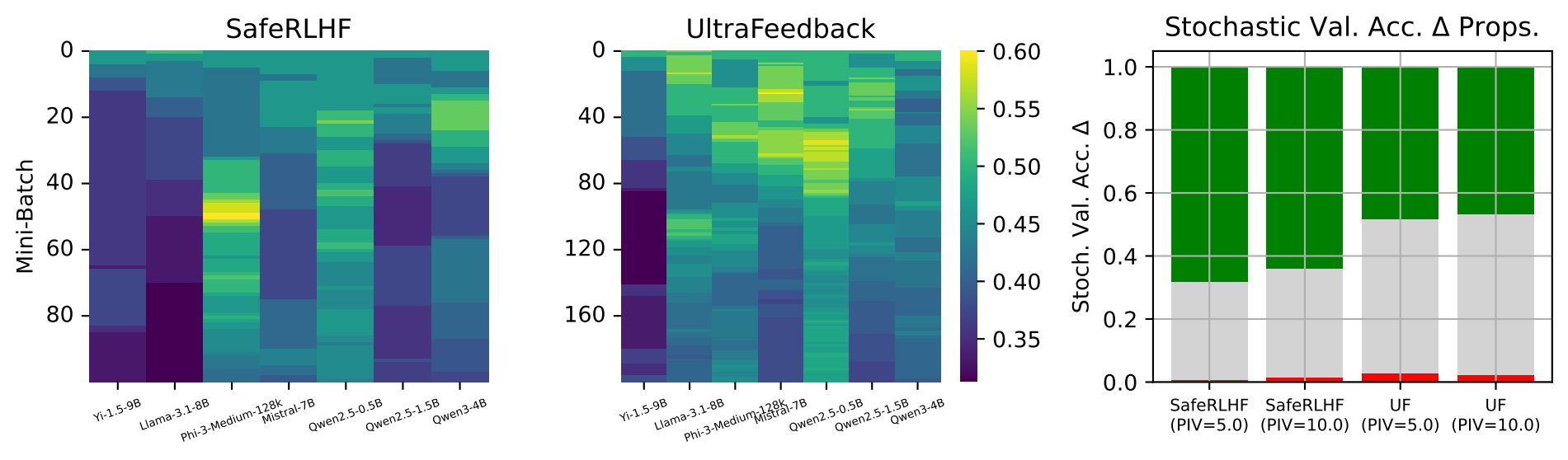
SafeRLHF
UltraFeedback
Mini-Batch
0
20
40
60
80
0
40
80
120
160
0.60
0.55
0.50
0.45
0.40
0.35
Yi-1.5-9B
Llama-3.1-8B
Phi-3-Medium-128k
Mistral-7B
Qwen2.5-0.5B
Qwen2.5-1.5B
Qwen3-4B
Stochastic Val. Acc. Δ Props.
Stoch. Val. Acc. Δ
1.0
0.8
0.6
0.4
0.2
0.0
SafeRLHF (PIV=5.0)
SafeRLHF (PIV=10.0)
UF (PIV=5.0)
UF (PIV=10.0)