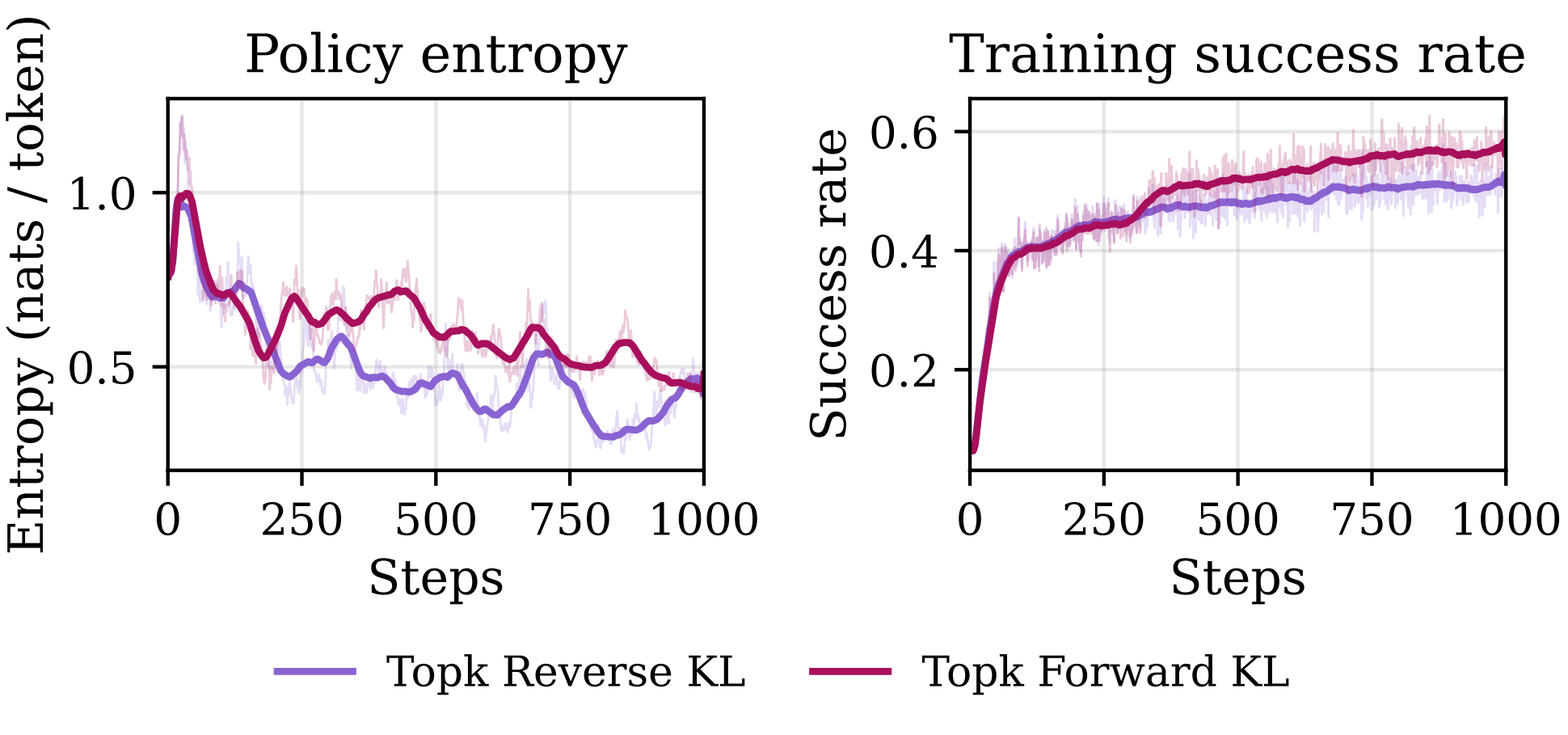

Policy entropy
Entropy (nats / token)
1.0
0.5
0
250
500
750
1000
Steps
Training success rate
Success rate
0.6
0.4
0.2
0
250
500
750
1000
Steps
Topk Reverse KL
Topk Forward KL