RL is Almost All Your Need in Post-Training

Intro

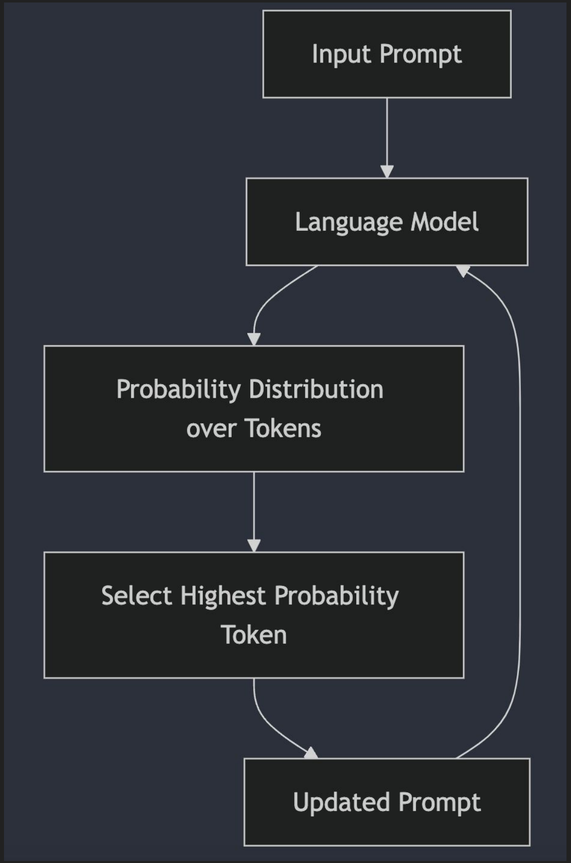
- Discuss the recent popular [DeepSeek-R1]() model
  - It is capable to doing reasoning and significantly boost foundation model's performance built upon on standard benchmark datasets.
  - It produces results close to the openai-ai o1 model.
  - It is open source (model weights) as compared to OpenAI's closed source models.
  - It shows that RL is ALMOST enough to boost foundation model performance in post training.
  - It shows that small models distilled from DeepSeek-R1 perform exceptionally well on benchmarks.

Main Topics

- Language Models (LM)
- Reinforcement Learning (RL)
- DeepSeek-R1

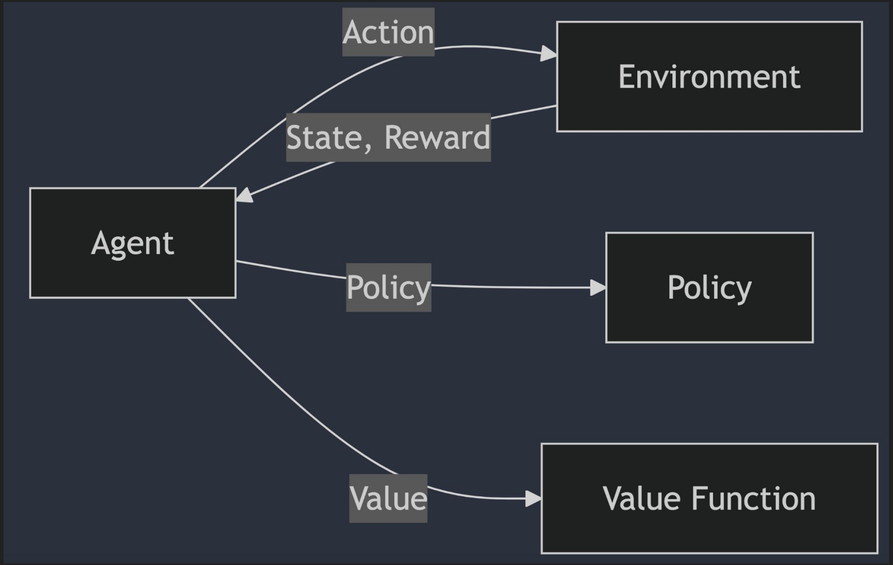
Language Models

- LMs are generative models with a simple objective: to predict the next token given an input prompt.
- Given a prompt, an LM generates a probability distribution over the next possible tokens.
- The token with the highest probability is selected and fed back into the LM as part of the prompt, and the process is repeated to generate text.


Input Prompt
Language Model
Probability Distribution over Tokens
Select Highest Probability Token
Updated Prompt

Reinforcement Learning (RL)

- **Agent**: The learner/decision-maker that interacts with the environment and takes actions
- **Environment**: The external system the agent operates in, which provides states and rewards
- **State**: The current situation/configuration the agent observes in the environment
- **Action**: The possible moves/decisions an agent can make in a given state
- **Reward**: The numerical feedback signal indicating how good an action was
- **Policy**: The strategy/mapping from states to actions that defines the agent's behavior
- **Value Function**: Estimates the expected future reward from being in a state or taking an action



DeepSeek-R1-Zero

- DeepSeek-V3 + Reinforcement Learning => DeepSeek-R1-Zero
    - Reinforcement learning itself may be sufficient to make a large enough foundational model (see this) able to achieve **self-verification, reflection, and generate long CoT**.
    - With a good foundational model, such as DeepSeek-V3, this approach is based to achieve performance similar to openai-o1 model on some reasoning tasks.
    - Choice of reinforcement learning algorithm may not matter that much based on limited study so far.
    - The "aha moment" moment mentioned in the paper reminds that in Psychology, there is a the "aha!" experience is defined as the sudden, unexpected comprehension of a solution to a problem that comes with an ease of understanding the solution. **This fact that purely RL is able develop learning behavior similar to human learning is remarkable, if there is no such pattern in training data.**
    - On a separate side, people has reported that they are able to reproduce DeepSeek-R1-Zero (self verification, search ability, aha moment) in CountDown Game (a reasoning task) using only 3B base LM, and it only costs < $30. **This shows RL is the right direction for boosting reasoning ability for base model and it does not cost much if the base model is small over a few reasoning tasks.**

DeepSeek-R1

- DeepSeek-V3 + Cold Start + Reinforcement Learning => DeepSeek-R1
    - The cold start refers to teaching the basic CoT thought process and other constraints in order to avoid the readability and language fixed issue observed by just using reinforcement learning itself.
    - DeepSeek R1 is close to openai-o1's performance, but overall may still perform a bit worse. However, it is probably the model with closest performance to openai-o1 in term of reasoning skills.
    - It is interested to see how the training data for cold start and RL is collected. May be one approach is construct representative samples for the corresponding reasoning tasks, query closed source API to get the responses (with CoT and final answer in them) and use them to train the model - this is similar to what distillation is doing.
    - Small open source models, such as Llama and Qwen, fine-tuned from DeepSeek-R1 with reasoning samples is able to achieve significant performance boost on some reasoning tasks compared to close source large models - this opens door to many applications with requires running smaller model with limited hardware resources.

Model-Hardware Co-Design For Efficiency Purpose

- To train the foundational model DeepSeek-V3, it only used 2048 GPUs and 11x less GPU hours compared to LLama 3 405B with a larger model size 671B.
- This significant less GPU resource usage indicates large amount of GPU resources (10K+) may not always be needed to training large foundational model. We can invest more on designing the current model architecture to make it hardware friendly (see The Hardware Lottery) and optimizing software (e.g optimized CUDA kernels) in order to achieve similar performance compared to other SOTA models.
- This current approach to invest heavily to buy a lot of hardwares (such as GPUs) for model training seems to be overkilled (see The Short Case for NVDA Stock). We also need to invest more on engineers who excel in designing model architecture and make the software run really efficiently in corresponding hardwares (model-hardware co-design)

References

- DeepSeek-R1 Tech Report
- DeepSeek-V3 Tech Report
- The Hardware Lottery
- Notes on Deepseek r1: Just how good it is compared to OpenAI o1
- DeepSeek R1-Zero in the CountDown game
- “aha!” experience
- The Short Case for NVDA Stock