揭秘LLM长链思考(Long CoT)： 如何让AI像人一样深度推理?

原创 唐国梁Tommy 唐国梁Tommy 2025年02月08日 20:37 广东

LLMs在各领域展现出惊人的能力，从数学推理到代码编写，几乎无所不能。这背后的一个关键技术就是**链式思考（Chain-of-Thought，CoT）**，它允许模型在给出最终答案前，先生成中间推理步骤。然而，如何让LLMs生成更长、更复杂的思考链，仍然是一个挑战。

今天跟大家分享一篇最新的论文（2025年2月5日发表），看看卡耐基梅隆大学的研究者们是如何解密LLM长链思考的奥秘的。这篇论文深入探讨了如何让LLM像人类一样进行深度推理，通过**更长、更复杂的思考过程**来解决难题。这不仅仅是简单地生成答案，而是让AI学会像人类一样逐步分析问题、反思错误、并迭代改进。

一、为什么我们需要长链思考?

- **LLM的推理能力**：大模型在数学和编程等领域展现出卓越的推理能力。它们通过**链式思考（Chain-of-Thought, CoT）**的方式，逐步生成中间推理步骤，最终得出答案，从而提高解决问题的准确性。

- **复杂任务的挑战**：面对**高度复杂的推理任务**，如数学竞赛、博士级科学问答和软件工程等，即使使用CoT，LLM仍然表现出不足。这些任务需要更深入的思考，包括识别和纠正错误、分解难题、以及尝试不同的方法。

- **长链思考的优势**：最近，OpenAI的o1和DeepSeek的R1模型展示了通过**扩展推理计算**，并采用**长CoT**来显著提升性能的能力。这种方法允许LLM进行更长、更结构化的推理过程，类似于人类的深度思考。

- **研究动机**：为了复制o1模型的性能，研究人员试图训练LLM生成长CoT。现有的方法大多依赖于**可验证的奖励信号**（例如，基于ground-truth答案的准确率），以避免强化学习（RL）中的奖励黑客现象。然而，我们对于**LLM如何学习和生成长CoT**的机制仍然缺乏全面的了解。这篇论文旨在系统地研究长CoT生成的内在机制，从而为优化训练策略提供指导。

二、论文主要思路与贡献

这篇论文通过大量的实验，揭示了以下四个关键发现：

- **有监督微调（SFT）并非必要，但能简化训练并提高效率。**虽然SFT不是生成长CoT的必要条件，但它可以作为RL的良好起点，帮助模型更快地学习。

- **推理能力随训练计算量增加而涌现，但并非必然。**这意味着需要精心设计奖励机制来稳定CoT长度的增长，并确保模型能够有效地进行深度推理。

- **可验证奖励信号的规模至关重要。**论文发现，利用带有过滤机制的、从网络提取的噪声解决方案具有巨大的潜力，尤其是在诸如STEM推理等超出分布（OOD）的任务中。

- **纠错等核心能力在基础模型中已存在，但通过RL有效地激励这些能力需要大量的计算。**同时，评估这些能力的涌现需要更加精细的方法。

总而言之，这篇论文不仅深入分析了长CoT的运作机制，还为如何高效地训练具有长CoT推理能力的LLM提供了实际指导。

三、方法详解：如何训练LLM生成长链思考?

这篇论文主要探讨了以下几种训练方法：

- **有监督微调（SFT）**：
  - **长CoT vs. 短CoT**：研究比较了使用长CoT和短CoT数据进行SFT的效果。结果表明，**长CoT SFT能够达到更高的性能上限**，并且更有潜力通过后续的RL进一步提升。
  - **SFT数据来源**：为了获取高质量的长CoT数据，论文比较了两种方法：通过提示短CoT模型生成基本动作，然后将它们组合成长的CoT轨迹，以及从现有的长CoT模型中提取轨迹。结果表明，**从现有长CoT模型中提取的轨迹**，可以更好地泛化并提升RL效果。

- **强化学习（RL）**：
  - **奖励函数设计**：论文发现，简单的奖励函数（如只奖励正确答案）会导致CoT长度不稳定。为了解决这个问题，论文引入了**余弦长度缩放奖励**，它根据CoT的长度调整奖励，并引入重复惩罚，从而稳定CoT的增长，并鼓励模型探索分支和回溯等推理行为。
  - **奖励信号**：论文强调了**可验证奖励信号**在稳定长CoT RL中的重要性。为了扩大可验证数据的规模，研究人员探索了使用**带有噪声的、从网络提取的解决方案**，并结合过滤机制来提高其质量。
  - **基模型的RL**：论文还研究了**直接从基模型进行RL训练**，而不是从SFT模型开始。结果表明，**从长CoT SFT模型初始化的RL**通常表现更好。同时，研究发现基模型中可能已经存在某些推理能力，但通过RL来激励这些能力需要更谨慎的设计。
  - **奖励的折扣因子**: 研究发现，不同类型的奖励和惩罚需要不同的**最优折扣因子**。例如，对于重复惩罚，**较低的折扣因子**能够更有效地进行惩罚，而对于正确性奖励，**较高的折扣因子**可以提高模型的性能。

- **模型结构**
  - **上下文窗口大小**：研究发现，模型可能需要**更多的训练样本**才能学会充分利用更大的上下文窗口。实验表明，虽然更大的窗口（8K）比更小的窗口（4K）表现更好，但更大的窗口（16K）却不如8K窗口，这表明更大的窗口需要更多的训练才能有效利用。

- **重复惩罚**：研究表明，当训练计算足够时，模型可能会通过重复来增加其CoT长度，从而出现**奖励攻击**。为了缓解这种情况，研究人员引入了**n-gram重复惩罚**，有效地减少了重复，并提高了模型的性能。

为了帮助大家理解，我将用一个简化的例子来说明这些概念：

- 假设你是一位老师，想教学生解一道难题
  - **短CoT** 就像老师只给出解题步骤，学生照搬即可。这适用于简单的问题，但对于复杂问题，学生可能无法理解背后的逻辑。
  - **长CoT** 就像老师引导学生逐步分析问题，鼓励学生尝试不同的解题思路，并在错误中学习。
  - **SFT** 就像老师先示范一些解题方法，帮助学生建立基础。
  - **RL** 就像老师根据学生的解题过程给出反馈，鼓励学生改进，并根据解题步骤的长度和质量给予不同的奖励。
  - **奖励函数** 就像老师的评分标准，包括解题正确性和思考过程。好的标准能够鼓励学生深度思考。
- **奖励的折扣因子** 就像老师在给评分时，更重视学生最近的表现（较低折扣因子）还是整体的解题过程（较高折扣因子）

通过这些方法，研究人员成功地训练出了能够生成更长、更复杂的CoT的LLM，从而提高了模型在复杂推理任务中的性能。

四、实验结果

论文通过大量实验，验证了以下结论：

- **SFT的扩展性**：使用长CoT进行SFT可以达到更高的性能上限，而短CoT则较早饱和。
- **SFT对RL的初始化作用**：使用长CoT进行SFT初始化，更容易通过RL进一步提升性能，而短CoT则提升不大。
- **长CoT数据的来源**：从现有的长CoT模型中蒸馏得到的长CoT数据，比通过动作提示框架构建的CoT数据，泛化能力更好，并且可以通过RL进一步提升。

Table 4. Performance of different models based on `Qwen2.5-Math-7B`. The SFT data here is distilled with rejection sampling from `QwQ-32B-Preview`.

Setup	MATH 500	AIME 2024	Theo. QA	MMLU Pro-1k	AVG
Base (0-shot)	52.0	13.3	17.1	2.4	21.2
(Direct) RL	77.4	23.3	43.5	19.7	41.0
SFT	84.0	24.4	42.2	38.5	47.3
SFT + RL	**85.9**	**26.9**	**45.4**	**40.6**	**49.7**

- **奖励设计的影响**：简单的奖励函数（例如，只奖励正确答案）会导致CoT长度不稳定，甚至出现奖励黑客现象（即通过重复生成token来增加CoT长度，而不是真正地解决问题）。
- **余弦奖励** 可以有效稳定CoT的长度，提高模型的训练效率和性能。

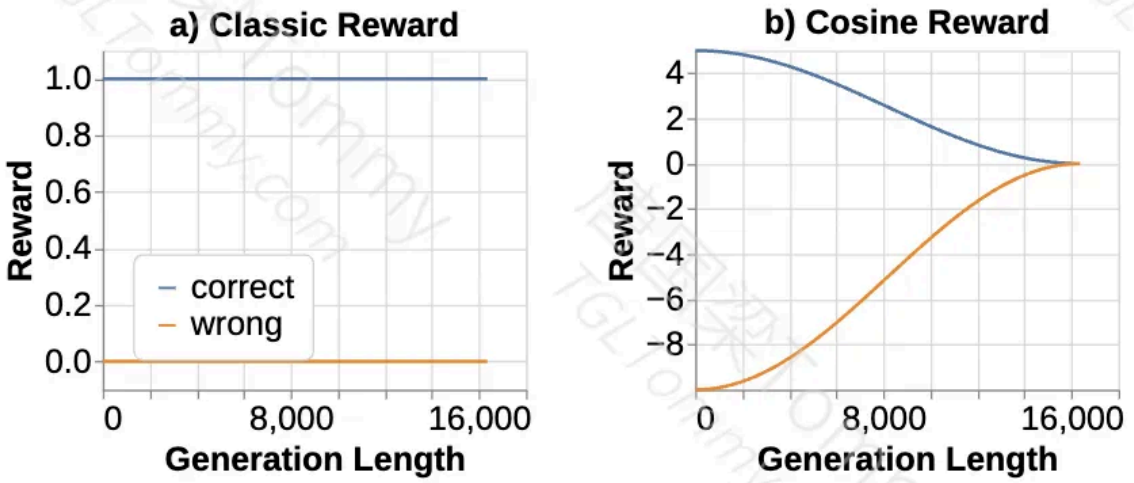

Figure 3. The Classic and Cosine Reward functions. The Cosine Reward varies with generation length.

- 奖励的超参数（如奖励的权重、折扣因子）会影响CoT的长度和模型的行为。

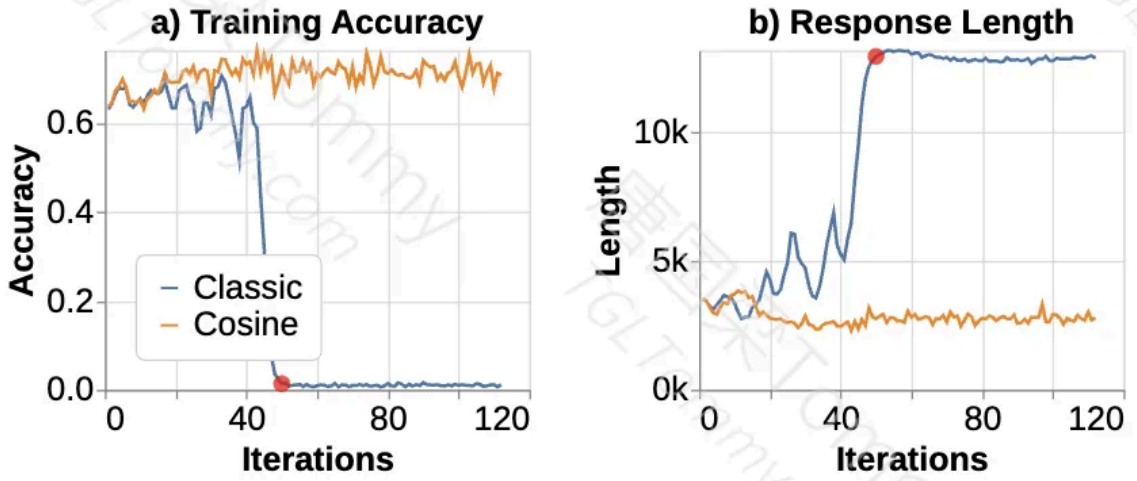

Figure 4. `Llama3.1-8B` trained with length shaping using the Cosine Reward exhibited more stable (a) training accuracy and (b) response length. This stability led to improved performance on downstream tasks (Figure 5). Red points on the charts indicate iterations where training accuracy dropped to near zero.

- **上下文窗口大小的影响**：模型需要更多的训练数据才能充分利用更大的上下文窗口。

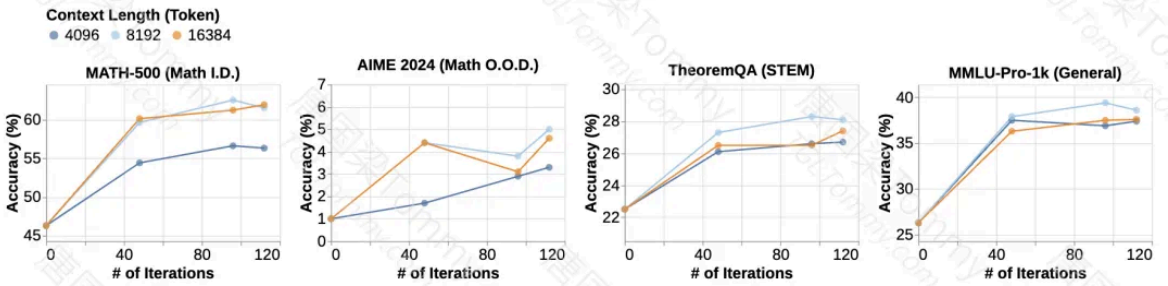

Figure 6. Performance of `Llama-3.1-8B` trained with different context window sizes. All experiments used the same number of training samples.

- **可验证奖励信号的扩展**：
  - 将带噪声的、从网络提取的解决方案（如WebInstruct）加入到SFT中，可以提高模型的泛化能力。
  - 使用规则验证器过滤带有短答案的数据集，可以有效利用噪声数据来提升RL的性能。

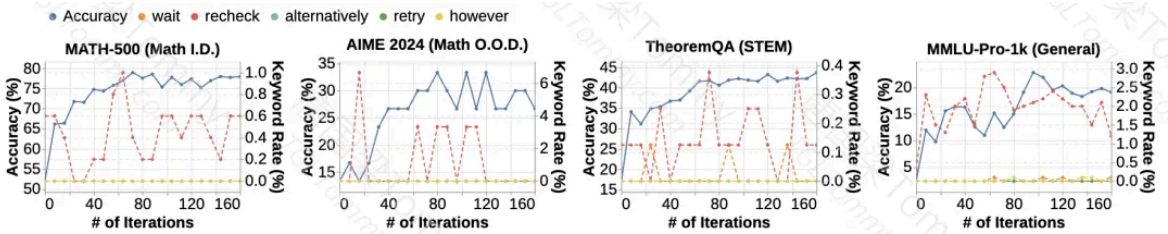

Figure 7. Dynamics of accuracies and reflection keyword rates on different benchmarks during our RL from the base model `Qwen2.5-Math-7B`. We do not see the keyword rates of "wait", "alternatively", and "recheck" get significantly improved during the RL training even though the accuracy is steadily increasing.

- **从基础模型开始的RL**：直接从基础模型开始RL训练，虽然可以提升性能，但是可能无法有效地激励长CoT的行为（如回溯、纠错），并且可能无法超过基础模型的输出

长度。

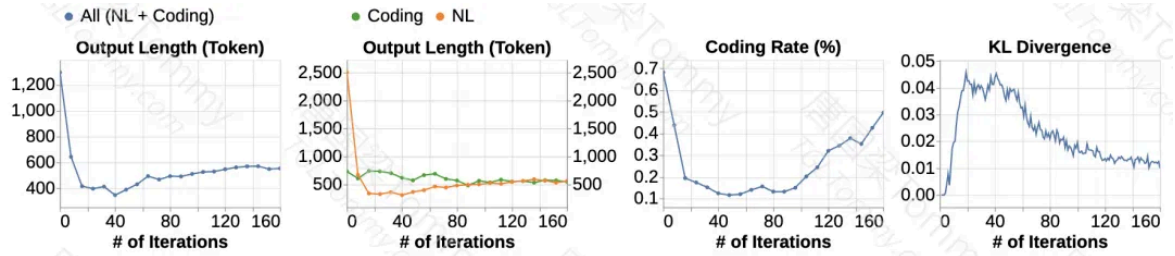

Figure 8. Dynamics of the output token lengths and the coding rate on MATH-500 and the KL divergence of the policy over the base model on MATH Lv3-5 (training data) during our RL from `Qwen2.5-Math-7B`.

- **长CoT模式的预训练数据**：互联网上的讨论论坛中存在大量的长CoT模式，表明这种能力可能在预训练阶段就被部分学习了。

总结

总的来说，这篇论文深入研究了LLM长链思考的机制，为如何训练具有深度推理能力的AI提供了重要的理论基础和实践指导。它不仅指出了训练长CoT的挑战和方法，还为未来的研究方向提供了深刻的见解。我们有理由相信，随着研究的深入，未来的AI将能够像人类一样进行更深入、更复杂的思考，从而更好地服务于社会和人类的发展。

这篇论文的研究告诉我们，**AI的进步不仅仅是技术上的突破，更是在理解人类思维模式基础上的创新。**通过对人类思维模式的深入理解，我们可以更好地设计AI，并让AI像人类一样进行深度思考和推理。

参考文献

论文名称: Demystifying Long Chain-of-Thought Reasoning in LLMs

第一作者: 卡耐基梅隆大学

论文链接: https://arxiv.org/abs/2502.03373v1

发表日期: 2025年2月5日

GitHub：https://github.com/eddycmu/demystify-long-cot.git