Exam review

Logistics

- ~40q, all multiple choice
- primarily cover 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

method

- full model tuning
- PEFT
  - ↳ LoRA
  - ↳ prompt tuning

data

- ↳ single task
- ↳ instruction tuning
- ↳ RAFT vs. best of N sampling

Limitations

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

RL:

- RLHF
  - reward model trained on human pref
    - Bradley-Terry model
- vs.
- RLVR
  - verifiable reward
    - typically done on math / code
    - harder to perform reward hacking

- Algorithms
  - REINFORCE → reward: $r(x,y) - b(x)$ (Advantage)
  - 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

$$\rho_i = \frac{P_\theta(y_i|x)}{P_{\theta_{old}}(y_i|x)} \Bigg\} \text{ trust region } \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)$$

↳ PPO to GRPO

↳ keep all the clipping / policy ratio

↳ remove value fn

$$A(x,y) = \frac{r(x,y) - \mu_x}{\sigma_x}$$ (μ_x → mean, σ_x → stdev)

↳ group normalization

↳ GRPO + verifiable reward + No KL term

↳ only have to store one big model in GPU mem. P_θ

↳ RLVR can incentivize "thinking" "reasoning"

↳ Prompt

response: reasoning chain "thoughts"

<think> ... </think>

<answer> 32 </answer>

↳ increasing test-time compute improves model perf