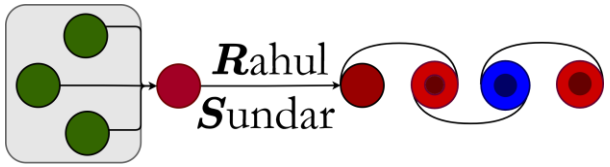
# PARENTING AI!

## How do you teach LLMs to behave?




PhD Scholar / Scientist (AI/ML)
Dept. of Aerospace Engineering, IIT Madras / **Verisk, India**


Founder, **Dhyuti**
An open source community initiative for AI in Science and Engineering

**@Ankura AI Meets – 2nd November, 2025**

# Your key takeaways

- Some **practical use cases** where finetuning/alignment become necessary
  *Explain RLHF and LLM Alignment in 6 levels!*

- Build intuition for **Reinforcement Learning**
  *With or w/o Human Feedback!*

- Build intuition for **finetuning LLMs with RLHF**
  *Understand how to parent an LLM (LLM alignment)!*

*Disclaimer – This session will be as devoid of equations and code as possible. A detailed, tailored and guided walkthrough can be scheduled for later depending on interest.*

# Large Language Models (LLMs)

- Everyone has heard about it somewhere!
- Everyone has used it somewhere!

Gemini

ChatGPT

Grok

SARVAM - M

**India’s sovereign Indic LLM**

Dhyuti

# Why do we need to align an LLM?

GPT-3 giving dangerous instructions vs **ChatGPT refusing politely!**

*(Grok vs ChatGPT is another example for responses related to verbal profanity!)*

Can Profanity be acceptable?

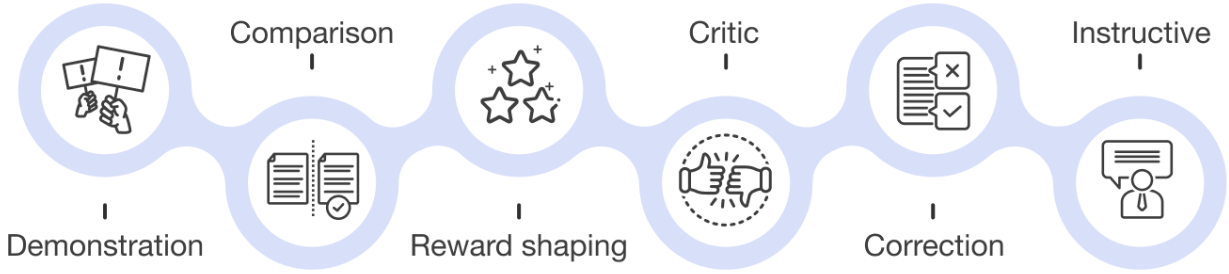
LLMs still work in a human world!

Human preferences, habits, values, and ethics become important!

# From good manners to moral reasoning — RLHF is how AI grows up with us!

| If you are a | Analogy |
|---|---|
| **Kid** | The AI is like a talking robot that learns **good manners** when people tell it what's nice or rude. |
| **Teen** | It's like training a gamer — they must **play smart and play fair** to win praise. |
| **College Student** | The model reads millions of books, but RLHF teaches it **what humans mean and prefer**. |
| **Professional** | Like performance reviews at work — feedback helps the **AI learn what success really looks like**. |
| **Scientist** | RLHF converts statistical prediction into moral alignment — **teaching machines to reason within human values.** |
| **Grandma /pa** | It's like teaching a grandchild to **speak kindly and behave well — not just talk, but talk with heart**. |

# What can good parenting of LLM do? - Applications of RLHF




LLMs still work in a human world.

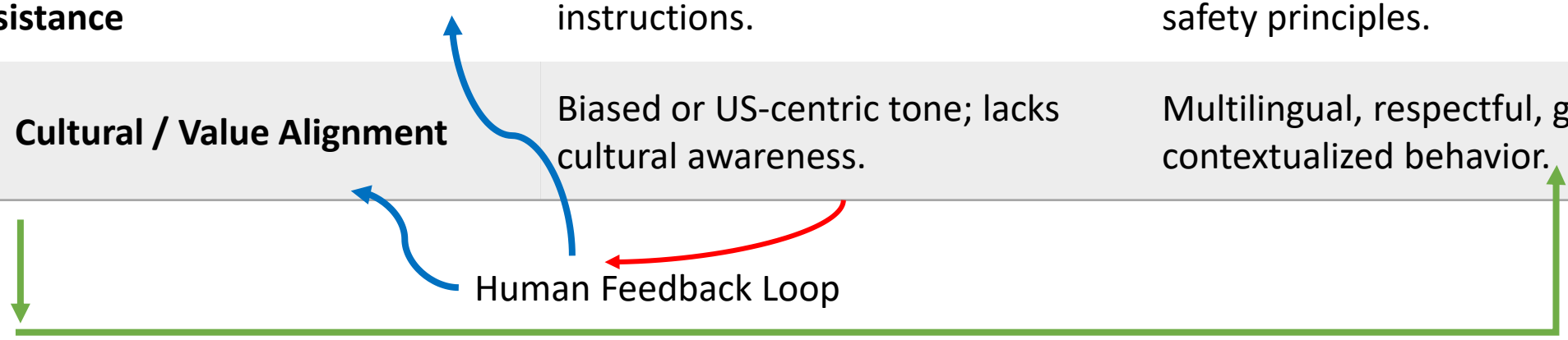
**Feedback and handholding becomes important!**

LLMs are worse than teenagers and quite powerful if untethered!

# Applications Layer 1 — Guarding the Boundaries of AI Behavior

*With great power comes great responsibility! LLMS have to behave well!*
*From unfiltered power to responsible intelligence.*

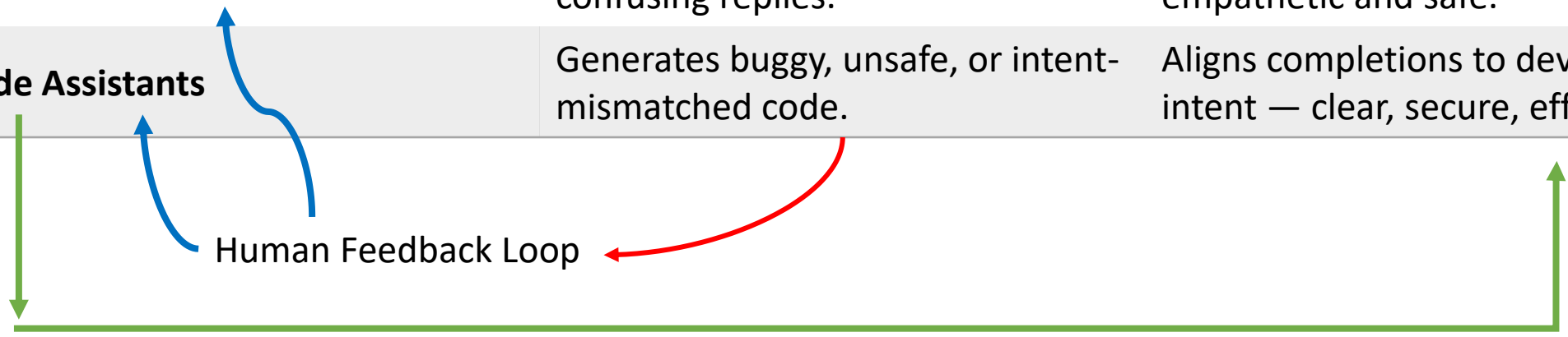
| | Before RLHF | After RLHF |
|---|---|---|
| 🚫 **Red-Teaming / Jailbreak Resistance** | Produces unsafe, unethical instructions. | Politely refuses, explains why; cites safety principles. |
| 🌍 **Cultural / Value Alignment** | Biased or US-centric tone; lacks cultural awareness. | Multilingual, respectful, globally contextualized behavior. |



# Applications Layer 2 — How RLHF Makes AI Conversations Human

***From customer service point of view, you want the LLMs to go from smart talkers to empathetic companions!***

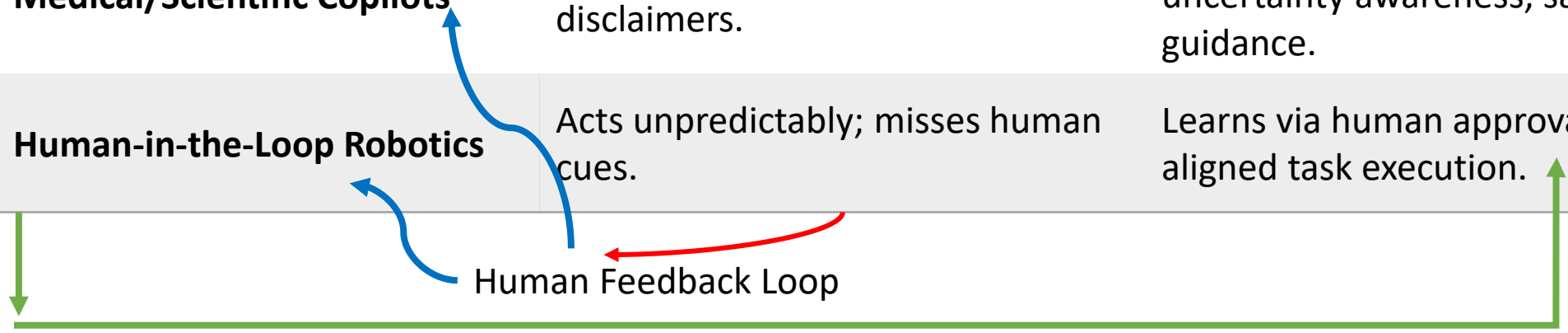
| | Before RLHF | After RLHF |
|---|---|---|
| **Conversational AI** | Tone: factual but rude; unsafe or confusing replies. | Tone: helpful, honest, harmless — empathetic and safe. |
| **Code Assistants** | Generates buggy, unsafe, or intent-mismatched code. | Aligns completions to developer intent — clear, secure, efficient. |



# Applications Layer 3 — Teaching AI Scientific Responsibility

*You do not want gibberish when communicating Data or scientific facts! From confident guessers to careful collaborators!*

| | Before RLHF | After RLHF |
|---|---|---|
| **Medical/Scientific Copilots** | Overconfident or hallucinates; lacks disclaimers. | Rewards factual precision, uncertainty awareness, safe guidance. |
| **Human-in-the-Loop Robotics** | Acts unpredictably; misses human cues. | Learns via human approval — safer, aligned task execution. |

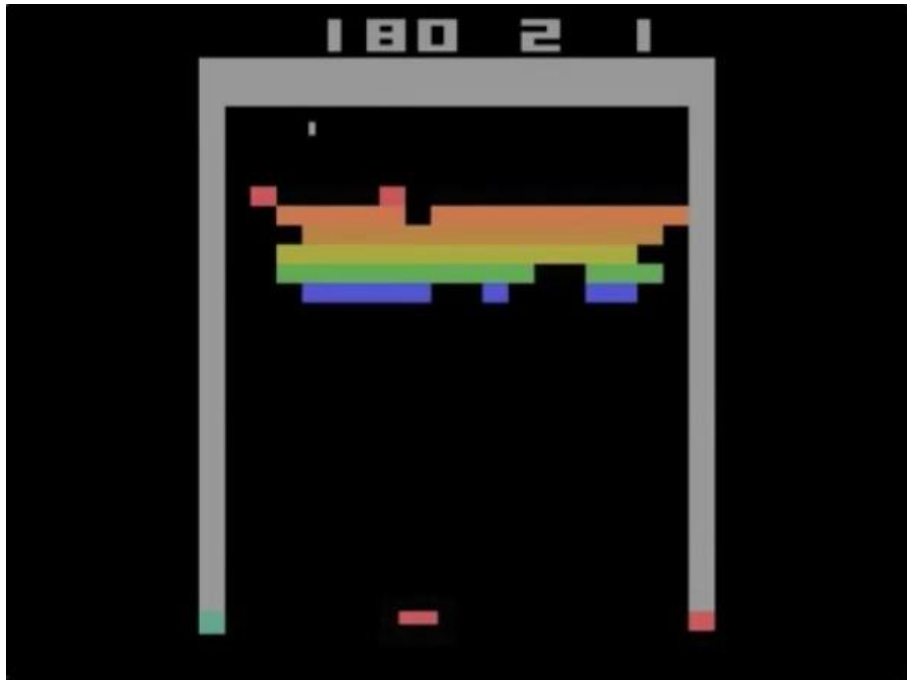


# How do we build intuition for RL and RLHF?

## Some background and flashback on RL first!

# Reinforcement learning (RL) breakthrough

Atari 2600 – Breakout Gameplay by RL *DeepMind beat records with expert level game play!*

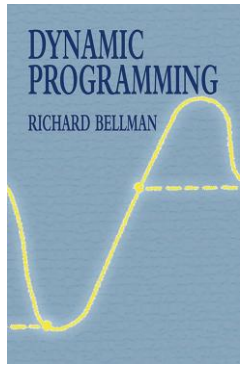


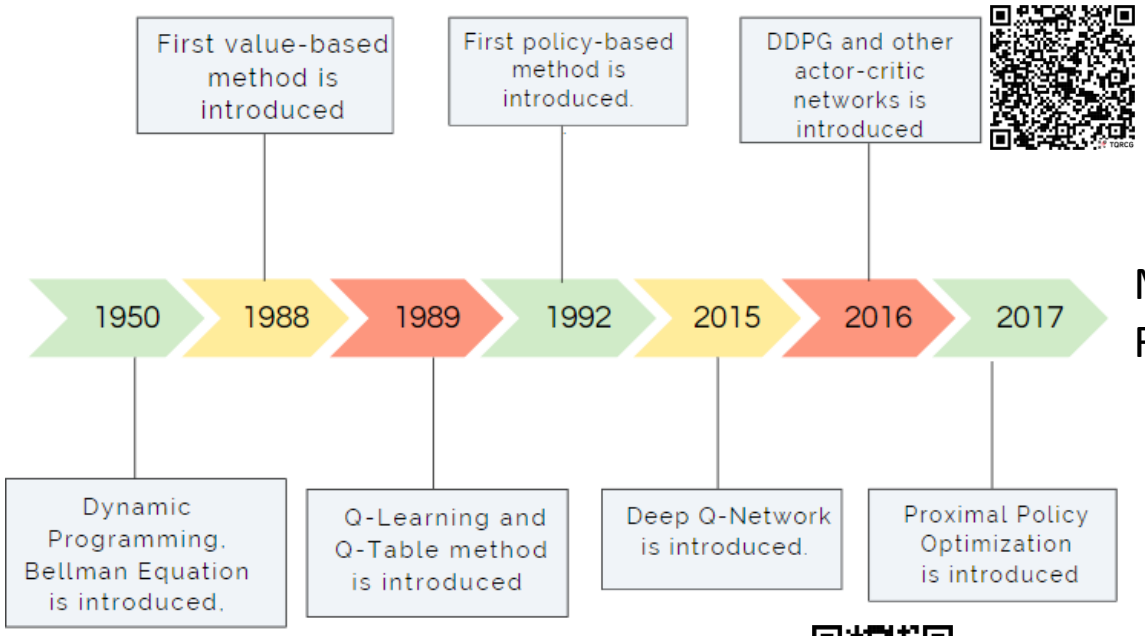

The game of Go is the most challenging of classic games. Despite decades of effort, prior methods had only achieved amateur level performance. We developed a deep RL algorithm that learns both a value network (which predicts the winner) and a policy network (which selects actions) through games of self-play. Our program AlphaGo combined these deep neural networks with a state-of-the-art tree search. In October 2015, AlphaGo became the first program to defeat a professional human player. In March 2016, AlphaGo defeated Lee Sedol (the strongest player of the last decade with an incredible 18 world titles) by 4 games to 1, in a match that was watched by an estimated 200 million viewers.

# Evolution of Reinforcement learning (RL)

The Go To course by the OG Prof Richard Sutton -

Dynamic Programming




Modern Reinforcement

2018-2025

Advanced RLHF, RLAIF, Actor Critic Frameworks *DPO, IPO< KPO, NLPO, A2C and TRPO*

Proximal Policy Optimisation Algorithms, Schulman et al. (2017)

# How do we build intuition for RL?

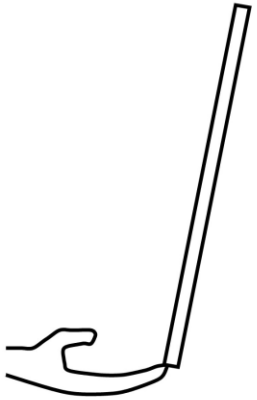
Some familiar examples!

# Intuition behind Reinforcement learning (RL)

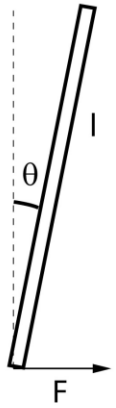
*How did you learn to ride a bicycle without support?*

*Have you ever tried balancing a ruler on your finger tip?*

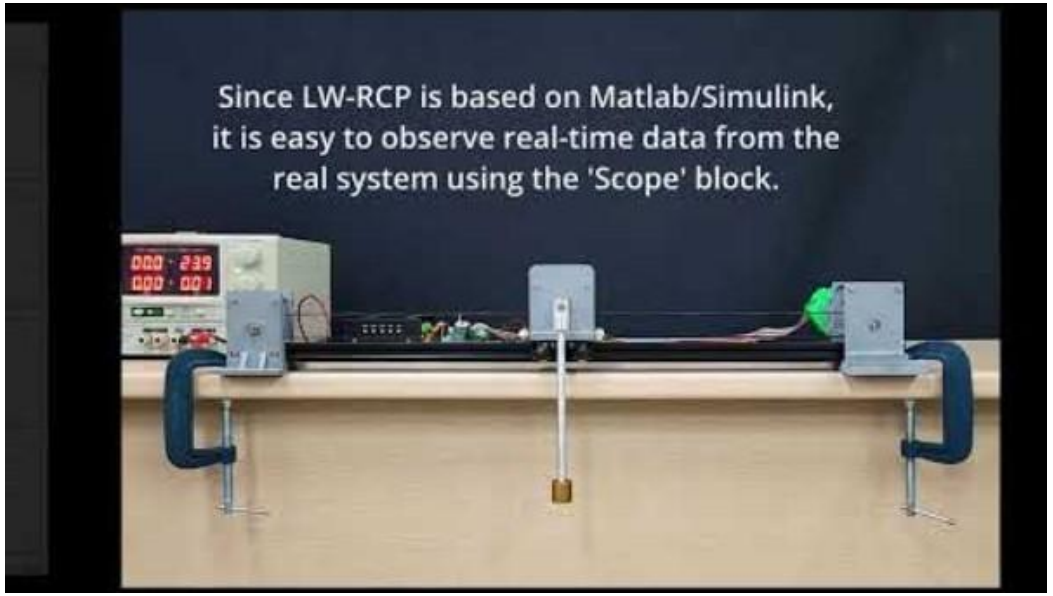
Segway



Balancing a stick




Inverted pendulum

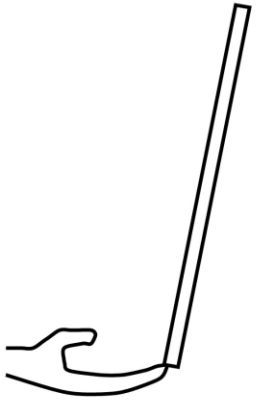



***Explore versus exploit!***

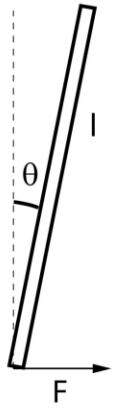
# Intuition behind Reinforcement learning (RL)

*Have you ever tried balancing a ruler on your finger tip?*

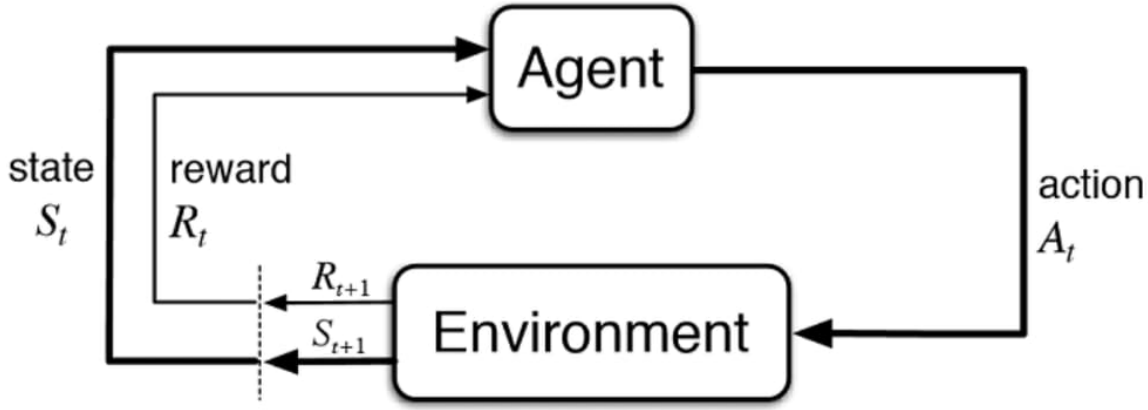
Segway



Balancing a stick




Inverted pendulum

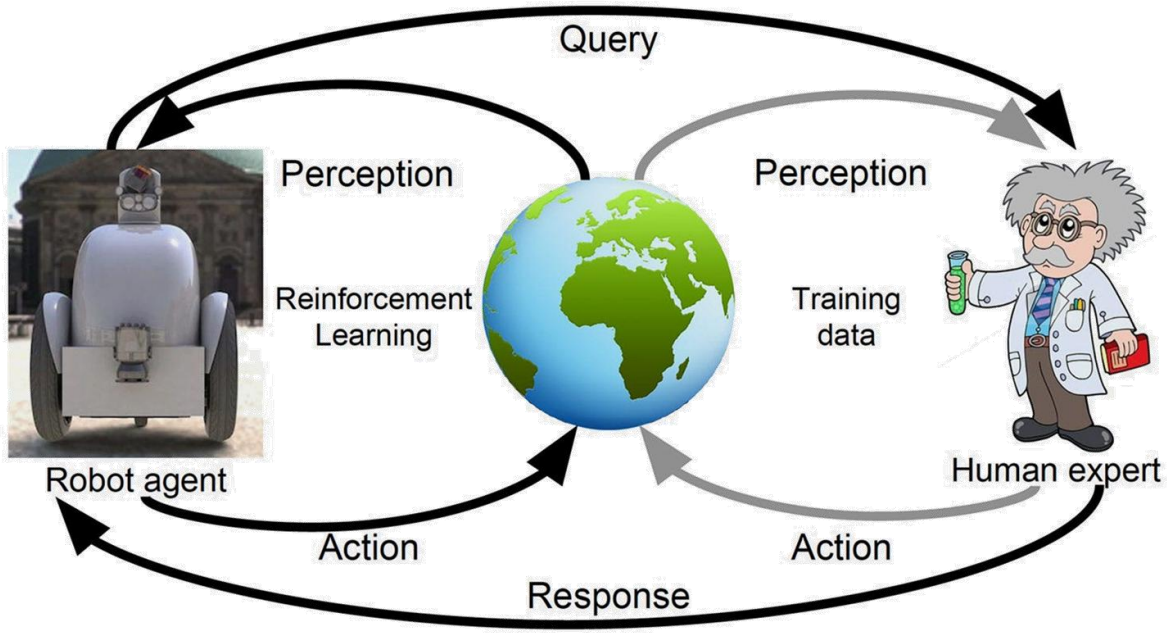



*You were the **Agent**, The ruler – **Object/System**, Gravity + Air – **Environment**, Your actuations based on your knowledge of Gravity – **The Agent Policy***

# Intuition behind Reinforcement learning (RL)

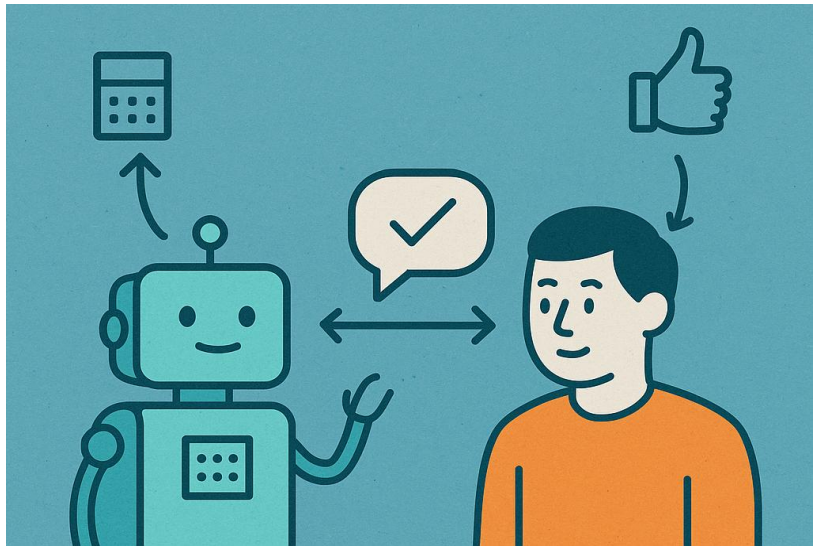
RL in “***Action***”

To train a robot to ***defend*** itself

# Intuition - Reinforcement learning with Human Feedback (RL)

Have you ever liked or disliked a ChatGPT response?



***For the geeks like me here:***

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback **(Bai et al. (2022), https://arxiv.org/abs/2204.05862**

# How do we build intuition for RL and RLHF?

## The building blocks!

# Intuition - Reinforcement learning with Human Feedback (RL) (No math)

1 LLM **Supervised Fine-tuning (SFT)** on curated examples
(Curate for the behaviour you want)
2 **Reward Model Training**
(ranking responses based on human preferences)
3 **Proximal Policy Optimisation or other RL Optimization**
(Optimise the LLM weights)

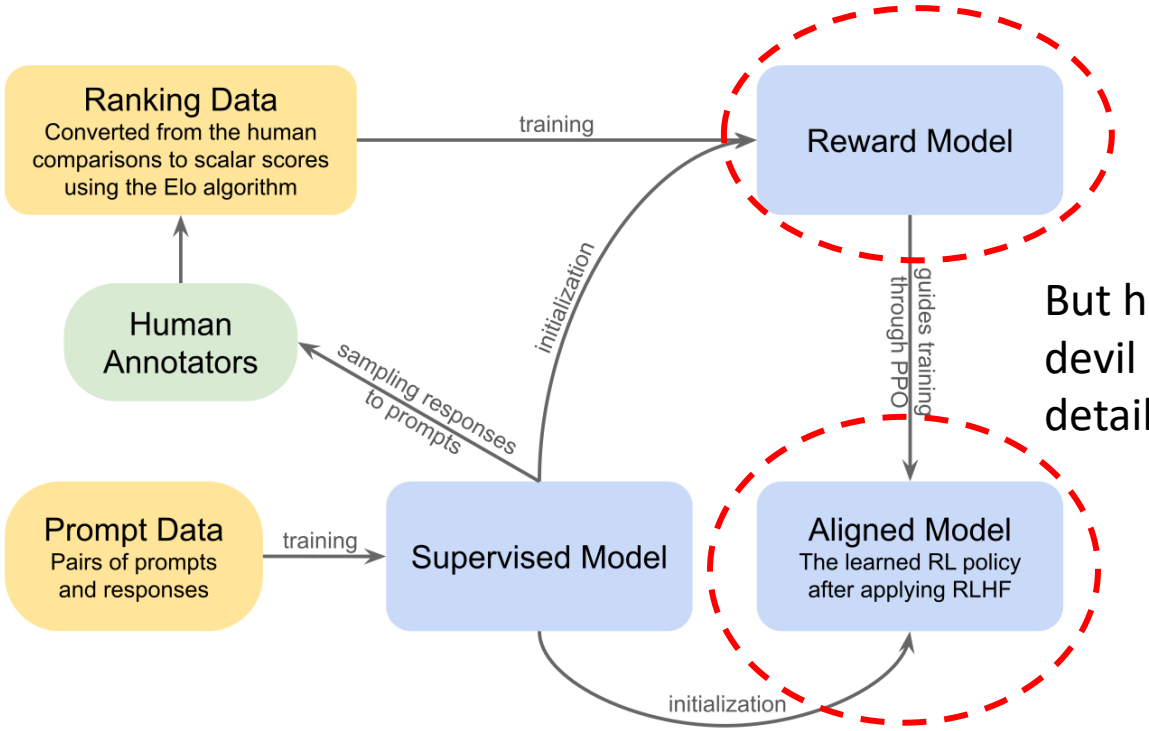
*"Humans rank model answers → we train a reward model → RL nudges the LLM toward high-reward behaviour."*

"Pretraining teaches *what is probable*; RLHF teaches *what is preferable*."

**Analogy:** Like teaching a kid by showing graded examples instead of laying out explicit rules.

# Intuition - Reinforcement learning with Human Feedback (RL) (No math)

*With the reward model, constrain the model to behave the way you want it to!*
*Parent it with care & patience! LLMs are the worst yet smart teenagers!*

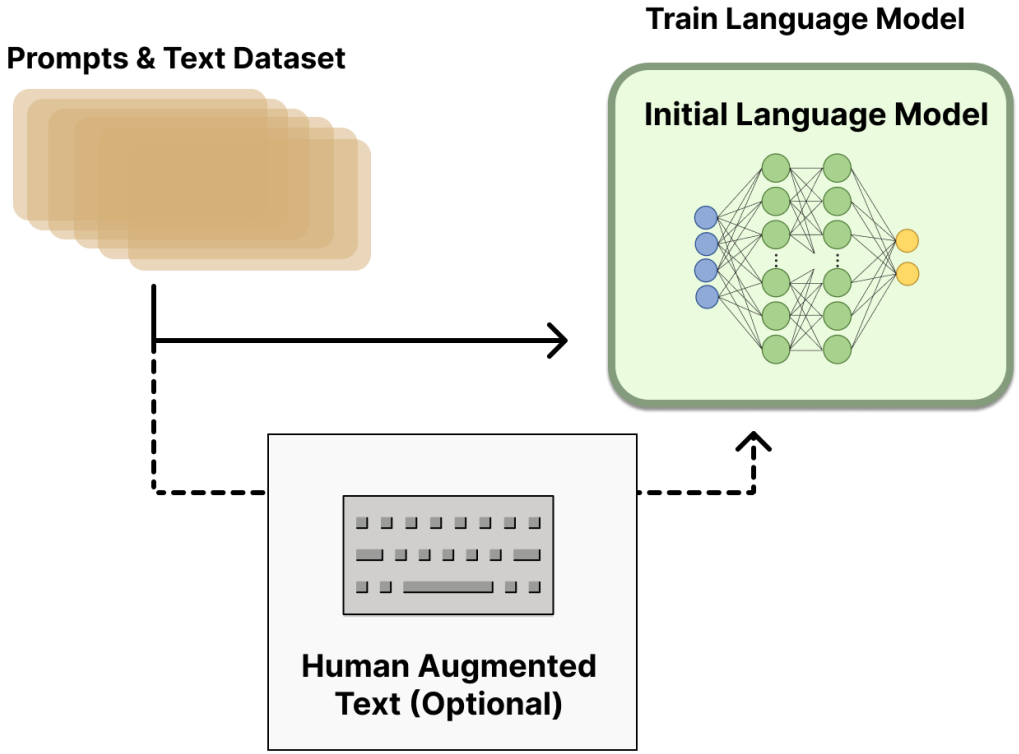



But how? The devil lies in the details! – PPO?

Dhyuti

# RLHF based finetuning

Pretrain LLM, do supervised finetuning, and freeze it

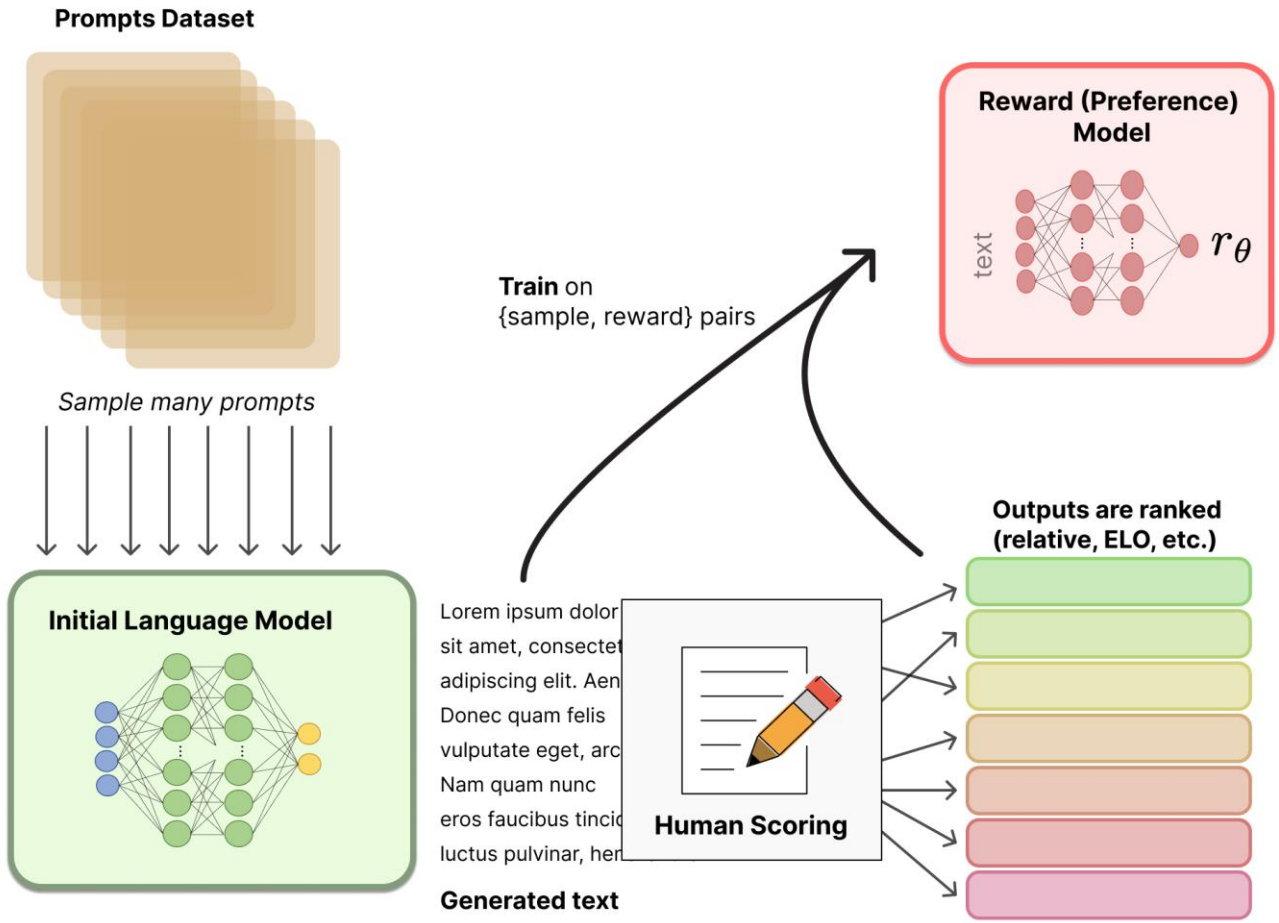



Source:
*https://huggingface.co/blog/rlhf*

# RLHF based finetuning

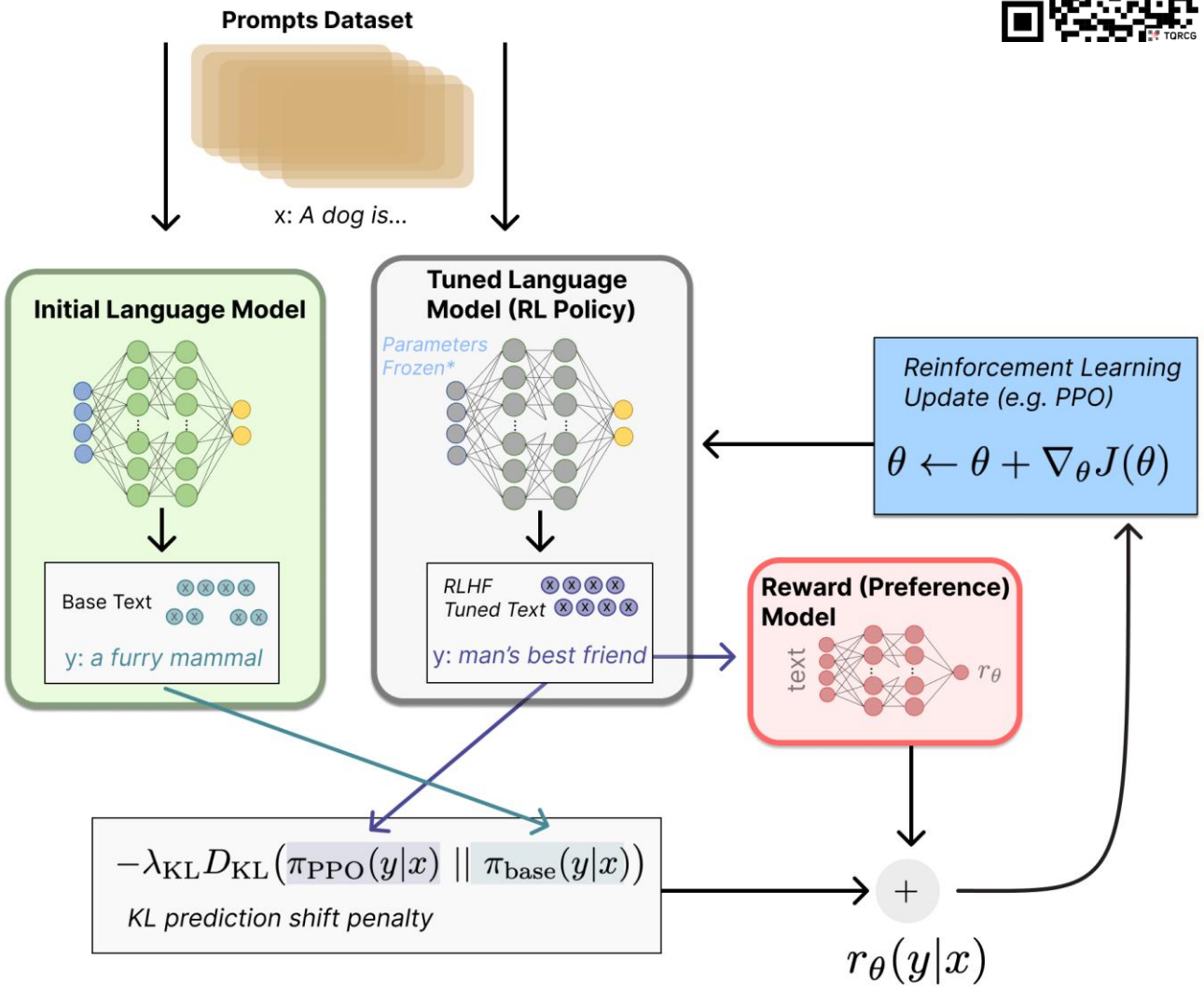
Train a reward model based on Human Preferences

Source: *https://huggingface.co/blog/rlhf*



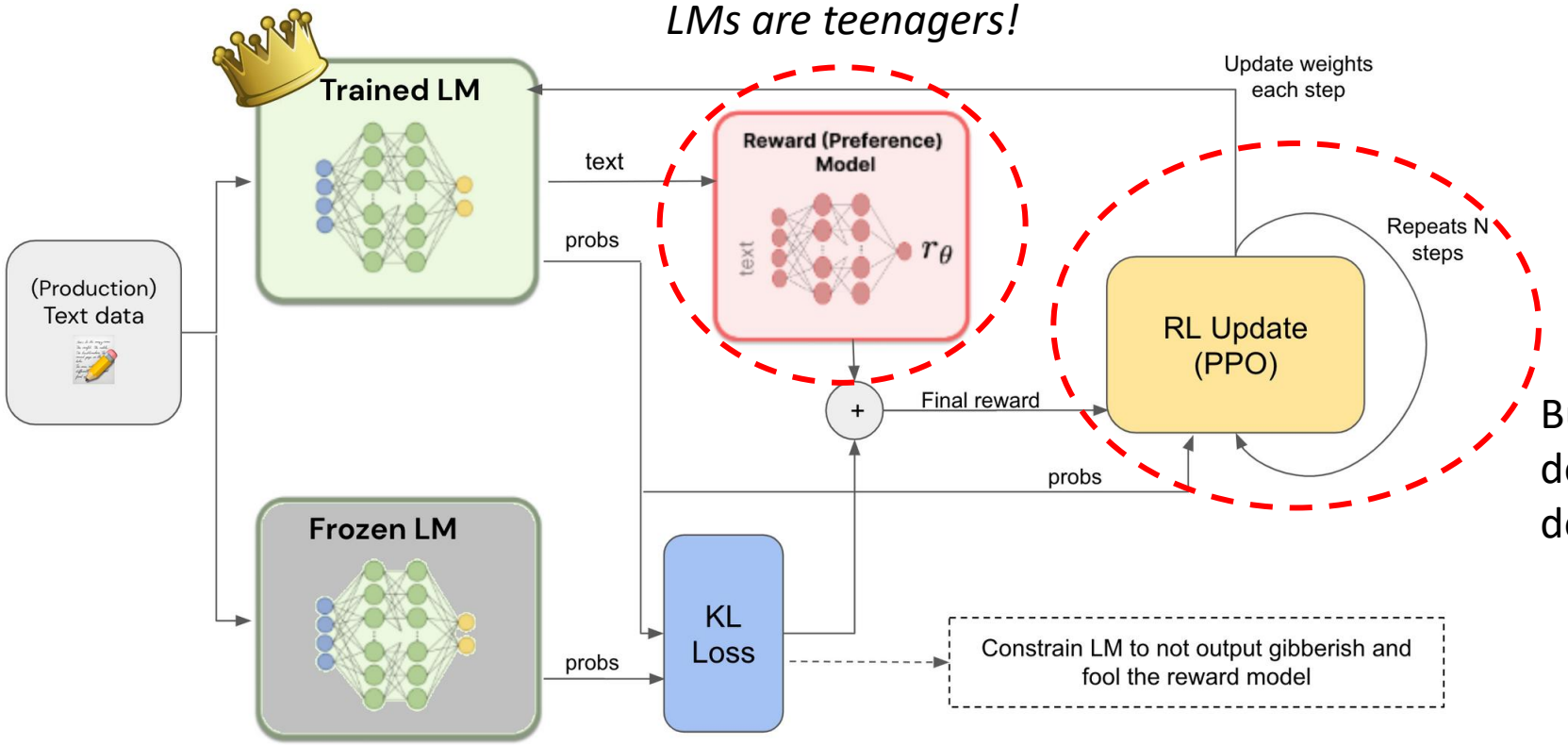
# RLHF based finetuning

Source: *https://huggingface.co/blog/rlhf*



# Intuition - Reinforcement learning with Human Feedback (RL)

*With the reward model, constrain the model to behave the way you want it to! Parent it with patience! LMs are teenagers!*




But how? The devil lies in the details! – PPO?

Dhyuti

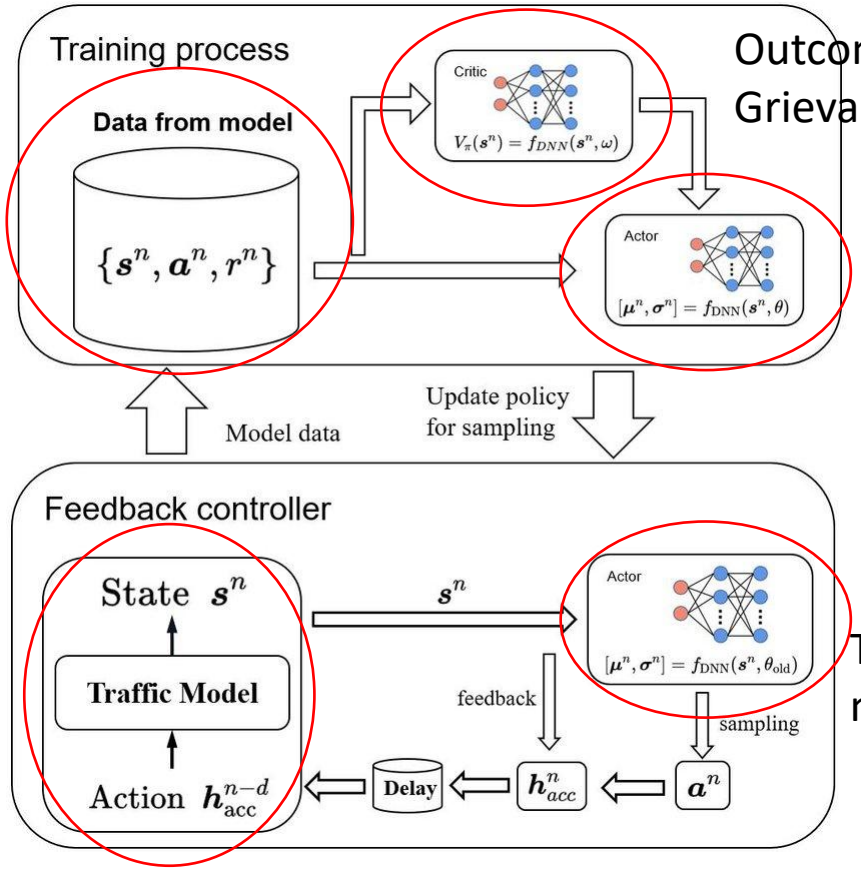
# How do we build intuition for RLHF?

## Building blocks of Proximal Policy Optimisation (PPO)

# PPO – Proximal Policy Optimisation

**Objective of PPO - How do I learn how to reward the right actions?**

Existing traffic control model – *For how much time do we wait near IKEA?*




Outcome favourable or not/ Grievances!

The Traffic policeman regulates the traffic -> based on a prior policy

The Traffic flow around IKEA as a model

The Traffic policeman tests out new policy

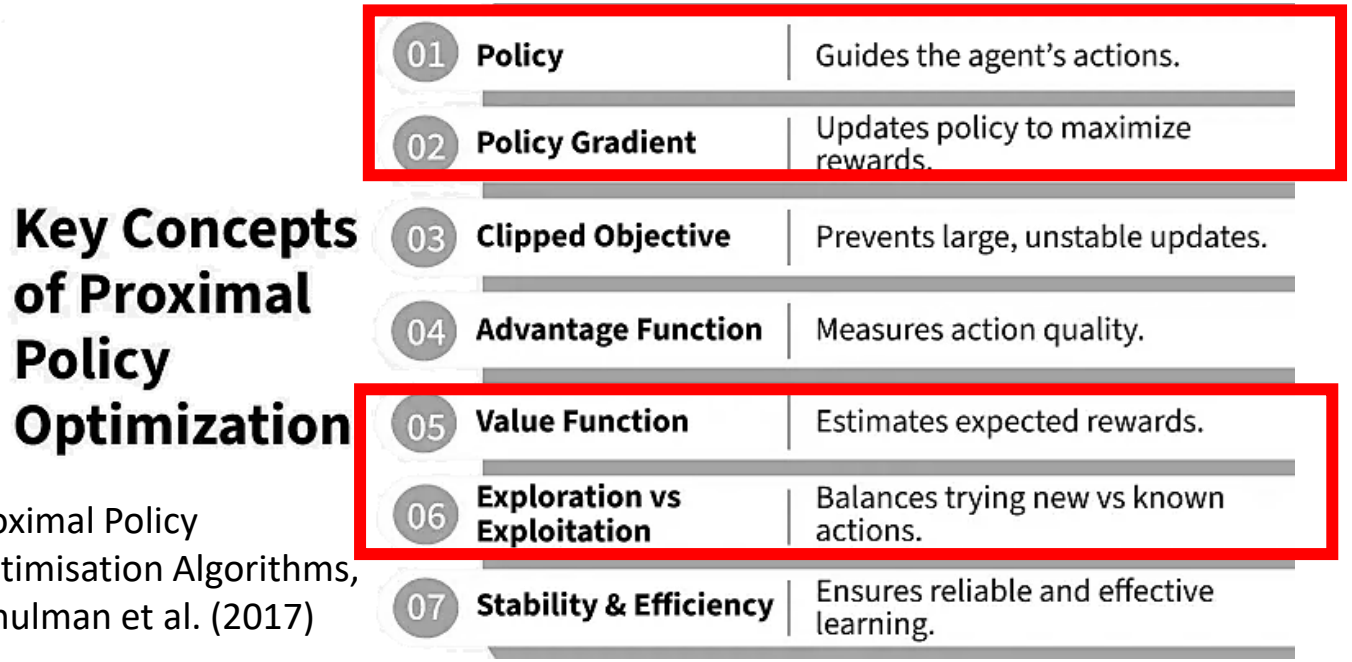
Explore versus Exploit! The loop above continues until we settle down to something viable!

# PPO – Proximal Policy Optimisation

Objective of PPO - How do I learn how to reward the right actions?

Actor – Critic based framework

Remember the keywords inside the boxes!




Proximal Policy Optimisation Algorithms, Schulman et al. (2017)

# Putting it together - Reinforcement learning with Human Feedback (RL)

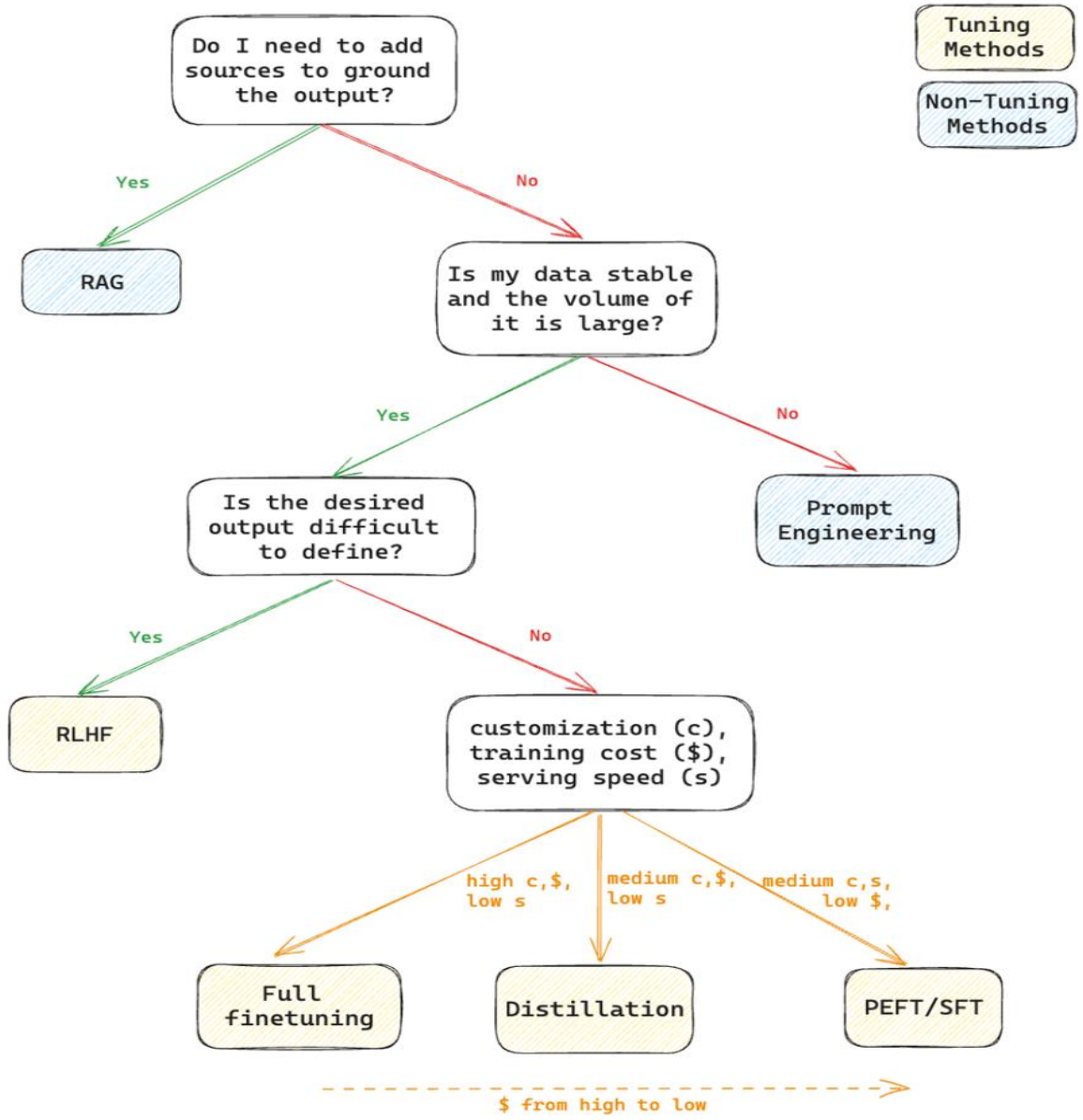
| Stage | Input | Process | Output | Key Goal |
|---|---|---|---|---|
| **1. Supervised Fine-Tuning (SFT)** | Pre-trained LLM, Curated prompt-response pairs (human-generated) | Supervised learning on demonstration data | SFT Model | Adapt LLM to task format, instill initial desired behavior/style |
| **2. Reward Model (RM) Training** | SFT Model, Diverse prompts, Human preference labels (rankings/comparisons) | SFT model generates responses; Humans rank/compare; Train RM to predict human preference scores | Reward Model (RM) | Create an automated proxy for human judgment that provides scalar rewards |
| **3. RL Policy Fine-tuning (e.g., PPO)** | SFT Model (as initial policy), RM, Prompts | LLM (policy) generates responses; RM scores responses; PPO updates LLM policy to maximize RM scores | RL-aligned LLM (Final Model) | Iteratively refine LLM to generate outputs aligned with human preferences learned by the RM |

# But When to RLHF?

*(For the engineers and decision makers here)*

**Example tutorial**
**Finetune a Gemma - 2B model with RLHF**

https://colab.research.google.com/github/heartexlabs/RLHF/blob/master/tutorials/RLHF_with_Custom_Datasets.ipynb#scrollTo=UziQ7Gyjs-ra



# Your key takeaways

- Discussed some practical use cases where finetuning/alignment become necessary
  ***You can now explain RLHF and LLM Alignment in 6 levels!***

- We have built fundamental intuition for Reinforcement Learning
  ***With or w/o Human Feedback!***

- We have built intuition for finetuning LLMs with RLHF
  ***Understand how to parent an LLM (LLM alignment)!***

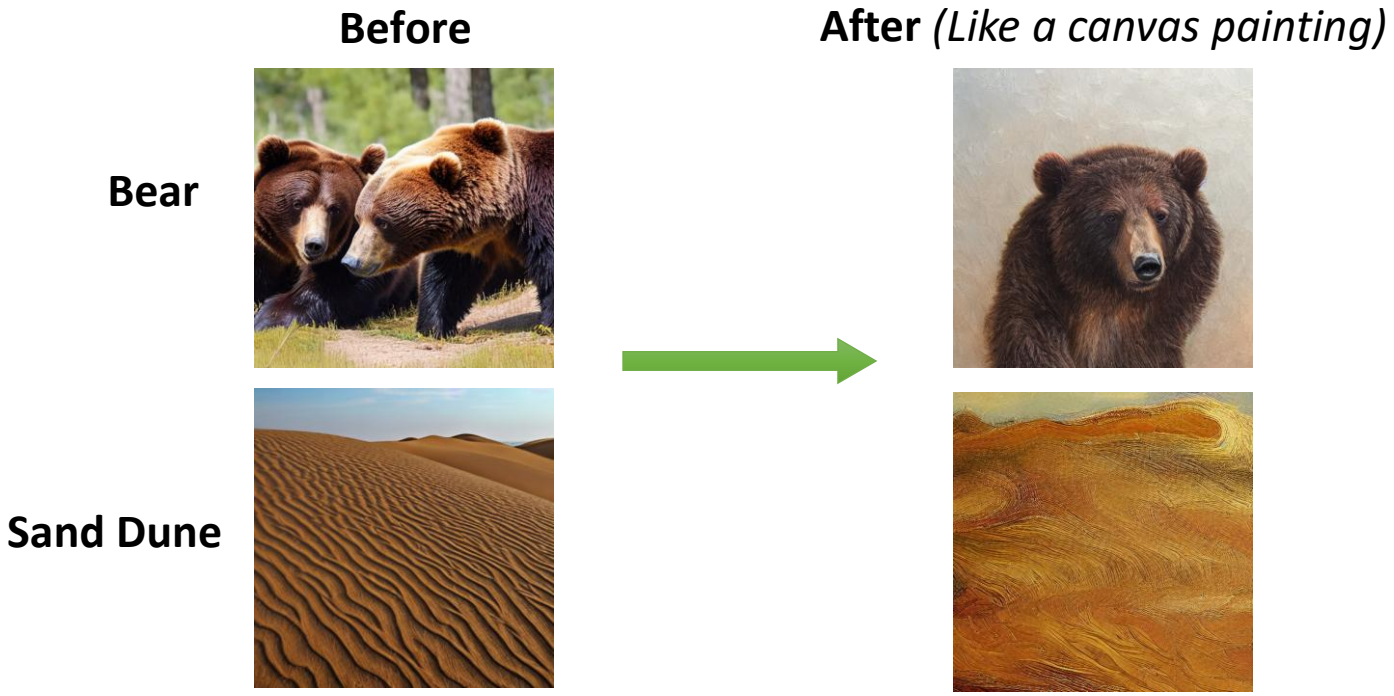
*What next? -> Taking RLHF and some low costs LLMS out for a spin for practical scenarios!*
***Calls for a deep dive hands-on session!***

# Explain RLHF to anyone!

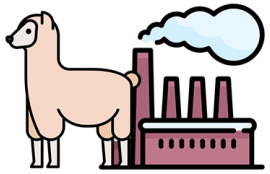
| Age Group | Analogy |
| --- | --- |
| 🧒 **Kid** | The AI is like a talking robot that learns good manners when people tell it what's nice or rude. |
| 👦 **Teen** | It's like training a gamer — they must play smart and play fair to win praise. |
| 🧑 **College Student** | The model reads millions of books, but RLHF teaches it what humans mean and prefer. |
| 👨‍💼 **Professional** | Like performance reviews at work — feedback helps the AI learn what success really looks like. |
| 👩‍🔬 **Scientist** | RLHF converts statistical prediction into moral alignment — teaching machines to reason within human values. |
| 👵 **Grandma** | It's like teaching a grandchild to speak kindly and behave well — not just talk, but talk with heart. |

# Beyond LLMS

- **RLHF** can be applied to **any AI model** that has to **align with Human preferences!**
- **Example:** Style transfer in Stable Diffusion models



# References

- **Tools and frameworks:** TRL, TRLX, RL4LMs, LLAMA Factory

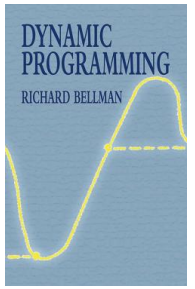

- **Code tutorials :** Hugging face , Kaggle
- **Key Research papers:**

Proximal Policy Optimisation Algorithms, Schulman et al. (2017)

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback **(Bai et al. (2022),** **https://arxiv.org/abs/2204.05862**

- **Educational material**

The OG course by Prof Richard Sutton




Dynamic Programming

# Thank you


**Let us connect on LinkedIn!**