Training LLM (LLM works)

- from last page we see that LLM are giant NN where the layers represent operations Like embedding, attention etc and the weights of that layer are used in the respective processes. these weights are known as parameters of model.

-NOTE : for a more efficient training process it is better to have the LLM predict all sequences of next words in a given price of text so if input : "the cat sat on the mat"

ie:

input	target
the	cat
the cat	sat
⋮	⋮
the cat sat on the	mat

→ even if the last prediction was "the dog" use the true next word "cat" in the training process and just penalize the model

RLHF (Reinforcement learning with human Feedback)

PPO is a new alt to RLHF! ←

- is a fine tuning step in the training process that helps the LLM feel helpfull, polite and align with human expectations. (beyond just predicting next word) and not be generic

How 3 Phase RLHF: works

1) Supererised fine tuning (SFT)
   - Humans write examples prompts and ideal responses

   → the model is trained to imitate these heppfull completions

2) Reward Model (RM) training
   - Humans are shown multiple responses to a prompt from the model

   → they rank them best to worst → a RM is trained to predict these rankings

3) Reinforcement learning (eg. PPO)
   - LLM generates responses

   → the RM scores each response

   → Model is fine tuned via RL to maximise the reward

   → this makes model human friendly