Accountable Clinical AI Requires More Than Accuracy

Author: Thomas F. Heston, MD, MSc

Affiliations: Department of Family Medicine, University of Washington, Seattle, WA; Department of Medical Education and Clinical Sciences, Washington State University, Spokane, WA

ORCID: 0000-0002-5655-2512

Citation: Heston TF. Accountable clinical AI requires more than accuracy. Internet Medical Journal. 2026;1(1):e19519377. doi:10.5281/zenodo.19519377.

Abstract

Large language models are approaching specialist-level performance in selected clinical tasks, but accuracy alone does not establish readiness for clinical deployment. This commentary argues that accountability, rather than raw performance, is now the central barrier to adoption. Recent evidence shows that clinically deployed large language models can perform radiology workflow tasks with high accuracy, yet important governance questions remain unresolved, including the provenance of inputs, the auditability of outputs, and the verification of downstream decision pathways. The present commentary proposes that accountability infrastructure should become a routine focus of clinical AI

evaluation alongside performance metrics. Distributed ledger and related audit technologies may offer one practical framework for tamper-resistant logging, verification, and oversight of model-mediated clinical decisions. Clinical studies should therefore report governance architecture in addition to accuracy, and medical education should treat prompt engineering as an operational clinical competency. The next phase of clinical AI is not merely accurate systems, but accountable ones.

Keywords: Clinical AI Governance, Large Language Models, Blockchain Healthcare, Radiology AI, Algorithmic Accountability.

The deployment of large language models in clinical workflows has reached a threshold where accuracy alone no longer constitutes the primary barrier to adoption — accountability does. A recent retrospective study demonstrated that an institutionally hosted Claude 3.5 model, augmenting magnetic resonance imaging examination requests, achieved 93.1% protocol accuracy across 608 outpatient examinations spanning the body, musculoskeletal, and neuroradiology subspecialties, performing comparably to board-certified general radiologists [1]. While this result is an important advance in artificial intelligence benchmarking, the authors note the study was not powered to statistically establish noninferiority or equivalence. Nevertheless, it demonstrates that a commercial large language model can operate within a real clinical radiology workflow with no clinically significant hallucinations identified on manual review. The question confronting the field is no longer whether large language models can perform clinical tasks competently, but

whether the infrastructure surrounding their deployment can ensure that such performance is verifiable, auditable, and trustworthy over time.

The empirical foundation for this shift has been accumulating rapidly. Head-to-head comparisons between large language models and physicians demonstrate impressive capabilities; for instance, on a standardized knowledge assessment concerning acute kidney injury, models consistently performed at or above the competence of various medical professionals, including junior- to mid-level physicians [2]. A scoping review of 35 studies examining large language models in emergency medicine found that the dominant applications — triage, diagnostic decision support, and documentation — remain confined largely to non-agentic large language models, with only three studies exploring multi-agent systems and very few explicitly evaluating workflow-level safety or longitudinal trustworthiness [3]. Meanwhile, research on bias in clinical decision support has revealed that prompt engineering strategies, such as chain-of-thought reasoning, reduce biased outcomes more effectively than medical fine-tuning of the underlying model architecture [4]. This finding carries a practical implication: the quality of a clinical large language model deployment depends as much on how the model is prompted and governed as on which model is selected. Prompt engineering — the structured design of instructions that guide model behavior — has been identified as a foundational competency for clinicians working with these systems, yet it remains largely absent from formal medical training [5].

The most urgent issue now is not so much one of performance but of provenance. The study reports that improvements in protocol accuracy arose primarily from salient

electronic medical record data that clinicians had omitted from examination requests — infection history, tumor history, and prior surgical details [1]. The model's value, in other words, is derived from synthesizing the clinical context that was already documented but not surfaced by existing workflows. This is precisely the scenario where accountability infrastructure becomes critical. When a large language model integrates data from multiple sources to generate a clinical recommendation, the reasoning chain connecting input data to output decision must be preserved, inspectable, and tamper-resistant. The current study employed a privately hosted model with temperature set to zero, ensuring deterministic outputs — a commendable methodological choice — but determinism at the model level does not address the broader question of whether the data fed into the model was complete, whether the output was reviewed before action, and whether any downstream modifications occurred. While a temperature of zero improves reproducibility, it does not eliminate algorithmic bias. These are governance questions, not performance questions, and they require infrastructure that the clinical large-language-model literature is currently actively addressing [6]. Blockchain-based distributed ledger systems designed for health data — already under active clinical evaluation — may provide one practical model for this kind of infrastructure to audit and verify model outputs [7,8].

Blockchain technology offers a plausible architectural approach to the accountability gap in clinical artificial intelligence. By using a distributed, immutable ledger, health systems can cryptographically record clinical interactions with large language models, protect training data for federated learning from poisoning attacks, and provide patients with highly usable access-control features. The convergence of these

technologies—clinical large language models providing decision support, blockchain providing verifiable audit trails, and federated learning enabling cross-institutional model improvement—has been proposed as a next step in digital health transformation and accountable telemedicine [7–10]. Moving beyond theoretical models, recent primary research has successfully implemented and evaluated this convergence in practice, including the deployment of decentralized digital hospital frameworks that use blockchain to audit and regulate large-language-model diagnostic outputs within clinical workflows [6].

Several concrete considerations support the feasibility of this integration. First, the use of a privately hosted large language model with deterministic settings demonstrates that institutional deployment can be controlled at the infrastructure level—the same infrastructure layer where blockchain logging would operate. Second, findings that model accuracy is comparable to radiologists across three subspecialties suggest that the performance question is sufficiently advanced for radiology protocol selection to serve as a reasonable test case for accountability infrastructure [1]. Third, the experience-dependent pattern observed in clinicians' trust of large language model outputs—where less experienced practitioners rate model recommendations more favorably than specialists—underscores the need for an external verification mechanism that does not depend on the reviewing clinician's level of expertise [11]. A blockchain-verified audit trail would provide such a mechanism, enabling retrospective quality assurance regardless of who reviewed the model's output at the point of care. The safety profile of large language models in clinical contexts remains an active area of investigation, with documented

inconsistencies in risk-stratification tasks highlighting the importance of systematic oversight mechanisms [12].

The path forward requires three developments. Clinical large language model studies should report not only accuracy metrics but also governance architecture: how inputs are sourced, how outputs are logged, who reviews them, and whether the audit trail is tamper-resistant. Blockchain researchers working in health information exchange should extend their frameworks to encompass large-language-model decision logging, not only patient data sharing. And medical education programs should incorporate prompt engineering as a clinical competency, recognizing that the quality of large language model outputs in patient care depends on the structured reasoning embedded in the prompts that clinicians use. These suggested developments are shown in Table 1.

Table 1. Accountability Infrastructure Checklist for Clinical AI Studies.

Domain	Reporting Requirement	Rationale
Input Provenance	Specify the source of clinical data (e.g., EMR, manual entry) and identify data types often omitted by clinicians (e.g., infection/tumor history).	Ensures the model's value is derived from synthesizing relevant clinical context.
Model Configuration	Report the model version and specific settings, such as setting the temperature to zero to ensure determinism.	Improves reproducibility and controls infrastructure-level behavior.
Prompt Governance	Detail the use of structured instructions, such as chain-of-thought reasoning, and describe the prompt design process.	Prompt engineering is a clinical competency that reduces biased outcomes more effectively than fine-tuning.
Output Review	Document whether outputs were reviewed by a clinician before action, and the expertise level of the reviewer.	Addresses trust differences between specialists and less-experienced practitioners.
Auditability	Describe the logging mechanism (e.g., blockchain-based distributed ledger) and whether the trail is tamper-resistant.	Preserves the reasoning chain and allows for retrospective quality assurance.
System Safety	Report performance on risk-stratification tasks and any multi-agent oversight mechanisms used.	Ensures longitudinal trustworthiness and identifies potential inconsistencies in high-stakes tasks.

The Hallinan study has demonstrated that a large language model can perform a clinical task at a specialist level in a real-world setting. The next challenge is developing systems that ensure such performance is not only accurate but also accountable — and that the evidence of its accountability is as permanent and verifiable as the clinical decisions it supports.

Declarations

The author reports no conflicts of interest. This study did not receive any external funding.

Large language models were used for language editing and formatting assistance; the author reviewed, verified, and is fully responsible for all content.

References

1. Hallinan JTPD, Leow NW, Low YX, Lee A, Ong W, Chan MDZ, et al. Initial Insights Into an Institutional Secure Large Language Model for Magnetic Resonance Imaging Examination Requests: Retrospective Study. J Med Internet Res. 2026;28: e82579. doi:10.2196/82579

2. Russ P, Bedenbender S, Einloft J, Meyer HL, Wenzel LT, Ganser A, et al. Potential of large language models for rapid clinical information support: evidence from acute kidney injury knowledge testing. Sci Rep. 2026;16: 11224. doi:10.1038/s41598-026-46846-7

3. Kim H, Jo S, Lim MH, Choi DH. From non-agentic large language models to multi-agent systems in emergency medicine: a scoping review. Clin Exp Emerg Med. 2026 [cited 10 Apr 2026]. doi:10.15441/ceem.26.136

4. Poulain R, Adiba FI, Fayyaz H, Beheshti R. Bias Patterns in the Application of LLMs for Clinical Decision Support. Del J Public Health. 2026;12: 54–67. doi:10.32481/djph.2026.03.10

5. Patil R, Heston TF, Bhuse V. Prompt Engineering in Healthcare. Electronics. 2024 [cited 26 July 2024]. Available: https://www.mdpi.com/2079-9292/13/15/2961

6. Sun L, Liu D, Wang M, Han Y, Zhang Y, Zhou B, et al. Taming Unleashed Large Language Models With Blockchain for Massive Personalized Reliable Healthcare. IEEE J Biomed Health Inform. 2025;29: 4498–4511. doi:10.1109/JBHI.2025.3528526

7. Guse R, Hu S, Thiebes S, Erler C, Caridia C, Stork W, et al. Patient Perceptions of Blockchain-Based Health Information Exchange: User-Centered Design Study. J Med Internet Res. 2026;28: e78849–e78849. doi:10.2196/78849

8. Liu J, Hu X. Blockchain meets AI in healthcare: a review of convergent technologies for digital health transformation. Front Blockchain. 2026;9: 1766092. doi:10.3389/fbloc.2026.1766092

9. Alruwaili E, Moulahi T. A robust and verifiable federated learning framework for preventing data poisonous threats in e-health. Front Public Health. 2026;14. doi:10.3389/fpubh.2026.1762346

10. Heston TF. Perspective Chapter: Integrating Large Language Models and Blockchain in Telemedicine. In: Heston TF, editor. A Comprehensive Overview of Telemedicine. London: IntechOpen; 2024. Available: https://www.intechopen.com/online-first/1176440

11. Fabi A, Egli CE, Wendelspiess SR, Griewing S, Haas Y, De Pellegrin L, et al. Exploring the Role of AI in Managing Treatment Recommendations for Lymphedema: International, Multidisciplinary, Multiprofessional Survey Study of Trust, Reliability, and Impact on Decision-Making. JMIR Med Inform. 2026;14: e80553–e80553. doi:10.2196/80553

12. Heston TF, Lewis LM. ChatGPT Provides Inconsistent Risk-Stratification of Patients With Atraumatic Chest Pain. medRxiv. 2023 [cited 31 Jan 2024]. doi:10.1101/2023.11.29.23299214