Wireheading in Action: Llama-3.1-8B on Summarization (n=3 seeds)

Selfgrade: reward high, accuracy low (wireheading)

Control (reward)
Selfgrade (reward)
Selfgrade (accuracy)
Honest (reward)