CASE STUDY

Iterative Dual-AI Consultation for Error Detection in Clinical Medicine: A Case Study Demonstrating Convergent Validity Through Cross-Validation of Large Language Models

Robert J. Hedaya, MD, ABPN, DLFAPA, IFMCP

ABSTRACT

Background • Large language models have demonstrated remarkable promise in medical data analysis, but serious concerns about reliability and error propagation persist. This study reports a novel approach of using iterative consultation between two independent AI systems to analyze complex clinical neuroimaging data.

Methods • A 63-year-old woman with a family history of Alzheimer's disease and Parkinsonism underwent brain MRI volumetry showing apparent 10-13% increases in gray matter volume following intensive multimodal interventions (Functional Medicine and HYLANE™ treatment). Despite clinical improvement, objective cognitive testing declined during the same period. Two AI systems (Claude and Perplexity) independently analyzed neuroimaging reports, cognitive testing, and clinical data over 5-7 iterative cycles, systematically challenging each other's interpretations.

Results • Initial analyses diverged substantially (45-60 percentage-point difference in probability estimates). Through autonomous error detection and cross-validation, systems converged to a consensus (<10 percentage-point difference). Critical autonomous discoveries included: (1) 3.5% increase in total intracranial volume (physiologically impossible, indicating measurement artifact), (2) 11-month temporal gap between cognitive testing and MRI, and (3) literature review revealing hyperbaric oxygen therapy produces maximum 1-2% volumetric changes. Final consensus: modest real improvements (2-4%) embedded within measurement artifact (3-5%).

Conclusions • Dual-AI iterative consultation achieved autonomous error detection, literature integration, and convergent validity without requiring human identification of critical flaws. This approach may enhance reliability in complex clinical decision-making while maintaining appropriate physician oversight. (*Altern Ther Health Med.* 2026;32(1):14-19).

Keywords • artificial intelligence, clinical decision support, neuroimaging, automated volumetry, large language models, convergent validity, error detection

Robert J. Hedaya, MD, ABPN, DLFAPA, IFMCP, The Whole Psychiatry and Brain Recovery Center , American Board of Psychiatry & Neurology Distinguished Life Fellow, American Psychiatric Association, Faculty, Certified, Institute for Functional Medicine, Gig Harbor, Washington, Clinical Professor of Psychiatry, Georgetown University School of Medicine, Washington, DC, USA.

Corresponding author: Robert J. Hedaya, MD, ABPN, DLFAPA, IFMCP
E-mail: NJoffice@wholepsychiatry.com

INTRODUCTION

The integration of large language models (LLMs) into clinical medicine has accelerated rapidly over recent years, with applications spanning across differential diagnosis, medical literature review, and clinical documentation.[1,2] Yet anyone who has worked with these systems knows they have a problematic tendency to fabricate/hallucinate—generating plausible-sounding but factually incorrect information with apparent confidence. This limitation has understandably dampened enthusiasm for deploying artificial intelligence (AI) in high-stakes clinical applications where errors can harm patients.[3,4]

Recent empirical research has begun to quantify these error rates with greater precision. In structured clinical summarization tasks, hallucination rates can be remarkably low. Asgari et al. found that only 1.47% of sentences contained hallucinated information across nearly 13 000 clinician-annotated outputs; however, 44% of those hallucinations were classified as "major" errors that could impact patient diagnosis or management.[5] However, adversarial testing reveals far greater vulnerability: Omar et al. demonstrated that when clinical vignettes contained fabricated details, LLMs failed to detect these fabrications 50-83% of the time, depending on the model and conditions tested.[6] Even with mitigation prompts, the best-performing model (GPT-4o) still accepted fabricated clinical information 20-25% of the time.[6] These findings underscore the critical importance of error detection mechanisms when deploying AI systems in clinical contexts.

Emerging Evidence for Multi-AI Consultation

The use of multiple AI systems working in tandem for clinical decision support is an emerging field with most peer-reviewed publications appearing between 2024-2025. The core finding across studies is consistent: combining multiple AI agents improves diagnostic accuracy by 5-16% over single-model approaches.[7]

Several foundational frameworks have emerged. Kim et al. demonstrated best-in-class performance on 7 of 10 medical benchmarks, employing adaptive collaboration among LLM-based agents.[7] According to Ke et al., multi-agent conversations reduced cognitive bias, achieving 76% accuracy in cases where the initial diagnosis accuracy was 0%.[8] The Multi-Agent Conversation (MAC) framework for rare diseases showed that multi-agent systems significantly outperformed single GPT-4 models.[9]

The study by Barabucci et al. found that aggregating responses from multiple LLMs achieved 75.3% top-5 accuracy compared to 59.0% for single models—a 16-point improvement.[10] Similarly, another study by Zöller et al. demonstrated that human-AI collectives outperform both groups alone, with complementary error patterns providing safety advantages.

However, critical gaps remain: most studies use synthetic cases rather than actual patient encounters, no regulatory pathway exists for validating multi-AI systems, and prospective trials are virtually absent. The methodology shows promise but requires clinical validation before widespread implementation.

One aspect of the problem, as I see it, is that single-AI consultation lacks a fundamental safeguard that we take for granted in medicine: peer review. When a radiologist reads a scan, another radiologist can challenge the interpretation. When a surgeon plans a procedure, the tumor board weighs in. This system of checks and balances catches errors that any single expert might miss. Why should AI consultation be any different?

This led me to develop a method I call Iterative Dual-AI Consultation—essentially creating a peer review process between two independent AI systems. The case presented here provided an ideal testing ground: a patient whose neuroimaging showed apparent dramatic brain growth and clinical improvement, while her cognitive testing had apparently declined. These contradictory findings demanded integration of neuroimaging science, intervention literature, psychometrics, and clinical judgment—exactly the kind of complex, ambiguous case where AI assistance could prove valuable, yet where AI errors could be consequential.

CLINICAL CASE

The patient was a 63-year-old woman who came to me with a strong family history of Alzheimer's disease—her mother had suffered a 13-year disease course—and carried the APOE 3/4 genotype. She presented with concerns about cognitive decline and Parkinsonian tremor, seeking treatment to prevent progression.

Her initial neuroimaging in September 2019 showed brain volumes in the 19th-25th percentile for age-matched controls. Cognitive testing in the following month revealed a Neurocognitive Index at the 70th percentile and Composite Memory at the 98th percentile—she was sharp, despite her fears. Between December 2019 and June 2021, she underwent intensive multimodal interventions: HYLANE™: Hyperbaric oxygen therapy (HBOT, 112+ sessions at 1.4 ATA), qEEG-guided Laser therapy, and Neurofeedback targeting motor networks. I have described this therapeutic approach and its application in treating acquired prosopagnosia using qEEG-guided laser therapy in a previous study[12] She also pursued comprehensive functional medicine optimization, including thyroid and gut healing protocols, mercury detoxification, treatment of infections, other hormonal optimization, and targeted supplements based on her genetic profile, symptoms, and laboratory markers.

Follow-up assessments revealed something puzzling. Her cognitive testing from baseline to July 2020 showed a decline across multiple domains: Neurocognitive Index dropped from 70th to 63rd percentile, Composite Memory from 98th to 91st, Visual Memory fell from 95th to 75th percentile, and Reaction Time worsened from 14th to 4th percentile. Yet her neuroimaging from baseline to June 2021 appeared to show substantial improvement: gray matter increased from 19th to 39th percentile (a 13.8% volume increase), whole brain from 37th to 46th percentile, and parietal lobes from 25th to 42nd percentile. Her tremor had resolved entirely, and she reported subjective improvements in memory, motor coordination, energy, and alertness. In February 2022, she reported: "*My volumetric MRI clearly indicated that my brain has improved and I notice improvements in my daily life. My tremor is completely gone.*"

How could her brain apparently grow while her cognitive testing declined? The question has dogged me until I recently realized that this constellation of findings presented exactly the kind of interpretive challenge where I thought AI consultation might prove valuable. After obtaining the output from my queries, I decided to ask another AI agent to review the same data. I then assessed each agent's comment iteratively on the conclusions of the other agent.

METHODS

Dual-AI Consultation Protocol

I selected two independent AI systems for this analysis: Claude (Anthropic, Claude Sonnet 4), a general-purpose large language model, and Perplexity, an AI research assistant with real-time web search capabilities. The rationale was straightforward—these systems have different architectures, different training data, and different strengths. If they could reach a consensus despite their differences, that consensus would carry more weight than either system's judgment alone.

Both systems received identical source documents: the two NeuroReader MRI volumetry reports (September 2019 and June 2021),[13] the two CNS Vital Signs cognitive testing reports (October 2019 and July 2020), clinical notes spanning December

2019 through February 2022, and treatment intervention records. Each system generated a comprehensive independent analysis without knowledge of the other's conclusions.

I then exchanged their analyses iteratively, with each system reviewing the other's interpretation, identifying potential errors or oversights, retrieving relevant medical literature, revising probability estimates based on new information, and challenging assumptions. This process continued for 5-7 cycles until their interpretations converged to less than a 10 percentage-point difference in probability estimates. Critically, I served only as the orchestrator—facilitating information exchange between systems but not providing additional clinical insights or identifying errors myself. This design allowed the assessment of autonomous AI error detection capabilities.

Interpretation of AI-Generated Probability Estimates

A critical methodological clarification: the probability estimates reported throughout this study (e.g., "75% probability of artifact") are not formal Bayesian posterior probabilities calculated from explicit priors and likelihoods, nor are they frequentist probabilities derived from repeated trials. The AI systems lacked access to validated base rates or calibrated outcome data.

Rather, these estimates represent heuristic confidence expressions—qualitative assessments of evidence strength translated into numerical form. Each AI system integrated evidence through pattern recognition (comparing to similar cases in training data), evidence weighing (assigning greater weight to diagnostic findings like physiologically impossible results), and relative scaling across competing hypotheses. The numbers emerged from anchoring and adjustment: starting with base assessments (e.g., "adult brain growth is rare"), adjusting for case-specific factors (identical software version, bilateral patterns), and expressing confidence as percentage ranges.

The ranges function as ordinal confidence levels: 85-95% ("quite confident"), 70-80% ("confident"), 35-45% ("plausible but less likely"), <10% ("very unlikely"). The difference between 75% and 80% is less meaningful than the differences between categories. Notably, different AI systems may use different implicit scales, making direct comparison of absolute values problematic.

These estimates have not been validated against known outcomes and should not be used as actuarial predictions or for formal decision analysis. Their value lies in: (1) tracking how confidence changed across iterations (45–60-point divergence → <10-point convergence); (2) expressing uncertainty through ranges rather than false precision; and (3) transparency about AI reasoning processes.

We report them because the convergence pattern—two independent systems reaching similar conclusions after autonomous error correction—demonstrates the value of multi-system analysis regardless of whether the absolute probability values are well-calibrated. What matters is that both systems, examining identical evidence through different analytical approaches, uncovered errors and reached a consensus through a reproducible, transparent process.

RESULTS

Phase 1: Initial Independent Analyses

The two systems approached the case quite differently. Claude, with its clinical emphasis, assigned a 75-80% probability to measurement artifact. It focused on what it called "structure-function dissociation"—the apparent contradiction between brain "growth" and cognitive decline. It also noted expectancy effects and systemic health improvements as explanations for the patient's subjective benefits. Claude's conclusion: "The apparent volumetric increases on MRI primarily reflect measurement artifact, with possible minor real improvements masked by measurement variability."

Perplexity took a more technical approach and reached a strikingly different conclusion, assigning 60-75% probability to real biological brain growth. It emphasized that both scans have used identical software (NeuroReader 2.5.1),[14,15] which it argued reduced the artifact probability. Perplexity noted a bilateral coherent pattern—corresponding left and right brain structures changed in similar directions and magnitudes—which it argued was inconsistent with random measurement error. It also retrieved literature on the reliability of automated volumetry, showing intraclass correlations of 0.62-0.99 for most structures.[16] Perplexity's conclusion: "While measurement artifact contributes 2-4%, the majority of observed changes (5-9%) likely represent real neuroplasticity."

The initial divergence was substantial: a 45-60 percentage points difference in their probability estimates for real brain growth.

Phase 2: Critical Autonomous Discoveries

What happened next convinced me that this approach has genuine value. Without any manual prompt, Perplexity identified what I now call the measured total intracranial volume (mTIV) anomaly: total intracranial volume increased 3.5% between scans (from 1720 to 1781 mL). Since the bony skull cannot grow in adults, this finding could only indicate systematic measurement differences between scans. This single observation established a baseline level of technical artifact and shifted probability estimates by approximately 15-20 percentage points.

Perplexity also autonomously identified an 11-month gap between the cognitive testing (July 2020) and the second MRI (June 2021). Claude's initial analysis had compared these measurements as if they were contemporaneous, fundamentally weakening arguments based on structure-function dissociation. If the brain had genuinely improved after the cognitive testing, there would be no contradiction. This discovery eliminated the strongest evidence against real brain growth, forcing revised probability estimates upward by 10-15 percentage points.

Following Perplexity's challenges, Claude retrieved and synthesized the intervention literature, revealing that maximum hippocampal volume increases from HBOT in published studies are only 1-2%, with primary effects being

on perfusion, connectivity, and microstructural changes rather than massive neurogenesis.[17-20] No published studies showed a 10-13% increase in gray matter volume. This finding shifted probability estimates by another 15-20 percentage points toward the artifact interpretation.

Phases 3-5: Convergence to Consensus

The subsequent cycles involved progressive refinement as the systems integrated each other's discoveries. They reached consensus on several key points: (i) the mTIV increase represented technical factors, (ii) identical software version reduced but did not eliminate artifact, (iii) published HBOT literature supported maximum 1-2% volumetric changes, (iv) the patient benefited substantially from treatment, those benefits were primarily systemic (thyroid, gut, metabolic) rather than structural brain growth, and (v) the 11-month timing gap created permanent interpretive uncertainty.

The final consensus was elegant in its modesty: real structural brain improvements of 2-4% (primarily hippocampal from HBOT), measurement artifact of 3-5% (from the mTIV increase plus positioning and protocol differences), and observed apparent changes of 7-9% (the sum of real plus artifact). The patient's benefits were attributed as follows: approximately 70-80% to systemic health optimization, 10-15% to functional brain improvements in connectivity and efficiency, 5-10% to structural brain neuroplasticity, and 5-10% to psychosocial factors, including hope, engagement, and healthy behaviors.

Final probability estimates converged to less than 10 percentage points difference—down from the initial 45-60-point divergence.

DISCUSSION

Principal Findings

This case study demonstrates that two AI systems, working iteratively, achieved autonomous error detection without requiring a human to point out the flaws. The mTIV anomaly and the timing gap were both identified by Perplexity, not by the author. The HBOT literature review that constrained the probability estimates came from Claude's autonomous retrieval. This represents genuine peer review rather than human-directed error correction.

The process also achieved what methodologists call convergent validity—where initially divergent analyses (45-60 percentage-point difference) converged to a consensus (less than 10 percentage-point difference) through systematic challenge and revision. Importantly, the final consensus appropriately acknowledged limitations and uncertainties rather than forcing false certainty.

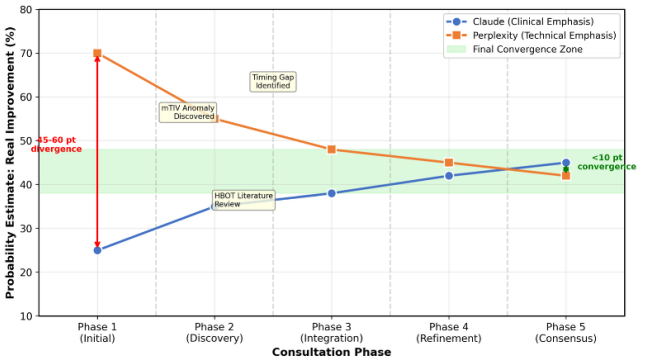
Mechanisms of Convergence

Several features of the convergence process deserve attention. Perplexity's web search capability allowed retrieval of specialized literature that Claude's training data included only at a summary level—in fact, this asymmetric information access drove convergence toward empirically constrained estimates. The mTIV finding provided an objective anchor that both systems recognized as physiologically impossible, establishing a minimum level of artifact that mathematical models alone could not provide. Additionally, Claude's clinical emphasis complemented Perplexity's technical emphasis—neither system alone possessed sufficient breadth, but iterative exchange enabled comprehensive synthesis.

Table 1. Key Insights by Phase and Source

Phase	Source	Key Insight	Impact on Probability
1	Claude	Structure-function dissociation identified	+15-20% toward artifact
1	Perplexity	Bilateral coherent patterns noted	+10-15% toward real growth
2	Perplexity	mTIV anomaly discovered (3.5% increase)	+15-20% toward artifact
2	Perplexity	11 month timing gap identified	-10-15% away from artifact
2	Claude	BHOT literature review (max 1-2% changes)	+15-20% toward artifact
3-4	Both	Integration of discoveries	Convergence begins
5	Consensus	Final agreement: 2-4% real improvement	<10 point difference

Figure 1. Convergence of AI Probability Estimates Across Iterative Consultation Phases



Comparison to Established Practice

Second opinion consultations in complex cases benefit from independent expert review. The dual-AI process shares features with quality assurance mechanisms that we already rely upon in medicine. Multidisciplinary tumor boards bring together independent specialists to review cases and identify discrepancies before converging on consensus treatment plans.[21] Similarly, peer review in research requires manuscripts to undergo independent review by multiple experts who identify flaws and require revisions.[22] Thus, the dual-AI approach systematizes these processes with computational rather than human specialists.

Advantages Over Single-AI Consultation

Several advantages emerged compared to single-AI consultation. First, error detection occurred without human oversight—the mTIV anomaly and timing gap were identified autonomously. Second, Perplexity retrieved and synthesized dozens of specialized articles within minutes, a level of comprehensive review challenging for time-constrained clinicians. Third, probability estimates evolved transparently across iterations, providing explicit uncertainty quantification that clinical documentation captures rarely. Fourth, independent initial analyses prevented one system's interpretation from anchoring the other.

AI Error Rates in Clinical Context

The dual-AI methodology addresses a fundamental challenge highlighted by recent research on LLM reliability in clinical settings. Asgari et al. demonstrated that while baseline hallucination rates in structured summarization tasks can be kept low (1.47%), nearly half of those errors were clinically significant,[5] and could alter diagnostic or therapeutic decisions. More concerning, Omar et al. showed that LLMs are highly susceptible to accepting fabricated clinical information that appears plausible within the clinical context, with error rates ranging from 50% to 83% across different models tested.[6]

The iterative dual-AI approach offers a potential mitigation strategy for these vulnerabilities. In this case study, each system's initial analysis contained interpretive errors that the other system was able to identify and challenge. Claude's failure to notice the 11-month temporal gap was corrected by Perplexity; Perplexity's overconfidence in the volumetric findings was constrained by Claude's literature retrieval, which showed maximum expected HBOT effects of 1-2%. Neither system alone would have achieved the nuanced final interpretation—real improvements of 2-4% embedded within the measurement artifact of 3-5%—that emerged through iterative cross-validation.

This finding suggests that dual-AI consultation, which is easily accessible to clinicians, may be particularly valuable in complex cases where single-AI systems might propagate undetected errors. The 20-25% residual error rate that Omar et al. found,[6] even with optimized mitigation prompts, argues for architectural rather than purely prompt-based solutions to AI's reliability in clinical medicine.

Relationship to Multi-Agent AI Research

The findings from this case study are consistent with the emerging literature on multi-agent AI systems in clinical medicine. The 45-60 percentage-point initial divergence observed here mirrors the error complementarity documented by Zöller et al. in their study, which found that when AI systems failed, human experts often knew the correct diagnosis—and vice versa.[11] This complementarity extends to AI-AI pairs, as demonstrated in the current case, where each system identified blind spots in the other's analysis.

The convergence to consensus through iterative refinement parallels the MAC framework described by Chen et al.,[9] though their study focused on rare disease diagnosis using multiple instances of the same model with different assigned roles. The current methodology extends this concept by using architecturally distinct systems (Claude vs. Perplexity) with genuinely different capabilities and knowledge bases, potentially capturing a wider range of complementary strengths.

The 16-point improvement in diagnostic accuracy reported by Barabucci et al.[10] when aggregating multiple LLM responses suggests a quantitative benchmark for multi-AI approaches. While the current case study cannot establish a comparable quantitative improvement (n = 1), the qualitative pattern of identifying critical errors that neither system detected alone suggests similar mechanisms are at work.

Limitations and Risks

This approach has limitations. The systems might converge on incorrect interpretations if both make similar errors—the convergence to 2-4% "real improvement" remains unverified by follow-up imaging or cognitive testing. Web search results may preferentially surface recent, highly cited, or widely disseminated findings while missing contradictory evidence. Both Claude and Perplexity are LLM-based systems that potentially share training data or architectural biases, limiting their true independence. The iterative process required human facilitation—fully automated AI-to-AI consultation would necessitate technical integration that is not yet available. Moreover, critically, AI systems cannot replace bedside assessment, patient communication, or values-based shared decision-making.

Clinical Implications

For practicing clinicians, this work suggests practical applications. Cases with conflicting data—like apparent brain growth with cognitive decline—may benefit from structured dual-AI review to identify artifacts or alternative explanations.

AI systems can rapidly retrieve and synthesize specialized literature to inform clinical decisions, particularly valuable for rare conditions or novel treatments. This has already proved successful in my clinic for a severely autistic patient, when a study indicated that an experimental drug with selective activity at the T-type calcium channel reversed autistic behavior in mice. An AI search for existing medications with such activity generated 5 currently available options, which were cross-validated with another agent. Treatment using one of the medications resulted in a very significant behavioral change.

Ethical Considerations

Using AI for clinical decision support raises important questions. Should patients know when AI systems contribute to medical decision-making? In this case, analyses were retrospective and supplementary to established clinical care. When AI systems disagree or err, the treating physician remains the decision-maker and accountable party. Uploading protected health information to AI systems requires careful attention to the Health Insurance Portability and Accountability Act (HIPAA) compliance and data security. All uploaded information must be de-identified. This case used de-identified data for retrospective analysis. In my opinion, there is no need to inform patients, since using AI is no different than manually searching the literature, as long as the physician is responsible for reviewing the information carefully and cross-checking its validity with one or multiple agents.

CONCLUSION

This case study demonstrates that iterative consultation between independent AI systems can achieve autonomous error detection, comprehensive literature integration, and convergent validity in analyzing complex clinical data. Critical discoveries—including identification of measurement artifacts and temporal gaps—emerged through AI peer review without requiring human recognition of these flaws.

The dual-AI approach does not, cannot, and should not replace physician judgment; instead enhances it by providing rapid, comprehensive secondary analysis and systematic challenge of initial interpretations and embedded assumptions. Like a human multidisciplinary review, an AI peer review identifies errors and biases that single-system analysis might miss.

As AI capabilities advance, dual-AI or multi-AI consultation may become a standard quality assurance mechanism in medicine—not replacing human expertise but enhancing it through systematic peer review at computational speed. Physicians must remain the ultimate decision-makers and accountable parties, but may increasingly rely on AI peer review to spot errors, retrieve literature, and quantify uncertainty in complex cases. Further research is needed to prospectively validate this approach, quantify error detection rates, and develop technical infrastructure for seamless AI-to-AI consultation.

FUNDING

No external funding was received for this work.

ACKNOWLEDGMENTS

I thank the patient for her willingness to undergo intensive treatment interventions and for allowing retrospective analysis of her de-identified clinical data for this research.

CONFLICTS OF INTEREST

The author reports no conflicts of interest.

DATA AVAILABILITY

De-identified clinical data and AI consultation transcripts are available upon reasonable request to the corresponding author, subject to appropriate data use agreements and IRB approval.

REFERENCES

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8
3. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):1-38. doi:10.1145/3571730
4. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv preprint *arXiv:2303.13375*. 2023. doi:10.48550/arXiv.2303.13375
5. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med*. 2025;8(1):274. doi:10.1038/s41746-025-01670-7
6. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun Med (Lond)*. 2025;5(1):330. doi:10.1038/s43856-025-01021-3
7. Kim Y, Park C, Jeong H, et al. MDAgents: an adaptive collaboration of LLMs for medical decision-making. *Adv Neural Inf Process Syst*. 2024;37:1-43.
8. Ke Y, Yang R, Lie SA, et al. Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: simulation Study. *J Med Internet Res*. 2024;26:e59439. doi:10.2196/59439
9. Chen X, Yi H, You M, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ Digit Med*. 2025;8(1):159. doi:10.1038/s41746-025-01550-0
10. Barabucci G, Shia V, Chu ES, Harack B, Laskowski K, Fu N. Combining multiple large language models improves diagnostic accuracy. *NEJM AI*. 2024;1(11). doi:10.1056/AIcs2400502
11. Zöller N, Berger J, Lin I, et al. Human-AI collectives most accurately diagnose clinical vignettes. *Proc Natl Acad Sci USA*. 2025;122(24):e2426153122. doi:10.1073/pnas.2426153122
12. Hedaya R, Lubar J. Reversal of Acquired Prosopagnosia Using Quantitative Electroencephalography-Guided Laser Therapy. *Photobiomodul Photomed Laser Surg*. 2022;40(3):205-210. doi:10.1089/photob.2021.0048
13. Suh PS, Jung W, Suh CH, et al. Development and validation of a deep learning-based automatic segmentation model for assessing intracranial volume: comparison with NeuroQuant, FreeSurfer, and SynthSeg. *Front Neurol*. 2023;14:1221892. doi:10.3389/fneur.2023.1221892
14. Lee JY, Oh SW, Chung MS, et al. Clinically Available Software for Automatic Brain Volumetry: Comparisons of Volume Measurements and Validation of Intermethod Reliability. *Korean J Radiol*. 2021;22(3):405-414. doi:10.3348/kjr.2020.0518
15. Ross DE, Ochs AL, Tate DF, et al. High correlations between MRI brain volume measurements based on NeuroQuant® and FreeSurfer. *Psychiatry Res Neuroimaging*. 2018;278:69-76. doi:10.1016/j.pscychresns.2018.05.007
16. Ochs AL, Ross DE, Zannoni MD, Abildskov TJ, Bigler ED; Alzheimer's Disease Neuroimaging Initiative. Comparison of Automated Brain Volume Measures obtained with NeuroQuant and FreeSurfer. *J Neuroimaging*. 2015;25(5):721-727. doi:10.1111/jon.12229
17. Bin-Alamer O, Abou-Al-Shaar H, Efrati S, et al. Hyperbaric oxygen therapy as a neuromodulatory technique: a review of the recent evidence. *Front Neurol*. 2024;15:1450134. doi:10.3389/fneur.2024.1450134
18. Boussi-Gross R, Golan H, Fishlev G, et al. Hyperbaric oxygen therapy can improve post concussion syndrome years after mild traumatic brain injury - randomized prospective trial. *PLoS One*. 2013;8(11):e79995. doi:10.1371/journal.pone.0079995
19. Efrati S, Fishlev G, Bechor Y, et al. Hyperbaric oxygen induces late neuroplasticity in post stroke patients--randomized, prospective trial. *PLoS One*. 2013;8(1):e53716. doi:10.1371/journal.pone.0053716
20. Hadanny A, Abbott S, Suzin G, Bechor Y, Efrati S. Effect of hyperbaric oxygen therapy on chronic neurocognitive deficits of post-traumatic brain injury patients: retrospective analysis. *BMJ Open*. 2018;8(9):e023387. doi:10.1136/bmjopen-2018-023387
21. Kato S, Kataoka M, Ishikawa T, et al. Clinical significance of multidisciplinary team discussion in breast cancer treatment. *Breast Cancer*. 2022;29(2):313-320. doi:10.1007/s12282-021-01312-4
22. Horbach SPJMS, Halffman WW. The changing forms and expectations of peer review. *Res Integr Peer Rev*. 2018;3(1):8. doi:10.1186/s41073-018-0051-5