# L*AR*GE L*AN*GU*AG*E MODELS C*AN* SELF-IMPROVE

# 大语言模型可以自我改进

**太长不看版：**

大型语言模型（LLM）在知识蒸馏和自我改进技术的助力下，实现了性能的大幅提升。蒸馏技术使得小型模型能够模拟大型模型的推理能力，甚至在某些情况下能够超越大型模型的表现，这在资源受限的环境中尤为重要。同时，自我改进方法使LLM能够利用自我生成的标签提升推理准确性，无需依赖真实标签，进一步展现了其强大的泛化能力。研究还发现，经过自我改进后，模型在采样温度选择上有所调整，以适应更确定的输出分布，这为提高模型的随机性和泛化能力提供了新的思路。这些改进不仅提高了模型在推理任务上的表现，还为未来的模型优化提供了新的方向，如结合生成数据和监督数据以进一步提升性能。总之，LLM在蒸馏和自我改进技术的推动下，为自然语言处理领域的发展注入了新的活力，开启了新的可能性。

简单来说，大型语言模型（LLM）在变得更聪明和更懂事的帮助下，变得更厉害了。变得更聪明的方法让小型模型也能像大型模型一样聪明，甚至在某些时候比大型模型还厉害，这对于资源有限的环境很重要。而变得更懂事的方法让LLM能够通过自己生成的问题和答案来提高推理的准确性，而不需要真实的答案。研究还发现，经过自我改进后，模型在选择如何随机回答时也会调整，以更好地适应答案的不确定性。这些改进让模型在推理任务上做得更好，也为未来让模型更厉害提供了新的思路，比如结合生成的数据和真实的数据来进一步提升性能。同时针对未标记的问题。用到了链式思考（*Chain-of-Thought*）提示和自我一致性（*self-consistency*）的技巧。

**摘要：**

本文描述了一项关于大型语言模型（LLMs）的研究。大型语言模型在各种任务中表现出色，但对其进行微调通常需要大量的监督数据。然而，人类可以通过自我思考，不依赖外部输入来提高推理能力。这项工作的主要目的是展示大型语言模型也能够仅使用未标记的数据集进行自我改进。研究者们使用预训练的大型语言模型，通过链式思考和自我一致性（*Chain-of-Thought prompting and self-consistency*）为未标记的问题生成“高置信度”的带理由的答案。然后，他们使用这些自我生成的答案作为目标输出，对LLM进行微调。这种方法提高了一个540B参数的大型语言模型的通用推理能力，例如在GSM8K上的准确率从74.4%提高到了82.1%，在DROP上从78.2%提高到了83.0%，在*OpenBookQA*上从90.0%提高到了94.4%，在*AN*LI-A3上从63.4%提高到了67.9%。所有这些提高都是在没有任何真实标签的情况下实现的。此外，研究者们还进行了消融研究（*ablation studies*），结果显示在推理上的微调对于自我改进至关重要。这项研究的结果表明，大型语言模型具有自我改进的能力，这可能会推动该领域的研究进一步发展。

1. **引言**

与人类可以通过自我思考来提高推理能力而无需外部输入不同，LLMs通常需要大量的监督数据进行微调。文章中在这里提出一个问题：LLMs是否也可以像人类一样，在没有外部输入的情况下进行自我改进？

为了回答这个问题，作者们提出了一种方法，使用预训练的LLM来生成“高置信度”的、带有推理依据的答案，针对未标记的问题。这里用到了链式思考（*Chain-of-Thought*）提示和自我一致性（*self-consistency*）的技巧。

使用这些自我生成的答案作为目标输出，对LLM进行微调。这意味着LLM是在学习如何从其自身的输出中学习，从而改进其推理能力。

*Scaling*已经使大型语言模型（LLM）能够在一系列自然语言处理（NLP）任务上实现最先进的性能 。更重要的是，当LLM被扩展到数千亿参数 时，出现了新的能力：环境中少样本学习使LLM有可能 在一个没有通过少数例子训练过的任务中表现良好；思维链（CoT）提示展示了LLM在不同任务中的推理能力；自我一致性，通过自我评估多种推理路径进一步提高了性能。

尽管在大型文本语料库上训练的模型具有令人难以置信的能力，从根本上提高模型性能超过少量基线，仍然需要对大量高质量监督数据集进行微调。 部分研究策划了数十个基准NLP数据集，以提高未知任务的零样本任务性能；部分研究收集了不同文本指令集的许多人类答案，以便更好地将其模型与人类指令对齐。虽然在收集高质量的监督数据集方 面付出了大量的努力，但相反，人类的大脑能够进行元认知过程，在那里模型可以在没有外部输入的情况下完善模型自己的推理能力。

在本文中，作者研究了LLM如何在没有监督数据的情况下自我提高其推理能力。研究表明，仅使用来自多个NLP任务数据集的输入序列（没有地面真实输出序列），一个预先训练的LLM能够提高域内和域外任务的性能。作者的方法如图1所示：作者首先使用少样本思维链（CoT）作为提示对多个预测进行抽样，使用多数投票过滤“高可信度”预测，最后对这些高可信度预测的LLM进行微调。结果模型在贪心和多路径评估中都显示了改进的推理能力。作者将以这种方式进行微调的模型称为语言模型自改进。

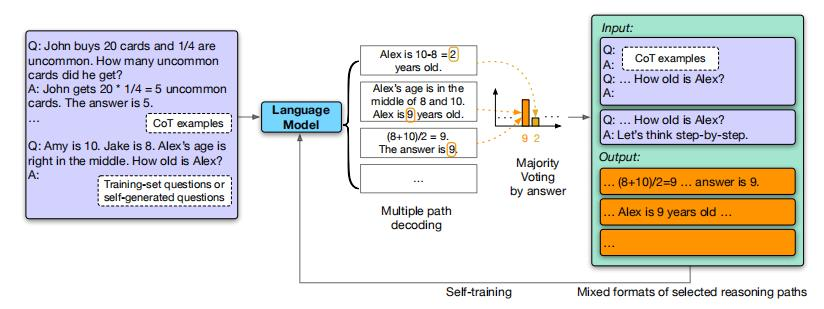


图1：作者的方法的概述。以思维链（CoT）示例作为演示，语言模型为每个问题生成多个CoT推理路径和答案（温度T > 0）。最一致的答案是通过多数投票选择的。导致大多数答案的“高可信度”CoT推理路径被混合格式增强，作为最终的训练样本，并反馈给模型进行微调。

语言模型自改进*(LMSI*). 这与人类大脑有时会学习的方式类似：给定一个问题，多次思考，得出不同的可能结果，得出应该如何解决这个问题的结论，然后从中学习或记忆自己的解决方案。作者使 用预先训练过的P*aL*M-540B LLM对作者的方法进行了实证验证，其中作者的方法不仅提高了训练任务的性能 (74.4%→82.1%on GSM8K, 78.2%→83.0% on DROP, 90.0%→94.4% on OpenBookQA, *an*d 63.4%→67.9% on *AN*LI-A3)，但也增强了域外（OOD）测试任务（*AQ*UA，战略QA，MNLI），在许多任务中实 现了最先进的性能，而不依赖于监督的真值答案。最后，作者对自我生成的额外输入问题和少样本CoT提示进行了初步研究，这可以进一步减少模型自我改进所需的人力，并对作者方法的重要超参数进行消融研究。

1. **方法**

方法的概述如图1所示: 作者给出了一个预先训练的大语言模型（LLM）*M*和一个只有问题的训练数据集*Dtrain = {xi}Di=1*与少数镜头的思想链（CoT）的例子。作者采用采样温度T > 0的多路径解码来生成m个推理路径和答案*{ri1, ri2, ..., rim}*对于每个问题*xiin Dtrain*，并使用多数投票（自我一致性）来选择最一致、最高可信度的答案。然后，作者保留所有导致最一致答案的推理路径，应用混合格式的提示和答案进行增强，并对这些自生成的推理-答案数据的模型进行微调

* 1. **生成和过滤多个推理路径**

自我一致性对推理任务带来了巨大的改进 (e.g. 56.5% →74.4% on GSM8K test set)与多样化解码之间的差距表明，利用自选择的高置信度推理路径作为训练数据，有进一步提高*M*的推理能力的潜力。

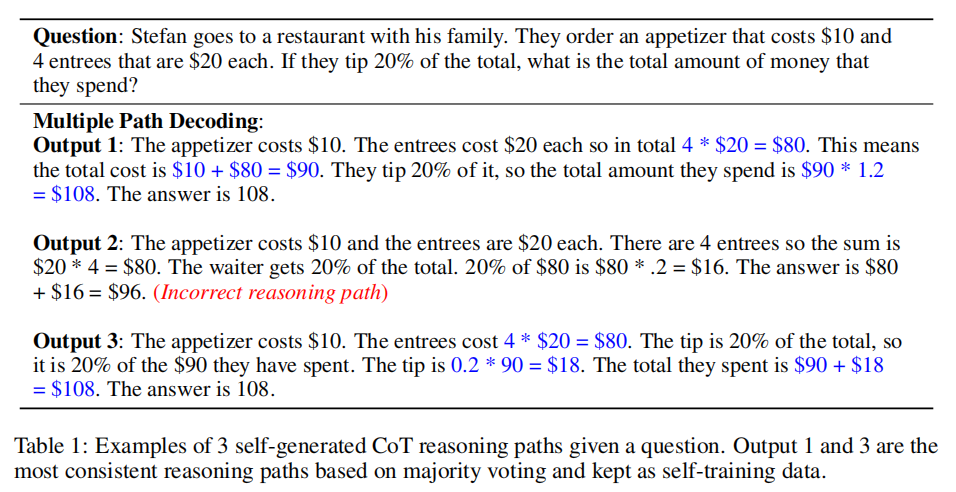


表1：给出一个问题的3个自生成的CoT推理路径的例子。输出1和输出3是基于多数投票的最一致的推理路径，并作为自我训练数据保存。

对于每个训练问题xi，作者采样m个CoT推理路径，记为{ri1, ri2, . . . , rim}（示例见表1）。作者应用相同的输出解析与“答案是”来生成他们的预测*an*swers {yi1; yi2; ...; yim}.最一致的答案，不一定是正确的答案，由多数投票选择，记为

对于所有的训练问题，作者过滤到达y˜的CoT推理路径，作为最终放入自我训练数据的答案，记Dself−consistent = {xi , r˜i}，其中r˜i = {rij |1 ≤ j ≤ m, yij= ˜yi} 因为作者不使用任何标签来过滤掉y˜i ≠ yi，重要的是，自生成的CoT推理路径大多是可靠的，而不正确的答案不会损害模型的自我改进。作者在图2中绘制了GSM8K训练集中每个问题的自生成CoT路径的准确性和置信度之间的关系。置信度是导致y˜的CoT路径数除以总路径数m。y轴表示y˜在一定置信度下的精度。圆圈区域和色暗表示在一定心下的问题数量。作者可以观察到，置信度高的答案更有可能是正确的，这意味着当一个问题有许多一致的CoT路径时，那么相应的y˜更有可能是正确的。另一方面，当y˜错误时，可能有较少的CoT路径支持，对训练样本的噪声较小。

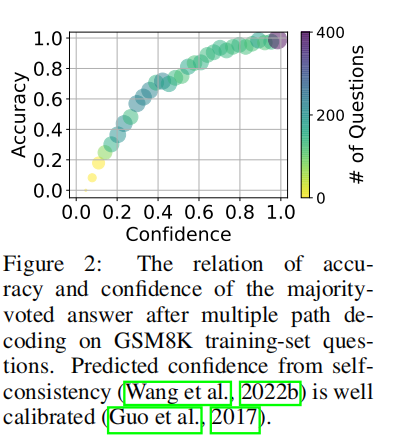


图2：GSM8K训练集问题在多路径解码 后的多数投票答案的准确性和置信度的关系。来自自我一致性的预测置信度得到了很好的校准。

* 1. **采用混合形式的训练**

为了防止语言模型对特定的提示或回答风格进行过拟合，作者为每种推理路径创建了四种不同的格式，并混合在自我训练数据中，如表2所示。在第一种格式中，在新问题之前准备几个思维链例子（问题和推理路径导致最终正确的答案），而语言模型输出被训练为与经过过滤的CoT推理路径相同。在第二种格式中，作者使用问题的例子及其直接答案作为标准提示，并且语言模型的输出也应该只包含直接答案。第三种和第四种格式类似于第一种和第二种格式，除了没有给出问答对的例子，因此模型将学习以上下文零射击的方式自己思考。在第三种格式中，作者希望模型输出CoT推理，而不附加包含CoT推理的示例，作者附加了“让模型一步一步地思考”。在输入序列的最后，指导语言模型逐步生成*CoT*推理路径。然后使 用训练样本的混合格式对预先训练好的语言模型*M*进行微调。

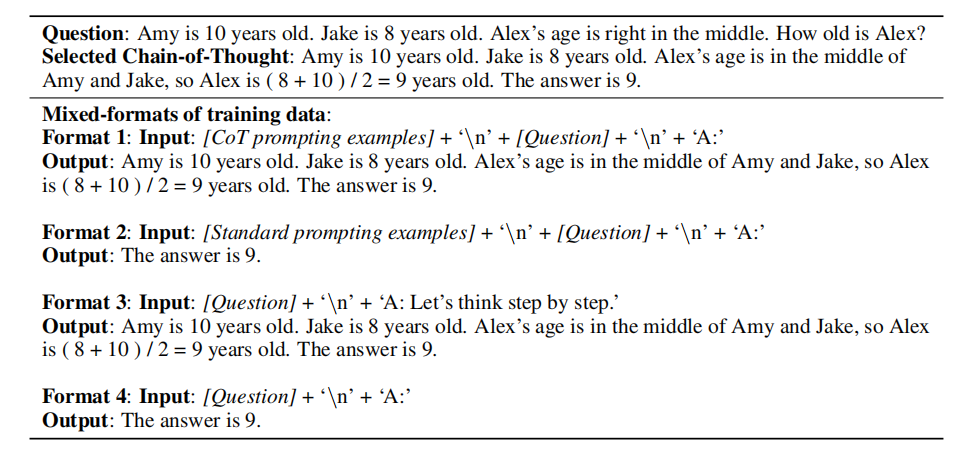


表2：一个如何将推理路径扩展为四种格式的训练数据的例子，这些数据具有不同的提示（输入）和答案风格（输出）。具体而言，附录A.2列出了每项任务使用的CoT提示示例。标准提示示例与CoT提示实例是相同的问答对，只是删除了推理。

* 1. **生成问题和提示**

给定一组训练问题和一些人工编写的思想链（CoT）示例作为提示，作者提出的方法可以实现模型的自我完善。然而，当训练问题或CoT示例的数量有限时，作者的方法可能无法生成足够的训练样本用于语言模型自训练。从网上收集问题需要人力资源。为了进一步减少人力资源，作者研究了如何自行生成更多的培训问题以及示例提示。

问题生成（*Question Generation*）先前的工作讨论了通过使用LLM生成不同的训练样本来增强少样本数据。然而，这些方法是为分类任务而设计的，并且需要为每个少样本中的例子人工标注。作者使用一种简单而有效的方法来生成域内问题的各种问题（没有基本事实答案）。具体来说，作者随机选择几个现有的问题，以随机顺序将它们连接起来作为输入提示，并让语言模型生成连续的序列作为新问题。作者重复这个过程以获得一大组新问题，然后使用自洽性只保留具有高度自信答案的问题。然后将这些问题用作自己生成的训练问题。

提示生成（*Prompt Generation* ）给定一组问题，人类可以编写CoT示例作为推理路径，从而得出最终答案。在没有手动提示的零样本设置中，作者可以使用模型本身生成这些CoT路径。，作者以“A：让模型一步一步地思考”作为答案，并让语言模型生成连续的推理路径。然后，作者使用这些生成的推理路径作为少镜头CoT提示的示例。

1. **实验**

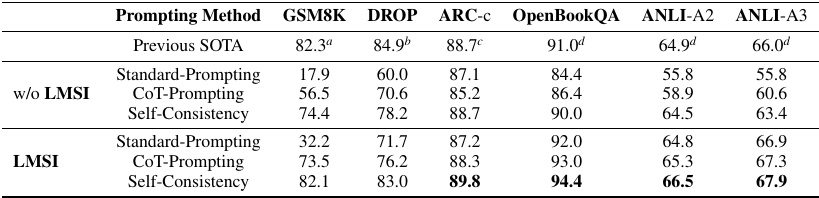
**实验设置：**实验所用的任务及数据集：作者展示了他们的方法在在三种类型任务（算数推理、常识推理、自然语言推理）上的有效性。算数推理包括数学问题集GSM8K与要求数值推理的阅读理解基准DROP常识推理包含OpenBookQA数据集和*AI*2推理挑战赛数据集；自然语言推理使用了对抗性NLI（*AN*LI）子集，*AN*LI-A2和*AN*LI-A3。实验的模型、训练设置和超参数：作者基于前人（的研究，选择了一个基于自回归Tr*an*sformer的语言模型上进行实验。

**实验目标：**证明本文提出的自我提升方法的有效性

**实验过程：**首先，作者在每个单独的数据集（任务）上应用实验方法并报告结果。然后，作者合并来自所有数据集的生成数据，并训练一个模型来研究该模型在未见数据集上的泛化能力，如同（Wei等人，2021）中所做的那样。除了使用生成的CoT推理路径的结果外，他们还展示了关于生成输入问题和少样本提示的研究。最终进行对模型大小和超参数的消融研究。

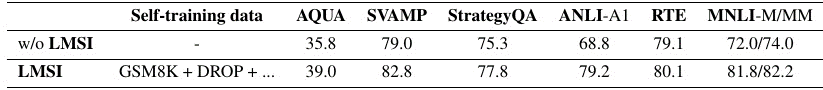
1. **结果**
   1. **主要结果**

作者在下表中列出了使用P*a*LM-540B模型在LMSI前后的结果。对于每个模型，在测试时，作者在所有六个数据集上应用了三种不同的提示方法：标准提示、CoT提示和自我一致性提示。下表中还有作者的结果与之前的不同数据集上通过不同方法取得的SOTA的比较。

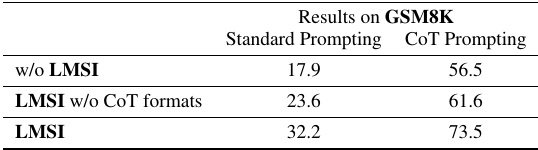


可以观察到经过LMSI之后，所有三种提示方法的性能都大幅提升。比较自我一致性与LMSI配合自我一致性，作者观察到显著改进（如上表）。作者认为这表明他们提出的方法相当有效。此外，LMSI的单路径CoT提示性能接近甚至优于未经LMSI的模型的多路径自我一致性性能，表明LMSI确实帮助语言模型从多个一致的推理路径中学习。在GSM8K数据集上，LMSI接近于DiVeRSe方法，该方法使用多样化提示和投票验证器来集成100条输出路径。相反，作者只使用32条输出路径来自我生成训练样本，并与LMSI自我一致。在DROP数据集上，LMSI接近于OPERA方法，该方法使用标签进行训练。另一方面，作者的方法仅利用训练集中的问题，不使用标签。

**针对未知任务进行多任务自我训练：**为了证明LMSI的泛化能力，作者进行了自监督训练实验，使用了上述六个数据集的训练集问题混合体（称为领域内任务），然后使用相同的模型检查点在六个领域外（OOD）任务上进行评估。在这些任务中，*AQ*UA、Str*at*egyQA和RTE与任何领域内任务都显著不同。这三项任务都有自己的少样本提示。从上表中，作者认为观察到LMSI在所有OOD任务上都取得了更高的准确率，表明语言模型的整体推理能力得到了提升。



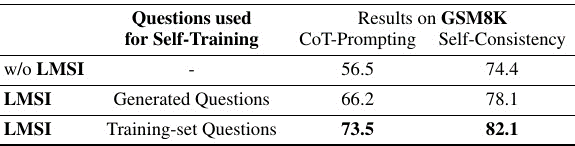
**用思维链（Ch*ai*n-of-Thought）格式训练的重要性：**与仅使用直接答案训练相比，作者证明了使用思维链格式训练语言模型的重要性。由上表可知，没有CoT格式，语言模型仍然可以自我改进，但与使用所有四种格式相比，性能提升的幅度大幅下降。



消融研究：在GSM8K数据集上以w/或w/o Cot推理路径作为训练格式。

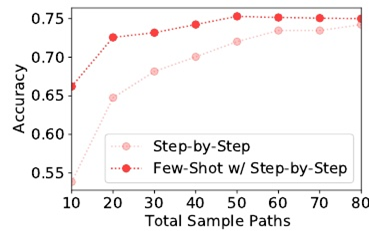
* 1. **突破自我提升的极限**

**自生成问题：**作者进一步探索了在目标领域中只有有限训练问题的小样本设置。在GSM8K上，作者抽取10个真实问题作为小样本，并使用语言模型通过第3.3节中的方法生成更多训练问题。然后，作者使用这些生成的问题自监督语言模型，结果如下。结果表明，使用自生成的问题仍然可以提高语言模型的推理能力，但使用真实的训练集问题会获得更好的结果。

****

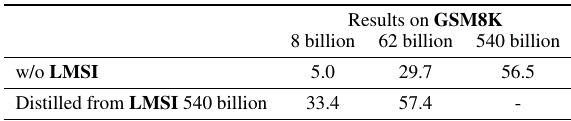
在自生成或训练集题目上进行自我训练后GSM8K测试集上的准确率。

**自生成少样本CoT提示：**作者应用逐步方法来使用语言模型生成CoT示例。如上图所示，使用自生成的逐步CoT示例进行少样本提示，显著优于逐步基线。并且几乎匹配了人类编写的少样本CoT的性能。作者指出。尽管提示示例的准确性有限（贪婪逐步为43.0%），“少样本w/逐步”的强劲表现很可能来自于利用更多样化的CoT提示进行多路径解码，在40条路径下，它使用了20个生成提示模板，每个模板有4个CoT示例，即总共80个生成的CoT示例，相比之下，在Wei等人中使用了8个人类编写的示例。，要注意的是由于作者没有使用训练问题或少样本CoT示例，74.2%也标志着GSM8K上新的最先进的零样本表现。

****

* 1. **蒸馏到更小的模型**

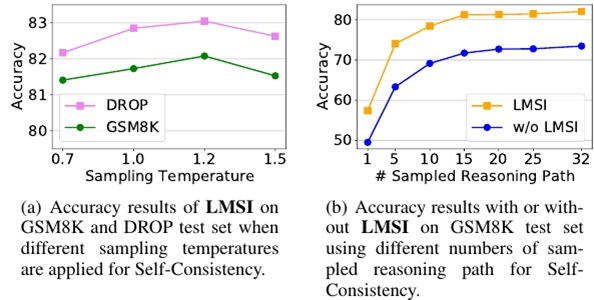
作者主要采用了以下几种方法探索其是否会蒸馏到更小的模型：1. 如Hinton等人（2015）以及Zelikm*an*等人（2022）中的蒸馏方法。2. 作者使用由P*aL*M-540B模型生成的相同训练样本集，但在体积更小的模型上进行微调（分别是P*aL*M-8B和P*aL*M-62B）。

****

从P*a*LM-540B模型到小模型的蒸馏。作者看到，蒸馏后的小模型比大一层的模型表现更好。

由上表中展示的CoT提示的结果，作者认为这意味着对于计算资源有限的下游应用，来自大模型的推理知识可以用来大幅增强小模型，以实现有竞争力的性能。

* 1. **超参数研究  
      自改进后的采样温度研究：**作者应用LMSI研究了多路径解码的采样温度T变化的效果。上图展示了在GSM8K和DROP数据集上的结果。如图，T=1.2对两个数据集都最有益，并被用于所有数据集上的LMSI自一致性方法中。模型自我改进后的最优T比自我改进前的最优T=0.7（W*an*g等人，2022b）要大（作者认为原因是训练模型后，输出分布的熵减少了）。

****

**自一致性的采样推理路径数量：**作者研究了自一致性的采样推理路径数量m在应用LMSI后是否大幅影响准确性。在上图中展示了有或没有LMSI的模型在GSM8K测试集上的准确度。对于这两种情况，设置m=15已经能达到相当好的准确度，使用更大的m只会带来边际改进。作者还注意到，经过自我改进后，使用5条路径进行自一致性就可以超过没有自我改进的模型使用32条路径的性能。因此，有了一个好的改进模型，当应用到实际应用场景时可以节省巨大的计算资源。

1. **局限性与讨论**
2. 3.2里的模型里有一个let's think step by step 的提示，这个提示下有没有可能会误导模型的判断，本文作者未有详细阐述。
3. 在问题的生成里的自我一致性方法筛选问题的时候，会不会导致一些不够自信的问题被筛掉，从而导致问题的严谨的程度不够高。

以上即本团队在深入阅读和理解文章后暂时提出的疑问。