Safety Constraints in Medical RLHF

Why Safety Constraints?

Medical AI requires hard constraints to prevent harmful outputs that could endanger patients, regardless of reward optimization.

Multi-Layer Safety Architecture

Layer 1: Hard Medical Rules

Never violate established clinical guidelines (e.g., contraindications, age restrictions)

Layer 2: Dosage Limits

Enforce safe medication dosing ranges based on patient factors (weight, age, renal function)

Layer 3: Contraindication Checking

Prevent dangerous drug interactions and allergic reactions

Layer 4: Appropriate Scope

Stay within system's trained competency domain