Case Study: Treatment Recommendation System

System Overview

A real-world example of RLHF applied to an AI system that recommends treatment options for chronic disease management.

Implementation Details

- Base Model: Fine-tuned medical LLM (e.g., Med-PaLM)
- Preference Data: 50,000 comparisons from 200 physicians
- Reward Model: Transformer-based classifier on treatment quality
- Policy Optimization: PPO with safety constraints
- Deployment: Staged rollout over 6 months

Results

- Accuracy: 15% improvement in treatment appropriateness
- Safety: 40% reduction in contraindication errors
- Satisfaction: 85% physician approval rating
- Efficiency: 30% reduction in time to formulate treatment plan
- Adherence: 12% increase in patient treatment adherence