Letter to the Editor

Toward Retrieval-Grounded Evaluation for Conversational Large Language Model–Based Risk Assessment

Yihan Hu, BSc, MPhil

MRC Epidemiology Unit, University of Cambridge, Cambridge, England, United Kingdom

Corresponding Author:

Yihan Hu, BSc, MPhil
MRC Epidemiology Unit, University of Cambridge
The Old Schools, Trinity Ln
Cambridge, England CB2 1TN
United Kingdom
Phone: 44 07526543793
Email: yh623@cam.ac.uk

Related Articles:
Comment in: https://ai.jmir.org/2026/1/e91981
Comment on: https://ai.jmir.org/2025/1/e67363

JMIR AI 2026;5:e90759; doi: 10.2196/90759

Keywords: personalized risk assessment; large language model; artificial intelligence; conversational AI; COVID-19

We read with great interest the paper by Roshani et al [1] on a generative large language model (LLM)–powered conversational app for pediatric COVID-19 severity risk assessment. The study makes a timely and valuable contribution by demonstrating how white-box LLMs (LLaMA2 [Meta], T0 [BigScience], and Flan-T5 [Google]) can be fine-tuned in low-data settings and deployed within an end-to-end mobile workflow. We also commend the authors for emphasizing local deployment and interpretability through attention-based feature importance.

We would like to highlight one methodological consideration that may affect how readers interpret the reported accuracy and the system's readiness for clinical or public-facing use. In the study, model outputs are operationalized as a binary risk score derived from the token probabilities of "yes" versus "no," and performance is primarily summarized using the area under the receiver operating characteristic curve (AUC) [1]. While this is an appropriate discriminative metric for outcome prediction within the dataset, the system is presented as a conversational interface intended to support real-time risk assessment. In such settings, a key safety concern extends beyond misclassification to the generation of fluent but unverifiable or incorrect clinical statements. Recent empirical evaluations of retrieval-augmented LLMs in guideline-grounded clinical tasks underscore the importance of grounding and safety-focused assessment in medical settings [2].

Recent advances in retrieval-augmented generation (RAG) challenge the implicit assumption that response fluency and stable token-level scoring sufficiently capture clinical reliability. Under an LLM-only paradigm, outputs may appear consistent and plausible, yet factual correctness is difficult to audit in the absence of explicit source grounding. By contrast, RAG enables models to condition responses on external authoritative resources, such as clinical guidelines or curated medical databases, thereby supporting source-linked verification and more stringent evaluation [3]. Importantly, this represents not merely an architectural refinement but a shift in what can be meaningfully evaluated.

We therefore suggest that future work in this area would benefit from a retrieval-grounded sensitivity analysis alongside conventional AUC reporting. For example, the same conversational pipeline could be evaluated under two conditions: (1) the current LLM-only setting and (2) an evidence-grounded setting in which responses are generated with explicit citations to a prespecified clinical corpus, such as pediatric COVID-19 guidance from the Centers for Disease Control and Prevention or equivalent institutional protocols. Evaluation could then incorporate evidence-grounded correctness, citation faithfulness, and robustness to retrieval constraints. In addition, subgroup-level audits—using metrics such as recall or F_1-score stratified by demographic or social determinant variables—could help identify whether aggregate performance masks safety-relevant disparities across populations.

We acknowledge that retrieval-grounded approaches introduce practical challenges, including retrieval latency and the need for ongoing corpus curation. Nonetheless, in

high-stakes pediatric and public health applications, these trade-offs may be justified by gains in verifiability and trustworthiness. We commend the authors for advancing conversational LLM–based risk assessment and hope these considerations help further strengthen evaluation rigor and clinical interpretability in future deployments.

Conflicts of Interest

None declared.

References

1. Roshani MA, Zhou X, Qiang Y, et al. Generative large language model-powered conversational AI app for personalized risk assessment: case study in COVID-19. JMIR AI. Mar 27, 2025;4:e67363. [doi: 10.2196/67363] [Medline: 40146990]
2. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. Apr 5, 2025;8(1):187. [doi: 10.1038/s41746-025-01519-z] [Medline: 40185842]
3. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. Preprint posted online on May 22, 2020. [doi: 10.48550/arXiv.2005.11401]

Abbreviations

AUC: area under the receiver operating characteristic curve
LLM: large language model
RAG: retrieval-augmented generation

Edited by Andrew Coristine; This is a non–peer-reviewed article; submitted 03.Jan.2026; final revised version received 07.Jan.2026; accepted 17.Feb.2026; published 12.Mar.2026

Please cite as:
Hu Y
Toward Retrieval-Grounded Evaluation for Conversational Large Language Model–Based Risk Assessment
JMIR AI 2026;5:e90759
URL: https://ai.jmir.org/2026/1/e90759
doi: 10.2196/90759