

Human-AI Collaboration for Co-Managing Epilepsy in Real Time

Corresponding Author: Mr Zack Goldblum

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

A. This manuscript presents a genuinely novel and technically impressive proof-of-concept platform for bidirectional patient-device communication in epilepsy management, with clear translational vision. The study's small sample ($n=13$), inpatient-only setting, predominantly qualitative/feasibility nature, incomplete statistical reporting, and fundamental limitations in seizure detection evaluation—including non-significant primary results in the prospective annotation arm—require substantial revision. The sample size issue is a mild weakness given the complexity of the undertaking - it is still a substantial contribution. The clinical problem is compelling and well-articulated: the authors correctly identify that current FDA-approved epilepsy devices (RNS, VNS, DBS) have no real-time bidirectional patient communication, and the manuscript convincingly frames this as a missed opportunity. The seizure detection models employed (SPaRCNet, ONCET, WaveNet) are explicitly acknowledged as 'not reflecting state-of-the-art performance for detecting seizures.' The baseline false alarm rate of 0.58/hour (IQR 0.35-0.77) while clinically unacceptable for an alert-based system—this translates to approximately 14 false alarms per day, is a great start to what I think is the future in our field.

B. The Human in the Loop concept of course is not novel but the context here is relatively original.

C. The seizure annotation feature was only implemented for 9 of 13 patients, and only 6 had confirmed events suitable for fine-tuning. All statistical tests for the primary innovation rest on $n=6$, which is insufficient to support the broad claims made in the abstract, introduction, and conclusion. This sample size provides inadequate power to detect clinically meaningful differences and makes the Wilcoxon signed-rank test results uninterpretable for inferential purposes. The manuscript describes multiple outcomes (sensitivity, F1, false alarms/hour, AUROC, AUPRC), multiple fine-tuning conditions (push-button, app annotations, clinician), and multiple evaluation time windows without pre-specification. This multiplicity, combined with one-sided hypothesis tests selected post-hoc in the direction of expected improvement, creates substantial risk of Type I error. The entire conversational and reasoning architecture depends on a HIPAA-secure instance of GPT-4o-mini, a proprietary model subject to unilateral updates by OpenAI. The manuscript does not describe model versioning, does not address what happens to validated system behavior when the underlying model is updated, and does not discuss the regulatory pathway for a medical device whose core decision-support layer is controlled by a third party. This is a fundamental reproducibility and translational concern. The authors also do not explicitly state whether missing survey data was excluded (listwise deletion), imputed, or analyzed as-is. We need to discuss potential bias from differential response rates by time-of-day and provide sensitivity analyses showing impact of missing data (e.g., assuming worst-case non-responders as negative cases). A small point - the study design is described as 'prospective observational' but the seizure fine-tuning evaluation was performed as a 'pseudo-prospective' post-hoc analysis for safety reasons. This distinction—that the fine-tuned models were never actually deployed in real-time during patient sessions—is critical to interpreting the results and is currently buried in the Methods section.

D/E. The fine-tuned seizure detection models were never actually deployed during patient sessions - the evaluation was performed post-hoc as a 'pseudo-prospective' simulation for safety reasons. This is a material distinction from true prospective performance, but the abstract and Results describe performance as if the system was evaluated in real-time deployment. Readers will be misled about the actual clinical readiness of the system. The System Usability Scale was completed by only 5 of 13 patients (38%), yet the mean score of 83/100 is used to characterize the system as 'excellent' in the abstract. The unvalidated 10-item visual analog scale surveys for mood, anxiety, and cognition are presented as clinically meaningful patient-reported outcomes without comparison to validated epilepsy-specific instruments. Also, the

push-button alarm validation (n=11) is presented as a stronger result but conflates two different processes: push-button alarms reflect patient perception of seizure occurrence (with known variability in postictal memory), not ground-truth EEG-confirmed events. Using these as training labels without reporting agreement with expert EEG annotation introduces label noise that is not quantified. Also, response latency is reported using median and IQR with a Mann-Whitney U test ($p=9.89 \times 10^{-21}$), and effect size is mentioned as 'rank-biserial correlation' but the actual effect size value is not reported in the results text.

F. For a prospective study - there is very little of the a-priori thinking described. Add sample size justification in addition to above. Add 95% CIs around median effect etc.

G. References appropriate

H. No concerns with clarity and context

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

Goldblum et al. present a very interesting concept of combining an LLM with ongoing EEG as a proof-of-concept for implantable conversational devices. They show that patient responses (annotations, patient-in-the-loop) can be used to improve the performance of seizure detection models, with further improvements by targeted clinician annotations (clinician-in-the-loop). This is very innovative research and nudges the field forward. I do not have any major concerns about this work.

A. Summary of the key results:

- Abstract: "An AI agent ingests ... seizure risk extracted from brain activity". Were we calculating seizure risk?

B. Originality and significance:

- Novel.

C. Data & methodology: validity of approach, quality of data, quality of presentation:

- Only 15.5% seizures detected are true seizures, which is very low and can lead to alert fatigue among patients with a high false positivity. Was that noted/measured?
- "and labeled 17.2% (n=10) as uncertain. All uncertain events (n=9)". Is uncertain 9 or 10?
- S4. Why are the stress/fatigue questions being repeated twice? Both questions are asking the same thing (how stressed are you/how calm are you, how fatigued are you/how much energy do you have). Similarly in S3 (how anxious are you/how tense are you)

D. Appropriate use of statistics and treatment of uncertainties:

- No concerns.

E. Conclusions: robustness, validity, reliability.

- No concerns.

F. Suggested improvements: experiments, data for possible revision.

- Fig 2: "Continuous EEG signals are acquired via clinical workstations and relayed through a cloud-hosted virtual machine to Azure Event Hubs for ingestion" What is the data bandwidth needed- is that feasible at a population level?

G. References: appropriate credit to previous work?

- Yes.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions.

- No concerns

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

This is a case series of a concept – that patients might be able to assist in training closed-loop seizure-suppression systems, by conversation with an LLM that in turn has control over device parameters and ability to interpret intracranial physiology. A core thesis is that "weak labels", i.e. potentially-inaccurate patient recall/labeling of seizure events (prompted by a conversation agent or, here, mostly by a push button) could still be quite useful. That thesis is reasonably substantiated and important across clinical neuromodulation. The overall study was technically complex due to the number of data handling steps and systems being interfaced, and presents a really interesting proof of concept towards a technology that could have

broad applications in enabling better/more rapid/more scalable development of closed-loop neurotechnology. It genuinely is a conversation-changing article, and in that context, Nature Medicine likely will be a reasonable venue (modulo comments below).

Other strengths of the report include:

- A reasonably well justified overall design, including some thoughts about layering LLMs to ensure accurate message handling and safety
- Specific considerations and discussion of ethical approvals for an AI-heavy study
- Discussion and demonstration of an ML pipeline that is oriented around “not making matters worse”, e.g. freezing key layers to prevent catastrophic forgetting

That said, the report as presented is substantially incomplete. There are major outcome details that are missing and need to be reported, as well as some methods details. These largely relate to not actually showing us how patient and LLM interacted. The framing is also deceptive. It is presented in the Abstract as an $n=13$ study of the approach, but based on the Methods, it appears that it may be as little as $n=2$ actually demonstrating the approach being claimed. Further, again based on the Methods, it appears that the core claim (improved seizure detection performance through real-time, LLM-driven event marking and tuning) is not actually done. First, improved seizure detection is not actually shown, and in the Discussion they explain why they probably cannot achieve it in this kind of study. Second, if I am reading this right, they never actually did real time interactive fine-tuning with the patient, and everything being plotted, especially in Figure 5, is all theoretical simulations of what might have happened.

This also highlights that there are many places where the text and figures are extremely unclear and I cannot be certain what was actually done or what is being plotted. This is not a complete scientific report.

These are major weaknesses. They are not fatal because the value of the report is more in demonstrating the technological proof of concept and starting a conversation, not in the specific clinical outcomes or the N (an $N=2$ pilot has value if the right data are shown). They still need to be corrected before this would be publishable, again because some of the claims are frankly false/deceptive as written. The article does not NEED those false claims in order to be interesting and high impact, and thus they need to come out.

Drilling into specifics,

1. It is very hard to identify exactly how many patients used the LLM system shown in Figure 2, or for what purpose. It is, however, not the $n=13$ that is claimed. In Figure 5, only 6 patients are described as having used the app for labeling. In the Methods, it is then said (p12-13) that seizure annotation was not actually implemented for 4 patients – which could mean there were $n=9$ (if this means push button) or possibly only $n=2$. What is needed is a table that shows, for each reported patient, what aspects of the study they actually participated in (using the LLM for surveys, to query data, using the push button, having fine-tuning, etc etc etc.). Further, the early claims in Abstract and Introduction need to specifically say how many patients actually used this full claimed system.
2. It appears that there was no actual adaptation of the model. If I read this right, what they actually did was have the patients interact with an untuned model for the full stay, but simulated what WOULD have happened if it were fine-tuned using their responses. (Methods p19). This is another place where the framing is deceptive, because this is very subtly buried after the claims of the article regarding fine tuning. They need to be up front and refer to this process as something like “simulated fine tuning” in the article body, abstract, etc.
3. The main effect actually demonstrated is a reduction in false alarms. This is in fact useful, but the claim “the system learned and re-trained itself to markedly improve seizure detection performance” is false both because of the above (the system did not re-train itself, there are several ways in which that statement is not true) and because there was not actually an improvement in seizure detection (sensitivity and/or PPV).
4. There is little to no data presented to explain how patients engaged with the LLM/chat interface and what the interface actually did well/poorly. Specific items of concern:
 - a. What types of data did patients request? How often? How many times did they need to re-prompt the LLM because it did not respond accurately? Did use of the LLM remain constant over the EMU stay, or increase/decrease? A pattern of diminishing engagement over the 4-5 days in the EMU would tell a very different story from increasing engagement.
 - b. Why was the patient response rate so low? What feedback did they collect that can explain this? I can somewhat understand it for surveys, and they discuss it for spikes, but I am very surprised that it is only 70% for seizure prompts. Further, how did it change over time? See above.
 - c. What were the inaccuracies in chat/data message routing claimed in Figure 3e?
 - d. What were the messages that were flagged by the LLM safety layer as unsafe, and why? If there are really only 6, then provide the text of the interactions.
 - e. Related, it is claimed that no messages from the LLM to the patient were marked as unsafe. If this is true, how do they know that the safety filter was effective? What messages WOULD have been flagged as unsafe? How did they test the patient-to-AI or AI-to-patient safety layers before deploying them? (Possibly they did not and relied on human review of the message chains, but again, then how do they know whether the AI safety review was valid if it never flagged a message?)
5. There is similarly minimal information on the specific design of this LLM architecture, including
 - a. What was done to prevent the LLMs from having context buffer leaks? How was reliable, sustained, and consistent interaction with the patient maintained?
 - b. How was each LLM layer prompted? Were any of these models specifically trained, or were they “out of the box” GPT implementations? In my opinion, the prompts themselves need to be visible to readers/reviewers. I know that I can see some of this in graph.py and did so, but not every reader wants to dig four folders deep in GitHub.
 - c. Why GPT? Possibly this is due to licensing/convenience factors, but then say so, and comment on the specific properties of the 4o model that may be better/worse for this application.
6. The LLM annotations actually appear to perform worse than the push button (median F1 and false alarm rate appear to be worse in Figure 5c than 5b). This needs to be explicitly acknowledged and discussed, because although the “Weak labels work” claim is valid, the fact that they work less well than spontaneous patient self-report suggests a problem that needs to

be remedied.

a. As a side note, it is not clear why there are only 4 patients in the Sensitivity plot for 5c.

7. Figure 5 seems to be claiming some kind of test set performance, but it is very unclear exactly what is training, validation, and test set in each of these figures. Different squares in 5A are highlighted pink, but we are never told what this pink means. Similarly, what exactly the test set is for the figures is not clear, for instance, it would appear that for the blue-pink curves in 5B-C, a model fine-tuned on all of the historical data is being re-run on all of that data minus the first block. That would be classic data leakage. Possibly that is not what was being done, but much clearer descriptions and labeling are needed to evaluate whether this is technically valid.

8. Given that only 2 patients are SEEG/iEEG, we need to know if the performance is different or the acceptability is different in patients undergoing those procedures, which can be substantially more burdensome than scalp EEG. Throughout the paper, those patients need to be visually distinguishable in figures, e.g. through symbols or coloring. This is critical because the use case, right at the start of the paper, is basically for LFP/iEEG signals from intracranial recordings, and we need to know how the acceptability and interaction went for those patients.

9. A large amount of the study data processing is performed in various cloud infrastructures, and apparently involves passing health information and unique medical record numbers through those clouds. What was done to ensure security? How were risks audited and discussed? Were these specific risks considered by the novel AI policy committee described herein?

These are important because there is currently a very fractured landscape among different institutions regarding what is/is not permitted on cloud platforms, and having a clear example here may aid other investigators.

There are lesser concerns that are more related to interest and clarification.

10. Nearly 40% of seizure messages were false positives (Figure 3c). How did patients react to that? Did they find it useful to be able to flag those? If there is information in the messages or the usability free text responses, provide it. (This relates to the point above about needing to show actual patient-LLM interactions, not stylized figure versions.)

11. The Discussion appropriately talks about how similar systems might be useful for tuning patient-specific classifiers for other brain phenomena, e.g. motor problems in Parkinson's or various psychiatric symptoms. This is valid, and a good extension of the "weak labels" claim. However, it also highlights the only reason this demonstration works: due to the availability of large, public, and extremely well curated data sets, there are broadly available seizure classifiers to form the foundation. If not for a decade or more of competitions and data sharing, as well as the availability of very clear ground truth, this study would be impossible (the concept of having a model where most layers could be frozen and a few fine-tuned would not be realistic). They need to be far more explicit that this was a necessary condition that does not yet exist in any other domain of neuro-medicine (one could argue about Parkinsonian tremor).

12. In the Methods, the seizure detection algorithms are described with a variety of specific parameters. Are these defaults? Chosen based on prior experience? Other? I have no quibble with the chosen parameters, but they need to explain WHY they were chosen.

13. Why, on page 9, is the number of uncertain events given as both $n=10$ and $n=9$?

14. On p9, a sensitivity of 0 is reported – is this for the physiologic detection of confirmed events, or the patients' own awareness of the events? It is very hard to understand how anything has a sensitivity of AND a precision of 1.0.

15. Was the clinician-in-the-loop tuning (simulated tuning) performed on top of the simulated patient-driven fine tuning, or is this a separate study?

16. On p2, design considerations are derived from "discussions with clinicians, caregivers, and patients". Some documentation of these discussions (e.g., citations to prior work) or literature citations to other studies about lived experience of persons with epilepsy would be helpful.

Finally, as a fellow Asimov fan, I must object to the invocation of the Three Laws at the end. They do not apply here. This is not a system that can experience a threat to its own existence, nor is this type of subsumptionist architecture built into the system they describe. There certainly are questions about how to design a human-interactive system for safety, but this article needs to stick to the basic facts of what they have built and what it does/does not do.

(Remarks on code availability)

I think it's a partial code drop. It doesn't contain the info to really replicate the system, just the chatbot. It does contain some basic install instructions that I am confident I could follow, but without available data and other components, it would be hard to exercise the full system.

Reviewer #4

(Remarks to the Author)

The manuscript by Goldblum et al. details a proof-of-concept study of a platform integrating artificial intelligence (AI) agents derived from a large language model (LLM) with seizure detection and ecological momentary assessments (EMAs). The platform interacts with the patient using natural language, leveraging the strength of the LLM and uses patient inputs from the EMAs to fine-tune seizure detection. The authors provide different measures of patient interaction with the platform and performance improvement of seizure detection based on patient inputs. This is a novel concept and has the potential to impact how care is delivered when using device-based treatments especially in scenarios that require device calibration over longer time frames beyond just the surgical procedure. It is also a timely study given current interest in AI agents in the context of healthcare and tries to envision a scenario where these technical advances help improve patients' life in a meaningful and accessible manner.

While this idea has promise, there are limitations that dampen enthusiasm for publication in Nature Medicine. First, my interpretation of the data presented is that it is very early stage data, showing promise but not yet clinical performance that has near term translational potential. Second, the user experience (which is a critical factor for this proposed approach) is

largely unstudied either before or after the current system implementation, leading to questions about the impact of the system even in an ideal implementation. Specific comments follow:

- The title is misleading, leading the reader to believe that this manuscript is mostly about implantable devices. While it is appreciated that the technology could be extended to an implantable device, the current manuscript is primarily with participants using surface EEG and only includes implantable electrodes with externalized hardware (clinical Natus systems) in 2 out of the 13 participants. On a related note, without taking anything away from the impressive engineering to make this system feasible, none of the implanted devices currently approved by the FDA for epilepsy are able to stream continuously without requiring specialized interfaces such as clinician programmer tablets. While this obviously may change in the future, this is a limitation in the present reality that severely limits the applicability of this platform in its presented form and this should be addressed. Overall, I would say that the writing in this manuscript needs to be edited to be careful not to overstate a claim on what has actually been achieved in the current project.

- In its current form, the utility of the platform is critically dependent on the performance of the seizure detection algorithms. Given the low baseline performance of the algorithms and short time frame of the study (4 days), it is unclear if the authors' approach will lead to improvements that are meaningful for patients. For example, a false alarm rate of about 0.2/hr that they are able to achieve by fine-tuning their seizure detection algorithm with user input, will have about 5 false alarms a day, which is still burdensome from a patient perspective. Also, the improvement seems to plateau after day 1, suggesting the models aren't adapting over time.

** A clarification question: It is not clear if the chat interface had option for patients to indicate that they were experiencing a seizure or if the patients had to use the push button. In case it is the latter, how is the seizure prediction performance reported affected by this 'missing' data?

- One potential option for the authors to address the two concerns listed above would be to use seizure prediction algorithms or features used in currently available implanted devices (e.g., NeuroPace RNS) and demonstrate this framework can improve performance and user experience.

- Multiple places in the manuscript refer to lived experience input, but in very qualitative terms (e.g., "Based upon our experience and discussions with clinicians, caregivers, and patients...", "discussions with patients indicate"). Given that the premise of this project is rooted in making a system that can "focus on what matters most to [patients]", it is a significant limitation that the lived experience voice is treated so informally. The manuscript would be significantly strengthened with data showing what features of a system epilepsy patients and caregivers actually want, or showing how the subjects in this study actually viewed the system. The subjects were queried about "usability", but did they actually find the interactions beneficial, helpful, positive, or supportive? How did they feel about a system that was learning from their brain data and behavior?

- The safety framework of the AI agent is not sufficiently detailed. Given the potential for hallucinations by LLMs, a clinical setting with a vulnerable population introduces an ethical imperative to disclose fully and clearly the safety aspects of this system. As just one example, it is not even clear to me if the authors verified or have a mechanism to verify if the information provided by the LLM was accurate (i.e., reflected accurate values from the SQL database).

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed almost every single point I raised. I have no additional criticisms.

(Remarks on code availability)

N/a

Reviewer #2

(Remarks to the Author)

The authors have made substantial revisions and have addressed all of my previous comments. I do not have further comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

This revised article does the vast majority of what I and other reviewers asked for. It tones down excessive claims, provides

a clearer description of what was done, and provides more details on the LLM architecture to support evaluation and possible replication. It remains interesting and conversation-provoking. I do have some minor comments that I believe can be handled through the editorial process without another round of peer review.

1. Patients appear to have used the system primarily for general chatting, with only 11% of their messages being queries to the system (around line 218). Comment on this further in the Discussion, please. This is an important point, because it highlights that there will be some substantial UI/UX work needed for this to achieve the vision of a co-manager/patient assistant.

2. Closely related, Figure Extended 3 shows very limited engagement with seizure prompts, and it looks like this falls off during the stay (the right side of the graph is mostly white). This suggests that novelty drove early engagement, but utility was not there to sustain it. This is again fine for proof of concept but needs explicit comment in the Discussion.

3. The decision to make the agent architecture non-stateful is logical, and probably led to more controlled/predictable behavior. On the other hand, it also likely would reduce helpfulness of a real-world system, because it cannot know a user's communication preferences well with only 200 messages of history. Many of the anecdotes around AI/LLM as useful companion to a human involve months of co-adaptation (of course, so do many of the stories of problems). Same comment – touch on this in the Discussion in terms of next steps and technology development that is still needed.

4. Revised figure 5 is good, but I notice that the two forms of delta sensitivity/F1 plots are on very different Y-axes, and having them on the same axis would highlight the substantial difference in the effects of the two retraining regimes.

(Remarks on code availability)

The revised availability is much better. I cannot run it, but then again, nobody could. I do think this would be sufficient to attempt replication.

Reviewer #4

(Remarks to the Author)

I appreciate the effort the authors went to in responding to the reviews and revising the manuscript. I would agree that there have been many improvements in the technical clarity of what was actually done and reducing the degree that the results were being overstated.

That all being said, I still find this manuscript to feel largely like a much more preliminary report than I would expect to see in Nature Medicine. Acknowledging the enormous amount of effort that the authors went to in building this system, the reality is that I think the demonstrated benefit remains limited. The authors essentially acknowledge that the approach is not yet able to demonstrate performance with translational potential, citing the slow enrollment rates in the EMU. I can fully appreciate this, but the limited data collection rate does not in and of itself elevate the impact of the results. In my view, it would be more appropriate to publish this proof of concept in a more specialized journal and then publish a report demonstrating clear clinical impact in Nature Medicine when enough data has been accumulated to clearly demonstrate the value.

The one place where I feel the authors have the opportunity to add clear impact with the enrollment numbers they have is in a formalized study of the user experience and desires. While I appreciate that the authors added a section to the discussion on this topic and agree that it is a welcome addition to the manuscript, the section primarily contains a few selected quotes from the exit survey (and citations to supporting literature). To be a substantial contribution, this section would need to contain an analysis that is more rigorous and comprehensive.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)

This revised manuscript has been substantially strengthened and addresses the major concerns raised during the previous round of review. The authors have appropriately tempered their claims, consistently framed the adaptive analyses as pseudo-prospective, improved the statistical reporting (including confidence intervals, effect sizes, and multiple-testing correction), and significantly expanded the discussion of governance, safety, and ethical oversight. The manuscript now provides a more balanced interpretation of the findings and a clearer acknowledgement of its limitations, including the proof-of-concept nature of the study, the small sample size, and the constraints of the inpatient EMU setting.

The work presents a genuinely novel platform for bidirectional patient–AI interaction that has the potential to influence the future design of intelligent neurotechnologies. The technical implementation is rigorous, the integration of patient- and clinician-in-the-loop model adaptation is compelling, and the manuscript is transparent about the current limitations and the need for future outpatient validation.

I have only a few minor suggestions. First, some of the more visionary statements in the Discussion and Conclusion (e.g. "With the current pace of innovation in computing, deep learning, and 648 neuroscience, it is only a matter of time before these kinds of interactions become

649 commonplace") could be further moderated to better reflect the current level of evidence. Second, while the governance framework is now well described, the Discussion could briefly acknowledge broader human factors relevant to long-term deployment, such as automation bias, trust calibration, clinician responsibility, and the implications of relying on evolving foundation models in regulated clinical environments.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Editor's Comments:

Your Article, "An Implantable Device that Converses with Patients and Learns to Co-Manage Epilepsy" has now been seen by 4 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Medicine but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

Please note that we find the approach to be very exciting and are OK with this being a proof-of-concept study. However, we do ask that the authors significantly tone down overstatements, caveat findings and explicitly address the potential safety/ethical considerations that may be associated with this approach (Reviewer #4 highlights some of these points). Please note that we will likely recruit an ethics referee in the next round of review, so please provide sufficient details here. As for other points, please could the authors provide the IRB-approved protocol with the revised manuscript and also please could the authors remove the "vignette box/story" from the paper.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

- * Please highlight or track all changes in the manuscript text file, which should be prepared in Word Docx format
- * Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.
- * If relevant, please ensure that your manuscript complies with our [clinical research reporting guidance](#).

* If you have not done so already please begin to revise your manuscript so that it conforms to our [policy and formatting instructions](#). Refer also to any guidelines provided in this letter.

* Include a revised version of any required reporting checklist. It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review. A revised checklist is essential for re-review of the paper.

Nature journals have recently announced an update to our guidance on reporting on sex and gender in research studies ([see here](#)). We strongly encourage researchers to follow the '[Sex and Gender Equity in Research – SAGER – guidelines](#)' and to include sex and gender considerations for studies involving humans, vertebrate animals and cell lines where relevant to the topic of study (an overview can be found [here](#)). Authors should use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms.

When preparing your revised manuscript, please be aware of our guidance on [Sex and Gender reporting](#)). Please note that:

1. If the research findings apply to only one sex or gender, that must be indicated in the title and/or abstract.

- 2a. For studies involving vertebrates animal and cell lines- The Reporting Summary should include whether sex was were considered in the study design.

- 2b. For studies involving human research participants- The Reporting Summary should include whether sex and/or gender was considered in the study design and whether sex and/or gender of participants was determined based on self-report or assigned (and methodology used).

3. Data should be reported disaggregated for sex and gender where this information has been collected and consent has been obtained for reporting and sharing individual-level data; disaggregated numbers for individual experiments must be provided in the source data as appropriate whereas overall numbers may be provided in the Nature Portfolio Reporting Summary.

Information on the 3 points above should be included in the revised manuscript and detailed in the cover letter.

In addition, please note that if sex- and gender-based analyses have been performed a priori, results should be reported regardless of positive or negative outcome. We discourage conducting post hoc sex- and gender-based analysis if the study design is insufficient (for example, low sample size) to enable meaningful conclusions.

If no sex- and gender-based analyses have been performed, please indicate the reasons for the lack of these analyses in the Reporting Summary.

When submitting the revised version of your manuscript, please also pay close attention to our [Digital Image Integrity Guidelines](#) and to the following points below:

- that unprocessed scans are clearly labelled and match the gels and western blots presented in figures.
- that control panels for gels and western blots are appropriately described as loading on sample processing controls
- all images in the paper are checked for duplication of panels and for splicing of gel lanes.

Finally, please ensure that you retain unprocessed data and metadata files after publication, ideally archiving data in perpetuity, as these may be requested during the peer review and production process or after publication if any issues arise.

Response:

We are grateful for the editor's enthusiasm and for the opportunity to revise. Below we address each editorial request; all corresponding changes are tracked in the revised manuscript.

Toning down overstatements and caveating findings. We have substantially softened overstatements throughout, in line with the reviewers' concerns. Key changes include: reframing the seizure-detection results around a reduction in false alarms with preserved sensitivity rather than an "improvement in overall seizure detection" (Abstract, Results, Discussion); consistently describing the fine-tuning analysis as a pseudo-prospective simulation rather than live, self-retraining deployment (Abstract, Results, Discussion); changing "excellent" usability to "highly usable" in the Abstract and qualifying the System Usability Scale result by its n=5 responder rate wherever it appears; and adding explicit caveats regarding

baseline detector performance, residual false-alarm burden (2–4/day), and the artificial, time-limited nature of the EMU setting (Limitations). We detail numerous other changes in our response to Reviewer #1, comment C.

Safety and ethical considerations. Anticipating review by an ethics referee, we have considerably expanded the safety and ethics disclosures. The Methods now detail the multi-layered input/output moderation architecture, pre-deployment red-teaming, daily human review of all patient–LLM exchanges, auditing of every data-agent numerical claim against the underlying SQL query, and the multi-track institutional governance (IRB, the newly established AI Governance Committee, the Information Services Architecture Review Board, and a Security Risk Assessment). The Discussion’s Ethical Considerations section addresses PHI handling within HIPAA-compliant infrastructure and the contractual prohibition on model training or data retention. See our responses to Reviewer #3 (points 4, 5, and 9) and Reviewer #4.

IRB-approved protocol. The IRB-approved protocol (University of Pennsylvania IRB #849276) is provided with this submission as requested.

Removal of the vignette box/story. Per the editor’s request, the vignette box/story has been removed from the manuscript. We have also removed the closing invocation of Asimov’s Three Laws noted by Reviewer #3.

Tracking of changes. All changes to the manuscript text are tracked, and this point-by-point document accompanies the revised manuscript.

Reporting summary and demographics reporting. A completed Nature Portfolio Reporting Summary is included with the submission. This proof-of-concept study was not designed or powered for sex- or gender-based analyses, and none were performed given the small sample (n=13). Participant sex was self-reported and is disaggregated in Table 1 (7 female, 6 male). The findings are not specific to one sex or gender, so no sex/gender qualifier is required in the title or abstract. We have added a statement to the Methods (page 22, line 12): “Sex and gender were not used as variables in the study design, and the study was not powered for sex- or gender-based analyses; participant sex was self-reported (Table 1)...” Self-reported race and ethnicity are now reported in Table 1 (race: White 11/13, Black or African American 2/13; ethnicity: Not Hispanic or Latinx 13/13). We have additionally added Code availability and Data availability sections to the manuscript end-matter.

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

A. This manuscript presents a genuinely novel and technically impressive proof-of-concept platform for bidirectional patient-device communication in epilepsy management, with clear translational vision.

Response:

[Thank you for your comments.](#)

The study's small sample (n=13), inpatient-only setting, predominantly qualitative/feasibility nature, incomplete statistical reporting, and fundamental limitations in seizure detection evaluation—including non-significant primary results in the prospective annotation arm—require substantial revision. The sample size issue is a mild weakness given the complexity of the undertaking - it is still a substantial contribution.

Response:

[We acknowledge that as a proof-of-concept study the sample size is relatively small. We are planning a second phase of data collection with an improved patient interface and additional features.](#)

The clinical problem is compelling and well-articulated: the authors correctly identify that current FDA-approved epilepsy devices (RNS, VNS, DBS) have no real-time bidirectional patient communication, and the manuscript convincingly frames this as a missed opportunity.

Response:

[Thank you.](#)

The seizure detection models employed (SPaRCNet, ONCET, WaveNet) are explicitly acknowledged as 'not reflecting state-of-the-art performance for detecting seizures.' The baseline false alarm rate of 0.58/hour (IQR 0.35-0.77) while clinically unacceptable for an alert-based system—this translates to approximately 14 false alarms per day, is a great start to what I think is the future in our field.

Response:

We agree that the false alarm rate in the current study is not yet clinically acceptable. We now discuss this point on page 19, line 13. "However, the false alarm rate achieved after fine-tuning (IQR 0.081-0.168) corresponds to 2-4 false alarms per day and may still be burdensome for some patients, necessitating further algorithm improvement before real-world application. In the current implementation, notifications were limited to high confidence events with high seizure probability or longer duration. A similar approach could be extended to allow patients to customize notification configurations, suppressing alerts for brief or less severe events." This likely reflects the use of relatively simple baseline models, and we expect that performance can be easily improved with more advanced approaches (page 19, line 40). "Adopting a better-performing baseline model would almost certainly improve overall performance and is straightforward to implement given the modular design of the platform."

B. The Human in the Loop concept of course is not novel but the context here is relatively original.

Response:

Thank you.

C. The seizure annotation feature was only implemented for 9 of 13 patients, and only 6 had confirmed events suitable for fine-tuning. All statistical tests for the primary innovation rest on $n=6$, which is insufficient to support the broad claims made in the abstract, introduction, and conclusion. This sample size provides inadequate power to detect clinically meaningful differences and makes the Wilcoxon signed-rank test results uninterpretable for inferential purposes. The manuscript describes multiple outcomes (sensitivity, F1, false alarms/hour, AUROC, AUPRC), multiple fine-tuning conditions (push-button, app annotations, clinician), and multiple evaluation time windows without pre-specification. This multiplicity, combined with one-sided hypothesis tests selected post-hoc in the direction of expected improvement, creates substantial risk of Type I error.

Response:

Thank you for your comments. We acknowledge that our sample size was relatively small, particularly for the fine-tuning experiment with patient annotations. To better quantify uncertainty despite the limited sample size, we now report rank biserial effect sizes (r) and 95% confidence intervals for key paired comparisons evaluated using Wilcoxon signed-rank tests and corrected multiple comparisons using the Benjamini–Hochberg (BH) procedure. These are now included on page 31, line 30 of the manuscript. "Rank biserial effect sizes (r) and 95% confidence intervals (CI) were reported for all Mann-Whitney U and Wilcoxon tests. Multiple comparisons were

corrected using the Benjamini–Hochberg (BH) procedure, with adjusted p-values < 0.05 considered statistically significant.”

We would also like to clarify that while we reported multiple metrics, statistical tests were only performed for F1 score and false alarm rates under 3 settings. The scores from multiple evaluation windows were aggregated into a single median value per patient for statistical analysis. This is now clarified on page 31, line 23. "The median performance difference across all fine-tuned models was aggregated for each patient. These medians were compared to zero using a one-sided Wilcoxon signed-rank test, with the alternative hypothesis set according to the expected direction of effect. "

"Statistical tests were limited to false alarm rate and F1 score to reduce multiple comparisons. "

We have also revised our statements in abstract, introduction, and conclusion to ensure we accurately represent the work done in this study and removed any potential embellishments. Specifically, we made the following changes: "to teach each other" → "to exchange information" (Abstract), "automatically retrains itself to improve performance" → "uses their feedback to retrain its seizure detector" (Abstract), "learns to co-manage epilepsy" → "learns to help patients manage their epilepsy" (Introduction), "devices to co-manage disease" → "devices to engage with their condition" (Discussion), "useful in better understanding and managing their condition" → dropped "and managing" (Discussion), "demonstrating" → "suggesting" (more appropriate for n=5, Discussion), "resulted in significant improvement in overall seizure detection performance" → "significantly improved F1 score, while preserving detection sensitivity" (Discussion). Additionally, several new caveats were added to the Limitations section.

The entire conversational and reasoning architecture depends on a HIPAA-secure instance of GPT-4o-mini, a proprietary model subject to unilateral updates by OpenAI. The manuscript does not describe model versioning, does not address what happens to validated system behavior when the underlying model is updated, and does not discuss the regulatory pathway for a medical device whose core decision-support layer is controlled by a third party. This is a fundamental reproducibility and translational concern.

Response:

Regarding the GPT-4o-mini dependency, we now detail the exact pinned model version with additional information about the implementation, *“We chose GPT-4o-mini (version: 2024-08-01-preview) as the underlying LLM because...”* Methods (page 23, line 37), *“Conversation context was managed by...”* Methods (page 27, line 18), *“To mitigate hallucination and unsafe content in a clinical setting...”* Methods (page 27, line 33). This pinned model version was used for all study participants, alleviating the concern that updates to the underlying model may have affected system behavior. We also added more information to the Study Design section about what was required for this study to pass institutional regulations around clinical AI, *“The AI Governance Committee’s review focused on...”* Methods (page 22, line 23).

The authors also do not explicitly state whether missing survey data was excluded (listwise deletion), imputed, or analyzed as-is. We need to discuss potential bias from differential response rates by time-of-day and provide sensitivity analyses showing impact of missing data (e.g., assuming worst-case non-responders as negative cases).

Response:

We acknowledge that the participant count in this study was too small to draw any conclusions from survey answers, so we only reported engagement metrics in this manuscript. The purpose was to demonstrate the feasibility of sending questionnaires to patients in this conversational context and linking it to neurophysiology. We further clarified this point in the Limitations section, *“The in-app surveys themselves were designed for low-burden, high-frequency ecological momentary assessment (EMA) during the EMU stay; they are not validated epilepsy-specific patient-reported outcome instruments, and we report them as feasibility and engagement signals rather than clinical endpoints...”* (page 18, line 25), and also now show daily survey response rates for all participants in Extended Data Fig. 3. Further details are included in our response to D/E (below).

A small point - the study design is described as 'prospective observational' but the seizure fine-tuning evaluation was performed as a 'pseudo-prospective' post-hoc analysis for safety reasons. This distinction—that the fine-tuned models were never actually deployed in real-time during patient sessions—is critical to interpreting the results and is currently buried in the Methods section.

Response:

We have now clarified that our fine-tuning was implemented in a pseudo-prospective manner in the Abstract *“In a pseudo-prospective simulation, algorithms for detecting seizures optimized their precision ...”* (page 2, line 18), Results *“We collected real-time patient feedback on seizure detections as weak labels and simulated prospective rolling*

fine-tuning to improve seizure detection without requiring clinician input.” (page 10, line 38), and Discussion *“In a pseudo-prospective simulation, the system re-trained ...”* (page 15 line 10).

D/E. The fine-tuned seizure detection models were never actually deployed during patient sessions - the evaluation was performed post-hoc as a 'pseudo-prospective' simulation for safety reasons. This is a material distinction from true prospective performance, but the abstract and Results describe performance as if the system was evaluated in real-time deployment. Readers will be misled about the actual clinical readiness of the system.

Response:

We have now clarified that our fine-tuning was implemented in a pseudo-prospective manner in the Abstract *“In a pseudo-prospective simulation, algorithms for detecting seizures optimized their precision ...”* (page 2, line 18), Results *“We collected real-time patient feedback on seizure detections as weak labels and simulated prospective rolling fine-tuning to improve seizure detection without requiring clinician input.”* (page 10, line 38), and Discussion *“In a pseudo-prospective simulation, the system re-trained ...”* (page 15 line 10).

The System Usability Scale was completed by only 5 of 13 patients (38%), yet the mean score of 83/100 is used to characterize the system as 'excellent' in the abstract. The unvalidated 10-item visual analog scale surveys for mood, anxiety, and cognition are presented as clinically meaningful patient-reported outcomes without comparison to validated epilepsy-specific instruments.

Response:

We modified the System Usability Scale language in the abstract to *“highly usable”* and throughout the manuscript when we mention the “excellent” usability range, it is qualified by the n=5 responder rate.

Regarding the visual-analog-scale survey items: we appreciate the reviewer's concern but want to clarify that the manuscript does not present these surveys as validated patient-reported outcomes (PROs) or as clinically meaningful endpoints. In the Results, patient responses to these items are reported only as feasibility and engagement signals (response rate, response latency, and qualitative direction of responses) rather than as inferential measures of mood, anxiety, or cognition; we draw no clinical conclusions from the rating values themselves. The items are described in Methods as in-app surveys, and we do not use the term "patient-reported outcomes" anywhere in the manuscript. To remove any remaining ambiguity, we have 1) added a Methods

sentence (page 28, line 22), “Survey items are brief, author-constructed visual-analog-scale (VAS) ratings designed for high-frequency in-app administration; they are not validated epilepsy-specific patient-reported outcome instruments and are reported here only as feasibility and engagement signals (response rates and directionality).” and 2) added a Limitations sentence, Discussion (page 18, line 25), “The in-app surveys themselves were designed for low-burden, high-frequency ecological momentary assessment (EMA) during the EMU stay; they are not validated epilepsy-specific patient-reported outcome instruments, and we report them as feasibility and engagement signals rather than as clinical endpoints, with future studies needed to benchmark these brief in-app ratings against validated tools such as QOLIE-31, NDDI-E, and PROMIS measures.”

Also, the push-button alarm validation (n=11) is presented as a stronger result but conflates two different processes: push-button alarms reflect patient perception of seizure occurrence (with known variability in postictal memory), not ground-truth EEG-confirmed events. Using these as training labels without reporting agreement with expert EEG annotation introduces label noise that is not quantified.

Response:

We agree that patient push button alarms may be inaccurate and we have reported their agreement with clinicians' annotations on page 11, line 5 “These push button events showed an overall sensitivity of 0.66 (41/62) and a precision of 0.60 (50/84) (Extended Data Figure 2).”

We would also like to clarify that we intentionally used patient-provided reports or annotations to demonstrate that even imperfect weak labels can achieve performance improvements during model fine-tuning, without requiring clinician input. This is now discussed on page 15, line 37. “Although annotation inaccuracies may arise from missed or delayed responses, postictal amnesia preventing patient input, or subclinical seizures that go unrecognized, these imperfect labels nonetheless proved effective for fine-tuning, consistent with prior studies demonstrating the utility of weak labels in model training.”

Also, response latency is reported using median and IQR with a Mann-Whitney U test ($p=9.89 \times 10^{-21}$), and effect size is mentioned as 'rank-biserial correlation' but the actual effect size value is not reported in the results text.

Response:

We thank the reviewer for identifying this omission. Consistent with the effect-size reporting we have added throughout (see our response to comment C above), we now

report the rank-biserial effect size for the latency comparison in the Results text alongside the Mann-Whitney U statistic and p-value. The data agent responded substantially more slowly than general chat (median difference 6.747 seconds (IQR 2.532-5.115; 95% CI [4.990,8.651])) owing to the additional SQL-generation and database-querying steps.

F. For a prospective study - there is very little of the a-priori thinking described. Add sample size justification in addition to above. Add 95% CIs around median effect etc.

Response:

We appreciate this suggestion and have added a sample size justification in the Methods under Study Design (page 22, line 9). "Given the proof-of-concept nature of the platform and the goal of demonstrating per-patient feasibility rather than a population-level effect, we pre-specified a target of 10–15 EMU patients based on feasibility and patient availability during the study period, rather than by a formal power calculation."

We now added 95% confidence intervals as well as rank biserial effect sizes (r) for key paired comparisons evaluated using Wilcoxon signed-rank tests (page 31, line 29). "Rank biserial effect sizes (r) and 95% confidence intervals (CI) were reported for all Mann-Whitney U and Wilcoxon tests."

G. References appropriate

Response:

Thank you.

H. No concerns with clarity and context

Response:

Thank you.

Reviewer #2 (Remarks to the Author):

Goldblum et al. present a very interesting concept of combining an LLM with ongoing EEG as a proof-of-concept for implantable conversational devices. They show that patient responses (annotations, patient-in-the-loop) can be used to improve the performance of seizure detection models, with further improvements by targeted clinician annotations (clinician-in-the-loop). This is very innovative research and nudges the field forward. I do not have any major concerns about this work.

Response:

Thank you.

A. Summary of the key results:

- Abstract: “An AI agent ingests ... seizure risk extracted from brain activity”. Were we calculating seizure risk?

Response:

Thank you for your question. We computed seizure probability, not seizure-risk forecasts. We revised this sentence to use the more accurate term "seizure probability".

B. Originality and significance:

- Novel.

Response:

Thank you.

C. Data & methodology: validity of approach, quality of data, quality of presentation:

- Only 15.5% seizures detected are true seizures, which is very low and can lead to alert fatigue among patients with a high false positivity. Was that noted/measured?

Response:

This is a valid concern. Alert fatigue was not something we explicitly measured in our participants. During the consent process, it was explained that this system is a research prototype and that erroneous seizure detections were to be expected – also that they should refer to their clinical care team for true seizure detections. So participants had the expectation that they may receive frequent seizure alerts, many of which were likely to be false positives. While we cannot make claims about the extent of alarm fatigue, and do not do so in the manuscript, we did not receive complaints from participants in this regard. We added additional detail to the “Patient recruitment” section (page 5, line 10): “They were informed that this system is a research prototype in which false positive seizure detections were anticipated, and that all clinical seizure concerns should be directed to their care team.”

- “and labeled 17.2% (n=10) as uncertain. All uncertain events (n=9)”. Is uncertain 9 or 10?

Response:

Thank you for pointing this out. The inconsistent uncertain event count was due to ID mismatches, which erroneously rejected annotations for two earlier participants. We have carefully examined the collected patient annotations and confirmed the event count for each category. We now use $n=8$ consistently in the Patient engagement section and have updated the associated contents. We additionally rejected 6 annotations for the fine-tuning analysis. This is detailed in Results (page 9, line 35), *“Six annotations from 3 patients (2 “yes,” 3 “no,” and 1 “uncertain”) were deemed invalid because of timestamp mismatches caused by technical issues and were excluded from downstream analyses, resulting in a final set of 7 true seizures, 20 false positives, and 7 uncertain annotations.”* and page 11, line 22, *“The seizure annotation feature was not implemented for the first 2 participants, with 6 of 11 implemented patients having confirmed events (true seizures or false positives). The raw seizure alerts showed a sensitivity of 0.136 (8/59) and a precision of 0.138 (8/58). Although alert sensitivity was limited by our choice of a simple, easily deployable detection model, patient-confirmed annotations demonstrated relatively high accuracy, showing a sensitivity of 1.0 (5/5), a precision of 0.71 (5/7), and a negative predictive value (NPV) of 1.0 (20/20) among annotated events ($n=39$), leading to an overall sensitivity of 0.085 (5/59). All uncertain events ($n=7$) were marked as non-seizure by experts.”*

- S4. Why are the stress/fatigue questions being repeated twice? Both questions are asking the same thing (how stressed are you/how calm are you, how fatigued are you/how much energy do you have). Similarly in S3 (how anxious are you/how tense are you)

Response:

The two items per construct were intentionally phrased with positive- and negative-valenced framings to allow internal-consistency checks. However, as you pointed out, we missed this positive/negative valence pairing for the anxiety construct, “Right now, how tense do you feel?” should have been “Right now, how relaxed do you feel?” We will incorporate this change for future participants.

D. Appropriate use of statistics and treatment of uncertainties:

- No concerns.

Response:

Thank you.

E. Conclusions: robustness, validity, reliability.

- No concerns.

Response:

Thank you.

F. Suggested improvements: experiments, data for possible revision.

- Fig 2: “Continuous EEG signals are acquired via clinical workstations and relayed through a cloud-hosted virtual machine to Azure Event Hubs for ingestion” What is the data bandwidth needed- is that feasible at a population level?

Response:

Thank you for raising this point. In our study, raw EEG data could be sampled at up to 1024 Hz and were automatically downsampled to 512 Hz when transmitted to the virtual machine via the Natus-provided SDK. From the virtual machine, data was transmitted once per second to Azure Event Hub, which imposes an ingress limit of 1 MB/s or 1000 events/s per throughput unit. Our current configuration provides 4 throughput units, with automatic scaling up to 12 units during periods of high input volume. To address data size constraints, we applied gzip compression and base64 encoding. Under this configuration, a 1-second scalp EEG packet is approximately 0.03–0.04 MB, whereas a 250-channel SEEG packet is approximately 0.35 MB. Based on these estimates, the current system can support at least 10 SEEG patients or 100 scalp EEG patients concurrently under stable operation. The framework can be easily scaled up through additional throughput units or higher service tiers supporting up to 20 MB/s ingress. Additionally, ambulatory EEG systems typically use fewer electrodes than EMU recordings and therefore generate substantially lower data volumes per second. This information is now included in the Methods (page 24, line 7) and Discussion (page 17, line 16).

G. References: appropriate credit to previous work?

- Yes.

Response:

Thank you.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions.

- No concerns

Response:

Thank you.

Reviewer #3 (Remarks to the Author):

This is a case series of a concept – that patients might be able to assist in training closed-loop seizure-suppression systems, by conversation with an LLM that in turn has control over device parameters and ability to interpret intracranial physiology. A core thesis is that “weak labels”, i.e. potentially-inaccurate patient recall/labeling of seizure events (prompted by a conversation agent or, here, mostly by a push button) could still be quite useful. That thesis is reasonably substantiated and important across clinical neuromodulation. The overall study was technically complex due to the number of data handling steps and systems being interfaced, and presents a really interesting proof of concept towards a technology that could have broad applications in enabling better/more rapid/more scalable development of closed-loop neurotechnology. It genuinely is a conversation-changing article, and in that context, Nature Medicine likely will be a reasonable venue (modulo comments below).

Other strengths of the report include:

- A reasonably well justified overall design, including some thoughts about layering LLMs to ensure accurate message handling and safety
 - Specific considerations and discussion of ethical approvals for an AI-heavy study
 - Discussion and demonstration of an ML pipeline that is oriented around “not making matters worse”, e.g. freezing key layers to prevent catastrophic forgetting
- That said, the report as presented is substantially incomplete. There are major outcome details that are missing and need to be reported, as well as some methods details. These largely relate to not actually showing us how patient and LLM interacted. The framing is also deceptive. It is presented in the Abstract as an n=13 study of the approach, but based on the Methods, it appears that it may be as little as n=2 actually demonstrating the approach being claimed. Further, again based on the Methods, it appears that the core claim (improved seizure detection performance through real-time, LLM-driven event marking and tuning) is not actually done. First, improved seizure detection is not actually shown, and in the Discussion they explain why they probably cannot achieve it in this kind of study. Second, if I am reading this right, they never actually did real time interactive fine-tuning with the patient, and everything being plotted, especially in Figure 5, is all theoretical simulations of what might have happened.

This also highlights that there are many places where the text and figures are extremely unclear and I cannot be certain what was actually done or what is being plotted. This is not a complete scientific report.

These are major weaknesses. They are not fatal because the value of the report is more in demonstrating the technological proof of concept and starting a conversation, not in the specific clinical outcomes or the N (an N=2 pilot has value if the right data are

shown). They still need to be corrected before this would be publishable, again because some of the claims are frankly false/deceptive as written. The article does not NEED those false claims in order to be interesting and high impact, and thus they need to come out.

Drilling into specifics,

1. It is very hard to identify exactly how many patients used the LLM system shown in Figure 2, or for what purpose. It is, however, not the $n=13$ that is claimed. In Figure 5, only 6 patients are described as having used the app for labeling. In the Methods, it is then said (p12-13) that seizure annotation was not actually implemented for 4 patients – which could mean there were $n=9$ (if this means push button) or possibly only $n=2$. What is needed is a table that shows, for each reported patient, what aspects of the study they actually participated in (using the LLM for surveys, to query data, using the push button, having fine-tuning, etc etc etc.). Further, the early claims in Abstract and Introduction need to specifically say how many patients actually used this full claimed system.

Response:

Thank you for your suggestions. We have now detailed the functions (surveys, seizure annotations, and queries) implemented for each patient in Table S1 (and a per-patient, per-day breakdown in Extended Data Fig. 3).

The sample size for patient-annotation fine-tuning was reported incorrectly in the original manuscript because of an ID mismatch between annotations and seizure events. To clarify, the seizure-annotation workflow was not implemented for the first two patients, leaving $n = 11$ patients eligible for this analysis. Of these, five did not provide usable seizure annotations, so the final sample size for seizure fine-tuning with system-derived annotations was $n = 6$. This is clarified on page 11, line 22, “The seizure annotation feature was not implemented for the first 2 participants, with 6 of 11 implemented patients having confirmed events (true seizures or false positives).” To mitigate this limited sample size, we additionally used push-button alarms, which were available for all enrolled patients.

We have now explicitly reported the sample size for each analysis in the Abstract (page 2, line 16) “We demonstrate this platform in 13 patients undergoing inpatient video-EEG monitoring for epilepsy (scalp $n=11$; intracranial $n=2$) and validate its performance. In a pseudo-prospective simulation, algorithms for detecting seizures optimized their precision using either patient self-reports ($n=13$) or annotations through the platform ($n=6$) without expert intervention ...” and Introduction (page 4, line 30) “In the following sections, we describe a platform that performs these functions, and report our results from testing it prospectively on 13 patients (annotation $n=11$; data query $n=8$) admitted

to the Hospital of the University of Pennsylvania epilepsy monitoring unit (EMU) for continuous video and scalp (n=11) or iEEG (n=2) monitoring.

2. It appears that there was no actual adaptation of the model. If I read this right, what they actually did was have the patients interact with an untuned model for the full stay, but simulated what WOULD have happened if it were fine-tuned using their responses. (Methods p19). This is another place where the framing is deceptive, because this is very subtly buried after the claims of the article regarding fine tuning. They need to be up front and refer to this process as something like “simulated fine tuning” in the article body, abstract, etc.

Response:

Your interpretation was correct. We have now clarified this point in the Abstract (page 2, line 18). “In a pseudo-prospective simulation, algorithms for detecting seizures optimized their precision ...”, Results (page 10 line 38), “We collected real-time patient feedback on seizure detections as weak labels and simulated prospective rolling fine-tuning to improve seizure detection without requiring clinician input.” and Discussion (page 15, line 10) “In a pseudo-prospective simulation, the system learned ...”

3. The main effect actually demonstrated is a reduction in false alarms. This is in fact useful, but the claim “the system learned and re-trained itself to markedly improve seizure detection performance” is false both because of the above (the system did not re-train itself, there are several ways in which that statement is not true) and because there was not actually an improvement in seizure detection (sensitivity and/or PPV).

Response:

We have revised the statement to: “the system learned and re-trained from imperfect patient annotations alone, markedly reducing false alarm rates while preserving baseline sensitivity ...”

4. There is little to no data presented to explain how patients engaged with the LLM/chat interface and what the interface actually did well/poorly. Specific items of concern:

a. What types of data did patients request? How often? How many times did they need to re-prompt the LLM because it did not respond accurately? Did use of the LLM remain constant over the EMU stay, or increase/decrease? A pattern of diminishing engagement over the 4-5 days in the EMU would tell a very different story from increasing engagement.

b. Why was the patient response rate so low? What feedback did they collect that can explain this? I can somewhat understand it for surveys, and they discuss it for spikes,

but I am very surprised that it is only 70% for seizure prompts. Further, how did it change over time? See above.

c. What were the inaccuracies in chat/data message routing claimed in Figure 3e?

d. What were the messages that were flagged by the LLM safety layer as unsafe, and why? If there are really only 6, then provide the text of the interactions.

e. Related, it is claimed that no messages from the LLM to the patient were marked as unsafe. If this is true, how do they know that the safety filter was effective? What messages WOULD have been flagged as unsafe?

f. How did they test the patient-to-AI or AI-to-patient safety layers before deploying them? (Possibly they did not and relied on human review of the message chains, but again, then how do they know whether the AI safety review was valid if it never flagged a message?)

Response:

a. To further clarify the patient-AI engagement metrics presented in Figure 2, we added message counts to the textual description of Figure 3a in the Results (page 9, line 17), *“Patients also utilized it for clinical data retrieval, directing 11.5% of their queries toward insights about seizure events (5.9%, n=24), sleep metrics (3.7%, n=15), and interictal spike rates (1.9%, n=8).”* We also added a supplemental figure (Extended Data Fig. 3) that shows the daily engagement with the application across all participants. This figure details 1) how often participants sent messages to the LLM, 2) how often they asked questions about their seizures, sleep, and spikes, 3) how often they responded to seizure and spike events, and 4) how often they completed morning, afternoon, and evening surveys. This shows how engagement evolved over the course of enrollment in the study. Furthermore, seizure event annotation performance is broken down at a per-participant level.

b. Thank you for this comment. We address this comment in two areas: 1) In the Discussion, Ethical Considerations section (page 20, line 14), we explain that, *“Patients were instructed to press the seizure alarm button for any suspected or impending event, and to interact with the AI only afterward. This protocol is reflected in our results: all patients self-reported seizures through the clinical system, whereas fewer did so through our platform. These two sources of events were analyzed and reported separately. Reduced engagement with the platform may also reflect the realities of the inpatient setting, where patients are frequently enrolled in other research studies, receiving concurrent clinical care, recovering in post-ictal states, or attending to visiting family—all of which compete for their attention and limit interaction.”* 2) In Extended Data Fig. 3, the plot shows relatively consistently within-patient daily engagement, with larger engagement variations between patients.

c. The erroneously routed messages are now detailed in the Supplemental Materials: Table S9.

d. The six messages flagged for moderation are now detailed in the Supplemental Materials: Table S8.

e. This is an important point since a moderation node that never fires is indistinguishable from one that is broken. We validated the LLM moderation system in simulated environments prior to study initiation and through continuous, daily review once the study was in progress. We now describe our approach in the "Conversational agent" section of the Methods (page 28, line 1), "Pre-deployment red-teaming was conducted iteratively as part of agent system development. Investigators interacted with the chat interface and attempted to elicit unsafe or off-protocol behavior..." This section also specifies the types of messages and responses that would elicit moderation, and for full transparency, the system prompts for the moderation nodes (and all other nodes) are detailed in the Supplemental Materials.

f. We have expanded the "Conversational agent" section of the Methods (page 28, line 1) with details of how both safety layers were tested before and during deployment, "Pre-deployment red-teaming was conducted iteratively as part of agent system development. Investigators interacted with the chat interface and attempted to elicit unsafe or off-protocol behavior, including: prompts intended to alter the agent's directive or persona, prompt-injection attempts, requests to discuss off-topic subjects, requests to produce harmful or hurtful content, and attempts to coerce the data agent into returning information outside its intended scope..."

5. There is similarly minimal information on the specific design of this LLM architecture, including

a. What was done to prevent the LLMs from having context buffer leaks? How was reliable, sustained, and consistent interaction with the patient maintained?

b. How was each LLM layer prompted? Were any of these models specifically trained, or were they "out of the box" GPT implementations? In my opinion, the prompts themselves need to be visible to readers/reviewers. I know that I can see some of this in graph.py and did so, but not every reader wants to dig four folders deep in GitHub.

c. Why GPT? Possibly this is due to licensing/convenience factors, but then say so, and comment on the specific properties of the 4o model that may be better/worse for this application.

Response:

We have expanded the manuscript's description of the LLM architecture. The full set of node-level system prompts is now provided in the Supplementary Materials so that reviewers and readers need not consult the source repository.

a. The "Conversational agent" subsection of the Methods (page 27, line 18) now explains states that each patient was provisioned a separate application account and conversation thread, with database access constrained server-side to the patient's unique identifier, such that no patient's history or data could enter another's session. We also clarify that, because the LLM is stateless between calls, each node's system prompt was re-injected on every turn rather than relying on model memory, keeping agent behavior consistent even as older messages aged out of the fixed 200-message sliding window. This window also bounded the input size to prevent token-limit overflow that could otherwise truncate context over a long conversation.

b. As now stated in the Methods (Overview, page 23, line 37), we did not train or fine-tune the LLM and did not employ rolling LLM-generated summarization; each node used a dedicated, fixed system prompt around an off-the-shelf GPT-4o-mini.

c. The section referenced in our response to B (above) also includes information that addresses this question. We selected GPT-4o-mini because it was the only LLM available at the time that satisfied all of the following criteria: HIPAA-compliance, native tool-calling, low latency, and low per-token cost for continuous use. While its smaller capacity makes it less suited to open-ended clinical reasoning; the agent's role was deliberately scoped to information delivery and structured retrieval rather than autonomous clinical judgment, and the modular architecture allows the model to be substituted as more capable HIPAA-compliant options emerge. Swapping in a state-of-the-art LLM would likely improve performance in every capacity.

6. The LLM annotations actually appear to perform worse than the push button (median F1 and false alarm rate appear to be worse in Figure 5c than 5b). This needs to be explicitly acknowledged and discussed, because although the "Weak labels work" claim is valid, the fact that they work less well than spontaneous patient self-report suggests a problem that needs to be remedied.

a. As a side note, it is not clear why there are only 4 patients in the Sensitivity plot for 5c.

Response:

We now report the fine-tuned performance using push button alarms and that using LLM annotations on page 11, line 39 "*The fine-tuning performance using annotations through the system were comparable with that using push button alarms on*

corresponding patients, which were 0.107 (IQR 0.082-0.147) for false alarms and 0.130 (IQR 0.025-0.258) for F1 score. The median differences were 0.02 (IQR -0.027 to 0.058, 95%CI [-0.037,0.061]) for false alarms and -0.041 (IQR -0.076 to -0.010, 95%CI [-0.123,-0.006]) for F1 score. We failed to detect any statistically meaningful difference between the two groups (false alarms: BH-adjusted $p=0.688$, $r=0.238$; F1 score: BH-adjusted $p=0.088$, $r=-0.867$). Notably, the performance using push button alarms were slightly poorer compared to that using patient annotations (FA: 0.107 vs 0.104; F1: 0.130 vs 0.143). We additionally discussed this point in Discussion on page 15, line 42 “Notably, re-training performance using system annotations was comparable, or slightly better than, re-training using patient self-reports. Given the low sensitivity of system annotations, this potentially suggests that only rejection of false alarms can significantly improve model performance.”

To clarify, both patient self-reports (push-button events) and system-based annotations were considered as weak labels. We now report the accuracy and precision of both sources in Figure Sx.

We have updated the results in Figure 5c due to ID mismatch issues (see point 13). Now the number of values in 5c should be 5, 6, and 6. The sensitivity values for two patients overlapped, leaving only four distinct data points visible.

7. Figure 5 seems to be claiming some kind of test set performance, but it is very unclear exactly what is training, validation, and test set in each of these figures. Different squares in 5A are highlighted pink, but we are never told what this pink means. Similarly, what exactly the test set is for the figures is not clear, for instance, it would appear that for the blue-pink curves in 5B-C, a model fine-tuned on all of the historical data is being re-run on all of that data minus the first block. That would be classic data leakage. Possibly that is not what was being done, but much clearer descriptions and labeling are needed to evaluate whether this is technically valid.

Response:

Our fine-tuning schema was as follows: when fine-tuning was triggered, the corresponding 3-hour block was used as fine-tuning data (yellow squares), and the fine-tuned model was evaluated on subsequent blocks (green squares). We then simulated prospective deployment of the fine-tuned models and evaluated their effects on performance across the entire EMU stay. Each fine-tuned model (e.g., Fine-tune 1 in Fig. 5a) was applied only to subsequent data, excluding the training data, as indicated by the pink squares in the corresponding row. When a newer fine-tuned model became available (e.g., Fine-tune 2 in Fig. 5a), it became the active model and was similarly applied only to future data, as indicated by the pink squares in its row. The combined performance of the most recently available fine-tuned models across the stay was then

compared with that of the baseline model (blue/pink circles). We have clarified the meaning of the pink and blue squares in the figure caption and revised the figure to improve clarity. “Left: The baseline model was fine-tuned sequentially every three hours when predefined criteria were met (left). Yellow squares indicated time blocks used for fine-tuning. Middle: We then evaluated the performance change from baseline for each fine-tuned model on its corresponding test data (future timeblocks labeled as green squares) and took the median across models. Right: We also compared the performance across the full EMU stay simulating prospective fine-tuning, where each model was deployed on future data (pink squares) until the next fine-tuned model became available, against baseline (blue squares). Blue and pink squares indicate the time blocks during which the baseline or fine-tuned models were active.”

8. Given that only 2 patients are SEEG/iEEG, we need to know if the performance is different or the acceptability is different in patients undergoing those procedures, which can be substantially more burdensome than scalp EEG. Throughout the paper, those patients need to be visually distinguishable in figures, e.g. through symbols or coloring. This is critical because the use case, right at the start of the paper, is basically for LFP/iEEG signals from intracranial recordings, and we need to know how the acceptability and interaction went for those patients.

Response:

Thank you for your suggestion. We now highlight SEEG patient performance using red dots and lines in Figure 5b–d (fine-tuning performance) and Extended Data Fig. 2 (baseline performance). We also provide per-patient details in Extended Data Table 1 and Extended Data Fig. 3 to illustrate the characteristics, response rates, and annotation accuracy for SEEG patients.

9. A large amount of the study data processing is performed in various cloud infrastructures, and apparently involves passing health information and unique medical record numbers through those clouds. What was done to ensure security? How were risks audited and discussed? Were these specific risks considered by the novel AI policy committee described herein? These are important because there is currently a very fractured landscape among different institutions regarding what is/is not permitted on cloud platforms, and having a clear example here may aid other investigators. There are lesser concerns that are more related to interest and clarification.

Response:

We thank the reviewer for raising data security, and we agree that a concrete worked example is valuable given the fragmented institutional landscape. The manuscript now describes the security architecture and its governance across the Methods (Study Design, page 22, line 16) and the Discussion (Ethical Considerations, page 20, line 23).

10. Nearly 40% of seizure messages were false positives (Figure 3c). How did patients react to that? Did they find it useful to be able to flag those? If there is information in the messages or the usability free text responses, provide it. (This relates to the point above about needing to show actual patient-LLM interactions, not stylized figure versions.)

Response:

The reviewer raises an important question about the patient experience of false-positive detections. We did not collect subjective, per-event reactions to false positives by design. Once a patient labeled a detection as a true or false event, the interaction concluded, and the platform did not prompt for further feedback on that event. Only following true-positive seizure detections are patients asked to provide additional context surrounding the flagged event. We therefore cannot report directly whether patients found the flagging function useful for false-positive events. We can, however, report the behavioral evidence. Patients actively labeled detections rather than ignoring them, and the rejections shown in Fig. 3c represent this flagging behavior. We interpret this as evidence that patients engaged with the confirm/reject mechanism, though we are careful to distinguish engagement from endorsement. Something that likely kept the burden of false positives low is that patients were informed at enrollment that this is a research prototype in which false-positive detections were anticipated. We now detail that this was framed to the patient prior to the beginning of the study in the “Patient recruitment” section of the Results (page 5, line 10), “They were informed that this system is a research prototype in which false positive seizure detections were anticipated, and that all clinical seizure concerns should be directed to their care team.” We will incorporate a mechanism to better capture this kind of feedback in future work.

Response:

11. The Discussion appropriately talks about how similar systems might be useful for tuning patient-specific classifiers for other brain phenomena, e.g. motor problems in Parkinson’s or various psychiatric symptoms. This is valid, and a good extension of the “weak labels” claim. However, it also highlights the only reason this demonstration works: due to the availability of large, public, and extremely well curated data sets, there are broadly available seizure classifiers to form the foundation. If not for a decade or more of competitions and data sharing, as well as the availability of very clear ground truth, this study would be impossible (the concept of having a model where most layers could be frozen and a few fine-tuned would not be realistic). They need to be far more explicit that this was a necessary condition that does not yet exist in any other domain of neuro-medicine (one could argue about Parkinsonian tremor).

Response:

We thank the reviewer for this insightful observation, with which we agree. The reviewer is correct that our demonstration rests on a precondition that is not yet broadly available: well-validated, pretrained seizure-detection models, which themselves exist only because of large-scale, publicly shared, expertly curated EEG datasets. We have revised the "expanding platform" section of the Discussion to state this explicitly rather than leave it implicit, to make clear that this is a necessary condition that is currently met for seizure detection but not for most other domains of neuro-medicine (page 17, line 7), "We emphasize, however, that a critical enabler of the present demonstration was the availability of well-validated, pretrained seizure-detection models. Such models exist because of more than a decade of large-scale, publicly shared, expertly curated EEG datasets. This precondition is not yet met in other domains of neuro-medicine, particularly in neuropsychiatric disorders. A platform, such as the one described here, offers a route to building the curated, ground-truth-labeled datasets that these domains lack by systematically capturing patient and clinician annotations during routine use. In this sense, it is not only a beneficiary of such foundations but also a means of establishing them where they do not yet exist."

12. In the Methods, the seizure detection algorithms are described with a variety of specific parameters. Are these defaults? Chosen based on prior experience? Other? I have no quibble with the chosen parameters, but they need to explain WHY they were chosen.

Response:

Thank you for your suggestion. Most parameters, such as window size, were kept at their default values, consistent with the original publication. Others were determined based on physiological considerations, such as seizures typically lasting longer than 10 seconds, or were empirically selected using validation results from an independent EMU dataset. We have now justified the parameter choices in the Methods section on page 24, line 36, "Per model design, we applied a 10-second window with 2-second overlap on bipolar data as inputs ... The parameters were empirically selected based on validation results from an independent EMU dataset ... Consecutive detections separated by less than 2 seconds were merged, and merged events shorter than 10 seconds were discarded, as seizures typically last longer than 10 seconds ... We followed the inference pipeline described in the original publication ... Sensitivity analyses demonstrated minimal performance differences across scaler fitting durations from 1 to 20 minutes ... operates on data downsampled to 256 Hz by default ... A sample was classified as seizure if more than 10% of channels were predicted as seizure, with the threshold determined based on an independent dataset."

13. Why, on page 9, is the number of uncertain events given as both n=10 and n=9?

Response:

This was due to ID mismatches when evaluating, which erroneously rejected annotations for two earlier participants. We have carefully examined the collected patient annotations and confirmed the event count for each category. We now use n=8 consistently in the Patient engagement section and have updated all related contents. We additionally rejected 6 annotations for the fine-tuning analysis. This is detailed on page 9, line 35, *“Six annotations from 3 patients (2 “yes,” 3 “no,” and 1 “uncertain”) were deemed invalid because of timestamp mismatches caused by technical issues and were excluded from downstream analyses, resulting in a final set of 7 true seizures, 20 false positives, and 7 uncertain annotations.”*

Page 11, lines 22, *“The seizure annotation feature was not implemented for the first 2 participants, with 6 of 11 implemented patients having confirmed events (true seizures or false positives). The raw seizure alerts showed a sensitivity of 0.136 (8/59) and a precision of 0.138 (8/58). Although alert sensitivity was limited by our choice of a simple, easily deployable detection model, patient-confirmed annotations demonstrated relatively high accuracy, showing a sensitivity of 1.0 (5/5), a precision of 0.71 (5/7), and a negative predictive value (NPV) of 1.0 (20/20) among annotated events (n=39), leading to an overall sensitivity of 0.085 (5/59). All uncertain events (n=7) were marked as non-seizure by experts.”*

14. On p9, a sensitivity of 0 is reported – is this for the physiologic detection of confirmed events, or the patients’ own awareness of the events? It is very hard to understand how anything has a sensitivity of AND a precision of 1.0.

Response:

Thank you for pointing this out. The accuracy value instead of precision was incorrectly reported. We have corrected these values and other relevant contents on page 11, line 24, *“The raw seizure alerts showed a sensitivity of 0.136 (8/59) and a precision of 0.138 (8/58). Although alert sensitivity was limited by our choice of a simple, easily deployable detection model, patient-confirmed annotations demonstrated relatively high accuracy, showing a sensitivity of 1.0 (5/5), a precision of 0.71 (5/7), and a negative predictive value (NPV) of 1.0 (20/20) among annotated events (n=39), leading to an overall sensitivity of 0.085 (5/59). All uncertain events (n=7) were marked as non-seizure by experts.”*

We would like to clarify that we now report the performance at two levels, for both automated system alerts and patient feedback. During annotation, the system generated an alert when a seizure was automatically detected, with a 30-minute refractory period. Patients then labeled each alert as “seizure,” “false alarm,” or

“uncertain.” The performance of the automated alerts was evaluated relative to ground-truth seizures, whereas the performance of patient annotations was measured relative to the true labels of alerts.

15. Was the clinician-in-the-loop tuning (simulated tuning) performed on top of the simulated patient-driven fine tuning, or is this a separate study?

Response:

This was a separate study, as the current study was primarily intended as a proof of concept. In the next iteration, we plan to combine clinician-in-the-loop and patient-in-the-loop fine-tuning through model distillation or by simply stacking one model on top of the other. This has been added to the Discussion section (page 16, line 16). *"While this clinician-in-the-loop training was performed separately from patient-in-the-loop fine-tuning in this proof-of-concept study, future iterations will aim to integrate both approaches through model distillation or model stacking."*

16. On p2, design considerations are derived from “discussions with clinicians, caregivers, and patients”. Some documentation of these discussions (e.g., citations to prior work) or literature citations to other studies about lived experience of persons with epilepsy would be helpful.

Response:

We thank the reviewer for this suggestion. We have revised the passage in the Introduction (page 4, line 3) to ground the proposed design features in the published literature on the lived experience and stated priorities of people with epilepsy and their caregivers. The design rationale is now anchored in this literature together with the decades of clinical experience of our group, rather than on the informal discussions referenced in the original text. We hope this provides the documentation the reviewer requested.

Finally, as a fellow Asimov fan, I must object to the invocation of the Three Laws at the end. They do not apply here. This is not a system that can experience a threat to its own existence, nor is this type of subsumptionist architecture built into the system they describe. There certainly are questions about how to design a human-interactive system for safety, but this article needs to stick to the basic facts of what they have built and what it does/does not do.

Response:

Thank you. This section has been removed.

Reviewer #3 (Remarks on code availability):

I think it's a partial code drop. It doesn't contain the info to really replicate the system, just the chatbot. It does contain some basic install instructions that I am confident I could follow, but without available data and other components, it would be hard to exercise the full system.

Response:

Our code is organized into five components, four of which we can release publicly:

1. Edge device: real-time streaming of EEG data from the bedside to Azure Event Hubs. Portions of this codebase depend on the Natus real-time streaming SDK, which is confidential and cannot be publicly released; the device code can be shared under a usage agreement with Natus.
2. Real-time processing and feature computation (Databricks): streaming ingestion, preprocessing, and clinical / ML feature extraction implemented as Delta Live Tables pipelines.
<https://github.com/penn-cnt/Pioneer-Databricks-Release>
3. Pioneer chat interface: the patient-facing conversational application, including the LangGraph data agent, moderation pipeline, scheduled prompts and surveys, and web-push notifications.
<https://github.com/penn-cnt/Pioneer-Chat-Release>
4. Offline seizure-detection model fine-tuning: per-patient fine-tuning and evaluation pipelines for SPaRCNet, WaveNet, and ONCET seizure detection models.
<https://github.com/penn-cnt/Pioneer-Finetune-Release>
5. Analysis and manuscript figure generation: notebooks reproducing the statistics and figures reported in the main manuscript and supplement.
<https://github.com/penn-cnt/Pioneer-Analysis-Release>

All four public repositories are released under GNU GPL v3.0 and include setup instructions, dependency lists, and citation information. These repositories are now listed in a Code availability statement, accompanied by a Data availability statement, in the revised manuscript's end-matter.

Reviewer #4 (Remarks to the Author):

The manuscript by Goldblum et al. details a proof-of-concept study of a platform integrating artificial intelligence (AI) agents derived from a large language model (LLM) with seizure detection and ecological momentary assessments (EMAs). The platform interacts with the patient using natural language, leveraging the strength of the LLM and uses patient inputs from the EMAs to fine-tune seizure detection. The authors provide

different measures of patient interaction with the platform and performance improvement of seizure detection based on patient inputs. This is a novel concept and has the potential to impact how care is delivered when using device-based treatments especially in scenarios that require device calibration over longer time frames beyond just the surgical procedure. It is also a timely study given current interest in AI agents in the context of healthcare and tries to envision a scenario where these technical advances help improve patients' life in a meaningful and accessible manner.

While this idea has promise, there are limitations that dampen enthusiasm for publication in Nature Medicine. First, my interpretation of the data presented is that it is very early stage data, showing promise but not yet clinical performance that has near term translational potential. Second, the user experience (which is a critical factor for this proposed approach) is largely unstudied either before or after the current system implementation, leading to questions about the impact of the system even in an ideal implementation. Specific comments follow:

- The title is misleading, leading the reader to believe that this manuscript is mostly about implantable devices. While it is appreciated that the technology could be extended to an implantable device, the current manuscript is primarily with participants using surface EEG and only includes implantable electrodes with externalized hardware (clinical Natus systems) in 2 out of the 13 participants. On a related note, without taking anything away from the impressive engineering to make this system feasible, none of the implanted devices currently approved by the FDA for epilepsy are able to stream continuously without requiring specialized interfaces such as clinician programmer tablets. While this obviously may change in the future, this is a limitation in the present reality that severely limits the applicability of this platform in its presented form and this should be addressed. Overall, I would say that the writing in this manuscript needs to be edited to be careful not to overstate a claim on what has actually been achieved in the current project.

Response:

Thank you for your comments. We have gone through the paper in detail to ensure we accurately represent the work done in this study and removed any potential embellishments. Specifically, we made the following changes:

We have revised the title to “AI for Real-Time Patient-Brain Communication and Co-Managing Epilepsy.”

We agree that only a small number of patients with implanted electrodes were included. This was primarily due to their relative scarcity compared with scalp EEG patients, as well as the short data collection period. During the study, we encountered approximately

one intracranial patient every several weeks. We plan to continue enrolling intracranial patients in the next phase of data collection to increase the sample size.

We briefly discussed the incapacabilities of currently FDA-approved devices in data streaming in the Introduction (page 3, line 24), "While some devices can store and transmit small amounts of data asynchronously, none can share it in real-time."

Notably, some emerging reduced-montage sub-scalp devices that we mentioned (page 3, line 33), such as the Minder System (Epiminder) and the SubQ 24/7 implant (UNEED), were recently FDA-approved or CE-marked and do have continuous data streaming capacity, "New minimally-invasive EEG devices are poised to dramatically change seizure management with their ability to continuously collect data from patients during normal daily life. These devices vary from small, battery-powered single-channel recording electrodes affixed to the scalp with an adhesive, to subgaleal wires and miniaturized electronics placed under the scalp, to wearable devices. Some devices can transmit continuous EEG to smartphones and subsequently to the cloud..."

The Minder system uses the Minder Wearable to receive and store EEG recordings transmitted from the Minder Implant, then transfers this data wirelessly via low power Bluetooth (BLE) to the Minder App running on a smartphone, and finally to the cloud. The SubQ device stores data temporarily on a recorder and uploads to the cloud during charging. These devices offer a potential technical path for real-time continuous data streaming and a promising interface to our system, where we can host our system locally in the smartphone for feature calculation and model fine-tuning in the cloud. We elaborated this point in the Discussion (page 16, line 16), "This study could extend to the outpatient setting with a design that distributes processing tasks between a smartphone and the cloud. In this model, basic functions like seizure detection and warnings would be run locally by low-overhead agents that leverage dedicated neural processing units in modern smartphones. This allows for the use of quantized language and deep learning models to ensure continuity of care regardless of internet connectivity. Computationally intensive tasks, such as model re-tuning, could then be prioritized for the cloud when connection is available. This hybrid architecture is particularly practical for low-channel-count hardware, such as recently FDA-approved subgaleal EEG systems, where reduced data dimensionality simplifies on-device inference and the infrastructure supports continuous streaming from the implant to a smartphone."

- In its current form, the utility of the platform is critically dependent on the performance of the seizure detection algorithms. Given the low baseline performance of the algorithms and short time frame of the study (4 days), it is unclear if the authors' approach will lead to improvements that are meaningful for patients. For example, a false alarm rate of about 0.2/hr that they are able to achieve by fine-tuning their seizure

detection algorithm with user input, will have about 5 false alarms a day, which is still burdensome from a patient perspective. Also, the improvement seems to plateau after day 1, suggesting the models aren't adapting over time.

Response:

We agree that the baseline seizure detection performance is not yet clinically acceptable, and we now acknowledge this as a limitation (page 19, line 12), "However, the false alarm rate achieved after fine-tuning (IQR 0.081-0.168) corresponds to 2-4 false alarms per day and may still be burdensome for some patients, necessitating further algorithm improvement before real-world application. In the current implementation, notifications were limited to high confidence events with high seizure probability or longer duration. A similar approach could be extended to allow patients to customize notification configurations, suppressing alerts for brief or less severe events."

We also believe that current performance limitations could be easily mitigated by substituting more capable baseline detectors (page 19, line 36), "Adopting a better-performing baseline model would almost certainly improve overall performance and is straightforward to implement given the modular design of the platform."

We acknowledge that it is difficult to assess the long-term effectiveness of our fine-tuning framework based on the current EMU results. While performance may seem plateau in some patients, this could be influenced by the relatively stable patient state in the restricted EMU setting, the absence of seizures during a given day, and that the fine-tuning methods have not yet been fully optimized. Additionally, the false alarm rate decreased to a relatively low level (below 0.1) after the initial fine-tuning step, leaving limited room for further improvement. Consequently, the performance curve appears to plateau thereafter. Notably, the primary objective of the current study was mainly to evaluate the feasibility of using weak patient-derived annotations to improve model performance. As a future direction, we plan to evaluate the system on long-term monitoring data from other sub-scalp implant systems, such as the 24/7 EEG SubQ devices from UNEEG, and to further optimize the fine-tuning methods. However, these investigations were beyond the scope of the current proof-of-concept study. We now discuss this point on page 19, lines 40, "Additionally, fine-tuning performance over the relatively short duration of an EMU stay may not fully reflect the effects that would be observed in ambulatory settings. While performance in some patients appeared to plateau during later fine-tuning runs, this may have been influenced by the relatively stable patient state in the restricted EMU setting, the absence of seizures during a given day, and the fact that the fine-tuning methods have not yet been fully optimized. Evaluation on long-term monitoring data from existing implantable systems, such as the NeuroPace RNS System, is warranted to determine whether the platform can achieve clinically meaningful improvements in patient care. However, these investigations were

beyond the scope of the current proof-of-concept study, which was intended primarily to establish the feasibility of the platform.”

** A clarification question: It is not clear if the chat interface had option for patients to indicate that they were experiencing a seizure or if the patients had to use the push button. In case it is the latter, how is the seizure prediction performance reported affected by this ‘missing’ data?

Response:

The current chat interface does not include a feature for patients to directly report seizures. Instead, we used patient push-button events as the source of patient-initiated seizure reporting. We intentionally did not introduce an additional reporting mechanism, as patients were already instructed to use the push button in routine clinical practice. Adding another method could have caused confusion and affected clinical practice or increased reporting burden through duplicate entries. However, such a feature could be readily implemented in ambulatory settings where workflow constraints differ. The impact of missing patient-reported data is partially reflected in the difference in fine-tuning performance when using push-button labels compared with annotations collected through the system. The manuscript now explains this choice in following sections: “Patient recruitment” (page 5, line 10) and “Ethical considerations” (page 20, line 10). In future iterations, the agent could potentially replace the standard clinical seizure event button, although this would require regulatory approval.

- One potential option for the authors to address the two concerns listed above would be to use seizure prediction algorithms or features used in currently available implanted devices (e.g., NeuroPace RNS) and demonstrate this framework can improve performance and user experience.

Response:

Thank you for this suggestion. We addressed the concerns above by clarifying that (1) devices with continuous data-streaming capabilities that could potentially interface with our platform are already receiving FDA and CE mark approval; (2) the false alarm rate could potentially be reduced through the use of more advanced models, and we plan to examine the long-term effects of the platform on continuous recordings from sub-scalp devices. We appreciate your suggestion on testing our platform in algorithms or features of existing clinically deployed devices. We believe this represents an important future direction and we are planning such experiments as next steps. We believe that the ability to annotate and query data, even from devices that periodically upload data to the cloud, such as the RNS device, could improve patient care. In this proof-of-concept study, our primary objective was to demonstrate the feasibility of using weak patient-

derived annotations to improve model performance. Although the performance improvements were limited over the relatively short duration of EMU stays, a more comprehensive investigation of fine-tuning strategies and performance optimization was beyond the scope of the current study and remains an important area for future work. We have now clarified this point in the Discussion section on page 20, line 3, "Evaluation on long-term monitoring data from existing implantable systems, such as the NeuroPace RNS System, is warranted to determine whether the platform can achieve clinically meaningful improvements in patient care. However, these investigations were beyond the scope of the current proof-of-concept study, which was intended primarily to establish the feasibility of the platform."

- Multiple places in the manuscript refer to lived experience input, but in very qualitative terms (e.g., "Based upon our experience and discussions with clinicians, caregivers, and patients...", "discussions with patients indicate"). Given that the premise of this project is rooted in making a system that can "focus on what matters most to [patients]", it is a significant limitation that the lived experience voice is treated so informally. The manuscript would be significantly strengthened with data showing what features of a system epilepsy patients and caregivers actually want, or showing how the subjects in this study actually viewed the system. The subjects were queried about "usability", but did they actually find the interactions beneficial, helpful, positive, or supportive? How did they feel about a system that was learning from their brain data and behavior?

Response:

We thank the reviewer for this important observation. We agree that the lived-experience perspective is central to the premise of this work and that our original manuscript conveyed it too informally. We have made several changes to address this directly. First, to clarify the basis for our design priorities, we revised the Introduction (page 4, line 3) to ground the six proposed device features in the published literature on the lived experience and stated priorities of people with epilepsy and their caregivers, alongside our group's decades of clinical experience. Second, we recognize that our original treatment of patient experience relied heavily on the System Usability Scale, which speaks to whether the interface was easy to use but not to whether patients found it beneficial or supportive. We now report the qualitative exit-survey responses directly in the Discussion (page 18, line 39), including patients' own descriptions of how the system helped them track their daily reports, interpret their sleep data, and validate the relationship between their episodes and emotions. Third, these responses revealed a pattern that we believe speaks directly to the reviewer's concern about the context of our findings: patients were notably more equivocal about the awareness the system provided during their brief admission than about the value they anticipated from

everyday use. We have therefore added a paragraph to the Limitations section (page 19, line 42) noting that the EMU is an artificial, time-limited environment and that the information the platform could surface there constitutes a feasibility demonstration rather than the day-to-day decision support most valuable in ambulatory life. Finally, we did not directly assess how patients feel about a device that continuously learns from their brain activity and behavior, though we agree it is an important question, and is something we will investigate in future work.

- The safety framework of the AI agent is not sufficiently detailed. Given the potential for hallucinations by LLMs, a clinical setting with a vulnerable population introduces an ethical imperative to disclose fully and clearly the safety aspects of this system. As just one example, it is not even clear to me if the authors verified or have a mechanism to verify if the information provided by the LLM was accurate (i.e., reflected accurate values from the SQL database).

Response:

We thank the reviewer for raising this point. We fully agree that deploying an LLM-based system in a clinical setting with a vulnerable population carries an ethical imperative to disclose its safety architecture completely and clearly. We have ensured the manuscript now addresses this directly, and we summarize the relevant mechanisms here.

The generated SQL query and its result are logged alongside every response, so that each numerical claim can be audited against the underlying database. We used this logging to quantify accuracy: as reported in Results, the data agent executed retrieval tasks with 92.1% accuracy (n=35/38), which we determined by inspecting the SQL query generated for each data-agent turn and verifying both that the retrieved data and the synthesized response correctly answered the patient's question (described in the “Conversational agent” subsection of the Methods (page 27, line 33)). The broader safety framework is now described in full in this section as well.

We also describe the pre-deployment red-teaming through which the AI agent system was hardened. Investigators iteratively attempted to elicit unsafe or off-protocol behavior, including persona-altering prompts, prompt injection, off-topic and harmful-content requests, and attempts to coerce the *data agent* into returning data outside its intended scope. The system prompts, tool-calling logic, and moderation rubric were refined until such probes no longer succeeded (Methods, page 28, line 1). We also note that the agent's role was deliberately scoped to information delivery and structured data retrieval rather than autonomous clinical judgment, and that all patients were informed

that the system was a research prototype and that clinical concerns should be directed to their care team (“Ethical considerations”, page 20, line 12).

Reviewers' Comments

Reviewer #1 (Remarks to the Author):

The authors have addressed almost every single point I raised. I have no additional criticisms.

Reviewer #1 (Remarks on code availability):

N/a

Response:

Thank you.

Reviewer #2 (Remarks to the Author):

The authors have made substantial revisions and have addressed all of my previous comments. I do not have further comments.

Response:

Thank you.

Reviewer #3 (Remarks to the Author):

This revised article does the vast majority of what I and other reviewers asked for. It tones down excessive claims, provides a clearer description of what was done, and provides more details on the LLM architecture to support evaluation and possible replication. It remains interesting and conversation-provoking. I do have some minor comments that I believe can be handled through the editorial process without another round of peer review.

Response:

Thank you.

1. Patients appear to have used the system primarily for general chatting, with only 11% of their messages being queries to the system (around line 218). Comment on this

further in the Discussion, please. This is an important point, because it highlights that there will be some substantial UI/UX work needed for this to achieve the vision of a co-manager/patient assistant.

Response:

We thank the reviewer for this important point. We have added comment on this in the Discussion (page 15, line 15): “We anticipate that sustained engagement will derive from delivering clear day-to-day value and from letting patients configure the AI’s behavior to their needs and preferences. Such personalization was deliberately constrained in this proof of concept and, for an adaptive assistive technology, is likely to develop over months of use rather than a brief inpatient stay. A rigorous, prospective study of user experience and needs is a priority for our planned outpatient deployment.”

2. Closely related, Figure Extended 3 shows very limited engagement with seizure prompts, and it looks like this falls off during the stay (the right side of the graph is mostly white). This suggests that novelty drove early engagement, but utility was not there to sustain it. This is again fine for proof of concept but needs explicit comment in the Discussion.

Response:

We agree that sustained engagement over longer horizons is an important question. We have incorporated this discussion point into the changes detailed in our response to the above point.

3. The decision to make the agent architecture non-stateful is logical, and probably led to more controlled/predictable behavior. On the other hand, it also likely would reduce helpfulness of a real-world system, because it cannot know a user’s communication preferences well with only 200 messages of history. Many of the anecdotes around AI/LLM as useful companion to a human involve months of co-adaptation (of course, so do many of the stories of problems). Same comment – touch on this in the Discussion in terms of next steps and technology development that is still needed.

Response:

We appreciate this thoughtful observation. The reviewer is correct that our deliberately constrained architecture, while favoring predictable and auditable behavior, limits the achievable personalization. We have added comment on this in the Discussion, also part of the above section.

4. Revised figure 5 is good, but I notice that the two forms of delta sensitivity/F1 plots are on very different Y-axes, and having them on the same axis would highlight the substantial difference in the effects of the two retraining regimes.

Response:

Thank you for your suggestion. We have revised Figure 5 to use consistent axis ranges for delta sensitivity/F1 in panels b and c.

Reviewer #3 (Remarks on code availability):

The revised availability is much better. I cannot run it, but then again, nobody could. I do think this would be sufficient to attempt replication.

Response:

Thank you for your efforts to replicate the study. We acknowledge that replication was not feasible using the current codebase because it depends on specialized hardware (e.g., Natus acquisition systems and the associated real-time streaming SDK) and secure cloud infrastructure and configurations in Azure and Databricks. These components are integral to the real-time data acquisition and processing pipeline and cannot be reproduced solely from the released source code. We have revised the code availability statement to explicitly describe these dependencies and their impact on reproducibility. Page 28, line 10, “The fifth component, edge-device EEG streaming to Azure Event Hubs, depends on the proprietary Natus NeuroWorks real-time streaming SDK, which is confidential and cannot be publicly released; this code can be shared under a usage agreement with Natus. Replication of the complete framework also requires secure Azure and Databricks cloud infrastructure and configuration, which are institution-specific and are therefore not fully reproducible from the released source code alone.”

Reviewer #4 (Remarks to the Author):

I appreciate the effort the authors went to in responding to the reviews and revising the manuscript. I would agree that there have been many improvements in the technical clarity of what was actually done and reducing the degree that the results were being overstated.

That all being said, I still find this manuscript to feel largely like a much more preliminary report than I would expect to see in Nature Medicine. Acknowledging the enormous amount of effort that the authors went to in building this system, the reality is that I think

the demonstrated benefit remains limited. The authors essentially acknowledge that the approach is not yet able to demonstrate performance with translational potential, citing the slow enrollment rates in the EMU. I can fully appreciate this, but the limited data collection rate does not in and of itself elevate the impact of the results. In my view, it would be more appropriate to publish this proof of concept in a more specialized journal and then publish a report demonstrating clear clinical impact in Nature Medicine when enough data has been accumulated to clearly demonstrate the value.

The one place where I feel the authors have the opportunity to add clear impact with the enrollment numbers they have is in a formalized study of the user experience and desires. While I appreciate that the authors added a section to the discussion on this topic and agree that it is a welcome addition to the manuscript, the section primarily contains a few selected quotes from the exit survey (and citations to supporting literature). To be a substantial contribution, this section would need to contain an analysis that is more rigorous and comprehensive.

Response:

We thank the reviewer for recognizing the improvements in technical clarity and acknowledging the engineering effort involved. The central contribution of this work is, to our knowledge, the first demonstration of a bidirectional, patient-facing platform that lets people converse with their neurological device in real time and that adapts its detection algorithms from patient feedback without expert intervention. We fully agree that demonstrating clear clinical impact at scale is the essential next step, and that is precisely the aim of the planned outpatient deployment. We note, however, that a fully powered, formal UX analysis is constrained by the current sample size and by the artificial, time-limited EMU environment, and we believe it is most meaningfully conducted in the ambulatory setting where the platform's day-to-day value can actually be assessed.

Reviewer #5 (Remarks to the Author):

This revised manuscript has been substantially strengthened and addresses the major concerns raised during the previous round of review. The authors have appropriately tempered their claims, consistently framed the adaptive analyses as pseudo-prospective, improved the statistical reporting (including confidence intervals, effect sizes, and multiple-testing correction), and significantly expanded the discussion of governance, safety, and ethical oversight. The manuscript now provides a more balanced interpretation of the findings and a clearer acknowledgement of its limitations, including the proof-of-concept nature of the study, the small sample size, and the constraints of the inpatient EMU setting.

The work presents a genuinely novel platform for bidirectional patient–AI interaction that has the potential to influence the future design of intelligent neurotechnologies. The technical implementation is rigorous, the integration of patient- and clinician-in-the-loop model adaptation is compelling, and the manuscript is transparent about the current limitations and the need for future outpatient validation.

Response:

Thank you for your comment.

I have only a few minor suggestions. First, some of the more visionary statements in the Discussion and Conclusion (e.g. "With the current pace of innovation in computing, deep learning, and neuroscience, it is only a matter of time before these kinds of interactions become commonplace") could be further moderated to better reflect the current level of evidence.

Response:

We thank the reviewer. We have moderated the visionary language in the Discussion and Conclusion to better reflect the current level of evidence. In particular, the statement that "it is only a matter of time before these kinds of interactions become commonplace" has been moderated to: "...such interactions may become increasingly common over time," (page 16, line 39).

Second, while the governance framework is now well described, the Discussion could briefly acknowledge broader human factors relevant to long-term deployment, such as automation bias, trust calibration, clinician responsibility, and the implications of relying on evolving foundation models in regulated clinical environments.

Response:

We agree and have added this. The Discussion now acknowledges broader human-factors considerations relevant to long-term deployment: "Beyond these safeguards, long term deployment surfaces broader human-factors considerations that warrant careful attention. Patients may over-rely on automated systems, and how an interface represents the AI's confidence and its limitations will shape whether patients weigh its outputs appropriately. Questions of clinical responsibility also become more complex once medical decisions are informed in part by an AI. How that responsibility is allocated remains to be established as these systems move toward routine care. Underlying all of this are rapidly advancing AI models, raising the question of how to

continually evaluate and ensure trustworthy behavior in medical settings,” (page 16, line 15).