Ethical Considerations in Medical RLHF

Ethical Imperatives

RLHF in healthcare must navigate complex ethical terrain, balancing innovation with patient safety and equity.

Key Ethical Concerns

- Bias & Fairness: Ensure equitable performance across demographics
- Transparency: Make AI reasoning understandable to clinicians
- Accountability: Clear responsibility when errors occur
- Privacy: Protect patient data in feedback loops
- Autonomy: Preserve patient and physician decision-making
- Beneficence: Ensure AI improves outcomes for all patients

Bias Mitigation

- Diverse Annotators: Include experts from various backgrounds
- Stratified Evaluation: Test performance across demographic groups
- Fairness Metrics: Monitor for disparate impact
- Bias Audits: Regular third-party fairness assessments
- Corrective Action: Retrain on underrepresented groups