Original Paper

Verification Behavior and Metacognitive Calibration in AI-Assisted Clinical Reasoning Among Physiotherapy Students: Cross-Sectional, 3-Group Comparative Study

Doğukan Kurç, PhD; Emre Dansuk, PhD; Ayşe Nur Tunalı, Prof Dr

Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Istanbul Medipol University, Istanbul, Turkey

Corresponding Author:
Doğukan Kurç, PhD
Department of Physiotherapy and Rehabilitation
Faculty of Health Sciences
Istanbul Medipol University
Ataturk Street no: 40/16
Istanbul, 34815
Turkey
Phone: 90 5376689085
Email: dkurc@medipol.edu.tr

Abstract

Background: Generative AI is part of everyday study in the health professions. Safe use is thought to depend less on how much students use it than on how they handle its output: checking it, reasoning soundly when it is wrong, and keeping confidence aligned with accuracy. Which behaviors are associated with better performance is untested in physiotherapy education.

Objective: We tested (1) whether status in a structured "AI in Healthcare" elective (no course exposure, currently taking, or completed) was associated with performance on a clinical case examination (basic knowledge, case-based reasoning, and hallucination-trap items) and with metacognitive calibration, and (2) among students who used AI during the examination, whether self-reported cross-checking of AI output or the quantity of AI use was associated with performance.

Methods: In this cross-sectional study, 137 (71%) of the 193 invited physiotherapy students (no course exposure: n=67; currently taking: n=43; completed: n=27) completed a researcher-developed 30-item open-resource examination with item-level confidence ratings and an AI-use survey (cross-check frequency; number of AI-assisted questions). Prespecified primary outcomes were the hallucination-trap score, case-based reasoning, and the calibration gap. The per-participant Brier score was the confirmatory calibration test. Group contrasts used Kruskal-Wallis tests and heteroscedasticity-consistent regression adjusted for age, year, grade point average, and prior AI use. Among 108 AI users, cross-check frequency and AI-use quantity were modeled jointly, then adjusted for demographics, course group, and basic-knowledge score.

Results: Cross-check frequency was associated with total score (β=2.98, 95% CI 1.27-4.68; P<.001), persisted after adjustment for demographics and course group (β=2.81, 95% CI 1.26-4.36; P<.001) and after excluding fast completions, and attenuated once basic-knowledge score was added (fully adjusted β=1.08, 95% CI –0.09 to 2.24; P=.07). AI-use quantity was not positively associated with performance (β=–0.26, 95% CI –0.53 to 0.00; P=.05). Calibration differed by group on the per-participant Brier score (H_2=10.23; P=.006) and expected calibration error (H_2=12.31; P=.002), both worst in the completed group (calibration gap: β=0.078, 95% CI 0.023-0.132; P=.005; accuracy at maximum confidence: 70.5%, 95% CI 44.8%-93.2%; 61 responses from 14 participants). Currently taking students outscored no-course-exposure peers on case-based reasoning (β=1.01, 95% CI 0.21-1.81; P=.01) and hallucination-trap items (β=0.92, 95% CI 0.06-1.79; P=.04), but not after excluding 11 completions under 5 minutes; basic knowledge did not differ (P=.74).

Conclusions: In an open-resource clinical examination, performance was associated with how students verified AI output, not how much they used it. The clearest deficit was metacognitive calibration rather than accuracy. These prespecified associations identify verification behavior and confidence calibration as measurable candidate targets for AI-literacy instruction. Because the design is cross-sectional and verification was self-reported, their instructional value requires testing in randomized or longitudinal designs.

(JMIR Med Educ 2026;12:e105832) doi: 10.2196/105832

XSL•FO
RenderX

KEYWORDS

AI; artificial intelligence; clinical reasoning; education; large language model; LLM; medical; metacognition; patient safety; physical therapy specialty; undergraduate; verification behavior

Introduction

Generative AI—particularly large language models (LLMs)—has moved from a niche research interest to a routine element of health professions education within a remarkably short period [1,2]. Students now use tools such as ChatGPT, Gemini, and Claude during coursework, clinical reasoning exercises, and self-study [3-5]. This rapid adoption has prompted health care faculties worldwide to introduce dedicated AI-literacy courses, yet evidence on whether such instruction produces safe, clinically translatable AI use remains limited [6-8]. A 2024 Best Evidence Medical Education (BEME)–aligned scoping review synthesizing 278 studies highlighted both the accelerating pace of adoption and the fragmented state of the evidence base, particularly for outcomes beyond knowledge acquisition [9].

Recent randomized trials have begun to quantify the pedagogical impact of AI integration. Brügge and colleagues [10] showed that ChatGPT-generated structured feedback improved medical students' clinical decision-making relative to AI-simulated history-taking alone. Hui et al [11] reported that ChatGPT-assisted problem-based learning (PBL) outperformed traditional PBL on postcourse knowledge, and Chan et al [12] reported that students rated a ChatGPT-based history-taking simulator favorably for learning. A systematic review by Feigerlova et al [13] nonetheless concluded that the evidence base remains methodologically heterogeneous, with small samples, short follow-up, and few validated outcome frameworks—and, critically, little attention to whether interventions teach students to verify AI output.

In physiotherapy specifically, Ergezen Sahin et al [14] conducted the first randomized controlled trial of AI-assisted PBL, showing that AI-PBL produced comparable or superior knowledge and clinical-competence gains relative to traditional PBL in undergraduate physiotherapy students at the same institution as this study. That work established the short-term efficacy of structured AI integration but did not examine the behaviors through which students convert AI use into safe performance, nor their metacognitive calibration when using AI tools—both central to this study.

Three considerations motivate a shift in emphasis from whether students use AI to how they use it. First, most evaluations have focused on overall accuracy rather than students' ability to reject plausible-but-incorrect, AI-generated content ("hallucinations"), arguably the most clinically consequential skill in an AI-assisted workflow [15,16]. Second, the verification behaviors that plausibly separate safe from unsafe AI use—cross-checking outputs against independent sources—have rarely been measured directly. Third, metacognitive calibration, the correspondence between a learner's confidence and actual accuracy, has received little empirical attention in AI-education research despite its central role in patient-safety frameworks [17,18].

Physiotherapy provides a particularly informative setting for these questions. Clinical reasoning in physiotherapy integrates pattern recognition, biomechanical reasoning, and patient-specific adaptation—domains where AI tools may be both helpful (rapid information retrieval) and hazardous (plausible but incorrect recommendations) [14,19]. Physiotherapy students are also heavy everyday users of generative AI, so formal AI education is layered onto preexisting informal habits rather than introduced to novices.

We organize these constructs into a simple conceptual framework grounded in metacognitive and confidence-calibration theory [17,18]. We treat AI-literacy instruction—operationalized here as course exposure, because the only AI-literacy scale we administered proved unreliable in this sample (see Methods)—as a distal input that may shape 2 proximal, safety-relevant processes when students reason with AI. These are the behavioral process of verification (cross-checking AI output against independent sources) and the metacognitive process of calibration (the correspondence between confidence and accuracy). Within this framework, performance on hallucination-trap items (plausible-but-incorrect and AI-typical content) is the most safety-sensitive outcome and case-based reasoning is the general-competence outcome, whereas basic factual knowledge should be comparatively insensitive to AI-specific instruction. We expected positive associations of course exposure and verification behavior with the AI-sensitive outcomes, and expected calibration to constitute a distinct vulnerability not guaranteed by accuracy. A cross-sectional design can test these associations but cannot establish the directional or mediating paths the framework implies. We therefore treat the framework as organizing rather than confirmatory (a schematic appears in Figure S1 in Multimedia Appendix 1).

This study examines how exposure to a structured "AI in Healthcare" course relates to three sets of outcomes in second- and third-year undergraduate physiotherapy students: (1) performance on a 30-item clinical case examination spanning basic clinical knowledge, case-based reasoning, and hallucination-trap items; (2) metacognitive calibration, operationalized through the confidence-accuracy relationship; and (3) AI-use behavior during the examination, including cross-check frequency and usage volume. We distinguish students currently enrolled in the course from those who completed it in a previous semester and those who never enrolled, enabling a 3-group comparison at a single time point.

Accordingly, we addressed the following four research questions (RQs): (RQ1) Is active course enrollment associated with better case-based reasoning and hallucination-trap performance than no course exposure? (RQ2) Do completed-course students differ from no-course-exposure students on these outcomes? (RQ3) Is metacognitive calibration associated with course status, and in what direction? (RQ4) Among students who use AI during assessment, is the quantity of AI use or the quality of AI use

XSL•FO
RenderX

(self-reported cross-checking) associated with performance? Because the design compares distinct cohorts at one time point, we framed all between-group findings as associations rather than as evidence of change over time.

Methods

Ethical Considerations

The study was approved by the Non-Interventional Clinical Research Ethics Committee of Istanbul Medipol University (decision number 713; April 30, 2026). All participants provided 3 separate electronic informed consents (study participation, data processing under the Turkish Personal Data Protection Law, and voluntary examination participation) before data collection. Data were pseudonymized using a participant code with no direct identifiers stored. Participants received no compensation for taking part in this study.

Study Design and Setting

This was a cross-sectional, comparative, single-institution study conducted at the Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Istanbul Medipol University. Data were collected on May 2-3, 2026, through a custom web-based study platform developed by the authors. Reporting follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for cross-sectional studies [20] (Multimedia Appendix 2).

The "AI in Healthcare" Course and Exposure Groups

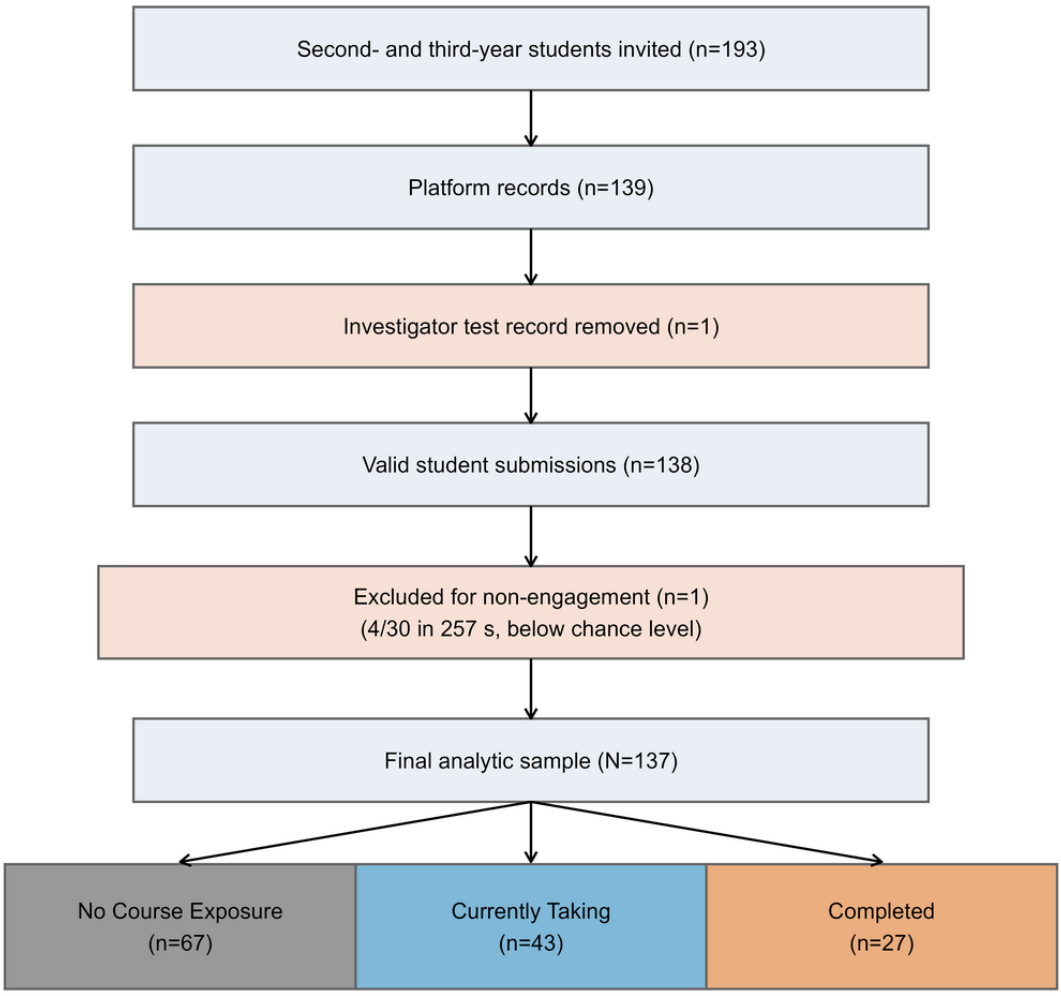
"AI in Healthcare" (Sağlıkta Yapay Zekâ; course code PTR2215211) is a program elective carrying 2 European Credit Transfer and Accumulation System (ECTS) credits offered in the spring semester to students of both the Turkish- and English-language physiotherapy programs, delivered as 14 weekly 2-hour sessions of lecture and question-and-answer teaching and assessed by a midterm (40%) and a final examination (60%). The delivered curriculum covers the foundations of AI and machine learning in health care; health data management and ethics; AI in medical imaging, diagnosis, clinical documentation, and predictive analytics; robotics, wearable technology, and telehealth in physiotherapy; legal and data-privacy aspects; and, in the sessions on generative AI, prompt engineering, cross-checking AI output against independent sources, and critical appraisal of generated content. Exposure status was taken from the enrollment item of the demographic form and defined as 3 groups: currently taking (enrolled in the course during data collection), completed (completed the course in an earlier semester), and no course exposure (never enrolled). At data collection (May 2-3, 2026), currently taking students had completed approximately 11 of the 14 teaching weeks, including the sessions on generative AI, prompt engineering, and cross-checking of AI output. Prior everyday AI use was near-universal in all three groups, so the exposure variable indexes curricular exposure only. The study examination was not part of course assessment, participation was voluntary, and responses were pseudonymized before analysis.

Participants

All second- and third-year undergraduate physiotherapy students enrolled during the spring 2026 semester (N=193) were invited through course announcements and email. Inclusion criteria were active enrollment, age ≥18 years, and electronic informed consent. Exclusion criteria, specified before analysis, were failure to complete any of the 3 assessment stages and evidence of nonengagement, operationalized as an examination score at or below the chance level for 5-option items (≤6/30) combined with a completion time inconsistent with item-by-item reading (<5 minutes). Of 139 platform records, 1 was an investigator test record and was removed before defining valid submissions; of the remaining 138 valid student submissions, 1 met the nonengagement criterion (4/30 in 257 seconds), leaving a final analytic sample of 137 (71% of those invited; Figure 1). The nonengagement rule was specified before analysis and applied without modification; it was the only participant exclusion.

XSL•FO
RenderX

Figure 1. Participant flow. Of 193 invited students, 139 platform records were created. One investigator test record was removed, leaving 138 valid student submissions. One submission met the prespecified nonengagement criterion, yielding the final analytic sample of 137 students (71% of those invited) across 3 AI-course exposure groups (no course exposure, currently taking, or completed).



Measures

Demographic and Background

Demographic and background data (age, gender, year of study, grade point average [GPA; 0-4.0], prior AI tool use, and daily technology and AI use) were collected via a structured form.

Medical Artificial Intelligence Readiness Scale for Medical Students

The 22-item Medical Artificial Intelligence Readiness Scale for Medical Students (MAIRS-MS) [21] assesses readiness for medical AI on a 5-point Likert scale; the Turkish version was used (internal consistency: α=.925).

AI Attitude Scale

The 25-item AI Attitude Scale (AIS) [22] assesses attitudes across interaction, widespread effect, and anxiety subdomains (α=.867).

Artificial Intelligence Literacy Scale

The 12-item Artificial Intelligence Literacy Scale (AILS) [23,24] was administered, but 2 of its 4 subscales showed unacceptable internal consistency in this sample (awareness: α=.167; use: α=.363). Because an aggregate score derived from unstable subdimensions would compromise construct validity, the AILS was excluded from all analyses and is reported only for transparency (Multimedia Appendix 1).

30-Item Clinical Case Examination

A researcher-developed examination assessed clinical reasoning under open-resource conditions across 3 content categories of 10 items each: basic clinical knowledge (core physiotherapy facts), case-based reasoning (clinical vignettes requiring integrative reasoning), and hallucination trap (items built around documented LLM failure modes in physiotherapy—fabricated protocols, outdated guidelines, and country-specific regulatory content [15,16]—in which the most plausible-sounding option is incorrect). Because students could answer with AI tools, the internet, lecture notes, or their own knowledge, the hallucination-trap score indexes the ability to reject plausible-but-incorrect, AI-typical content, whatever resource was consulted. It is not a measure of detecting errors in a particular AI output, and we use the term hallucination-trap score throughout. Items were written from a blueprint of 10 items per category and reviewed by the 3 authors, all physiotherapy faculty, with each hallucination-trap item mapped a priori to a documented failure mode. No external expert panel

XSL•FO
RenderX

or formal pretest was conducted. Each item offered five options with answer-key rotation. Students had 30 minutes and were permitted to use any resources (notes, internet, or AI tools) to reflect real-world workflows. Item difficulty, discrimination, and category- and total-scale reliability were computed (Results; Multimedia Appendix 1). The full instruments (English and original Turkish) appear in Multimedia Appendix 3.

Confidence Ratings and Calibration

For each item, participants rated confidence in their chosen answer on a 5-point scale anchored at 1=not at all sure, 2=slightly sure, 3=moderately sure, 4=quite sure, and 5=completely sure, after any self-directed resource or AI use for that item. No correctness feedback was given during the examination. To express confidence as a subjective probability of being correct, we applied the affine mapping (confidence – 1)/4, so that levels 1-5 correspond to 0, 0.25, 0.50, 0.75, and 1.00—the conventional linear anchoring for evenly spaced Likert confidence. The item-level calibration gap was |mapped confidence – correct|, averaged within participants (higher = worse). We additionally computed a signed bias (mapped confidence – accuracy), the Brier score with its Murphy decomposition (reliability, resolution, and uncertainty), the expected calibration error (ECE), the area under the receiver operating characteristic curve (AUC) for confidence discriminating correct from incorrect responses, and accuracy at each confidence level [17]. Because the mapping is an assumption, we repeated the calibration-gap and Brier group comparisons under an alternative chance-floored mapping (levels 1-5 → 0.20, 0.40, 0.60, 0.80, 1.00) appropriate to 5-option items. Metrics formed by pooling responses across participants (eg, accuracy at a given confidence level and AUC) are descriptive, response-level statistics. Participant-level inference rested on the per-participant calibration gap and Brier score.

Postexamination AI-Use Survey

Immediately after the examination, participants reported the resources used, the number of questions for which AI was used (recorded in 7 bands: 0, 1-5, 6-10, 11-15, 16-20, 21-25, and 26-30, and scored at the band midpoint of 0, 3, 8, 13, 18, 23, and 28 for analysis), how often they cross-checked the AI's answers (a single item: never, rarely, sometimes, mostly, and always, scored 1-5); perceived usefulness, and open-ended reflections on prompt strategy and views on AI in physiotherapy education. Throughout, verification behavior denotes this single self-reported cross-check item, and AI-use quantity denotes the midpoint-coded band.

Statistical Analysis

Analyses were performed in Python 3.12 (*pandas*, *scipy*, and *statsmodels*) with R 4.6 for figures. Most outcomes departed from normality (Shapiro-Wilk), so nonparametric methods were used. The ethics protocol reported an a priori G*Power calculation for the planned comparison of course-exposed with unexposed students (d=0.5; α=.05; 80% power), specifying a minimum of 52 students per group and a target of 60 per group after allowing for incomplete records. The final sample comprised 137 students. The small completed-course group (n=27) is interpreted using effect sizes and CIs. Primary outcomes and the exclusion rule were specified before analysis. The analysis plan was not publicly registered, and the analysis code is provided in Multimedia Appendix 4. Primary outcomes were the hallucination-trap score, case-based reasoning, and the calibration gap, each a distinct prespecified hypothesis tested at α=.05. We did not apply a family-wise correction across these 3 outcomes but report which contrasts survive a conservative Bonferroni adjustment. The verification analysis (cross-check frequency and AI-use quantity in relation to total and hallucination-trap scores) was prespecified and is treated as confirmatory within its 2-outcome family. The calibration decomposition beyond the calibration gap and the per-participant Brier test (Murphy components, ECE, AUC, and accuracy by confidence level) are descriptive and uncorrected and are read as convergent rather than independent evidence. Total and basic-knowledge scores were secondary. Because the total score is the sum of the 3 subscales, any model adjusting the total score for the basic subscale is partly an outcome-on-component regression. We therefore use the nonoverlapping case-plus-hallucination subtotal as the competence-adjusted outcome. All 137 participants sat an examination in which AI use was permitted. Of these, 108 reported using AI, and only the verification analyses are restricted to them. Cross-check frequency and AI-use quantity were defined only for these 108 students. Demographic and outcome variables were complete (no imputation).

Three-group differences were tested with Kruskal-Wallis *H* tests, with epsilon-squared (ε^2 = H/[N – 1]) effect sizes; significant omnibus tests were followed by Mann-Whitney *U* pairwise comparisons with Holm correction within each outcome family. Because groups differed in age and year, group contrasts were reestimated with heteroscedasticity-consistent (type 3; HC3)–robust ordinary least squares (OLS) adjusted for age, year, GPA, and prior AI use (no course exposure as reference). Among AI users, the verification analysis is reported as 8 HC3 models. Models 1-5 and 7 follow the analyses described in the submitted version. The full competence-adjusted model (model 6), the fast-completion sensitivity model (model 8), and standardized coefficients were added during revision. Model 1 includes cross-check frequency alone or AI-use quantity alone. Model 2 includes both predictors jointly and is the prespecified primary model. Model 3 adds age, year, GPA, and prior AI use, and model 4 additionally includes course group. Models 5 and 6 use the nonoverlapping case-plus-hallucination subtotal as the outcome and adjust for basic knowledge. Model 5 also includes AI-use quantity and GPA, whereas model 6 uses the full model 4 covariate set. Model 7 models the hallucination-trap score with AI-use quantity, GPA, and case-based-reasoning score as covariates. Model 8 repeats model 2 after excluding completions under 5 minutes. Collinearity was checked with the variance inflation factor (VIF). Because cross-check frequency (per level) and AI-use quantity (per question) are on different scales, standardized coefficients (per SD of each predictor) are also reported. Scale-examination relationships used Spearman correlations and rank-based partial correlations adjusting for age, GPA, and year. Prior AI use was reported by 129/137 (94.2%; near-universal) participants and was retained

XSL•FO
RenderX

in the HC3 OLS but dropped from the rank-based partials because its near-zero variance offers no adjustment leverage. Robustness was assessed by reestimating the group contrasts and the verification models after excluding (1) a single bulk-response participant (30/30 in 129 seconds) and (2) all completions under 5 minutes. Response-level calibration statistics (accuracy by confidence level and AUC) are descriptive and were accompanied by participant-cluster percentile bootstrap 95% CIs (2000 resamples; seed and code in Multimedia Appendix 4). Per-participant ECE was computed over the 5 confidence levels and compared across groups in the same way as the per-participant Brier score. Open-ended responses were summarized descriptively. Two-tailed α=.05 was used, with Holm correction within each outcome's pairwise comparisons.

Results

Participant Characteristics

Participants averaged 21.3 (SD 2.2) years; 97 (70.8%) were women. GPA (mean 2.62, SD 0.69) and gender did not differ across groups (GPA: Kruskal-Wallis H_2=1.96; P=.38; gender: χ^2_4=7.5; P=.11). Groups differed in age (Kruskal-Wallis H_2=7.52; P=.02) and year distribution (χ^2_2=6.8; P=.03), with currently taking students younger and more concentrated in year 2—consistent with progression through an elective. Age and year were therefore adjusted for in regression models. Prior AI use was near-universal (129/137, 94.2% overall), and 108 (78.8%) participants reported using AI during the examination (no course exposure: n=47, 70.1%; currently taking: n=37, 86%; completed: n=24, 88.9%; χ^2_2=6.0; P=.05); of the 29 nonusers, 25 used no resource, 3 used lecture notes, and 1 used lecture notes and internet search. Group characteristics appear in Table 1.

Table 1. Participant characteristics by AI-course exposure group. *P* values were obtained from Kruskal-Wallis H (age and grade point average [GPA]) and chi-square tests (gender, year, and AI use during examination).

Characteristic	No course exposure (n=67)	Currently taking (n=43)	Completed (n=27)	All (n=137)	*P* value
Age (years), mean (SD)	21.7 (2.4)	20.8 (2.1)	21.2 (1.7)	21.3 (2.2)	.02
Women, n (%)	43 (64.2)	33 (76.7)	21 (77.8)	97 (70.8)	.11
GPA, mean (SD)	2.67 (0.68)	2.64 (0.62)	2.48 (0.83)	2.62 (0.69)	.38
Year 2 students, n	58	42	21	121	.03
Year 3 students, n	9	1	6	16	.03
Prior AI use, n (%)	60 (89.6)	43 (100)	26 (96.3)	129 (94.2)	—[a]
Used AI during examination, n (%)	47 (70.1)	37 (86)	24 (88.9)	108 (78.8)	.05

[a]Not tested: all currently taking students reported prior AI use.

Reliability and Psychometric Validity of the Examination

Internal consistency was excellent for the MAIRS-MS (α=.925) and AIS (α=.867). The 30-item examination was highly reliable (Kuder-Richardson Formula 20 [KR-20]=0.916), with strong category-level reliability for basic clinical knowledge (KR-20=0.814), case-based reasoning (0.861), and the hallucination-trap domain (0.832). Mean item difficulty was 0.788 (SD 0.111) and mean discrimination was 0.491 (SD 0.120), with all items having a discrimination index of at least 0.20. The 3 content categories were moderately intercorrelated (basic-case: ρ=0.47; basic-hallucination: ρ=0.41; case-hallucination: ρ=0.60). After disattenuation for category reliabilities, the case-hallucination correlation rose to ρ=0.71, indicating substantial shared variance. We therefore treat the hallucination-trap domain as internally consistent and provisional rather than as a factorially distinct, fully validated construct—a formal factor analysis is precluded by the 30-item/137-participant ratio for dichotomous data. Within the hallucination-trap domain, 9 of 10 items showed strong corrected item-total correlations (0.49-0.72); only item Q21 was weak (difficulty=0.336; corrected item-total correlation 0.090 within the hallucination-trap domain and 0.133 against the 30-item total) and functioned as a low-discrimination sentinel item. Item Q7 (basic knowledge) showed a ceiling effect (difficulty=0.920). Because 2 AILS subscales were unreliable (awareness: α=.167; use: α=.363), the AILS was excluded from analysis, and we accordingly treat AI literacy as operationalized only through course exposure rather than as a validly measured latent trait.

Group Comparisons of Primary Outcomes (RQ1, RQ2)

Significant omnibus group differences emerged for hallucination-trap score (H_2=10.07; P=.007; ε^2=0.074), case-based reasoning (H_2=8.03; P=.02; ε^2=0.059), and the calibration gap (H_2=10.02; P=.007; ε^2=0.074). The total score difference was not significant (H_2=5.69; P=.06) and basic knowledge did not differ (H_2=0.61; P=.74). For the hallucination-trap score, currently taking students (mean 8.60, SD 1.68) scored higher than both completed (mean 7.26, SD 2.57; U=775; Holm-adjusted P=.03) and no-course-exposure (mean 7.22, SD 2.72; U=1908; Holm-adjusted P=.01) students, who did not differ from each other (U=909; Holm-adjusted P=.97). For case-based reasoning, currently taking students (mean 9.07, SD 1.50) exceeded no-course-exposure peers (mean

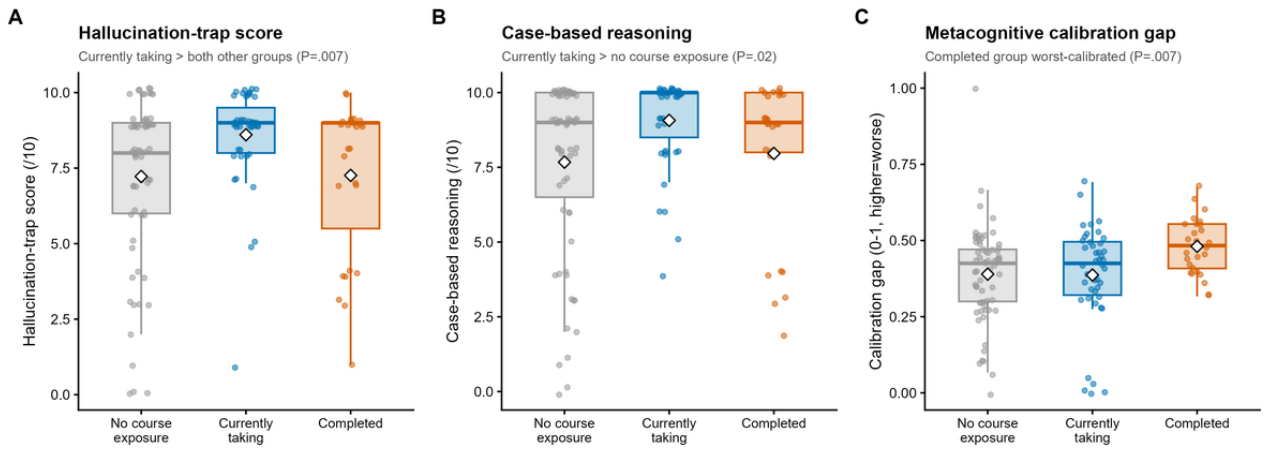
XSL•FO
RenderX

7.67, SD 2.93; *U*=1861; Holm-adjusted *P*=.02). For the calibration gap, completed-course students (mean 0.48, SD 0.09) showed larger gaps than both no-course-exposure students (mean 0.39, SD 0.15; *U*=1275; Holm-adjusted *P*=.006) and currently taking students (mean 0.39, SD 0.17; *U*=374; Holm-adjusted *P*=.03). With 3 primary outcomes, a conservative Bonferroni threshold (0.0167) is met by the hallucination-trap (*P*=.007) and calibration-gap (*P*=.007) omnibus tests but not by case-based reasoning (*P*=.02), which we therefore read as borderline (Table 2; Figure 2).

Table 2. Three-group comparisons (mean, SD) with Kruskal-Wallis tests and ε2 effect sizes. Higher calibration gap = worse calibration.

Outcome	Maximum value	No course exposure	Currently taking	Completed	*H*	*P* value	ε^2
Total score	30	22.46 (7.13)	25.86 (4.13)	23.00 (6.80)	5.69	.06	0.042
Basic knowledge	10	7.57 (2.68)	8.19 (2.00)	7.78 (2.78)	0.61	.74	0.004
Case-based reasoning	10	7.67 (2.93)	9.07 (1.50)	7.96 (2.62)	8.03	.02	0.059
Hallucination trap	10	7.22 (2.72)	8.60 (1.68)	7.26 (2.57)	10.07	.007	0.074
Calibration gap	1	0.39 (0.15)	0.39 (0.17)	0.48 (0.09)	10.02	.007	0.074

Figure 2. Primary outcomes by group: (A) hallucination-trap score, (B) case-based reasoning, and (C) metacognitive calibration gap. Boxplots show median and IQR. Points represent individual participants, and white diamonds indicate means. Colors denote AI-course exposure group: gray = no course exposure, blue = currently taking, orange = completed. *P* values are from Kruskal-Wallis omnibus tests (panel subtitles) with Holm-corrected pairwise comparisons.



Adjusted Group Contrasts

In HC3-robust models adjusted for age, year, GPA, and prior AI use, currently taking students scored higher than no-course-exposure peers on total (β=2.50, 95% CI 0.35-4.64; *P*=.02), hallucination-trap (β=0.92, 95% CI 0.06-1.79; *P*=.04), and case-based reasoning (β=1.01, 95% CI 0.21-1.81; *P*=.01) outcomes. In standardized terms these active-exposure associations were modest (standardized β=0.39-0.40). The calibration gap did not differ for currently taking students (β=–0.008; *P*=.81). Completed-course students showed no detectable performance difference on any examination outcome—total β=0.06 (95% CI –3.05 to 3.17), case β=0.09 (95% CI –1.12 to 1.31), hallucination-trap β=–0.14 (95% CI –1.24 to 0.96)—but the wide intervals reflect limited precision in this small group (n=27) rather than established equivalence. Their only significant contrast was a larger calibration gap (β=0.078, 95% CI 0.023-0.132; *P*=.005). A rank-based partial correlation supported the active-exposure association with case-based reasoning after adjustment for age, GPA, and year (*r*=0.231; *P*=.007; hallucination-trap: *r*=0.232; *P*=.007). Full models appear in Table 3.

Table 3. Heteroscedasticity-consistent (type 3)–robust ordinary least squares contrasts vs the no-course-exposure reference, adjusted for age, year, grade point average, and prior AI use.

Outcome	Currently taking, β (95% CI)	*P* value	Completed, β (95% CI)	*P* value
Total score	2.50 (0.35 to 4.64)	.02	0.06 (–3.05 to 3.17)	.97
Case-based reasoning	1.01 (0.21 to 1.81)	.01	0.09 (–1.12 to 1.31)	.88
Hallucination trap	0.92 (0.06 to 1.79)	.04	–0.14 (–1.24 to 0.96)	.80
Calibration gap	–0.008 (–0.076 to 0.059)	.81	0.078 (0.023 to 0.132)	.005

XSL•FO
RenderX

Metacognitive Calibration (RQ3)

Calibration analyses pointed to the completed group as the most miscalibrated, with the precision limits expected for n=27. The confirmatory, participant-level test—the per-participant Brier score—differed across groups (Kruskal-Wallis H_2=10.23; P=.006), being poorest (highest) for completed students (mean 0.306, SD 0.104) vs currently taking (mean 0.225, SD 0.120) and no course exposure (mean 0.233, SD 0.143); per-participant ECE was also highest in the completed group (mean 0.373, SD 0.124; mean 0.293, SD 0.161; and mean 0.264, SD 0.158; Kruskal-Wallis H_2=12.31; P=.002; completed vs no course exposure adjusted β=0.100, 95% CI 0.037-0.162; P=.002). Convergent exploratory, response-level metrics pointed the same way: the Murphy reliability (calibration-error) component was largest for the completed group (0.133 vs 0.112 and 0.066) and resolution was 0.021 in the no-course-exposure group and 0.006 in both other groups, and confidence discriminated correct from incorrect responses least well (AUC=0.57, 95% CI 0.45-0.68 vs AUC=0.64, 95% CI 0.57-0.71 and AUC=0.71, 95% CI 0.66-0.75). The most-cited descriptive figure—accuracy at maximum confidence—was 95.0% (currently taking; 260 responses from 32 participants) and 90.8% (no course exposure; 295 responses from 36 participants) but 70.5% (completed); we emphasize, however, that the completed-group value rests on only 61 maximum-confidence responses from 14 students and is correspondingly imprecise (participant-bootstrap 95% CI 44.8%-93.2%). It should therefore be read as a hypothesis-generating descriptive statistic, not a precise estimate (Figure 3; Table 4).

Figure 3. Metacognitive calibration by group. (A) Calibration curves: observed accuracy (y) against stated confidence (x). The dashed line indicates perfect calibration. Point size is proportional to the number of responses. Colors denote AI-course exposure group (gray = no course exposure, blue = currently taking, orange = completed). (B) Accuracy on responses marked "completely sure" (confidence = 5), with per-group response and participant counts (no course exposure: 295 responses from 36 participants, currently taking: 260 from 32, completed: 61 from 14); the completed-group estimate is imprecise (Table 4).

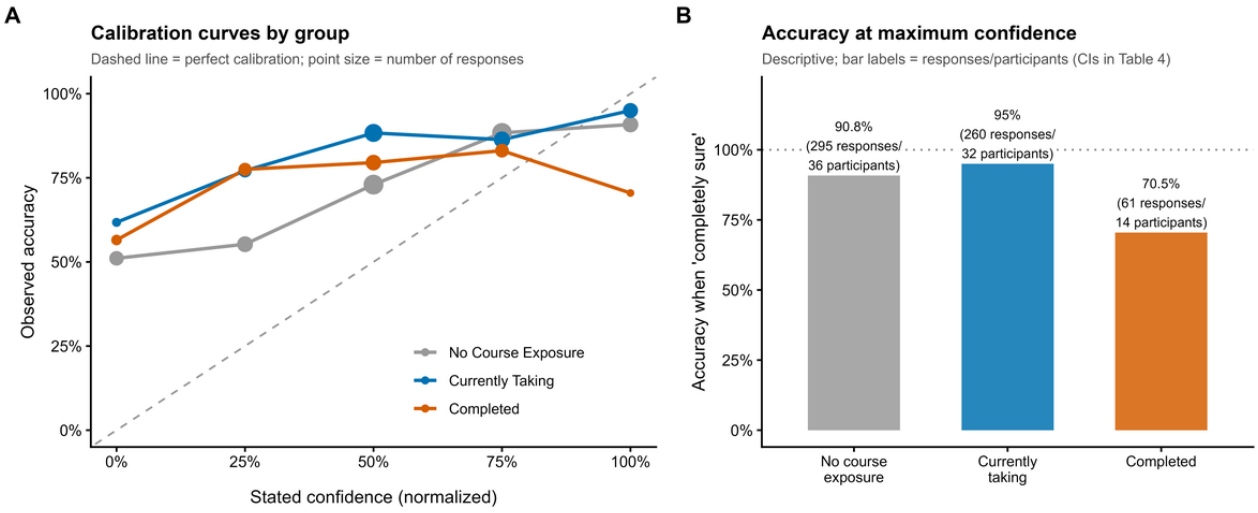


Table 4. Per-group calibration metrics. Lower Brier, reliability, and expected calibration error (ECE) values indicate better calibration, whereas higher resolution and the area under the receiver operating characteristic curve (AUC) values indicate better performance. Calibration gap, signed bias, Brier score, and per-participant ECE are reported as participant-level means (SDs). Reliability, resolution, pooled ECE, and AUC are response-level descriptives. Per-participant ECE and the per-participant Brier Kruskal-Wallis test (P=.006) are the participant-level comparisons. Accuracy at maximum confidence is a response-level proportion with a participant-cluster percentile bootstrap 95% CI (2000 resamples).

Group	Calibration gap, mean (SD)	Signed bias, mean (SD)	Brier, mean (SD)	Reliability, pooled	Resolution, pooled	ECE, pooled	ECE, per participant, mean (SD)	AUC (95% CI)	Accuracy at maximum confidence, % (responses / participants; 95% CI)
No course exposure	0.39 (0.15)	–0.20 (0.20)	0.233 (0.143)	0.066	0.021	0.225	0.264 (0.158)	0.71 (0.66-0.75)	90.8 (295/36; 85.8-94.0)
Currently taking	0.39 (0.17)	–0.26 (0.18)	0.225 (0.120)	0.112	0.006	0.277	0.293 (0.161)	0.64 (0.57-0.71)	95.0 (260/32; 90.6-97.6)
Completed	0.48 (0.09)	–0.28 (0.26)	0.306 (0.104)	0.133	0.006	0.322	0.373 (0.124)	0.57 (0.45-0.68)	70.5 (61/14; 44.8-93.2)

This pattern reflected a loss of confidence-accuracy coupling rather than classic overconfidence: average signed bias was negative in all groups (currently taking: mean –0.26, SD 0.18; completed: mean –0.28, SD 0.26; no course exposure: mean –0.20, SD 0.20), indicating net underconfidence, and the proportion with an overconfident profile did not differ

XSL•FO
RenderX

significantly (currently taking: 0/43, 0%; completed: 3/27, 11.1%; no course exposure: 4/67, 6%; χ^2_2=4.4; *P*=.11). The group difference was robust to the confidence-to-probability mapping: under the chance-floored alternative the calibration-gap ordering was preserved (completed worst; Kruskal-Wallis H_2=8.07; *P*=.02; Brier H_2=7.14; *P*=.03). Because the completed group is also the smallest and most self-selected (students who elected the course in an earlier semester), we cannot exclude that its calibration gap partly reflects cohort composition rather than a calibration deficit per se (see Limitations).

Verification Behavior vs Usage Volume Among AI Users (RQ4)

Among the 108 participants who used AI during the examination, cross-check frequency was associated with both total score (model 2: β=2.98, 95% CI 1.27-4.68 per level; *P*<.001) and hallucination-trap score (β=1.03, 95% CI 0.33-1.73; *P*=.004); the joint model explained R^2=0.198 (adjusted R^2=0.183; $F_{2,105}$=5.84; *P*=.004). The association was essentially unchanged after adjustment for demographics (model 3: β=2.64, 95% CI 0.96-4.33; *P*=.002) and additionally for course group (model 4: β=2.81, 95% CI 1.26-4.36; *P*<.001). The number of AI-assisted questions was not positively associated with performance: alone it was null (model 1: β=–0.015, 95% CI –0.24 to 0.21 per question; *P*=.90), and in the joint model it was weakly negative with an interval reaching zero (β=–0.26, 95% CI –0.53 to 0.00; *P*=.051). On a common scale, 1 SD of cross-check frequency corresponded to 0.52 SD of total score (95% CI 0.22-0.81) and 1 SD of AI-use quantity to –0.28 SD (95% CI –0.56 to 0.00); the 2 predictors were moderately correlated (ρ=0.47; VIF=1.35).

Adjustment for basic knowledge attenuated the verification association. In the previously reported specification, the association with the case-plus-hallucination subtotal was β=1.38 (95% CI 0.17-2.59; *P*=.03; model 5), adjusting for AI-use quantity, GPA, and basic knowledge. With the full covariate set, the estimate was β=1.08 (95% CI –0.09 to 2.24; *P*=.07; model 6). Thus, the estimate was smaller after competence adjustment and its precision depended on the covariate set. The association was not specific to hallucination-trap items: after adjustment for case-based reasoning, cross-check frequency was no longer associated with hallucination-trap score (model 7: β=0.28; 95% CI –0.27 to 0.84; *P*=.32), and the currently taking advantage on hallucination-trap items likewise did not survive adjustment for case-based reasoning (β=0.27; *P*=.42). Verification and active course enrollment were therefore associated with general clinical reasoning, of which hallucination-trap performance is a correlated, safety-weighted facet (disattenuated case-hallucination: ρ=0.71), rather than a separable hallucination-specific ability (Table 5).

The cross-check relationship was monotonic over its informative range (Spearman ρ=0.38; *P*<.001): mean total scores rose from 19.7/30 (SD 8.1) among students who never cross-checked (n=21) to 26.5/30 (SD 3.7) at the middle level (sometimes: n=50); the gradient rests on the never-to-sometimes range, as the highest level was sparsely populated (mean 24.0/30, SD 10.4; n=3; Figure 4).

XSL•FO
RenderX

Table 5. Heteroscedasticity-consistent (type 3)–robust regression among AI users (n=108): verification behavior (self-reported cross-check frequency, scored 1-5) vs AI-use quantity (number of AI-assisted questions, band midpoint). Model 2 is the prespecified primary model (adjusted R2=0.183; F2,105=5.84; *P*=.004; variance inflation factor=1.35). Models 5 and 6 use the case-plus-hallucination subtotal as the outcome to avoid adjusting the total score for one of its own components; model 7 tests specificity to hallucination-trap items; model 8 repeats model 2 after excluding completions under 5 minutes.

Model	Predictor	Covariates	Total score, β (95% CI)	*P* value	Hallucination-trap score, β (95% CI)	*P* value
1	Cross-check frequency (per level)	None	2.15 (0.72 to 3.59)	.003	0.77 (0.18 to 1.35)	.01
1	AI-assisted questions (per question)	None	–0.015 (–0.24 to 0.21)	.90	0.00 (–0.08 to 0.08)	.97
2	Cross-check frequency	AI-assisted questions	2.98 (1.27 to 4.68)	<.001	1.03 (0.33 to 1.73)	.004
2	AI-assisted questions	Cross-check frequency	–0.26 (–0.53 to 0.00)	.051	–0.08 (–0.18 to 0.01)	.08
3	Cross-check frequency	Model 2 + age, year, GPA, prior AI use	2.64 (0.96 to 4.33)	.002	0.82 (0.13 to 1.50)	.02
4	Cross-check frequency	Model 3 + course group	2.81 (1.26 to 4.36)	<.001	0.89 (0.24 to 1.54)	.007
5	Cross-check frequency	AI-assisted questions, GPA, basic-knowledge score (subtotal outcome)[a]	1.38 (0.17 to 2.59)	.03	—[b]	—
6	Cross-check frequency	Model 4 covariates + basic-knowledge score (subtotal outcome)[a]	1.08 (–0.09 to 2.24)	.07	—	—
7	Cross-check frequency	AI-assisted questions, GPA, case-based reasoning score	—	—	0.28 (–0.27 to 0.84)	.32
8	Cross-check frequency	AI-assisted questions; completions <5 min excluded (n=97)	2.97 (1.05 to 4.89)	.002	1.12 (0.30 to 1.95)	.008
2[c]	Cross-check frequency (per SD)	AI-assisted questions	0.52 (0.22 to 0.81)	<.001	0.42 (0.14 to 0.71)	.004
2[c]	AI-assisted questions (per SD)	Cross-check frequency	–0.28 (–0.56 to 0.00)	.051	–0.21 (–0.45 to 0.03)	.08

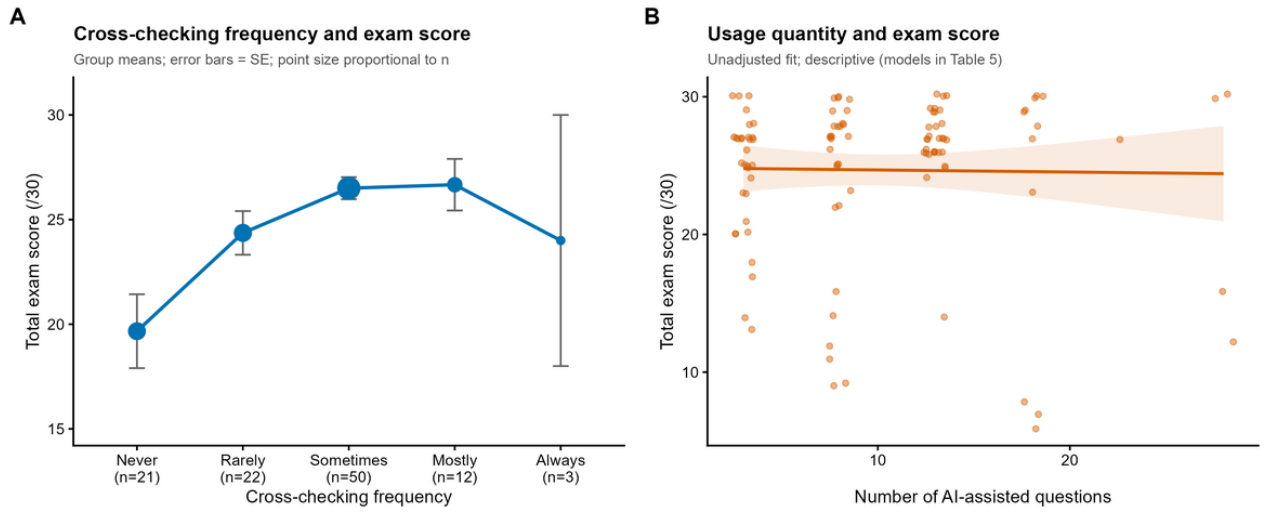
[a]Outcome is the case-plus-hallucination subtotal (0-20); the coefficient is identical for the total score when basic knowledge is a covariate. Model 5 is the specification reported in the submitted version; model 6 is the fully adjusted competence-baseline estimate.

[b]Not applicable.

[c]Standardized coefficients: outcome and predictors *z*-scored within the 108 AI users.

XSL•FO
RenderX

Figure 4. AI-use quality vs quantity among AI users (n=108). (A) Total exam score by cross-check frequency (error bars = SE; point size proportional to n; the highest level, "always," is n=3). (B) Number of AI-assisted questions vs total score, showing no positive association. These are descriptive; the model sequence appears in Table 5.



Self-Report Scales and Examination Accuracy

Readiness and attitude scales correlated with one another (MAIRS-MS–AIS: ρ=0.49) but not with examination accuracy (MAIRS-MS: ρ=0.12; *P*=.16; AIS: ρ=0.12; *P*=.17). In contrast, mean confidence was strongly associated with accuracy (ρ=0.50; *P*<.001) and the calibration gap was negatively associated with accuracy (ρ=–0.24; *P*=.005). The AIS Interaction subscale tracked self-reported confidence (ρ=0.21; *P*=.02) but not actual accuracy (ρ=0.09; *P*=.28). Positive attitudes toward AI tracked subjective certainty rather than objective performance. Because the calibration gap is an arithmetic function of confidence and correctness, the confidence-accuracy and gap-accuracy correlations are partly structural. The independent evidence for the calibration group difference is the between-group per-participant Brier comparison (*P*=.006), not these coupled within-sample correlations.

Qualitative Findings

Rapid thematic coding of open-ended responses (AI users) identified writing clear, concise prompts as the most common strategy, followed by reporting no strategy; "asking AI to check my answer" and "comparing across tools" appeared as explicit verification strategies, providing convergent behavioral evidence for the quantitative cross-check finding. Views on AI in physiotherapy education were predominantly positive but conditional, emphasizing the need for verified, controlled use and concern about overreliance.

Sensitivity Analyses

The calibration finding was robust. The larger calibration gap in the completed group persisted after excluding the single bulk-response participant (β=.079; *P*=.004; n=136) and after excluding all completions under 5 minutes (β=.075, 95% CI 0.017-0.132; *P*=.01; n=126), and the per-participant Brier difference likewise persisted (H_2=9.32; *P*=.009). The performance advantages of currently taking students were directionally stable but not robust to the exclusion of fast completions. A total of 11 participants completed the examination in under 5 minutes (no course exposure=6, currently taking=4, completed=1); all 11 had used AI. Excluding the bulk-response participant alone left the contrasts essentially unchanged (total: β=2.38; *P*=.03; hallucination-trap: β=0.88; *P*=.048), whereas excluding all 11 attenuated them to nonsignificance (total: β=1.98, 95% CI –0.30 to 4.27; *P*=.09; case-based reasoning: β=0.84, 95% CI –0.01 to 1.69; *P*=.052; hallucination-trap: β=0.82, 95% CI –0.09 to 1.73; *P*=.08); the omnibus tests for hallucination-trap score (H_2=6.42; *P*=.04) and calibration gap (H_2=8.63; *P*=.01) remained significant and that for case-based reasoning did not (H_2=5.54; *P*=.06). The verification association was unaffected; among the 97 remaining AI users, cross-check frequency remained associated with total score (β=2.97, 95% CI 1.05-4.89; *P*=.002) and hallucination-trap score (β=1.12, 95% CI 0.30-1.95; *P*=.008; model 8). We therefore read the group performance differences as precision-limited and sensitivity-dependent and give interpretive weight to the calibration and verification findings. Finally, removing the one weak hallucination-trap item (Q21) and rescoring the domain on the remaining 9 items left the group difference unchanged in direction and significance (Kruskal-Wallis H_2=9.67; *P*=.008; currently taking vs no course exposure adjusted: β=0.86; *P*=.04).

Discussion

Principal Findings

This 3-group, single–time point comparison yielded 4 convergent associations, and we state at the outset how far they reach. The association tests were prespecified hypothesis tests, but the causal and curricular inferences that can be drawn from them are hypothesis-generating, and the data do not by themselves justify a curriculum or assessment recommendation. First, examination performance in an AI-permitted assessment was associated with verification behavior, not with usage volume. Self-reported cross-check frequency was associated with overall performance; the association was unchanged by adjustment for demographics and course group and by excluding

XSL•FO
RenderX

fast completions, and it attenuated when basic knowledge was added, with the fully adjusted interval including zero. AI-use quantity showed no positive association. The association was not specific to hallucination-trap items, because it did not survive adjustment for case-based reasoning, and is therefore best read as an association with general clinical reasoning. One interpretation consistent with the data is that students who routinely cross-check draw on the same operations—source triangulation, internal consistency checking, and rejecting plausible-but-unsupported claims [25]—that case vignettes also reward; the design cannot confirm that these operations occur or that verification is a domain-general epistemic habit rather than a hallucination-specific skill. The disattenuated case-hallucination correlation (ρ=0.71) and the loss of the hallucination-specific signal after adjustment are at least compatible with this account. Hallucination-trap items appear to load on the same latent reasoning competence as clinical cases.

Second, metacognitive calibration, rather than accuracy, was the dimension on which students were weakest. Signed bias was negative in every group, calibration gaps were large in every group (0.39-0.48), and confidence discriminated correct from incorrect responses only moderately. On the confirmatory per-participant Brier test, and convergently on per-participant ECE, calibration was poorest in the completed-course group; this between-group contrast was robust to both sensitivity exclusions but rests on 27 students and is estimated imprecisely. Three interpretations of the completed-group result are mutually compatible and not adjudicable in a cross-sectional design: self-selection, because students who elected the course in an earlier semester are a small, nonrandom subset. Another possibility is an overfamiliarity or illusion-of-knowing account, under which prior coursework might be associated with fluency mistaken for accuracy and a weaker confidence-accuracy coupling [17]; and the unskilled-and-unaware pattern, in which weaker performers fail to recognize the limits of their own competence [18,26]. Consistent with a coupling-loss rather than a uniform-overconfidence account, the completed group's deficit is better described as degraded confidence-accuracy correspondence than as global overconfidence. We read these results through a metacognitive, confidence-calibration lens [17,18]: accuracy and the calibration of confidence to that accuracy are distinct constructs, and in an AI-assisted workflow calibration may be especially important [27], because unwarranted certainty about a fluent wrong answer could allow an LLM error to propagate to a patient-safety event [28]—a pathway beyond the scope of this study.

Third, basic factual knowledge did not differ by exposure group (H_2=0.61; P=.74), as the framework predicted. This null is consistent with, but does not establish, the absence of large generic cohort or ability differences behind the AI-sensitive results, and it was not powered as an equivalence test (basic-knowledge difficulty was also somewhat ceiling-restricted (mean 0.780, SD 0.075), making this a coarse rather than fully sensitive control). The associations between active course enrollment and case-based reasoning and hallucination-trap performance, by contrast, were precision-limited and sensitivity-dependent. They attenuated to nonsignificance when 11 completions under 5 minutes were excluded.

Fourth, self-reported readiness and attitudes were not associated with examination accuracy, which suggests that attitude-only evaluations of AI curricula may miss performance-relevant variance. The hallucination-trap domain behaved as an internally consistent but provisional construct: its high disattenuated correlation with case-based reasoning and the loss of hallucination-specific signal after adjustment indicate that it is, at present, best understood as a safety-weighted facet of clinical reasoning rather than a psychometrically separable ability. Methodological strengths include a behaviorally meaningful outcome set, robust (HC3) and effect-size-oriented reporting, prespecified primary outcomes and exclusion criteria, a basic-knowledge negative control, prespecified sensitivity analyses, and open data and code.

Comparison With Prior Work

These findings extend the same-institution randomized trial of AI-PBL in physiotherapy [14], which established short-term efficacy, by pointing to a learner behavior—verification—that is associated with better examination performance (a proxy we interpret, cautiously, as relevant to safer AI use); whereas that trial asked whether structured AI integration helps, this study asks which learner process is associated with better performance and identifies cross-checking rather than usage volume. The verification-over-quantity result sharpens a theme running through the AI-in-education evidence base. The BEME-aligned scoping review of this literature emphasized accelerating adoption alongside a fragmented outcome base that rarely reaches beyond knowledge acquisition [9], and these cross-check findings speak directly to that gap by tying performance to a behavior rather than to exposure alone. It aligns most closely with the structured-feedback trials: Brügge et al [10] found that structured feedback loops, not mere LLM exposure, drove reasoning gains, and the present association between verification and performance—paired with the absence of a positive association for usage quantity—is an observational parallel, not causal evidence, to that contrast between active, checked engagement and passive volume. Hui et al [11] reported higher postcourse knowledge scores with ChatGPT-assisted PBL, and Chan et al [12] reported favorable student ratings of simulated history-taking. Neither study isolated the contribution of AI-use quantity.

Feigerlova et al [13] made the point most explicitly, noting that few AI-education interventions teach students to verify AI output. Consistent with this, a cross-sectional survey of medical students found that 62.3% (165/265) recognized that ChatGPT answers require verification, indicating that cross-checking is acknowledged but unevenly practiced among trainees [29]. This study adds a behavioral correlate to that picture: verification was not merely endorsed but was the variable whose association with performance persisted after adjustment for demographics and course group, whereas usage volume showed none. The hallucination findings connect to the misinformation-risk literature, which frames LLM outputs as double-edged and prone to fabricated, outdated, or jurisdiction-specific errors [15,16]. A comparative study measured hallucination rates

XSL•FO
RenderX

ranging from roughly one-third to over half of medical question-answering responses [30]. Our result that hallucination-trap performance was not separable from general case-based reasoning (disattenuated: ρ=0.71) qualifies the implicit assumption in that literature that hallucination recognition is a stand-alone competence, suggesting instead that it is a safety-weighted expression of clinical reasoning.

The calibration result is where this study departs most from prior emphasis. The AI-education trials [10-12,14] and reviews [9,13] foreground accuracy and knowledge gain and are largely silent on whether learners' confidence is calibrated to that accuracy. By contrast, our data suggest that the more salient vulnerability may lie in metacognitive calibration rather than accuracy, connecting the long-standing medical-education concern with confidence-accuracy dissociation and the unskilled-and-unaware pattern [17,18,31] to AI-assisted reasoning, where unwarranted certainty about plausible-but-wrong AI output is itself a patient-safety risk [15,16]. Finally, the dissociation between self-reported readiness and attitude scales and objective accuracy echoes the calibration tradition at the construct level [17,18] and cautions that the survey-based outcomes common in this literature [29] may index disposition rather than the verification and calibration behaviors that may be more closely tied to safe AI use.

The associations reported here were tested as prespecified hypotheses, but they are hypothesis-generating with respect to causal and curricular inference, and we therefore make no curriculum or assessment recommendation on the basis of these data. What the data do is specify, with effect sizes and precision, 2 candidate instructional targets. These are verification behavior, whose association with performance was robust to adjustment for demographics and course group and to the exclusion of fast completions but attenuated, to an estimate of borderline precision, once a competence baseline was added, and metacognitive calibration, which differed by course status on the confirmatory participant-level Brier test. Both targets are operationalized by the instruments provided in Multimedia Appendix 3 and can be tested directly. Four testable hypotheses follow from the specific limits of this design. First, whether explicit verification training (structured cross-checking routines) improves AI-assisted reasoning relative to a matched active-AI comparator. This requires a randomized or stepped-wedge trial that follows the same cohort over time, so that the temporal ordering this study cannot establish becomes identifiable. Second, whether verification measured as an observed process—logged source consultations, prompt-and-recheck sequences, or answer revision after cross-checking—shows the same association as the single self-report item used here, which is open to reverse causation. Third, whether per-item confidence-accuracy feedback tightens confidence-accuracy correspondence [32], with the per-participant Brier score and calibration gap as prespecified, participant-level end points. Fourth, whether pairing hallucination-style items with item-level confidence ratings yields assessment information about calibration that accuracy-only scoring misses, and whether this information, rather than attitudinal self-report, tracks safe AI use. These questions are timely because structured AI-literacy training remains uncommon relative to student AI use—in one survey only 5.3% of medical students reported curricular AI education despite widespread informal use [33]—and published AI curricula for medical learners remain scarce and largely atheoretical [34].

Limitations

Several limitations temper interpretation. First and foremost, the cross-sectional design compares distinct cohorts at a single time point and cannot establish change, decay, or causation. Observed differences may reflect cohort effects, self-selection, prior exposure, or unmeasured confounding, and longitudinal follow-up is needed. Relatedly, the 3 exposure categories are a composite proxy that conflates instructional dose, recency, and self-selection, which the design cannot separate, so group differences attach to the exposure category as a whole rather than to instruction per se. Second, verification was measured with a single, postexamination, self-reported item and is therefore open to reverse causation. More competent students may both verify more and score higher. Adjustment for a competence baseline attenuated the association, and the fully adjusted CI included zero. Third, the completed-course group was small (n=27) and self-selected; its calibration findings are imprecise (eg, accuracy at maximum confidence, with a 95% CI of 44.8%-93.2%), may partly reflect cohort composition, and the study was not powered to establish equivalence for the completed-group performance nulls, which should be read as "no detectable difference," not "no difference." Fourth, the 30-item examination is a researcher-developed, study-specific instrument that was not validated by an external expert panel, pretested, or compared with a criterion measure. The validity evidence reported here (content blueprint and failure-mode mapping, internal consistency, item discrimination, and known-groups and confidence-accuracy relations) is therefore internal to this study, and the hallucination-trap domain's discriminant validity from case-based reasoning is unproven (disattenuated: ρ=0.71). Neither the verification association nor the active-exposure advantage on hallucination-trap items remained after adjustment for case-based reasoning, so both relate to general clinical reasoning rather than to a hallucination-specific ability. Fifth, all calibration metrics depend on a confidence-to-probability mapping; reassuringly, the group ordering was preserved under an alternative chance-floored mapping, and the confirmatory basis is the participant-level Brier test rather than the response-level descriptives, which are uncorrected and read as convergent evidence only. Sixth, AI literacy as a latent trait was not validly measured (the AILS was unreliable), so "AI literacy" here denotes course exposure only. Seventh, the performance advantages of currently taking students were sensitive to excluding very fast responders and are precision-limited. Eighth, multiple-choice items may understate the ability to reject plausible-but-incorrect content; open-ended formats are a valuable direction. Finally, participation was 71% (137/193) of those invited and was concentrated in year 2; the cited same-institution randomized trial [14] raises a possibility of participant overlap and shared local context that we could not formally exclude, and generalizability beyond a single elective at one institution is limited.

XSL•FO
RenderX

Conclusions

Among physiotherapy students sitting a realistic, open-resource clinical examination, performance was associated with verification behavior rather than with the quantity of AI use, and metacognitive calibration, not accuracy, was the dimension on which students were weakest, most markedly in the small completed-course group. These are prespecified, participant-level associations from 137 (71%) of 193 invited students, with effect sizes and precision reported. They identify verification and calibration as candidate instructional targets to be tested in longitudinal and multi-institution designs, not as curriculum or assessment recommendations, which the cross-sectional, self-report design cannot justify.

Acknowledgments

The authors thank the physiotherapy students who participated and the Department of Physiotherapy and Rehabilitation at Istanbul Medipol University for logistical support.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

A deidentified participant-level dataset with free-text responses and absolute time stamps removed, a data dictionary, and the analysis code are provided as Multimedia Appendix 4. The full dataset, including qualitative free-text responses (time stamps still removed), is available from the corresponding author on reasonable request, subject to institutional data-sharing agreements and the Turkish Personal Data Protection Law. Requests should be directed to the corresponding author (dkurc@medipol.edu.tr) and will be reviewed by the institutional ethics committee; approved requests require a signed data-sharing agreement.

Generative AI Use

Generative AI tools (Claude Fable 5 [Anthropic Inc] and ChatGPT GPT-5.5 [OpenAI]) were used to develop the Python and R analysis and figure code from author-specified analyses and to provide English-language editing. ChatGPT was additionally used to review the draft for clarity. The tools did not generate, alter, or interpret data; all code was executed, and all results were verified by the authors, and the code is provided in Multimedia Appendix 4. All authors take full responsibility for the content of the manuscript.

Authors' Contributions

Conceptualization: DK
Methodology: DK, ED, ANT
Software: DK
Formal analysis: DK
Data curation: DK
Investigation: ED
Supervision: ANT
Validation: ANT
Writing – original draft: DK
Writing – review & editing: ED, ANT
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods and psychometrics — conceptual framework schematic (Figure S1); item-level difficulty, discrimination, and reliability; calibration detail by group; and Artificial Intelligence Literacy Scale reliability.
[DOCX File , 161 KB-Multimedia Appendix 1]

Multimedia Appendix 2

STROBE checklist for cross-sectional studies.
[DOCX File , 38 KB-Multimedia Appendix 2]

XSL•FO
RenderX

Multimedia Appendix 3

Researcher-developed study instruments (English translation and original Turkish): demographic form, 30-item clinical case examination, item-level confidence rating, and postexamination AI-use survey.
[DOCX File , 35 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Deidentified participant-level dataset (free-text and absolute timestamps removed), data dictionary, and analysis code, including the revision analyses (r1_additions.py).
[ZIP File (Zip Archive), 48 KB-Multimedia Appendix 4]

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6):887. [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
3. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ. 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
4. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med. 2024;99(1):22-27. [FREE Full text] [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
5. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ. 2023;9:e48785. [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
6. Charow R, Jeyakumar T, Younus S, Dolatabadi E, Salhia M, Al-Mouaswas D, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ. 2021;7(4):e31043. [FREE Full text] [doi: 10.2196/31043] [Medline: 34898458]
7. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach. 2023;45(6):574-584. [FREE Full text] [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
8. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
9. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024;46(4):446-470. [FREE Full text] [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
10. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ. 2024;24(1):1391. [FREE Full text] [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
11. Hui Z, Zewu Z, Jiao H, Yu C. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. BMC Med Educ. 2025;25(1):50. [FREE Full text] [doi: 10.1186/s12909-024-06321-1] [Medline: 39799356]
12. Chan KY, Yuen TH, Co M. Using ChatGPT for medical education: the technical perspective. BMC Med Educ. 2025;25(1):201. [FREE Full text] [doi: 10.1186/s12909-025-06785-9] [Medline: 39920711]
13. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ. 2025;25(1):129. [FREE Full text] [doi: 10.1186/s12909-025-06719-5] [Medline: 39871336]
14. Ergezen Sahin G, Aras Bayram G, Sanchez Sierra A, Akdemir S, Kurc D, Tarakci D, et al. Effects of artificial intelligence based physiotherapy educational approach in developing clinical reasoning skills: a randomized controlled trial. BMC Med Educ. 2025;25(1):1378. [FREE Full text] [doi: 10.1186/s12909-025-07926-w] [Medline: 41068907]
15. 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. 2023;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
16. Alkaissi H, McFarlane S. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023;15(2):e35179. [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
17. Koriat A. The self-consistency model of subjective confidence. Psychol Rev. 2012;119(1):80-113. [doi: 10.1037/a0025648] [Medline: 22022833]
18. Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. Acad Med. 2001;76(10 Suppl):S87-S89. [doi: 10.1097/00001888-200110001-00029] [Medline: 11597883]
19. Edwards I, Jones M, Carr J, Braunack-Mayer A, Jensen GM. Clinical reasoning strategies in physical therapy. Phys Ther. 2004;84(4):312-330. [Medline: 15049726] [doi: 10.1093/ptj/84.4.312]

XSL•FO
RenderX

20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147(8):573-577. [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
21. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ. 2021;21(1):112. [FREE Full text] [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
22. Süleymanoğulları M, Özdemir A, Tekin A. Yapay Zekâ Ölçeği: Geçerlik ve güvenirlik çalışması [Artificial Intelligence Scale: a validity and reliability study]. Educ Sci Sport. 2024;6(1):13-27. [FREE Full text]
23. Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: an exploratory review. Comput Educ Artif Intell. 2021;2:100041. [doi: 10.1016/j.caeai.2021.100041]
24. Polatgil M, Güler A. Yapay zekâ okuryazarlığı ölçeğinin Türkçe'ye uyarlanması: adaptation of Artificial Intelligence Literacy Scale into Turkish. Sosyal Bilimlerde Nicel Araştırmalar Dergisi. 2023;3(2):99-114. [FREE Full text]
25. Wineburg S, McGrew S. Lateral reading and the nature of expertise: reading less and learning more when evaluating digital information. Teach Coll Rec. 2019;121(11):1-40. [doi: 10.1177/016146811912101102]
26. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol. 1999;77(6):1121-1134. [doi: 10.1037//0022-3514.77.6.1121] [Medline: 10626367]
27. Lee D, Pruitt J, Zhou T, Du J, Odegaard B. Metacognitive sensitivity: the key to calibrating trust and optimal decision making with AI. PNAS Nexus. 2025;4(5):pgaf133. [FREE Full text] [doi: 10.1093/pnasnexus/pgaf133] [Medline: 40417078]
28. 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):121-127. [FREE Full text] [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
29. Alkhaaldi SMI, Kassab CH, Dimassi Z, Oyoun Alsoud L, Al Fahim M, Al Hageh C, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ. 2023;9:e51302. [FREE Full text] [doi: 10.2196/51302] [Medline: 38133911]
30. Wang D, Ye J, Li J, Liang J, Zhang Q, Hu Q, et al. Enhancing large language models for improved accuracy and safety in medical question answering: comparative study. JMIR Med Educ. 2025;11:e70190. [FREE Full text] [doi: 10.2196/70190] [Medline: 41329953]
31. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA. 2006;296(9):1094-1102. [doi: 10.1001/jama.296.9.1094] [Medline: 16954489]
32. Staal J, Katarya K, Speelman M, Brand R, Alsma J, Sloane J, et al. Impact of performance and information feedback on medical interns' confidence-accuracy calibration. Adv Health Sci Educ Theory Pract. 2024;29(1):129-145. [FREE Full text] [doi: 10.1007/s10459-023-10252-9] [Medline: 37329493]
33. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ. 2024;10:e51247. [FREE Full text] [doi: 10.2196/51247] [Medline: 38180787]
34. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ. 2024;10:e54793. [FREE Full text] [doi: 10.2196/54793] [Medline: 39023999]

Abbreviations

AILS: Artificial Intelligence Literacy Scale
AIS: AI Attitude Scale
AUC: area under the receiver operating characteristic curve
BEME: Best Evidence Medical Education
ECE: expected calibration error
ECTS: European Credit Transfer and Accumulation System
GPA: grade point average
HC3: heteroscedasticity-consistent (type 3)
KR-20: Kuder-Richardson Formula 20
LLM: large language model
MAIRS-MS: Medical Artificial Intelligence Readiness Scale for Medical Students
OLS: ordinary least squares
PBL: problem-based learning
RQ: research question
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VIF: variance inflation factor

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 29.Jun.2026; peer-reviewed by SS Shah, A El Khatib; comments to author 08.Oct.2026; revised version received 11.Oct.2026; accepted 29.Oct.2026; published 27.Nov.2026

Please cite as:
Kurç D, Dansuk E, Tunalı AN
Verification Behavior and Metacognitive Calibration in AI-Assisted Clinical Reasoning Among Physiotherapy Students: Cross-Sectional, 3-Group Comparative Study
JMIR Med Educ 2026;12:e105832
URL: https://mededu.jmir.org/2026/1/e105832
doi: 10.2196/105832
PMID:

©Doğukan Kurç, Emre Dansuk, Ayşe Nur Tunalı. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.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 JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX