Instructions: Select the best answer(s). Multi-select where noted.

Q1. **(Multi-select)** Which are **true** about pretraining in LLMs?

A. Maximizes log-likelihood of a token sequence given its prefix

B. Results in instruction-following behavior directly

C. Produces base models for downstream fine-tuning

D. Optimizes human preference rankings

Q2. Instruction tuning is most effective when:

A. The model classifies instruction types

B. Fine-tuned on diverse input-output instruction pairs

C. The model is distilled first

D. Loss is replaced with an RL reward signal

Q3. **(Multi-select)** In RLHF, the reward model is trained by:

A. Cross-entropy on token prediction

B. Learning scalar rewards from preference comparisons

C. Binary classification between preferred and dispreferred responses

D. RL training from scratch

Q4. PPO is used **after**:

A. Pretraining and before instruction tuning

B. Reward model training from preferences

C. Direct preference optimization

D. Few-shot prompting

Q5. PPO uses this to avoid destructive updates:

A. KL regularization

B. Entropy bonus

C. Value baseline

D. Clipped probability ratio

Q6. DPO avoids reward modeling by:

A. Actor-critic methods

B. Contrastive loss on preference pairs

C. Sampling rewards until saturation

D. KL-minimization against deterministic function

Q7. **(Multi-select)** Advantages of DPO over PPO-based RLHF:

A. Simpler, no rollouts required

B. Maximum likelihood training from preferences

C. Token-level reward shaping

D. End-to-end gradient optimization from preference pairs

Q8. What is the key idea behind Group Relative Preference Optimization (GRPO) introduced by DeepSeek?

A. Using PPO with dynamic temperature scaling

B. Token-level reward shaping via causal masking

C. Learning from groups of ranked responses rather than only pairs

D. Regularizing with KL divergence to the base model

Q9. In DPO, which objective is correct?

A. $\log \pi_\theta(y^+) - \log \pi_\theta(y^-)$

B. $\mathrm{KL}(\pi_\theta \parallel \pi_{\mathrm{ref}})$

C. $-\log\left(\dfrac{e^{\beta \log \pi_\theta(y^+)}}{e^{\beta \log \pi_\theta(y^+)} + e^{\beta \log \pi_\theta(y^-)}}\right)$

D. Both A and C

Q10. Which of the following is a drawback of Direct Preference Optimization (DPO) compared to PPO?

A. Requires separate reward model training

B. Involves costly sampling rollouts

C. Lacks token-level reward shaping flexibility

D. Suffers from off-policy instability