Original Paper

Hallucinated Evidence and Real-World Risk—Physicians' Intent to Act on AI-Generated Abstracts in Emergency Medicine and Trauma Surgery: Cross-Sectional, Vignette-Based Diagnostic Accuracy Study

Hassan Tarek Hakam[1,2,3], BSc, MD; Nicolas Vuillerme[2,4], Prof Dr; Felix Muehlensiepen[2,5,6], PhD; Nikolai Ramadanov[1,3], MD, PhD; Jonathan Lettner[1,3], MD; Marko Ostojic[7,8], MD, PhD; Roland Becker[1,3], Prof Dr Med; Robert Prill[1,3], PhD

[1]Center of Orthopaedics and Trauma Surgery, Universitätsklinikum Brandenburg an der Havel, Brandenburg an der Havel, Brandenburg, Germany

[2]Laboratoire d'Informatique de Grenoble (LIG), SANGRIA, National Center for Scientific Research (CNRS), Grenoble Institute of Technology, Université Grenoble Alpes, Grenoble, France

[3]Faculty of Health Sciences, Medizinische Hochschule Brandenburg Theodor Fontane, Brandenburg an der Havel, Brandenburg, Germany

[4]Institut Universitaire de France, Paris, Île-de-France, France

[5]Center for Health Services Research, Medizinische Hochschule Brandenburg Theodor Fontane, Rüdersdorf bei Berlin, Brandenburg, Germany

[6]Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charité - Universitätsmedizin Berlin, Berlin, Berlin, Germany

[7]Sports Traumatology Division, Traumatology Department, University Hospital, Zagreb, Croatia

[8]Osteon Orthopaedics and Sports Medicine Clinic, Mostar, Bosnia and Herzegovina

Corresponding Author:
Hassan Tarek Hakam, BSc, MD
Center of Orthopaedics and Trauma Surgery
Universitätsklinikum Brandenburg an der Havel
Hochstr. 29
Brandenburg an der Havel, Brandenburg, 14770
Germany
Phone: 1 015904797505
Email: hassantarek.hakam@mhb-fontane.de

Abstract

Background: Large language models (LLMs) are increasingly used in academic writing and can generate linguistically polished but factually incorrect scientific content. Such misinformation may distort evidence appraisal and, if translated into clinical practice, compromise patient care.

Objective: This study aimed to estimate physicians' likelihood of misperceiving false information embedded in LLM-generated abstracts as true and reporting an intent to use that information in future practice. We also explored which features were associated with perceived credibility and intended clinical action.

Methods: We conducted a cross-sectional, vignette-based survey of physicians working in emergency medicine and trauma surgery across 4 hospitals in Germany. Each participant assessed 4 abstracts generated using GPT-5 Thinking mode, of which 2 contained intentionally false claims supported by plausible explanations and 2 contained accurate claims. The primary outcome was the proportion of physicians reporting an intent to use or act on recommendations contained in abstracts with false claims. Secondary outcomes included perceived accuracy, rejection of accurate claims, and ratings of precision, structure, and grammar.

Results: Twenty physicians provided 80 ratings of 4 abstracts. Overall, physicians reported that they would act on 68% (54/80) of abstracts. Using abstract truth status as the reference, the decision to act had a sensitivity of 78% (31/40), specificity of 43% (17/40), positive predictive value of 57% (31/54), and negative predictive value of 65% (17/26). The positive likelihood ratio was 1.35 (95% CI 0.98-1.85), the negative likelihood ratio was 0.53 (95% CI 0.27-1.04), and the diagnostic odds ratio was 2.55 (95% CI 0.96-6.73). True and false abstracts received similar ratings for perceived accuracy (mean 3.15, SD 0.89 vs mean 2.88, SD 1.04; P=.21), precision (mean 3.15, SD 0.89 vs mean 2.93, SD 1.00; P=.29), structure (mean 3.23, SD 0.77 vs mean 3.20, SD 0.88; P=.89), and grammar (mean 3.40, SD 0.59 vs mean 3.43, SD 0.68; P=.86). Abstracts that physicians intended to act on were rated substantially higher for perceived accuracy, precision, and structure than those they would not act on (all P<.001), whereas grammar ratings did not differ (P=.80).

XSL•FO
RenderX

Conclusions: Physicians frequently reported an intent to act on LLM-generated abstracts containing false claims, and true and false abstracts were not reliably distinguished on perceived quality measures. Perceived accuracy, precision, and structure were strongly associated with intended action, whereas grammar was not. These findings highlight the risk that persuasive scientific presentation may increase the credibility and potential clinical influence of fabricated or inaccurate AI-generated evidence.

(J Med Internet Res 2026;28:e91095) doi: 10.2196/91095

KEYWORDS

large language models; misinformation; survey; content analysis; emergency medicine; clinical decision-making

Introduction

Large language models (LLMs) are neural networks trained with the simple but powerful objective of next-token prediction: given preceding context, they learn a probability distribution over candidate words (or subword tokens) and generate the most plausible continuation [1-3]. Through this autoregressive training on large text corpora—and, in newer systems, paired multimodal data—LLMs can compose, condense, and transform language and, in some settings, interface with allied models for automatic speech recognition, optical character recognition, and even handwriting transcription [4-8]. Beyond fluent text generation, these capabilities enable structured reasoning over documents, retrieval-augmented synthesis, and task automation [9-11].

In health care, early adoption has centered on clinical documentation, patient communication, decision support, and research assistance. Examples include drafting discharge summaries, summarizing extensive chart histories, retrieving and synthesizing evidence, and producing case-specific recommendations. Emergency medicine pilots further explore triage support, imaging protocol suggestions, and real-time protocol reminders [12-19].

Traumatology offers a particularly rich continuum of opportunities for human-in-the-loop LLM support. In the prehospital and emergency department (ED) phases, models could structure triage notes, surface red-flag checklists, and compress prior records for rapid handovers [12,17,20]. During trauma bay assessment, they may provide protocol reminders, role-based task lists, and time-stamped action logs aligned with Advanced Trauma Life Support (ATLS) principles [19,21-24]. For diagnostics, LLMs can suggest appropriate imaging bundles, convert free-text radiology reports into structured injury lists, and automatically detect discrepancies across documentation sources [25-27]. In the operative setting, LLMs might draft preoperative briefs, prepare implant checklists, and generate patient-facing consent documents [28-34]. During inpatient care and rehabilitation, LLMs could assist with daily progress notes, targeted consult requests, discharge instructions, and individualized recovery plans [35-42]. At the systems level, LLMs may help prefill trauma registries, prepare morbidity and mortality reviews, and generate adherence checklists for critical protocols [43-49].

However, the aforementioned benefits are inseparable from risks. LLM outputs can be confidently wrong—"hallucinated"—and may include fabricated facts or citations [50-54]. In clinical contexts, such errors can foster automation bias, degrade diagnostic reasoning, and undermine scientific integrity [52,55,56]. Prior evaluations show that human readers often struggle to distinguish AI-generated prose from human-written prose, blurring lines of acceptable use and intensifying concerns about overtrust [57-60]. Accordingly, any clinical deployment of LLMs must remain strictly human-supervised, be bound to local guidelines and provenance, protect patient data, and be prohibited from autonomous, high-risk decision-making [61-66]. Early clinical integrations should target low-risk, auditable tasks—documentation, summarization, and protocol reminders—where efficiency and usability can be measured without compromising safety.

Against this backdrop of rapidly expanding clinician and patient use of LLMs for medical communication despite well-documented hallucinations and overtrust of AI-generated health advice, a critical unanswered question is not simply whether LLMs can generate persuasive clinical narratives, but whether this persuasiveness can shift clinicians' diagnostic or management decisions when the underlying content is false [67]. This study was specifically designed to address that gap by evaluating physicians' willingness to act on LLM-generated misinformation. The primary end point is the proportion of physicians indicating an intent to base clinical decision-making on hallucinated, false evidence embedded in seemingly authoritative text. Secondary end points include physicians' perceived precision, accuracy, grammatical soundness, and structural quality of the same material. By linking perceived quality to intent to act, our work aims to quantify the diagnostic and decision-making hazards posed by LLM hallucinations and to inform guardrails, training, and deployment policies for safe use in trauma care and beyond.

Methods

Overview

Four research-style abstracts were generated using ChatGPT (GPT-5 Thinking mode; OpenAI) on May 1, 2026, following a standardized, prespecified prompt workflow. Prompts provided the clinical scenario and instructed the model to produce a structured abstract (Background, Methods, Results, and Conclusions; 200-300 words; neutral scientific tone; and no citations). Two abstracts were produced under a true or data-consistent instruction set, restricting content to the provided source information and explicitly prohibiting unsupported additions. Two abstracts were produced under a false or misleading instruction set that required an incorrect yet clinically plausible conclusion and the integration of 2 to 4 predefined false or misleading claims ("hallucinations"), without indicating within the text that deception had been introduced. Given the clinical context, these false claims were selected to represent

XSL•FO
RenderX

misinformation with potential for clinically meaningful harm if interpreted as actionable guidance. The complete generation protocol, including the stepwise table, full prompt texts, and condition-specific constraints, is provided in Multimedia Appendix 1.

The table in Multimedia Appendix 2 summarizes, for each abstract, (1) the scenario description, (2) the principal claims, (3) the conclusions presented, and (4) an adjudication of factuality (true vs false) with a brief rationale explaining why the statements are accurate or erroneous. Full, detailed versions of all 4 abstracts are provided in German in Multimedia Appendix 3. A translation of all 4 abstracts is provided in Multimedia Appendix 4.

The 4 abstracts were presented in randomized order within a standardized, paper-based questionnaire. Randomization was computer-generated without replacement to minimize order and carryover effects. The instrument did not disclose the origin (LLM vs human) or the factual status (true vs false) of any abstract.

After reading each abstract, participants completed quality ratings on a 4-point Likert-type scale (1="poor" and 4="excellent") across 4 prespecified domains—precision, factual accuracy, grammatical quality, and structural coherence—and provided an overall quality score. Subsequently, they answered a dichotomous clinical-intent item. To test the hypothesis that participants' intent to follow a recommendation is chiefly driven by their agreement with the abstract's conclusion, we elicited—immediately after each abstract—a 2-point agreement rating (1="disagree" and 2="agree") with the concluding recommendation.

We specifically asked clinicians to rate each abstract's accuracy, precision, grammatical quality, and structural coherence because work in health communication and online health information shows that perceived message accuracy and completeness are core components of credibility, which in turn is a strong predictor of intentions to follow recommended health behaviors [68-70]. At the same time, research on processing fluency demonstrates that texts that are easier to read, that is, grammatically correct, stylistically polished, and coherently structured, are more likely to be judged as true and to elicit greater confidence in subsequent decisions, even when their content is objectively false [71-73]. Finally, clinicians frequently rely on abstracts alone to inform diagnostic and treatment decisions, and variations in abstract reporting quality are known to affect the accuracy of those decisions [74-77]. Together, these literatures suggest that abstracts perceived as more accurate, precise, grammatically polished, and coherent will be seen as more credible and easier to process and are therefore more likely to increase clinicians' intent to act on the information they contain.

To enhance content validity, participants were explicitly encouraged to propose edits and corrections to the questionnaire items and abstracts. Free-text fields after each abstract and at the end of the survey invited rewording suggestions and the identification of ambiguities, typographical errors, and clinical inaccuracies. All comments were recorded verbatim and reviewed by 2 investigators; minor editorial corrections were implemented for subsequent iterations, while content-altering changes were deferred until completion of the primary analyses.

Participants were physicians working at primary, secondary, and tertiary trauma centers. Specifically, clinicians were recruited from the University Hospital Brandenburg, Charité – Universitätsmedizin Berlin (Campus Benjamin Franklin), the Ernst von Bergmann Clinic, and the Evangelical Hospital Ludwigsfelde in Germany.

Eligible participants were physicians who had completed at least 6 months of training in the ED and were directly involved in acute care, with emphasis on trauma surgery and emergency medicine. Physicians from other specialties (eg, internal medicine) were also eligible if they had participated in the acute management of patients with trauma during their ED training. The full inclusion and exclusion criteria are summarized in Textbox 1.

The primary outcome was the proportion of respondents indicating willingness to follow recommendations in abstracts that contained fabricated ("hallucinated") claims. Secondary outcomes were domain-specific and overall quality ratings for all abstracts. Per protocol, prespecified attention or comprehension checks were embedded in the questionnaire; responses failing these checks or having substantial missing data were excluded from analysis.

To minimize demand characteristics, participants were masked to the study's specific hypothesis and to the truth status of the abstracts. The cover explanation stated that the study evaluated the structural and grammatical quality of research abstracts and its association with physicians' stated willingness to follow the embedded recommendations. Immediately after completing the questionnaire, all participants were debriefed: we disclosed that some abstracts contained intentionally fabricated claims, clarified which recommendations were true vs false, and provided the correct information.

Analyses were performed in SPSS (version 29.0; IBM Corp). Given the preliminary nature of this study and the small sample size, analyses were primarily descriptive and exploratory. The full analysis set included eligible respondents who passed attention checks and completed at least 3 of 4 abstracts. For the binary clinical-intent outcome, we summarized responses using contingency tables and standard diagnostic performance measures, including sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratios, and diagnostic odds ratios, each with 95% CIs. For the 4-point rating outcomes, results are presented as mean (SD) values by group. Exploratory between-group comparisons were performed using 2-sample 1-tailed *t* tests. Because each physician rated multiple abstracts, these exploratory comparisons did not fully account for within-physician clustering and should therefore be interpreted cautiously as descriptive rather than confirmatory. Missing data were minimal and were handled by complete-case analysis. All reported *P* values are 2-sided and are presented for exploratory interpretation only.

Because this initial study included only the first 20 physician questionnaires, the present report should be interpreted as a preliminary, exploratory, hypothesis-generating analysis

XSL•FO
RenderX

intended to assess instrument feasibility, randomization performance, and the plausibility and direction of observed effects rather than to provide definitive hypothesis-confirming results or broadly generalizable estimates.

Textbox 1. Eligibility criteria for study participation.

Inclusion criteria

- Licensed physicians (residents, fellows, or attending physicians) with direct responsibilities in acute care (eg, emergency department [ED], trauma bay, operating room, or intensive care unit consultations)
- ≥6 months of cumulative experience in an interdisciplinary ED
- Documented or self-reported participation in the acute management of patients with trauma during training (eg, participation in an Advanced Trauma Life Support primary or secondary survey, resuscitation team role, or operative or consultative role)
- Current clinical practice of ≥20 clinical hours/week in acute care (ED, trauma or acute care surgery, anesthesia or critical care, surgery, or acute internal medicine)
- Recent exposure through direct involvement in ≥10 trauma activations or ≥12 ED or trauma shifts within the past 12 months
- Professional proficiency in German sufficient to read and appraise scientific abstracts
- Age ≥18 years, provision of informed consent, and agreement not to use external aids (search engines or large language models [LLMs]) while completing the questionnaire

Exclusion criteria

- Nonphysician personnel (eg, nurses, paramedics, physician assistants, or medical students) or personnel in nonclinical roles
- <6 months of cumulative ED experience
- No participation in acute trauma care during training
- Exclusively administrative, research-only, outpatient-only, or telemedicine-only roles without direct acute trauma involvement in the past 12 months
- No direct involvement in acute trauma care in the past 12 months
- Inability to read scientific German at a professional level
- Age <18 years, refusal to provide consent, or use of external aids during the questionnaire contrary to instructions
- Prior participation in pilot testing or prior exposure to the study abstracts or materials
- Failure to meet data quality safeguards, including failure of attention or comprehension checks, an incomplete questionnaire (>10% key items missing), duplicate submission, or an implausible completion time (below the prespecified threshold)
- Direct involvement in the development or marketing of the evaluated LLMs or a relationship with the study team that could bias responses

Ethical Considerations

This cross-sectional, vignette-based survey enrolled licensed physicians who completed a paper-based questionnaire. Participation was voluntary, limited to adults (aged ≥18 years), and based on written informed consent. Under applicable institutional or national regulations, anonymous survey research involving health care professionals and collecting no directly identifying or patient data does not require formal ethics committee review (78,79). No directly identifying personal data (eg, name, date of birth, or contact details) were collected; responses were recorded anonymously and are reported only in aggregated or anonymized form. To minimize expectancy and demand characteristics, participants were not informed during questionnaire completion of the specific study hypothesis or that some presented abstracts contained deliberate inaccuracies; immediately after participation, all respondents were debriefed, informed of the fabricated elements, and provided with corrected information. No patient data were collected or used. Study data were stored securely with access restricted to the research team and will be deleted or destroyed 5 years after completion of the final analyses and publication. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Results

Approximately 50 physicians were approached for participation. Only 20 (40%) returned completed questionnaires and were included in the final analysis. No returned questionnaires were excluded for failing attention or comprehension checks, having incomplete responses, or failing to meet the prespecified eligibility criteria.

Twenty physicians from several surgical and internal medicine specialties with varying levels of seniority participated in the study. Their specialties, training levels, and clinical and ED experience are summarized in Table 1. Given the small sample, all effect estimates should be interpreted cautiously as preliminary and primarily descriptive.

Each physician rated 4 research abstracts, 2 containing true information and 2 containing false information, resulting in an equal number of ratings for true and false abstracts. Overall, this provided a balanced dataset for comparing responses to correct and incorrect information (Tables 2-5).

XSL•FO
RenderX

Table 1. Characteristics of participating physicians (N=20).

Characteristics	Physicians
Primary specialty, n (%)[a]	
Trauma surgery	7 (35)
Internal medicine	5 (25)
General surgery	5 (25)
Orthopedic surgery	2 (10)
Intensive care	1 (5)
Level of training, n (%)	
Resident	9 (45)
Attending	8 (40)
Chief	3 (15)
Years of clinical experience, median (IQR)	7 (3.9-15.5)
Years of emergency department experience, median (IQR)	5 (2.8-10.5)

[a]Primary specialty was defined as the first specialty listed for each physician.

Table 2. Contingency table of truth status vs decision to act.

Decision to act[a]	True abstract (gt[b]=1)	False abstract (gt=0)
Would act (rec[c]=1)	31 (true positive)	23 (false positive)
Would not act (rec=0)	9 (false negative)	17 (true negative)

[a]"Would act" was derived by dichotomizing the original 4-point clinical-intent response as follows: "No" or "Probably would not follow"=would not act; "Probably would follow" or "Yes"=would act.

[b]gt: ground truth.

[c]rec: recommendation.

Table 3. Diagnostic performance of physicians' willingness to act on the abstract.

Measures	Counts, n/N	Estimate	95% CI
Sensitivity (act on true abstracts)	31/40	78%	65%-90%
Specificity (withhold on false abstracts)	17/40	43%	27%-58%
Positive predictive value	31/54	57%	44%-71%
Negative predictive value	17/26	65%	47%-84%
Positive likelihood ratio	—[a]	1.35	0.98-1.85
Negative likelihood ratio	—	0.53	0.27-1.04
Diagnostic odds ratio	—	2.55	0.96-6.73

[a]Not applicable.

XSL•FO
RenderX

Table 4. Perceived quality ratings by truth status of the abstract[a].

Parameters	False abstracts, mean (SD)	True abstracts, mean (SD)	*P* value[b]
Accuracy	2.88 (1.04)	3.15 (0.89)	.21
Precision	2.93 (1)	3.15 (0.89)	.29
Structure	3.2 (0.88)	3.23 (0.77)	.89
Grammar	3.43 (0.68)	3.4 (0.59)	.86

[a]Values are mean ratings on a 4-point scale.

[b]*P* values were calculated using 2-sample *t* tests without accounting for clustering.

Table 5. Perceived quality ratings by willingness to act on the abstract[a].

Parameters	Would not act, mean (SD)	Would act, mean (SD)	*P* value[b]
Accuracy	1.92 (0.63)	3.54 (0.61)	<.001
Precision	2.08 (0.80)	3.5 (0.61)	<.001
Structure	2.65 (1.06)	3.48 (0.50)	<.001
Grammar	3.38 (0.70)	3.43 (0.60)	.80

[a]Values are mean ratings on a 4-point scale.

[b]*P* values were calculated using 2-sample *t* tests without accounting for clustering.

For the analysis of diagnostic performance metrics, the 4-point clinical-intent item (“No,” “Probably would not follow,” “Probably would follow,” and “Yes”) was dichotomized a priori into “would not act” (“No” or “Probably would not follow”) and “would act” (“Probably would follow” or “Yes”). Using the true status of the abstract as the reference, physicians’ willingness to clinically act on the information showed greater sensitivity to true abstracts than to false ones.

In other words, physicians tended to act on many abstracts that were true (31/40, 78%), but also on a substantial proportion of those that were false (23/40, 58%), and their decision not to act did not reliably exclude false information.

Perceived accuracy, precision, structure, and grammatical soundness of the abstracts are presented by truth status in Table 4. Ratings for true abstracts were, on average, slightly higher than those for false abstracts for accuracy and precision, while structure and grammar were not significantly different. Formal comparisons did not show statistically significant differences between true and false abstracts on any of the rating dimensions.

Perceived quality ratings displayed a different pattern when abstracts were grouped by physicians’ willingness to act (Table 5). Abstracts that physicians indicated they would act on were consistently rated more favorably for accuracy, precision, and structure than those that they would not act on, whereas ratings of grammatical soundness did not differ significantly.

These findings suggest that physicians’ decisions to act are much more strongly aligned with their subjective quality judgments than with the actual truth status of the abstracts.

CIs were wide for several measures, reflecting limited precision (n=20); results should therefore be interpreted as hypothesis-generating and primarily informative about the direction and plausible magnitude of effects.

Discussion

In this preliminary sample, physicians frequently indicated a willingness to follow recommendations derived from AI-generated medical text even when the underlying content was intentionally false (“hallucinated”). This readiness to act coexisted with consistently favorable ratings of superficial quality—structure, grammar, and perceived precision or accuracy—which suggests that formal features of presentation may unduly influence trust and clinical intent. In other words, persuasive linguistic polish and coherent narrative scaffolding appeared to function as heuristic cues for credibility, potentially overriding the latent uncertainty inherent in LLM outputs. Notably, free-text responses were sparse and provided little insight into respondents’ reasoning, limiting our ability to identify the specific cues (eg, authoritative tone, internal consistency, or familiar phrasing) that triggered endorsement. Taken together, these results are consistent with automation bias dynamics: clinicians may overweight well-formed AI prose and underweight verification, creating a pathway by which hallucinated content could shape decisions. This highlights the need for explicit guardrails (eg, provenance and citation requirements, structured uncertainty displays, and local guideline binding), as well as training that disentangles rhetorical quality from evidentiary quality in AI-assisted clinical communication.

Our findings align with a growing body of work showing that clinicians modify decisions in response to AI guidance—and that the direction of change depends on the accuracy of the advice [80,81]. Prior evaluations of decision support and imaging AI report overall gains when recommendations are correct, but measurable decrements when they are incorrect, consistent with automation bias and anchoring [82]. Recent empirical work illustrates how LLMs can both shape clinicians’ decisions and simultaneously rely on weak or fabricated

XSL•FO
RenderX

evidence. In a randomized clinical trial of diagnostic reasoning, Goh et al [83] showed that GPT-4 produced polished differential diagnoses and management rationales that performed at least comparably to those of physicians using standard online resources, underscoring how authoritative such output can appear. A plausible mechanism is processing fluency, whereby text that is easy to read and internally coherent is processed with lower subjective effort, which can be misattributed to truthfulness and credibility (ie, an illusory truth or truth effect) even when objective accuracy is unchanged [84,85]. In a complementary randomized vignette study, Goh et al [86] found that exposure to ChatGPT's recommendations for patients with chest pain led clinicians to frequently revise their initial triage and treatment decisions, generally improving guideline concordance but clearly demonstrating that chatbot text can sway clinical judgment. At the same time, studies of LLM evidentiary reliability raise serious concerns: in an evaluation of migraine prophylaxis literature, Moskatel and Zhang [87] observed that ChatGPT often offered nuanced-sounding appraisals of drug efficacy while supplying predominantly hallucinated or incorrect references, and Chelli et al [88] likewise reported high hallucination rates and poor reference accuracy when ChatGPT and Bard (now Gemini) were asked to reconstruct systematic review bibliographies. Taken together, these findings suggest that clinicians may be influenced by the rhetorical authority of LLM-generated text even when the underlying evidentiary support is thin or partially fabricated, highlighting the need for careful verification and critical appraisal of chatbot outputs in clinical practice [83,86-88].

Building on these findings, our study adopted a complementary paradigm: we presented clinicians with LLM-style abstracts rather than point-of-care prompts, blinded them both to the provenance of the text (human vs model-generated) and to its veracity, and then quantified the "diagnostic accuracy" of their intent to act by measuring sensitivity to true vs false content. In contrast to prior work by Goh et al [83,86], which examined how ChatGPT's advice alters concrete diagnostic and management decisions in vignette-based tasks, our design isolates the rhetorical pull of polished language itself, independent of any explicit endorsement of the tool as a decision aid [87,88]. The central result—that coherent, well-written summaries reliably increased clinicians' willingness to act, even when key claims were false—converges with and helps explain the mixed performance seen in previous LLM decision support experiments: if clinicians are differentially swayed by linguistic fluency rather than evidentiary strength, then LLMs can appear beneficial in some scenarios (where their fluent output happens to align with guidelines) yet propagate error in others (where equally fluent but incorrect content is accepted at face value) [1,89-97].

At the same time, our design addresses gaps left by previous studies. First, we directly link perceived textual quality (precision, grammar, and structure) to intent to act, showing that rhetorical features can function as credibility heuristics independent of content validity [90,93,98]; most prior work has focused on answer correctness rather than the persuasive mechanics of the presentation itself [94-96]. Second, by seeding targeted hallucinations into plausible abstracts, we model a realistic failure mode for LLM-assisted documentation and literature synthesis—contexts in which clinicians routinely encounter long-form prose rather than single-turn suggestions [97,99,100]. Finally, our diagnostic accuracy framing (treating clinician behavior as the index test against the truth of the abstract) provides a transferable metric set—sensitivity, specificity, and balanced accuracy—for comparing interventions aimed at mitigating overtrust (eg, citations by default, uncertainty displays, or guideline binding) [96,101-104]. Taken together, these contributions help reconcile why LLMs can appear "useful" in aggregate yet still produce pockets of harmful influence when they are incorrect—an insight that should inform both deployment guardrails and future trials.

If left unchecked, this pattern could harm clinical practice in several concrete ways. First, persuasive but incorrect AI prose may anchor clinicians to unsafe actions (eg, universally targeting normotension in hemorrhagic trauma or deferring cranial computed tomography [cCT] in anticoagulated head injury), increasing the risk of missed injuries, delayed hemostasis, or preventable deterioration. Second, repeated exposure to fluent but unreliable summaries can erode evidence-based habits, normalizing citation-free assertions and encouraging cognitive off-loading in lieu of verification; over time, this may deskill clinicians in appraisal and differential diagnosis [105,106]. Third, system-level propagation of AI-generated text into notes, order sets, or handovers can amplify errors—a single hallucination can cascade across teams and shifts, inflating unnecessary resource use (imaging, consults, and admissions) or, conversely, delaying indicated interventions [107,108]. Fourth, because LLMs can encode and reproduce training-data biases, overtrust may widen disparities (eg, undertriage of older or marginalized patients if the narrative deemphasizes red flags common in those groups) [109-111]. Finally, reliance on unverified AI recommendations increases medicolegal exposure and undermines a culture of safety: when documentation appears authoritative, deviations from guidelines may go unnoticed until an adverse event occurs, at which point audit trails reveal that decisions were influenced by unverifiable sources [112-114].

The findings should be interpreted in light of the vignette-based design and the outcome assessed. Our primary end point reflects clinicians' stated intent to rely on or act on information presented in LLM-generated abstracts, which represents a clinically meaningful proximal indicator of downstream impact rather than a direct measure of bedside behavior. Although intent is a recognized predictor of subsequent action, translation into practice may be moderated by contextual factors such as time pressure, team oversight, workflow integration, accountability structures, and opportunities for verification. Accordingly, these results are best viewed as evidence that misleading LLM outputs can plausibly influence decision-relevant judgments under controlled conditions, while underscoring the need for follow-up studies using higher-fidelity approaches (eg, standardized patient encounters, simulation-based assessments, or embedded workflow evaluations) to directly observe behavioral effects and patient-relevant consequences.

This study has several strengths. We used clinically realistic, research-style abstracts in German, randomized their order, and masked participants to both provenance and truth status,

XSL•FO
RenderX

allowing us to isolate the persuasive effect of textual features on physicians' intent to act.

Data quality safeguards (eligibility screening, attention checks, and exclusion rules) further bolstered internal validity, and systematic debriefing upheld ethical transparency. Important limitations temper interpretation. The preliminary sample (n=20) is underpowered for precise estimates and subgroup inferences; the vignette design measures stated intent rather than observed bedside behavior; and sparse free-text responses restricted insight into the specific linguistic cues driving trust. As a pilot study, these findings should be viewed as hypothesis-generating, and replication with larger, more diverse samples is needed to support robust inference and improve generalizability. Generalizability may be limited by the single language and acute care context, by reliance on a single LLM version and prompt specification, and by potential mode effects from paper-based administration. Finally, although SPSS supported our core models, small cluster sizes and software constraints required analytic simplifications for some ordinal and resampling procedures, and unmeasured factors (eg, topic familiarity and local guideline awareness) could still influence responses.

Furthermore, this study has some limitations related to reporting standards. This study applied a diagnostic accuracy framework to physicians' judgment of LLM-generated abstracts; however, it does not fully adhere to the STARD (Standards for Reporting Diagnostic Accuracy Studies) 2015 reporting guideline. STARD was developed for patient-level clinical diagnostic test accuracy studies involving a defined target condition, patient flow, and real-world testing pathways, which are not present in a vignette-based judgment experiment. Several STARD items (eg, patient recruitment, disease severity distribution, test execution in clinical care, and patient flow diagrams) were therefore not applicable. Nevertheless, we adopted core STARD principles where conceptually appropriate, including an explicit definition of the index test and reference standard, prespecification of accuracy measures, and transparent reporting of sensitivity, specificity, and predictive values. Consequently, the reported diagnostic accuracy metrics should be interpreted as measures of judgment discrimination rather than clinical diagnostic performance. In addition, because we did not perform a formal postsurvey manipulation check to assess whether participants suspected AI authorship or the presence of intentionally false content during completion, we cannot quantify the extent to which such awareness may have influenced judgment criteria and the resulting discrimination metrics. Finally, because exploratory comparisons did not account for repeated ratings within physicians, inferential results should be interpreted cautiously.

Acknowledgments

This work is dedicated to the memory of Dr Mark Msallem, whose commitment to patient care and research continues to inspire the authors.

Data Avilability

The anonymized questionnaire data used in this study are stored in the office of the head of research of the Department of Orthopedics and Traumatology at the University Hospital of Brandenburg (RP). The data will be destroyed 5 years after completion of the final analyses and publication of the respective results.

Funding

The publication fees for this article were covered by the Library of Medizinische Hochschule Brandenburg CAMPUS GmbH.

Authors' Contributions

HTH is the primary author of this study and contributed to the study design, questionnaire development, and data collection. NV, RB, and RP supervised the project and contributed to the study design. FM, NR, and JL contributed to data collection. NR and MO performed the statistical analyses. RP supervised the preparation of the final manuscript and provided input on referencing and wording.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompts and constraints to generate the abstracts.
[DOCX File , 21 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Overview of abstract scenarios, principal claims, prompting approaches, conclusions, and adjudicated factuality with rationale.
[DOCX File , 37 KB-Multimedia Appendix 2]

XSL•FO
RenderX

Multimedia Appendix 3

German version of the questionnaire.
[DOCX File , 30 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Translation of the abstracts into English.
[DOCX File , 21 KB-Multimedia Appendix 4]

References

1. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med. Feb 2024;177(2):210-220. [doi: 10.7326/M23-2772] [Medline: 38285984]
3. Bradshaw TJ, Tie X, Warner J, Hu J, Li Q, Li X. Large language models and large multimodal models in medical imaging: a primer for physicians. J Nucl Med. Feb 03, 2025;66(2):173-182. [FREE Full text] [doi: 10.2967/jnumed.124.268072] [Medline: 39819692]
4. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst. Feb 17, 2024;48(1):22. [FREE Full text] [doi: 10.1007/s10916-024-02045-3] [Medline: 38366043]
5. Zolnoori M, Vergez S, Xu Z, Esmaeili E, Zolnour A, Anne Briggs K, et al. Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA Open. Dec 10, 2024;7(4):ooae130. [FREE Full text] [doi: 10.1093/jamiaopen/ooae130] [Medline: 39659993]
6. Lyu KM, Lyu RY, Chang HT. Real-time multilingual speech recognition and speaker diarization system based on whisper segmentation. PeerJ Comput Sci. Mar 29, 2024;10:e1973. [FREE Full text] [doi: 10.7717/peerj-cs.1973] [Medline: 38660177]
7. Ganesh S, Bhatti PT, Alkhalaf M, Gupta S, Shah AJ, Tridandapani S. Combining optical character recognition with paper ECG digitization. IEEE J Transl Eng Health Med. May 25, 2021;9:1900309. [FREE Full text] [doi: 10.1109/JTEHM.2021.3083482] [Medline: 34235006]
8. Lun KC. Pen-based computing: is it good for health? Technol Health Care. Jan 01, 1994;2(1):27-33. [doi: 10.3233/THC-1994-2103] [Medline: 25273804]
9. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y). Mar 01, 2024;5(3):100943. [FREE Full text] [doi: 10.1016/j.patter.2024.100943] [Medline: 38487804]
10. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. Apr 01, 2025;32(4):605-615. [FREE Full text] [doi: 10.1093/jamia/ocaf008] [Medline: 39812777]
11. Tripathi S, Sukumaran R, Cook TS. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. J Am Med Inform Assoc. May 20, 2024;31(6):1436-1440. [doi: 10.1093/jamia/ocad258] [Medline: 38273739]
12. Williams CY, Bains J, Tang T, Patel K, Lucas AN, Chen F, et al. Evaluating large language models for drafting emergency department discharge summaries. medRxiv. Preprint posted online on April 4, 2024. [FREE Full text] [doi: 10.1101/2024.04.03.24305088] [Medline: 38633805]
13. Tung JY, Gill SR, Sng GG, Lim DY, Ke Y, Tan TF, et al. Comparison of the quality of discharge letters written by large language models and junior clinicians: single-blinded study. J Med Internet Res. Jul 24, 2024;26:e57721. [FREE Full text] [doi: 10.2196/57721] [Medline: 39047282]
14. Liu S, Wright AP, Mccoy AB, Huang SS, Genkins JZ, Peterson JF, et al. Using large language model to guide patients to create efficient and comprehensive clinical care message. J Am Med Inform Assoc. Aug 01, 2024;31(8):1665-1670. [FREE Full text] [doi: 10.1093/jamia/ocae142] [Medline: 38917441]
15. Ozmen BB, Singh N, Shah K, Berber I, Singh D, Pinsky E, et al. Development of a novel artificial intelligence clinical decision support tool for hand surgery: HandRAG. J Hand Microsurg. Jun 11, 2025;17(4):100293. [FREE Full text] [doi: 10.1016/j.jham.2025.100293] [Medline: 40606653]
16. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res. Jan 12, 2024;26:e48996. [FREE Full text] [doi: 10.2196/48996] [Medline: 38214966]
17. Williams CY, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open. May 01, 2024;7(5):e248895. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.8895] [Medline: 38713466]

XSL•FO
RenderX

18. Hallinan JT, Leow NW, Ong W, Lee A, Low YX, Chan MD, et al. MRI spine request form enhancement and auto protocoling using a secure institutional large language model. Spine J. Mar 2025;25(3):505-514. [doi: 10.1016/j.spinee.2024.10.021] [Medline: 39536908]
19. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform. May 10, 2024;12:e53787. [FREE Full text] [doi: 10.2196/53787] [Medline: 38728687]
20. Rovati L, Privitera D, Finch AS, Litell JM, Brogan AM, Tekin A, et al. Development of an Emergency Department Safety Checklist through a global consensus process. Intern Emerg Med. Jun 2025;20(4):1203-1213. [doi: 10.1007/s11739-024-03760-y] [Medline: 39352462]
21. Mastrianni A, Kulp L, Mapelli E, Sarcevic A. Understanding digital checklist use through team communication. Ext Abstr Hum Factors Computing Syst. Apr 2020;2020. [FREE Full text] [doi: 10.1145/3334480.3382817] [Medline: 32747878]
22. Kulp L, Sarcevic A, Cheng M, Zheng Y, Burd RS. Comparing the effects of paper and digital checklists on team performance in time-critical work. In: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY. Association for Computing Machinery; 2019.
23. Montagna S, Croatti A, Ricci A, Agnoletti V, Albarello V, Gamberini E. Real-time tracking and documentation in trauma management. Health Informatics J. Mar 2020;26(1):328-341. [FREE Full text] [doi: 10.1177/1460458219825507] [Medline: 30726161]
24. Zargaran E, Spence R, Adolph L, Nicol A, Schuurman N, Navsaria P, et al. Association between real-time electronic injury surveillance applications and clinical documentation and data acquisition in a South African trauma center. JAMA Surg. May 16, 2018;153(5):e180087. [FREE Full text] [doi: 10.1001/jamasurg.2018.0087] [Medline: 29541765]
25. Kanemaru N, Yasaka K, Okimoto N, Sato M, Nomura T, Morita Y, et al. Efficacy of fine-tuned large language model in CT protocol assignment as clinical decision-supporting system. J Imaging Inform Med. Dec 2025;38(6):4336-4348. [doi: 10.1007/s10278-025-01433-6] [Medline: 39909993]
26. Woźnicki P, Laqua C, Fiku I, Hekalo A, Truhn D, Engelhardt S, et al. Automatic structuring of radiology reports with on-premise open-source large language models. Eur Radiol. Apr 2025;35(4):2018-2029. [doi: 10.1007/s00330-024-11074-y] [Medline: 39390261]
27. Sun C, Teichman K, Zhou Y, Critelli B, Nauheim D, Keir G, et al. Generative large language models trained for detecting errors in radiology reports. Radiology. May 2025;315(2):e242575. [doi: 10.1148/radiol.242575] [Medline: 40392090]
28. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. JAMA Surg. Aug 01, 2024;159(8):928-937. [doi: 10.1001/jamasurg.2024.1621] [Medline: 38837145]
29. Cheng T, Li Y, Gu J, He Y, He G, Zhou P, et al. The performance of ChatGPT in day surgery and pre-anesthesia risk assessment: a case-control study of 150 simulated patient presentations. Perioper Med (Lond). Nov 21, 2024;13(1):111. [FREE Full text] [doi: 10.1186/s13741-024-00469-6] [Medline: 39574189]
30. Alba C, Xue B, Abraham J, Kannampallil T, Lu C. The foundational capabilities of large language models in predicting postoperative risks using clinical notes. NPJ Digit Med. Feb 11, 2025;8(1):95. [FREE Full text] [doi: 10.1038/s41746-025-01489-2] [Medline: 39934379]
31. Tabuchi H, Ishitobi N, Deguchi H, Nakaniida Y, Tanaka H, Akada M, et al. Large-scale observational study of AI-based patient and surgical material verification system in ophthalmology: real-world evaluation in 37 529 cases. BMJ Qual Saf. Jun 19, 2025;34(7):433-442. [FREE Full text] [doi: 10.1136/bmjqs-2024-018018] [Medline: 39613452]
32. Deol ES, Henning G, Basourakos S, Vasdev RM, Sharma V, Kavoussi NL, et al. Artificial intelligence model for automated surgical instrument detection and counting: an experimental proof-of-concept study. Patient Saf Surg. Jul 21, 2024;18(1):24. [FREE Full text] [doi: 10.1186/s13037-024-00406-y] [Medline: 39034409]
33. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open. Oct 02, 2023;6(10):e2336997. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
34. Ali R, Connolly ID, Tang OY, Mirza FN, Johnston B, Abdulrazeq HF, et al. Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. NPJ Digit Med. Mar 08, 2024;7(1):63. [FREE Full text] [doi: 10.1038/s41746-024-01039-2] [Medline: 38459205]
35. Kernberg A, Gold JA, Mohan V. Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: comparative study. J Med Internet Res. Apr 22, 2024;26:e54419. [FREE Full text] [doi: 10.2196/54419] [Medline: 38648636]
36. Hartman V, Zhang X, Poddar R, McCarty M, Fortenko A, Sholle E, et al. Developing and evaluating large language model-generated emergency medicine handoff notes. JAMA Netw Open. Dec 02, 2024;7(12):e2448723. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.48723] [Medline: 39625719]
37. Small WR, Austrian J, O'Donnell L, Burk-Rafel J, Hochman KA, Goodman A, et al. Evaluating hospital course summarization by an electronic health record-based large language model. JAMA Netw Open. Aug 01, 2025;8(8):e2526339. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.26339] [Medline: 40802185]

XSL•FO
RenderX

38. Wada A, Tanaka Y, Nishizawa M, Yamamoto A, Akashi T, Hagiwara A, et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. NPJ Digit Med. Jul 02, 2025;8(1):395. [FREE Full text] [doi: 10.1038/s41746-025-01802-z] [Medline: 40604147]
39. Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med. May 09, 2025;8(1):263. [FREE Full text] [doi: 10.1038/s41746-025-01684-1] [Medline: 40346344]
40. Huang T, Safranek C, Socrates V, Chartash D, Wright D, Dilip M, et al. Patient-representing population's perceptions of GPT-generated versus standard emergency department discharge instructions: randomized blind survey assessment. J Med Internet Res. Aug 02, 2024;26:e60336. [FREE Full text] [doi: 10.2196/60336] [Medline: 39094112]
41. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). Oct 29, 2024;11:1477898. [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
42. Qiang S, Zhang H, Liao Y, Zhang Y, Gu Y, Wang Y, et al. Application of large language models in stroke rehabilitation health education: 2-phase study. J Med Internet Res. Jul 22, 2025;27:e73226. [FREE Full text] [doi: 10.2196/73226] [Medline: 40694436]
43. Hersh WR, Leen TK, Rehfuss PS, Malveau S. Automatic prediction of trauma registry procedure codes from emergency room dictations. Stud Health Technol Inform. 1998;52 Pt 1:665-669. [Medline: 10384538]
44. Choi J, Chen Y, Sivura A, Vendrow EB, Wang J, Spain DA. TraumaICD bidirectional encoder representation from transformers: a natural language processing algorithm to extract injury international classification of diseases, 10th edition diagnosis code from free text. Ann Surg. Jul 01, 2024;280(1):150-155. [doi: 10.1097/SLA.0000000000006107] [Medline: 37753654]
45. Tignanelli CJ, Silverman GM, Lindemann EA, Trembley AL, Gipson JC, Beilman G, et al. Natural language processing of prehospital emergency medical services trauma records allows for automated characterization of treatment appropriateness. J Trauma Acute Care Surg. May 2020;88(5):607-614. [FREE Full text] [doi: 10.1097/TA.0000000000002598] [Medline: 31977990]
46. Kobritz M, Patel V, Rindskopf D, Demyan L, Jarrett M, Coppa G, et al. Practice-based learning and improvement: improving morbidity and mortality review using natural language processing. J Surg Res. Mar 2023;283:351-356. [doi: 10.1016/j.jss.2022.10.075] [Medline: 36427445]
47. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005;12(4):448-457. [FREE Full text] [doi: 10.1197/jamia.M1794] [Medline: 15802475]
48. Fast D, Adams LC, Busch F, Fallon C, Huppertz M, Siepmann R, et al. Autonomous medical evaluation for guideline adherence of large language models. NPJ Digit Med. Dec 12, 2024;7(1):358. [FREE Full text] [doi: 10.1038/s41746-024-01356-6] [Medline: 39668168]
49. Srinivasan A, Berkowitz J, Friedrich NA, Kivelson S, Tatonetti NP. Large language model analysis of reporting quality of randomized clinical trial articles: a systematic review. JAMA Netw Open. Aug 01, 2025;8(8):e2529418. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.29418] [Medline: 40875232]
50. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus. Sep 05, 2023;15(9):e44720. [FREE Full text] [doi: 10.7759/cureus.44720] [Medline: 37809168]
51. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. Oct 18, 2024;11:e57400. [FREE Full text] [doi: 10.2196/57400] [Medline: 39423368]
52. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus. May 19, 2023;15(5):e39238. [FREE Full text] [doi: 10.7759/cureus.39238] [Medline: 37337480]
53. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. Sep 07, 2023;13(1):14045. [FREE Full text] [doi: 10.1038/s41598-023-41032-5] [Medline: 37679503]
54. Emsley R. ChatGPT: these are not hallucinations - they're fabrications and falsifications. Schizophrenia (Heidelb). Aug 19, 2023;9(1):52. [FREE Full text] [doi: 10.1038/s41537-023-00379-4] [Medline: 37598184]
55. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. May 2023;307(4):e222176. [doi: 10.1148/radiol.222176] [Medline: 37129490]
56. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA. Dec 19, 2023;330(23):2275-2284. [FREE Full text] [doi: 10.1001/jama.2023.22295] [Medline: 38112814]
57. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med. Apr 26, 2023;6(1):75. [FREE Full text] [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
58. Makiev KG, Asimakidou M, Vasios IS, Keskinis A, Petkidis G, Tilkeridis K, et al. A study on distinguishing ChatGPT-generated and human-written orthopaedic abstracts by reviewers: decoding the discrepancies. Cureus. Nov 21, 2023;15(11):e49166. [FREE Full text] [doi: 10.7759/cureus.49166] [Medline: 38130535]

XSL•FO
RenderX

59. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 01, 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
60. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology. Apr 2023;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
61. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ. Apr 01, 2020;98(4):251-256. [FREE Full text] [doi: 10.2471/BLT.19.237487] [Medline: 32284648]
62. Baldassarre A, Padovan M. Regulatory and ethical considerations on artificial intelligence for occupational medicine. Med Lav. Apr 24, 2024;115(2):e2024013. [FREE Full text] [doi: 10.23749/mdl.v115i2.15881] [Medline: 38686573]
63. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci. Mar 15, 2024;19(1):27. [FREE Full text] [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]
64. Denecke K, May R, LLMHealthGroup, Rivera Romero O. Potential of large language models in health care: Delphi study. J Med Internet Res. May 13, 2024;26:e52399. [FREE Full text] [doi: 10.2196/52399] [Medline: 38739445]
65. Good machine learning practice for medical device development: guiding principles. U.S. Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles [accessed 2025-10-15]
66. Mello MM, Cohen IG. Regulation of health and health care artificial intelligence. JAMA. May 27, 2025;333(20):1769-1770. [doi: 10.1001/jama.2025.3308] [Medline: 40095599]
67. Rosen KL, Sui M, Heydari K, Enichen EJ, Kvedar JC. The perils of politeness: how large language models may amplify medical misinformation. NPJ Digit Med. Nov 06, 2025;8(1):644. [FREE Full text] [doi: 10.1038/s41746-025-02135-7] [Medline: 41198821]
68. Alotaibi KA. Perceived credibility of public health campaigns and its impact on infection control behaviors: mediating roles of health literacy and motivation. J Multidiscip Healthc. Jun 03, 2025;18:3153-3163. [FREE Full text] [doi: 10.2147/JMDH.S520357] [Medline: 40486256]
69. Sbaffi L, Rowley J. Trust and credibility in web-based health information: a review and agenda for future research. J Med Internet Res. Jun 19, 2017;19(6):e218. [FREE Full text] [doi: 10.2196/jmir.7579] [Medline: 28630033]
70. Zhao Y, Zhang L, Zeng C, Chen Y, Lu W, Song N. Factors influencing online health information credibility: a meta-analysis. Aslib J Inf Manag. Mar 14, 2024;77(4):701-732. [doi: 10.1108/AJIM-05-2023-0155]
71. Wang WC, Brashier NM, Wing EA, Marsh EJ, Cabeza R. On known unknowns: fluency and the neural mechanisms of illusory truth. J Cogn Neurosci. May 2016;28(5):739-746. [FREE Full text] [doi: 10.1162/jocn_a_00923] [Medline: 26765947]
72. Henderson EL, Simons DJ, Barr DJ. The trajectory of truth: a longitudinal study of the illusory truth effect. J Cogn. Jun 08, 2021;4(1):29. [FREE Full text] [doi: 10.5334/joc.161] [Medline: 34164597]
73. Reber R, Unkelbach C. The epistemic status of processing fluency as source for judgments of truth. Rev Philos Psychol. Dec 2010;1(4):563-581. [FREE Full text] [doi: 10.1007/s13164-010-0039-7] [Medline: 22558063]
74. Barry HC, Ebell MH, Shaughnessy AF, Slawson DC, Nietzke F. Family physicians' use of medical abstracts to guide decision making: style or substance? J Am Board Fam Pract. 2001;14(6):437-442. [FREE Full text] [Medline: 11757886]
75. Marcelo A, Gavino A, Isip-Tan IT, Apostol-Nicodemus L, Mesa-Gaerlan FJ, Firaza PN, et al. A comparison of the accuracy of clinical decisions based on full-text articles and on journal abstracts alone: a study among residents in a tertiary care hospital. Evid Based Med. Apr 2013;18(2):48-53. [FREE Full text] [doi: 10.1136/eb-2012-100537] [Medline: 22782923]
76. Nascimento DP, Ostelo RW, van Tulder MW, Gonzalez GZ, Araujo AC, Vanin AA, et al. Do not make clinical decisions based on abstracts of healthcare research: a systematic review. J Clin Epidemiol. Jul 2021;135:136-157. [doi: 10.1016/j.jclinepi.2021.03.030] [Medline: 33839242]
77. Fontelo P. Can journal abstracts alone be used for clinical decision making? Rapid response to: innovations in publishing BMJ research. BMJ. Jun 11, 2009. URL: https://www.bmj.com/rapid-response/2011/11/02/can-journal-abstracts-alone-be-used-clinical-decision-making [accessed 2026-09-15]
78. Universitätsklinikum Leipzig. Ethics Committee. Waiver – Guidance on the requirement for ethics committee consultation. Leipzig University Hospital URL: https://www.uniklinikum-leipzig.de/einrichtungen/ethik-kommission/Seiten/Waiver.aspx?utm
79. Ludwig-Maximilians-Universität München, Faculty of Medicine. Ethics Committee. Guidance on studies not requiring mandatory ethics consultation. LMU Munich URL: https://www.med.lmu.de/de/fakultaet/wer-wir-sind/gremien-kommissionen/ethikkommission/wissenswertes/?utm
80. Rezazade Mehrizi MH, Mol F, Peter M, Ranschaert E, Dos Santos DP, Shahidi R, et al. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. Sci Rep. Jun 07, 2023;13(1):9230. [FREE Full text] [doi: 10.1038/s41598-023-36435-3] [Medline: 37286665]
81. Golchin K, Roudsari A. Study of the effects of Clinical Decision Support System's incorrect advice and clinical case difficulty on users' decision making accuracy. Stud Health Technol Inform. 2011;164:13-16. [Medline: 21335681]

XSL•FO
RenderX

82. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med. Feb 19, 2021;4(1):31. [FREE Full text] [doi: 10.1038/s41746-021-00385-9] [Medline: 33608629]
83. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. Oct 01, 2024;7(10):e2440969. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
84. Reber R, Schwarz N. Effects of perceptual fluency on judgments of truth. Conscious Cogn. Sep 1999;8(3):338-342. [doi: 10.1006/ccog.1999.0386] [Medline: 10487787]
85. Hasher L, Goldstein D, Toppino T. Frequency and the conference of referential validity. J Verbal Learn Verbal Behav. Feb 1977;16(1):107-112. [doi: 10.1016/S0022-5371(77)80012-1]
86. Goh E, Bunning B, Khoong E, Gallo R, Milstein A, Centola D, et al. ChatGPT influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes. medRxiv. Preprint posted online on November 27, 2023. [FREE Full text] [doi: 10.1101/2023.11.24.23298844] [Medline: 38076944]
87. Moskatel LS, Zhang N. The utility of ChatGPT in the assessment of literature on the prevention of migraine: an observational, qualitative study. Front Neurol. Aug 17, 2023;14:1225223. [FREE Full text] [doi: 10.3389/fneur.2023.1225223] [Medline: 37662036]
88. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. J Med Internet Res. May 22, 2024;26:e53164. [FREE Full text] [doi: 10.2196/53164] [Medline: 38776130]
89. Dechêne A, Stahl C, Hansen J, Wänke M. The truth about the truth: a meta-analytic review of the truth effect. Pers Soc Psychol Rev. May 2010;14(2):238-257. [doi: 10.1177/1088868309352251] [Medline: 20023210]
90. Alter AL, Oppenheimer DM. Uniting the tribes of fluency to form a metacognitive nation. Pers Soc Psychol Rev. Aug 2009;13(3):219-235. [doi: 10.1177/1088868309341564] [Medline: 19638628]
91. Fazio LK, Rand DG, Pennycook G. Repetition increases perceived truth equally for plausible and implausible statements. Psychon Bull Rev. Oct 2019;26(5):1705-1710. [doi: 10.3758/s13423-019-01651-4] [Medline: 31420808]
92. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med. Mar 22, 2025;8(1):175. [FREE Full text] [doi: 10.1038/s41746-025-01543-z] [Medline: 40121370]
93. Ketron S. Investigating the effect of quality of grammar and mechanics (QGAM) in online reviews: the mediating role of reviewer crediblity. J Bus Res. Dec 2017;81:51-59. [doi: 10.1016/j.jbusres.2017.08.008]
94. Urda-Cîmpean AE, Leucu a DC, Drugan C, Du u AG, Călinici T, Drugan T. Assessing the accuracy of diagnostic capabilities of large language models. Diagnostics (Basel). Jun 29, 2025;15(13):1657. [FREE Full text] [doi: 10.3390/diagnostics15131657] [Medline: 40647657]
95. Aththanagoda AK, Kulathilake KA, Abdullah NA. Precision and personalization: how large language models redefining diagnostic accuracy in personalized medicine - a systematic literature review. IEEE J Biomed Health Inform (Forthcoming). Jun 30, 2025. [doi: 10.1109/JBHI.2025.3584179] [Medline: 40587361]
96. Tran VT, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, et al. Sensitivity, specificity and avoidable workload of using a large language models for title and abstract screening in systematic reviews and meta-analyses. medRxiv. Preprint posted online on December 17, 2023. [FREE Full text] [doi: 10.1101/2023.12.15.23300018]
97. Zhu Z, Zhang Y, Zhuang X, Zhang F, Wan Z, Chen Y, et al. Can we trust AI doctors? A survey of medical hallucination in large language and large vision-language models. In: Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg, PA. Association for Computational Linguistics; 2025.
98. Okuhara T, Ishikawa H, Ueno H, Okada H, Kato M, Kiuchi T. Influence of high versus low readability level of written health information on self-efficacy: a randomized controlled study of the processing fluency effect. Health Psychol Open. Feb 12, 2020;7(1):2055102920905627. [FREE Full text] [doi: 10.1177/2055102920905627] [Medline: 32110424]
99. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel). Mar 10, 2025;13(6):603. [FREE Full text] [doi: 10.3390/healthcare13060603] [Medline: 40150453]
100. Tilton AK, Caplan BE, Cole BJ. Generative AI in consumer health: leveraging large language models for health literacy and clinical safety with a digital health framework. Front Digit Health. Aug 26, 2025;7:1616488. [FREE Full text] [doi: 10.3389/fdgth.2025.1616488] [Medline: 40933812]
101. Kostick-Quenet K, Lang BH, Smith J, Hurley M, Blumenthal-Barby J. Trust criteria for artificial intelligence in health: normative and epistemic considerations. J Med Ethics. Jul 23, 2024;50(8):544-551. [doi: 10.1136/jme-2023-109338] [Medline: 37979976]
102. Lindenmeyer A, Blattmann M, Franke S, Neumuth T, Schneider D. Towards trustworthy AI in healthcare: epistemic uncertainty estimation for clinical decision support. J Pers Med. Jan 31, 2025;15(2):58. [FREE Full text] [doi: 10.3390/jpm15020058] [Medline: 39997335]
103. Pardoux É, Kerasidou A. Compliance with clinical guidelines and AI-based clinical decision support systems: implications for ethics and trust. Sci Eng Ethics. Nov 13, 2025;31(6):34. [doi: 10.1007/s11948-025-00562-z] [Medline: 41231280]

XSL•FO
RenderX

104. Qazi IA, Ali A, Khawaja AU, Akhtar MJ, Sheikh AZ, Alizai MH. Automation bias in large language model assisted diagnostic reasoning among AI-trained physicians. medRxiv. Preprint posted online on August 26, 2025. [FREE Full text] [doi: 10.1101/2025.08.23.25334280]
105. Natali C, Marconi L, Dias Duran LD, Cabitza F. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. Artif Intell Rev. Aug 27, 2025;58:356. [doi: 10.1007/s10462-025-11352-1]
106. Khan FA. AI in clinical diagnostics: is overreliance eroding clinical expertise? PLOS Digit Health. Aug 4, 2025;4(8):e0000959. [FREE Full text] [doi: 10.1371/journal.pdig.0000959] [Medline: 40758655]
107. Al Bahrani B, Medhi I. Copy-pasting in patients' electronic medical records (EMRs): use judiciously and with caution. Cureus. Jun 15, 2023;15(6):e40486. [FREE Full text] [doi: 10.7759/cureus.40486] [Medline: 37461761]
108. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA. Dec 19, 2023;330(23):2255-2257. [doi: 10.1001/jama.2023.22557] [Medline: 38112824]
109. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns (N Y). Oct 08, 2021;2(10):100347. [FREE Full text] [doi: 10.1016/j.patter.2021.100347] [Medline: 34693373]
110. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open. Dec 01, 2023;6(12):e2345050. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]
111. Agarwal R, Bjarnadottir M, Rhue L, Dugas M, Crowley K, Clark J, et al. Addressing algorithmic bias and the perpetuation of health inequities: an AI bias aware framework. Health Policy Technol. Mar 2023;12(1):100702. [doi: 10.1016/j.hlpt.2022.100702]
112. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med. Jan 31, 2024;124(7):287-290. [FREE Full text] [doi: 10.1515/jom-2023-0229] [Medline: 38295300]
113. Terranova C, Cestonaro C, Fava L, Cinquetti A. AI and professional liability assessment in healthcare. A revolution in legal medicine? Front Med (Lausanne). Jan 8, 2024;10:1337335. [FREE Full text] [doi: 10.3389/fmed.2023.1337335] [Medline: 38259835]
114. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q. Sep 2021;99(3):629-647. [FREE Full text] [doi: 10.1111/1468-0009.12504] [Medline: 33822422]

Abbreviations

ATLS: Advanced Trauma Life Support
cCT: cranial computed tomography
ED: emergency department
LLM: large language model
STARD: Standards for Reporting Diagnostic Accuracy Studies

Edited by Y Hong; submitted 08.Jan.2026; peer-reviewed by Y Zhang, Y Wang; comments to author 27.Jan.2026; revised version received 26.Apr.2026; accepted 25.May.2026; published 13.Nov.2026

Please cite as:
Hakam HT, Vuillerme N, Muehlensiepen F, Ramadanov N, Lettner J, Ostojic M, Becker R, Prill R
Hallucinated Evidence and Real-World Risk—Physicians' Intent to Act on AI-Generated Abstracts in Emergency Medicine and Trauma Surgery: Cross-Sectional, Vignette-Based Diagnostic Accuracy Study
J Med Internet Res 2026;28:e91095
URL: https://www.jmir.org/2026/1/e91095
doi: 10.2196/91095
PMID:

©Hassan Tarek Hakam, Nicolas Vuillerme, Felix Muehlensiepen, Nikolai Ramadanov, Jonathan Lettner, Marko Ostojic, Roland Becker, Robert Prill. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX