LLM Training & Reinforcement Learning.

Pierre Lepagnol

2025-02-14

Sommaire

Qu'est-ce le Reinforcement Learning ?

Définitions & Basics

Un **agent** qui interagit avec l'**environnement** en prenant des **actions** de manière à maximiser la fonction de **valeur** (sommes des récompenses).

L'objectif du RL: apprendre la **politique optimale** qui maximise les récompenses futures par **essais et erreurs**.

Composants essentiels du RL: Actions

- **Discrètes** : Exemple - LLM, où chaque token est une action est une action spécifique.
- **Continues** : Exemple - contrôler un robot, où bouger le bras méchanique est une action continue.

Politique (π) & Fonction de valeur (V)

Politique

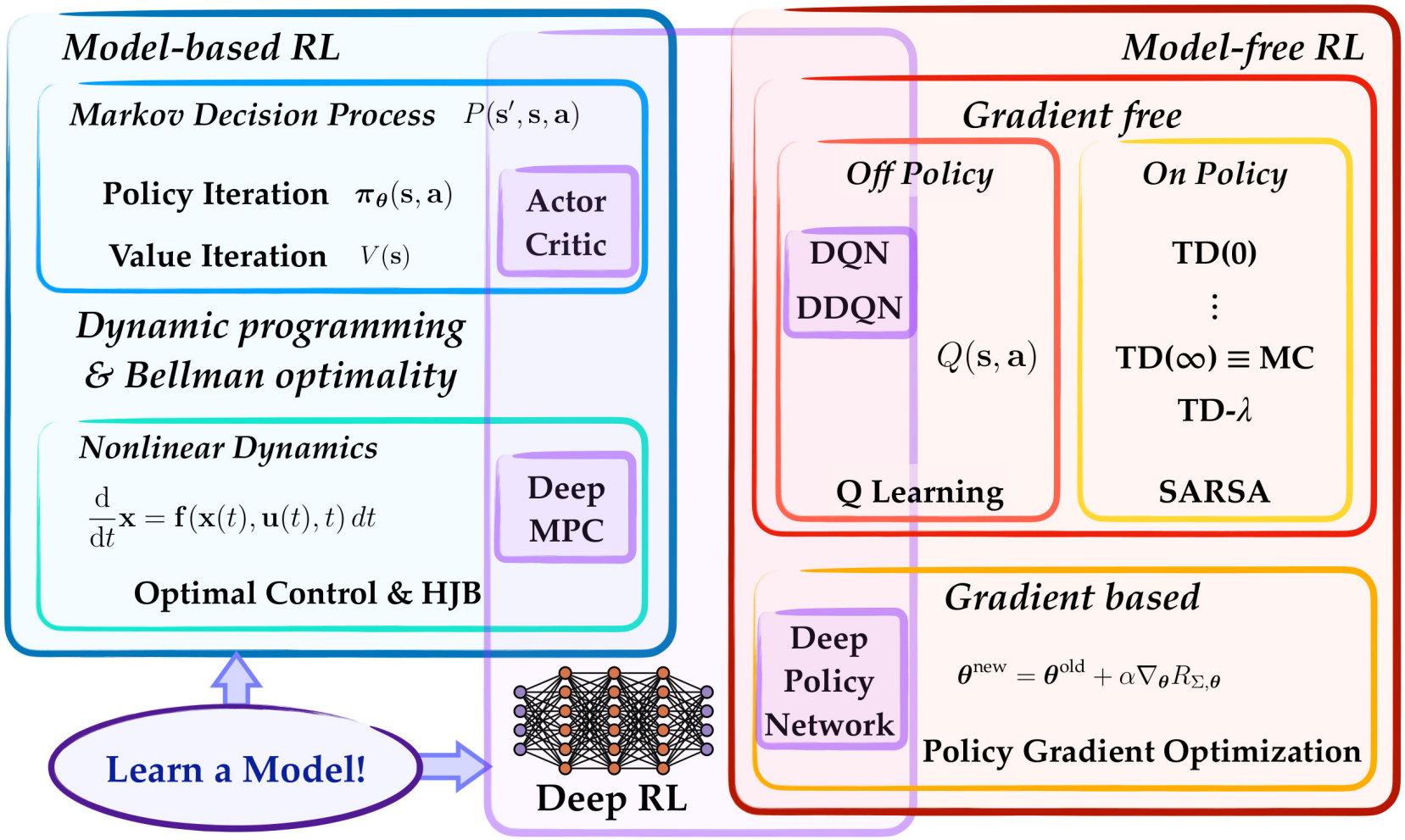
- **Stratégie** (ensemble de règles) que l'agent suit pour décider quelle action entreprendre dans un état donné: Exemple : Réseau de neurones, une liste de If-then-Else, etc
- Peut être **déterministe** ou **probabiliste**.
- Formalisme : $\pi(a|s)$, ce qui signifie la probabilité de prendre l'action **a** dans l'état **s**.

Fonction de valeur (V)

- Souvent la somme des **Somme des récompenses futures**
- Mesure à quel point il est **bon** d'être dans un état particulier.

Rappel Objectif: apprendre la **politique optimale** qui maximise les récompenses futures par **essais et erreurs**.

Differents algorithmes de RL



- Policy Gradient Optimization : TRPO, PPO, GRPO

Application du RL au LLMs

Application du RL au LLMs

Le RL est utilisé à 2 moments de l'apprentissage d'un LLM :

- Pre-training: Phase nécéssitant le plus de compute et de données.
- Post-Training (RL-HF/AI): Phase d'alignement pour assurer

Exemple de Deepseek-R1

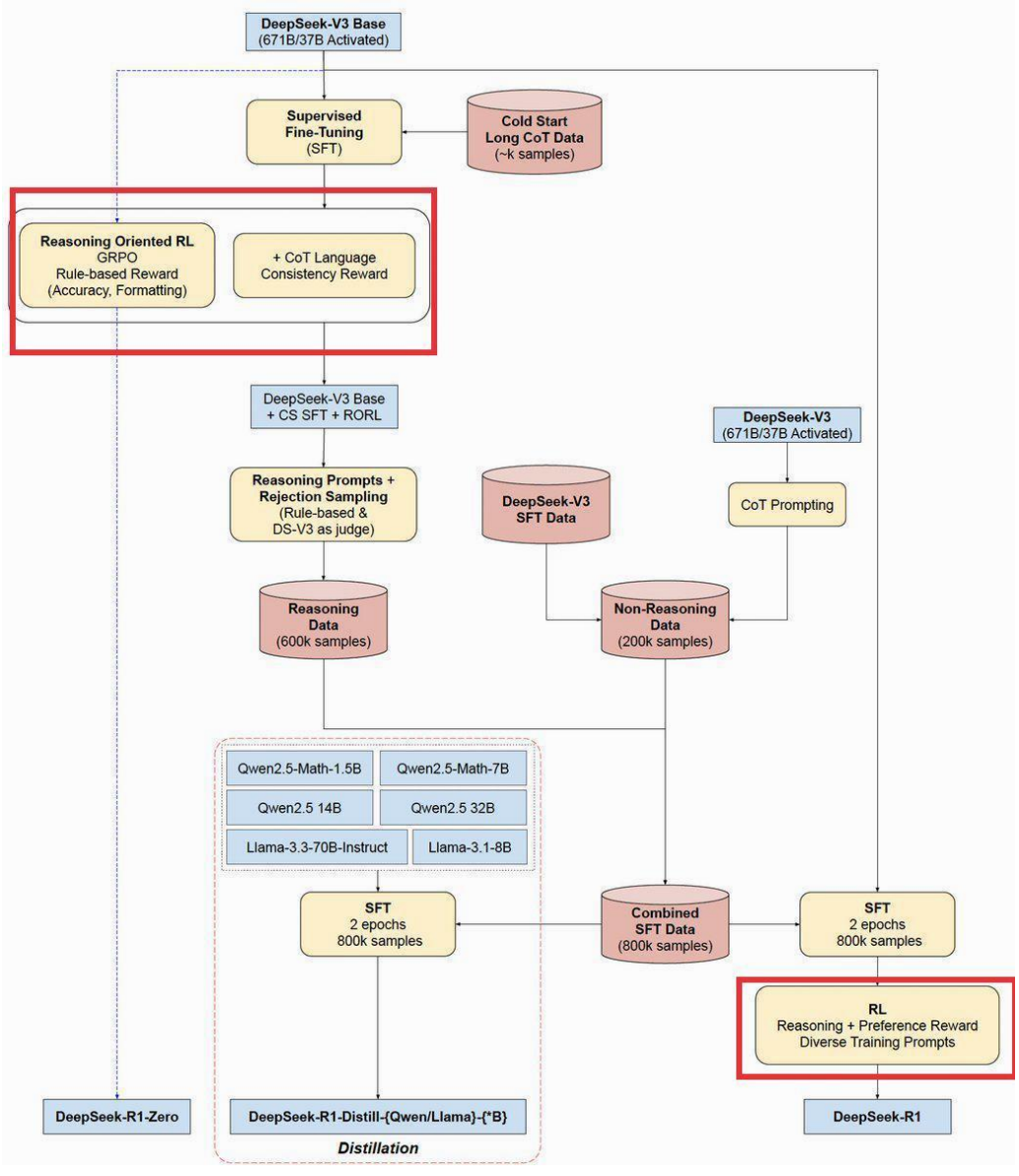


Figure 1: TrainingPipeline de DeepSeek R1

Différents Algo

- TRPO: Trust Region Policy Optimization
- PPO: Proximal Policy Optimization
- GRPO: Group Relative Policy Optimization

GRPO (Group Relative Policy Optimization) is a method used in reinforcement learning (RL) to help a model learn better by comparing different actions and making small, controlled updates using a group of observations

RL during Post-Training

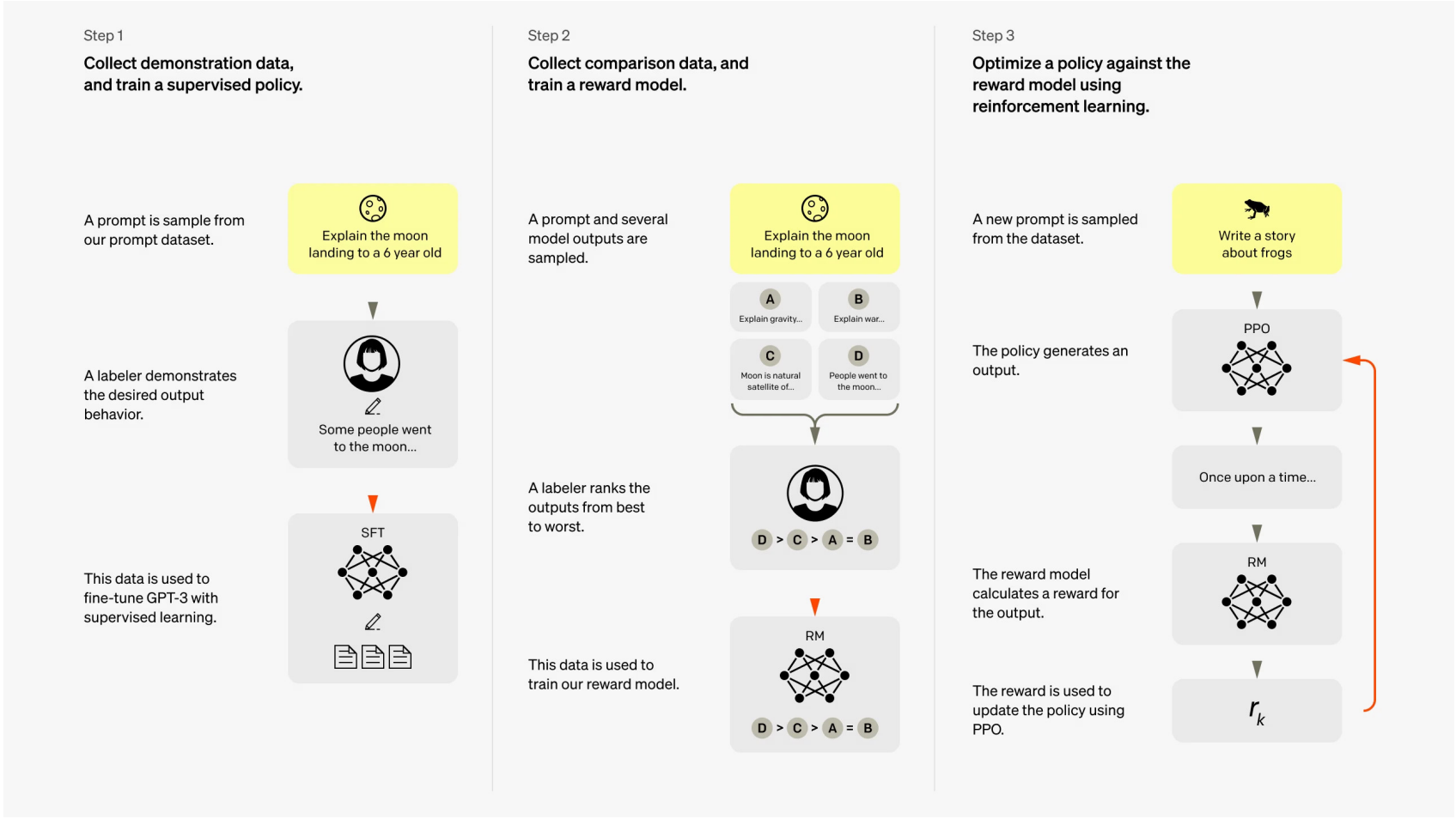


Figure 2: Pipeline Instruct GPT: RLHF

Alternative au RL: (DPO)

RL during Pre-Training

- GRPO is an improved version of Proximal Policy Optimization (PPO).

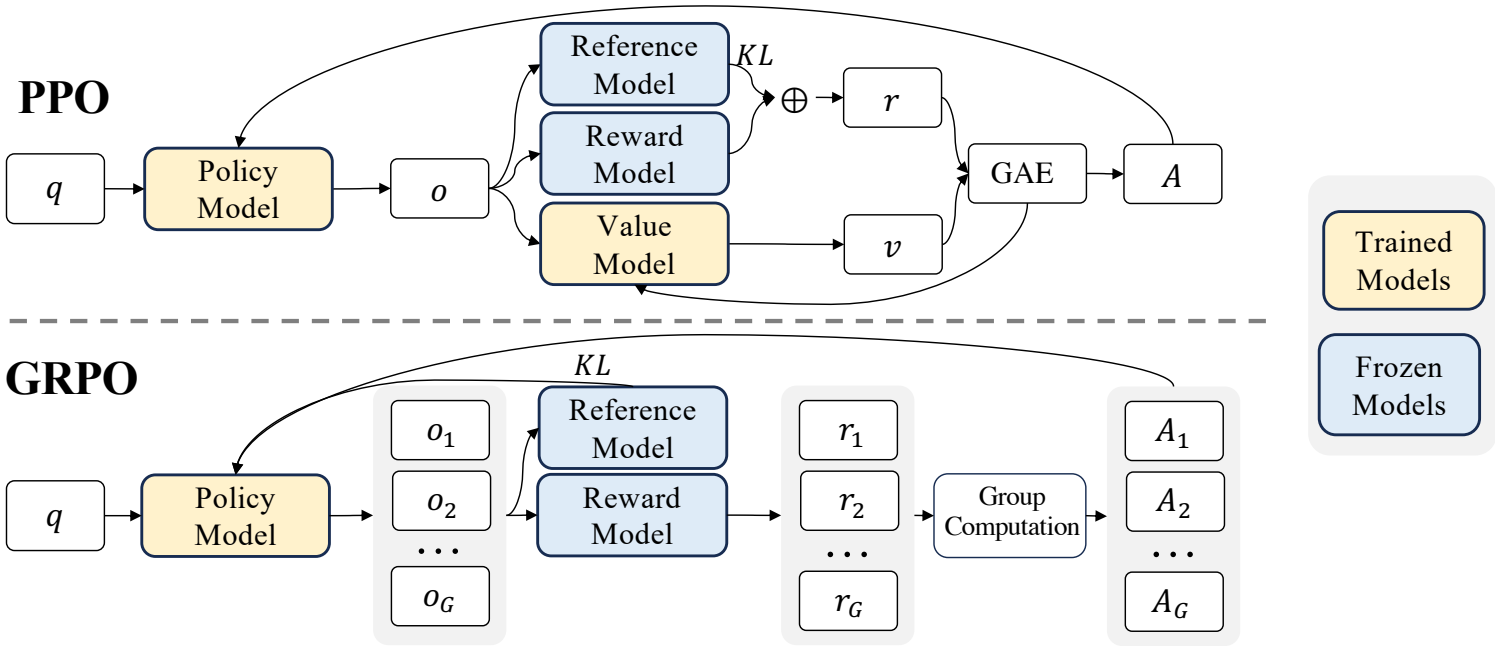


Figure 3: Comparaison PPO & GRPO issue de DeepSeekMath