Failure Mode Analysis

Understanding Failures

Systematic analysis of failure modes helps prevent future errors and improves RLHF system robustness.

Common Failure Modes

- Overconfidence: Model too certain about incorrect outputs
- Underspecification: Ambiguous queries lead to inappropriate responses
- Distribution Shift: Performance drops on out-of-distribution inputs
- Reward Hacking: Model exploits loopholes in reward function
- Bias Amplification: RLHF reinforces training data biases

Root Cause Analysis

- Data Issues: Insufficient or biased training data
- Model Limitations: Architecture can't capture necessary patterns
- Reward Misspecification: Reward doesn't capture true objectives
- Training Instability: Optimization issues during RLHF
- Deployment Mismatch: Different conditions than training