# Kimi k1.5: Scaling Reinforcement Learning with LLMs

2025年2月13日 1 分钟阅读

#强化学习 #LLM #Kimi #k1.5 #AI

> *MoonshotAI开源的Kimi k1.5项目，并对其技术原理、功能特点、应用前景和伦理风险进行了详细解读。*

## 全文摘要

本文介绍了一种新的多模态语言模型（LLM）——Kimi k1.5，该模型通过强化学习（RL）训练得到，并且在多个测试任务上取得了优异的表现。作者采用了长序列到短序列转换技术（CoT），并结合了长期奖励和改进的策略优化方法，使得Kimi k1.5能够有效地探索环境并学习探索行为。实验结果表明，Kimi k1.5在AIME、MATH500、Codeforces和MathVista等任务中均达到了最先进的性能水平，超过了现有的多模态语言模型和单模态语言模型。此外，文章还介绍了如何使用长序列到短序列转换技术提高短序列模型的性能，进一步证明了这种方法的有效性。总之，本文提出的Kimi k1.5是一种具有广泛适用性的高效多模态语言模型，为人工智能的发展提供了新的思路和方向。

## 论文方法

### 方法描述

本文主要介绍了Kimik1.5模型的开发过程，包括预训练、Vanilla supervised fine-tuning（SSFT）、long-CoT supervised fine-tuning（LCoT-SFT）以及reinforcement learning（RL）四个阶段。其中，预训练采用了基于大规模中文文本语料库的transformer模型，并在大规模数据集上进行了fine-tuning；SSFT则是在大规模标注数据集上进行的supervised fine-tuning；LCoT-SFT则是通过构建小而高质

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#强化学习 #LLM #Kimi #AI

量的long-CoT warm-up数据集来进一步提升模型的表现；最后，RL则是通过与奖励信号相结合来进行强化学习，以进一步提高模型的性能。

## 方法改进

本文的主要贡献在于提出了新的方法来提高Kimik1.5模型的性能。具体来说，本文采用了以下几种方法：

使用自动过滤器来筛选问题并生成高质量的prompt集，从而避免了奖励黑客现象； 在long-CoT warm-up数据集中使用prompt工程来构造一个小型但高质量的数据集； 使用链式思考（CoT）方法来生成多个答案候选，并从中选择最佳答案； 使用长2短技术将长-CoT模型的知识转移给短-CoT模型。

## 解决的问题

本文解决了以下几个问题：

如何构建高质量的prompt集，以便更好地指导模型的学习？ 如何利用chain-of-thought方法生成多个答案候选，并从中选择最佳答案？ 如何将长-CoT模型的知识转移到短-CoT模型中，以提高模型的性能?

## 论文实验

本文主要介绍了三组实验，分别是模型性能比较、长上下文扩展训练、长到短迁移学习和ablation研究。

首先，在模型性能比较中，作者对Kimi k1.5长CoT和短CoT模型进行了测试，并与其他开源和专有模型进行了比较。结果显示，Kimi k1.5模型在文本、视觉和推理挑战方面表现出色，特别是在自然语言理解、数学、编码和逻辑推理方面具有显著优势。此外，作者还展示了通过长上下文扩展训练可以提高模型的长期推理能力，而长到短迁移学习可以通过使用强化学习算法来优化模型的短语效率。

其次，在长上下文扩展训练中，作者通过对小规模模型进行训练，观察了训练准确率和响应长度随迭代次数的变化趋势。结果表明，随着训练的进行，模型的响应长度和性能准确性都会增加，更复杂的挑战问题会呈现出更快的增长速度。同时，作者还发现模型输出上下文长度与解决问题的能力之间存在强相关性。最后，作者将模型扩展到了128k上下文长度，并观察到其在困难推理挑战方面的持续改进。

第三，在长到短迁移学习中，作者比较了四种不同的方法，包括DPO、最短拒绝采样、模型合并和提出的RL算法。结果表明，提出的RL算法在短模型上表现最佳，具有最高的词效性和性能。此外，所有Kimi k1.5系列模型都比其他模型更具优越性。

最后，在ablation研究中，作者探讨了模型大小和上下文长度的扩展以及使用负梯度的影响。结果表明，虽然更大的模型在开始时表现更好，但较小的模型可以通过利用经过RL优化的更长的上下文来实现相当的性能。然而，较大的模型通常比较小的模型具有更好的词效性。此外，作者还证明了使用ReST作为策略优化算法的效果不如使用负梯度的方法，而且选择适当的优化策略对于生成长上下文至关重要。最后，作者还演示了使用不同采样策略对模型性能的影响，结果表明使用课程采样方法可以显著提高模型的性能。

table_2

table_3

# 论文总结

## 文章优点

本文提出了一种训练大规模多模态预训练模型的方法，并在多个任务上进行了验证。该方法采用了优化长上下文（long-context）RL训练策略，结合了部分卷积等技术，实现了更高效的长期记忆。此外，作者还提出了多种改进的技巧，如结合多项式近似来实现在线反向传播等，进一步提高了模型性能。最后，作者通过实验表明，使用这种方法可以显著提高模型的性能，甚至不需要使用更复杂的技巧，如蒙特卡罗树搜索、值函数等。

## 方法创新点

本文的主要贡献在于提出了一种有效的训练大规模多模态预训练模型的方法，包括以下几个方面的创新：

优化长上下文（long-context）RL训练策略：作者通过对不同长度的上下文进行实验，发现较长的上下文能够带来更好的效果。因此，他们设计了一个基于深度强化学习的训练框架，以最大化预测下一个标记的概率为目标，同时考虑上下文的长度。 部分卷积（partial rollout）：由于计算资源有限，无法处理整个上下文序列，因此作者采用了部分卷积的技术，只考虑部分上下文序列，从而减少了计算量。 改进的技巧：除了上述主要方法外，作者还提出了一些其他技巧，如结合多项式近似来实现在线反向传播等，这些技巧都有助于提高模型的性能。

## 未来展望

尽管本文提出的训练方法已经取得了很好的效果，但仍然存在一些挑战和未来的研究方向。例如，如何更好地利用多模态数据，以及如何将这种方法扩展到更大规模的数据集上。此外，还需要进一步

研究如何将这种方法应用于不同的自然语言处理任务中，以便更好地发挥其潜力。总之，本文提出的训练方法为大规模多模态预训练模型的发展提供了新的思路和方法，具有重要的理论和实践意义。

# 论文10个问题及其回答

## 1. Kimi k1.5的核心创新是什么?

**回答**：Kimi k1.5的核心创新包括：

**长链式思维（long-CoT）扩展**：通过将上下文窗口扩展到128k tokens，显著提升了推理能力和模型性能。

**强化学习优化**：采用在线镜像下降算法、长度惩罚、采样策略等方法，优化了强化学习的效率和效果。

**多模态整合**：模型同时在文本和视觉数据上进行训练，实现跨模态推理能力。

**long2short方法**：将长链式模型的推理能力迁移到短链式模型中，提高了令牌效率和推理性能。

## 2. Kimi k1.5与现有模型相比的主要优势是什么?

**回答**：

在长链式推理任务中，Kimi k1.5在多个基准测试上达到了最先进的性能（如AIME、MATH-500等）。

在短链式推理任务中，Kimi k1.5显著超越了GPT-4o和Claude Sonnet 3.5等模型，性能提升高达550%。

通过强化学习和多模态训练，Kimi k1.5在文本和视觉任务中均表现优异。

## 3. 论文中提到的long2short方法如何优化短链式模型的性能?

**回答**： long2short方法通过以下步骤优化短链式模型：

**模型融合**：将长链式模型与短链式模型的权重平均融合。

**最短拒绝采样**：对同一问题采样多次，选择最短的正确解作为训练数据。

**DPO（直接偏好优化）**：生成正负样本对，通过偏好学习优化短链模型。

**long2short强化学习**：在标准强化学习的基础上引入长度惩罚，进一步提升令牌效率。

## 4. 强化学习中的部分回滚技术（Partial Rollouts）是如何工作的?

**回答**： 部分回滚技术通过以下方式优化长链式强化学习：

**固定输出令牌预算**：限制每次回滚的长度，超出部分存储到回放缓冲区中。

**分段处理**：长链式推理被分成多个片段，异步操作以提高计算效率。

**重复检测**：识别生成内容中的重复片段，提前终止计算并施加惩罚。

**缓冲区复用**：历史片段可重复使用，减少重新生成的计算开销。

## 5. 强化学习的提示集（Prompt Set）是如何构建的?

**回答**： 提示集的构建遵循以下原则：

**多样性覆盖**：涵盖STEM、编程、通用推理等多个领域。

**难度平衡**：包含易、中、难三种难度的题目。

**可评估性**：确保提示能够通过验证器客观评估，避免奖励黑客问题。 数据过滤和标签系统用于保证提示集的质量和多样性。

## 6. Kimi k1.5在多模态任务中的表现如何?

**回答**： Kimi k1.5在多模态任务中表现优异，例如：

在视觉推理基准（如MathVista和MATH-Vision）上，达到了最先进的准确率。

在文本-视觉任务中，模型展现了强大的跨模态推理能力。

通过OCR、图像-文本交错数据和合成数据，增强了模型的视觉理解能力。

## 7. Kimi k1.5的训练分为哪些阶段?

**回答**： Kimi k1.5的训练分为以下三个阶段：

**视觉-语言预训练阶段**：先单独训练语言模型，再逐步引入多模态数据。

**冷却阶段**：使用高质量数据进行微调，提升数学推理、知识问答和代码生成能力。

**长上下文激活阶段**：逐步扩展最大序列长度至131,072 tokens，以支持长文本推理任务。

## 8. 论文中提到的采样策略如何提高训练效率?

**回答**： 论文提出了两种采样策略：

**课程采样**：从简单任务开始训练，逐步过渡到复杂任务，提高初期训练效率。

**优先级采样**：根据模型在每个问题上的成功率调整采样概率，优先训练模型表现较差的问题。

## 9. Kimi k1.5在数学和代码任务中的表现如何?

**回答**： 在数学和代码任务中，Kimi k1.5表现卓越：

在MATH-500基准上，长链模型达到96.2的准确率，显著优于其他模型。

在LiveCodeBench基准上，短链模型达到47.3的Pass@1分数，超越大多数开源和专有模型。

自动化测试用例生成和奖励建模进一步提升了强化学习的效果。

## 10. 未来的研究方向有哪些?

**回答**： 论文指出以下未来研究方向：

提高长上下文强化学习的效率和可扩展性。

改进奖励分配策略，减少过度思考现象。

将long2short方法与长链式强化学习结合，迭代优化性能和令牌效率。

探索多模态任务中的更多应用场景。

## 参考

Kimi-k1.5

分享这篇文章

相关文章推荐

### OpenAI Model Sp...

OpenAI Model Spec 解读

### 字节跳动 OmniHu...

字节跳动开源的 OmniHuman-1...

### 计算机使用代理

计算机使用代理