

Letter to the Editor

Authors' Reply: Toward Retrieval-Grounded Evaluation for Conversational Large Language Model-Based Risk Assessment

Mohammad Amin Roshani¹, MS; Xiangyu Zhou¹, MS; Yao Qiang², PhD; Srinivasan Suresh³, MD; Steve Hicks⁴, MD; Usha Sethuraman⁵, MD; Dongxiao Zhu¹, PhD

¹Department of Computer Science, Wayne State University, Detroit, MI, United States

²Oakland University, Rochester, MI, United States

³UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, United States

⁴Department of Pediatrics, Pennsylvania State University, Hershey, PA, United States

⁵Children's Hospital of Michigan, Detroit, MI, United States

Corresponding Author:

Dongxiao Zhu, PhD
Department of Computer Science
Wayne State University
5057 Woodward Ave
Detroit, MI 48202
United States
Phone: 1 3135773104
Email: dzhu@wayne.edu

Related Articles:

Comment on: <https://ai.jmir.org/2025/1/e67363>

Comment on: <https://ai.jmir.org/2026/1/e90759>

JMIR AI 2026;5:e91981; doi: [10.2196/91981](https://doi.org/10.2196/91981)

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

We thank the authors of the letter to the editor [1] for their thoughtful perspective and for highlighting the potential value of retrieval-grounded sensitivity analyses alongside conventional area under the receiver operating characteristic curve (AUC) reporting. We welcome this opportunity to further clarify the design choices and evaluation scope of our published work.

As described in Roshani et al [2], the system explicitly provides 2 distinct user-facing interfaces: a clinician-facing interface and a patient-facing interface (these are shown in Figure 3 in Roshani et al [2]), each designed for a different purpose. We agree that subgroup-level audits are essential; accordingly, our clinician-facing interface already supports stratified performance review (eg, AUC by demographic variables), laying the groundwork for more systematic fairness analyses in future versions.

In contrast, the patient-facing interface prioritizes interpretability and usability rather than aggregate performance metrics. It presents both a binary classification (severe vs nonsevere) and an individualized continuous risk score derived from the model's logit output, where higher values

indicate greater severity. This output is further complemented by attention-based feature importance to support transparent, conversational risk assessment. Notably, the patient-facing interface produces no free text in clinical language, so the risk of hallucinated clinical statements is not applicable to our system's current design.

Regarding retrieval grounding, the letter raises an important direction for future evaluation. We note that retrieval-based augmentation requires the availability of stable, domain-specific corpora. In the published study, our primary objective was to investigate the feasibility and performance of generative white-box large language models (LLMs) in low-data, emergent-disease settings. As such, we primarily considered a novel infection scenario, in which trusted external knowledge sources may be sparse, incomplete, or evolving. This was the case for pediatric COVID-19 during its early emergence, when authoritative, age-specific guidelines were limited, making LLM-only approaches more practical at the time. In this context, the system is intended to provide an initial, clinician-informed risk assessment, leveraging a small number of curated cases incorporated

during fine-tuning, rather than relying on external retrieval. We therefore presented the reported system as an initial in-house baseline, designed to function in early-stage or data-limited settings.

As clinical knowledge bases mature, the same conversational pipeline can naturally be extended to incorporate

explicit evidence grounding and source citation, enabling more personalized and evidence-supported risk assessment. We appreciate the authors' comments in highlighting this direction and view it as complementary to, rather than in conflict with, the scope and objectives of the published work.

Funding

Research reported in this publication was supported by the Eunice Kennedy Shriver Institute of Child Health and Human Development of the National Institute of Health under awards R61HD105610 and R33HD105610.

Conflicts of Interest

SH is named as a co-inventor on a patent for the diagnostic use of salivary RNA in neurologic disorders. He previously served as a scientific advisory board member for Quadrant Biosciences and Spectrum Solutions. All other authors declare no conflicts of interest.

References

1. Hu Y. Toward retrieval-grounded evaluation for conversational large language model-based risk assessment. JMIR AI. 2026;5:e90759. [doi: [10.2196/90759](https://doi.org/10.2196/90759)]
2. 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](https://doi.org/10.2196/67363)] [Medline: [40146990](https://pubmed.ncbi.nlm.nih.gov/40146990/)]

Abbreviations

AUC: area under the receiver operating characteristic curve

LLM: large language model

Edited by Andrew Coristine; This is a non-peer-reviewed article; submitted 30.Jan.2026; accepted 17.Feb.2026; published 12.Mar.2026

Please cite as:

Roshani MA, Zhou X, Qiang Y, Suresh S, Hicks S, Sethuraman U, Zhu D

Authors' Reply: Toward Retrieval-Grounded Evaluation for Conversational Large Language Model-Based Risk Assessment JMIR AI 2026;5:e91981

URL: <https://ai.jmir.org/2026/1/e91981>

doi: [10.2196/91981](https://doi.org/10.2196/91981)

© Dongxiao Zhu, Mohammad Amin Roshani, Xiangyu Zhou, Srinivasan Suresh Yao Qiang, Steve Hicks, Usha Sethuraman. Originally published in JMIR AI (<https://ai.jmir.org>), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.