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

Reward (= Self-Grade)
Actual Accuracy
Grade Inflation (Wireheading)

Model learns to inflate grades
while task performance stagnates