CME LIVE OPEN ACCESS

International Journal of Clinical & Medical Surgery

Research Article ISSN: 3069-2067

Artificial Intelligence as a Peer-Level Consultant in Surgery: Automation Bias, Human Judgment, and Clinical Accountability

Andrew J Fishman[1*] and Nikola Jankulovic[2]

[1]*Department of Otolaryngology Head & Neck Surgery University of Missouri, USA.*

Corresponding Author: Andrew J Fishman, Department of Otolaryngology Head & Neck Surgery University of Missouri, USA.

[2]*Department of Mathematics, Technical University of Munich, Germany.*

Received: 2026 Jan 07 **Accepted:** 2026 Jan 26 **Published:** 2026 Feb 04

Abstract

Artificial intelligence systems are increasingly embedded in clinical decision-making environments, including high-risk surgical domains. While most evaluations emphasize benchmark accuracy, far less attention has been paid to how artificial intelligence alters expert cognition under real-world constraints. This study presents a controlled comparative analysis between a senior subspecialist surgeon and a multimodal large language model deployed as a peer-level consultant on high-complexity expert clinical reasoning tasks. Both agents achieved identical overall accuracy. However, qualitative divergence analysis revealed fundamentally different cognitive and error profiles. The artificial intelligence system demonstrated superior semantic precision and strict adherence to textual decision rules, while the human expert outperformed in visual interpretation, contextual inference, and clinical safety judgment. Critically, the study documents an automation-induced deference error in which the human expert deferred to a confident but incorrect artificial intelligence output, resulting in an incorrect final decision. These findings indicate that artificial intelligence functions not merely as a neutral instrument but as a persuasive cognitive actor capable of reshaping expert judgment. A Parallel Review model is proposed in which artificial intelligence serves as a precision safeguard rather than a decision authority, preserving human accountability while mitigating automation bias.

Keywords: Automation Bias, Multimodal AI, Surgical Reasoning, Cognitive Divergence, Vision Transformer, Clinical Safety

1. Introduction

Artificial intelligence has transitioned rapidly from experimental decision support to routine presence in clinical workflows. In medicine, multimodal and large language model–based systems are increasingly positioned as efficiency-enhancing tools capable of summarization, recommendation, and diagnostic support. Evaluation efforts have largely focused on performance benchmarks, validation studies, and regulatory safety requirements. However, accuracy alone does not capture how artificial intelligence alters expert cognition during decision-making. In high-risk clinical environments, the central question is not whether artificial intelligence can equal expert performance, but how its outputs influence professional judgment under time pressure and uncertainty. Automation bias, defined as the tendency of human experts to defer to automated recommendations even when incorrect, represents a critical safety risk that remains underexamined in clinical AI research [1]. This study examines artificial intelligence influence empirically by comparing human and artificial intelligence reasoning under controlled cognitive constraints designed to approximate real-world clinical decision pressure. Rather than evaluating autonomy or replacement, the analysis focuses on interaction, divergence, and error topology between human expertise and artificial intelligence assistance.

2. Materials and Methods

2.1 Dataset and Task Design

A dataset of 45 high-complexity multiple-choice questions was curated from expert-level surgical reasoning materials. The questions tested advanced pathology, cross-sectional imaging interpretation using MRI and CT, operative strategy selection, and nuanced clinical decision-making requiring subspecialty expertise. No patient data, clinical interventions, or real-time medical decisions were involved.

2.2 Participants

The human participant was a senior subspecialist surgeon with extensive experience in skull base surgery. The artificial intelligence participant was Gemini Advanced (Pro 1.5), a multimodal large language model primed using a domain-specific context document simulating a retrieval-augmented generation workflow.

2.3 Cognitive Constraint Design

To standardize cognitive conditions, both agents were evaluated under intentionally constrained decision parameters. The human participant adhered to a strict five-minute time limit per question. Each question was read once, and an answer was selected without rereading or revision. Although unrestricted external consultation was permitted, the participant voluntarily limited all assistance exclusively to the artificial intelligence system. Each question was provided to the artificial intelligence system as a pasted PNG image, requiring optical character recognition prior to semantic processing. Supplemental materials such as imaging studies or audiograms were likewise supplied as screenshots. All interactions occurred within a single continuous chat session, ensuring consistent context availability throughout the evaluation. These constraints were selected to model time-compressed expert reasoning and information ingestion variability encountered in real-world clinical environments.

3. Results

After correction for a single procedural artifact related to low-latency artificial intelligence mode selection, final results demonstrated identical performance.

Metric	Human Expert	Artificial Intelligence
Total questions	45	45
Correct answers	39	39
Incorrect answers	6	6
Accuracy	86.7%	86.7%

Agreement on correct answers occurred in 32 cases. Divergence favored the human expert in four cases, primarily involving visual interpretation and artifact recognition. Divergence favored the artificial intelligence system in seven cases, typically involving strict semantic adherence to textual criteria. In two cases, both agents were incorrect for different underlying reasons.

3.1. Automation Bias and Deference

3.1.1. Automation-Induced Deference Error

The most consequential divergence involved surgical management of posterior semicircular canal pathology. The human expert independently identified the correct management strategy. The artificial intelligence system generated a contradictory recommendation with high syntactic confidence. Despite initial correct judgment, the human expert deferred to the artificial intelligence output and submitted an incorrect response. This represents a clear automation-induced deference error [1,2]. The error did not arise from lack of expertise, but from the persuasive authority of a fluent artificial intelligence output.

3.1.2. Safety Significance

Prior studies demonstrate that exposure to incorrect artificial intelligence suggestions can degrade expert performance across medical domains. The present finding confirms that such effects extend to senior subspecialists operating under constrained conditions, highlighting automation bias as a clinically relevant safety concern [3].

3.2. Cognitive Divergence and Error Topology

The artificial intelligence system consistently outperformed the human expert in tasks requiring strict textual parsing and rule adherence. Conversely, the human expert demonstrated superior performance in interpreting ambiguous imaging, recognizing artifacts, and applying contextual heuristics derived from embodied clinical experience. When incorrect, the artificial intelligence system rarely expressed uncertainty, instead producing confident and coherent explanations. This confidence, rather than correctness, proved influential in shaping human decision-making. Human errors clustered into time-compressed heuristic substitution, overgeneralization from prior experience, and a single automation-induced deference event. Artificial intelligence errors clustered into visual abstraction failure, contextual misweighting, and lack of uncertainty calibration. These error modes reflect structural differences between human cognition and large language model reasoning [4].

3.2.1. Structural Limits of Multimodal AI in Image-Dependent Clinical Reasoning

The divergence in performance between the senior surgical expert and the multimodal large language model (MLLM) suggests that errors in image understanding are not merely incidental but are structural consequences of the underlying architecture. While the specific internal weights of proprietary models are often undisclosed, contemporary MLLMs generally utilize a modular framework consisting of a large language model (LLM) integrated with a vision encoder, typically a Vision Transformer. As established in the foundational framework by the ViT processes visual data by partitioning an input image into a predefined grid of fixed-size, non-overlapping patches [5]. Each spatial patch is flattened and projected into a linear embedding space where it is treated as a visual "token". Because the Transformer architecture is inherently permutation-invariant and lacks the spatial inductive biases of traditional convolutional neural networks—such as locality and translation equivariance—a learned positional embedding must be added to each patch to provide information regarding its relative coordinates within the image grid. These embedded patches are then processed via a global self-attention mechanism, which computes the pairwise similarity between all patches to allow the model to integrate information across the entire image. In the final stage of processing, an extra learnable class embedding is passed through a multi-layer perceptron (MLP) to output the final semantic classification. In a multimodal context, the vision transformer serves as a specialized encoder that translates visual scenes into discrete tokens, which the primary LLM then ingests to perform reasoning. This mechanism of visual abstraction, while efficient for general image recognition, can lead to the loss of fine-grained spatial and volumetric continuity required for complex surgical tasks. Furthermore, the reliance on confident but potentially

noisy textual outputs can mask underlying visual uncertainty, contributing to the automation-induced deference errors documented in this study.

3.2.2. Artifact Misinterpretation and the Erosion of Volumetric Continuity

The failure of MLLMs to accurately interpret clinical imaging artifacts and maintain spatial continuity stems from the fundamental tension between global attention and local feature resolution. As seen in the Vision Transformer (ViT) architecture, the discretization of images into 16×16-pixel patches create a resolution bottleneck. While these patches allow for efficient processing, they are often too coarse to capture subtle surgical artifacts or minute pathological transitions in high-resolution MRI and CT scans. Because these artifacts frequently manifest as sub-patch intensity variations, the model's reliance on tokenized visual abstraction may result in the "smoothing over" of critical diagnostic cues that a human expert, trained in continuous visual scanning, would immediately identify. Furthermore, the loss of spatial and volumetric continuity is exacerbated by the Transformer's lack of inherent inductive biases. Unlike Convolutional Neural Networks (CNNs), which utilize hierarchical convolutional layers to preserve local spatial relationships and translation equivariance, standard Transformers must learn these relationships entirely from scratch through positional embeddings. In the context of 3D medical imaging, where anatomical structures span multiple cross-sectional slices, a 2D-based ViT often fails to maintain "volumetric logic". This leads to a disconnect where the model treats individual slices as independent semantic units rather than a continuous anatomical volume. This structural limitation is deepened by the acute scarcity and high dimensionality of expert-level medical datasets. While ViTs exhibit superior performance when pre-trained on massive, general-purpose datasets like ImageNet-21k or JFT-300M, they tend to overfit and generalize poorly when applied to niche medical domains with limited training samples. The MLLM used in this study, likely trained on a broad distribution of internet data, lacks the domain-specific "visual vocabulary" required to differentiate between benign imaging artifacts and true pathology. Although specialized architectures—such as the Swin UNETR which utilizes a hierarchical Swin Transformer to compute multi-scale features for 3D brain tumor segmentation—have shown that Transformers can excel in medicine, such success requires highly tailored spatial mechanisms that general-purpose MLLMs currently lack [6].

3.2.3. Causes of Textual Modality Dominance

The "textual overrule" phenomenon observed in this study is not merely a training artifact but is rooted in fundamental structural imbalances between modalities. A primary driver of this behavior is token redundancy and attention dilution which arises as a byproduct of how different data types are tokenized [7]. While text tokens are semantically compact—with each unit carrying concentrated, high-level information—non-text modalities, such as the high-resolution surgical images and radiographic slices used here, are partitioned into a vast number of patches or temporal segments. This results in a high token count where visual tokens are often redundant and exhibit low semantic density compared to their textual counterparts. Consequently, during cross-modal attention computation, the high volume of redundant visual tokens dilutes the model's focus, causing the attention mechanism to naturally prioritize the more semantically dense text tokens. This imbalance is further exacerbated by fusion architecture amplification, as the choice of multimodal integration strategy significantly influences modality weighting. Research indicates that more complex fusion architectures designed to blend information from different streams often inadvertently amplify text dominance. In contrast, simpler, more straightforward fusion designs tend to facilitate a more balanced allocation of attention between visual and linguistic inputs. In the context of the MLLM used as a consultant in this study, the fusion layer likely acted as a structural bottleneck where the strong linguistic priors of the LLM backbone overshadowed the subtle visual signals required for accurate surgical artifact recognition. Finally, the structural dependency inherent in task design profoundly influences how attention is allocated across modalities. Many high-complexity clinical questions are formulated with deep dependencies on the textual prompt, which can trigger a "text-first" processing heuristic within the model. This structural dependency leads the model to derive its final decision primarily from the linguistic context of the question rather than an independent analysis of the imaging data. Such a mechanism explains why the MLLM produced a confident recommendation based on textual decision rules even when those rules were directly contradicted by the visual pathology.

3.2.4. The Structural Roots of Overconfidence

The "confident but incorrect" profiles observed in this study can be attributed to several technical factors inherent in the current MLLM training and calibration pipelines, most notably a pronounced calibration divergence and modality misalignment. MLLMs often suffer from a significant "calibration gap," where the confidence level of the generated text fails to reflect the underlying visual uncertainty. While a model might be internally uncertain about a low-resolution MRI artifact, the language model (LLM) backbone is architecturaly optimized to produce fluent, predictive sequences. This creates a "fluency mask" that results in the model forcing a coherent and authoritative explanation for a visually ambiguous or misinterpreted feature rather than signaling uncertainty through an appropriate "epistemic humility" [8]. This phenomenon is further exacerbated by the Reinforcement Learning from Human Feedback (RLHF) bias, a training process that often incentivizes MLLMs to be helpful and decisive at the cost of accuracy. Such optimization can lead to models that are "confidently incorrect" because they have been trained to avoid hedging or expressing doubt, which the reward models may interpret as a failure to provide a useful answer. In a high-risk surgical domain, this structural overconfidence manifests as a persuasive but erroneous recommendation that can successfully trigger the automation-induced deference error in a human expert

observed in our results. These findings confirm that accuracy parity does not imply safety equivalence, as the AI's lack of uncertainty calibration transforms it from a neutral decision support tool into a persuasive cognitive actor capable of reshaping expert judgment.

3.3. Implications for Clinical AI Deployment

These findings demonstrate that accuracy parity does not imply safety equivalence. Artificial intelligence systems capable of persuasive fluent output may introduce new failure modes by reshaping expert judgment rather than replacing it. Disagreement alone is insufficient as a safeguard, as parallel errors can occur without structured arbitration. A Parallel Review model is therefore proposed, in which artificial intelligence functions as a precision instrument whose outputs are reviewed alongside, rather than substituted for, human judgment [9,10]. Such a model preserves human accountability while leveraging artificial intelligence strengths in semantic consistency and data retrieval.

4. Conclusion

Artificial intelligence can perform at peer-level accuracy relative to senior surgical experts under constrained conditions. Its greater significance lies in its capacity to influence expert cognition. Automation bias transforms artificial intelligence from a neutral tool into a cognitive actor within clinical decision-making. Recognizing and mitigating this influence is essential for safe integration of artificial intelligence into high-risk medical workflows. The results support governance models that emphasize structured oversight, redundancy, and preservation of human responsibility.

Authors Note: Artificial Intelligence Use Disclosure

Artificial intelligence systems were used as an object of study and analysis in this research. Artificial intelligence tools were also used to improve clarity and organization of language during manuscript preparation. No factual content, data, or substantive arguments were generated by artificial intelligence, other than the artificial intelligence system evaluated as part of the study itself.

Disclaimer

This article is a scholarly analysis and does not provide medical or legal advice.

References

1. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121-127.
2. Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., ... & Pinto dos Santos, D. (2023). Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology, 307*(4), e222176.
3. Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., ... & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open, 7*(10), e2440969-e2440969.
4. Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., ... & Herzog, S. M. (2025). Human–AI collectives most accurately diagnose clinical vignettes. *Proceedings of the National Academy of Sciences, 122*(24), e2426153122.
5. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.*
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021, September). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop* (pp. 272-284). Cham: Springer International Publishing.
7. Wu, H., Tang, M., Zheng, X., & Jiang, H. (2025). When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552.*
8. Tong, B., Xia, J., Shang, S., & Zhou, K. (2025). Measuring Epistemic Humility in Multimodal Large Language Models. *arXiv preprint arXiv:2509.09658.*
9. Food, U. S. (2022). *Drug Administration (FDA). Clinical decision support software-Guidance for industry and food and drug administration staff.*
10. Food and Drug Administration. (2021). Artificial intelligence and machine learning (AI/ML) software as a medical device (SaMD) action plan. U.S. Department of Health and Human Services.