AdaptThink：推理模型可以学习何时思考

张佳杰、林念义、侯磊、冯玲、李娟子清华大学

摘要

最近，大型推理模型通过采用类人的深度思维，在各种任务上取得了令人印象深刻的性能。然而，漫长的思考过程大大增加了推理开销，使效率成为一个关键的瓶颈。在这项工作中，我们首先证明了NoThinking，它促使推理模型跳过思考并直接生成最终解决方案，在性能和效率方面都是相对简单任务的更好选择。受此启发，我们提出了AdaptThink，这是一种新的RL算法，用于教导推理模型根据问题难度自适应地选择最佳思维模式。具体来说，AdaptThink有两个核心组成部分：（1）一个约束优化目标，鼓励模型在保持整体性能的同时选择NoThinking；（2）重要性抽样策略，在政策培训期间平衡思考和非思考样本，从而实现冷启动，并允许模型在整个培训过程中探索和利用这两种思维模式。我们的实验表明，AdaptThink显著降低了推理成本，同时进一步提高了性能。值得注意的是，在三个数学数据集上，AdaptThink将DeepSeek-R1-DistillQwen-1.5B的平均响应长度缩短了53%并通过以下方式提高其精度2.4%强调了自适应思维模式选择对优化推理质量和效率之间平衡的承诺。我们的代码和模型可以在https://github上找到。

1简介

大型推理模型的最新进展，如OpenAI o1（OpenAI，2024）和DeepSeekR1（DeepSeek AI，2025），在处理复杂任务方面表现出了非凡的能力。给定一个问题，这些模型首先参与一个长链的思考——也称为思考——在那里，它们迭代地探索不同的方法，并伴随着反思、回溯和自我验证。随后，他们会生成一个最终解决方案，其中只包含正确的步骤和要呈现给用户的答案。虽然长时间思考过程显著提高了模型的推理能力，但它也大大增加了推理开销和延迟（Qu等人，2025；Sui等人，2015）。特别是，对于用户期望快速、近乎即时的响应的一些简单查询，这些模型通常会产生过度的思考，包括不必要的详细步骤或重复的尝试，从而导致次优的用户体验（Chen等人，2024；Shen等人，2025）。

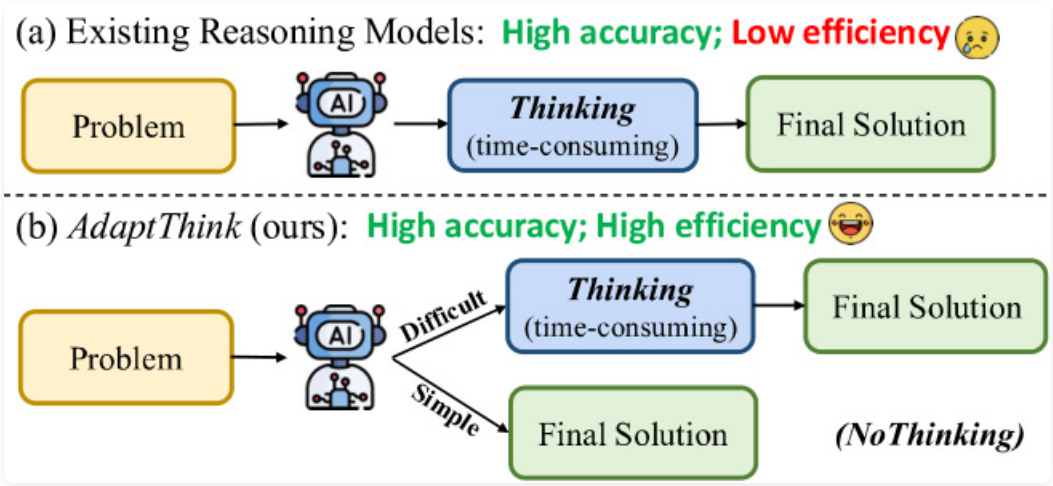

图1:AdaptThink使模型能够根据问题难度在思考或无思考模式之间进行自适应选择，从而提高推理效率，同时进一步提高整体性能。

现有的提高推理效率的努力主要集中在减少模型响应的长度上，要么通过在强化学习（RL）中引入基于长度的奖励（Arora和Zanette，2025；Team等人，2025），用偏好对进行微调以惩罚较长的响应（Chen等人，2024；Shen等人，2025；Luo等人，2025a），要么通过合并推理和非推理模型（Wu等人，2025）。然而，这些方法仍然将思维应用于所有情况，无论思维本身是否对每个问题都是必要的。在

这项工作中，我们从最近引入的NoThinking方法（Ma等人，2025）中获得了灵感，该方法允许推理模型跳过思维过程，通过伪思维过程的提示直接生成最终解决方案。具体来说，我们通过用空思维片段（即“”）提示模型来进一步简化方法。我们在第3节中的试点研究表明，在相对简单的问题上（高达高中竞争水平），NoThinking的表现与thinking相当，甚至更好，同时显著减少了代币的使用；只有当问题足够困难时，thinking的好处才会显现出来。

鉴于这一观察，我们很好奇：推理模型能否根据输入问题的难度自适应地选择思考或无思维模式，从而在不牺牲甚至提高性能的情况下实现更有效的推理？为此，我们提出了AdaptThink，这是一种新颖的RL算法，用于教导推理模型何时思考。具体来说，AdaptThink有两个核心组成部分：（1）一个约束优化目标，鼓励模型选择NoThinking，同时确保整体性能不会下降；（2）在政策培训过程中平衡思考和非思考样本的重要性抽样策略，从而克服冷启动的挑战，并允许模型在整个培训过程中探索和利用这两种思维模式。

我们的实验表明，AdaptThink有效地使推理模型能够根据问题难度自适应地选择最佳思维模式，与现有方法相比，大大降低了推理成本，同时持续提高了模型的准确性。例如，在GSM8K、MATH500和AIME2024上，AdaptThink将DeepSeek-R1-Distill-Qwen-1.5B的平均响应长度缩短了50.9% 63.5%，以及44.7%，并通过以下方式提高其准确性4.1%，1.4%，以及1.6%分别。这些显著的结果证实了困难适应思维模式选择作为促进推理性能和效率之间权衡的有前景的范式的潜力。

总之，我们的主要贡献如下：（1）我们简化了NoThinking方法，并证明了它在性能和效率方面优于Thinking，适用于更简单的任务；（2）我们提出了AdaptThink，这是一种新的RL算法，它使推理模型能够根据问题难度自适应地选择最优思维模式，从而大大降低了推理成本，进一步提高了性能；（3）我们进行了广泛的实验来验证AdaptThink的有效性。

2相关工作

大型推理模型。最近的前沿大型推理模型（LRM），如OpenAI o1（OpenAI，2024）、DeepSeek-R1（DeepSeek AI 2025）和QwQ（Qwen Team，2025），已经开发出在问题解决中使用类人深度思维的能力，通过在得出最终解决方案之前生成长链思维。这种高级能力通常是通过大规模强化学习获得的，具有经过验证的奖励或对提取的推理痕迹进行微调。尽管性能很有希望，但漫长的思考过程会带来大量的推理成本和延迟。因此，人们提出了各种方法来进行更有效的推理。

LRM的高效推理。大多数现有的提高LRM效率的方法都侧重于减少模型响应中的令牌使用。一些方法将基于长度的奖励纳入RL，以激励更简洁的反应（Arora和Zanette，2025；Team等人，2025）或实现对反应长度的精确控制（Aggarwal和Welleck，2025）。其他方法通过最佳N采样获得的具有长度相关偏好对的模型进行微调（Luo等人。，2025aShen等人，2025）或通过后处理（Chen等人，2024）。此外，有几项研究采用无需训练的方法来缩短响应时间，采用模型合并（Team等人，2025；Wu等人，2025）或提示（Han等人，2024；Muennighoff等人，2025，Fu等人，2015；Xu等人，2020）等技术。尽管如此，这些方法仍然对所有问题都采用了长期思考，而最近的NoThinking方法（Ma等人，2025）允许推理模型绕过长期思考，通过提示直接输出最终解决方案，在低令牌预算环境中实现了与thinking相当的性能。在这项工作中，我们进一步证明，即使有足够的代币预算，NoThinking在使用更少的代币的情况下，也可以在简单问题上超越Thinking。这一观察促使我们提出了AdaptThink来教授推理模型，以根据问题难度自适应地选择最佳思维模式，这是高效推理的一个新方向。

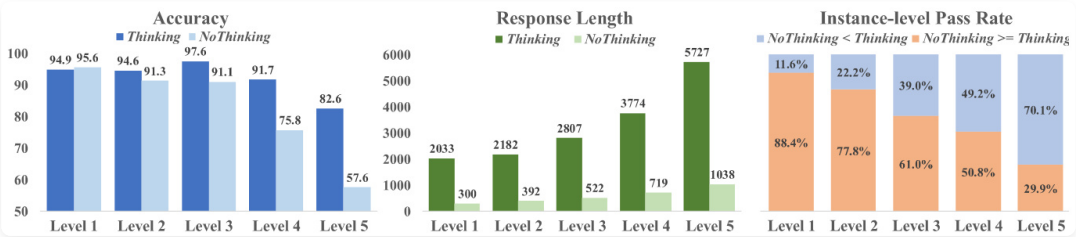

图2:MATH500数据集不同难度级别下使用思维和无思维模式的DeepSeek-R1-Distill-Qwen-7B的比较。

3动机

3.1初步

考虑一个推理模型，其参数化为θ并表示为π_θ.给予提示$x = [x_1, \ldots, x_n,]$，在哪里$[x_1, \ldots, x_n]$代表问题和是启动思维过程的特殊标记，模型生成响应$y = [y_1, \ldots, y_l, , y_{l+2}, \ldots, y_2$在这里，$[y_1, \ldots, y_l]$与思维相对应，思维是一个由不断探索、反思和自我验证组成的长链思维。标记标志着思考的结束，$[y_{l+2}, \ldots, y_m]$，表示最终解决方案，仅包括解决问题的正确步骤和最终答案。从概率论的角度来看，反应y是从条件概率分布中提取的样本$\pi_\theta(\cdot|x)$.由于y以自回归的方式生成条件概率$\pi_\theta(y|x)$可以分解为：

$$ \pi_{\boldsymbol{\theta}}(y|x)=\prod_{t=1}^{m}\pi_{\boldsymbol{\theta}}(y_{t}|x,y_{ # 3.2简单问题最好不要思考 当前的推理模型，如OpenAI-ol和DeepSeek-R1，在所有问题上都应用了长时间思维（称为思维模式）。虽然增强了模型的推理能力，但漫长的思维过程往往会导致不必要的计算开销，特别是对于一些简单的问题，这些问题也可以通过非推理模型（例如GPT-4o和Qwen-2.5-Induce）无需思考即可解决。最近，Ma等人（2025）提出了NoThinking方法，该方法使推理模型能够绕过长时间思考，通过虚假的思维过程“好吧，我想我已经完成了思考。”来直接生成最终解决方案，并发现它在低代币预算设置中仍然有效。在这项工作中，我们通过为模型提供空思维（即强制执行第一个生成的代币）来进一步简化NoThinking$y_1 =$;然后，我们进行了一项试点研究，从问题难度的角度比较思考和无思考，并有足够的代币预算（16K）。 具体来说，我们利用MATH5OO（Lightman等人，2O24）数据集进行试点研究，因为它将问题分为五个难度级别。对于每个问题，我们使用DeepSeek-R1-DistillQwen-7B分别使用Thinking和NoThinking生成16个响应。然后，我们分析了五个难度级别的准确性、响应长度和实例级通过率。如图2所示，尽管该模型是使用长思维数据训练的，但NoThinking在相对简单的问题（1到3级）上仍能达到与Thinking相当的准确性，甚至在最简单的1级问题上略优于Thinking。与此同时，NoThinking反应的平均长度明显短于Thinking反应。此外，与Nothinking相比，Thinking仅将不到一半的问题的实例级通过率从1级提高到4级。总的来说，这些发现表明，思考只会给具有挑战性的问题带来显著的好处，而无思考在准确性和效率方面都是更简单问题的更好选择。这促使我们从新的角度探索高效推理：教推理模型根据问题难度自适应地选择思考或无思维模式，从而在保持甚至提高整体性能的同时降低推理成本。为此，我们提出了AdaptThink，这是一种新颖的RL算法，可以教推理模型何时思考。 # 1自适应思维 我们的AdaptThink算法由两个重要部分组成：（1）约束优化