Safety-Critical RLHF

Medical Error Prevention

RLHF systems in healthcare must have multiple safety layers to prevent errors that could harm patients.

Safety Mechanisms

- Pre-Deployment Testing: Extensive validation before clinical use
- Conservative Defaults: Err on side of caution when uncertain
- Human-in-the-Loop: Require expert confirmation for critical decisions
- Redundant Checks: Multiple independent safety verifications
- Fail-Safe Modes: Graceful degradation when errors detected

Critical Failure Modes

- Medication Errors: Wrong drug, dose, or timing
- Diagnostic Misses: Failing to identify serious conditions
- Inappropriate Reassurance: Downplaying concerning symptoms
- Scope Violations: Advising beyond system competency
- Harmful Content: Suggesting dangerous treatments