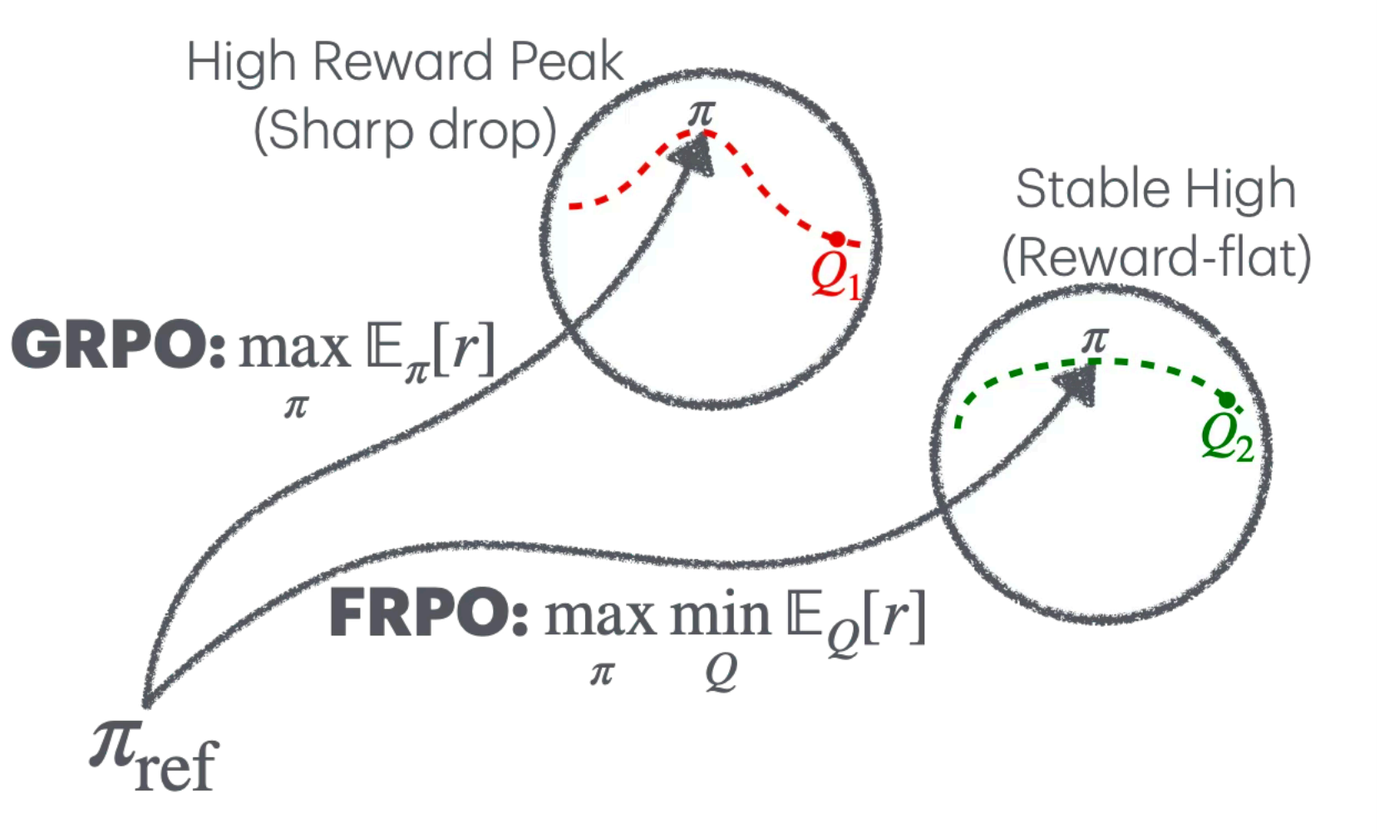
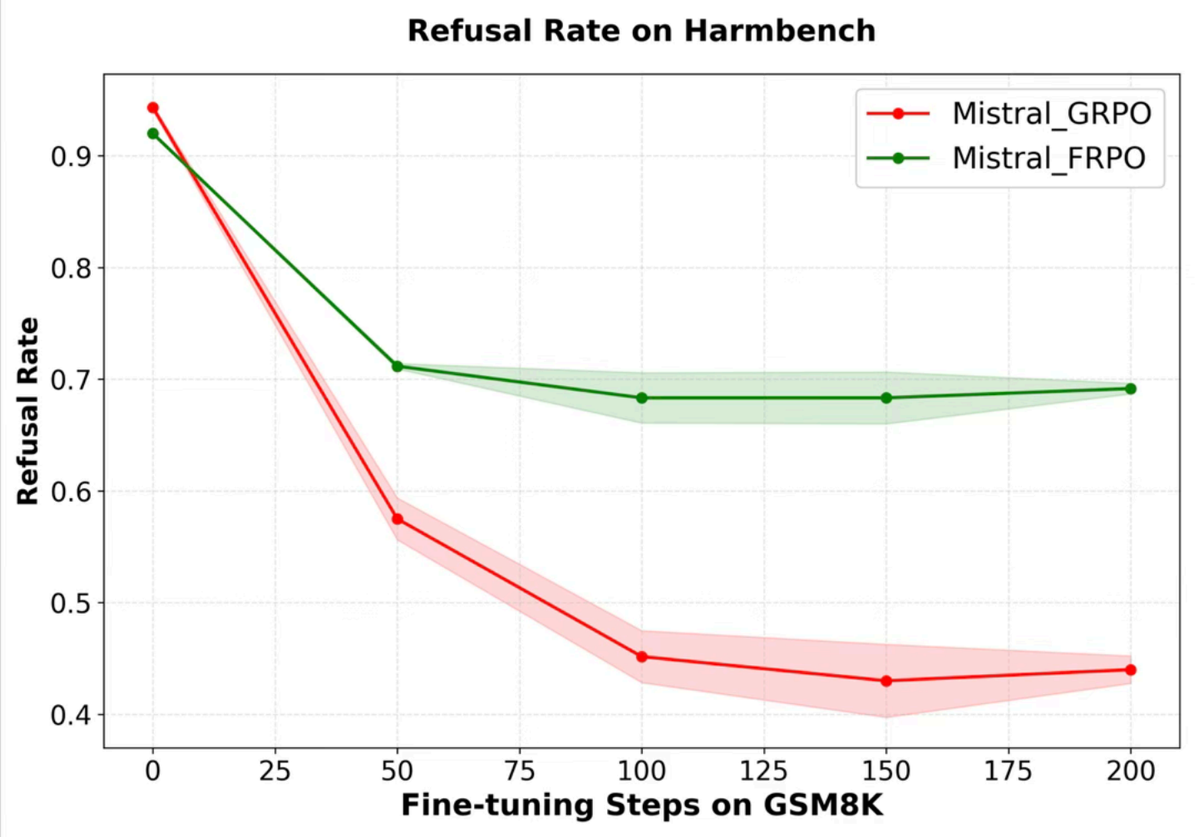

High Reward Peak
(Sharp drop)
Stable High
(Reward-flat)
GRPO: $\max_{\pi} \mathbb{E}_{\pi}[r]$
FRPO: $\max_{\pi} \min_{Q} \mathbb{E}_{Q}[r]$
$\pi_{\text{ref}}$
$\pi$
$Q_1$
$\pi$
$Q_2$


Refusal Rate on Harmbench
Mistral_GRPO
Mistral_FRPO
Refusal Rate
0.9
0.8
0.7
0.6
0.5
0.4
0
25
50
75
100
125
150
175
200
Fine-tuning Steps on GSM8K