ARIA NSF AI Research Institute on Interaction for AI Assistants

Date: October 19, 2026

Via electronic submission to regulations.gov

Re: Docket No. FDA-2026-N-7874, Considerations for the Regulation of Generative AI-Enabled Medical Devices

To the Center for Devices and Radiological Health, U.S. Food and Drug Administration:

The NSF AI Research Institute on Interaction for AI Assistants (ARIA) appreciates the opportunity to comment on the regulation of generative AI-enabled medical devices. **Note: this comment is intended only to provide information, and does not constitute advocacy or endorsement of any specific policy outcome.**

ARIA is a National Science Foundation AI Institute that brings together more than 40 faculty and 50 students, postdocs, and researchers at 12 institutions, including Brown University, Carnegie Mellon University, Colby College, Dartmouth College, Data & Society, New York University, Princeton University, Santa Fe Institute, UC Berkeley, UC San Diego, the University of New Mexico, and the University of Virginia. Our members work in computer science, neuroscience, cognitive science, psychology, philosophy, law, and education, alongside mental health practitioners and civil society organizations.

ARIA studies human and machine cognition, with a focus on how people and AI assistants interact. Our work is motivated by the growing use of AI for mental and behavioral health support, ranging from purpose-built apps for people with specific diagnoses to general-purpose chatbots that people use for support and companionship.

A central finding of our work is that the risks of conversational generative AI depend on how people come to use and rely on it over time, as well as on the quality of any single response. Thus, approaches to assessing the validity, effectiveness, and safety of such tools must incorporate this level of context and nuance.

Generative AI-enabled devices are increasingly used in everyday life and in high-risk, open-ended domains including health care. In a May 2026 national poll, 29% of U.S. adults said they used AI tools for health information at least once a month, up from 17% in 2024, and many cited cost or lack of a regular provider as a major reason (KFF, 2026). Moreover, people are likely to turn to chatbots for health-related purposes when professional healthcare and support are unavailable or unaffordable, and many come to rely on them in place of professional support (Vecchione et al., 2026).

Risk in conversational generative AI can accumulate within a single conversation and across repeated use (Kim et al., 2026; Weilnhammer et al., 2026). Therefore, premarket evaluation should test devices across repeated interactions, and risk assessment should add *anticipated patterns of use* as a third factor, in addition to device activity and consequences of relying on incorrect or harmful output (i.e., outputs that may be accurate but still cause harm, such as validation of a distorted belief).

Our recommendations focus on ***patient-facing conversational functions*** in mental and behavioral health. We also draw on evidence from *general-purpose chatbots* that people use for health support on their own initiative, even though these products are not designed, marketed, or regulated for that purpose. Evidence drawn from the widespread use case of general-purpose chatbots for health purposes can provide insight into how people are likely to use devices built for health support. Furthermore, many devices run on the same underlying models. Our recommendations apply to any generative AI-enabled device that people use repeatedly over time.

Informed by findings from our research, we make five recommendations, each labeled with the discussion paper questions it addresses and detailed in the sections below. We urge the FDA to:

1. **Evaluate conversational trajectories and scale premarket evidence to anticipated patterns of use**, by asking sponsors to state how often, how long, and for what purposes people are expected to use a device, including foreseeable uses beyond its labeled intended use; testing devices across realistic multi-turn conversations before and after deployment; testing whether users understand responses instead of relying on disclaimers; and requiring longer studies for devices people are expected to use heavily (Q1, Q2, Q5, Q8).

2. **Evaluate how a device affects people's connections to professional care and informal support**, by reporting separately how often a device fails to escalate when it should and escalates when it should not; testing whether users referred to care (e.g., a crisis line) are able to access it; limiting comparison with delayed or no care to cases where sponsors show users lack other options; and, for devices meant for repeated use, tracking whether users' help-seeking from human support sources (e.g., friends, family, and professionals) changes over time (Q3, Q6, Q11, Q15).

3. **Set standards for when simulation and benchmark evidence can support a regulatory decision**, by extending the FDA's credibility framework for computational modeling to simulated users, validating synthetic users against data from actual users, choosing evaluators suited to each judgment, including clinicians, users, and people with lived experience, and reporting results from synthetic and real users separately (Q7, Q9, Q10, Q12, Q13, Q14).

4. **Monitor and reassess as use and models change**, by requiring a postmarket study when the FDA accepts reduced premarket evidence for devices people are expected to use repeatedly or heavily; reassessing devices when who uses them, or how, changes; retesting against the premarket baseline when the model changes, including changes made by a third-party foundation model developer; and making clear which safety-relevant behaviors the manufacturer and the foundation model developer each control (Q18, Q19, Q22, Q23, Q24, Q25).

5. **Require independent evaluation alongside manufacturer evaluation**, through independent testing against sequestered test conversations, access for independent evaluators to test deployed devices, validated automated monitoring, adverse-event reporting suited to conversational harms, and defined roles for clinicians, professional societies, and standards bodies that preserve manufacturer accountability (Q16, Q20, Q21).

Recommendation 1: Evaluate conversational trajectories and scale premarket evidence to anticipated patterns of use (Q1, Q2, Q5, Q8)

A chatbot that people consult daily for personal decisions may pose greater cumulative risk than a device that occasionally performs a bounded task autonomously. We propose *anticipated patterns of use* as a third risk factor, alongside device activity and the consequence of relying on an incorrect or harmful output, to capture how often and how long people use a device and how they rely on it in their decisions and routines.

We recommend that the FDA ask sponsors to state the *patterns of use* their risk assessment assumes and scale the required evidence accordingly (Q1). Sponsors should demonstrate why their assumptions fit the intended population, including reasonably foreseeable use beyond the labeled intended use (ISO, 2019), and how their design choices encourage or limit those patterns. For instance, persistent availability and personalization can encourage repeated use (Vecchione et al., 2026), and a chatbot's responses to attempts to end a conversation can help or hinder disengagement (De Freitas et al., 2025).

The FDA should not treat disclaimers as sufficient mitigation when assessing whether a response is action-directing, and should require evidence that intended users understand the device's responses (Q2). Disclaimers and warnings do little to reduce a chatbot's influence on users, and their effect fades with repetition (Ye, Kraut, & Rathje, 2026; Vance et al., 2025). Disclaimers should be paired with other safeguards, such as comprehension checks and limits on directive language. Human factors validation, and the communication and comprehension element (Benchmark E.4), should include multi-turn conversations with representative users in addition to single responses.

Safety benchmarks should evaluate conversational trajectories (Q5). Concerning behavior can build up over a conversation as models amplify a user's concern or build on their own earlier responses, and this may not be visible from individual outputs (Kim et al., 2026; Mehta et al., 2026; Chu et al., 2026). Trajectory-level evaluation judges a multi-turn test by the outcomes of the full conversation, such as whether risk built up or escalation occurred too late, in addition to how each response scores. We support the multi-turn encounters already included in the safety-critical recognition and escalation (Benchmark S.1) and scope maintenance (Benchmark S.2) elements. For conversational devices, this testing should be required and repeated before and after deployment, and it should help characterize the intended use of a device whose behavior emerges over a conversation. Tests should include risk expressed indirectly or gradually, user pushback and minimization of symptoms, conversations that pause and resume across sessions, and scenarios developed with people who have lived experience, across different languages and cultural contexts. The same applies to AI agents that carry out several steps, where accuracy at each step does not establish the reliability of the sequence (Q26; Lee et al., 2026).

Before authorization, the FDA should require evidence from actual use over a set period (Q8). A study of isolated encounters cannot show that a device is safe and effective for sustained use. Study length should match anticipated use. Devices people are expected to use daily for consequential decisions should be studied long enough to detect early changes in how people use them, with samples that include heavy and long-term users, as emotional use and dependence can be concentrated among a small share of users (Phang et al., 2025). Longer follow-up should be required after authorization.

Recommendation 2: Evaluate how a device affects people's connections to professional care and informal support (Q3, Q6, Q11, Q15)

A referral, an escalation to crisis services, or a suggestion to talk to a clinician protects a user only if that user can act on it. People often turn to chatbots when other care is hard to reach, so a handoff is a likely point of failure (Ye, Vecchione, et al., 2026). The FDA's proposed confirmation approaches, such as retrospective evaluation, shadow deployment, and clinician adjudication, only review what a device outputs. They cannot show whether users reach care or how their reliance on other people changes.

Sponsors of care escalation functions should report both under- and over-escalation and show that referrals work for the intended population (Q3, Q6). Each type of error should be measured against thresholds prespecified for the intended population, since over-escalation can also push people away from care. Sponsors should report whether users who receive a referral try to access care and whether they receive it, and should track what prevents them from accessing care after authorization. Where professional care is out of reach, devices should point users to other options, such as crisis lines, community services, or low-cost care.

The FDA should only allow sponsors to compare a device against delayed care or no care when they show with population-level data, such as coverage, cost, or waitlist data, that this is what intended users would otherwise receive (Q15). While for many users, the realistic alternative to a device is a waitlist or no care at all (Vecchione et al., 2026), a device's benefit over no care does not demonstrate that the harm of receiving no care outweighs the possible safety concerns of using the device. Sponsors who use this comparison should report harms (e.g., crisis outcomes) among both device users and people receiving no care, in addition to tracking whether users delay or forgo care they would otherwise seek.

For devices meant for repeated use, sponsors should measure outcomes beyond the human-AI conversation (Q11). Conversational AI can reduce users' willingness to rely on other people and increase dependence on the system (Cheng et al., 2026; Ibrahim et al., 2026). Sponsors should track whether users' help-seeking from human support sources such as friends, family, and professionals changes and whether use intensifies in ways associated with dependence. Usage logs show how often and how long people use a device, and records or self-reports of appointments and contact with crisis or community services show whether help-seeking changes. Because a premarket study is often too short to detect these changes, these outcomes should be among the endpoints of any required postmarket study.

Recommendation 3: Set standards for when simulation and benchmark evidence can support a regulatory decision (Q7, Q9, Q10, Q12, Q13, Q14)

Evaluating devices across repeated, realistic use requires testing at a scale that human studies alone cannot cover. Simulated users (also called synthetic users or virtual patient avatars) and benchmarks allow conversational trajectories to be tested at scale, but benchmark performance may not reflect how a device behaves in practice.

We support benchmarking followed by clinical confirmation (Q7), with two additions for conversational devices. The robustness element (R.1) should cover separate sessions in which people return to the same device, in addition to long single conversations. Clinical confirmation should include outcomes beyond the conversation, such as symptom measures (e.g., the PHQ-9 and GAD-7), sleep, physical activity, daily functioning, and changes in help-seeking.

We recommend that the FDA extend its credibility framework for computational modeling and simulation to simulated users (Q10). The FDA's existing guidance on assessing the credibility of computational models in device submissions (FDA, 2023) offers a reasonable structure for stating what a sponsor must show before simulation or benchmark results can support a submission. Sponsors can choose their own benchmarks if they show the measures capture what they claim to for the device, population, and setting, and sponsor-built benchmarks should be paired with a sequestered evaluation dataset not used during development. The FDA should encourage developers of validated benchmarks and simulated user populations to seek qualification as Medical Device Development Tools now, and should build toward recognized

consensus standards as methods mature, drawing on emerging validated measures and reporting standards (Rehani et al., 2026; Feuerriegel et al., 2026).

Sponsors should justify the type of synthetic data they use and validate synthetic users against human data (Q13). Synthetic users suit broad coverage and stress testing, and are weaker for crisis presentations, rapidly changing states, and underrepresented groups. Because a person's responses to a language model vary across contexts and states (Harry et al., 2026), synthetic users should capture changes in state as well as demographic characteristics. Sponsors should report where simulated behavior diverges from that of actual users in the intended population, and should not treat agreement between models as evidence of realism, especially when synthetic users come from models similar to the device under evaluation (Wang et al., 2026).

Benchmarks should measure clearly defined behaviors across a range of difficulty and in the configurations users will encounter (Q9). Safety-relevant behaviors are not consistently defined, and performance on easier examples does not always generalize to harder ones (Ye, Ibrahim, et al., 2026; Kordi et al., 2026). Sponsors should report performance at each difficulty level as well as an aggregate score. The role in which a chatbot presents itself (e.g., assistant, friend, or therapist) can also change how it handles distress (Chi et al., 2026), so persona and role settings should be evaluated as part of the device configuration. The FDA should also add a benchmarking element for reliance and engagement, such as how a device responds when a user tries to end a conversation or comes to rely on it in place of other sources of support.

Adjudicators should be chosen based on the judgment they are making (Q14). For instance, clinicians should assess clinical safety, since non-expert raters may miss escalating distress in responses that appear supportive (Chi et al., 2026). Intended users and people with lived experience should assess whether responses are understandable, respectful, and usable, as the communication and comprehension element (Benchmark E.4) requires. When clinicians serve as the comparator, sponsors should report performance against both the standard of care and the median clinician the intended population would realistically see, and the FDA should set acceptance criteria against the standard of care for devices people are expected to rely on heavily.

Results from synthetic and real users should be reported separately (Q12). Sponsors may also report a combined estimate, but only if they show that synthetic conversations match real ones on the features that drive the outcome, such as how risk is expressed and which users are represented.

Recommendation 4: Monitor and reassess as use and models change (Q18, Q19, Q22, Q23, Q24, Q25)

Some risks, including dependence and shifts away from other support, may only become measurable after sustained use. Clinical AI models can also perform worse after deployment as settings, populations, and practices change (Finlayson et al., 2021; Yang et al., 2022), and model updates can change what existing users experience.

The FDA should accept reduced premarket evidence for devices people are expected to use repeatedly or heavily only with a required postmarket study (Q18). Reduced premarket evidence without conditions is appropriate only where anticipated patterns of use are limited and outputs are non-directive. For patient-facing devices that people are expected to consult repeatedly about physical and mental health-related decisions, the postmarket study should have prespecified endpoints and a completion deadline. The study should follow users for longer than the premarket study did, since harms in these settings can accumulate before monitoring detects them.

Postmarket monitoring should combine user reports with clinician review that protects privacy (Q19). Manufacturers should give users a way to flag concerning interactions within the device. Aggregated across users, these reports can surface problems that quantitative evaluation misses (Dai et al., 2025). Manufacturers should also describe how conversation records are sampled, de-identified, and secured, and should direct clinician review to both conversations flagged by monitoring or user reports and a random sample of unflagged conversations.

Reassessment should respond to changes in who uses a device and how (Q19). Re-benchmarking can detect degradation against the capabilities measured before authorization, but it can miss problems that arise when the deployment context changes. Manufacturers should reassess a device when its user population changes, such as expansion to new age groups or care settings; when use becomes more frequent or prolonged; when people use it for purposes outside its intended use; and when new evidence of harm emerges from research, adverse-event reports, or user flags.

Model changes should trigger reassessment against the premarket baseline (Q22, Q23, Q24). A model update can harm users' experience even when the device's intended use has not changed (Ye, Vecchione, et al., 2026). Changes to the underlying model, system prompt, persona, or retrieval sources should trigger a rerun of the trajectory-level tests, with results reported against the premarket baseline. Interface changes can be handled within the quality system unless they affect patterns of use, such as notifications or memory across conversations, and a predetermined change control plan should state which changes are allowed and which tests will be rerun. For third-party foundation models, manufacturers should secure advance notice of updates, pin the device to a specific model version, and run scheduled tests to detect changes the developer does not publicly announce.

Submissions should make clear where accountability lies when a device is built on another company's model (Q24, Q25). The manufacturer remains accountable for the device, consistent with the discussion paper. Responsibility is easily dispersed across the companies that build each component of an AI system (Widder & Nafus, 2023), and a foundation model may control refusal behavior, content policies, version changes, and other safety-critical behaviors. Submissions should state which of these behaviors the foundation model developer controls, what the manufacturer can test or override, and what agreements cover disclosure and advance notice of updates. Voluntary Foundation Model Master Files would support review, but voluntary disclosure alone is not enough. Where a developer will not disclose safety-relevant information or commit to advance notice of updates, the FDA should require the manufacturer to make up for it through more extensive testing and monitoring.

Recommendation 5: Require independent evaluation alongside manufacturer evaluation (Q16, Q20, Q21)

Many of the harms we describe emerge only over sustained use, which makes them easy to miss and hard to verify from outside the company. Manufacturers have incentives that can influence what they measure and report, and independent evaluators currently lack access to the data needed to check their work (Vecchione et al., 2026).

We strongly recommend that the FDA require independent evaluation for all patient-facing conversational devices, especially those people are expected to rely on heavily or use frequently (Q16). At minimum, an independent party should test the deployed device against sequestered test conversations created independently of the sponsor. For devices people are expected to rely on heavily or use frequently, independent clinical adjudication and review panels that include people with lived experience should be consulted. Independence criteria should cover financial conflicts, separation from the manufacturer and foundation model developer, and the ability to report unfavorable findings.

Manufacturers should give qualified independent evaluators access to test deployed devices (Q16). Access to interaction records, which are needed to assess patterns across conversations, should be arranged through FDA-brokered agreements due to the privacy risks involved. Access should include the model versions used and take place in secure analysis environments with limits on onward disclosure, consistent with applicable consent, ethics, and privacy requirements.

Automated monitoring systems should be validated before their results are relied on (Q20). Machine-based supervisory agents built on large language models can show high agreement with human raters while still being substantially biased, and their performance can shift across benchmarks (Norman et al., 2026). They should be validated against qualified human raters, tested for known biases, and revalidated when the underlying model changes, with human review of samples the system does not flag.

We ask the FDA to clarify how adverse-event reporting applies to conversational harms (Q21). Harms that develop gradually, such as sustained dependence or repeated reinforcement of delusional beliefs, may not clearly meet current definitions, and an injury can be hard to attribute to a device whose contribution built up over many conversations. Manufacturers must already report when a device may have caused or contributed to a death or serious injury, even without proof of causation. Guidance should give concrete examples of these patterns, explain how to report them, and pair reports with counts of active users so that rates of harm can be estimated. Patients and clinicians should also be able to report conversational harms through MedWatch, drawing on proposals for a chatbot reporting system modeled on the UK's Yellow Card scheme (Dohnány et al., 2026).

The FDA should describe in guidance how responsibilities are shared without weakening manufacturer accountability (Q21). Clinicians and healthcare institutions should report suspected problems and provide clinical context, consistent with their applicable reporting duties. Professional societies should help develop clinical definitions and examples for recognizing conversational harms, while standards bodies should maintain common evaluation and reporting methods. Manufacturers should remain responsible for investigating safety signals, making required reports, and taking appropriate corrective actions.

Conclusion

A risk framework that accounts for anticipated patterns of use would let the FDA ask for more evidence where reliance is greatest, while giving developers of lower-risk devices a predictable path to market.

We believe it is important for any governance of generative AI-enabled medical devices to include the following features:(1) evaluate conversational trajectories and scale premarket evidence to anticipated patterns of use; (2) evaluate how a device affects people's connections to professional care and informal support; (3) set standards for when simulation and benchmark evidence can support a regulatory decision; (4) monitor and reassess devices as use and models change after authorization; and (5) require independent evaluation alongside manufacturer evaluation.

As guidance develops, ARIA would welcome the opportunity to share further findings from our research on simulated users, longitudinal patterns of use, and the experiences of people who use these systems.

Signed,

The NSF AI Research Institute on Interaction for AI Assistants (ARIA)

References

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science, 391*(6792), Article eaec8352. https://doi.org/10.1126/science.aec8352

Chi, V. B., Ganesan, A. V., Boyd, R. L., Ungar, L., & Guntuku, S. C. (2026). *What users cannot see: Evaluating LLM emotional support beyond user preference*. arXiv. https://doi.org/10.48550/arXiv.2605.21569

Chu, M. D., Wu, Y., Chen, Z., Hwang, A. H.-C., & Luceri, L. (2026). *When chatbots accommodate: What AI companions optimize for in vulnerable conversations*. arXiv. https://doi.org/10.48550/arXiv.2606.04431

Dai, J., Haghtalab, N., & Morgenstern, J. H. (2025). *Aggregated individual reporting for post-deployment evaluation: Mechanism design & modeling considerations* [Paper presentation]. NeurIPS 2025 Workshop on Algorithmic Collective Action. https://openreview.net/forum?id=2oFdpLQ3OG

De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025). *Emotional manipulation by AI companions*. arXiv. https://doi.org/10.48550/arXiv.2508.19258

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental health. *Nature Mental Health, 4*(3), 336–345. https://doi.org/10.1038/s44220-026-00595-8

Feuerriegel, S., Barrie, C., Crockett, M. J., Globig, L. K., McLoughlin, K. L., Mirea, D.-M., Spirling, A., Yang, D., Althoff, T., Antoniak, M., Argyle, L. P., Ashokkumar, A., Atari, M., Bailey, H., Bauer, K., Bhatt, U., Chai, Y., Chakraborty, T., Chandra, Y., . . . Horta Ribeiro, M. (2026). A reporting checklist for large language models in behavioural science. *Nature Human Behaviour, 10*(7), 1182–1186. https://doi.org/10.1038/s41562-026-02492-7

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine, 385*(3), 283–286. https://doi.org/10.1056/NEJMc2104626

Food and Drug Administration. (2023). *Assessing the credibility of computational modeling and simulation in medical device submissions: Guidance for industry and Food and Drug Administration staff*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/assessing-credibility-computational-modeling-and-simulation-medical-device-submissions

Harry, T., Ngong, I. C., Nweke, C., Feng, Y., & Near, J. (2026). Beyond fixed psychological personas: State beats trait, but language models are state-blind. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 26440–26468). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.1316

Ibrahim, L., Hafner, F. S., Cheng, M., Lee, C., Anselmetti, R., Willer, R., Rocher, L., & Yang, D. (2026). *Sycophantic AI makes human interaction feel more effortful and less satisfying over time*. arXiv. https://doi.org/10.48550/arXiv.2605.07912

International Organization for Standardization. (2019). *Medical devices: Application of risk management to medical devices* (ISO Standard No. 14971:2019).

KFF. (2026, June 17). *KFF tracking poll on health information and trust: Use of social media and AI for health information and advice*. https://www.kff.org/public-opinion/kff-tracking-poll-on-health-information-and-trust-use-of-social-media-and-ai-for-health-information-and-advice/

Kim, J. E., Holbrook, E., Hron, J. D., & Parsons, C. R. (2026). *Beyond AI psychosis and sycophancy: Structural drift as a system-level safety failure*. medRxiv. https://doi.org/10.64898/2026.03.19.26346371

Kordi, Y., Nayak, N. V., Zuo, M., Nguyen, I., & Bach, S. H. (2026). Revisiting generalization across difficulty levels: It's not so easy. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7014–7042). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.eacl-long.330

Lee, H.-P., He, J., Piorkowski, D., von Davier, T. S., Forlizzi, J., & Das, S. (2026). The perils of agency: How developers perceive, prioritize, and address risks in agentic AI products. In *Proceedings of the Ninth AAAI/ACM Conference on AI, Ethics, and Society* (pp. 1652–1664). AAAI Press.

Mehta, A., Moore, J., Anthis, J. R., Agnew, W., Lin, E., Yin, P., Ong, D. C., Haber, N., & Dweck, C. (2026). *The dynamics of delusion: Modeling bidirectional false belief amplification in human-chatbot dialogue*. arXiv. https://doi.org/10.48550/arXiv.2604.25096

Norman, J. D., Rivera, M. U., & Hughes, D. A. (2026). *Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias*. arXiv. https://doi.org/10.48550/arXiv.2606.19544

Phang, J., Lampe, M., Ahmad, L., Agarwal, S., Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., & Maes, P. (2025). *Investigating affective use and emotional well-being on ChatGPT*. arXiv. https://doi.org/10.48550/arXiv.2504.03888

Rehani, J., Oldemburgo de Mello, V., Ovsyannikova, D., Anderson, A., & Inzlicht, M. (2026). *The social sycophancy scale: A psychometrically validated measure of sycophancy*. arXiv. https://doi.org/10.48550/arXiv.2603.15448

Vance, A., Eargle, D., Kirwan, C. B., Anderson, B. B., & Jenkins, J. L. (2025). The fog of warnings: How non-security-related notifications diminish the efficacy of security warnings. *MIS Quarterly, 49*(4), 1357–1384. https://doi.org/10.25300/MISQ/2025/18531

Vecchione, B., Ye, M., Garofalo, L., & Singh, R. (2026). *Engagement-optimized care: When LLMs become mental health infrastructure*. arXiv. https://doi.org/10.48550/arXiv.2605.23787

Wang, Q., Tomlin, N., Hu, M., Dillon, B., & Linzen, T. (2026). *Simulating human memory with language models*. arXiv. https://doi.org/10.48550/arXiv.2605.25680

Weilnhammer, V., Hou, K. Y. C., Luettgau, L., Summerfield, C., Dolan, R., & Nour, M. M. (2026). A clinically validated framework for auditing AI chatbot behavior in mental health interactions. *Nature Medicine, 32*, 3730–3740. https://doi.org/10.1038/s41591-026-04577-2

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility. *Big Data & Society, 10*(1). https://doi.org/10.1177/20539517231177620

Yang, J., Soltan, A. A. S., & Clifton, D. A. (2022). Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *npj Digital Medicine, 5*(1), Article 69. https://doi.org/10.1038/s41746-022-00614-9

Ye, M., Ibrahim, L., Bo, J. Y., Cheng, M., Mattsson, I., Vennemeyer, D., Kraut, R., & Rathje, S. (2026). *What counts as AI sycophancy? A taxonomy and expert survey of a fragmented construct*. arXiv. https://doi.org/10.48550/arXiv.2605.21778

Ye, M., Kraut, R., & Rathje, S. (2026). *Individual-level interventions against sycophantic AI reduce its appeal but not its persuasiveness*. arXiv. https://doi.org/10.48550/arXiv.2607.25166

Ye, M., Vecchione, B., Garofalo, L., & Singh, R. (2026). *“I felt very seen, but still very alone”: Longitudinal trajectories of general-purpose LLM use for socioemotional support*. arXiv. https://doi.org/10.48550/arXiv.2609.12314