Internet Medical Journal
ISSN 1093-7935 · medjournal.com · Est. 1996

The AI Consultation Report as a Format for Documenting Clinical AI Failure Modes

Thomas F. Heston

Affiliation	Department of Family Medicine, University of Washington, Seattle, USA
Affiliation	Department of Medical Education and Clinical Sciences, Elson S. Floyd College of Medicine, Washington State University, Spokane, USA
ORCID	0000-0002-5655-2512
Published	13 September 2026
DOI	10.5281/zenodo.22740633
Article type	Perspective
Citation	Heston TF. The AI Consultation Report as a Format for Documenting Clinical AI Failure Modes. Internet Medical Journal. 2026;1:e22740633

Abstract

Controlled evaluations of large language models (LLMs) in medicine measure how often a system answers a clinical question correctly, but they cannot describe what happens when a clinician consults one about a specific patient. No current reporting instrument lets an individual clinician document a consultation, and isolated published accounts cannot be compared or pooled because no format defines them. The artificial intelligence (AI) consultation report is proposed as a defined article type: a first-person account specifying the clinical impasse, the tool, version, and approximate date, the question as posed, the output as returned, the clinician's decision, and the outcome. Accumulated, such reports offer a mechanism for identifying clinical AI failure modes before they can be measured, on the model of spontaneous reporting in pharmacovigilance.

Keywords

ai consultation report, clinical ai failure modes, large language models, generative artificial intelligence, clinical decision making, signal detection, medical publishing

Controlled evaluation cannot capture what happens when a clinician consults a large language model (LLM) about a specific patient in the middle of a workday. Evaluation studies fix the prompt, model version, case material, and reader, while bedside consultations vary on all four. The consequence is a literature that reports how often these systems answer clinical questions correctly and says little about how clinicians actually use them, where the output misleads, and what happens next. That gap warrants a defined publication format, here called the artificial intelligence (AI) consultation report: a first-person, structured account by a clinician of a single consultation with a generative model during real patient care, submitted whether the model helped or failed. The Internet Medical Journal established this type and solicits it directly; the format is described here so any journal publishing clinical AI work can adopt or adapt it.

The instruments now available for reporting clinical AI work target studies rather than practice. TRIPOD-LLM provides 19 main items and 50 subitems covering the design, conduct, and reporting of LLM research in health care [1]. Editorial criteria for evaluating LLM manuscripts address clinical research contexts specifically, calling for detailed specification of model versions and tunable parameters, repeated identical requests, alternative prompts, and characterization of incorrect responses [2]. Professional-society guidance separates patient-facing applications, clinician-facing applications, and background institutional systems, and emphasizes human oversight and systematic validation [3]. Each of these is well constructed for its purpose, and each raises the standard of the work it governs. None of them, though, gives an individual clinician a way to report a single consultation.

That route matters because model behavior varies enough between identical calls that an aggregate figure cannot describe a single encounter. An evaluation of ChatGPT-4 on simulated atraumatic chest-pain cases found that the model returned a different risk assessment 45% to 48% of the time when presented with a fixed TIMI or HEART score, and that a majority of five runs agreed on a diagnostic category only 56% of the time on a 44-variable model [4]. The pattern is not specific to one research group: repeated queries of 30 atrial fibrillation guideline questions across three current models produced only moderate to substantial agreement between runs, and the best-performing combination matched the guideline's recommendation class 60.3% of the time [5]. Models also differ measurably in unprompted response patterns [6]. A reported accuracy rate therefore characterizes a distribution, not the answer a particular clinician received on a particular afternoon. Individual accounts of clinical LLM use have been published [7] but without a defined format, they cannot be compared or pooled.

Given that variability, the reporting elements of an AI consultation report follow from what a reader must know to interpret the encounter. Six are proposed. First, the clinical situation and the point at which the clinician's own reasoning reached its limit, because the question put to the model is meaningful only against the impasse that produced it. Second, the tool, version, and approximate date, because models differ and change over time (Heston & Gillette, 2025). Third, the exact form of the question, because how a prompt is constructed shapes what is returned [8]. Fourth, the output quoted or closely paraphrased, since summarizing a wrong answer usually removes the feature that made it persuasive. Fifth, what the clinician did with the output and why. Sixth, the outcome and what the clinician would do differently. The account is written in the first person because it focuses on the clinician's reasoning and describes an actual encounter; constructed examples belong to a different genre.

The purpose of accumulating such reports is signal detection, and the precedent is pharmacovigilance [9]. Spontaneous reporting systems remain the cornerstone of post-marketing drug safety surveillance despite well-recognized limitations [10]. They persist because a failure mode must be described before it can be measured, and because rare or context-dependent failures do not reliably appear in the populations trials enroll. The same asymmetry applies to clinical AI failure modes: a benchmark can quantify a failure that someone has already named, but it cannot discover one that no one has reported. A corpus of consultation reports is proposed as a source of candidate failure modes for later systematic study, not as a substitute for that study. Its limitations are those of any spontaneous reporting system, and they are substantial, including the absence of a denominator, selection toward memorable cases, and reconstruction after the outcome is known.

Those limitations bound what the format can establish, not whether it is worth establishing. The Internet Medical Journal has defined the AI consultation report as an article type, requires the six elements, and states plainly that accounts of unhelpful, misleading, or confidently incorrect output are as welcome as accounts of assistance. Without that provision, the corpus will fill with successes and will misdescribe the technology to the readers who most need an accurate description. Every journal that adopts the format enlarges the shared corpus, and every clinician who contributes to it helps name a failure mode before a colleague encounters it unwarned. Reports are anonymized stringently enough that the patient cannot be identified; a case that cannot be told without identifying detail belongs in a conventional case report with written consent. Clinicians are already consulting these systems about real patients, and the open question is whether the profession collects what they learn or leaves it in the corridor.

Declarations

Funding: This study did not receive any external funding.

Conflicts of Interest: The author reports no conflicts of interest.

Data Availability: Not applicable.

Research Ethics Statement: Not applicable. This perspective did not involve human subjects research, animal research, or protected health information.

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

References

1. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31: 60–69. doi:10.1038/s41591-024-03425-5

2. Perlis RH, Fihn SD. Evaluating the Application of Large Language Models in Clinical Research Contexts. JAMA Netw Open. 2023;6: e2335924. doi:10.1001/jamanetworkopen.2023.35924

3. Wong EYT, Verlingue L, Aldea M, Franzoi MA, Umeton R, Halabi S, et al. ESMO guidance on the use of Large Language Models in Clinical Practice (ELCAP). Ann Oncol. 2025;36: 1447–1457. doi:10.1016/j.annonc.2025.09.001

4. Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PloS One. 2024;19: e0301854. doi:10.1371/journal.pone.0301854

5. Li Z, Yan C, Cao Y, Gong A, Li F, Zeng R. Evaluating performance of large language models for atrial fibrillation management using different prompting strategies and languages. Sci Rep. 2025;15: 19028. doi:10.1038/s41598-025-04309-5

6. Heston TF, Gillette J. Large language models demonstrate distinct personality profiles. Cureus. 2025;17: e84706. doi:10.7759/cureus.84706

7. Ye Y, Sarkar S, Bhaskar A, Tomlinson B, Monteiro O. Using ChatGPT in a clinical setting: A case report. MedComm – Future Med. 2023;2: e51. doi:10.1002/mef2.51

8. Heston TF, Khun C. Prompt Engineering in Medical Education. Int Med Educ. 2023;2: 198–205. doi:10.3390/ime2030019

9. Meyboom RHB, Egberts ACG, Edwards IR, Hekster YA, De Koning FHP, Gribnau FWJ. Principles of Signal Detection in Pharmacovigilance: Drug Saf. 1997;16: 355–365. doi:10.2165/00002018-199716060-00002

10. Pacurariu AC, Straus SM, Trifirò G, Schuemie MJ, Gini R, Herings R, et al. Useful Interplay Between Spontaneous ADR Reports and Electronic Healthcare Records in Signal Detection. Drug Saf. 2015;38: 1201–1210. doi:10.1007/s40264-015-0341-5