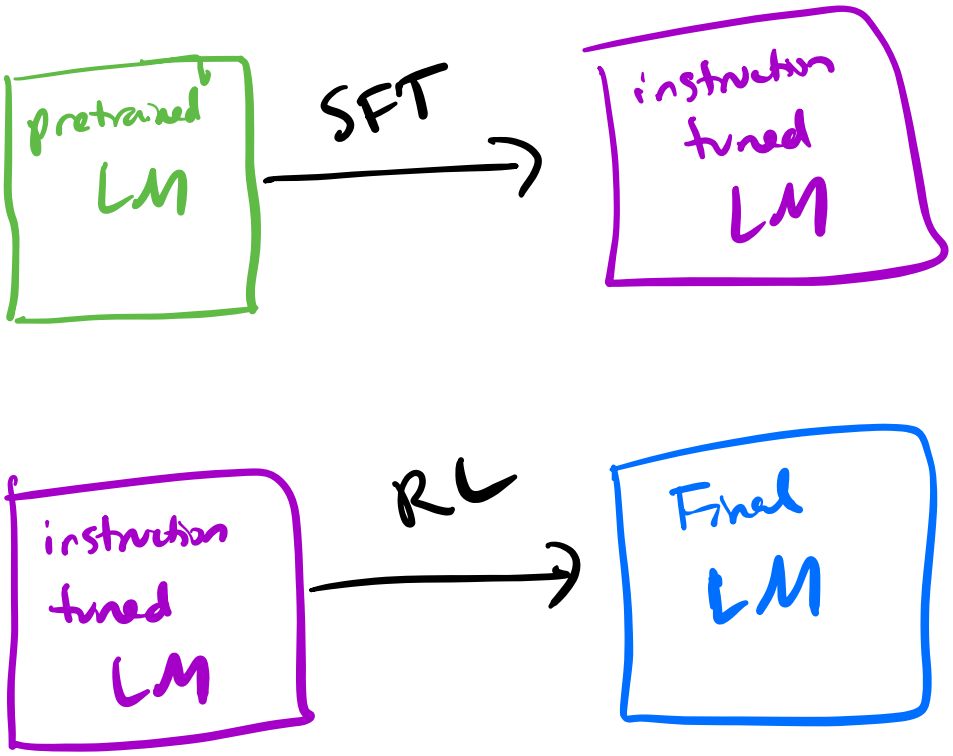
SFT to RL:



↳ instruction tuning requires a diverse dataset of instruction / input / output tuples

x y

$$L = \frac{1}{N} \sum_{i=1}^{N} -\log P(y_i \mid x_i)$$

↳ ground-truth output y is generally created by a human

issues w/ instruction tuning:

- expensive to create HQ instruction following datasets
  - synthetic data
- one acceptable y per input x
- no learning from negative examples
- distribution mismatch:
  - training time: expert HQ y (teacher forcing)
  - test time: model-gen y
    - exposure bias
- doesn't directly involve human prefs