MIDI 2025 Conference

Plan–Act–Verify: An Agentic AI Question Answering and Reasoning System Evaluated on the CURE-Bench Challenge

Abstract

Reliable therapeutic reasoning with Large Language Models (LLMs) remains challenging due to stale knowledge, limited patient context handling, and hallucinations.

We present a lightweight, plan-act-verify system evaluated on NeurIPS CURE-Bench challenge that combines a two-pass agentic pipeline with high-value biomedical tools. In Pass 1, a planner enumerates required facts and proposes specific tool calls; in Pass 2, a tool agent queries a curated allowlist spanning FDA labels, DailyMed, MedlinePlus, RxNav or RxNorm, OpenTargets, and PubChem. Retrieved evidence is distilled into short "Tool Facts" snippets (capped length, de-duplicated, and source-attributed) that ground a final multiple-choice decision constrained to a single line. We compare local and open-source models against commercial APIs and show that fine-tuning provides the most significant single boost, while tools add complementary facts for regulation-heavy questions. On the CURE-Bench evaluation, small local models score 0.410–0.507, GPT-4.1 reaches 0.520 and 0.543 with tools, fine-tuned GPT-4.1 attains 0.671, and fine-tuned GPT-4.1 with tools achieves 0.69564. Tool gains average +0.019 across model families, and our fine-tuning grid peaks at 88 out of 92 on validation. We incorporated a model that provides patient-specific insights and continuous updates from verified information sources to enhance reliability in clinical applications further. It also delivers consistent, measurable improvements in accuracy, supported by transparent evidence documentation and achieved at a modest cost.