SFT:

input comes from human, output comes from LM and the human rates the response.

This information is used for reward modelling (which we can think of as a classifier)

Then, you optimize a policy with that reward modeling.

self-instruct: synthetic generation

human-in-the-loop: human doing post-processing tasks.

- human preference dataset, human writes input, model outputs, human rates them.
- for the SFT dataset both input and output are written by an human

Distillation:

dSFT (generate multi-turn AI dialogues)

AIF (response generation and AI ranking)

dDPO (distillation of AI preferences)

4 different training techniques

1. Pretraining the LLM
2. In context learning (prompt-based learning)
3. Supervised fine-tuning
4. Reinforcement learning with Human Feedback

<u>evaluating a dataset:</u>

- evaluating instruction following/chatiness
- evaluating the reward modelling
- red-teaming (or checking response based on adversarial prompting)
- helpfulness (elo rating)
- can the model choose between a truthful and an untruthful response?. (HH reward model benchmarks)

Red Teaming LLMs:

- there's no leaderboard for this unfortunately.

using llms as an evaluator (gpt4 in particular)

↳ gpt 4 prefers gpt inputs to human inputs :)

↳ prefers more unique and longer responses

↳ doesn't align with math related responses.