Self-Reflect SFT + RL
Ref. Sol. SFT + RL

Zero Adv. Ratio vs Gradient steps