Preference Fine-Tuning (Part 1)

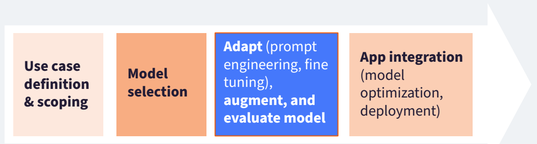
INTRODUCTION

Some models exhibit undesirable behavior:

- Generating toxic language
- Responding aggressively
- Providing harmful information

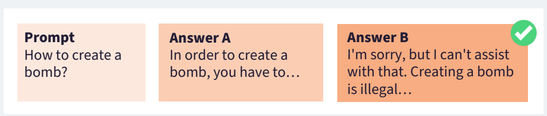
To ensure alignment between LLMs and human values, emphasis should be placed on qualities like helpfulness, honesty, and harmlessness (HHH).

Generative AI Project Lifecycle



→ Additional training with **preference data** can boost HHH in completions.

Preference data



The answers have been generated by the model we want to fine-tune and then assessed by human evaluators or an LLM.

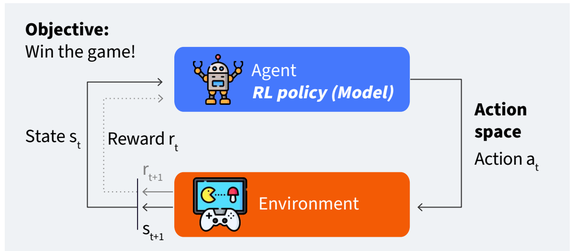
Two approaches:

- **Reinforcement Learning With Human Feedback (RLHF):** Preference data is used to train a reward model that mimic human annotator preferences, which then scores LLM completions for reinforcement learning adjustments.
- **Preference Optimization (DPO, IPO):** Minimize a training loss directly on preference data.

RLHF PRINCIPLES

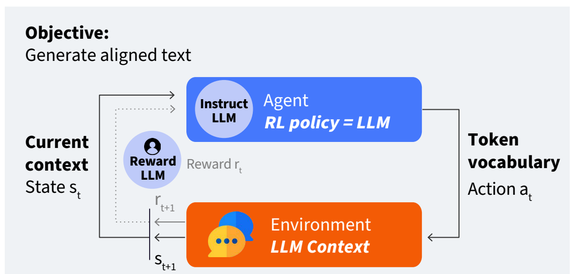
Reminder on Reinforcement Learning

Type of ML in which an agent learns to make **decisions** towards a specific goal by taking **actions** in an **environment**, aiming to **maximize some cumulative reward.**



Action space: All possible actions based on the current environment state.

In the context of LLMs...



Action: Text generation
Action space: Token vocabulary
State: Any text in the current context window

The **action** the model will take depends on:

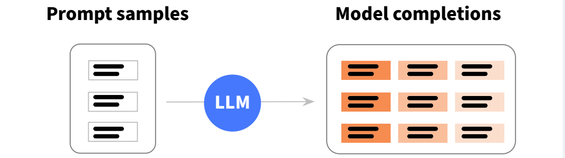
- The *prompt text* in the context
- The *probability distribution* across the vocabulary space

The **reward model** assesses **alignment** of LLM outputs with **human preferences.**
The reward values obtained are then used to **update the LLM weights** and **train a new human-aligned version,** with the specifics determined by the optimization algorithm.

COLLECTING HUMAN FEEDBACK

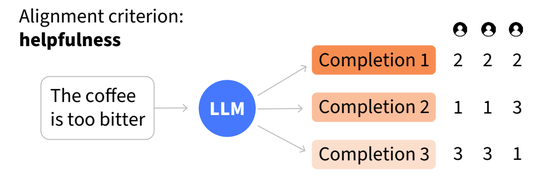
Steps

1. Choose a model and use it to curate a dataset for human feedback.



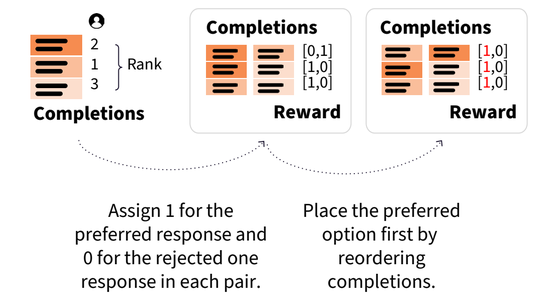
2. Collect feedback from human labelers (generally, thousands of people):

- Specify the model alignment criterion.
- Request that the labelers rank the outputs according to that criterion.



→ **Detailed instructions** improve response **quality** and **consistency**, resulting in labeled completions that reflect a consensus.

3. Prepare the data for training
Create pairwise training data from rankings for the training of the reward model.



REWARD MODEL

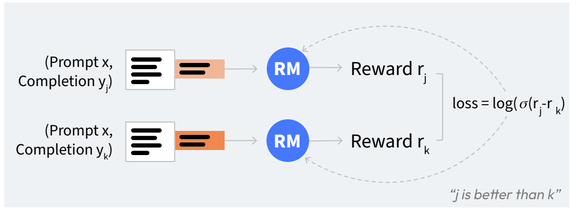
Objective:

To develop a model or system that accepts a text sequence and outputs a scalar reward representing human preference numerically.

Reward model training:

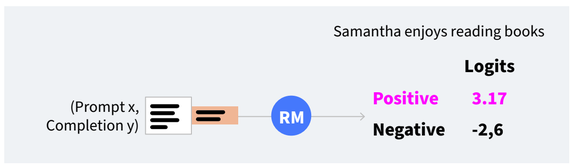
The reward model, often a **language model** (e.g., BERT), is trained using supervised learning on **pairwise comparison data** derived from human assessments of prompts.

Mathematically, it learns to prioritize the human-preferred completion while **minimizing the log sigmoid of the reward difference.**



Usage of the reward model:

Use the reward model as a binary classifier to assign reward values to prompt-completion pairs.



Reward value equals the **logits** output by the model.