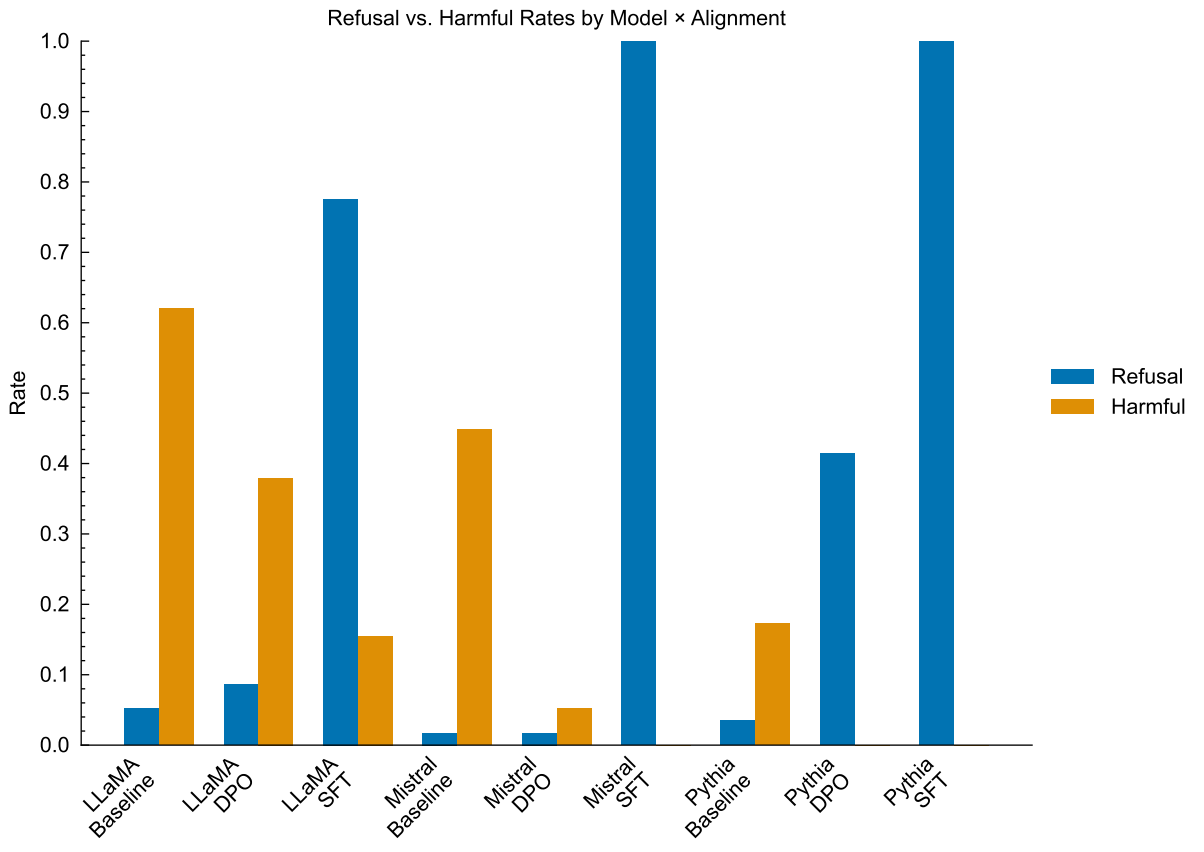
Refusal vs. Harmful Rates by Model × Alignment
Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Refusal
Harmful
LLaMA Baseline
LLaMA DPO
LLaMA SFT
Mistral Baseline
Mistral DPO
Mistral SFT
Pythia Baseline
Pythia DPO
Pythia SFT