Improving Outpatient Workflows through Structured History Collection with LLMs

Sophie C. Steinbuch[1,2,3], Luka de Vos-Hillebrand[2], Benjamin Wolf[1], Connor O'Neill-Dee[4], Iris Wu[5,6], Hannah Steinbuch[7], Zsófi Kulcsár[2,3], Avanish Ranjan[8], Ashish Verma[4], Jennifer Landsberg[2], Dimo Dietrich[3], Charles C. Hardin[6], Rakesh K. Jain[1], Sonu Subudhi[1,9,10,#]

[1]Steele Laboratories of Tumor Biology, Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, United States;
[2]Clinic for of Dermato-Oncology and Phlebology, University Hospital Bonn, University of Bonn, Bonn, Germany;
[3]Clinic and Polyclinic for Ear, Nose and Throat Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany;
[4]Department of Medicine (Nephrology), Boston University Chobanian & Avedisian School of Medicine, Boston, United States;
[5]Harvard-MIT Health Sciences and Technology Program, Harvard Medical School, Boston, United States;
[6]Medicine-Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, United States;
[7]Department of Internal Medicine, University Hospital Zürich, Zürich, Switzerland;
[8]Adobe, Boston, United States;
[9]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, United States;
[10]Broad Institute of MIT and Harvard, Cambridge, United States

Abstract

Comprehensive clinical history taking is central to diagnostic reasoning and treatment planning but remains one of the most time-consuming components of outpatient care. We evaluated whether a modular, model-independent agentic framework could provide reliable, structured pre-visit history collection across two successive generations of large language models. Across three evaluation phases involving 400 synthetic cases, 252 case reports, and 20 physician-constructed scenarios, we benchmarked six LLM backends using independent physician raters and a three-model LLM judge ensemble. Results demonstrated consistently high history recall and robust red-flag identification across all models and both LLM generations. Although individual models showed metric-specific advantages in history recall, diagnosis, and investigation recommendations, conversational structure rather than model selection was the primary determinant of history completeness. These findings support the feasibility of model-agnostic agentic frameworks for scalable, generation-stable clinical history collection, justifying prospective clinical evaluation in real-world pre-visit settings.

Background

- Clinical history-taking is foundational to diagnosis and treatment planning, but it is one of the most time-consuming parts of outpatient care.
- General-purpose LLMs used as unstructured assistants underperform in one study, correct identification of relevant conditions dropped from more than 94 percent (models working independently) to fewer than 35 percent when the same models were used as unstructured aids by physicians.
- Letting models freely ask follow-up questions without a structured framework has been shown to paradoxically reduce diagnostic accuracy.
- Purpose-built, structured conversational systems have matched or exceeded clinician performance in simulated settings, suggesting the limiting factor is conversational design rather than raw model capability.
- **Open question:** can a structured, model-agnostic framework deliver reliable, generation-stable history collection no matter which LLM powers it?

Objectives

We developed and evaluated a structured agentic framework that guides LLMs through modular, sequential clinical history collection in a pre-visit setting, generating EHR-compatible summaries, differential diagnoses, and investigation recommendations, and tested it across three evaluation phases and two LLM generations.

Methods/Algorithm

- Modular agentic system sequentially collects chief complaint, history of present illness, past medical history, medications, allergies, social and family history, and review of systems.
- Each response is evaluated for clinical relevance before the system advances; interim summaries preserve context; conversation ends when completeness criteria are met.
- At conclusion, the system generates: an EHR-compatible structured summary, a ranked differential diagnosis, a red-flag conditions list, and recommended initial investigations.
- Scoring used a standardized rubric across four history domains, diagnostic accuracy, and red-flag identification, with items scored 0 (not elicited), 0.5 (partial), or 1 (fully elicited).
- An ensemble of three locally hosted LLM judges (Qwen 2.5 14B, Mistral 7B, Llama 3.1 8B) was validated against independent physician scoring using Bland-Altman analysis, showing strong agreement (mean bias 0.04 to 0.06).

Results

Phase I: *Automated evaluation, synthetic cases*

- 200 synthetic cases; GPT-4o, Grok-3, Gemini-2.5-Flash-Lite; Llama 3.1 8B automated patient simulation.
- History recall high for all models: Grok-3 0.96, Gemini 0.95, GPT-4o 0.94.
- Red-flag elicitation uniformly strong (0.84-0.86) across models.
- Diagnostic any-match accuracy 0.82-0.89; investigations near-ceiling (≥0.97).

Phase II: *Physician-evaluated cases*

- 52 NEJM case reports + 20 constructed scenarios; physician actors; independent blinded raters (4 and 3).
- History recall: 0.87-0.89 (NEJM cases), 0.77-0.79 (constructed scenarios).
- Diagnosis tracked case design: 0.92-0.96 (constructed) vs 0.38-0.45 (complex NEJM cases).
- Red flags strong across both case types (0.66-0.90).
- LLM-judge ensemble closely matched physician scoring (Bland-Altman bias 0.04-0.06).

Phase III: *Newer-generation models*

- GPT-5-Mini, Grok-4.1-Fast, Gemini-3.1-Flash on 200 open-access + 200 synthetic cases.
- Near-ceiling history recall (0.96-0.99), an improvement over Phase I.
- Diagnosis tracked complexity: 0.72-0.78 (synthetic) vs 0.33-0.37 (open-access).
- No single model dominated every metric across history, diagnosis, and investigations.

Conclusion

- Conversational structure, not model selection, is the principal determinant of history completeness in agentic clinical AI.
- A structured, model-agnostic framework delivers generation-stable, clinician-validated history collection across 6 LLM backends and 2 generations.
- Diagnostic accuracy from history alone remains limited for complex, atypical presentations, supporting a triage and referral design goal rather than diagnosis.
- Cross-metric model dissociation motivates future task-adaptive routing strategies.

References

1. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613-2622. DOI: 10.1038/s41591-024-03097-1.
2. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. Nat Commun 2024;15(1):8236. DOI: 10.1038/s41467-024-52415-1.
3. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642(8067):451-457. DOI: 10.1038/s41586-025-08869-4.
4. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025. DOI: 10.1038/s41586-025-08866-7.

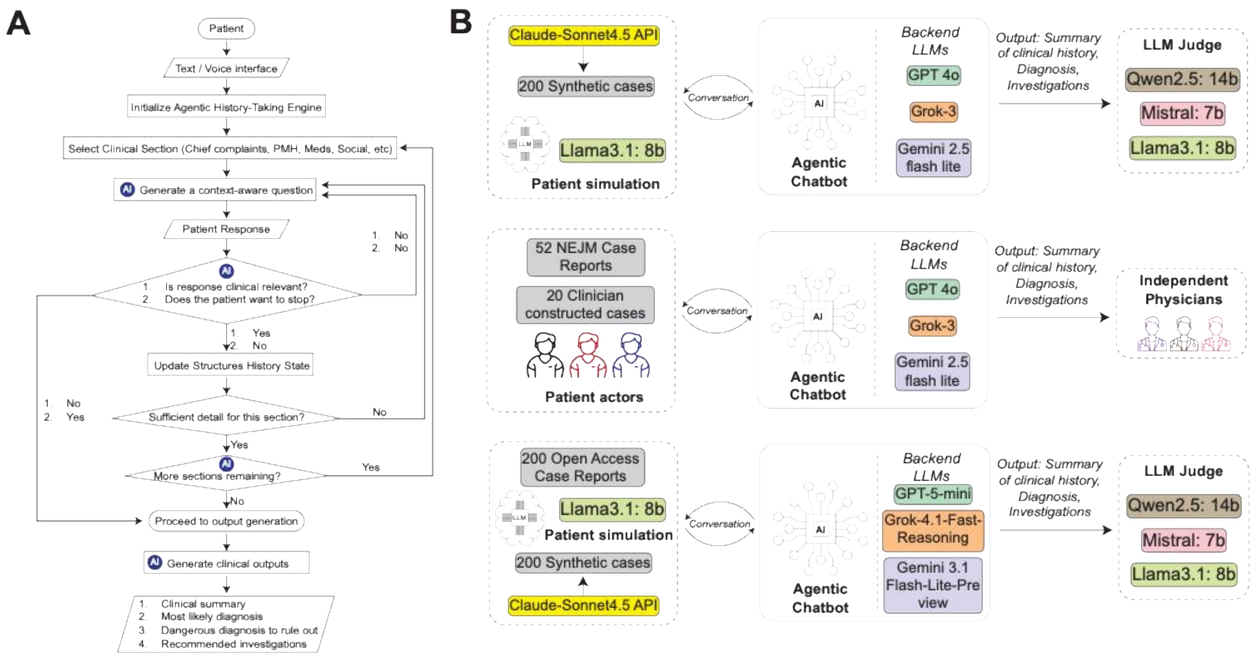


Figure 1. Three-phase evaluation framework and agentic system architecture. (A) Each phase paired a different generation of LLM backend with either automated or physician-based patient simulation. (B) The agentic engine steps sequentially through each history section, checks response relevance, updates a structured internal state, and produces the final clinical outputs.

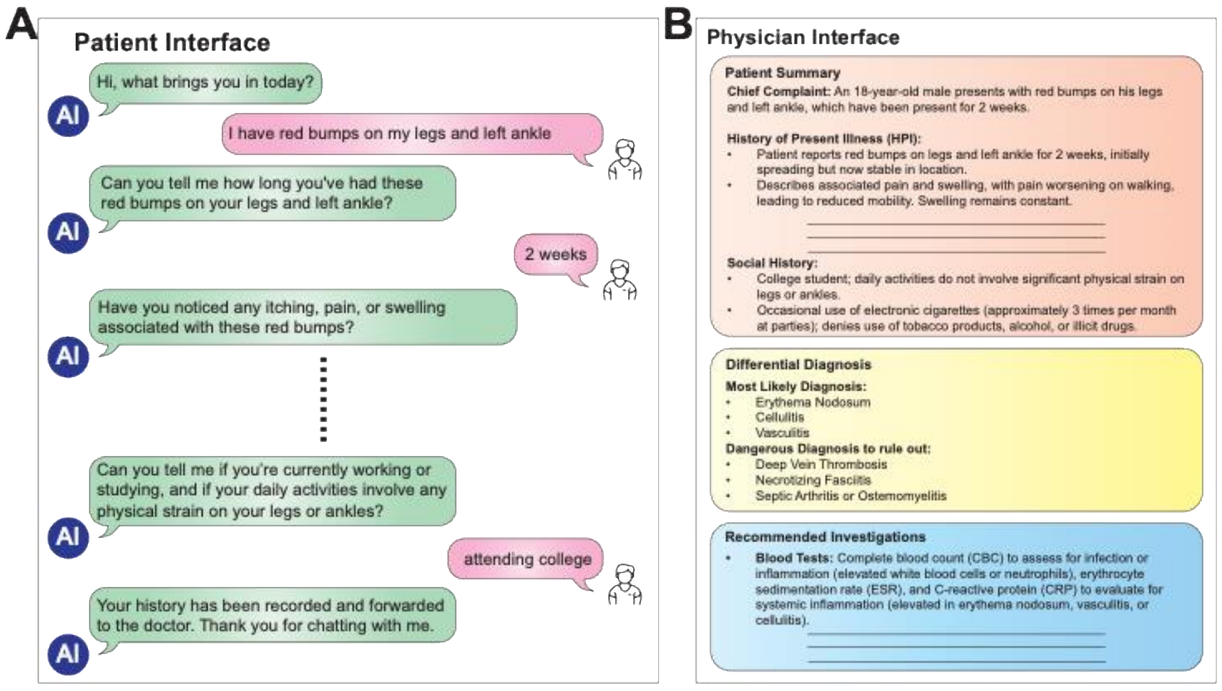


Figure 2. (A) Chatbot interaction. Example of a patient–chatbot dialogue, where the system dynamically asks follow-up questions to capture clinically relevant details. **(B) Structured history output.** Automatically generated, EHR-compatible history including chief complaint, history of present illness, past medical history, medications, allergies, and family history. **(C) Clinical reasoning support.** Chatbot-generated differential diagnosis, dangerous conditions to rule out, and recommended investigations based on the collected history.

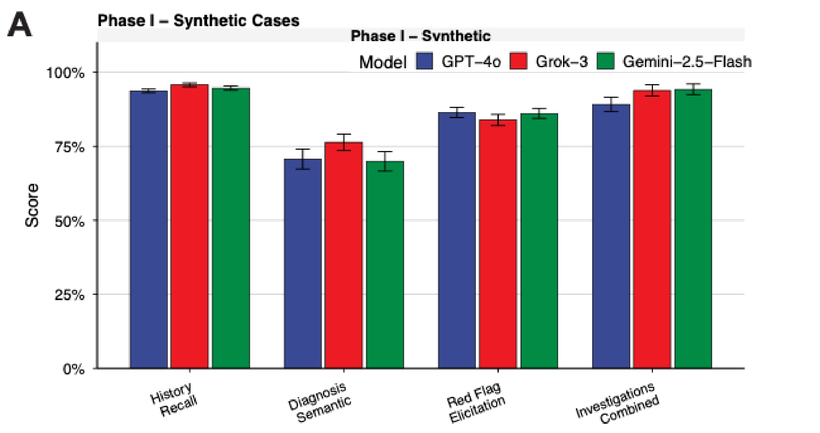


Figure 3. Phase I performance on synthetic cases. Grouped bar chart showing mean scores with 95% confidence intervals for history recall, diagnosis semantic similarity, red flag elicitation, and investigations combined across GPT-4o, Grok-3, and Gemini-2.5-Flash on 200 synthetic cases.

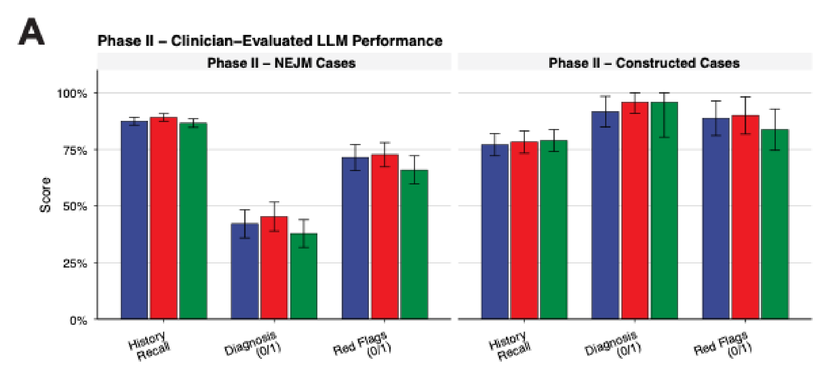


Figure 4. Phase II clinician-evaluated performance on published case reports and constructed scenarios. (A) Grouped bar chart showing mean scores with 95% confidence intervals for history recall, diagnosis, and red flags across GPT-4o, Grok-3, and Gemini-2.5-Flash, faceted by NEJM cases and constructed cases. **(B)** Paired scatter plots showing case-level history recall for each pairwise model comparison within NEJM cases; the dashed diagonal represents perfect agreement and paired Wilcoxon p-values are shown for each comparison.

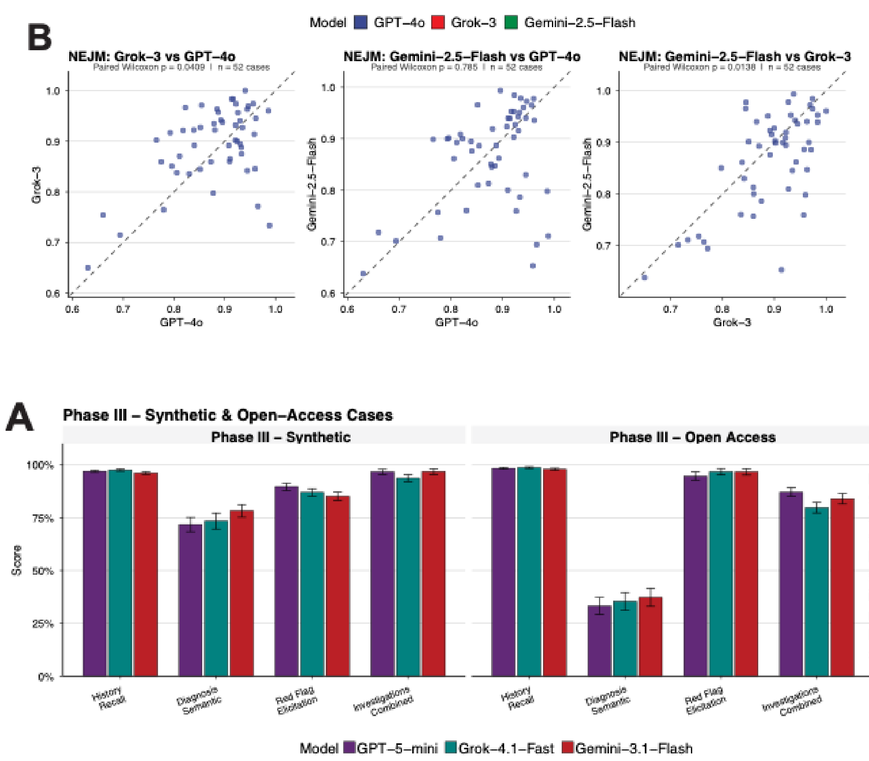


Figure 5. Phase III performance on synthetic and open-access published cases. Grouped bar chart showing mean scores with 95% confidence intervals for history recall, diagnosis semantic similarity, red flag elicitation, and investigations combined across GPT-5-mini, Grok-4.1-Fast, and Gemini-3.1-Flash, faceted by Phase III synthetic cases and open-access cases.