Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Advice Overrides Reasoning: Diagnostic Deference to AI- and Faculty-Labelled Recommendations in Undergraduate Dental Education

Dusan Surdilovic

surdilovic74@gmail.com

RAK College of Dental Sciences, RAK Medical and Health Sciences University

Raghavendra Manjunath Shetty

Ajman University

Sabrin Ali Azim

Al Ain University

Fadia Awadalkreem

RAK College of Dental Sciences, RAK Medical and Health Sciences University

B. S.M.S. Siriwardena

RAK College of Dental Sciences, RAK Medical and Health Sciences University

Research Article

Keywords: dental education, clinical reasoning, artificial intelligence, automation bias, diagnostic accuracy, undergraduate dental students, decision support, cognitive forcing

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10208256/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

AI-supported recommendations are increasingly discussed in dental education, but it remains unclear whether students maintain independent diagnostic judgment when external advice conflicts with their own decisions. This study examined diagnostic deference to AI-labelled and faculty-labelled advice among undergraduate dental students.

Methods

In a repeated-measures vignette study, 80 fourth- and fifth-year dental students completed 16 standardized pediatric and preventive dentistry cases, generating 1,280 trial-level observations. For each case, students gave an initial answer and confidence rating, reviewed an accurate or inaccurate externally labelled recommendation, and then gave a final answer. The primary outcome was harmful diagnostic deference: an initially correct answer becoming incorrect after inaccurate advice. The main secondary outcome was beneficial correction. Adjusted analyses used logistic regression with two-way cluster-robust standard errors clustered by participant and case.

Results

Overall accuracy increased from 57.9% to 66.4% after advice. Harmful diagnostic deference occurred in 112/374 eligible trials (29.9%), whereas beneficial correction occurred in 230/284 eligible trials (81.0%). Harmful deference was similar for AI-labelled and faculty-labelled advice (29.2% versus 30.5%), and adjusted models showed no clear association between advice label and either outcome.

Conclusions

External advice improved diagnostic accuracy when accurate, but inaccurate advice reversed nearly one-third of initially correct decisions. The findings do not indicate a uniquely harmful AI-label effect. Dental curricula should train students to verify and justify whether clinical recommendations should be accepted or challenged.

Introduction

Artificial intelligence (AI), particularly large language model (LLM)-based generative AI, is rapidly entering dental education. El-Hakim et al. mapped applications, challenges, and gaps in AI use across dental education [1]. International survey data show substantial educator interest in AI chatbots and LLMs, while also identifying the need for clearer implementation guidance and training [2]. A guidance-oriented review of generative AI integration emphasized structured curricular use, governance, and faculty

preparation [3]. Reviews focused on ChatGPT, large language models, and multimodal foundation models similarly describe educational opportunities while warning against unstructured adoption [4]. Large-language-model reviews also highlight applications in feedback, case simulation, and learning support [5]. Foundation-model discussions emphasize the need for careful curricular integration and oversight [6]. The emerging consensus is therefore not that AI should be excluded, but that students and faculty need structured approaches for using AI-supported recommendations without weakening clinical judgment.

In clinical-stage undergraduate education, students are expected to move beyond factual recall toward independent diagnostic reasoning, justification of treatment decisions, and recognition of uncertainty. AI tools may support these goals through rapid explanations, alternative hypotheses, formative feedback, and simulated cases [3]. However, the same properties that make AI-generated advice attractive–fluency, immediacy, and apparent authority–may also make incorrect advice difficult for students to challenge. This concern connects to automation bias, the tendency to over-rely on automated recommendations in a way that reduces independent verification or vigilance [7]. Verification complexity may further increase the risk that users accept automated recommendations without sufficient checking [8]. In healthcare, AI-driven clinical decision support may introduce harm when users accept incorrect recommendations or fail to detect system limitations [9]. Recent empirical work likewise shows that users can be vulnerable to false agreement with erroneous AI recommendations [10].

Despite a growing literature on AI in dentistry, an important gap remains. Much of this literature focuses on applications, perceived usefulness, educator attitudes, student readiness, technical performance of LLMs, or broad ethical and curricular implications. Broad dental-education reviews have mapped these applications and gaps [1]. Educator survey work has described readiness and implementation concerns [2]. ChatGPT-focused reviews have emphasized educational applications and limitations [4]. LLM-focused reviews have similarly examined prospects and challenges in dental education [5]. Fewer studies directly test what students do when their own clinical judgment conflicts with externally presented advice. This distinction matters because the educational risk is not merely that AI may provide an inaccurate answer; the more consequential risk is that a learner may abandon an initially correct decision after exposure to incorrect advice that appears credible.

The present study addresses this gap by examining diagnostic deference to externally labelled advice among undergraduate dental students. Using a repeated-measures, vignette-based design, students first made independent clinical decisions across standardized pediatric and preventive dentistry cases, then reviewed recommendations labelled either AI-generated or faculty-generated, and finally made a post-advice decision. The primary outcome was harmful diagnostic deference, defined as changing from an initially correct answer to an incorrect final answer after exposure to inaccurate advice. Beneficial correction, defined as changing from an initially incorrect answer to a correct final answer after accurate advice, was the main secondary outcome, capturing the dual effect of external advice. Three questions guided the study: to what extent does external advice change diagnostic decisions; how frequently does inaccurate advice produce harmful deference among initially correct students; and are deference

patterns different when the same decision-support context is labelled AI-generated rather than faculty-generated?

Methods

Study design

This study used a repeated-measures, vignette-based design with manipulated advice labels and advice accuracy to examine diagnostic deference to externally labelled clinical advice among undergraduate dental students. It measured whether students maintained or changed their initial clinical decisions after exposure to a recommendation labelled either as AI-generated or faculty-generated. Clinical vignettes were used because they allow standardized presentation of identical clinical information across participants and are suitable for examining clinical judgment under controlled conditions [11]. Prior validation work also supports the utility of vignettes as a method for measuring quality and decision-making in comparison with other assessment approaches [12]. Reporting followed relevant STROBE principles for transparent description of participant selection, variables, outcome definitions, bias control, study size, and statistical methods [13].

Setting and participants

The study was conducted among undergraduate dental students at Gulf Medical University. Eligible participants were clinical-stage undergraduate dental students who had received prior instruction in pediatric dentistry, preventive dentistry, dental trauma, pulp therapy, and behavior guidance. Participants were grouped analytically as fourth-year and fifth-year undergraduate dental students. Participation was voluntary and did not affect course grades, clinical assessment, or academic progression. A total of 80 undergraduate dental students were included in the final cleaned analysis dataset. Each participant completed 16 standardized case-based decision tasks, yielding 1,280 trial-level observations, where each trial represented one participant's response to one clinical vignette.

Allocation and sequence versions

Participants completed one of four pre-specified sequence versions, which distributed case order and advice-accuracy conditions across the instrument. Each participant completed either the AI-labelled or the faculty-labelled source condition, yielding 37 students in the AI-labelled condition and 43 in the faculty-labelled condition. Baseline comparability of the source-label groups was assessed descriptively using undergraduate level, sequence version, initial diagnostic accuracy, initial confidence, and advice-accuracy distribution. Because a fully auditable individual-level randomization record matching the final analytic sample was not available, source-label comparisons were interpreted as adjusted between-group comparisons rather than definitive causal estimates of the effect of the source label.

Ethical approval

Ethical approval was obtained from the institutional research ethics committee at Gulf Medical University before data collection. The study involved educational decision-making tasks using standardized hypothetical cases and did not involve patients, clinical interventions, radiographic exposure, or collection of identifiable patient data. Participants provided informed consent before participation. Responses were coded using participant identifiers, and the analysis dataset did not contain directly identifiable student information. Participants were informed that the study concerned clinical reasoning and advice-supported decision-making; after completion, they were debriefed that some recommendations had intentionally been accurate and others inaccurate to allow measurement of diagnostic deference.

Case-vignette instrument

The instrument consisted of 16 standardized pediatric and preventive dentistry case vignettes covering four domains: caries risk and preventive decision-making; pulp therapy and deep caries management; dental trauma and emergency reasoning; and behavior guidance/communication. Each vignette contained a brief clinical scenario, a focused decision question, one gold-standard answer, and one plausible incorrect answer. Caries-risk and preventive cases were aligned with American Academy of Pediatric Dentistry guidance on caries-risk assessment and management [14]. Pulp-therapy cases were aligned with AAPD guidance for primary and immature permanent teeth [15]. Behavior-guidance cases were aligned with AAPD best-practice guidance for pediatric dental patients [16]. Trauma-related cases were aligned with International Association of Dental Traumatology guidance for traumatic dental injuries, including the distinct management principles for primary and permanent dentitions [17].

Instrument validation

Before administration, the vignette bank and advice stimuli were reviewed by three faculty experts in pediatric and preventive dentistry. Experts independently assessed each vignette for clinical plausibility, clarity, alignment with accepted pediatric dentistry principles, suitability for undergraduate learners, and defensibility of the gold-standard and plausible-wrong options. Case validity was rated on a 1-4 scale. Cases were locked only when the mean validity rating was at least 3.50/4.00 and no expert flagged a serious clinical error; disagreements or weaker ratings were addressed through wording revision and re-review before the case bank was locked. All 16 cases met the predefined validation threshold. The full case bank, advice stimuli, and validation summary are provided as Additional file 1.

Advice stimuli

For each case, two types of external recommendation label were used: AI-labelled advice and faculty-labelled advice. Correct advice supported the gold-standard answer, and inaccurate advice supported the plausible incorrect answer. Advice statements were written in a concise, standardized format to reduce variation in length, tone, and persuasive force across cases. The study therefore tested deference to externally labelled advice under controlled conditions and was not designed as a diagnostic performance evaluation of a live AI system.

Procedure

For each case, participants first read the vignette and selected an initial clinical decision, then rated their confidence on a 1-5 ordinal scale, with higher values indicating greater confidence. They were then shown the assigned advice stimulus—labelled either AI-generated or faculty-generated and containing either accurate or inaccurate advice according to the predefined case standard—and selected a final answer and rated final confidence on the same 1-5 scale. A rationale score was recorded on a 1-5 ordinal scale as a descriptive summary of the quality of the clinical justification; because rationale scoring was not supported by documented inter-rater reliability data, it was not used as a covariate in the primary adjusted models.

Variables and coding

The primary unit of analysis was the trial-level observation, defined as one participant's response to one case. The final dataset contained participant identifier, case identifier, undergraduate level, initial answer, initial correctness, initial confidence, recommendation source, advice accuracy, final answer, final correctness, final confidence, rationale score, sequence version, and group allocation. Undergraduate level was coded as fourth-year or fifth-year undergraduate dental student. Initial and final answers were coded as gold-standard or plausible wrong, and initial and final correctness were recalculated directly from the answer fields. Rationale score was retained for descriptive reporting only.

Outcome definitions

The primary outcome was harmful diagnostic deference, defined as a trial in which the participant initially selected the gold-standard answer, was then exposed to inaccurate advice, and subsequently selected the plausible wrong answer as the final response; eligibility was restricted to trials in which the initial answer was correct and the advice was inaccurate. The main secondary outcome was beneficial correction, defined as a trial in which the participant initially selected the plausible wrong answer, was exposed to correct advice, and subsequently selected the gold-standard answer; eligibility was restricted to trials in which the initial answer was incorrect and the advice was correct. Additional secondary outcomes included initial accuracy, final accuracy, strict answer switching, confidence change, and rationale score.

Data cleaning and quality control

Initial and final correctness were recalculated from the raw answer fields, and harmful diagnostic deference and beneficial correction were then recalculated using predefined logical rules. Manually entered outcome flags were not used as primary analytic variables when they conflicted with recalculated logical definitions. Completeness was assessed at participant and case level, and participants were retained only if they had complete responses for all 16 cases. Duplicate participant-case rows were checked. The final cleaned dataset included 80 participants and 1,280 complete trial-level observations.

Statistical analysis

Descriptive statistics summarized participant distribution, initial accuracy, final accuracy, strict switching, harmful diagnostic deference, beneficial correction, confidence change, and rationale scores. Categorical variables were summarized as frequencies and percentages; continuous or ordinal variables as means and standard deviations or medians and interquartile ranges, depending on distribution. Because each participant completed multiple cases and each case was completed by multiple participants, trial-level observations were not treated as independent. Correlated trial-level responses were addressed using a regression framework consistent with longitudinal and clustered-data principles [18]. The handling of repeated responses followed established approaches for correlated educational and clinical data [19]. Two-way clustering by participant and case was used because both learners and vignettes contributed repeated observations, following multiway cluster-robust inference principles [20].

For the primary outcome, the dependent variable was harmful diagnostic deference among eligible trials, and the model included recommendation source, undergraduate level, initial confidence, and sequence version (harmful diagnostic deference ~ recommendation source + undergraduate level + initial confidence + sequence version), with standard errors clustered by participant and case. The analogous model for beneficial correction included the same covariates and clustering. Adjusted odds ratios, 95% confidence intervals, and Wald p values were reported. Rationale score was summarized descriptively but excluded from adjusted models because it was not supported by documented inter-rater reliability and could over-adjust for reasoning quality that is closely tied to the response itself.

Results

Study sample and dataset completeness

The final cleaned dataset included 80 undergraduate dental students and 1,280 complete trial-level observations (16 cases per participant), with complete responses for all 16 cases and no duplicate participant-case combinations. Of the 80 students, 44 (55.0%) were fourth-year and 36 (45.0%) were

fifth-year; 37 (46.3%) completed AI-labelled advice conditions and 43 (53.8%) completed faculty-labelled conditions. Trial-level characteristics are summarized in Table 1.

Table 1. Participant and trial-level characteristics

Characteristic	Value
Participants	80
Trial-level observations	1,280
Cases per participant	16
Fourth-year undergraduate students	44 (55.0%)
Fifth-year undergraduate students	36 (45.0%)
AI-labelled advice group	37 participants; 592 trials
Faculty-labelled advice group	43 participants; 688 trials
Correct advice trials	651 (50.9%)
Inaccurate advice trials	629 (49.1%)
Sequence versions	V1: 25; V2: 16; V3: 19; V4: 20 participants

Note. Percentages are calculated using participant-level or trial-level denominators as appropriate.

Descriptive balance checks showed similar initial diagnostic accuracy in AI-labelled and faculty-labelled conditions (56.8% versus 58.9%), similar initial confidence (mean 3.78 versus 3.74), and a similar proportion of correct-advice trials (50.7% versus 51.0%). Sequence-version distribution was also broadly balanced. Undergraduate level was less evenly distributed between source-label groups, with fifth-year students comprising 21 of 37 AI-labelled participants and 15 of 43 faculty-labelled participants; undergraduate level was therefore retained as an adjusted covariate in the inferential models.

Overall diagnostic accuracy and answer switching

Initial diagnostic accuracy was 57.9% (741/1,280) and rose to 66.4% (850/1,280) after exposure to external advice, an absolute increase of 8.5 percentage points. Strict answer switching occurred in 27.4% of trials (351/1,280): 230 trials moved from an initially incorrect to a final correct answer and 121 from an initially correct to a final incorrect answer, while the remaining trials were unchanged (Figure 1).

Primary outcome: harmful diagnostic deference

Of 374 trials eligible for harmful diagnostic deference, 112 showed the outcome, a rate of 29.9% (95% CI 25.5-34.8%). Approximately three in ten initially correct decisions therefore became incorrect after

exposure to inaccurate external advice.

Secondary outcome: beneficial correction

Of 284 trials eligible for beneficial correction, 230 showed the outcome, a rate of 81.0% (95% CI 76.0-85.1%), indicating that correct advice frequently—though not always—helped students revise initially incorrect decisions.

AI-labelled versus faculty-labelled advice

Harmful diagnostic deference was descriptively similar for AI-labelled and faculty-labelled advice (50/171, 29.2% versus 62/203, 30.5%), and beneficial correction was also broadly comparable across advice labels (Table 2). These data do not support a strong descriptive separation between AI-labelled and faculty-labelled advice as sources of deference in this dataset (Figure 2).

Table 2. Harmful diagnostic deference and beneficial correction by advice source and student level

Outcome / subgroup	Numerator / eligible denominator	Rate
Harmful deference, overall	112 / 374	29.9%
Harmful deference, AI-labelled advice	50 / 171	29.2%
Harmful deference, faculty-labelled advice	62 / 203	30.5%
Harmful deference, fourth-year students	66 / 200	33.0%
Harmful deference, fifth-year students	46 / 174	26.4%
Beneficial correction, overall	230 / 284	81.0%
Beneficial correction, AI-labelled advice	113 / 135	83.7%
Beneficial correction, faculty-labelled advice	117 / 149	78.5%
Beneficial correction, fourth-year students	133 / 173	76.9%
Beneficial correction, fifth-year students	97 / 111	87.4%

Note. Harmful deference was calculated only among trials with initially correct answers followed by inaccurate advice. Beneficial correction was calculated only among trials with initially incorrect answers followed by correct advice.

Adjusted models

In adjusted logistic regression models with two-way cluster-robust standard errors, advice label was not associated with harmful diagnostic deference (AI-labelled versus faculty-labelled: adjusted OR 1.05, 95% CI 0.67-1.65; p = 0.836) or with beneficial correction (adjusted OR 1.29, 95% CI 0.59-2.81; p = 0.517). Higher initial confidence showed a directionally protective but non-significant association with harmful deference (adjusted OR 0.57 per 1-point increase, 95% CI 0.30-1.07; p = 0.080), and fifth-year status was not significantly associated with either outcome. Full estimates are reported in Table 3.

Table 3. Adjusted logistic regression models with two-way cluster-robust standard errors

Outcome	Predictor	Reference	Adj. OR	95% CI	p value
Harmful diagnostic deference	AI-labelled advice	Faculty-labelled advice	1.05	0.67-1.65	0.836
Harmful diagnostic deference	5th-year undergraduate	4th-year undergraduate	0.85	0.44-1.64	0.624
Harmful diagnostic deference	Initial confidence	per 1-point increase	0.57	0.30-1.07	0.080
Beneficial correction	AI-labelled advice	Faculty-labelled advice	1.29	0.59-2.81	0.517
Beneficial correction	5th-year undergraduate	4th-year undergraduate	1.66	0.68-4.06	0.264
Beneficial correction	Initial confidence	per 1-point increase	1.07	0.71-1.61	0.760

Note. Models were estimated on eligible trial sets only. Standard errors were two-way cluster-robust, clustered by participant and case, and models adjusted for sequence version. Rationale score was excluded from adjusted models and summarized descriptively only.

Undergraduate level

Fifth-year students showed higher descriptive initial accuracy (62.0% versus 54.5%) and final accuracy (69.9% versus 63.5%) than fourth-year students, but undergraduate level was not statistically significant in the adjusted models.

Confidence and rationale scores

Mean confidence increased from 3.75 (SD 0.90) before advice to 4.00 (SD 0.89) after advice (mean change +0.25, SD 0.53), with similar increases for AI-labelled (+0.25, SD 0.54) and faculty-labelled (+0.24, SD 0.52) trials. The mean rationale score was 3.52 (SD 1.23); because it was recorded as a descriptive

ordinal summary and no inter-rater reliability estimate was available, it was not used for inferential claims.

Case-level variability

Case-level change was heterogeneous. Initial accuracy ranged from 47.5% to 72.5% and final accuracy from 46.3% to 82.5% across cases; among eligible trials, case-level harmful deference ranged from 5.9% to 62.5% and beneficial correction from 47.6% to 100.0%. This variability indicates that deference depended on the clinical content and plausibility of the advice stimulus, not only on the participant (Figure 3).

Discussion

Principal findings

This study examined whether undergraduate dental students maintained or changed diagnostic decisions after exposure to externally labelled clinical advice. The finding was double-edged: advice improved overall accuracy, but inaccurate advice produced measurable harmful deference. Final accuracy rose from 57.9% to 66.4%, and beneficial correction occurred in 81.0% of eligible trials; however, harmful diagnostic deference occurred in 29.9% of eligible trials, meaning that roughly one-third of initially correct decisions became incorrect after inaccurate advice. Adjusted analyses showed no clear difference between AI-labelled and faculty-labelled advice for either outcome. The study should therefore not be read as evidence that AI-labelled advice is uniquely dangerous, but as evidence that credible-looking external advice can move learners in both productive and harmful directions.

Comparison with previous literature

These findings align with the automation-bias literature. Goddard et al. showed that decision-support systems can improve performance yet generate new errors when users accept incorrect recommendations without sufficient verification [7]. Lyell and Coiera further emphasized that verification complexity can make such errors more likely when users must evaluate complex automated advice [8]. Clinical AI commentaries have similarly warned that assistive AI can create harm when incorrect recommendations are accepted as authoritative [9]. Empirical work on AI-decision support also shows that users may agree with erroneous AI recommendations despite their inaccuracy [10]. Recent dental education work places this finding in a broader curricular context. El-Hakim et al. described the expanding range of AI applications in dental education [1]. Uribe et al. showed that dental educators are already engaging with AI chatbots and LLMs but need clearer implementation structures [2]. Uribe, Maldupa, and Schwendicke argued for structured integration of generative AI rather than informal adoption [3]. ChatGPT-focused dental-education reviews describe potential educational uses but also limitations requiring oversight [4]. LLM-focused reviews similarly emphasize the need to balance

educational opportunity with critical appraisal [5]. Foundation-model discussions reinforce the need for faculty guidance and governance [6]. In the broader medical education literature, Sokol et al. argued that AI should support rather than replace clinical reasoning [21]. Schwartzstein similarly questioned whether AI can genuinely substitute for human reasoning in clinical education [22]. Izquierdo-Condoy et al. warned that poorly integrated generative AI may undermine cognitive autonomy [23]. The present study adds behavioural evidence to this discussion by showing that students changed decisions after advice exposure even when their initial answer was correct. The absence of a clear AI-specific effect is notable: it shifts the focus from a simple warning against AI toward a broader problem in which learners defer to authority cues, whether attached to an AI system or to a faculty label.

Interpretation

Deference appeared to vary by learner level and case content, but these patterns should be interpreted cautiously. Fifth-year students showed descriptively higher initial and final accuracy than fourth-year students, but undergraduate level was not statistically significant in the adjusted models. Case-level change was also substantial, with some vignettes showing large accuracy gains after advice and others showing deterioration, indicating that deference depended on the clinical content, ambiguity, and plausibility of the advice stimulus and not only on the learner. Higher initial confidence was directionally protective against harmful deference, although the estimate was imprecise.

Educational implications

The practical implication is that students should use advice through an explicit verification routine rather than as a shortcut to an answer. A defensible sequence is to make an independent decision, state confidence, identify the supporting evidence, review the recommendation, and then justify accepting or rejecting it. This sequence is consistent with cognitive forcing strategies in clinical decision-making [24]. It is also consistent with experimental evidence that cognitive forcing functions can reduce overreliance on incorrect AI recommendations [25]. AI literacy in dental education should therefore extend beyond prompt writing, privacy, and academic integrity to include repeated practice in checking both AI-labelled and faculty-labelled advice against clinical findings, guidelines, and the student's own rationale. Faculty feedback can follow the same pattern, asking students to justify why a recommendation should be accepted or rejected rather than only confirming whether an answer is correct.

Strengths and limitations

This study has several strengths. It moved beyond perception-based research by measuring actual decision change after exposure to external advice. The repeated-measures design generated 1,280 trial-level observations from 80 undergraduate dental students, allowing each participant to contribute multiple standardized diagnostic decisions. The study used predefined outcome definitions, a faculty-

reviewed vignette bank, recalculated logical outcomes, and inferential analyses that accounted for clustering by both participant and case. The distinction between harmful diagnostic deference and beneficial correction also allowed the dual effect of advice to be examined, rather than treating advice use as uniformly positive or negative.

Several limitations should be considered. The study was conducted at a single institution, which limits generalizability to other dental schools, curricula, and learner populations. Clinical vignettes provide experimental control and standardization for studying clinical decision-making [11]. However, vignettes cannot fully reproduce real patient encounters, time pressure, communication demands, clinical supervision, or the consequences of diagnostic error, and prior validation work shows that different assessment methods capture overlapping but not identical aspects of clinical quality [12]. The study tested advice labels and standardized advice statements rather than a live AI system; therefore, the findings should be interpreted as evidence about deference to externally labelled recommendations, not as an evaluation of real-world AI diagnostic performance.

Although source-label and sequence versions were pre-specified, a fully auditable individual-level randomization record matching the final analytic sample was not available. AI-versus-faculty comparisons should therefore be interpreted as adjusted between-group comparisons rather than definitive causal estimates of the effect of the source label, and residual confounding cannot be excluded. Undergraduate level was less evenly distributed between source-label groups, although it was retained as an adjusted covariate in the inferential models. Response options were simplified into gold-standard and plausible-wrong choices, which supported clean outcome definitions but did not capture the full complexity of clinical reasoning. Finally, rationale score was retained only descriptively because inter-rater reliability data were unavailable.

Future research should test this paradigm in larger, multi-institutional samples with fully documented individual-level randomized or crossover designs. Further studies should examine moderators such as prior AI exposure, confidence calibration, case ambiguity, and baseline diagnostic ability, and should evaluate whether structured verification prompts can reduce harmful deference while preserving beneficial correction.

Conclusion

External advice can both strengthen and distort undergraduate dental students' diagnostic decisions. Correct advice substantially improved initially incorrect answers, whereas inaccurate advice produced harmful diagnostic deference in approximately one-third of eligible trials, with no clear AI-specific effect. The educational task is therefore to train students not merely to use AI or faculty recommendations, but to verify, justify, and challenge external advice before allowing it to override their own clinical reasoning.

Abbreviations

AI: Artificial intelligence; BMC: BioMed Central; CI: confidence interval; LLM: large language model; OR: odds ratio; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate

Ethical approval was obtained from the institutional research ethics committee at Gulf Medical University before data collection. The study was conducted in accordance with the Declaration of Helsinki. All participants provided informed consent before participation.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. The de-identified trial-level dataset and analysis outputs are provided as Additional file 2.

Competing interests

The authors declare that they have no competing interests.

Funding

No specific funding was received for this work.

Authors' contributions

DS conceptualised the study, developed the study design and outcome definitions, supervised data cleaning and analysis, interpreted the findings, and drafted and revised the manuscript. RMS, SAA, FA, and BSMS contributed to clinical content review, interpretation of the findings, and critical revision of the manuscript for intellectual content. All authors read and approved the final manuscript.

Acknowledgements

AI-assisted tools were used to support language editing, formatting, and manuscript preparation. Eligibility decisions, analysis, interpretation, and final responsibility for the manuscript remained with the human authors. No AI tool is listed as an author.

Authors' information

Not applicable.

References

1. El-Hakim M, Anthonappa R, Fawzy A. Artificial Intelligence in Dental Education: A Scoping Review of Applications, Challenges, and Gaps. Dentistry J. 2025;13(9):384. 10.3390/dj13090384.
2. Uribe SE, Maldupa I, Kavadella A, El Tantawi M, Chaurasia A, Fontana M, et al. Artificial intelligence chatbots and large language models in dental education: Worldwide survey of educators. Eur J Dent Educ. 2024;28(4):865–76. 10.1111/eje.13009.
3. Uribe SE, Maldupa I, Schwendicke F. Integrating Generative AI in Dental Education: A Scoping Review of Current Practices and Recommendations. Eur J Dent Educ. 2025;29(2):341–55. 10.1111/eje.13074.
4. Aura-Tormos JI, Llacer-Martinez M, Torres-Osca I. Educational Applications of ChatGPT in University-Based Dental Education. Syst Rev Eur J Dent Educ. 2026;30(2):644–60. 10.1111/eje.70011.
5. Kong M, Fok EHW, Yiu CKY. A Scoping Review of Large Language Models in Dental Education: Applications, Challenges, and Prospects. Int Dent J. 2025;75(6):103854. 10.1016/j.identj.2025.103854.
6. Claman D, Sezgin E. Artificial Intelligence in Dental Education: Opportunities and Challenges of Large Language Models and Multimodal Foundation Models. JMIR Med Educ. 2024;10:e52346. 10.2196/52346.
7. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121–7. 10.1136/amiajnl-2011-000089.
8. Lyell D, Coiera E. Automation bias and verification complexity: A systematic review. J Am Med Inform Assoc. 2017;24(2):423–31. 10.1093/jamia/ocw105.
9. Khera R, Simon MA, Ross JS. Automation Bias and Assistive AI: Risk of Harm From AI-Driven Clinical Decision Support. JAMA. 2023;330(23):2255–7. 10.1001/jama.2023.22557.
10. Kucking F, Hubner U, Przysucha M. Automation Bias in AI-Decision Support: Results from an Empirical Study. Stud Health Technol Inform. 2024;316:1124–8. 10.3233/SHTI240871.
11. Evans SC, Roberts MC, Keeley JW, Blossom JB, Amaro CM, Garcia AM, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J

Clin Health Psychol. 2015;15(2):160–70. 10.1016/j.ijchp.2014.12.001.

12. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of three methods for measuring quality. JAMA. 2000;283(13):1715–22. 10.1001/jama.283.13.1715.
13. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, for the STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology statement: Guidelines for reporting observational studies. Ann Intern Med. 2007;147(8):573–7. 10.7326/0003-4819-147-8-200710160-00010.
14. American Academy of Pediatric Dentistry. Caries-risk assessment and management for infants, children, and adolescents. In: The Reference Manual of Pediatric Dentistry. Chicago, IL: American Academy of Pediatric Dentistry; 2025–2026.
15. American Academy of Pediatric Dentistry. Pulp therapy for primary and immature permanent teeth. In: The Reference Manual of Pediatric Dentistry. Chicago, IL: American Academy of Pediatric Dentistry; 2025–2026.
16. American Academy of Pediatric Dentistry. Behavior guidance for the pediatric dental patient. In: The Reference Manual of Pediatric Dentistry. Chicago, IL: American Academy of Pediatric Dentistry; 2025–2026.
17. Day PF, Flores MT, O'Connell AC, Abbott PV, Tsilingaridis G, Fouad AF, Cohenca N, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 3. Injuries in the primary dentition. Dent Traumatol. 2020;36(4):343–59. 10.1111/edt.12576.
18. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22. 10.1093/biomet/73.1.13.
19. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: An orientation. Am J Epidemiol. 2003;157(4):364–75. 10.1093/aje/kwf215.
20. Cameron AC, Gelbach JB, Miller DL. Robust inference with multiway clustering. J Bus Economic Stat. 2011;29(2):238–49. 10.1198/jbes.2010.07136.
21. Sokol K, Fackler J, Vogt JE. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. NPJ Digit Med. 2025;8:345. 10.1038/s41746-025-01725-9.
22. Schwartzstein RM. Clinical reasoning and artificial intelligence: Can AI really think? Trans Am Clin Climatol Assoc. 2024;134:133–45.
23. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: Enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res. 2025;27:e76340. 10.2196/76340.
24. Croskerry P. Cognitive forcing strategies in clinical decisionmaking. Ann Emerg Med. 2003;41(1):110–20. 10.1067/mem.2003.22.

25. Bucinca Z, Malaya MB, Gajos KZ. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW1):188. 10.1145/3449287

Figures

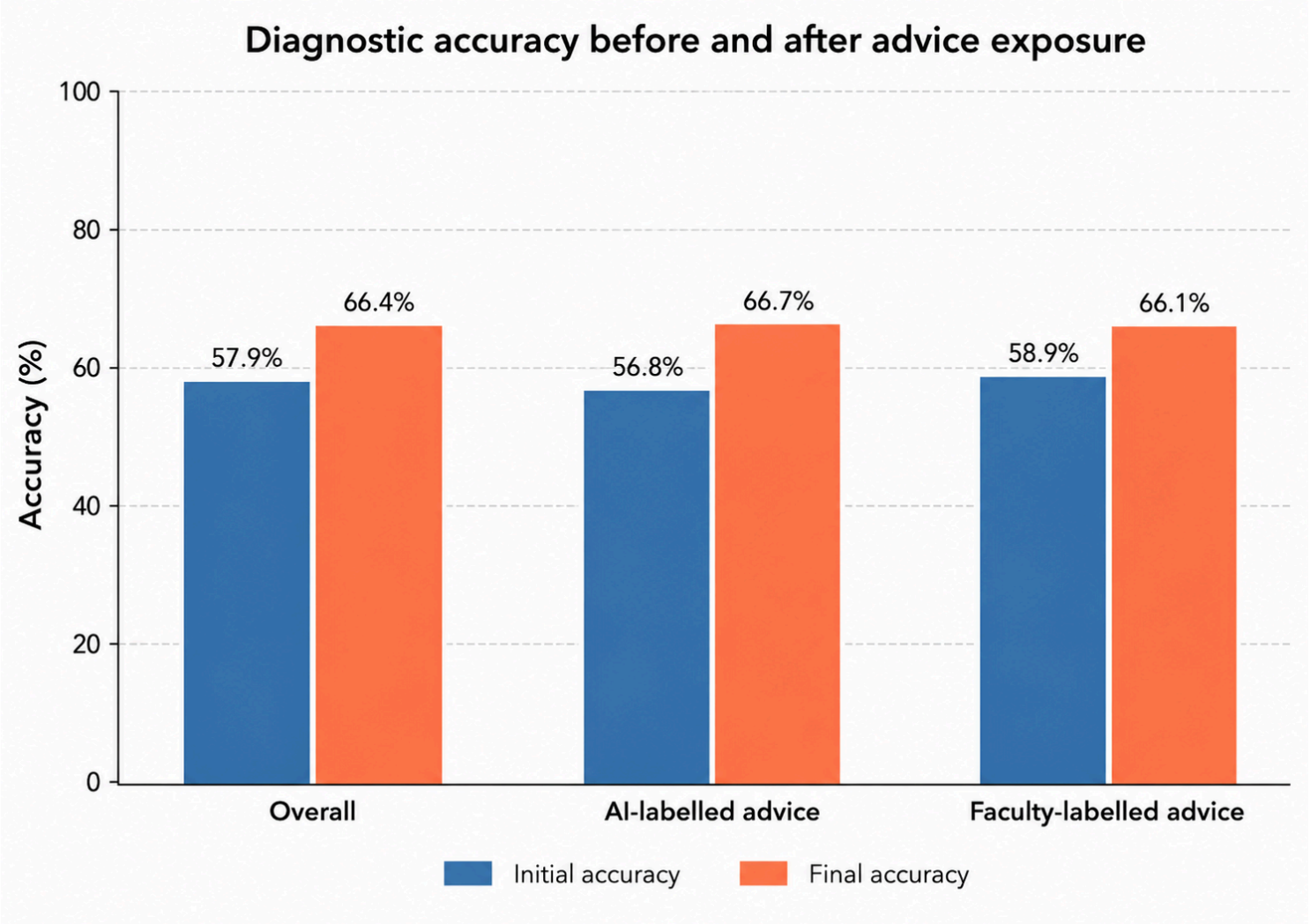


Figure 1

Diagnostic accuracy before and after advice exposure. Bars show initial and final diagnostic accuracy overall and according to advice label. Overall accuracy increased from 57.9% to 66.4%; in AI-labelled advice trials from 56.8% to 66.7%; and in faculty-labelled advice trials from 58.9% to 66.1%. Values are expressed as percentages of all trial-level observations within each subgroup.

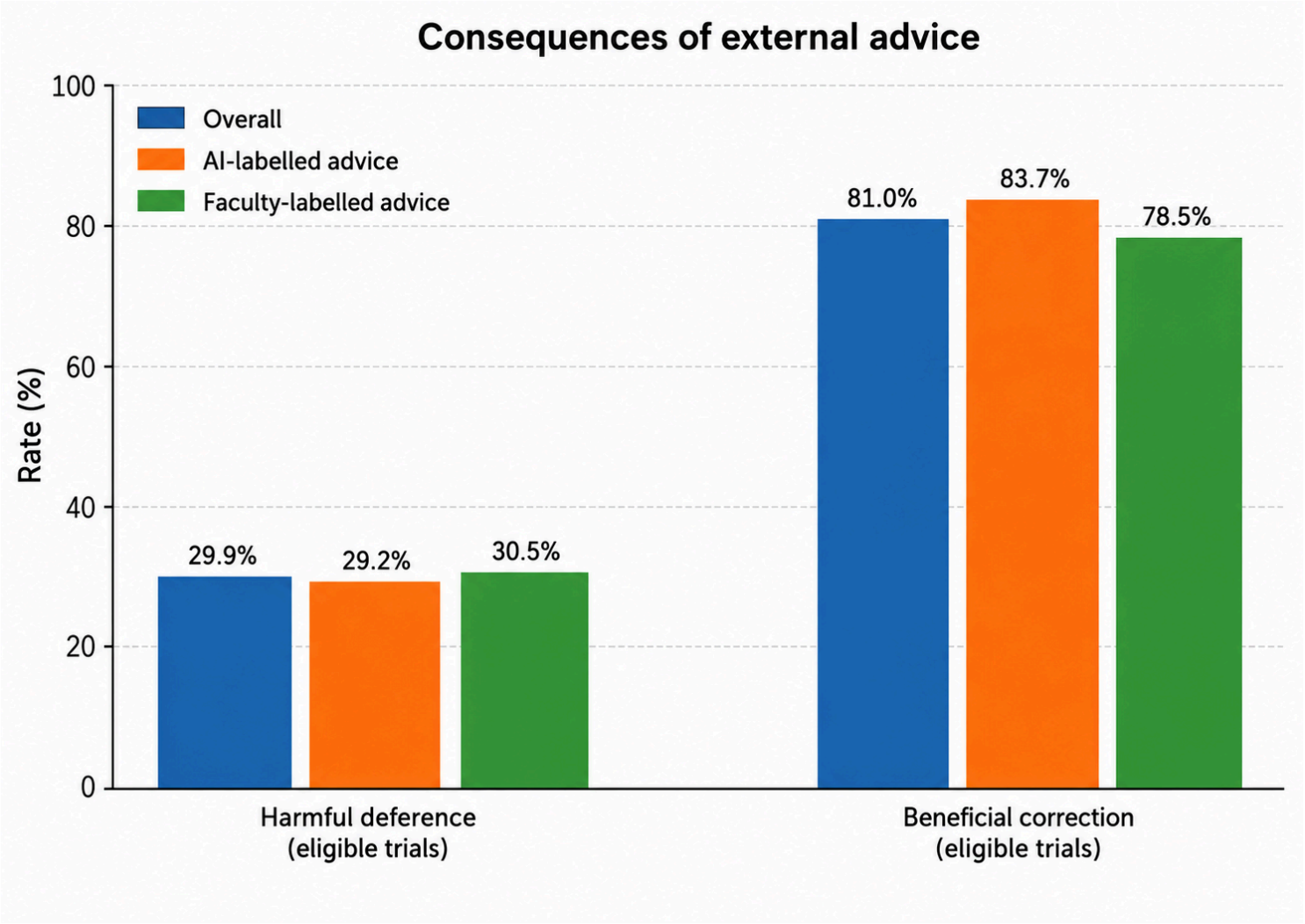


Figure 2

Consequences of external advice. The chart displays harmful diagnostic deference and beneficial correction overall and by advice label. Harmful diagnostic deference was calculated only among trials with an initially correct answer followed by inaccurate advice. Beneficial correction was calculated only among trials with an initially incorrect answer followed by correct advice. The figure shows that beneficial correction was substantially more frequent than harmful deference, whereas differences between AI-labelled and faculty-labelled advice were small.

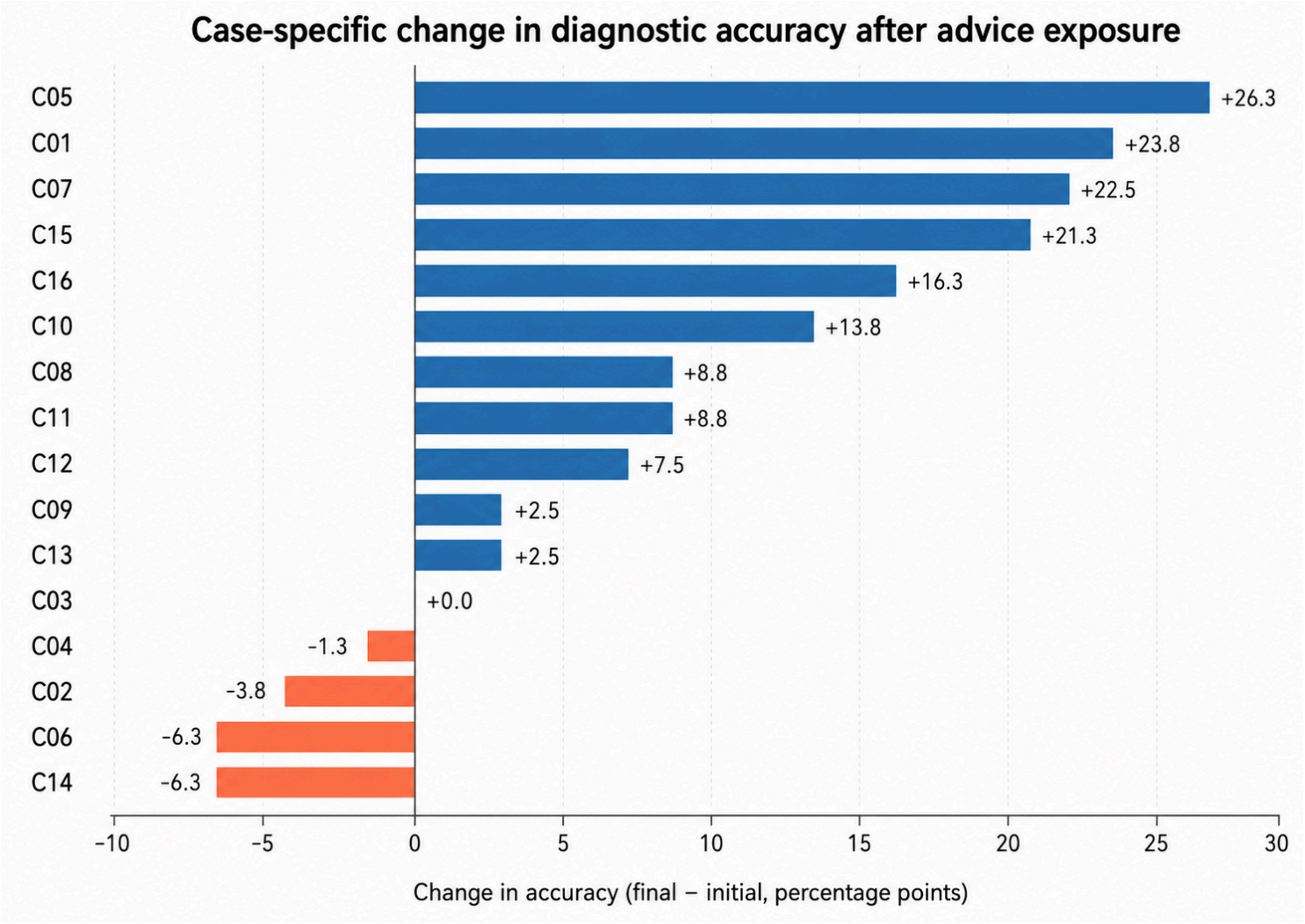


Figure 3

Case-specific change in diagnostic accuracy after advice exposure. Values represent final accuracy minus initial accuracy for each case, expressed in percentage points. Positive values indicate improved accuracy after advice exposure, whereas negative values indicate deterioration. The largest gains were observed for Cases C05, C01, C07, and C15, while Cases C06 and C14 showed the largest declines.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1CaseBankAdviceStimuliValidation.xlsx
- Additionalfile2DeidentifiedDatasetAnalysisPack.xlsx