DeepSeek QA

Utrecht University

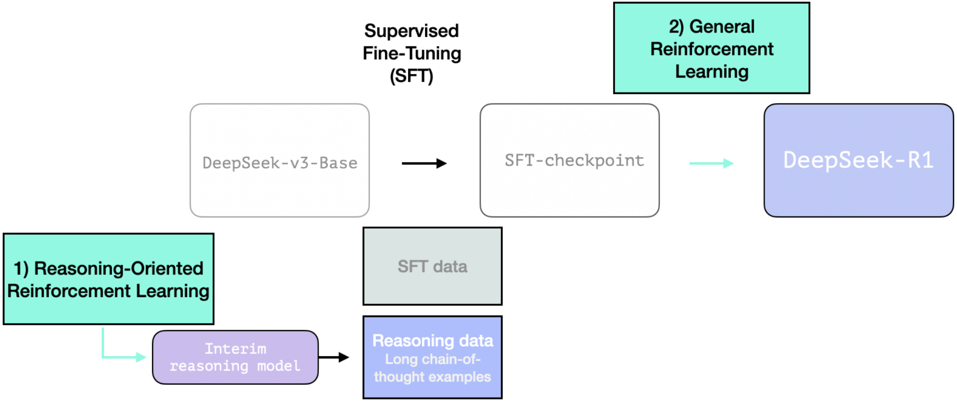
February 27, 2025

Universiteit Utrecht

Overview

- Training recipe
- Notable details
- Relevance & Discussion

DeepSeek Training recipe



Supervised Fine-Tuning (SFT)

2) General Reinforcement Learning

DeepSeek-v3-Base

SFT-checkpoint

DeepSeek-R1

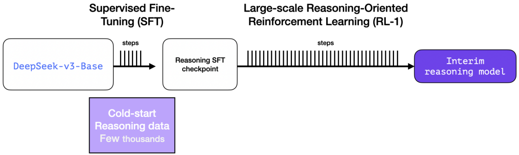
1) Reasoning-Oriented Reinforcement Learning

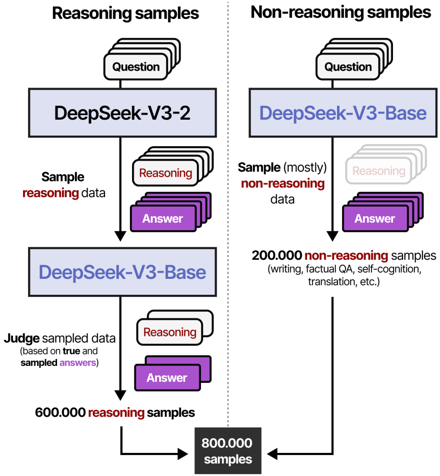
SFT data

Interim reasoning model

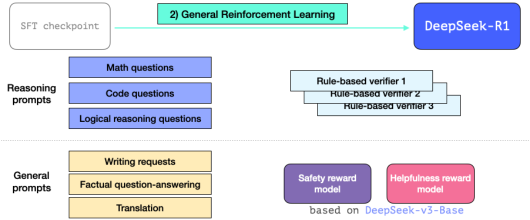
Reasoning data
Long chain-of-thought examples

Training: step 1 & 2


Supervised Fine-Tuning (SFT)
Large-scale Reasoning-Oriented Reinforcement Learning (RL-1)
DeepSeek-v3-Base
steps
Reasoning SFT checkpoint
steps
Interim reasoning model
Cold-start Reasoning data Few thousands


Reasoning samples
Non-reasoning samples
Question
DeepSeek-V3-2
Sample reasoning data
Reasoning
Answer
DeepSeek-V3-Base
Judge sampled data (based on true and sampled answers)
Reasoning
Answer
600.000 reasoning samples
Question
DeepSeek-V3-Base
Sample (mostly) non-reasoning data
Reasoning
Answer
200.000 non-reasoning samples (writing, factual QA, self-cognition, translation, etc.)
800.000 samples

RL process



SFT checkpoint

2) General Reinforcement Learning

DeepSeek-R1

Reasoning prompts

Math questions

Code questions

Logical reasoning questions

Rule-based verifier 1

Rule-based verifier 2

Rule-based verifier 3

General prompts

Writing requests

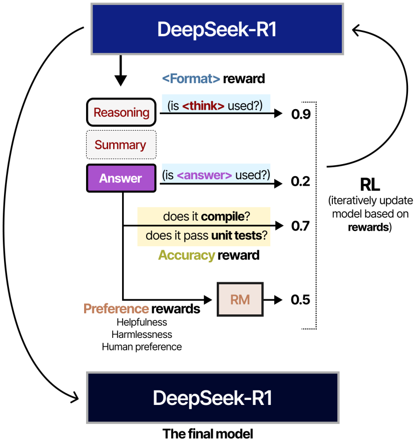
Factual question-answering

Translation

Safety reward model

Helpfulness reward model

based on DeepSeek-v3-Base



DeepSeek-R1

<Format> reward

Reasoning

(is <think> used?)

0.9

Summary

Answer

(is <answer> used?)

0.2

does it compile?

does it pass unit tests?

0.7

Accuracy reward

RL

(iteratively update model based on rewards)

Preference rewards

Helpfulness

Harmlessness

Human preference

RM

0.5

DeepSeek-R1

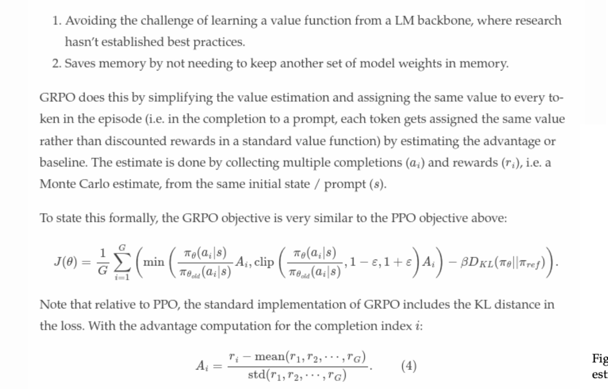
The final model

Details: Data?

What does the RL data look like?

1. Unreleased by DeepSeek
2. Reproduced by Open-R1 (and openly released).
3. https://huggingface.co/datasets/open-r1/OpenThoughts-114k-math

Details: GRPO



1. Avoiding the challenge of learning a value function from a LM backbone, where research hasn't established best practices.
2. Saves memory by not needing to keep another set of model weights in memory.

GRPO does this by simplifying the value estimation and assigning the same value to every token in the episode (i.e. in the completion to a prompt, each token gets assigned the same value rather than discounted rewards in a standard value function) by estimating the advantage or baseline. The estimate is done by collecting multiple completions (a_i) and rewards (r_i), i.e. a Monte Carlo estimate, from the same initial state / prompt (s).

To state this formally, the GRPO objective is very similar to the PPO objective above:

$$J(\theta) = \frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(a_i|s)}{\pi_{\theta_{old}}(a_i|s)}A_i, \text{clip}\left(\frac{\pi_\theta(a_i|s)}{\pi_{\theta_{old}}(a_i|s)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta D_{KL}(\pi_\theta||\pi_{ref})\right).$$

Note that relative to PPO, the standard implementation of GRPO includes the KL distance in the loss. With the advantage computation for the completion index i:

$$A_i = \frac{r_i - \text{mean}(r_1, r_2, \cdots, r_G)}{\text{std}(r_1, r_2, \cdots, r_G)}. \qquad (4)$$

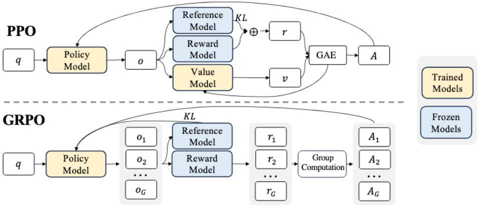


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

Relevance

Can our projects be relevant in this paradigm?

1. Downsampling SFT and reasoning data = data pruning?
2. GRPO: alternative distance functions to KL divergence.
3. Alternatives to coding as reasoning tasks?

Discussion

References:

1. https://arxiv.org/abs/2412.19437
2. https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1
3. https://www.interconnects.ai/p/deepseek-r1-recipe-for-o1
4. https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms