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.

Effect of Guided ChatGPT Use on Documented Clinical Reasoning and Its Retention Among Emergency Medicine Students: A Randomized Controlled Trial

Hayate Errifaiy

h.errifaiy@uca.ac.ma

Cadi Ayyad University

Hicham Maghtoum

Cadi Ayyad University

Manal Rhezali

Cadi Ayyad University

Naila Boukoub

Cadi Ayyad University

Hicham NEJMI

Cadi Ayyad University

Research Article

Keywords: Clinical reasoning, Generative artificial intelligence, ChatGPT, Medical education, Knowledge retention, Randomized controlled trial, Emergency medicine

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10463281/v1

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

Additional Declarations: No competing interests reported.

Effect of Guided ChatGPT Use on Documented Clinical Reasoning and Its Retention Among Emergency Medicine Students: A Randomized Controlled Trial

Authors

Hayate Errifaiy, MD[1,2] (Corresponding author),Hicham Elmaghtoum, MD[1,2],

Manal Rhezali, MD[1,2], Naila Boukoub, MD[2,3], Hicham Nejmi, MD[1,2]

Affiliations

1: Department of Anesthesia and Intensive Care Medicine, Emergency Department, Mohammed VI University Hospital of Marrakech, Marrakech, Morocco

2: Faculty of Medicine and Pharmacy, Cadi Ayyad University, Marrakech, Morocco

3: Medical Intensive Care Unit, Ar-Razi Hospital, Mohammed VI University Hospital of Marrakech, Marrakech, Morocco

Corresponding Author

Hayate Errifaiy, MD (corresponding author)
Department of Anesthesia and Intensive Care Medicine, Emergency Department
Mohammed VI University Hospital of Marrakech
Faculty of Medicine and Pharmacy, Cadi Ayyad University
Marrakech, Morocco
Email: h.errifaiy@uca.ac.ma Phone: +212 666 928 926

ABSTRACT

Background: Clinical reasoning is a core competency in emergency medicine, and its structured teaching remains challenging. Generative artificial intelligence (ChatGPT) may represent a useful educational tool, but its impact on learning consolidation beyond immediate benefit remains uncertain. We assessed whether guided use of ChatGPT improves immediate documented clinical reasoning performance without compromising its retention over time.

Methods: A single-center, two-arm, parallel-group, exploratory randomized controlled trial was conducted in the emergency department of Mohammed VI University Hospital, Marrakech, Morocco. Sixty medical students (5th–6th year) were randomized 1:1, using sequentially numbered, opaque, sealed envelopes generated by an independent investigator, to a ChatGPT group (simulated clinical cases with guided GPT-4o assistance) or a control group (no assistance); outcome assessors were blinded to group and time point, although participants could not be blinded to the intervention received. The primary outcome was the total Revised-IDEA (R-IDEA) score (/10), measured immediately after the intervention (T0) and again one month later under unassisted conditions (T1), analyzed by analysis of covariance (ANCOVA) adjusted for the T0 score.

Results: All 60 participants were followed at T0 and T1, with no loss to follow-up. At T0, the R-IDEA score was significantly higher in the ChatGPT group (median 9.50 vs 8.50; $p < 0.001$; $r = 0.96$). At T1, the pattern reversed: the ChatGPT group showed a significant decline (median $\Delta = -0.75$), whereas the control group improved ($\Delta = +1.00$; $p < 0.001$; $r = 0.96$). ANCOVA confirmed a group effect at T1 favoring the control group (adjusted difference = 0.48 points, 95% CI 0.16–0.80; $p = 0.004$; partial $\eta^2 = 0.136$). Perceived confidence and benefit in the ChatGPT group were not correlated with actual performance change.

Conclusions: Conclusions: Guided ChatGPT assistance improves immediate documented clinical reasoning performance, but this benefit does not persist at one month and is accompanied by poorly calibrated learner confidence. These findings support structured pedagogical oversight of generative AI in medical education, combined with active learning-consolidation strategies to ensure durable skill acquisition.

Trial registration: ISRCTN registry ISRCTN13261015, registered 3 August 2026. Retrospectively registered.

Keywords: Clinical reasoning; Generative artificial intelligence; ChatGPT; Medical education; Knowledge retention; Randomized controlled trial; Emergency medicine.

BACKGROUND

Clinical reasoning is a core competency in medical training, and particularly so in emergency medicine. It engages complex cognitive processes both analytical and non-analytical that allow clinicians to gather and interpret a patient's clinical data, generate diagnostic hypotheses, and guide therapeutic management [1,2]. In emergency medicine, these processes operate under especially demanding conditions, characterized by time pressure, diagnostic uncertainty, and often incomplete information. Structured teaching and assessment of clinical reasoning nevertheless remain difficult, owing in particular to limited supervision time and heterogeneous feedback provided to students; validated instruments for assessing documented clinical reasoning, such as the Revised-IDEA (R-IDEA) rubric, have been developed to address this need for objective measurement [3], building on earlier structured frameworks for assessing diagnostic reasoning in written documentation [4].

The emergence of generative artificial intelligence models, notably ChatGPT developed by OpenAI, offers a new tool with the potential to influence learning mechanisms in medical education. The clinical reasoning capabilities of these large language models have been evaluated in several studies [5], including formal assessments of their performance on medical licensing examinations [6], and their use as a source of pedagogical feedback for medical students has been explored in recent randomized controlled trials [7,8]. These developments raise the prospect of an

accessible, scalable, and personalized digital tutor, particularly relevant in settings where supervisory resources are limited.

While these tools may improve the structuring of reasoning in the short term, their impact on knowledge consolidation and cognitive autonomy over the medium term remains uncertain. Growing concerns have indeed been raised in the literature regarding cognitive offloading and "metacognitive laziness," mechanisms by which learners may delegate the effort of reflection to the tool, to the detriment of their cognitive engagement and self-regulation [9–11]. Moreover, most available studies have assessed immediate or short-term performance, most often under assisted conditions, and rarely distinguish the effect of immediate assistance from retention measured later under unassisted conditions [7–9]. Data also remain scarce in emergency medicine and in resource-limited settings.

It is therefore essential to determine whether the use of ChatGPT improves immediate performance without compromising knowledge retention over time. To this end, we designed a randomized controlled trial evaluating the effect of guided ChatGPT use on the quality of documented clinical reasoning among medical students, measured using the R-IDEA rubric [3] immediately after the intervention (T0) and again at one month under unassisted conditions (T1). This dual measurement design was specifically intended to distinguish the effect of immediate assistance from a potential durable learning effect. We hypothesized that ChatGPT assistance would improve immediate performance, with the persistence of this benefit over time constituting the central question of the study.

METHODS

1. Study design and setting

We conducted a single-center, two-arm, parallel-group, exploratory randomized controlled trial with two measurement time points, in the emergency department of Ar-Razi Hospital (Mohammed VI University Hospital, Marrakech). This was a minimal-risk educational study involving only simulated clinical cases, conducted over six weeks and preceded by a pilot study (see Section 3).

The primary objective was to evaluate the effect of guided ChatGPT use (GPT-4o, OpenAI) on the quality of documented clinical reasoning, measured at T0 (immediately after the intervention) and at T1 (one month later, under unassisted conditions), in order to distinguish the effect of immediate assistance from a durable learning effect. This trial is reported in accordance with the CONSORT 2010 statement for randomized controlled trials [27].

2. Sample size calculation

The required sample size was calculated a priori for the primary outcome (total R-IDEA score, comparison between independent groups). In the absence of local data and given the large effects reported in recent literature [7,8], a large effect size (Cohen's d = 0.80) was conservatively assumed. For $\alpha = 0.05$ (two-sided) and 80% power, the required sample was 26 participants per group, increased by 15% to anticipate loss to follow-up and the reduced efficiency of nonparametric tests, yielding 30 per group (n = 60). Sample size was calculated using G*Power software [12]. No interim analyses or stopping guidelines were planned, given the trial's short duration and fixed sample size.

3. Pilot study

A pilot study was conducted two weeks before enrollment among six students meeting the same eligibility criteria, to verify the feasibility of the protocol (clinical cases, R-IDEA rubric, duration, guided access to ChatGPT, questionnaire). As no difficulties were identified, the protocol was not modified; these six students were excluded from the final sample.

4. Participants and randomization

Sixty fifth- and sixth-year medical students, undertaking clinical rotations in emergency medicine or intensive care, were enrolled after providing free and informed consent. Non-inclusion criteria were refusal to participate and prior participation in the pilot study; absence from the one-month assessment constituted a secondary exclusion criterion.

Participants were randomized 1:1 into two groups of 30 (a ChatGPT group, solving cases with guided assistance from the tool, and a control group, without digital assistance). The allocation sequence was simple (unrestricted, with no stratification or blocking) and was generated and held by an independent investigator not otherwise involved in enrolment, intervention delivery, or outcome assessment, using sequentially numbered, opaque, sealed envelopes prepared in advance; envelopes were opened by the enrolling investigator only at the time of each participant's inclusion, concealing group allocation until assignment. Baseline comparability between the two groups was verified for age, sex, and year of study. Patients and the public were not involved in the design, conduct, or reporting of this study, which enrolled medical students as participants in a pedagogical intervention.

5. Intervention and procedure

The intervention consisted of a 30-minute session solving two simulated emergency medicine clinical cases (chest pain, respiratory distress) of intermediate complexity, identical for both groups and developed by two instructors before validation by a senior physician. The ChatGPT group had guided access to the tool (GPT-4o) for structuring their reasoning, generating

hypotheses, and prioritizing clinical elements, with standardized instructions asking them not to copy the tool's answers but to formulate their own reasoning; investigators provided no pedagogical assistance, their role being limited to ensuring protocol adherence. The control group solved the same cases without digital assistance.

At T0, immediately after the session, each participant solved, in writing and without assistance, two standardized cases distinct from those used in the session. At T1, one month later, a retention assessment was performed under the same conditions, with no access to ChatGPT for either group, using two new cases of comparable complexity, validated by consensus. Written productions were anonymized before scoring.

At the end of the intervention, participants in the ChatGPT group completed a 13-item, 5-point Likert-type satisfaction questionnaire exploring the perceived usefulness of the tool, decisional confidence, and motivation.

6. Primary outcome: the R-IDEA instrument

The quality of documented clinical reasoning was assessed using the R-IDEA (Revised-IDEA) rubric [3], validated across four domains: Interpretation (I, /4), Differential diagnosis (D, /2), Explanation (E, /2), and Argumentation (A, /2), for a total score out of 10 (higher = better quality). The rubric, previously used to assess clinical reasoning assisted by generative AI [5], was adapted for emergency medicine cases an adaptation tested without modification during the pilot study then translated into French for the purposes of this study and validated by consensus among investigators.

Each written production was scored independently by two anesthesiology and critical care physicians holding a postgraduate diploma in medical education, blinded to group and time point, using a scoring guide with anchors and a procedure for resolving disagreements. The score used was the mean of the two raters; inter-rater reliability was quantified at each time point (see Statistical analysis).

7. Statistical analysis

Analyses were performed using Jamovi (version 2.6.44, with a significance threshold of $p < 0.05$ (two-sided).

Qualitative variables are described as counts and percentages, and quantitative variables as mean ± standard deviation and median [Q1; Q3]. Baseline comparability between groups was tested using the χ^2 test (or Fisher's exact test) for categorical variables and the t-test or Mann-Whitney U test for age, depending on normality (Shapiro-Wilk test). As R-IDEA scores deviated significantly

from normality, notably owing to a ceiling effect, nonparametric tests were preferred, and scores are reported as median [Q1; Q3].

Inter-rater reliability for the total R-IDEA score was estimated using the intraclass correlation coefficient (ICC, two-way random-effects model, absolute agreement, single measurement) at T0 and T1, supplemented by raw percentage agreement and Cohen's kappa by domain, interpreted according to the thresholds of Koo and Li [13] (<0.50: poor; 0.50–0.75: moderate; 0.75–0.90: good; >0.90: excellent).

Between-group comparisons of R-IDEA scores (total and by domain) at T0, at T1, and of their change ($\Delta = T1 - T0$) were performed using the Mann-Whitney U test, and within-group changes using the Wilcoxon signed-rank test. Effect sizes are reported as rank-biserial correlation ($r \approx 0.10$: small; 0.30: moderate; ≥0.50: large) [14].

In accordance with recommendations for analyzing trials with pre- and post-intervention measurements [15], the primary retention analysis consisted of an ANCOVA on the total R-IDEA score at T1, with group as a fixed factor and the T0 score as a covariate, after verifying model assumptions (homogeneity of slopes, normality of residuals, Levene's test). Results are expressed as estimated marginal means (95% CI) and partial eta-squared (η^2p; 0.01: small; 0.06: medium; ≥0.14: large), with a nonparametric sensitivity analysis (comparison of Δ values). No subgroup analyses were pre-specified or performed.

Responses to the satisfaction questionnaire are described as counts and percentages by response category and compared, item by item, using the Mann-Whitney U test.

8. Ethical considerations

The study received a favorable opinion from the Marrakech University Hospital Ethics Committee (approval No. 046/2026, dated March 9, 2026). All procedures involving human participants were performed in accordance with the ethical standards of the institutional research committee and with the 1964 Declaration of Helsinki and its later amendments. All participants provided written free and informed consent; participation was voluntary, had no bearing on academic progression, and data were anonymized before analysis. This minimal-risk study involved only simulated clinical cases. No changes were made to the study protocol after the trial commenced. As this was a minimal-risk educational study, no harms or adverse events were anticipated or systematically assessed, and none were reported during the study. This trial was retrospectively registered with the ISRCTN registry, registration number ISRCTN13261015, on 3 August 2026 (see Trial registration, above). The full study protocol is available from the corresponding author upon reasonable request. For pedagogical fairness, an introductory session on ChatGPT was offered to the control group at the conclusion of the study.

RESULTS

1. Study population and baseline characteristics

Sixty students were enrolled and randomized into two groups of 30. Participants were enrolled between 12 and 22 April 2026. The T0 intervention and immediate assessment session was conducted for all participants on 30 April 2026, and the T1 retention assessment was conducted one month later, on 30 May 2026. All completed the intervention and the T0 and T1 assessments, with no loss to follow-up or missing data (Figure 1). The trial was completed as planned for all enrolled participants; no interim stopping rules were applied and the study was not terminated early. All 60 randomized participants were analysed in the group to which they were originally assigned (as-randomized). The groups were comparable at baseline in age, sex, and year of study (Table 1).

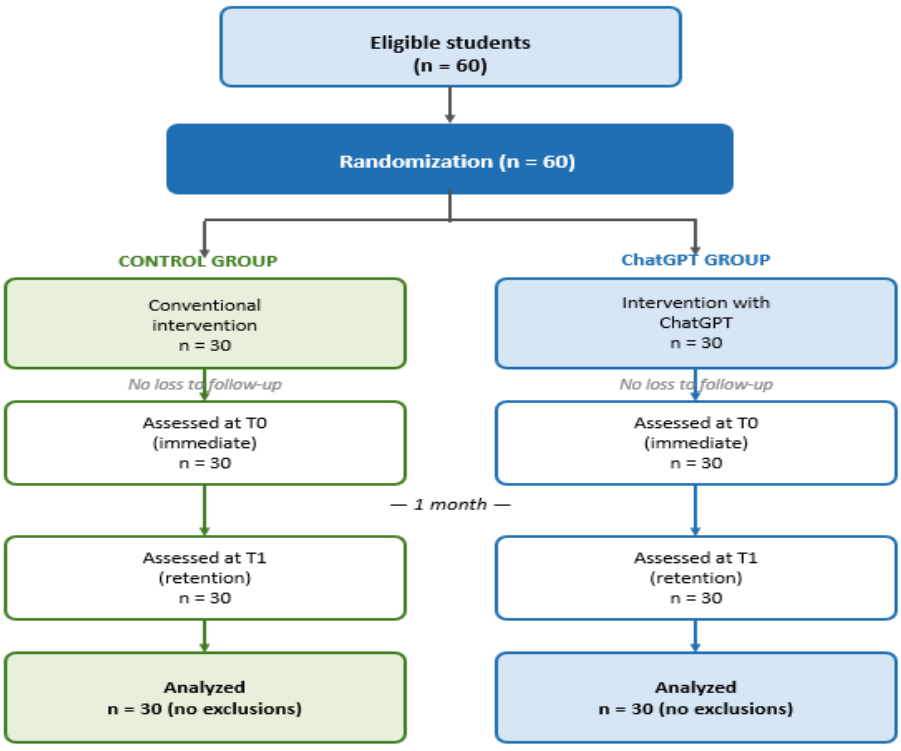


Figure 1. CONSORT flow : diagra All 60 randomized participants were assessed at T0 and T1 with no loss to follow-up or exclusion.

Table 1. Baseline characteristics of participants by randomization group.

Characteristic	ChatGPT group (n = 30)	Control group (n = 30)	p
Age (years), mean ± SD	22.8 ± 1.2	22.7 ± 1.1	0.768*
Female sex, n (%)	19 (63.3)	17 (56.7)	0.598**
Male sex, n (%)	11 (36.7)	13 (43.3)	
5th year, n (%)	21 (70.0)	24 (80.0)	0.371**
6th year, n (%)	9 (30.0)	6 (20.0)	

*SD: standard deviation. *Mann-Whitney U test. **Pearson's χ^2 test.*

2. Inter-rater reliability of R-IDEA scoring

Inter-rater reliability for the total R-IDEA score was moderate at T0 (ICC = 0.694) and poor at T1 (ICC = 0.105); this decline was attributable to a ceiling effect restricting score variance at T1 rather than to genuine scoring divergence: raw percentage agreement by domain remained high at both time points (65–95%), with the lowest kappa values corresponding to domains with near-zero variance (the kappa paradox) (Table 2). The score used for each participant was the mean of the two raters

Table 2. Inter-rater reliability of the R-IDEA rubric at T0 and T1.

Index	T0	T1
ICC total score (95% CI)	0.694 (0.535–0.805)	0.105 (−0.156–0.350)
Kappa Interpretation (I) [raw agreement]	0.707 [82%]	0.610 [80%]
Kappa Differential diagnosis (D) [raw agreement]	−0.03 [92%]*	0.375 [95%]
Kappa Explanation (E) [raw agreement]	0.503 [92%]	0.375 [95%]
Kappa Argumentation (A) [raw agreement]	0.584 [80%]	0.300 [65%]

*ICC: intraclass correlation coefficient (two-way random-effects model, absolute agreement, single measurement); CI: confidence interval; raw agreement: percentage of exact agreement between the two raters. *The kappa for domain D at T0 is artificially low despite 92% raw agreement, owing to near-absent variance (the kappa paradox).*

3. Immediate performance (T0)

At T0, the total R-IDEA score was significantly higher in the ChatGPT group (median 9.50 [9.50; 10.0] vs 8.50 [8.00; 8.88]; $p < 0.001$; $r = 0.96$), with a near-maximal effect size (Table 3). This superiority involved Interpretation, Argumentation, and, to a lesser extent, Explanation; Differential diagnosis, close to the maximum score in both groups, did not differ significantly. A

marked ceiling effect was observed in the ChatGPT group, whose upper quartile reached the maximum score of 10.

Table 3. R-IDEA scores at T0 (immediate performance) by group.

Score, median [Q1; Q3]	ChatGPT group	Control group	U	p*	r
Total score (/10)	**9.50 [9.50; 10.0]**	**8.50 [8.00; 8.88]**	**16.0**	**< 0.001**	**0.96**
Interpretation (I, /4)	4.00 [3.50; 4.00]	3.00 [3.00; 3.50]	155.5	< 0.001	0.65
Differential diagnosis (D, /2)	2.00 [2.00; 2.00]	2.00 [2.00; 2.00]	405.0	0.169	0.10
Explanation (E, /2)	2.00 [2.00; 2.00]	2.00 [1.63; 2.00]	330.0	0.003	0.27
Argumentation (A, /2)	2.00 [1.50; 2.00]	1.00 [1.00; 2.00]	210.0	< 0.001	0.53

**Mann-Whitney U test. r: rank-biserial correlation (effect size, absolute value); all significant differences at T0 favored the ChatGPT group. Domains D and E, with near-zero variance (ceiling effect), render the comparison of limited informativeness.*

4. One-month retention (T1) and score evolution

At T1, the pattern had reversed: the median total R-IDEA score was 9.00 [8.50; 9.00] in the ChatGPT group versus 9.25 [9.00; 9.50] in the control group ($p = 0.009$), favoring the control group. Within-group changes showed a significant decline in the ChatGPT group (median $\Delta = -0.75$) and a significant improvement in the control group (median $\Delta = +1.00$), a highly significant between-group difference ($p < 0.001$; $r = 0.96$), with individual distributions in the two groups showing virtually no overlap (Table 4, Figure 2). At the domain level, the improvement in the control group was driven by Interpretation, whereas the decline in the ChatGPT group involved Interpretation and Argumentation; domains already saturated at T0 (Differential diagnosis, Explanation) did not change significantly in either group.

Table 4. R-IDEA scores at T1 and T0→T1 change by group.

Total score	ChatGPT group	Control group	p between-groups*
T0, median [Q1; Q3]	9.50 [9.50; 10.0]	8.50 [8.00; 8.88]	< 0.001
T1, median [Q1; Q3]	**9.00 [8.50; 9.00]**	**9.25 [9.00; 9.50]**	**0.009**
Δ (T1 − T0), median [Q1; Q3]	**−0.75 [−1.00; −0.50]**	**+1.00 [0.50; 1.50]**	**< 0.001**
p within-group (Wilcoxon)	< 0.001	< 0.001	—

r within-group	1.00	0.97	—

**Mann-Whitney U test. Δ: change in total score between T0 and T1; Wilcoxon W: ChatGPT W = 325; control W = 6.00. r: rank-biserial correlation.*

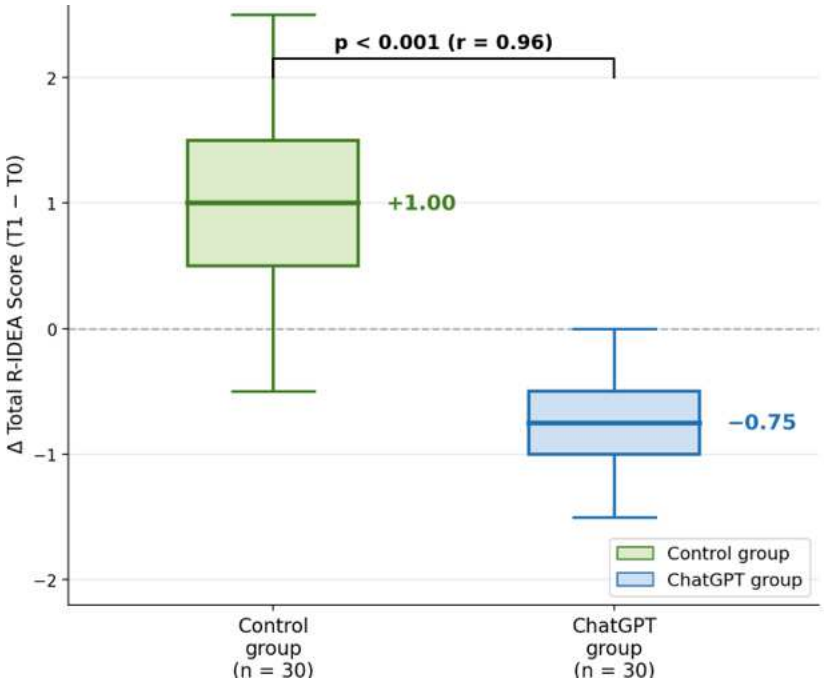


Figure 2. Distribution of the change in total R-IDEA score (Δ = T1 − T0) by group. The box represents the IQR [Q1; Q3]; the center line, the median; the whiskers, the extreme values. Control group: median Δ = +1.00 [+0.50; +1.50]; ChatGPT group: median Δ = −0.75 [−1.00; −0.50]. U = 17.5; p < 0.001; r = 0.96 (Mann-Whitney U test).

The primary ANCOVA analysis (total R-IDEA score at T1, adjusted for the T0 score) confirmed a significant group effect ($F(1,57) = 8.94$; $p = 0.004$; $\eta^2p = 0.136$, large effect), with estimated marginal means of 8.80 (95% CI: 8.61–8.99) in the ChatGPT group versus 9.28 (9.09–9.47) in the control group an adjusted difference of 0.48 points favoring the control group, with non-overlapping confidence intervals (Table 5). Model assumptions were satisfied, and the nonparametric sensitivity analysis confirmed this result.

Table 5. ANCOVA on the total R-IDEA score at T1, adjusted for the T0 score.

Parameter	Value
Group effect: F (df)	F(1, 57) = 8.94
p	0.004
η^2p	0.136

EMM ChatGPT group (95% CI)	8.80 (8.61–8.99)
EMM control group (95% CI)	9.28 (9.09–9.47)
Adjusted difference (control − ChatGPT), 95% CI	**+0.48 [0.16; 0.80]**
Group × T0 interaction (p)	0.346
Cohen's d (95% CI)	1.23 [0.375; 2.09]
Levene's test (p)	0.454
T0 covariate (p)	0.124

EMM: estimated marginal means; η²p: partial eta-squared; df: degrees of freedom; d: Cohen's d (post hoc comparison on estimated marginal means, t = 2.99). The T0 covariate did not reach significance (p = 0.124), reflecting the variance restriction related to the ceiling effect rather than an absence of association between initial and final performance.

5. Perception of the educational activity

All 60 participants (100%) completed the perception questionnaire. As internal consistency across the 13 items was insufficient to justify a composite score (Cronbach's $\alpha = 0.47$; McDonald's $\omega = 0.53$), each item was analyzed individually, without correction for multiple comparisons, in an exploratory approach.

Overall perception of the activity was favorable in both groups, with no significant difference on items related to the pedagogical design itself (Table 6). Six of the thirteen items differed significantly between groups: five favoring the ChatGPT group, with moderate-to-large effect sizes generation of relevant diagnostic hypotheses, confidence in analyzing clinical cases, appropriateness of cases to training level, adequacy of allotted time, and perceived improvement in clinical reasoning (all $p < 0.001$); and one favoring the control group the perception of having reasoned autonomously and in a structured manner ($p < 0.001$).

In an exploratory analysis, perceived benefit was positively correlated with the objective improvement in R-IDEA score in the control group ($\rho = 0.39$; $p = 0.034$), but not in the ChatGPT group (all $\rho \leq 0.22$; $p \geq 0.25$), suggesting a less well-calibrated perception of learning in the latter group: higher perceived confidence and benefit despite an objective decline in performance at one month, and reduced reasoning autonomy acknowledged by the participants themselves.

Table 6. Perception of the educational activity — between-group comparison of the 13 items.

Item	Control median [Q1;Q3]	ChatGPT median [Q1;Q3]	U	p*	r

The tool helped me better structure my clinical reasoning	4.00 [4.00;5.00]	5.00 [4.00;5.00]	373	0.204	0.17
The clinical cases were adapted to my training level	4.00 [4.00;5.00]	5.00 [5.00;5.00]	179	< 0.001	0.60
Knowledge acquired is transferable to other clinical situations	4.50 [4.00;5.00]	5.00 [4.00;5.00]	405	0.445	0.10
Instructions were clear and understandable	5.00 [4.00;5.00]	5.00 [5.00;5.00]	395	0.268	0.12
Time allocated to solve cases was appropriate	4.00 [3.00;4.00]	5.00 [4.00;5.00]	192	< 0.001	0.57
The CRAT rubric was easy to understand and use	4.00 [4.00;5.00]	5.00 [4.00;5.00]	384	0.263	0.15
I feel more confident in analyzing clinical cases after this activity	4.00 [3.00;4.00]	5.00 [4.25;5.00]	151	< 0.001	0.66
The activity allowed me to reason autonomously and in a structured way	5.00 [4.00;5.00]	4.00 [3.00;4.00]	222	< 0.001	0.51
The activity motivated me to deepen my clinical knowledge	4.00 [4.00;5.00]	5.00 [4.00;5.00]	337	0.066	0.25
I found the activity interesting and stimulating	4.00 [4.00;5.00]	4.00 [4.00;5.00]	411	0.501	0.09
The tool helped me generate relevant diagnostic hypotheses	4.00 [3.00;4.00]	4.00 [4.00;5.00]	143	< 0.001	0.68
The tool helped me better prioritize clinical elements	4.00 [4.00;5.00]	4.00 [4.00;5.00]	408	0.486	0.09
The tool contributed to improving my overall clinical reasoning quality	4.00 [3.00;4.00]	4.00 [4.00;5.00]	210	< 0.001	0.53

Mann–Whitney U test. Values are expressed as median [Q1–Q3]. r: effect size (rank-biserial correlation). Agreement percentages correspond to responses ≥4 on a 5-point Likert scale. Significance threshold: $p < 0.05$.

DISCUSSION

This randomized controlled trial reveals a marked dissociation between the immediate effect of ChatGPT assistance on the quality of documented clinical reasoning and its effect over time. At T0, the total R-IDEA score in the ChatGPT group was significantly higher than in the control group, with a near-maximal effect size. At one month, however, this benefit had reversed: the ChatGPT group declined significantly while the control group improved, a difference confirmed after adjustment for baseline score by analysis of covariance. This dual-measurement design suggests that the apparent benefit conferred by the tool reflects a transient boost to performance rather than genuinely consolidated learning.

The immediate superiority of the assisted group is consistent with recent randomized controlled trials evaluating ChatGPT as a feedback or facilitation tool in medical learning [7,8,16], as well as with a recent meta-analysis of randomized trials on generative AI–assisted medical teaching,

which reports a reproducible advantage for practical, short-term skills without a clear superiority for theoretical knowledge acquisition assessed over time [17]. Our results thus fit within an already documented trend: generative artificial intelligence improves performance under assisted conditions, without this benefit necessarily translating into durable learning gains.

The significant decline observed at one month in the ChatGPT group, contrasting with the improvement in the control group, is consistent with the cognitive offloading and "metacognitive laziness" hypotheses raised in the Introduction [9–11]. By delegating the structuring of reasoning, hypothesis generation, and prioritization of clinical elements to the tool, ChatGPT-group students may have been deprived of the active retrieval effort necessary for long-term memory consolidation, whereas the control group, required to fully mobilize their own reasoning capacities, may have benefited from a genuine learning effect through repeated unassisted practice, consistent with well-established evidence on the benefits of active retrieval practice for long-term retention [18,19]. This mechanism aligns with recent data suggesting that unstructured use of generative artificial intelligence in educational settings exposes learners to an erosion of critical thinking and cognitive autonomy [10,20,21], and echoes a recent randomized controlled trial that directly demonstrated, outside the medical field, reduced knowledge retention after ChatGPT-assisted learning compared with unassisted learning, with the tool described there as a "cognitive crutch" [22].

A second notable finding concerns the decoupling between subjective perception and objective performance : students in the ChatGPT group reported higher perceived confidence and benefit, uncorrelated with the actual change in their scores, while acknowledging reduced reasoning autonomy, whereas in the control group, perceived benefit was, by contrast, correlated with objective improvement. This dissociation evokes a confidence-calibration deficit already reported in human–AI interactions, consistent with broader evidence on complacency and bias in human reliance on automated systems [23], and reminiscent of the imperfect correspondence between self-assessed and observed competence long documented in health professions education [24], in which assistance perceived as effective can obscure, in the learner's eyes, the absence of genuine skill progression [25], a bias potentially detrimental to self-regulated learning.

Methodologically, the moderate inter-rater reliability at T0 and poor reliability at T1 are explained by a marked ceiling effect rather than by genuine scoring divergence, as evidenced by the raw percentage agreement remaining high at both time points; this observation, consistent with the kappa paradox described by Koo and Li [13], calls for caution when interpreting agreement indices when score variance is restricted. The primary ANCOVA analysis, robust to this phenomenon and consistent with the nonparametric sensitivity analysis, nonetheless reinforces the internal validity of the main result.

This study has several limitations. It was conducted at a single center, with a modest sample size, using simulated clinical cases whose ecological validity remains lower than that of real clinical situations. The retention follow-up, limited to one month, does not allow conclusions about longer-term evolution, nor about a possible catch-up effect in the ChatGPT group or, conversely, a widening of the gap between groups. The intensity and individual patterns of ChatGPT use were not quantified, precluding the establishment of a dose-effect relationship between the degree of tool use and the magnitude of the observed decline. Finally, since the assessment concerned written documented reasoning, the transferability of these findings to other dimensions of clinical competence, such as oral reasoning or real-time decision-making, remains to be demonstrated.

These findings argue for structured pedagogical oversight of generative artificial intelligence use in medical education, combining assisted access with active consolidation strategies active recall, spaced practice, reflective debriefing rather than unsupervised free use [20], particularly given documented concerns about the reliability and occasional inaccuracy of large language model outputs in clinical contexts [26]. Future studies, with larger samples, extended follow-up, and an objective measure of the level of tool use, appear necessary to define modalities for integrating these technologies that can reconcile immediate benefit with durable learning.

CONCLUSIONS

This randomized controlled trial shows that guided ChatGPT assistance significantly improves the immediate performance of documented clinical reasoning among medical students, but that this benefit is not maintained at one month: the assisted group declines while the unassisted group improves, resulting in a significantly lower retention score in the ChatGPT group. This dissociation between immediate gain and durable learning, combined with perceived confidence poorly calibrated to actual performance, cautions against equating a transient improvement obtained under digital assistance with genuine reinforcement of clinical reasoning skills.

In resource-limited training settings, where generative artificial intelligence represents a potentially valuable pedagogical aid, these findings call for guided and pedagogically structured use of these tools, combined with active learning-consolidation strategies, rather than unsupervised free access. Further studies, with larger samples, extended follow-up, and conducted in real clinical settings, appear necessary to clarify the modalities for integrating generative artificial intelligence that can reconcile immediate performance with the durable cognitive autonomy of future physicians.

DECLARATIONS

Ethics approval and consent to participate: The study received a favorable opinion from the Marrakech University Hospital Ethics Committee (approval No. 046/2026, dated March 9, 2026). All procedures involving human participants were performed in accordance with the ethical standards of the institutional research committee and with the 1964 Declaration of Helsinki and its later amendments. All participants provided written free and informed consent after being informed of the study's objectives and procedures.

Trial registration: This trial was registered with the ISRCTN registry, registration number ISRCTN13261015, on 3 August 2026 (https://doi.org/10.1186/ISRCTN13261015). As enrolment had already commenced at the time of registration, it was retrospectively registered.

Consent for publication: Not applicable (the manuscript contains no identifiable individual data).

Availability of data and materials: The datasets generated and/or analysed during the current study, as well as the study protocol, the R-IDEA scoring guide (adapted and translated version), and the simulated clinical case vignettes, are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: The authors declare that this study received no specific funding.

Authors' contributions:

Conceptualization: HE

Methodology: HN, HE

Data collection: HE,NB

Data analysis: HE

Writing original draft: HE

Writing review & editing: MR,HN, HE,NB

Supervision: HN,MR

Acknowledgements:

The authors thank the medical students who participated in this study, as well as the two evaluating physicians who independently and blindly scored the R-IDEA rubric. The authors also thank the staff of the Emergency Department of Ar-Razi Hospital, Mohammed VI University Hospital, Marrakech, for their support in conducting this study

REFERENCES

1. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005;39(1):98-106.

2. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003;78(8):775-780.

3. Schaye V, Miller L, Kudlowitz D, et al. Development of a clinical reasoning documentation assessment tool for resident and fellow admission notes: a shared mental model for feedback. J Gen Intern Med. 2022;37(3):507-512.

4. Baker EA, Ledford CH, Fogg L, Way DP, Park YS. The IDEA assessment tool: assessing the reporting, diagnostic reasoning, and decision-making skills demonstrated in medical students' hospital admission notes. Teach Learn Med. 2015;27(2):163-173.

5. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour RE, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med. 2024;184(5):581-583.

6. 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.

7. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. 2025;101(1195):458-463.

8. Kalam KA, Masoud FD, Muntaser A, Ranga R, Geng X, Goyal M. ChatGPT as a learning tool for medical students: results from a randomized controlled trial. Cureus. 2025;17(6):e85767.

9. Fan Y, Tang L, Le H, Shen K, Tan S, Zhao Y, et al. Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. Br J Educ Technol. 2025;56(2):489-530.

10. Gerlich M. AI tools in society: impacts on cognitive offloading and the future of critical thinking. Societies. 2025;15(1):6.

11. Lee HP, Sarkar A, Tankelevitch L, et al. The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). New York: ACM; 2025.

12. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175-191.

13. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163.

14. Kerby DS. The simple difference formula: an approach to teaching nonparametric correlation. Compr Psychol. 2014;3:11.IT.3.1.

15. Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323(7321):1123-1124.

16. Wang S, Zuo Y, Zou B, Liu G, Zhou J, Zheng Y, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach. 2024. doi:10.1080/0142159X.2024.2413023.

17. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. BMC Med Educ. 2025;25(1):1175.

18. Roediger HL 3rd, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. Psychol Sci. 2006;17(3):249-255.

19. Karpicke JD, Roediger HL 3rd. The critical importance of retrieval for learning. Science. 2008;319(5865):966-968.

20. 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.

21. Jose B, Cherian J, Verghis AM, Varghise SM, Mumthas S, Joseph S. The cognitive paradox of AI in education: between enhancement and erosion. Front Psychol. 2025;16:1550621.

22. Barcaui A. ChatGPT as a cognitive crutch: evidence from a randomized controlled trial on knowledge retention. Soc Sci Humanit Open. 2025;12:102287.

23. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010;52(3):381-410.

24. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. Acad Med. 2005;80(10 Suppl):S46-S54.

25. Bo JY, Wan S, Anderson A. To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). New York: ACM; 2025.

26. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res. 2025;14(1):e59823.

27. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.

Supplementary Files

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

- RIDEAAssessment.docx
- CONSORT2010Checklist.docx