

Letter to the Editor

Authors' Reply: Same Name, Moving Target: Model Version Drift in Syntheses of Large Language Models for Mental Health

Benjamin Johnson, BA (Hons); Janni Leung, PhD; Gary Chung Kai Chan, PhD

The School of Psychology, National Centre for Youth Substance Use Research, Brisbane, Queensland, Australia

Corresponding Author:

Benjamin Johnson, BA (Hons)
The School of Psychology
National Centre for Youth Substance Use Research
31 Upland Road
Brisbane, Queensland 4067
Australia
Phone: 61 429894154
Email: benjamin.johnson@uq.edu.au

Related Articles:

Comment on: <https://ai.jmir.org/2026/1/e87730>

Comment on: <https://ai.jmir.org/2026/1/e111307>

JMIR AI 2026;5:e112028; doi: [10.2196/112028](https://doi.org/10.2196/112028)

Keywords: large language models; generative artificial intelligence; generative AI; GenAI; mental health; primary health care; screening; clinical decision support; psychotherapy; digital health; eHealth; PRISMA

We thank Zhang [1] for highlighting model-version drift as an important issue in the synthesis of evidence on large language models (LLMs). We agree that labels such as “GPT-4” may refer to different model versions or snapshots with different behavior and performance.

This issue is not unique to LLM research. In health research, interventions included in systematic reviews and meta-analyses are rarely exactly the same across studies. LLMs add another layer of complexity because models reported under the same name may change over time.

In our review [2], we grouped models as closely as possible based on information reported in the primary studies. Our data extraction captured model version and evaluation date, as reported in Multimedia Appendix 2 of our original article. The appendix includes snapshot-level identifiers for several studies and an evaluation-date field for every included study. Where a primary study did not report an exact model version or evaluation date, we did not infer the missing detail; evaluation dates absent from the primary report were recorded as “not explicitly reported.” Model-version drift may contribute to the substantial between-study heterogeneity observed in our meta-analysis. High I^2 values and wide prediction intervals reflect substantial variability, but cannot be used to determine how much is attributable specifically to model version rather than prompting strategies,

populations, tasks, data sources, or outcome definitions. The pooled estimates therefore should not be interpreted as the performance of a single, stable model, but as an average estimate across several closely related models. Our finding that GPT-4 (OpenAI) had higher pooled sensitivity was descriptive: the model groups were pooled separately, no formal between-group test was performed, and the finding should not be interpreted as evidence that GPT-4 was superior to GPT-3.5.

Better reporting is needed. Future studies should report the exact model and version, date of testing, access route, prompts, and relevant model settings. TRIPOD-LLM (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-LLM) provides an important framework for improving reporting and reproducibility [3] and would allow future systematic reviews to better examine model-version effects.

However, improved reporting alone will not address the broader evidence gap. LLMs are already being used in practice while the evidence base remains limited and rapidly changing. A wider range of study designs, including pragmatic trials and well-designed observational studies using causal inference methods, will be needed to estimate the real-world effects of LLM use in mental health care [4]. Evidence synthesis will also need to become more timely.

AI-assisted screening and increasingly automated systematic review systems may support more frequent updating as new models and evidence emerge [5].

research and evidence-synthesis methods that can keep pace with rapidly changing technology.

The commentary therefore highlights a broader challenge: we need both better reporting of what is being evaluated and

Acknowledgments

There was no use of any generative AI tools or technologies in the preparation of this manuscript.

Funding

Preparation of this reply received no specific funding. BJ, JL, and GCKC were supported by the Australian National Health and Medical Research Council. The funder had no role in preparing or submitting the reply.

Authors' Contributions

Writing - original draft: BJ, JL, GCKC

Writing - review and editing: BJ, JL, GCKC

Conflicts of Interest

None declared.

References

1. Zhang Sr B. Same name, moving target: model version drift in syntheses of large language models for mental health. JMIR AI. 2026(5):e111307. [doi: [10.2196/111307](https://doi.org/10.2196/111307)]
2. Leung J, Johnson B, McRae K, et al. Generative large language models in mental health care settings: systematic review and meta-analysis. JMIR AI. Aug 31, 2026;5:e87730. [doi: [10.2196/87730](https://doi.org/10.2196/87730)] [Medline: [42684205](https://pubmed.ncbi.nlm.nih.gov/42684205/)]
3. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: [10.1038/s41591-024-03425-5](https://doi.org/10.1038/s41591-024-03425-5)] [Medline: [39779929](https://pubmed.ncbi.nlm.nih.gov/39779929/)]
4. Chan GCK, Lim C, Sun T, et al. Causal inference with observational data in addiction research. Addiction. Oct 2022;117(10):2736-2744. [doi: [10.1111/add.15972](https://doi.org/10.1111/add.15972)] [Medline: [35661462](https://pubmed.ncbi.nlm.nih.gov/35661462/)]
5. Chan GCK, He E, Leung J, Verspoor K. A comprehensive systematic review dataset is a rich resource for training and evaluation of AI systems for title and abstract screening. Res Synth Methods. Mar 2025;16(2):308-322. [doi: [10.1017/rsm.2025.1](https://doi.org/10.1017/rsm.2025.1)] [Medline: [41626972](https://pubmed.ncbi.nlm.nih.gov/41626972/)]

Abbreviations

LLM: large language model

TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model

Edited by Andrew Coristine; This is a non-peer-reviewed article; submitted 14.Sep.2026; final revised version received 17.Sep.2026; accepted 28.Sep.2026; published 15.Oct.2026

Please cite as:

Johnson B, Leung J, Chan GCK

Authors' Reply: Same Name, Moving Target: Model Version Drift in Syntheses of Large Language Models for Mental Health

JMIR AI 2026;5:e112028

URL: <https://ai.jmir.org/2026/1/e112028>

doi: [10.2196/112028](https://doi.org/10.2196/112028)

© Benjamin Johnson, Janni Leung, Gary Chung Kai Chan. Originally published in JMIR AI (<https://ai.jmir.org>), 15.Oct.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.