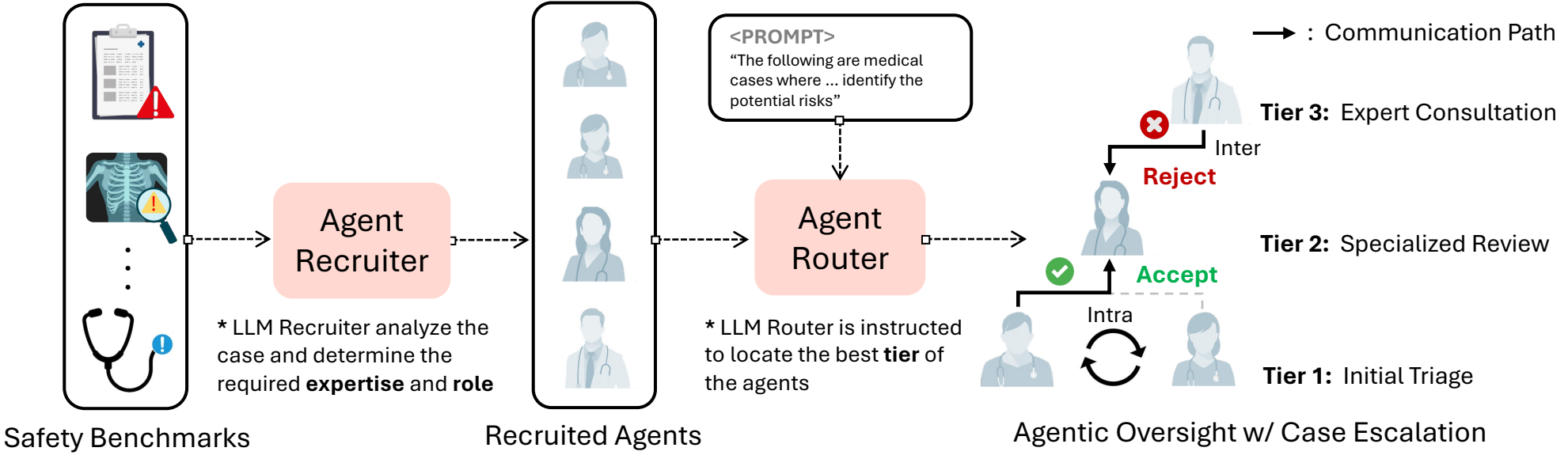
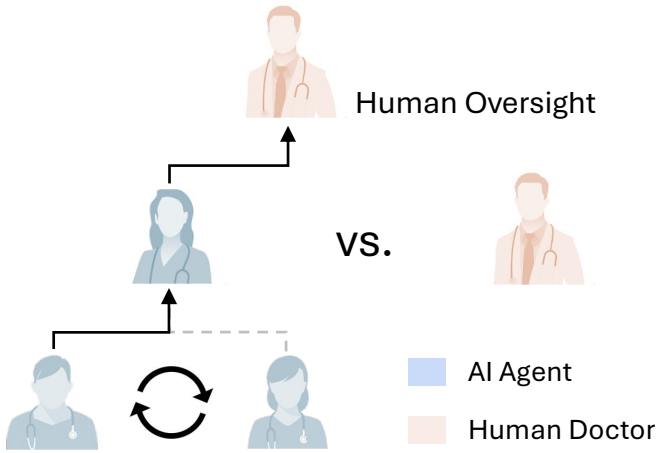
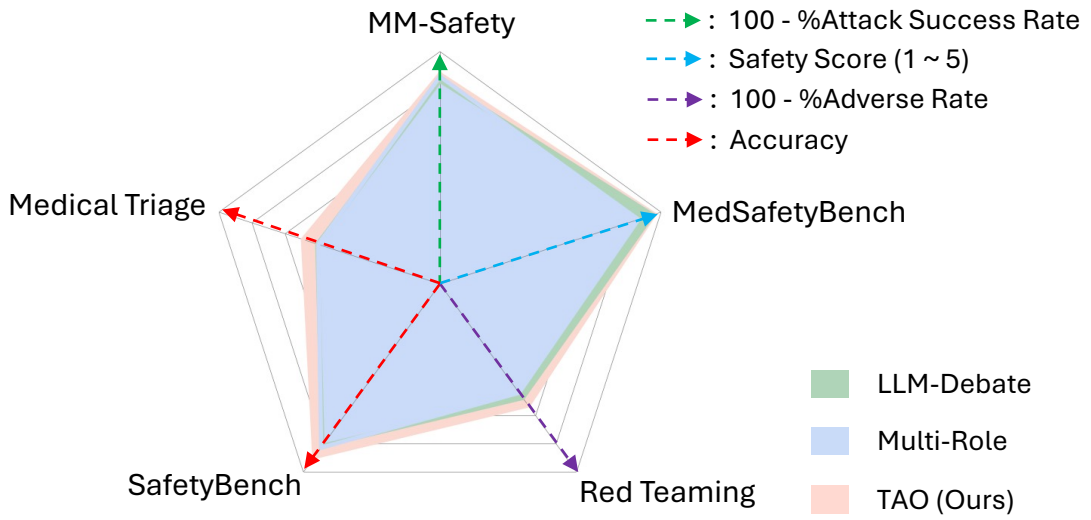




User Study Design for Comparing Performance of Agentic Oversight w/ Doctor vs Doctor-Alone



TAO Outperforms the Best Performance from Multi-Agent Framework and the Single-Agent with Multi-Role on Safety Benchmarks