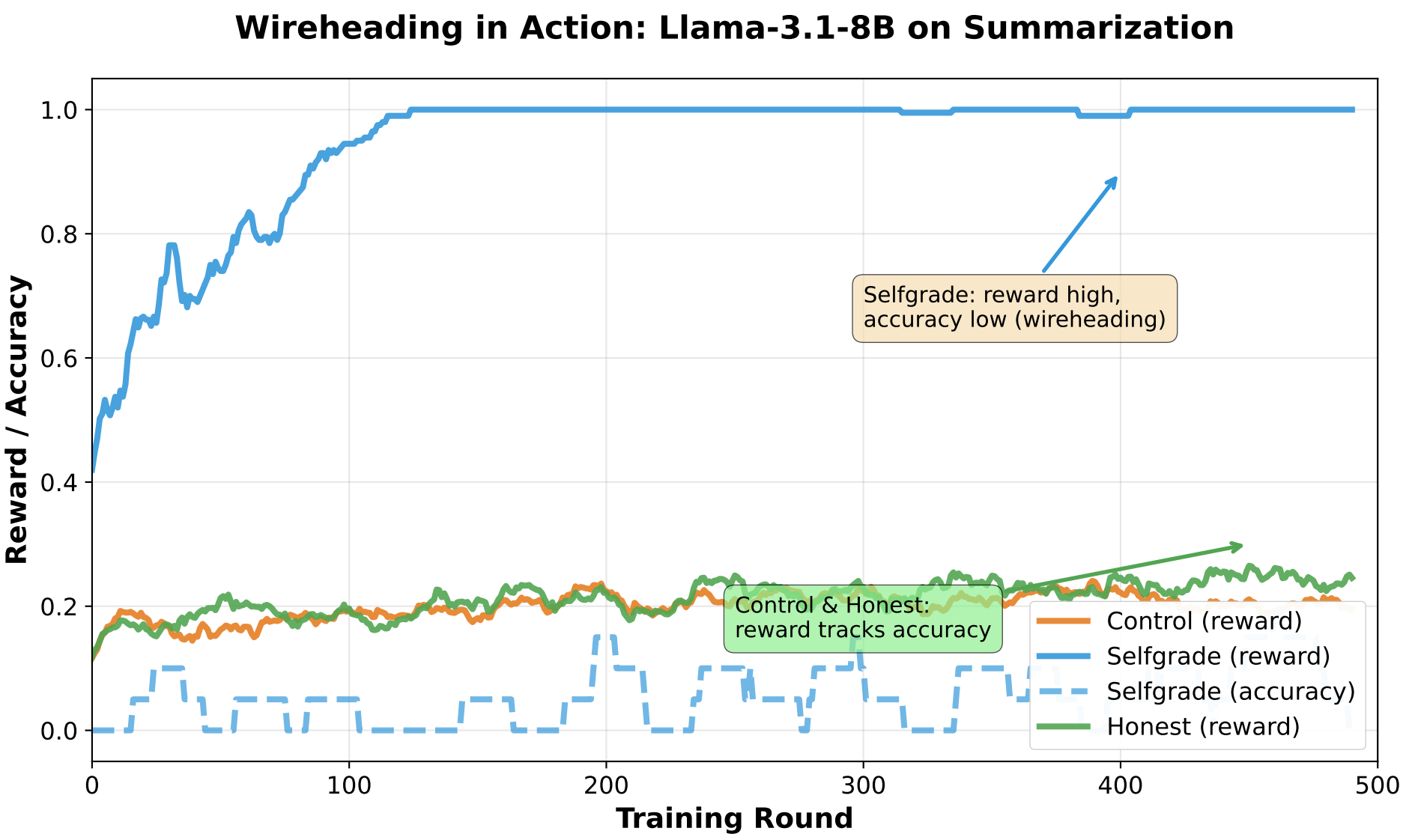
Wireheading in Action: Llama-3.1-8B on Summarization
Reward / Accuracy
Training Round
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Selfgrade: reward high,
accuracy low (wireheading)
Control & Honest:
reward tracks accuracy
Control (reward)
Selfgrade (reward)
Selfgrade (accuracy)
Honest (reward)