168飞艇全国统一开奖官网

EMCm7DuGMf9lBRLV

168飞艇全国统一开奖官网清华黄高团队：强化学习无法诱发新的推理能力？ | 今日热门论文

速览热门论文

1. 清华黄高团队：强化学习无法诱发新的推理能力

2. NodeRAG：利用异构节点构建基于图形的 RAG

3. 睡眠时计算：有效降低测试时计算要求

4. Meta 提出“感知编码器”，图像视频理解 SOTA

5. 不要蒸馏！CMU 团队提出「反蒸馏采样」

6. WORLDMEM：利用记忆库增强世界模拟

7. 利用「专家失败」提高 agent 微调性能

1. 清华黄高团队：强化学习无法诱发新的推理能力

研究表明，基于可验证奖励的强化学习（RLVR）有望提升大语言模型（LLM）在数学与编程任务中的推理能力。普遍观点认为，RLVR 能够推动模型持续自我优化，从而习得超出其基础模型能力的新型推理能力。

然而，清华大学自动化系黄高副教授团队对此提出了重新审视。他们通过在更大 k 值下测量 pass@k 指标，对该假设进行了系统评估，旨在探究不同模型系列与基准下的推理能力上限。

令人意外的是，RLVR 并未显著引入全新的推理模式。虽然在较小的 k 值（如 k=1）下，RL 训练模型在性能上优于其基础模型，但在更高的 k 值下，基础模型却能取得与 RL 模型相当甚至更优的 pass@k 分数。此外，RL 模型生成的大多数推理路径实际上已包含在基础模型的采样分布中，这表明 RL 模型的推理表现主要源自对基础模型能力的重加权，而非学习到新的能力。

进一步分析显示，RL 训练通过调整模型的输出分布，使其更倾向于采样能获得奖励的路径，从而提升生成正确答案的概率。这一机制在提高效率的同时，也导致模型在推理空间中的覆盖范围变窄。相似现象亦在 RLVR 训练的视觉推理任务中得以观察。

此外，研究还指出，相较于 RLVR，“知识蒸馏”（distillation）更有可能真正向模型注入新的知识，从而拓展其推理能力边界。

这些发现凸显了 RLVR 在提升 LLM 推理能力方面的局限性，促使我们必须从更根本的层面重新审视强化学习在推理能力塑造中的作用，并思考是否亟需更优的训练范式。

论文链接：https://arxiv.org/abs/2504.13837

2. NodeRAG：利用异构节点构建基于图形的 RAG

检索增强生成（RAG）使大语言模型（LLM）能够访问外部和私人语料库，从而在特定领域做出与事实一致的响应。通过利用语料库的固有结构，基于图的 RAG 方法通过建立知识图索引和利用图的结构特性，进一步丰富了这一过程。然而，目前基于图的 RAG 方法很少优先考虑图结构的设计。设计不当的图不仅会阻碍各种图算法的集成，还会导致工作流程不一致和性能下降。

为了进一步释放图在 RAG 中的潜力，来自哥伦比亚大学、宾夕法尼亚大学和里海大学的研究团队提出了 NodeRAG，这是一个以图为中心的框架，引入了异构图结构，可以将基于图的方法无缝、整体地集成到 RAG 工作流中。通过与 LLM 的能力紧密结合，该框架可确保端到端流程的充分内聚和高效。

实验证明，NodeRAG 与 GraphRAG 和 LightRAG 等以前的方法相比，不仅在索引时间、查询时间和存储效率方面具有性能优势，而且在多跳基准和使用最少检索 token 的开放式头对头评估中提供了更强的问题解答性能。

论文链接：https://arxiv.org/abs/2504.11544

3. 睡眠时计算：有效降低测试时计算要求

扩展测试时计算已成为大语言模型（LLM）解决棘手问题的关键要素，但同时也带来了高延迟和高推理成本。

在这项工作中，来自 Letta 和加州大学伯克利分校的研究团队提出了睡眠时计算（sleep-time compute），它允许模型在提出查询之前离线“思考”上下文：通过预测用户可能提出的查询并预先计算有用的数量，有效降低测试时的计算要求。

为了证明这一法的有效性，他们创建了两个推理任务的修改版本——Stateful GSM-Symbolic 和 Stateful AIME。他们发现，在这两个任务上，睡眠时计算可以将达到相同准确度所需的测试时计算量减少约 5 倍；通过调整睡眠时计算的规模，他们可以进一步提高这个两个任务的准确度，分别提高 13% 和 18%。

他们还提出了 Multi-Query GSM-Symbolic，通过在每个上下文中包含多个相关查询来扩展 GSM-Symbolic。通过使用 Multi-Query GSM-Symbolic，在同一上下文的相关查询中分摊睡眠时计算，他们可以将每次查询的平均成本降低 2.5 倍。

此外，他们还进行了其他分析，以了解睡眠时计算何时更有效，结果发现用户查询的可预测性与睡眠时计算的有效性密切相关。最后，他们进行了一项案例研究，将睡眠时计算应用到现实的代理 SWE 任务中。

论文链接：https://arxiv.org/abs/2504.13171

4. Meta 提出“感知编码器”，图像视频理解 SOTA

在这项工作中，Meta 团队提出了感知编码器（PE），这是一种通过简单的视觉语言学习训练出来的 SOTA 图像和视频理解编码器。

传统上，视觉编码器依赖于各种预训练目标，每个目标都是为特定的下游任务（如分类、字幕或定位）定制的。令人惊讶的是，在扩大精心调整的图像预训练方案并使用视频数据引擎进行改进后，他们发现，仅凭视觉语言对比训练就能为所有这些下游任务生成强大、通用的嵌入。唯一需要注意的是：这些嵌入都隐藏在网络的中间层中。为了将它们提取出来，他们提出了两种对齐方法，一种是用于多模态语言建模的语言对齐，另一种是用于密集预测的空间对齐。

连同核心对比检查点，PE 模型系列在各种任务中都取得了 SOTA，包括零样本图像和视频分类与检索；文档、图像和视频问答；以及检测、深度估计和跟踪等空间任务。

论文链接：https://arxiv.org/abs/2504.13181

5. 不要蒸馏！CMU 团队提出「反蒸馏采样」

模型在生成扩展推理轨迹的同时，会无意中产生丰富的 token 序列，从而促进模型的蒸馏。认识到这一点后，模型所有者可能会寻求既能限制提炼效果又不影响模型性能的采样策略。

在这项工作中，卡内基梅隆大学团队提出了“反蒸馏采样”（Antidistillation Sampling），通过策略性地修改模型的下一个 token 概率分布，这一方法可以毒化推理轨迹，使其蒸馏效果降低，同时保留模型的实用性。

论文链接：https://arxiv.org/abs/2504.13146

6. WORLDMEM：利用记忆库增强世界模拟

世界模拟因其能够模拟虚拟环境和预测行动后果而越来越受欢迎。然而，有限的时间上下文窗口往往导致无法保持长期一致性，特别是在保持三维空间一致性方面。

在这项工作中，南洋理工大学 S-Lab 团队提出了 WorldMem，这是一个利用由存储记忆帧和状态（如姿势和时间戳）的记忆单元组成的记忆库来增强场景生成的框架。这一方法采用了一种记忆注意力机制，可以根据记忆帧的状态有效提取其中的相关信息，因此即使在视角或时间存在明显偏差的情况下，也能准确重建之前观察到的场景。

此外，通过在状态中加入时间戳，这一框架不仅能模拟静态世界，还能捕捉其随时间的动态演变，从而在模拟世界中实现感知和互动。在虚拟和真实场景中进行的实验，验证了这一方法的有效性。

论文链接：https://arxiv.org/abs/2504.12369

7. 利用「专家失败」提高 agent 微调性能

大语言模型（LLM）已显示出作为 agent 的巨大潜力，在需要多轮推理和互动的任务中表现出色。拒绝采样微调（RFT）已成为将 LLM 微调为 agent 的有效方法：它首先模仿专家生成的成功轨迹，然后通过对成功的、自我生成的轨迹进行迭代微调，进一步提高 agent 技能。然而，由于专家（如 GPT-4）主要在较简单的子任务上取得成功，而 RFT 本身偏向于较简单的场景，因此许多复杂的子任务仍未解决，长期处于分布外（OOD）状态。

在研究这些具有挑战性的子任务时，来自加州大学洛杉矶分校的研究团队及其合作者发现，之前失败的专家轨迹往往能提供有价值的指导，例如计划和关键行动，从而显著提高 agent 的探索效率并获得关键技能。受这些观察结果的启发，他们提出了“探索专家失败”（EEF），它能从失败的专家轨迹中识别出有益的行动，并将其整合到训练数据集中。潜在的有害行为会被仔细排除，以防止模型学习过程受到污染。通过利用专家失败中的有利行动，EEF 成功解决了一些以前无法解决的子任务，并提高了 agent 微调性能。

值得注意的是，这一方法在 WebShop 中的胜率达到了 62%，超过了 RFT（53. 6%）和 GPT-4（35.6%）。

论文链接：https://arxiv.org/abs/2504.13145

整理：学术君

如需转载或投稿，请直接在公众号内留言

澳洲10二期计划

168澳洲幸运10官网查询结果幸运

澳彩大数据分析软件

澳洲十全天计划网页版

助赢计划永久免费版

澳洲幸运10开奖号码查询

加导师微信一天赚500

澳洲幸运十计划推荐app

澳洲幸运5官方历史查询开奖记录

幸运5稳赢的死方法

飞艇6码倍投计划表

澳洲幸运10官方开奖记录

如何分析澳洲幸运10

澳洲十历史开奖结果

168网官网

澳洲幸运5大数据分析软件

澳洲10期开奖号码

澳洲幸运10精准六码计划

澳洲10公式图