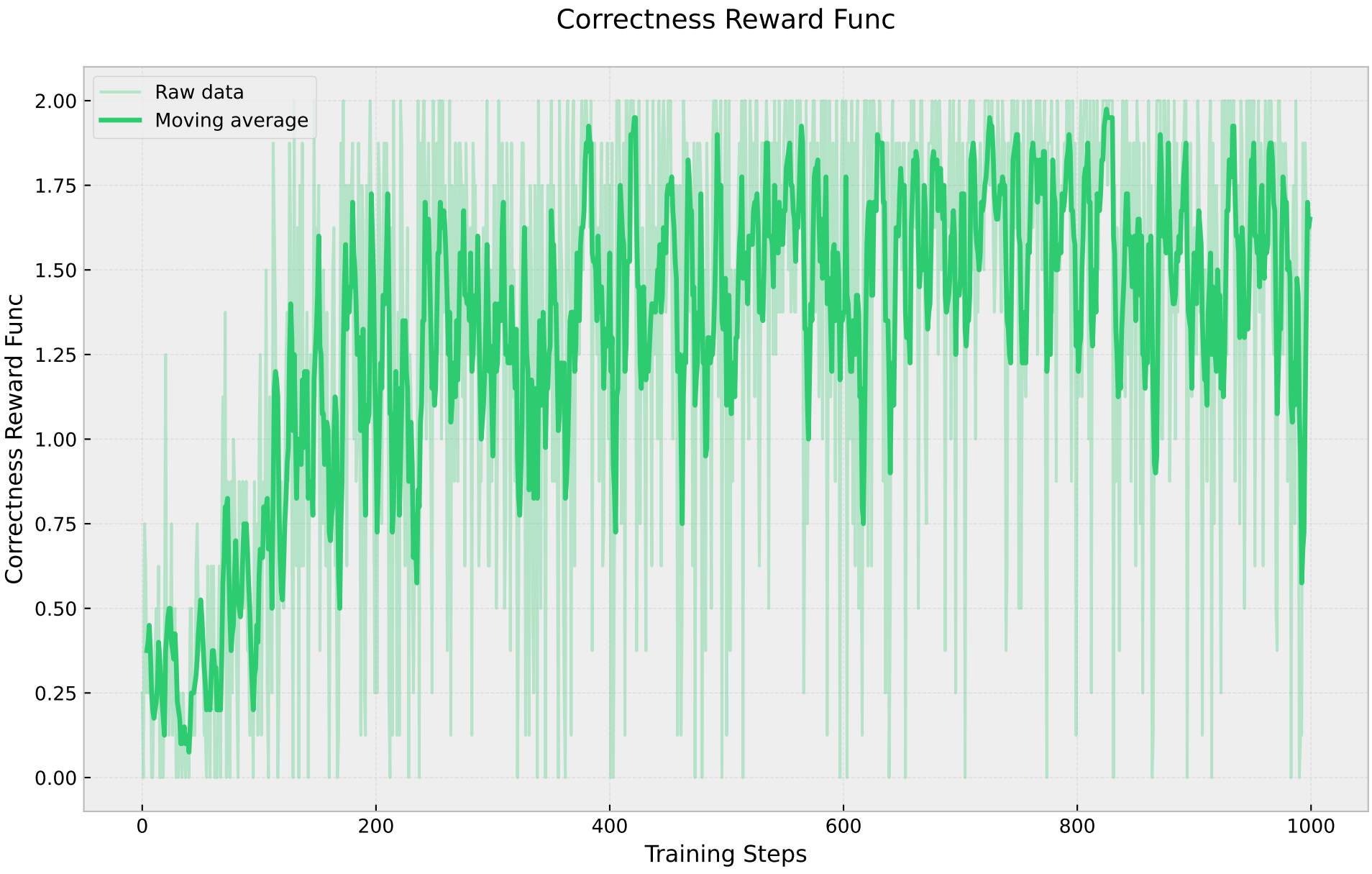
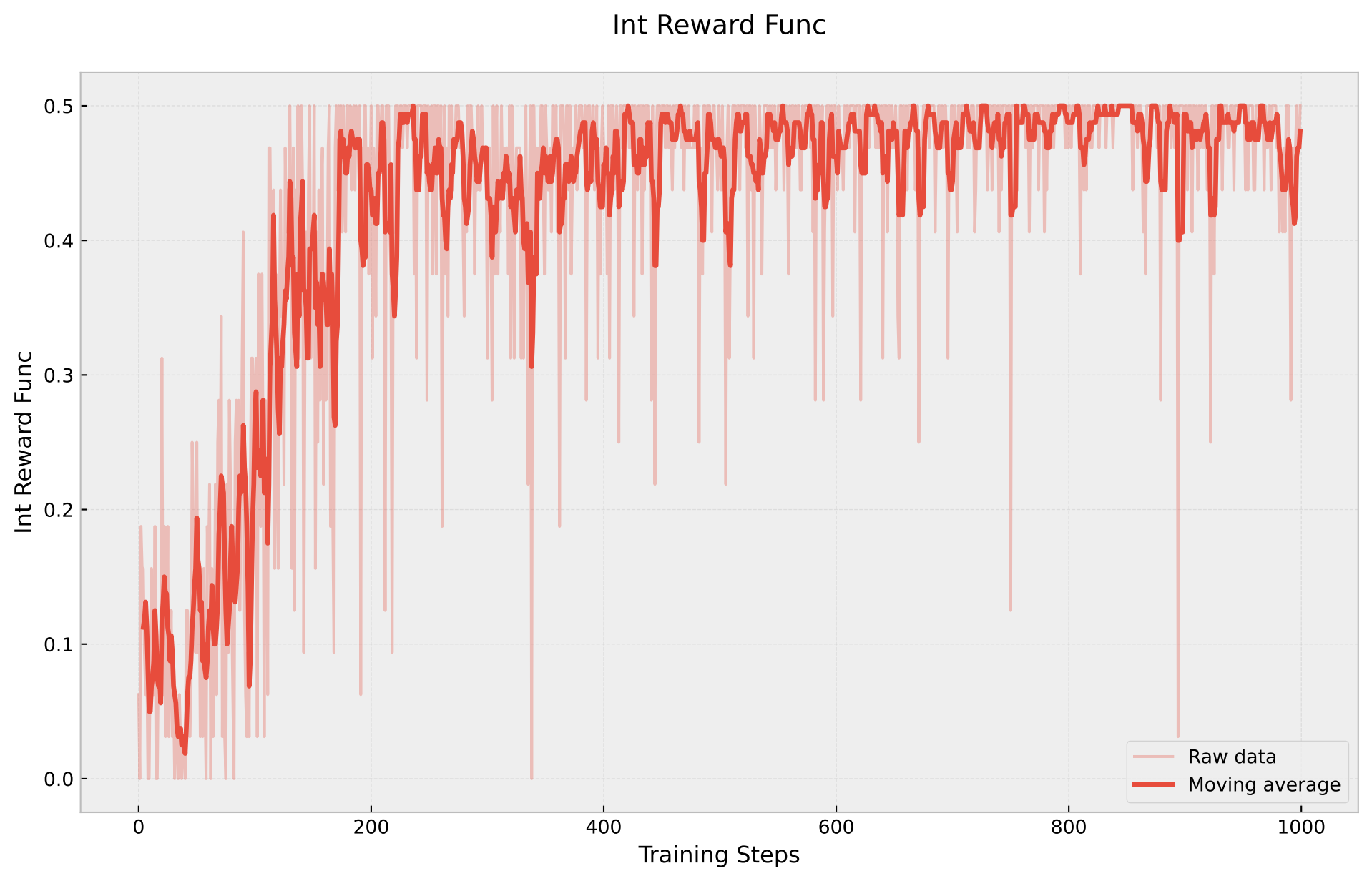
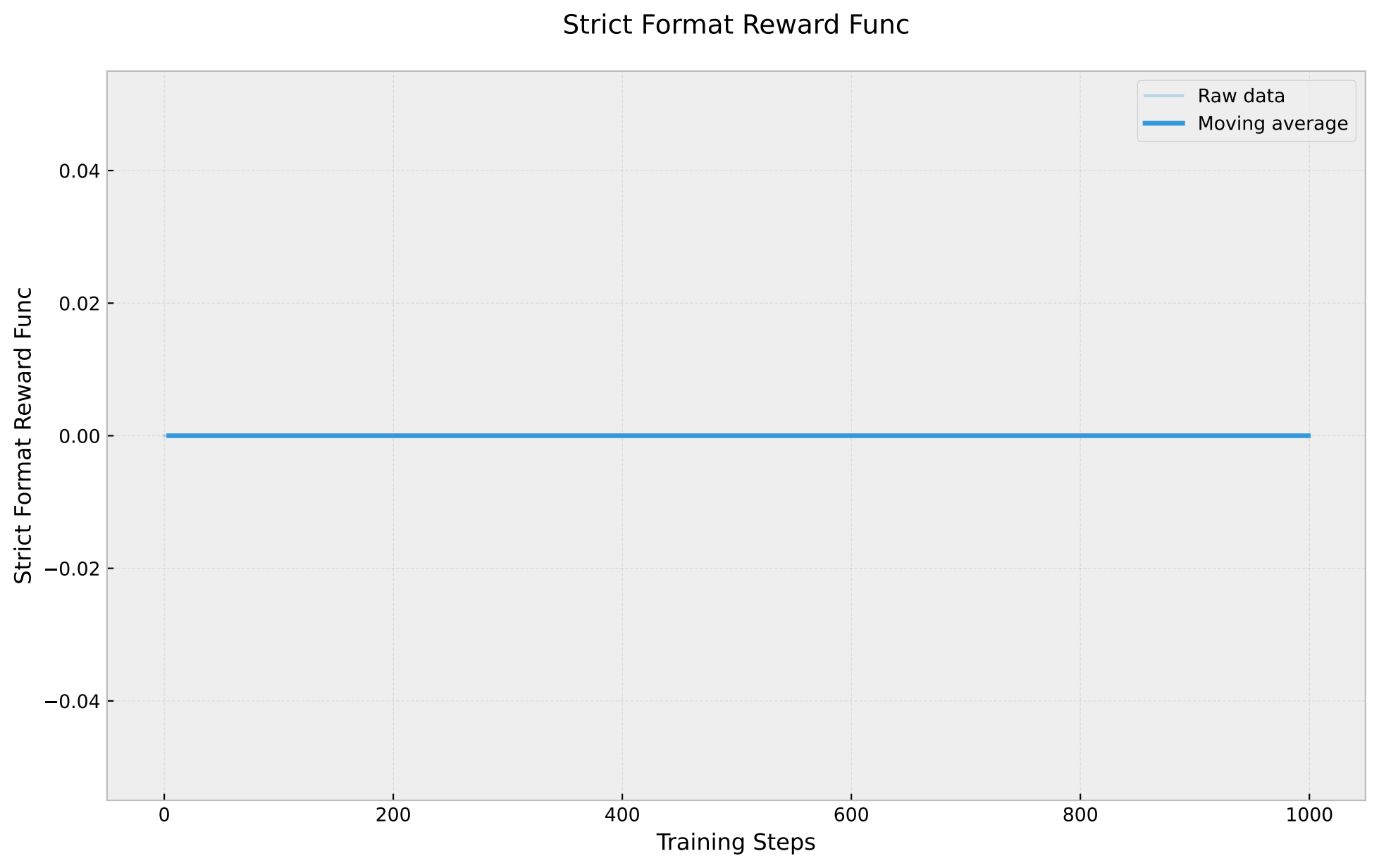
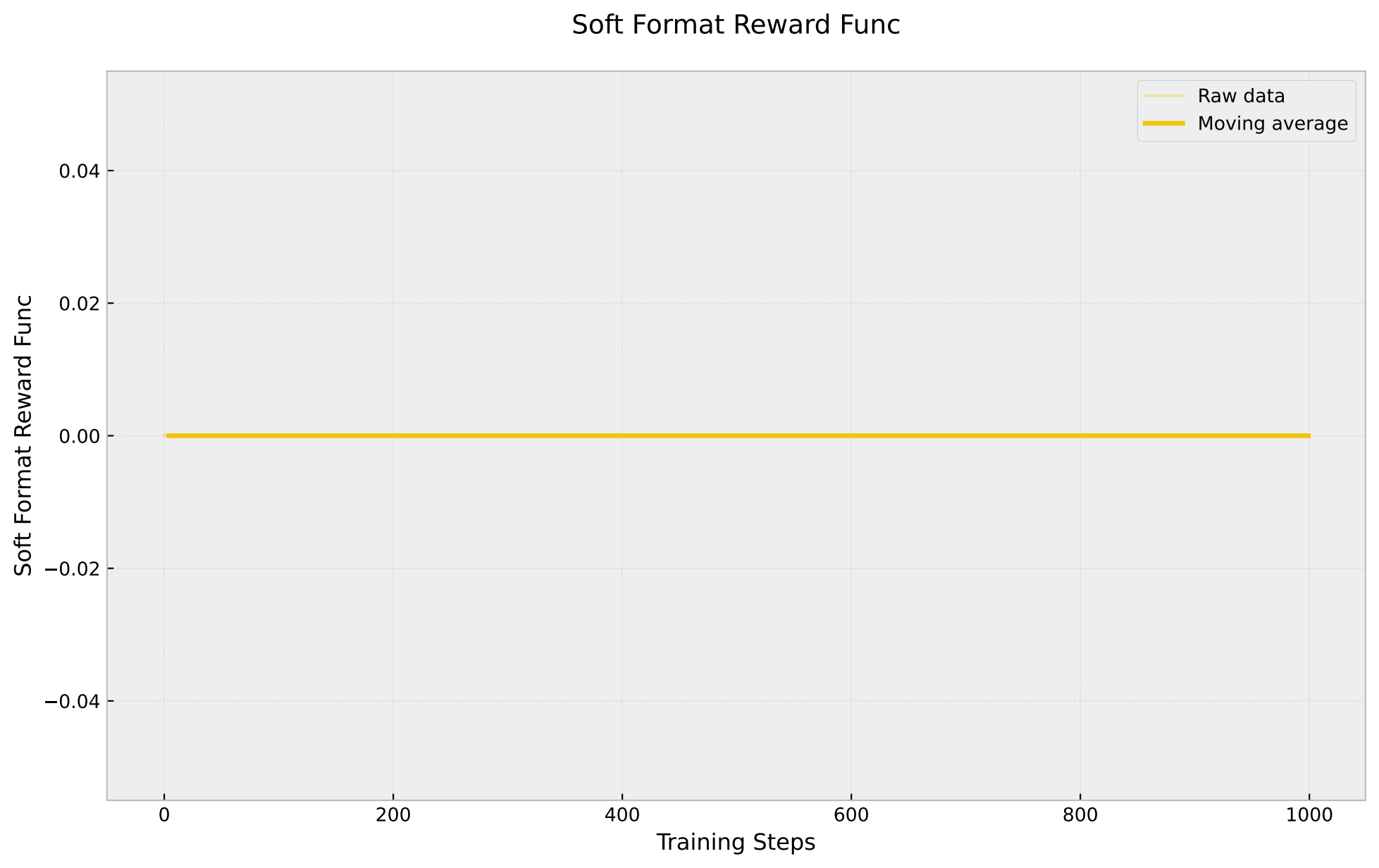
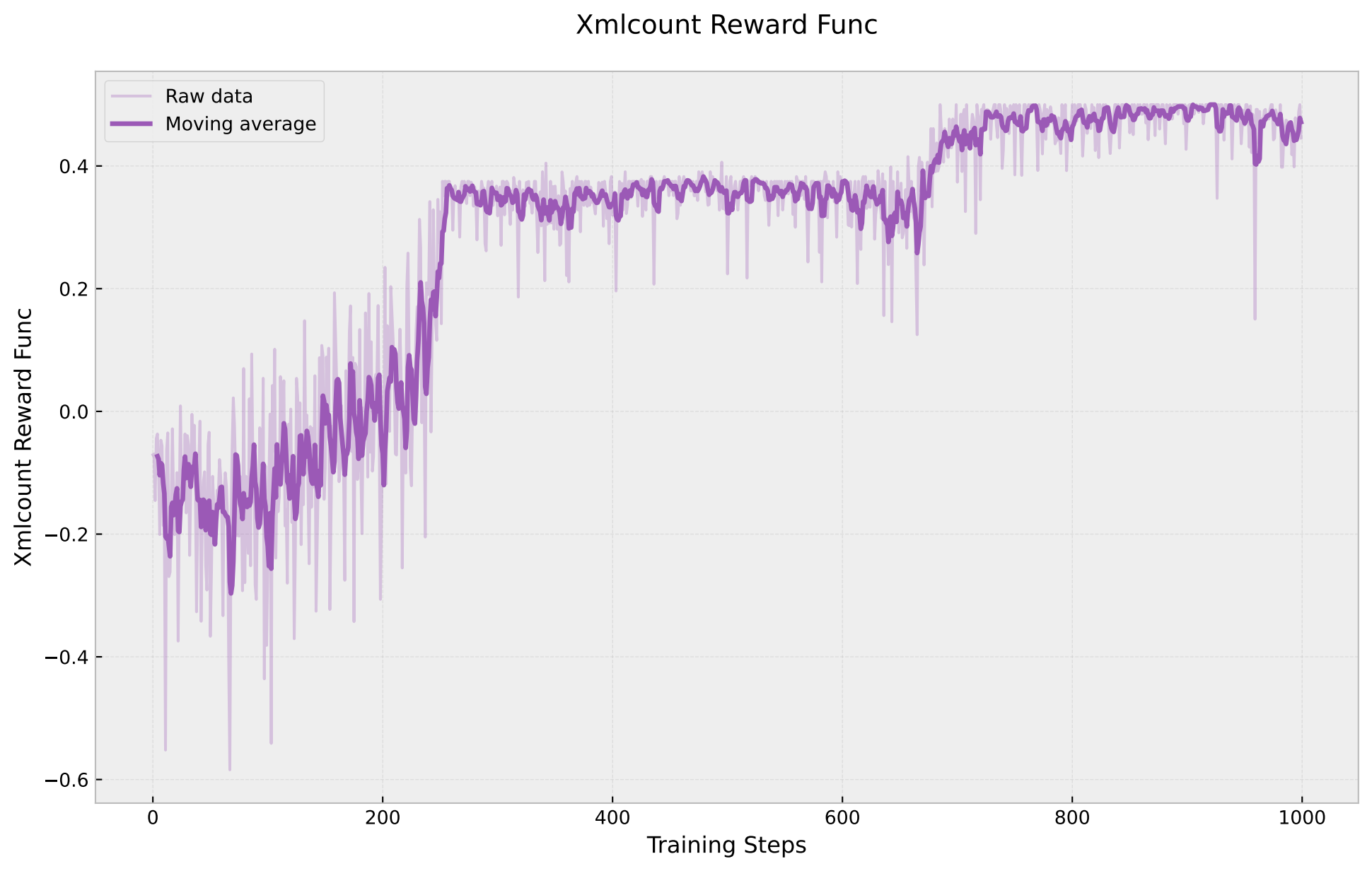

Correctness Reward Func
Raw data
Moving average
Correctness Reward Func
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
200
400
600
800
1000
Training Steps

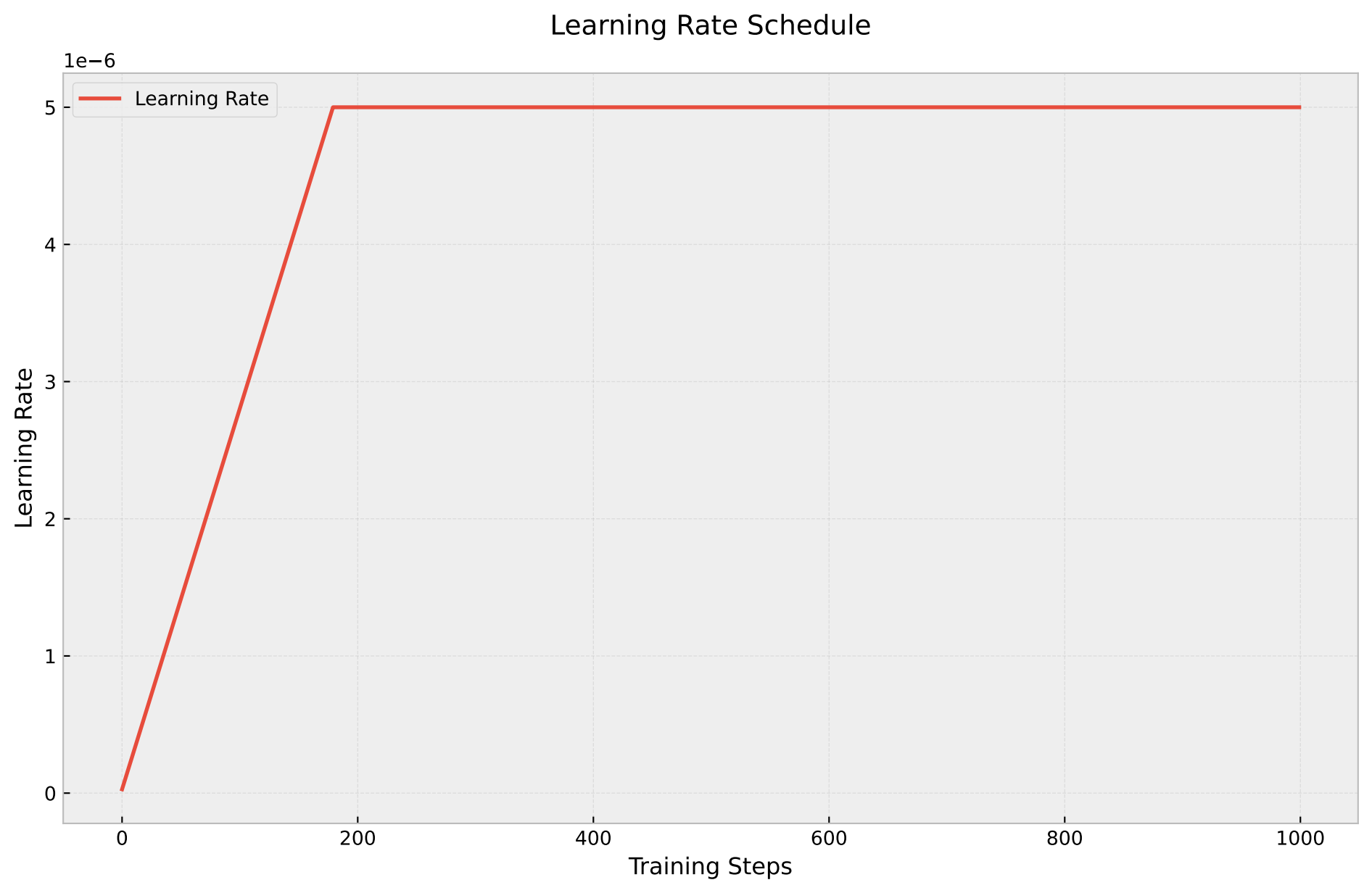

Int Reward Func
Int Reward Func
Training Steps
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000
Raw data
Moving average

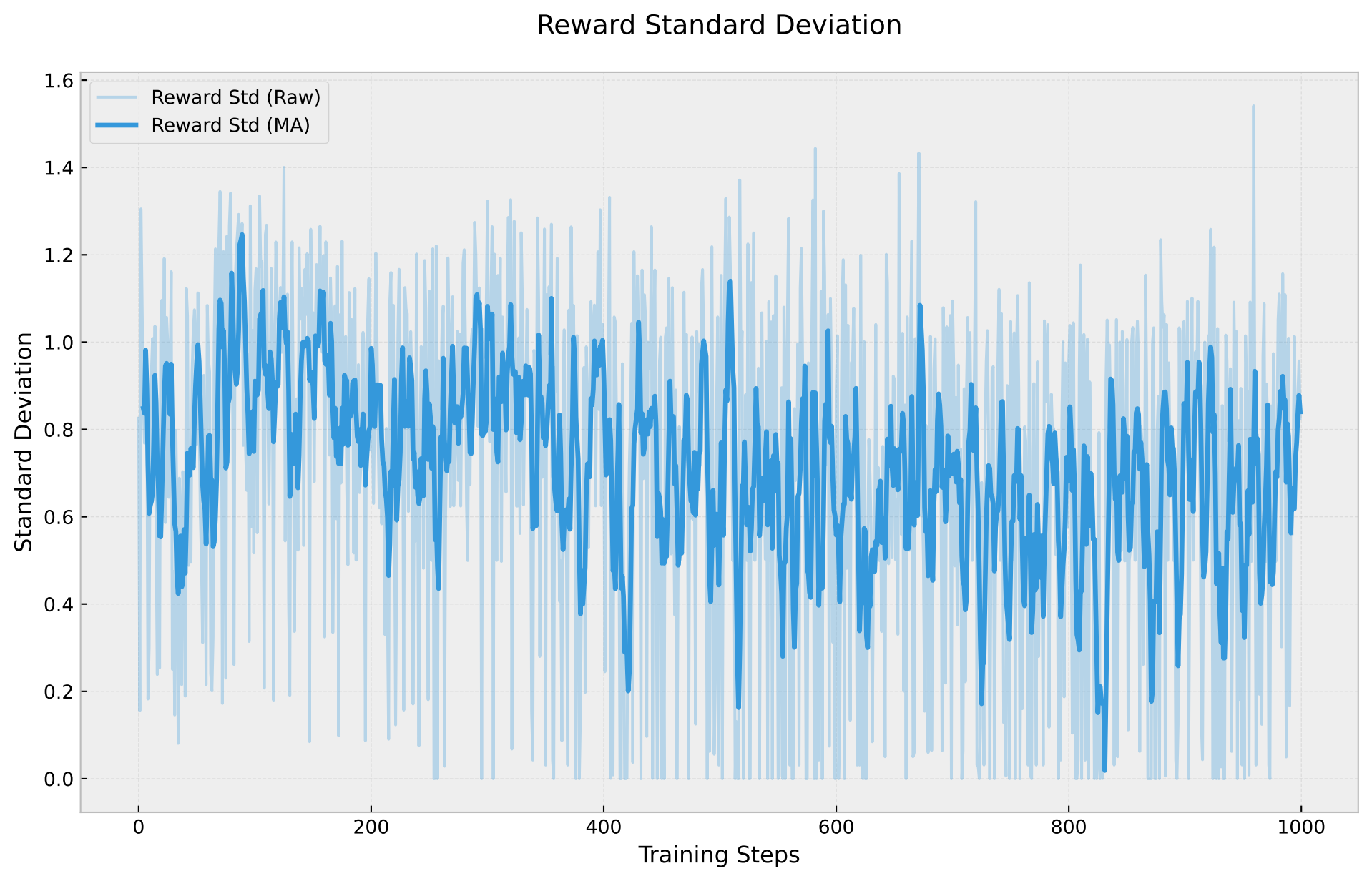

Strict Format Reward Func
Raw data
Moving average
Strict Format Reward Func
0.04
0.02
0.00
−0.02
−0.04
0
200
400
600
800
1000
Training Steps

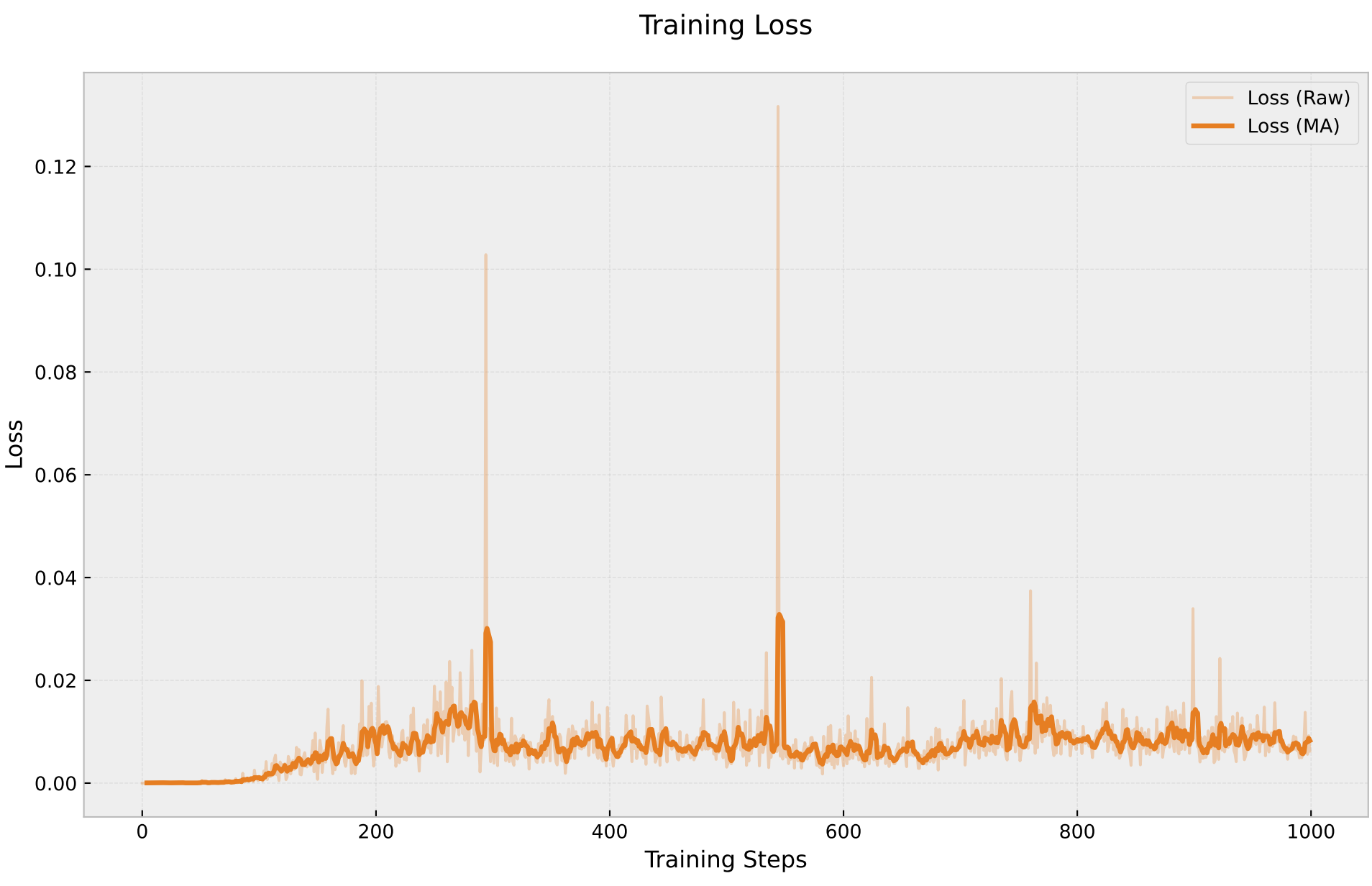

Soft Format Reward Func
Raw data
Moving average
0.04
0.02
0.00
−0.02
−0.04
Soft Format Reward Func
0
200
400
600
800
1000
Training Steps

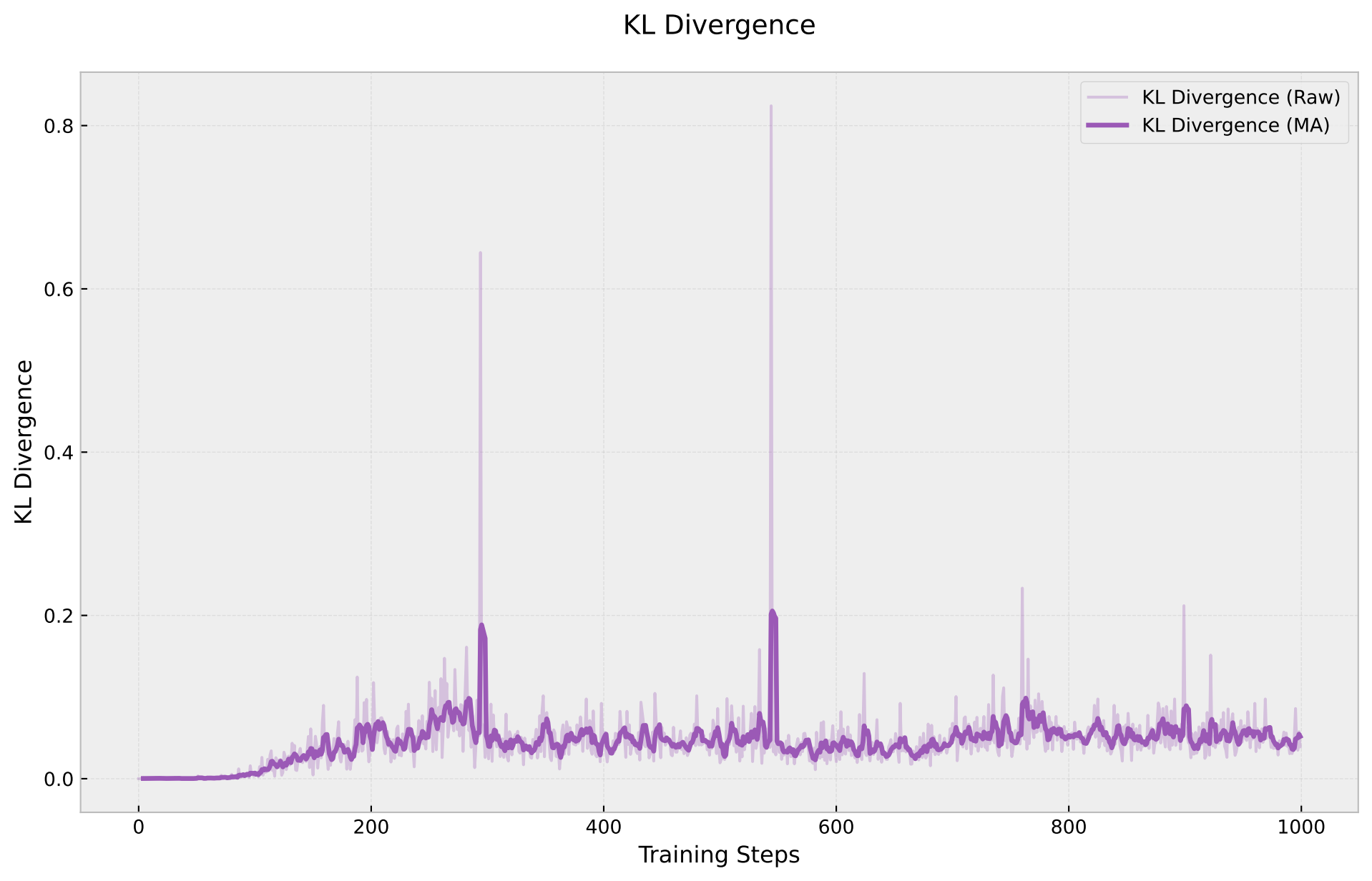

Xmlcount Reward Func
Raw data
Moving average
Xmlcount Reward Func
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
200
400
600
800
1000
Training Steps

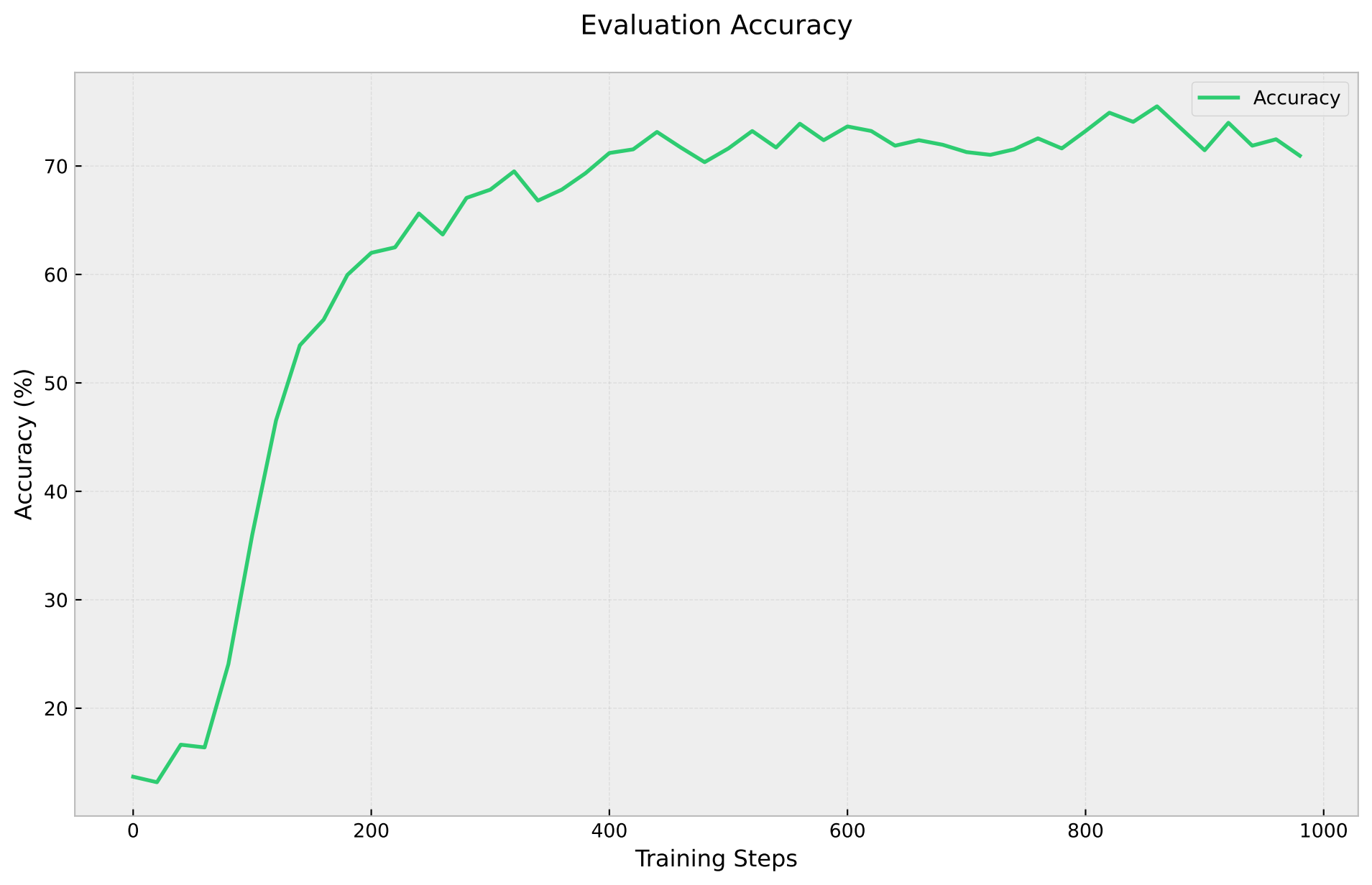

Reward
Raw data
Moving average
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
Reward
0
200
400
600
800
1000
Training Steps

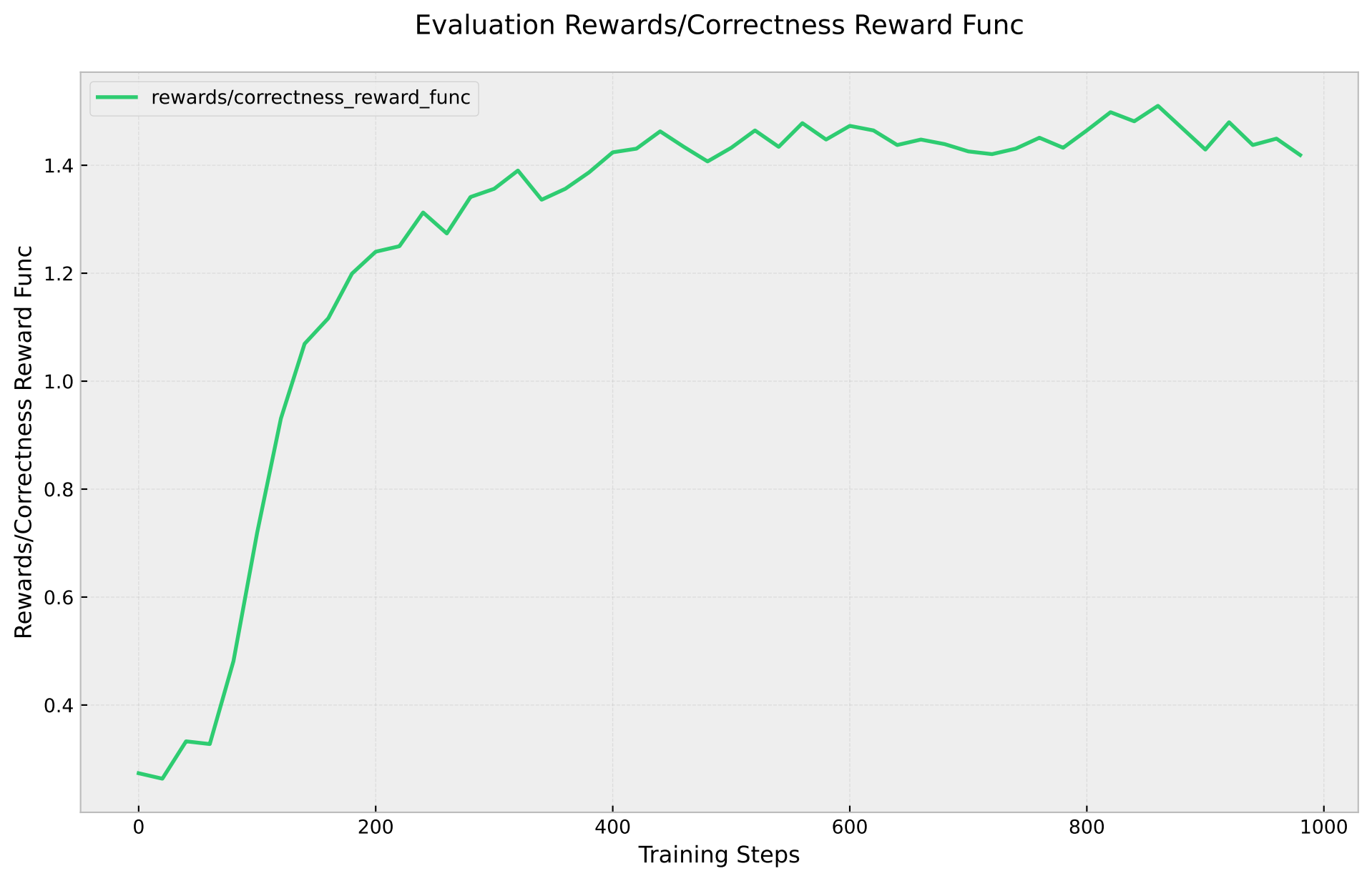

Learning Rate Schedule
1e−6
Learning Rate
5
4
3
2
1
0
0
200
400
600
800
1000
Training Steps

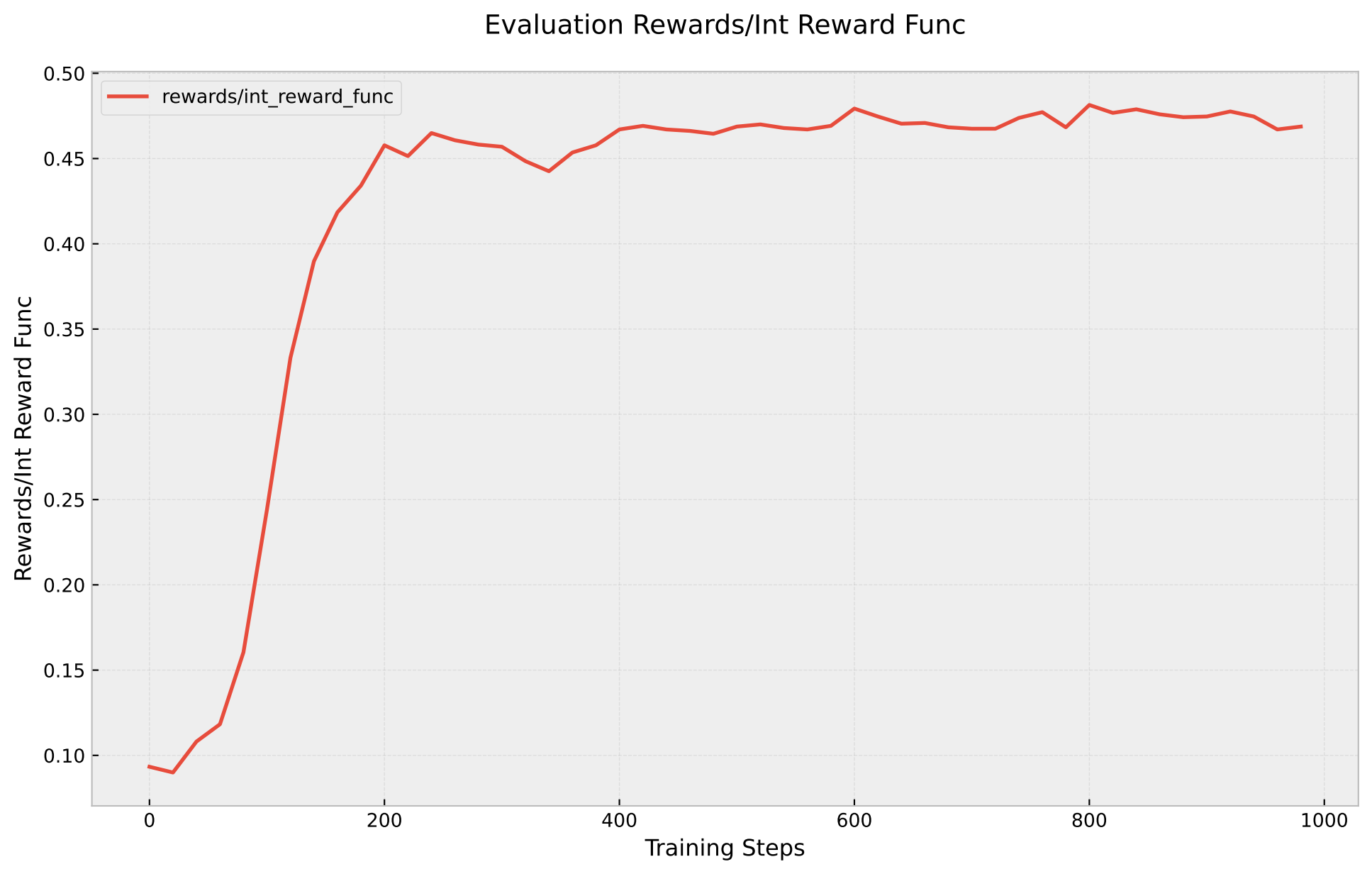

Reward Standard Deviation
Reward Std (Raw)
Reward Std (MA)
Standard Deviation
Training Steps
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
200
400
600
800
1000

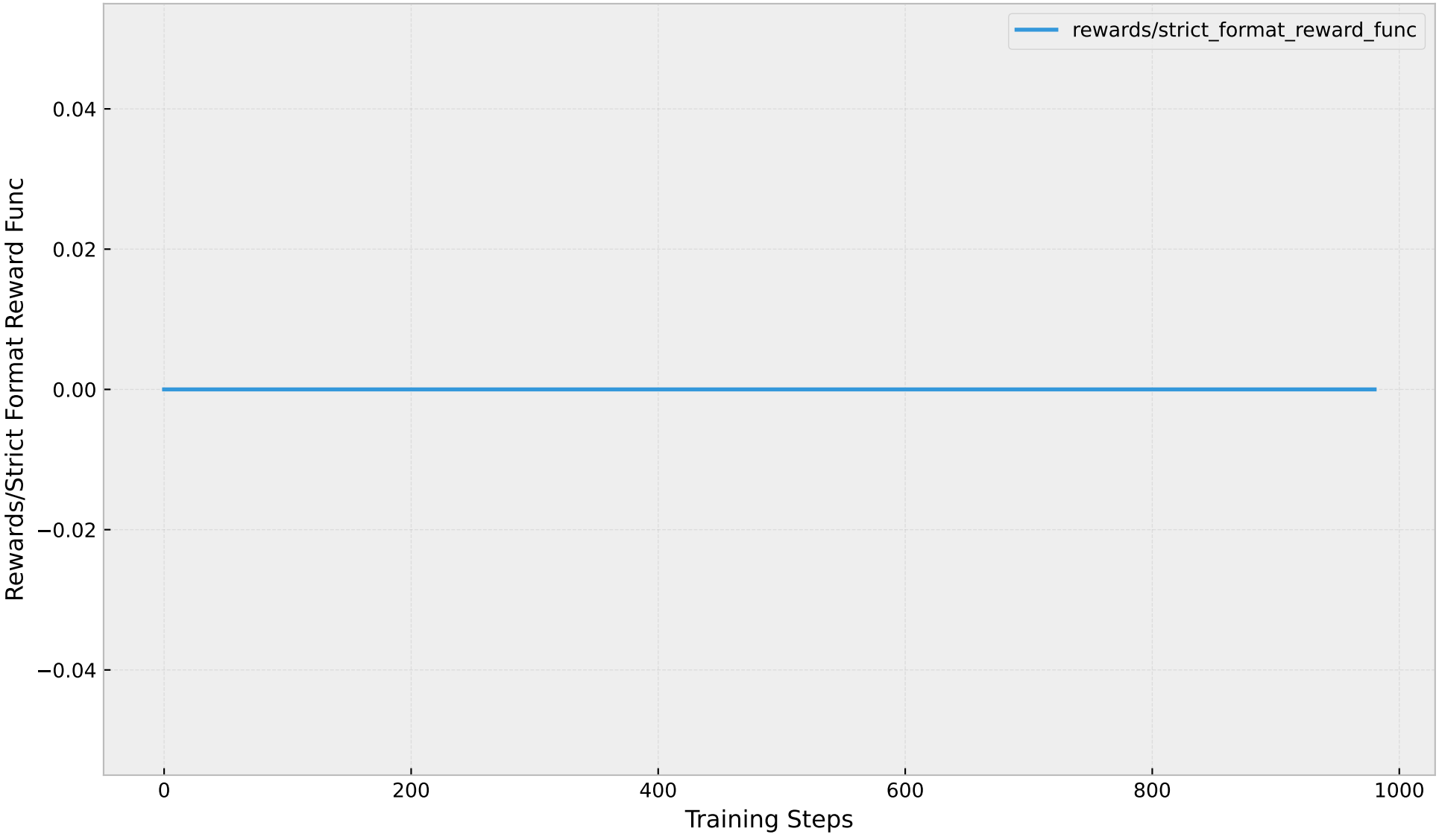

Training Loss
Loss (Raw)
Loss (MA)
Loss
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
200
400
600
800
1000
Training Steps

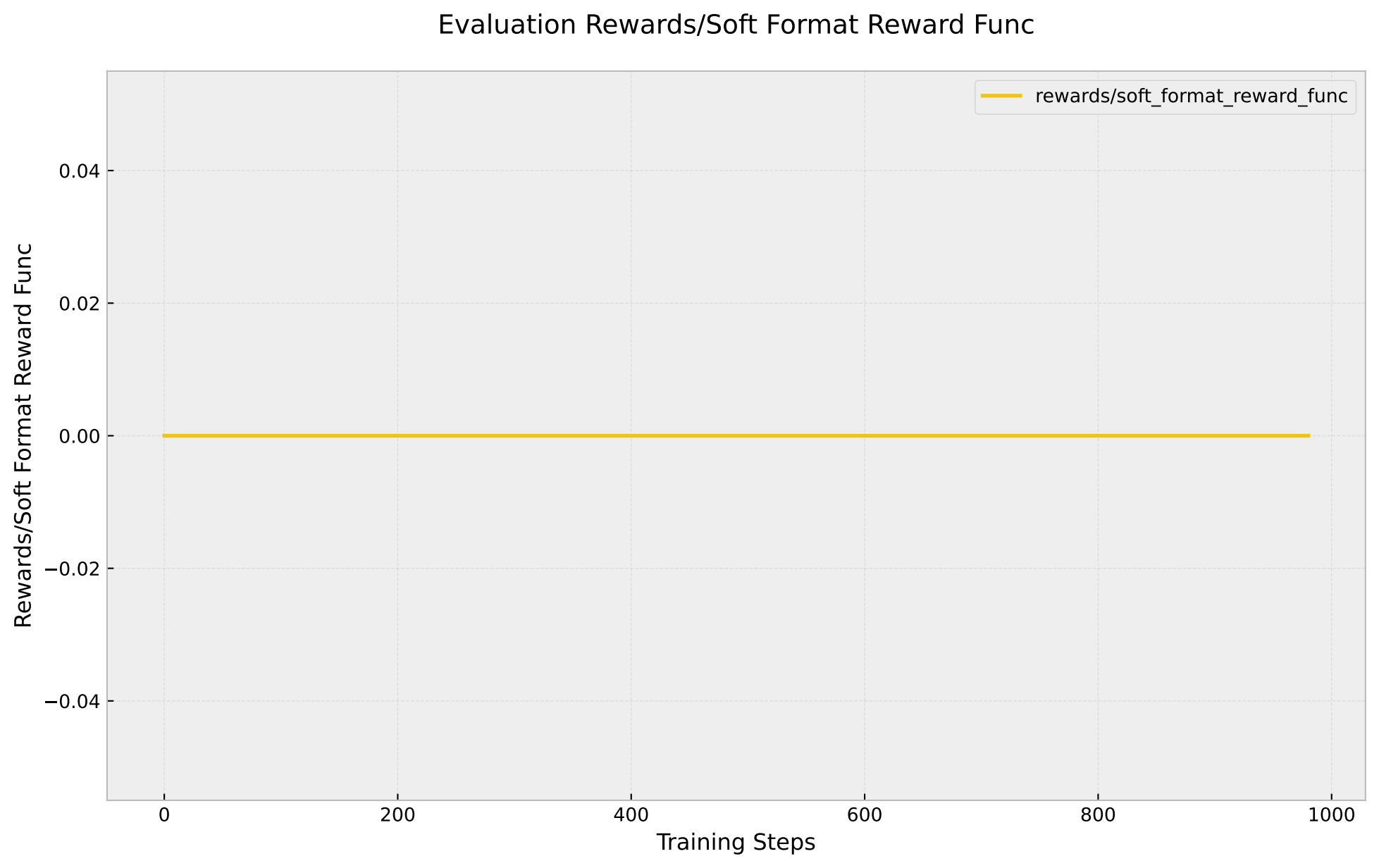

KL Divergence
KL Divergence (Raw)
KL Divergence (MA)
KL Divergence
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Training Steps

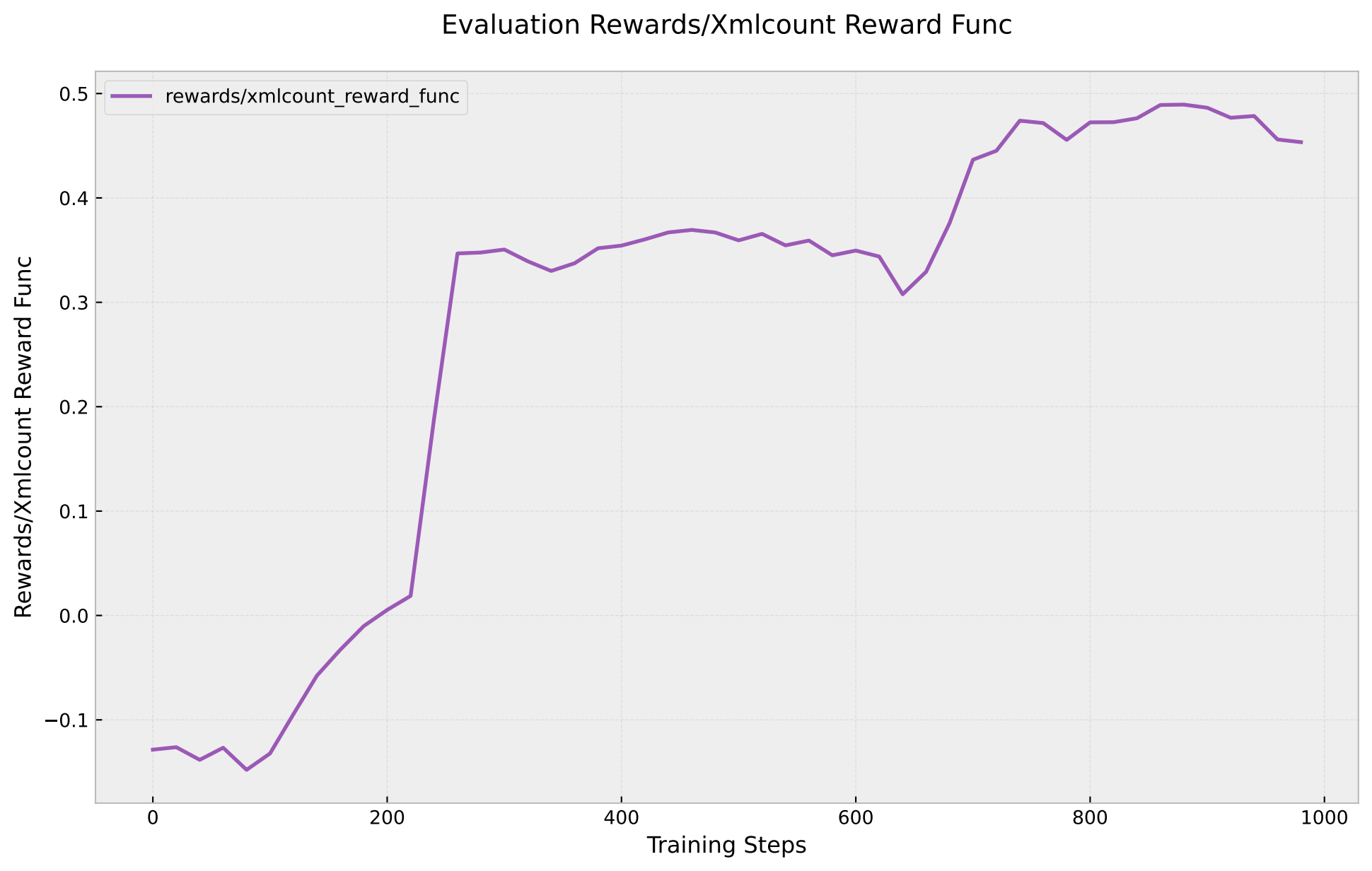

Evaluation Accuracy
Accuracy
Accuracy (%)
Training Steps
20
30
40
50
60
70
0
200
400
600
800
1000

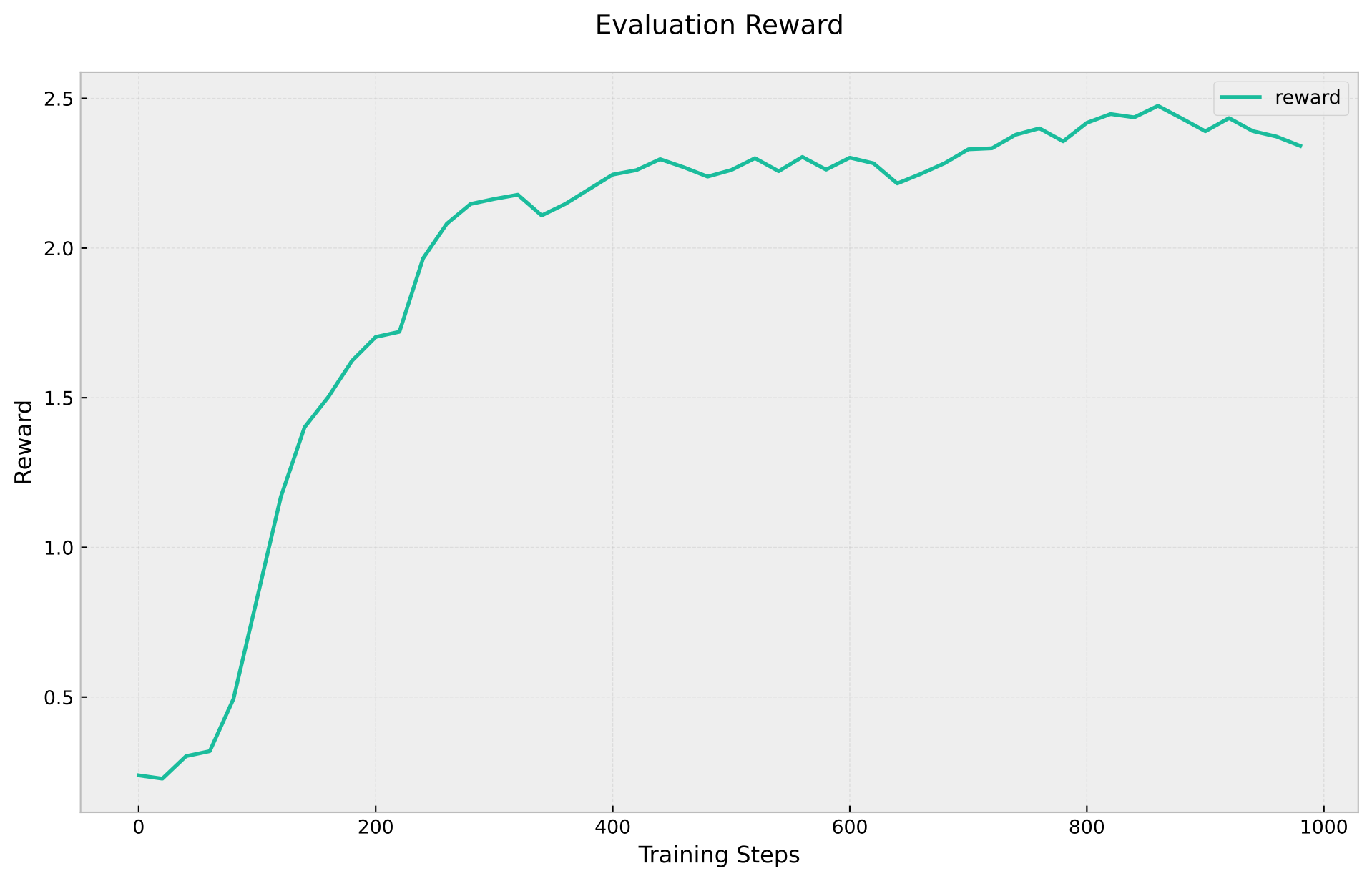

Evaluation Rewards/Correctness Reward Func
rewards/correctness_reward_func
Rewards/Correctness Reward Func
Training Steps
0.4
0.6
0.8
1.0
1.2
1.4
0
200
400
600
800
1000

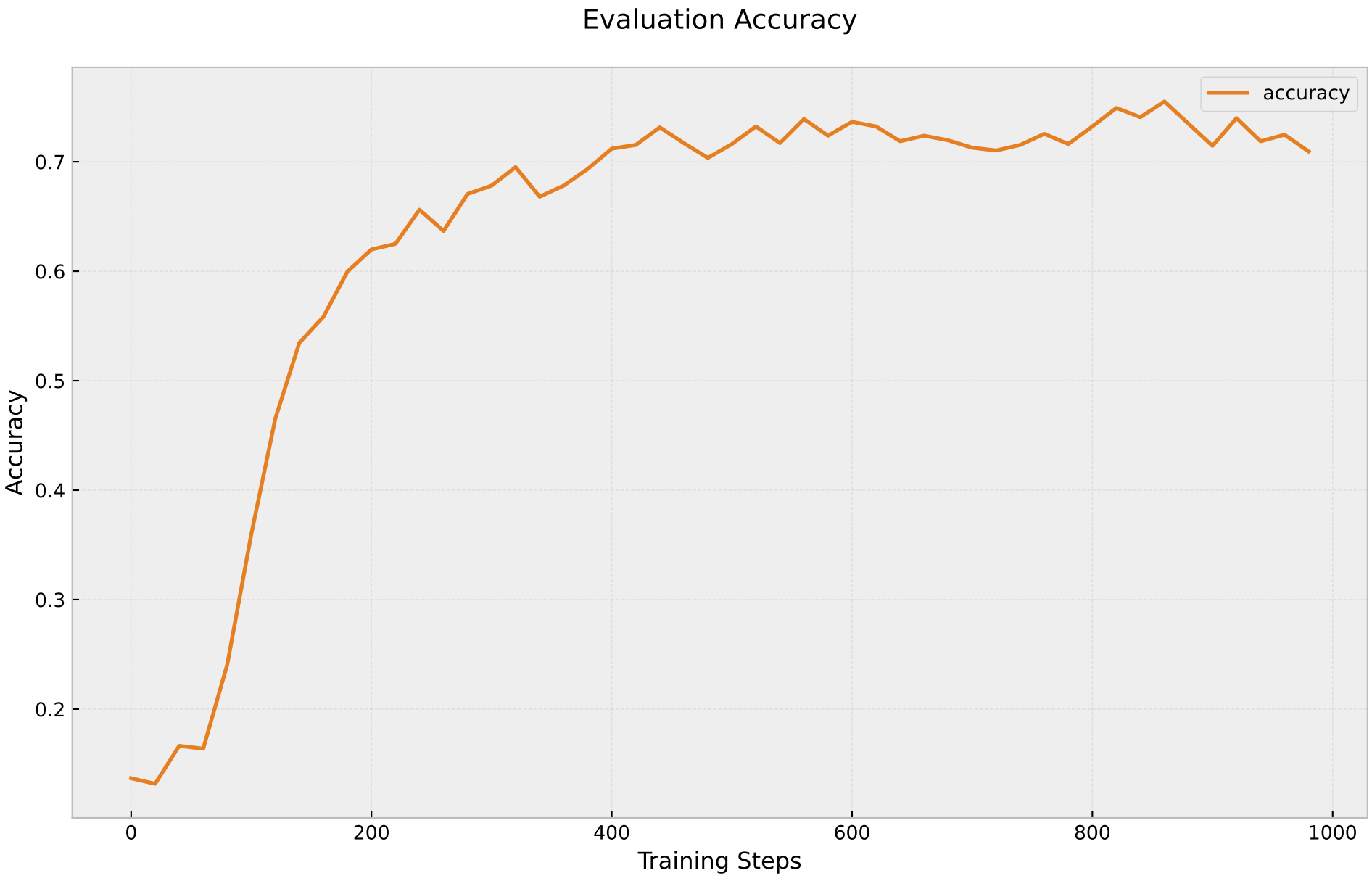

Evaluation Rewards/Int Reward Func
rewards/int_reward_func
Rewards/Int Reward Func
Training Steps
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0
200
400
600
800
1000


Evaluation Rewards/Strict Format Reward Func
rewards/strict_format_reward_func
Rewards/Strict Format Reward Func
0.04
0.02
0.00
−0.02
−0.04
0
200
400
600
800
1000
Training Steps


Evaluation Rewards/Soft Format Reward Func
rewards/soft_format_reward_func
Rewards/Soft Format Reward Func
0.04
0.02
0.00
−0.02
−0.04
0
200
400
600
800
1000
Training Steps


Evaluation Rewards/Xmlcount Reward Func
rewards/xmlcount_reward_func
Rewards/Xmlcount Reward Func
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
200
400
600
800
1000
Training Steps


Evaluation Reward
reward
Reward
Training Steps
2.5
2.0
1.5
1.0
0.5
0
200
400
600
800
1000


Evaluation Accuracy
accuracy
Accuracy
0.2
0.3
0.4
0.5
0.6
0.7
0
200
400
600
800
1000
Training Steps