A STRUCTURAL TAXONOMY OF LLM-BASED CLINICAL AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

1 INTRODUCTION

Large language models have rapidly entered clinical medicine, initially through applications such as single turn question answering, document summarization, and static decision support. Early medical uses largely treated these models as passive tools, evaluated through benchmark accuracy on curated datasets or examination style tasks. More recent work increasingly deploys large language models as agents that plan actions, invoke tools, coordinate with other agents, and interact with clinical environments over multiple steps. This transition from isolated model invocation to goal directed agentic systems marks a substantive change in the design space of medical artificial intelligence.

Clinical settings place unique demands on such systems. Medical decision making is constrained by safety requirements, accountability standards, auditability expectations, and the need for continuity across time. An agentic system in medicine is therefore expected not only to produce plausible outputs, but also to justify decisions, manage errors, respect authority boundaries, and maintain state across interactions. Under these conditions, system level design choices such as how agents coordinate, how decisions are verified, and how control is distributed become as important as the underlying language model itself.

Several recent surveys have begun to organize the growing literature on large language model based agents in medicine. Existing reviews distinguish between single agent and multi agent systems, discuss centralized versus decentralized coordination strategies, and introduce workflow or paradigm level categorizations of agent behavior. These efforts provide valuable overviews of agent capabilities and application domains, and they reflect increasing recognition that architectural considerations matter in clinical artificial intelligence.

At the same time, existing surveys do not provide a clinical deployment oriented structural taxonomy that jointly captures three dimensions that are central to reliable medical use. First, they do not systematize collaboration topology together with control distribution beyond coarse architectural categories. Second, they do not treat verification mechanisms with explicit veto or override strength as a first class organizing axis at the level of individual systems. Third, they do not classify agentic systems by temporal scope, ranging from episodic interactions to session level workflows and longitudinal care. As a result, current taxonomies offer limited support for reasoning about safety guarantees, auditability, and failure modes across different clinical deployment settings.

In this survey, we address this gap by organizing large language model based clinical agent systems through a structural lens. We classify systems according to their collaboration topology, verification mechanisms including veto and override behavior, and temporal scope of operation. By treating these dimensions as first class organizing axes rather than implementation details, this survey aims to clarify how architectural choices shape clinical reliability and to support principled design, evaluation, and regulation of agentic artificial intelligence in medicine.