Demo: Medbot, A Practical Tool Built by Clinicians for Clinicians to Leverage AI Agents to Enhance Clinical Practice

Anthony Lianjie Li, Niroj Bhandhari, Antonio Bandeira, Rahul Gorijavolu
Bloomberg School of Public Health
Johns Hopkins University
615 N Wolfe St, Baltimore, MD 21205, United States
lli163@jh.edu, nbhanda8@jh.edu abandei1@jh.edu, rgorija1@jh.edu

Aaron Ruikang Teo, Jasmin Xiaojin Zhang, Alvin Hein Aung Aung, Mark I-Cheng Chen
Office of Clinical Epidemiology and Analytics
Tan Tock Seng Hospital
11 Jln Tan Tock Seng, SG 308433
aaron.rk.teo@nhghealth.com.sg, xiaojin.zhang@nhghealth.com.sg
aung.hein.aung@nhghealth.com.sg, mark.ic.chen@nhghealth.com.sg

Abstract

Clinician burnout is a critical issue in modern healthcare, exacerbated by extensive time spent on Electronic Health Records (EHRs), high cognitive load, and numerous manual administrative tasks. While current AI solutions like ambient scribes and Retrieval-Augmented Generation (RAG) systems have shown promise, they do not fully address the complex, labor-intensive workflows inherent in clinical practice. We present Medbot, an agentic AI platform designed by and for clinicians to automate and enhance clinical workflows. Medbot provides a flexible interface for creating, deploying, and evaluating specialized AI agents that can access local datasets, utilize specific tools, and integrate seamlessly with existing EHRs via the FHIR standard. We demonstrate the practical application of Medbot through a real-world use case: an AI agent co-developed with specialists at Tan Tock Seng Hospital, Singapore, to improve the timely initiation and auditing of the sepsis care bundle. This demonstration showcases how Medbot's agentic architecture can directly address clinical needs, reduce administrative burden, and ultimately improve patient outcomes.

1 Background and Motivation

The modern clinical environment places enormous pressure on healthcare professionals, contributed by the complexity and intensity of interacting with EHRs. This can lead to widespread burnout and reducing time for direct patient care [15]. Furthermore, clinical practice demands a constant recall of vast amounts of information, from international guidelines to institution-specific protocols, while simultaneously managing a multitude of manual tasks like order entry, documentation auditing, and care coordination [12]. This is not uncommon in patients with complex life threatening conditions like sepsis.

Current AI innovations have attempted to alleviate these pressures. Ambient health technologies (e.g., Abridge [1], Nuance DAX [10]) aim to automate documentation by transcribing and summarizing

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

patient-clinician conversations. While useful, these systems often face resistance due to physician and patient discomfort with continuous recording [8]. Other solutions, such as Elsevier's Clinicalkey AI [5], Open Evidence [11], leverages RAG to provide clinical decision support by querying the latest medical literature. However, their utility is often constrained by the following factors:

- **Limited Knowledge Scope:** Most RAG tools are often limited to a finite set of clinical publication databases (e.g., JAMA, NEJM), to which these solutions have access due to data sensitivities and security requirements. They often lack access to crucial local context, such as a country's national guidelines or a hospital's specific antibiotic stewardship protocols, which are dictated by local antibiogram data.
- **Insufficient Research Rigor:** While modern LLMs can break down prompts and use web search tools[3], this process lacks the structured, reproducible methodology (e.g. systematic review) required for evidence-based medicine. Clinical questions are rigorously answered through systematic reviews, which involve a detailed process: framing the question using frameworks like PICO (Population, Intervention, Comparison, Outcome), defining strict inclusion/exclusion criteria, conducting standardized searches across specific databases, and performing multi-stage filtering and quality assessment of evidence. This structured approach is absent in general-purpose, tool-using LLMs.

The recent advancements in agentic AI, underscored by the development of multi-agent protocols like the Model Context Protocol (MCP) [4] and Universal Tool Calling Protocol (UTCP) [13], present a new paradigm. These systems allow for the creation of autonomous agents that can collaborate, use tools, and execute complex, multi-step workflows. This provides a powerful opportunity to automate the laborious clinical tasks that current AI solutions leave unaddressed. To seize this opportunity, we developed Medbot.

2 The Medbot Platform: An Agentic AI Workbench

Medbot is a platform designed to empower clinicians to build, deploy, and manage their own AI agents tailored to their specific workflows. It is built on the principle that the most effective clinical tools are those that can be customized and controlled by the end-users themselves. Its core features are designed for flexibility, integration, and clinical rigor.

2.1 Clinician-Centric Agent Creation

Medbot provides an intuitive, no-code/low-code interface for clinicians to define their own agents. The key components of an agent definition include:

a) **Datasets:** Clinicians can specify the knowledge corpus for an agent, restricting it to trusted sources, including institutional documents, national guidelines, or specific research databases.

b) **Tools:** Agents can be equipped with a variety of tools, from simple calculators and database query functions to complex APIs for interacting with other software.

c) **Knowledge Graphs:** Clinicians can define or upload existing knowledge graphs representing their domain expertise, enabling agents to reason about complex clinical relationships more effectively.

2.2 Customizable Knowledge and EHR Integration

To ensure contextual relevance, Medbot allows clinicians to upload their own appropriately de-identified clinical resources (e.g., hospital protocols in PDF format, departmental guidelines) to an on-premise database and generate vector embeddings on-the-fly. This local knowledge base can be used by agents to ground their responses and actions in the specific context of their practice environment.

Crucially, the platform is tightly integrated with modern EHRs to retrieve encounters and medication requests through the HL7 FHIR (Fast Healthcare Interoperability Resources) standard. This enables agents to securely read, process, and (with appropriate permissions) write clinical data, allowing them

to work seamlessly within existing clinical workflows. For instance, a clinician can edit a patient's note, and an agent can run in the background, analyzing the new information to provide real-time suggestions or alerts to the user.

2.3 LLM Agnosticism and Performance Evaluation

Medbot is designed to be LLM-agnostic. We believe clinicians and administrators should have the freedom to choose the foundational model that best suits their needs, whether for performance, cost, or data privacy reasons. The platform supports integration with various leading models (e.g., GPT-4 series, Gemini family, Claude series), preventing vendor lock-in. An integrated interface allows for the systematic evaluation of agent performance against predefined clinical benchmarks and expert-validated ground truths.

2.4 Deep Research Capabilities

Addressing the limitations of standard RAG, Medbot's "deep research" agents are designed to emulate the systematic review process. When tasked with a complex clinical question, the agent initiates a structured workflow:

1. **Deconstruction:** Breaks down the query into the PICO components.
2. **Strategy Formulation:** Defines inclusion/exclusion criteria and selects relevant databases.
3. **Standardized Search:** Executes reproducible search queries across the selected databases.
4. **Hierarchical Filtering:** Screens articles by title, abstract, and then full text, logging every step for transparency.
5. **Data Extraction & Synthesis:** Extracts data from the final article set and, where applicable, summarizes findings, including assessments of evidence quality.

This structured approach provides clinicians with auditable, evidence-based answers that align with the principles of evidence-based medicine.

2.5 Medbot Platform Architecture

The Medbot platform is implemented using TypeScript and the Next.js framework, providing a robust and scalable foundation for the web and companion desktop applications. Medbot's agentic capabilities are powered by LangGraph [14], which enables the creation and orchestration of complex AI agents tailored to clinical workflows. Each agent's interactions and reasoning steps are traced and monitored using LangSmith, facilitating detailed logging and performance evaluation. Structured data and embeddings are concurrently stored on a Postgres SQL database with an opensource Postgres Vector extension.

The web application frontend and backend are deployed on the Vercel platform, ensuring reliable, low-latency access for clinicians. Meanwhile, the AI agents themselves are hosted and managed on the LangSmith platform, which supports secure execution, tracing, and evaluation of agent workflows [9]. This architecture allows seamless integration between the user-facing interface and the underlying agentic AI infrastructure, supporting both real-time decision support and retrospective auditing in clinical environments.

3 Demonstration Use Case: Sepsis Bundle Initiation Agent

To showcase Medbot's real-world utility, we present a demonstration of an agentic AI sepsis module developed in collaboration with clinical experts from Tan Tock Seng Hospital (TTSH), a major tertiary hospital in Singapore.

3.1 The Clinical Challenge

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection. Timely intervention with the "sepsis care bundle"—a series of evidence-based tasks including

blood cultures, lactate measurement, antibiotic administration, and fluid resuscitation—is proven to significantly reduce mortality [2]. However, initiating this bundle is often delayed. The high cognitive load of managing critically ill patients, combined with the documentation burden in the EHR, means that the subtle signs of sepsis can be missed, leading to critical delays in care [6].

3.2 The Medbot Sepsis Agent

Using the Medbot platform, we co-developed and validated an agentic AI sepsis module with advanced internal medicine specialists and practicing nurse clinicians from TTSH. The agent is designed to perform two primary functions.

Function 1, Real-time Decision Support: If the agent is integrated with the EHR, the agent will continuously and securely monitor incoming clinical data for high-risk patients. This includes parsing new entries in clinical notes (using NLP), vital signs, and laboratory results. When criteria for sepsis (e.g. Systematic Inflammatory Response Syndrome (SIRS) criteria [7] combined with suspected infection) are met, the agent provides an immediate, non-intrusive alert with the reasons to the primary clinical team directly within the EHR interface, suggesting the initiation of the sepsis bundle. The agent's knowledge base is grounded in both international Surviving Sepsis Campaign guidelines and TTSH's sepsis specific management protocols.

Function 2, Retrospective Quality Auditing: The agent retrospectively analyzes clinical notes and timestamps from the EHR to audit the accuracy and timeliness of sepsis bundle implementation. It can automatically generate reports for quality improvement committees, identifying cases where bundle elements were delayed or missed, thus replacing a laborious manual audit process.

3.3 System in Action: A Demo Workflow

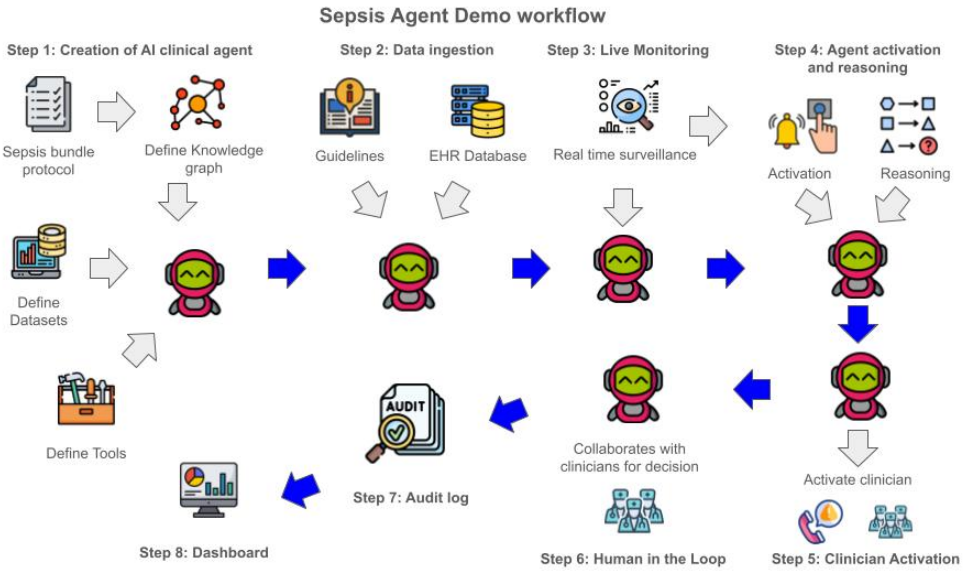

Figure 1: Demo workflow for Medbot Sepsis Agent

The demonstration is attached as supplementary material and is presented at the following video link (`https://www.youtube.com/watch?v=tpn0S2NAI34`). It will walk through part of the following workflow (see Figure 1):

1. **AI Agent creation by clinicians:** A brief overview of how the TTSH clinical team used Medbot's interface to define the sepsis agent, including selecting relevant datasets (e.g., institutional sepsis protocols), tools (e.g., EHR FHIR API access), and knowledge graphs (e.g., sepsis bundle initiation decision algorithm).

2. **Data Ingestion:** The Medbot sepsis agent connects to a de-identified instance of an EHR via its FHIR API.
3. **Live Monitoring:** The audience will see a simulated feed of clinical data (nurse's notes, lab results showing rising lactate, vital signs showing tachycardia and hypotension) for a mock patient.
4. **Agent Activation and Reasoning:** The sepsis agent processes this data in real-time using the knowledge graph defined by the clinician in the first step. Its log will be displayed, showing how it identifies keywords like "fever", "lethargy" and "cough", and correlates them with structured data (vital signs, white blood cell count) to meet the sepsis criteria defined in its knowledge base.
5. **Clinician Notification:** A prompt appears in the simulated EHR UI, stating: "Sepsis criteria met for Patient [ID]. Consider initiating a sepsis bundle. [Link to institutional protocol]"
6. **Human In the Loop:** The clinician (played by a team member) acknowledges the alert, reviews the AI agent's suggestions and justifications, and initiates the sepsis bundle orders directly from the EHR interface.
7. **Audit Log:** The agent logs all the details related to the event including the timestamps, clinical details, its reasoning log, the recommendation and the human interventions.
8. **Retrospective Quality Audit**: The demonstration will then switch to a dashboard view, showing how this data contributes to a retrospective audit of care quality, e.g. delays in bundle initiation, providing actionable insights for quality improvement. Finally, administrators can evaluate the performance of the AI agents and identify areas for further training or refinement.

This module is expected to enhance patient safety and standardize care by ensuring consistent, evidence-based sepsis management. By automating the review of clinical documentation and providing real-time prompts, the AI agent will reduce diagnostic delays, improve patient outcomes, and alleviate the administrative tasks contributing to physician burnout.

4 Limitations and mitigation strategies

Medbot, as an agentic AI platform for clinical workflows, faces several limitations inherent to both technical and practical deployment in healthcare environments. First, the accuracy and reliability of AI-driven decision support are constrained by the quality and completeness of input data. EHRs often contain unstructured, incomplete, or erroneous information, which can lead to missed alerts or incorrect recommendations. Additionally, the performance of Medbot agents is dependent on the underlying language models and tool integrations, which may not generalize well to all clinical specialties or adapt to rapidly evolving medical guidelines.

Second, integration with existing hospital systems and workflows presents significant challenges. Healthcare institutions vary widely in their IT infrastructure, data standards, and security protocols. Seamless interoperability with diverse EHR systems via FHIR, while technically feasible, may be hindered by legacy systems, inconsistent data formats, and institutional resistance to change. Furthermore, ensuring robust data privacy and compliance with regulations such as HIPAA or GDPR is critical, especially when agents access sensitive patient information or automate clinical actions.

To mitigate these limitations, we have considered the following strategies. Rigorous validation and continuous monitoring of agent outputs against expert-annotated ground truths can help maintain clinical accuracy. Modular design and configurable interfaces allow adaptation to local protocols and facilitate incremental integration with hospital IT systems. Privacy safeguards, such as on-premise deployment options, granular access controls, and audit logging, can address regulatory concerns. Finally, ongoing collaboration with clinicians for agent development and feedback ensures that Medbot remains clinically relevant, trustworthy, and responsive to real-world needs.

5 Conclusion and Future Work

Medbot is an agentic AI platform designed to bridge the gap between the potential of large language models and the practical, everyday needs of clinicians. It moves beyond passive documentation and

generic information retrieval to enable the creation of active, specialized AI agents that can automate complex workflows and provide context-aware, actionable support.

Our collaboration with Tan Tock Seng Hospital on the sepsis module demonstrates the viability and clinical value of this approach. By empowering clinicians to build and validate their own AI tools, we ensure that the solutions are relevant, trusted, and seamlessly integrated into practice. Moving forward, we plan to continue validating the Medbot platform and proving its value in other specialty domains, including surgery (e.g., for post-operative complication monitoring) and preventive medicine (e.g., for chronic disease screening and management).

Interested users can access Medbot at `https://medbot.me`.

References

[1] Abridge. Intelligence at the point of conversation. `https://www.abridge.com/`. Index page. Retrieved on 25th August 2025.

[2] Jonathan D Baghdadi, Robert H Brook, Daniel Z Uslan, Jack Needleman, Douglas S Bell, William E Cunningham, and Mitchell D Wong. Association of a care bundle for early sepsis management with mortality among patients with hospital-onset or community-onset sepsis. *JAMA internal medicine*, 180(5):707–716, 2020.

[3] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

[4] Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). *arXiv preprint arXiv:2505.02279*, 2025.

[5] Elsevier. Clinikey ai. `https://www.elsevier.com/products/clinicalkey/clinicalkey-ai`. Products, Clinicalkey. Retrieved on 25th August 2025.

[6] D Gilhooly, SA Green, C McCann, N Black, and SR Moonesinghe. Barriers and facilitators to the successful development, implementation and evaluation of care bundles in acute care in hospital: a scoping review. *Implementation Science*, 14(1):47, 2019.

[7] Kirsi-Maija Kaukonen, Michael Bailey, David Pilcher, D Jamie Cooper, and Rinaldo Bellomo. Systemic inflammatory response syndrome criteria in defining severe sepsis. *New England Journal of Medicine*, 372(17):1629–1638, 2015.

[8] Tiffany I Leung, Andrew J Coristine, and Arriel Benis. Ai scribes in health care: Balancing transformative potential with responsible integration. *JMIR Medical Informatics*, 13(1):e80898, 2025.

[9] Sarah McAvinue and Kapal Dev. Comparative evaluation of large language models using key metrics and emerging tools. *Expert Systems*, 42(2):e13719, 2025.

[10] Microsoft. Dragon medical one. `https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-medical-one`. Health Solutions, Clinical Workflow. Retrieved on 25th August 2025.

[11] Openevidence. Openevidence. `https://www.openevidence.com/`. Index page. Retrieved on 25th August 2025.

[12] Ann S O'Malley, Joy M Grossman, Genna R Cohen, Nicole M Kemper, and Hoangmai H Pham. Are electronic medical records helpful for care coordination? experiences of physician practices. *Journal of general internal medicine*, 25(3):177–185, 2010.

[13] UTCP. About us. `https://www.utcp.io/about/about-us`. UTCP.io, About. Retrieved on 25th August 2025.

[14] Jialin Wang and Zhihua Duan. Agent ai with langgraph: A modular framework for enhancing machine translation using large language models. *arXiv preprint arXiv:2412.03801*, 2024.

[15] Qi Yan, Zheng Jiang, Zachary Harbin, Preston H Tolbert, and Mark G Davies. Exploring the relationship between electronic health records and provider burnout: a systematic review. *Journal of the American Medical Informatics Association*, 28(5):1009–1021, 2021.