

# Exam review

## Logistics

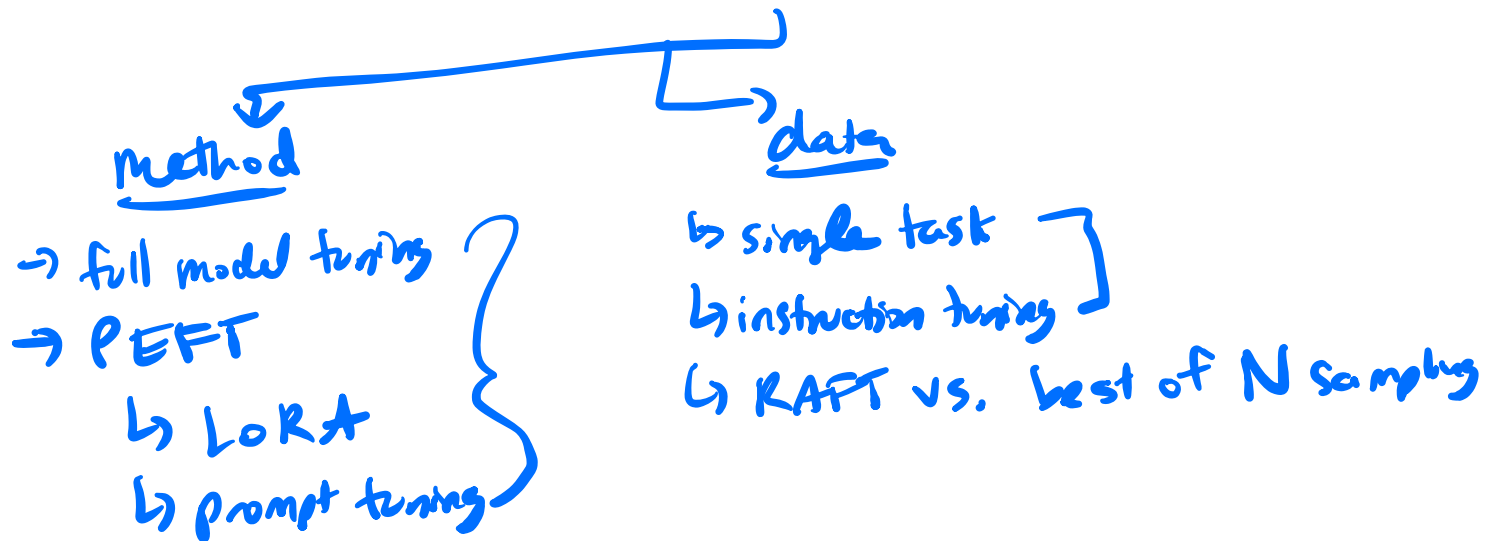
- ↳ ~40q, <sup>all</sup> multiple choice
- ↳ primarily lower post exam 1 material
- ↳ couple qs about HW2, none coding-related
- ↳ 8.5" x 11" cheat sheet, both sides, handwritten
- ↳ please circle your chosen answer

# Topics

↳ pretraining vs. post-training



SFT vs. RL



## Limitations

- ↳ teacher forcing / exposure bias
- ↳ <sup>only</sup> one reference per prompt
  - ↳ data collection is expensive
- ↳ don't learn from negative feedback

RL :

↳ RLHF



reward model  
trained on human  
prefs

↳ Bradley-Terry  
model

vs.

RLVR



verifiable reward

↳ typically done  
on math / code

↳ harder to perform  
reward hacking

↳ Algorithms

↳ REINFORCE → reward :

Advantage

$$\underbrace{r(x,y) - b(x)}$$

↳ PPO → value function

↳ GRPO → group normalization

↳ KL divergence penalty

↳ keep policy model from diverging  
too much from SFT model

↳ catastrophic forgetting

↳ clipping based on the policy ratio

$$\rho_i = \left[ \frac{p_{\theta}(y_i|x)}{p_{\theta_{old}}(y_i|x)} \right] \left. \vphantom{\frac{p_{\theta}(y_i|x)}{p_{\theta_{old}}(y_i|x)}} \right\} \begin{array}{l} \text{trust region} \\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) \end{array}$$

↳ PPO to GRPO

↳ keep all the clipping / policy ratio

↳ remove value for

$$A(x, y) = \frac{r(x, y) - \mu_x}{\sigma_x}$$

$\mu_x \rightarrow \text{mean}$   
 $\sigma_x \rightarrow \text{std}$

↳ group normalization

↳ GRPO + verifiable reward + No KL term

↳ only have to store one big model in GPU mem.  $p_{\theta}$

↳ RLVR can incentivize "thinking"  
"reasoning"

↳ Prompt  
response: <sup>reasoning chain</sup> <think> ... </think> <sup>"thought"</sup>  
<sup>↓</sup>  
<answer> 32 </answer>

↳ increasing test-time compute  
improves model perf