DeepScaleR_阅读

- 模型
  - 基座：Deepseek-R1-Distilled-Qwen-1.5B
  - 开源模型地址：https://huggingface.co/agentica-org/DeepScaleR-1.5B-Preview
- 数据
  - 数据量：40k
  - 来源：we compiled AIME problems from 1984-2023 and AMC problems prior to 2023, along with questions from the Omni-MATH and Still datasets, which feature problems from various national and international math competitions.
  - 数据预处理（答案抽取/去重/格式过滤）
    - Extracting Answers: For datasets such as AMC and AIME, we use `gemini-1.5-pro-002` to extract answers from official AoPS solutions.
    - Removing Redundant Questions: We employ RAG with embeddings from `sentence-transformers/all-MiniLM-L6-v2` to eliminate duplicate problems. To prevent data contamination, we also check for overlaps between the training and test sets.
    - Filtering Ungradable Questions: Some datasets, such as Omni-MATH, include problems that cannot be evaluated using `sympy` and require an LLM judge. Since using LLM judges may slow down training and introduce noisy reward signals, we apply an additional filtering step to remove these ungradable questions.
  - 开源数据地址：https://huggingface.co/datasets/agentica-org/DeepScaleR-Preview-Dataset
- 创新方法
  - we leverage a distilled model and introduce a novel iterative lengthening scheme for R
    - 为了加速收敛，降低训练成本，基座使用R1蒸馏的模型
    - 添加了长度限制
  - 在分析基座模型在AIME2024集合上的效果的时候(长的回复导致错误率提升/并且对CoT推理没有帮助)
    - On average, incorrect responses contained three times more tokens than correct ones (20,346 vs. 6,395)
    - Additionally, we observed in our evaluation logs that lengthy responses exhibit repetitive patterns, indicating that they do not contribute meaningfully to effective chain-of-thought (CoT) reasoning.
    - 尝试思路
      - Given this insight, we initiated training with an 8K context, achieving an initial AIME2024 accuracy of 22.9%—just 6% below the original model. This strategy proved effective: Over the course of training, mean training rewards increased from 46% to 58%, while average response length dropped from 5,500 to 3,500 tokens .
        - 疑问：这种是否是直接SFT的？因为效果下降了

        - More importantly, constraining output to 8K tokens led the model to utilize context more effectively. As shown in the table, our model generates significantly shorter responses for both correct and incorrect answers while surpassing the base model's AIME accuracy by 5%—with only one-third of the tokens.
            - 疑问：这里应该是用了RL，效果提升了5%，并且回复长度明显低于base
    - 在添加到16k上下文长度过程中，发现clip ratio增加了，表明模型输出的长度增加了
        - After approximately 1,000 steps, an interesting shift occurs for our 8K run: response length begins to increase again. However, this leads to diminishing returns—accuracy plateaus and eventually declines. At the same time, the response clipping ratio rises from 4.2% to 6.5%, indicating that more responses are being truncated at the context limit.
        - 尝试思路
            - 在1040步将max leangth设置为16k，效果提升至38%，训练了500steps
    - 在跑到1040+480=1520steps时候，将上下文长度添加到24k，200steps后达到43%准确率
        - After training an additional 500 steps on 16K context, we noticed performance beginning to plateau—mean training rewards converged at 62.5%, AIME Pass@1 accuracy hovered around 38%, and response length started to decline again. Meanwhile, the maximum response clipping ratio crept up to 2%.
        - To make the final push towards o1-level performance, we decided to rollout the 24k magic— increasing the context window to 24K. We take our 16K run's checkpoint at step 480, and relaunch a training run with 24K context window.
        - With the extended context window, the model finally broke free. After around 50 steps, our model finally surpass 40% AIME accuracy and eventually reaches 43% at step 200. The *24K magic* was in full effect!
- 训练方法
    - Reward
        - `1` – If the LLM's answer passes basic LaTeX/Sympy checks.
        - `0` – If the LLM's answer is incorrect or formatted incorrectly (e.g. missing `<think>`, `</think>` delimiters).
    - 训练步骤（先短后长）
        - First, we perform RL training with 8K max context for more effective reasoning and efficient training.
        - Next, we scale up training to 16K and 24K contexts so that the model can solve more challenging, previously unsolved problems.
    - 训练日志：https://wandb.ai/mluo/deepscaler-1.5b
- 实验中的关键想法
    - RL scaling can manifest in small models as well. `Deepseek-R1` demonstrates that applying RL directly on small models is not as effective as distillation. Their ablations shows that RL on `Qwen-32B` achieves 47% on AIME, whereas distillation alone reaches 72.6%. A common myth is that RL scaling only benefits large models. However, with high-quality SFT data distilled from larger models, smaller models can also learn to reason more effectively with RL. Our results confirm this: RL scaling improved AIME accuracy from 28.9% to 43.1%! These findings suggest that neither SFT nor RL alone is sufficient. Instead, by combining high-

quality SFT distillation with RL scaling, we can truly unlock the reasoning potential of LLMs.

    - **Iterative lengthening enables more effective length scaling.** Prior works [1, 2] indicate that training RL directly on 16K context yields no significant improvement over 8K, likely due to insufficient compute for the model to fully exploit the extended context. And a recent work [3] suggests longer response lengths consists of redundant self-reflection that leads to incorrect results. Our experiments are consistent with these findings. By first optimizing reasoning at shorter contexts (8K), we enable faster and more effective training in subsequent 16K and 24K runs. This iterative approach grounds the model in effective thinking patterns before scaling to longer contexts, making RL-based length scaling more efficient.

- 借鉴的工作
    - DeepScaleR surpasses recent academic works such as rSTAR, Prime, and SimpleRL, which are finetuned from 7B models.
- 原始文章路径
    - https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2