DeepSeek R1 论文解读

2025年2月10日 3 分钟阅读

#AI #深度思考 #DeepSeek-R1 #论文 #技术

本文介绍了深度求索（DeepSeek）公司推出的新一代推理模型DeepSeek-R1，并对其技术原理、主要贡献、论文方法、评估结果和局限性进行了详细解读。

引言

近年来，大型语言模型（LLMs）快速迭代和发展，正逐步缩小与通用人工智能（AGI）的差距。 其中，后训练技术已成为完整训练流程中的重要组成部分，它能够在推理任务、社会价值观对齐和用户偏好适应等方面提升模型性能，同时相较于预训练所需的计算资源更少。

OpenAI的o1系列模型率先引入了推理时扩展，通过增加思维链长度提升推理能力。这种方法在数学、编码和科学推理等各种任务中取得了显著进步。然而，有效扩展测试时计算的挑战仍然是一个悬而未决的问题。

DeepSeek-R1论文摘要

本文介绍了一种名为DeepSeek-R1的模型，它通过强化学习的方式训练，并具有出色的推理能力。该模型分为两个版本：DeepSeek-R1-Zero和DeepSeek-R1。 DeepSeek-R1-Zero在无监督下进行训练，表现出了强大的推理行为，但存在可读性差、语言混合等问题。为了解决这些问题并进一步提高推理性能，作者引入了多阶段训练和冷启动数据，并开发了DeepSeek-R1。实验结果表明，DeepSeek-R1的推理性能与OpenAI-o1-1217相当。此外，作者还开源了DeepSeek-R1-Zero、DeepSeek-R1以及基于Qwen和Llama的六个密集模型（1.5B、7B、8B、14B、32B、70B）。

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#AI #深度思考 #DeepSeek #技术

DeepSeek-R1的主要贡献

DeepSeek-R1是深智科技开发的新一代推理模型，其主要贡献在于：

> 纯强化学习实现推理能力提升

DeepSeek-R1-Zero是第一个在没有监督微调（SFT）的情况下，通过大规模强化学习（RL）训练的推理模型。 通过RL，DeepSeek-R1-Zero自然地展现出许多强大的推理行为，如自我验证、反思和生成长思维链等。 这是第一个验证LLMs推理能力可以通过纯RL激励的公开研究，为该领域的未来发展铺平了道路。 2. 多阶段训练流程 DeepSeek-R1引入多阶段训练流程，包括两个RL阶段和两个SFT阶段，旨在发现改进的推理模式并与人类偏好保持一致。 该流程包括：

> 冷启动：使用少量长思维链样本微调基础模型，解决RL训练早期不稳定的问题。
>
> 面向推理的强化学习： 使用规则奖励系统，重点提升模型在编码、数学、科学和逻辑推理等推理密集型任务上的能力。
>
> 拒绝抽样和监督微调： 利用RL训练的检查点收集SFT数据，涵盖推理和非推理领域，进一步增强模型的多功能性。
>
> 面向所有场景的强化学习： 结合规则奖励和奖励模型， 提升模型在所有场景下的实用性和安全性。
>
> 推理能力蒸馏至小型模型 DeepSeek-R1的推理能力可以被蒸馏到较小的密集模型中，例如Qwen和Llama系列。 使用DeepSeek-R1生成的推理数据微调这些模型，其性能显著提升，超越了现有开源模型，并可与o1-mini相媲美。

论文速读

论文方法

方法描述

该论文提出了两种方法来提高模型的推理能力：基于强化学习（RL）的训练和知识蒸馏。 首先，他们直接应用了强化学习到基础模型上，而没有使用监督微调作为预处理步骤。这种方法允许模型探索链式思维（CoT），以解决复杂问题，并开发出DeepSeek-R1-Zero。DeepSeek-R1-Zero展示了自我验证、反思和生成长链CoT等能力，标志着研究社区的一个重要里程碑。值得注意的是，这是第一个公开的研究，证明了LLMs的推理能力可以通过纯粹的RL激励实现，而不必依赖SFT。这一突破为未来的发展铺平了道路。 其次，

他们引入了一个管道来开发DeepSeek-R1。该管道包括两个RL阶段，旨在发现更好的推理模式并与人类偏好相一致，以及两个SFT阶段，作为模型推理和非推理能力的种子。他们相信这个管道将通过创建更好的模型来造福行业。

方法改进

对于第一种方法，他们通过RL训练提高了模型的推理能力。而对于第二种方法，他们利用已经发现的更大模型的推理模式来"蒸馏"更小的模型，从而提高了性能。

解决的问题

该论文的主要目标是提高模型的推理能力，使其能够更好地解决复杂的任务。他们通过RL训练和知识蒸馏两种方法实现了这一目标，并取得了显著的成果。这些成果不仅超越了之前公开的模型，而且也达到了o1-mini的水平。因此，他们的工作在推动自然语言处理领域的发展方面具有重要意义。

论文实验

本文主要介绍了作者在深度学习模型上所做的两个对比实验。第一个实验是使用纯强化学习训练的模型（DeepSeek-R1-Zero）和带有少量冷启动数据的模型（DeepSeek-R1）之间的比较。第二个实验是对小模型（如Qwen和Llama）进行知识推理能力的提高，通过直接对这些模型进行微调来实现。 在第一个实验中，作者将纯强化学习训练的模型（DeepSeek-R1-Zero）与带有少量冷启动数据的模型（DeepSeek-R1）进行了比较。作者采用了不同的奖励系统来引导模型的学习过程，并观察了模型在各种任务上的表现。结果表明，带有少量冷启动数据的模型（DeepSeek-R1）在大多数任务上都表现出更好的性能，特别是在需要更复杂推理的任务上。这表明，通过引入少量高质量的数据可以显著提高模型的性能。 在第二个实验中，作者通过对几个开源模型（如Qwen和Llama）进行微调来提高它们的知识推理能力。作者使用了相同的方法来收集和筛选数据，并对模型进行了微调。结果显示，经过微调后的小型模型（如Qwen和Llama）在许多任务上都表现出了比原始模型更好的性能。这进一步证明了微调方法的有效性，并为如何提高小型模型的性能提供了一种可行的方法。 总之，这两个实验都展示了强化学习和微调方法在深度学习模型中的有效性，并为进一步改进模型提供了有价值的见解。

论文总结

文章优点

● 本文提出了一种基于强化学习的方法来增强模型的推理能力，并在多个任务上取得了优异的表现。 ● 通过对比不同模型训练方式的结果，证明了模型大小对于性能的影响以及预训练的重要性。 ● 对

于未来的改进方向，作者提出了几个研究方向，包括增强模型的通用能力、解决多语言问题、优化提示工程等。

方法创新点

● 本文提出的强化学习方法可以有效地提高模型的推理能力，特别是在数学和科学领域。

● 利用大型数据集进行预训练，可以使模型更好地适应不同的任务和场景。 ● 通过对小模型进行知识蒸馏，可以在不增加计算资源的情况下进一步提升模型性能。

未来展望

● 在未来的研究中，可以通过增强模型的通用能力来扩展其应用范围。 ● 解决多语言问题可以使得模型更加普适，能够处理来自不同语言环境下的查询请求。 ● 优化提示工程可以帮助用户更准确地描述问题并获得更好的结果。

● 将强化学习方法应用于软件工程等领域，可以进一步拓展模型的应用范围。

DeepSeek-R1的评估结果

DeepSeek-R1在多个基准测试中取得了令人瞩目的成果， 总结如下:

> 推理任务: 在AIME 2024上，DeepSeek-R1的Pass@1得分为79.8%，略高于OpenAI-o1-1217。 在MATH-500上，DeepSeek-R1取得了97.3%的Pass@1得分，与OpenAI-o1-1217相当，并显著优于其他模型。 在代码竞赛任务中，DeepSeek-R1在Codeforces上获得了2029的Elo评级，超过了96.3%的人类参与者。
>
> 知识: 在MMLU、MMLU-Pro和GPQA Diamond等基准测试中，DeepSeek-R1取得了优异的成绩，显著优于DeepSeek-V3。DeepSeek-R1在MMLU上得分为90.8%，在MMLU-Pro上得分为84.0%，在GPQA Diamond上得分为71.5%。
>
> 其他: DeepSeek-R1在创造性写作、一般问答、编辑、摘要等方面也表现出色。 在AlpacaEval 2.0上，DeepSeek-R1的长度控制胜率为87.6%，在ArenaHard上为92.3%。 此外，DeepSeek-R1在需要长上下文理解的任务上也表现出色，显著优于DeepSeek-V3。

DeepSeek-R1的局限性和未来方向

虽然DeepSeek-R1取得了显著进展，但仍存在一些局限性， 未来工作将集中在以下方向：

通用能力: 目前，DeepSeek-R1在函数调用、多回合对话、复杂角色扮演和JSON输出等任务上的能力不足。未来将探索如何利用长思维链来增强这些领域的任务。

语言混合: DeepSeek-R1目前针对中文和英文进行优化，处理其他语言的查询时可能会出现语言混合问题。 未来将努力解决这一局限性。

提示工程: DeepSeek-R1对提示非常敏感，少样本提示会降低其性能。建议用户直接描述问题并使用零样本设置指定输出格式以获得最佳结果。

软件工程任务: 由于评估时间过长，影响了RL流程的效率，大规模RL尚未在软件工程任务中得到广泛应用。 未来版本将通过对软件工程数据进行拒绝抽样或在RL过程中加入异步评估来提高效率。

结论

DeepSeek-R1是深智科技在提升模型推理能力方面取得的重要里程碑。 DeepSeek-R1-Zero展示了纯RL方法的潜力，而DeepSeek-R1则结合了冷启动数据和迭代RL微调，取得了与OpenAI-o1-1217相当的性能。 DeepSeek-R1的开源和API将推动推理模型的发展，使其在更广泛的应用中发挥作用。

论文10问

针对DeepSeek-R1论文，采用“论文10问”的方法进行解读，可以帮助你更深入地理解其核心内容、创新之处以及潜在价值。

研究背景是什么?

近年来，大型语言模型 (LLM) 发展迅速，逐渐缩小了与通用人工智能 (AGI) 之间的差距。后训练已成为完整训练流程的重要组成部分，可以提高推理任务的准确性，符合社会价值观，并适应用户偏好，同时相对于预训练，所需的计算资源相对较少. OpenAI 的 o1 系列模型首先通过增加思维链推理过程的长度来

引入推理时缩放，在数学、编码和科学推理等各种推理任务中取得了显著的改进。然而，有效的测试时缩放的挑战仍然是研究界的一个开放性问题。

DeepSeek-R1 旨在**通过纯强化学习 (RL) 改进语言模型的推理能力**。目标是探索 LLM 在没有任何监督数据的情况下发展推理能力的潜力，重点关注它们通过纯 RL 过程的自我进化。

论文要解决什么问题?

DeepSeek-R1 旨在解决以下几个核心问题：

如何通过纯强化学习 (RL) 提升 LLM 的推理能力，而无需依赖监督微调 (SFT)?

是否可以通过引入少量高质量数据作为冷启动来进一步提高推理性能或加速收敛?

如何训练一个用户友好的模型，该模型不仅产生清晰连贯的思维链 (CoT)，而且还展示出强大的通用能力?

如何将大型模型的推理能力迁移到较小的模型中，从而实现更高效的推理?

如何解决 DeepSeek-R1-Zero 存在的诸如可读性差和语言混合等问题?

论文采用了什么方法?

为了解决上述问题，DeepSeek-R1 采用了以下关键方法：

DeepSeek-R1-Zero： 直接将强化学习 (RL) 应用于基础模型，而无需任何 SFT 数据。DeepSeek-R1-Zero 展示了自验证、反思和生成长 CoT 等能力。

DeepSeek-R1： 从使用数千个长思维链 (CoT) 示例进行微调的检查点开始应用 RL。该流程包括四个阶段：

冷启动 (Cold Start)： 构建并收集少量的长 CoT 数据，以微调模型，作为初始 RL 参与者。

面向推理的强化学习 (Reasoning-oriented Reinforcement Learning)： 在冷启动数据上微调 DeepSeek-V3-Base 后，应用与 DeepSeek-R1-Zero 相同的大规模强化学习训练过程。

拒绝采样和监督微调 (Rejection Sampling and Supervised Fine-Tuning)： 当面向推理的 RL 收敛时，利用生成的检查点来收集后续回合的 SFT（监督微调）数据。

所有场景的强化学习 (Reinforcement Learning for all Scenarios)： 实施了二级强化学习阶段，旨在提高模型的帮助性和无害性，同时改进其推理能力。

知识蒸馏 (Knowledge Distillation)： 使用 DeepSeek-R1 生成的推理数据微调了多个广泛用于研究社区的密集模型。

论文的关键结果是什么?

DeepSeek-R1 的关键结果包括：

DeepSeek-R1-Zero 证明了 LLM 的推理能力可以通过纯 RL 来激励，而无需 SFT。在 AIME 2024 基准测试中，DeepSeek-R1-Zero 的 pass@1 分数从 15.6% 提高到 71.0%，与 OpenAI-o1-0912 的性能相当。

DeepSeek-R1 在各种推理任务中实现了与 OpenAI-o1-1217 相当的性能。在 MATH-500 上，DeepSeek-R1 达到了 97.3% 的高分。在代码竞赛任务中，DeepSeek-R1 在 Codeforces 上获得了 2,029 Elo 评分，超过了 96.3% 的人类参与者。

通过知识蒸馏，可以将大型模型的推理能力迁移到较小的模型中，并且可以获得非常好的效果。例如，DeepSeek-R1-Distill-Qwen-7B 在 AIME 2024 上实现了 55.5% 的准确率，超过了 QwQ-32B-Preview。

DeepSeek-R1 在知识密集型基准测试（如 MMLU、MMLU-Pro 和 GPQA Diamond）上取得了出色的成果。

论文有哪些创新点?

DeepSeek-R1 的创新点主要体现在以下几个方面：

首次验证了 LLM 的推理能力可以通过纯 RL 来激励，而无需 SFT。

提出了一个用于开发 DeepSeek-R1 的流水线，该流水线结合了两个 RL 阶段和两个 SFT 阶段。

证明了大型模型的推理模式可以提炼到较小的模型中，从而获得比通过 RL 在小型模型上发现的推理模式更好的性能。

论文有哪些局限性?

尽管 DeepSeek-R1 取得了显著进展，但仍然存在一些局限性：

通用能力：在函数调用、多轮对话、复杂角色扮演和 json 输出等任务中的能力不如 DeepSeek-V3。

语言混合：目前针对中文和英文进行了优化，这可能会导致在处理其他语言的查询时出现语言混合问题。

提示工程：对提示很敏感，Few-shot 提示会持续降低其性能。

软件工程任务：在软件工程基准测试中没有表现出比 DeepSeek-V3 更大的改进。

论文有哪些潜在的应用价值?

DeepSeek-R1 的潜在应用价值包括：

教育领域：可以用于开发 AI 驱动的搜索和数据分析工具。

代码生成和软件工程：可以帮助开发人员完成各种编码任务。

通用问题解答：可以处理各种非考试导向的查询。

论文对未来的研究有什么启示?

DeepSeek-R1 的研究结果表明，强化学习在提升语言模型的推理能力方面具有巨大潜力。未来的研究可以集中在以下几个方面：

探索如何利用长 CoT 来增强函数调用、多轮对话和复杂角色扮演等任务。

解决语言混合问题，使模型能够更好地处理各种语言的查询。

改进提示工程，使模型对提示不那么敏感。

将大规模 RL 应用于软件工程任务，以提高模型在这些任务中的性能。

探索更有效的知识蒸馏方法，将大型模型的推理能力迁移到更小的模型中。

论文有哪些可以改进的地方?

根据论文的局限性，可以改进的地方包括：

提高通用能力，使其在函数调用、多轮对话等任务中达到与 DeepSeek-V3 相当的水平。

解决语言混合问题，使其能够更好地处理各种语言的查询。

改进提示工程，使其对提示不那么敏感。

将大规模 RL 应用于软件工程任务，以提高模型在这些任务中的性能。

你对这篇论文有什么评价?

DeepSeek-R1 是一篇具有重要意义的论文，它**展示了强化学习在提升语言模型推理能力方面的巨大潜力。**该论文的创新之处在于，它**首次验证了 LLM 的推理能力可以通过纯 RL 来激励，而无需 SFT。**此外，该论文还**提出了一个用于开发 DeepSeek-R1 的流水线，该流水线结合了两个 RL 阶段和两个 SFT 阶段。**DeepSeek-R1 的研究结果对未来的研究具有重要的启示作用，可以促进语言模型推理能力的进一步发展。当然, 论文也坦诚地指出了自身存在的局限性, 为后续改进指明了方向。

问与答QA

DeepSeek-R1-Zero如何通过纯强化学习涌现出强大的推理能力?

DeepSeek-R1-Zero 通过纯强化学习 (RL) 涌现出强大的推理能力，主要体现在以下几个方面：

● 直接应用强化学习：DeepSeek-R1-Zero 没有依赖于任何监督微调 (SFT) 作为初步步骤，而是直接将强化学习应用于基础模型。这种方法使得模型能够探索用于解决复杂问题的思维链 (CoT)，从而发展出 DeepSeek-R1-Zero。 ● 自然涌现推理行为：在训练过程中，DeepSeek-R1-Zero 自然涌现出许多强大而有趣的推理行为。经过数千次 RL 步骤后，DeepSeek-R1-Zero 在推理基准测试中表现出色。 ● 奖励机制：DeepSeek-R1-Zero 采用了基于规则的奖励系统，主要包括以下两种类型的奖励：

○ 准确性奖励：评估回答是否正确。例如，对于具有确定性结果的数学问题，要求模型以特定格式（例如，在框内）提供最终答案，从而实现对正确性的可靠的基于规则的验证。 ○ 格式奖励：强制模型将其思考过程置于 ‘’ 和 ‘’ 标签之间。 ● 训练模板：使用一个简单的模板来指导基础模型遵守指定的指令。该模板要求 DeepSeek-R1-Zero 首先产生一个推理过程，然后给出最终答案。 ● 自我进化过程：DeepSeek-R1-Zero 的自我进化过程展示了 RL 如何驱动模型自主提高其推理能力。通过直接从基础模型启动 RL，可以密切监控模型的进展，而不会受到监督微调阶段的影响。 ● “顿悟时刻”：在 DeepSeek-R1-Zero 的训练过程中观察到一个有趣的"顿悟时刻"。在这个阶段，模型学会通过重新评估其初始方法来为问题分配更多的思考时间。这种行为不仅证明了模型不断增长的推理能力，也证明了强化学习能够带来意想不到的和复杂的结果。 ● 性能提升：在强化学习 (RL) 训练过程中，DeepSeek-R1-Zero 在 AIME 2024 基准测试中表现出稳定且持续的性能提升。AIME 2024 的平均 pass@1 分数从最初的 15.6% 显著提高到 71.0%，达到了与 OpenAI-o1-0912 相当的性能水平。在 AIME 基准测试中采用多数投票时，DeepSeek-R1-Zero 的性能从 71.0% 提高到 86.7%，超过了 OpenAI-o1-0912 的性能。 ● 更长的思考时间：在训练过程中，DeepSeek-R1-Zero 的思考时间持续增加。模型通过利用更长的测试时计算来解决越来越复杂的推理任务，计算范围从生成数百到数千个推理 tokens，从而能够更深入地探索和完善其思考过程。 总的来说，DeepSeek-R1-Zero 通过纯强化学习，在没有监督数据的情况下，实现了强大的推理能力。

Distillation如何提升小模型的推理能力?

Distillation 通过将大型模型的推理能力转移到较小的模型中，从而提升小模型的推理能力。以下是 Distillation 如何实现这一点的详细解释： ● 知识转移：Distillation 涉及使用大型、性能优越的模型（称为"教师"模型）来训练较小的模型（称为"学生"模型）。教师模型已经学习了复杂的推理模式，并且可以生成高质量的推理数据。 ● 训练数据生成：教师模型，例如 DeepSeek-R1，用于生成大量的推理数据，包括解决问题的步骤和最终答案。这些数据随后被用于训练较小的模型。DeepSeek-R1 使用了约 80 万个样本进行提炼。 ● 监督式微调 (SFT)：较小的模型，例如 Qwen 和 Llama 系列，使用由 DeepSeek-R1 生成的数据进行微调。通过这种方式，较小的模型可以学习模仿大型模型的推理过程。 ● 性能提升：经过 distillation 后，较小的模型在推理基准测试中表现出显著的性能提升。例如，DeepSeek-R1-Distill-Qwen-7B 在 AIME 2024 上的得分超过了 GPT-4o-0513。DeepSeek-R1-14B 在所有评估指标上均优于 QwQ-32B-Preview。DeepSeek-R1-32B 和 DeepSeek-R1-70B 在大多数基准测试中显著超过了 o1-mini。 ● 与强化学习的比较：Distillation 是一种比直接在小型模型上应用强化学习 (RL) 更有效的方法。使用数学、代码和 STEM 数据对 Qwen-32B-Base 进行大规模 RL 训练后，得到的 DeepSeek-R1-Zero-Qwen-32B 的性能与 QwQ-32B-Preview 相当，但不如从 DeepSeek-R1 中提炼出的 DeepSeek-R1-Distill-Qwen-32B。这表明，将更强大的模型提炼成较小的模型可以产生出色的结果，而依赖于大规模 RL 的较小模型可能需要巨大的计算能力，甚至可能无法达到 distillation 的性能。 ● 实际应用：Distillation 使得开发具有竞争力的推理能力的小型模型成为可能，这些模型可以在资源受限的环境中部署。通过开放源代码 Distillation 模型（例如 DeepSeek-R1），研究团体可以进一步探索和改进小型模型的推理能力。 总而言之，Distillation 是一种有效的技术，可以通过将大型模型的知识和推理模式转移到较小的模型中，从而提高较小模型的推理能力。

DeepSeek-R1的方法论：强化学习如何驱动推理，有什么局限性?

DeepSeek-R1 通过强化学习 (RL) 驱动推理的方法，以及其局限性如下： 强化学习如何驱动推理能力： ● 多阶段训练流程：DeepSeek-R1 采用了一个包含两个强化学习阶段的训练流程，旨在发现改进的推理模式并与人类偏好对齐。此外，该流程还包含两个 SFT 阶段，为模型的推理和非推理能力提供基础。

● 冷启动数据：为了避免从基础模型开始 RL 训练时早期不稳定的冷启动阶段，DeepSeek-R1 构建并收集了少量的长 CoT（Chain-of-Thought，思维链）数据，以微调模型作为初始 RL actor。 ● 奖励信号与提示分布的结合：DeepSeek-R1 使用奖励信号和多样化的提示分布来训练模型，从而提高模型的帮助性和无害性，同时改进

其推理能力。对于推理数据，遵循 DeepSeek-R1-Zero 中概述的方法，利用基于规则的奖励来指导模型在数学、代码和逻辑推理领域的学习过程。对于通用数据，则采用奖励模型来捕捉复杂场景中人类的偏好。 ● 语言一致性奖励：为了缓解 CoT 中出现的语言混合问题，DeepSeek-R1 在 RL 训练期间引入了语言一致性奖励，该奖励计算为 CoT 中目标语言单词的比例。 ● 从 DeepSeek-R1 到小型密集模型的知识蒸馏：通过使用 DeepSeek-R1 提炼出的数据对 Qwen 和 Llama 等开源模型进行微调，结果表明这种直接的蒸馏方法显著增强了较小模型的推理能力。 DeepSeek-R1 的局限性： ● 通用能力不足：在函数调用、多轮对话、复杂角色扮演和 JSON 输出等任务中，DeepSeek-R1 的能力不如 DeepSeek-V3。 ● 语言混合问题：DeepSeek-R1 目前针对中文和英文进行了优化，这可能导致在处理其他语言的查询时出现语言混合问题。例如，即使查询使用的语言不是英语或中文，DeepSeek-R1 也可能使用英语进行推理和回答。 ● 对提示词敏感：DeepSeek-R1 对提示词比较敏感，少量样本提示会降低其性能。因此，建议用户直接描述问题并使用零样本设置指定输出格式，以获得最佳结果。 ● 软件工程任务的改进有限：由于评估时间较长，影响了 RL 过程的效率，因此大规模 RL 并未广泛应用于软件工程任务。因此，在软件工程基准测试中，DeepSeek-R1 相对于 DeepSeek-V3 并没有表现出巨大的改进。 ● 安全性奖励后的性能下降：在中文 SimpleQA 基准测试中，DeepSeek-R1 的性能不如 DeepSeek-V3，这主要是因为它在安全 RL 之后拒绝回答某些查询。在没有安全 RL 的情况下，DeepSeek-R1 可以达到 70% 以上的准确率。

强化学习在推理任务中的应用有哪些优势?

强化学习在推理任务中具有显著的优势，如之前的研究已经证明了这一点（Shao等人，2024年；Wang等人，2023年）。然而，这些研究严重依赖于监督数据，而收集监督数据需要花费大量时间。因此，我们探索了语言模型通过纯粹的强化学习过程自我进化以发展推理能力的潜力，并在此过程中没有使用任何监督数据。我们将简要概述我们的强化学习算法，然后介绍一些令人兴奋的结果，希望这能为社区提供有价值的见解。

DeepSeek-R1-Zero是通过什么方式实现的?

DeepSeek-R1-Zero是直接应用强化学习（RL）到基础模型上而不需要预先进行监督微调（SFT）。这种做法使模型能够探索链式思维（CoT），解决复杂的问题，并生成长的CoT，实现了自我验证、反思等能力。

DeepSeek-R1-Zero的限制是什么?

DeepSeek-R1-Zero的一个关键限制是其内容通常不适合阅读。响应可能混合多种语言或缺乏markdown格式以突出显示答案供用户查看。相比之下，在为DeepSeek-R1创建冷启动数据时，我们设计了一个可读模式，其中包括每个响应结尾的摘要，并过滤掉不友好的响应。

DeepSeek-R1的训练流程包括哪些阶段?

DeepSeek-R1的训练流程包括四个阶段：数据预处理、模型构建、模型训练和评估。在数据预处理阶段，我们首先收集高质量的数据集，并对其进行清洗和标注；在模型构建阶段，我们设计了一个基于Transformer的多任务学习框架，该框架可以同时完成推理和生成两个任务；在模型训练阶段，我们采用了随机梯度下降（SGD）算法进行优化，并使用Adam作为自适应学习率调整方法；最后，在评估阶段，我们通过计算准确率、召回率、F1值等指标来评估模型性能。

DeepSeek-R1-Zero在训练过程中出现了什么有趣的现象?

DeepSeek-R1-Zero在训练过程中出现了一个"啊哈时刻"，即在模型的中间版本中学习重新评估其初始方法以分配更多思考时间给一个问题的行为。这不仅是模型推理能力不断提高的表现，也是强化学习如何导致意外而复杂结果的一个引人入胜的例子。

在工程导向的编码任务中，OpenAI-o1-1217与DeepSeek-R1相比如何?

在工程导向的编码任务中，OpenAI-o1-1217在Aider上表现优于DeepSeek-R1，但在SWE Verified上的表现相当。

分享这篇文章

相关文章推荐

Pangu Deep Dive...

Pangu 相关论文的深度解析和...

DeepSeek R1 Paper...

A comprehensive review of t...

Cursor Rules 使...

Cursor 的 .cursor/rules 使...