年）进行语义过滤。受大语言模型的最大输入长度限制，我们将每个数据集中每个簇的记忆候选数量固定为 10。在测试阶段，为保证结果的稳定性，除非另有说明，我们使用贪心解码生成输出。为简化实验，我们分别在每个数据集上运行 MoT，并将跨数据集记忆召回留作未来工作。基线方法的结果来自我们的实现，并且与 MoT 使用相同的模板、答案解析和评估方法。

对于 AQuA、OpenBookQA、BoolQ、DROP、ANLI-A1、ANLI-A2 和 ANLI-A3，我们分别使用与 Wei 等人（2022c）中相同的少样本思维链示例、Zhou 等人（2022年）中相同的少样本思维链示例、Wang 等人（2022b）中相同的少样本思维链示例以及 Wang 等人（2022a）中相同的少样本思维链示例。对于其余没有公开发布人工思维链示例的数据集，我们从训练集中随机选择问题，并使用 ChatGPT 生成推理路径，从而获得相应的少样本思维链示例。我们在表-格8、9、10、12、11、13、14 和 15。

F 伦理声明

本文首次尝试让大语言模型基于记忆机制实现自我改进。所进行的实验仍处于安全环境中，即使用特定的无标签数据集，且大语言模型无法访问互联网或控制外部工具。因此，我们认为我们的方法和实验仍足够安全，不会对社会造成严重影响或不可挽回的后果。

大语言模型输入

我将我需提供一个目标问题和选择一个参考问题，其问题、思路或答案对你回答目标问题最有帮助。请注意，以下参考问答对是随机排列的，没有任何优先顺序。

目标问题：

机器 A 每 6 分钟产出一个悠悠球

它们需要多少分钟才能生产出 10 个悠悠球？答案选项：(A) 24 分钟 (B) 32 分钟 (C) 36 分钟 (D) 64 分钟 (E) 72 分钟

参考问题：1. 问题：两台机器 Y 和 Z 以恒定速率工作，生产相同物品。机器 Y 在机器 Z 生产 2 件物品所用的同一时间内生产 5 件物品。如果机器 Y 需要 9 分钟生产一批物品，那么机器 Z 生产同样数量的物品需要多少分钟？答案选项：(A) 6 (B) 9 (C)9 1/2 (D) 22 2 2 5 13 1. (E) / . 问题：两台机器 Y 和 Z 以恒定速率工作，生产相同物品。机器 Y 在机器 Z 生产 38 件物品所用的同一时间内生产 30 件物品。如果机器 Y 需要 19 分钟生产一批物品，那么机器 Z 生产同样数量的物品需要多少分钟？答案选项：(A) 6 (B) 9 (C) 9 1 2 3 2 15 13 1 / (D) (E) / .

问：两台机器 Y 和 Z 以恒定速率工作，生产相同物品。机器 Y 在机器 Z 生产 24 件物品的同一时间内生产 30 件物品。如果机器 Y 生产一批物品需要 36 分钟，那么机器 Z 生产同样数量的物品需要多少分钟？答案选项：(A) 60 (B) 90 (C) 9 1/2 (D) 45 (E) 13 1/2

4.

问：机器 A 单独以恒定速率工作时，10 分钟可生产 x 个箱子；机器 B 单独以恒定速率工作时，5 分钟可生产 2x 个箱子。机器 A 和机器 B 按各自的恒定速率同时工作，生产 10x 个箱子需要多少分钟？答案选项：(A) 13 分钟 (B) 14 分钟 (C) 15 分钟 (D) 16 分钟 (E) 20 分钟

5.

问：两台机器 Y 和 Z 以恒定速率工作，生产相同物品。机器 Y 在机器 Z 生产 21 件物品的同一时间内生产 23 件物品。如果机器 Y 生产一批物品需要 21 分钟，那么机器 Z 生产相同数量的物品需要多少分钟？答案选项：(A) 6 (B) 9 (C) 9 1/2 (D) 12 (E) 23

6.

问：机器 X 和机器 Y 分别以恒定速率生产瓶子。机器 X 在 6 小时内生产 k 个瓶子，机器 Y 在 12 小时内生产 k 个瓶子。机器 X 和机器 Y 同时工作，生产 12k 个瓶子需要多少小时？答案选项：(A) 8 (B) 12 (C) 15 (D) 48 (E) 24 7。

问：机器 X 和机器 Y 分别以恒定速率生产瓶子。机器 X 在 6 小时内生产 k 个瓶子，机器 Y 在 3 小时内生产 k 个瓶子。机器 X 和机器 Y 同时工作，生产 12k 个瓶子需要多少小时？答案选项：(A) 4 (B) 8 (C) 12 (D) 18 (E) 4 8。

问：机器 X 和机器 Y 分别以各自的恒定速率生产瓶子。机器 X 在 4 小时内生产 k 个瓶子，机器 Y 在 5 小时内生产 k 个瓶子。机器 X 和机器 Y 同时工作，需要多少小时才能生产 10 9 9 8 2 3 125 3 15 18 22 2 k 个瓶子？答案选项：(A) / (B) / (C) (D) (E) / 。

问：机器 A 以恒定速率单独工作，10 分钟生产 x 个箱子；机器 B 以恒定速率单独工作，5 分钟生产 2x 个箱子。如果机器 A 和机器 B 以各自的恒定速率同时工作，生产 6x 个箱子需要多少分钟？答案选项：(A) 3 分钟(B) 4 分钟(C) 5 分钟(D) 6 分钟(E) 12 分钟

10.

问：机器 A 每小时可以生产 350 个小部件，机器 B 每小时可以生产 250 个小部件。如果两台机器一起工作，要生产总共 900 个小部件需要多少时间？答案选项：(A) 1小时20分钟 (B) 1小时24分钟 (C) 1小时30分钟 (D) 1小时36分钟 (E) 1小时40分钟

以上哪个参考问题最有助于你回答目标问题？你必须且只能选择一个参考问题来回答目标问题。你的回复必须以以下格式结尾：“最有帮助的问题是问题 [编号]。”例如，如果你的答案是问题5，就必须以“最有帮助的问题是问题5。”结尾。

大语言模型输出

最有帮助的问题是问题10。

表6：LLM检索示例。

任务类别	Task	任务格式	无标签问题	测试问题	内存大小	指标
算术推理	AQuA	多项选择	97467	254	19334	准确率
	DROP	抽象式问答	42777	1000	17066	F1
NLI	ANLI-A1	分类	16946	1000	9721	准确率
	ANLI-A2	分类	45460	1000	25444	准确率
	ANLI-A3	分类	100459	1000	53402	准确率
常识推理	ComV	多项选择	10000	1000	9003	准确率
	OBQA	多项选择	4957	1000	3559	准确率
事实性推理	BoolQ	分类	9427	1000	6422	准确率
	FactCK	分类	6654	500	4782	准确率
	WikiQA	抽象式问答	19442	1000	9041	F1

表 7：所使用数据集及其评估指标的统计信息。

问： 约翰发现 15 个数的平均值是 40。如果给每个数都加上 10，那么这些数的均值是多少？答案选项：（A）50 （B）45 （C）65 （D）78 （E）64**答：** 如果给每个数都加上 10，那么这些数的均值也会增加 10。因此新的均值是 50。答案是（A）。**问：** 如果 a / b = 3/4 且 8a + 5b = 22，那么求 a 的值。答案选项：（A）1/2 （B）3/2 （C）5/2 （D）4/2 （E）7/2**答：** 如果 a / b = 3 4, = 4 3 8 + 5 4 3 = 22 8 + 20 3 = 22 44 3 /, = 22 那么 b a /。因此 a (a /)。这化简为 a a /，这意味着 a /。所以 a 等于 /。答案是（B）。**问：** 一个人以 20 千米/小时的速度行驶，并在 2.5 小时后到达目的地，那么路程是多少？答案选项：（A）53 千米 （B）55 千米 （C）52 千米 （D）60 千米 （E）50 千米**答：** 此人行驶的路程应为 20 千米/小时 * 2.5 小时 = 50 千米。答案是（E）。**问：** 输入从 1 到 500 的数字需要多少次击键？答案选项：（A）1156 （B）1392 （C）1480 （D）1562 （E）1788**答：** 从 1 到 9 有 9 个一位数。从 10 到 99 有 90 个两位数。从 100 到 500 有 401 个三位数。 9 + 90(2) + 401(3) = 1392。答案是（B）。

3 2

表 8：AQuA 的少样本思维链提示，摘自Wang 等人（2022b）。

Q: 海鹰队对阵旧金山49人队。第一节，海鹰队跑卫朱利叶斯·琼斯完成了一次27码的达阵冲球，防守截锋克雷格·特里尔还将一次掉球回攻9码达阵。第三节，49人队差点扳回局势：跑卫H. J.托雷斯向卢卡斯·内利送出一记12码达阵传球，马雷还踢进了一记32码任意球。最后一节，朱利叶斯·琼斯又完成了一次11码达阵。最短的达阵冲球和最长的达阵传球合计多少码?

A: 所有达阵跑球分别为：27码达阵跑球、9码达阵跑球、11码达阵跑球。27、9、11中最小的数字是9。因此，最短的达阵跑球距离为9码。所有达阵传球为：一次12码达阵传球。因此，最长的达阵传球距离为12码。所以，最短的达阵跑球距离与最长的达阵传球距离相加为 9 + 12 = 21 码。因此，答案是21码。

问题： 该县的人口分布如下：18岁以下占23.50%，18至24岁占8.70%，25至44岁占29.70%，45至64岁占24.70%，65岁及以上占13.30%。18岁以下人口所占比例比18至24岁人口高出多少个百分点?

答案： 根据这段文字，18岁以下人口占23.5%，18至24岁人口占8.7%。23.5% - 8.7% = 14 8 14 8 %。因此答案是.....

问题： 自20世纪70年代以来，美国政府一直在协商管理贸易协定，例如20世纪90年代的《北美自由贸易协定》、2006年的《多米尼加共和国—中美洲自由贸易协定》，以及若干双边协定。在欧洲，六个国家于1951年成立了欧洲煤钢共同体，该共同体于1958年发展为欧洲经济共同体。欧洲经济共同体的两个核心目标是发展共同市场（后来更名为单一市场），以及在成员国之间建立关税同盟。欧洲煤钢共同体存在了多少年?

A: 根据这段文字，欧洲煤钢共同体于1951年成立，并于1958年成为欧洲经济共同体。1958 - 1951 = 7。所以答案是7。

问题： 维京人队飞往美国银行体育场迎战卡罗莱纳黑豹队。第一节双方均未得分后，卡罗莱纳队凭借四分卫马特·摩尔的一记1码达阵传球，由全卫布拉德·胡佛接球，率先得分。又一个未得分的一节过后，马特·摩尔完成一记42码达阵传球，传给外接手史蒂夫·史密斯，卡罗莱纳队就此锁定胜局。双方有几节均未得分?

A: 第一节和第三节是未得分的一节。因此有2个未得分的一节。所以答案是 2.

表 9：用于 DROP 的少样本思维链示例，来自Zhou 等人（2022年）。

前提：“从概念上讲，撇脂竞争有两个基本维度——产品和地理。”根据这一前提，我们能否得出假设“产品和地理是撇脂竞争发挥作用的关键因素”这一结论？选项：- 是 - 否 - 无法判断**A:** 根据“撇脂竞争有两个基本维度”这一说法，我们无法推断这两个维度就是撇脂竞争发挥作用的原因。答案是无法判断。前提：“我们的一位成员将仔细执行你的指示。”根据这一前提，我们能否得出假设“我的团队成员将以极高的精确度执行你的命令”这一结论？选项：- 是 - 否 - 无法判断**A:** “one of”与“a member of”（一名成员）的意思相同，“carry out”与“execute”（执行）的意思相同，而“minutely”与“immense precision”（极高的精确度）的意思相同。答案是是。前提：“适合成年人和儿童的乐趣。”根据这一前提，我们能否得出假设“只适合儿童的乐趣”这一结论？选项：- 是 - 否 - 无法判断**A:**“成年人和儿童”与“只有儿童”相矛盾。答案是否。前提：“他转过身，对 Vrenna 微笑。”根据这一前提，我们能否得出假设“他对 Vrenna 微笑，而 Vrenna 正和她的母亲一起在他身后慢慢行走”这一结论？选项：- 是 - 否 - 无法判断 **A:** 前提中没有提及 “Vrenna 正在行走 ”。答案是无法判断。

表 10：用于自然语言推理任务的少样本思维链提示，来自 Wang 等人 （2022a） .

问： 毒药会伤害以下哪一种？ (A) 一棵树 (B) 一个机器人 (C) 一栋房子 (D) 一辆汽车**答：** 毒药会伤害生物，只有树是生物。答案是 (A)。**问：** 当你仔细观察一块 Marbel 时，你能看到 (A) 未来 (B) 微小的缺陷 (C) 颜色 (D) 另一面**答：** Marbel 不透明，所以你看不到另一面。Marbel 不一定有多种颜色。你会看到微小的缺陷。答案是 (B)。**问：** 食物在胃里被分解时，(A) 大脑需要时间来消化 (B) 花点时间消化我说的话 (C) 营养物质正在被分解 (D)《读者文摘》是一部作品集**答：** 在消化过程中，食物在胃里被分解。答案是 (C)。**问：** 太阳与以下哪项有关？ (A) 小狗学会新把戏 (B) 孩子长大成人并变老 (C) 花瓶里的花枯萎 (D) 植物发芽、开花和枯萎**答：** 太阳会影响生物的生长，比如植物。答案是 (D)。

表 11：OBQA 的少样本思维链示例，摘自Wang 等人（2022a）。

问：以下哪项陈述违背常识？（A）玫瑰花蕾吃毛毛虫（B）毛毛虫吃玫瑰花蕾**答：** 陈述（A）违背常识，因为它违反了自然食物链以及人们已知的玫瑰行为。玫瑰是植物，不能吃或吞食其他生物，包括毛毛虫。答案是 A。**问：** 以下哪项陈述违背常识？（A）他把房子扔进了垃圾桶（B）他把厨余垃圾扔进了垃圾桶**答：** 把房子扔进垃圾桶在物理上是不可能的。陈述（A）违背物理定律，因此不合逻辑。答案是 A。**问：** 以下哪项陈述违背常识？（A）因为他的汽车受损，他从电力公司获得了 1000 元人民币（B）因为他的汽车受损，他从保险公司获得了 1000 元人民币**答：** 电力公司赔偿某人的汽车损失在逻辑上说不通。由保险公司为汽车损失提供赔偿更为合理。答案是 A。**问：** 以下哪项陈述违背常识？（A）因为他的汽车受损，他从电力公司获得了 1000 元人民币（B）因为他的汽车受损，他从保险公司获得了 1000 元人民币**答：** 电力公司赔偿某人的汽车损失在逻辑上说不通。由保险公司为汽车损失提供赔偿更为合理。答案是 A。

表 12：ComV 的少样本思维链示例。

问：系统崩溃乐队有两位主唱吗？
答： 系统崩溃乐队目前由 Serj Tankian、Daron Malakian、Shavo Odadjian 和 John Dolmayan 组成。Serj 和 Daron 负责演唱，因此乐队确实有两位主唱。答案是肯定的。

问：伊朗和阿富汗说的是同一种语言吗？
答：伊朗和阿富汗都说印欧语系语言波斯语。答案是肯定的。

Q:大提琴和低音提琴是同一种乐器吗？
A: 演奏大提琴时需要坐着，将乐器置于双膝之间，而演奏低音提琴时则需要站着或坐在凳子上。答案是否定的。
Q:可以在埃普瑟姆火车站使用牡蛎卡吗？
A: 埃普瑟姆火车站服务于萨里郡的埃普瑟姆镇，并不在伦敦牡蛎卡区域内。
答案是否定的。

表 13：BoolQ 的少样本思维链示例，摘自Wang 等人（2022a）。

问： 2017年6月，有人提出了以下说法：大卫·劳合·乔治竞选首相屡次失利。这个说法正确还是错误？**答：** 大卫·劳合·乔治于1916年至1922年担任英国首相。在此之前，他还曾担任财政大臣和军需大臣。因此，说他竞选首相屡次失利是错误的。答案是错误。**问：** 2017年6月，有人提出了以下说法：乔治·哈里森于1966年第一次结婚。这个说法正确还是错误？**答：**乔治·哈里森于1966年1月21日与第一任妻子、模特帕蒂·博伊德结婚。答案是正确。**问：** 2017年6月，有人提出了以下说法：伍德罗·威尔逊没有经历过第一次世界大战。这个说法正确还是错误？**答：** 第一次世界大战期间，伍德罗·威尔逊担任美国总统，任期为1913年至1921年。答案是错误。**问：** 2008年4月17日，有人提出了以下说法：希拉里·克林顿从游说者那里获得了超过80万美元。这个说法正确还是错误？**答：** 根据 OpenSecrets.org（一家追踪政治资金的无党派研究机构）的数据，希拉里·克林顿在2008年总统竞选期间从游说者那里获得了超过80万美元的竞选捐款。答案是正确。

表 14：事实核查器的少样本思维链示例。

问： 2017年6月，有人提出了以下说法：大卫·劳合·乔治竞选英国首相屡屡失利。这个说法是真是假？**答：** 大卫·劳合·乔治于1916年至1922年担任英国首相。在成为首相之前，他还担任过财政大臣和军需大臣。因此，说他竞选英国首相屡屡失利是错误的。答案是错误。**问：** Aaron Swartz 的母语是什么？**答：** Aaron Swartz 出生于美国伊利诺伊州芝加哥。因此，他的母语很可能是英语，因为英语是美国的主要通用语言。答案是英语。**问：** 巴差提朴信奉什么宗教？**答：** 巴差提朴是一位佛教徒，因为佛教是泰国的主要宗教；在泰国成为君主立宪制国家之前，他是该国最后一位绝对君主。答案是佛教。**问：** 瓦莱塔位于哪个国家？**答：** 瓦莱塔是马耳他的首都。马耳他是位于地中海的一座小岛国。答案是马耳他。**问：** 加里·卡斯帕罗夫从事什么运动？**答：** 加里·卡斯帕罗夫曾是世界国际象棋冠军，因此他从事的运动是国际象棋。答案是国际象棋。

表 15：WikiQA 的少样本思维链示例。