

# Entrainement de LLM & Reinforcement Learning.

---

Pierre Lepagnol

2025-02-14

Un peu de contexte

Qu'est-ce le Reinforcement Learning ?

Application du RL au LLMs

RL Post-Training: Alignement

RL Pre-Training :

## Un peu de contexte

---

---

## Training language models to follow instructions with human feedback

---

Long Ouyang\*   Jeff Wu\*   Xu Jiang\*   Diogo Almeida\*   Carroll L. Wainwright\*  
Pamela Mishkin\*   Chong Zhang   Sandhini Agarwal   Katarina Slama   Alex Ray  
John Schulman   Jacob Hilton   Fraser Kelton   Luke Miller   Maddie Simens  
Amanda Askell<sup>†</sup>   Peter Welinder   Paul Christiano<sup>\*†</sup>  
Jan Leike\*   Ryan Lowe\*

OpenAI

- Papier de recherche : 4 Mars 2022
- ChatGPT : 30 Novembre 2022

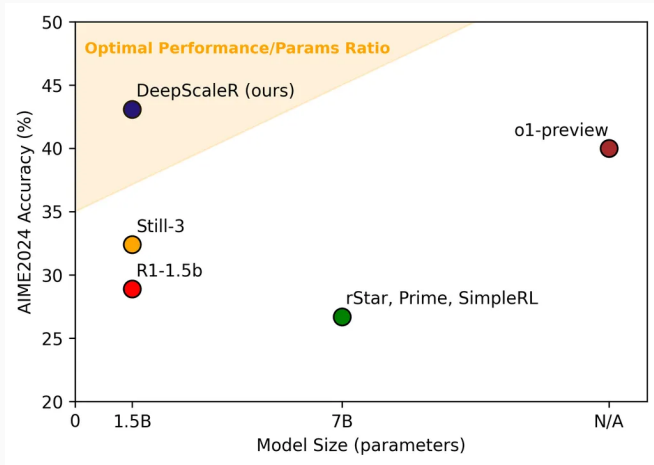


Downloads last month  
3,468,420



# Alternative OpenSource DeepScaleR

DeepScaleR: Surpassing O1-Preview with a 1.5B Model by Scaling RL

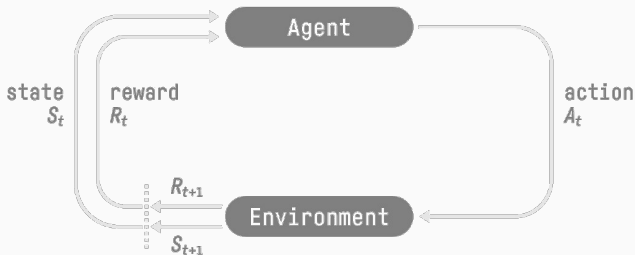


# Qu'est-ce le Reinforcement Learning ?

---

# Définitions & Basics

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

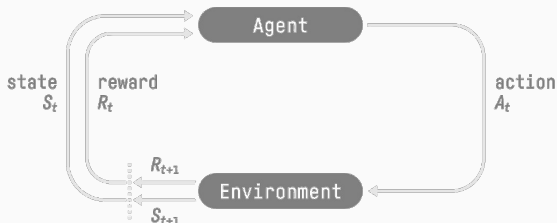


## Objectif du RL

Apprendre la **politique optimale** qui maximise les récompenses futures par **essais/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 mécanique est une action continue.

# Politique ( $\pi$ ) & 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**.

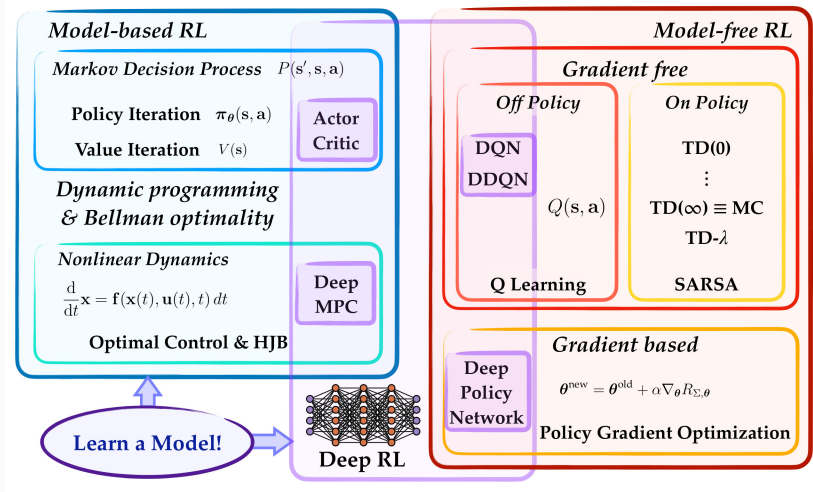
## Fonction de valeur ( $V$ )

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

Rappel - Objectif du RL

Apprendre la **politique optimale** qui maximise les récompenses futures par **essais/erreurs**.

# Different algorithms de RL



- Policy Gradient Optimization : TRPO, PPO, GRPO

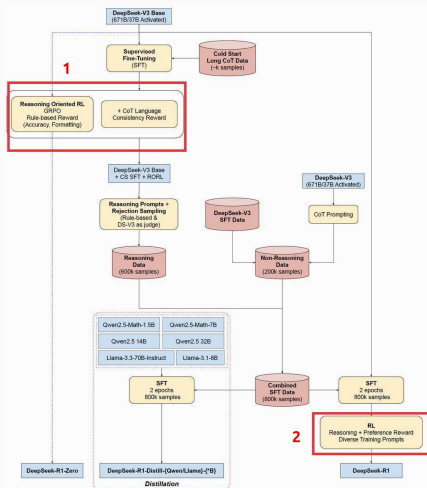
## **Application du RL au LLMs**

---

# Application du RL au LLMs

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

## Exemple de Deepseek-R1



1. *Post-Training* (RL-HF/AI) Phase d'alignement pour assurer **Reasoning + Preference RL**
2. *Pre-training*: Phase Compute & Data Intensive. **Reasoning Oriented RL**

## Différents Algo

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

Figure 1: Training Pipeline de DeepSeek R1

# RL Post-Training: Alignement

---

## Aligner le modèle

- Biasez le modèle pour lui faire adopter un comportement *préférable*.
- Etre *HH*: Helpful et Harmless.
- Besoin de récolter des jeux de données de préférence.



## Historique

- *Deep reinforcement learning from human preferences*, 2017
- *Fine-Tuning Language Models from Human Preferences*, 2020

# RL during Post-Training

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[ \frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left( \frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

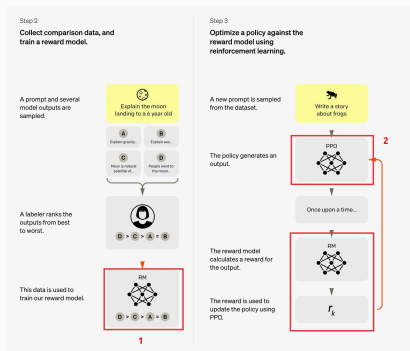


Figure 2: Pipeline Instruct-GPT: RLHF



# PPO - Proximal Policy Optimization

## Key Takeaways

- Objectif : Éviter les mises à jour trop grandes qui pourraient dégrader les performances.
- Modèle Acteur-Critique:
  - **Acteur**: modèle qui génère les actions (Policy Model - LLM)
  - **Critique**: évalue la qualité des actions - estime la valeur des récompenses futures (Value Model)<sup>1</sup>

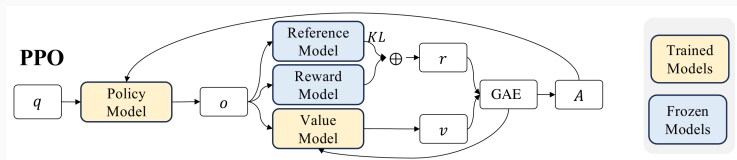


Figure 3: Diagramme de PPO

<sup>1</sup>Generalized Advantage Estimation (GAE)

# DPO : Direct Preference Optimization

## Key Takeways

- Approche alternative au RLHF
- Pas de Reward Model ni de Value Model
- Plus simple à implémenter
- Résultats similaires en performance

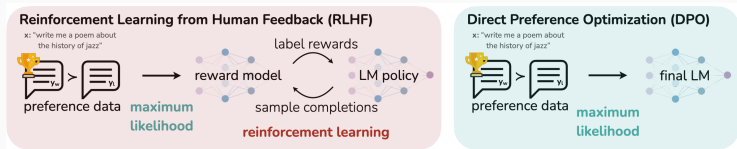


Figure 4: Illustration de DPO

## RL Pre-Training :

---

## Découvrir les Chain of Thoughts

- Laisser le modèle découvrir la meilleure façon de raisonner
- Se passer de grande quantités de données pour le SFT
- Question sous-jacente: Peut-on simplement récompenser le modèle pour sa précision et le laisser découvrir par lui-même la meilleure façon de penser ?

## Key Takeways

- Approche RL alternative à PPO
- Pas de Reward Model → Vérification fiable basée sur des règles.
- Pas de Value Model → Group-based Comparison. (gain de mémoire)

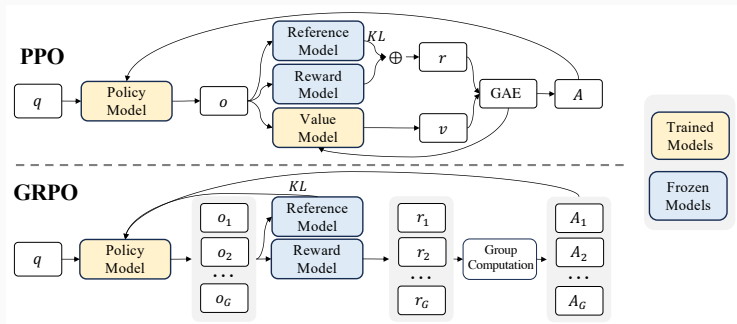


### DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao<sup>1,2,†</sup>, Peiyi Wang<sup>1,3,†</sup>, Qihao Zhu<sup>1,3,†</sup>, Runxin Xu<sup>1</sup>, Junxiao Song<sup>1</sup>  
Xiao Bi<sup>1</sup>, Haowei Zhang<sup>1</sup>, Mingchuan Zhang<sup>1</sup>, Y.K. Li<sup>1</sup>, Y. Wu<sup>1</sup>, Daya Guo<sup>1,2</sup>

<sup>1</sup>DeepSeek-AI, <sup>2</sup>Tsinghua University, <sup>3</sup>Peking University

# GRPO: Group Relative Policy Optimization

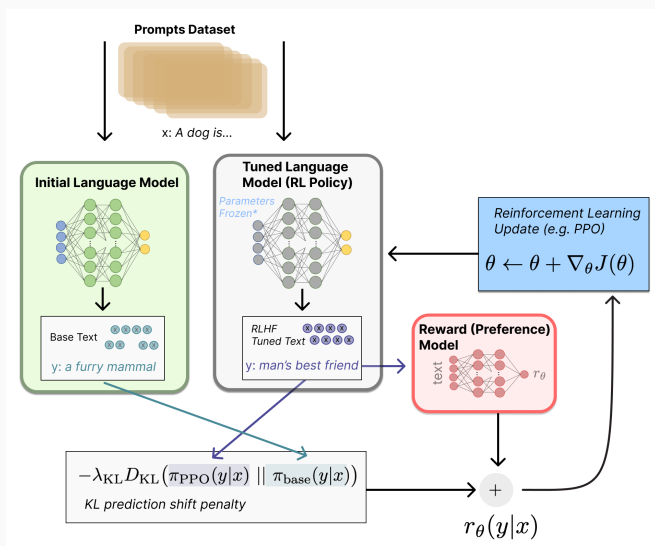


**Figure 5:** Comparaison PPO & GRPO issue de DeepSeekMath

- Récompenses de précision (Accuracy rewards)
- Récompense de format.  
*Pas de neural reward model car il pourrait souffrir de reward hacking.*

- Cours de la Fac de washington
- Video associée au Cours
- The 37 Implementation Details of Proximal Policy Optimization
- Video Yannick Kilcher : GRPO Explained
- Blog Hugging Face sur le RLHF
- Blog: DeepSeek's Lies: A Closer Look at GRPO Implementation

# Illustration Supplémentaire RLHF



**Figure 6:** Illustration du RLHF - Blog HuggingFace