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.

Human-first learning: scaffolding reasoning before AI assistance improves diagnostic accuracy

Yizhou Fan

fyz@pku.edu.cn

Peking University https://orcid.org/0000-0003-2777-1705

Luzhen Tang

The University of Hong Kong

Mengyu Xia

Peking University

Jiaqi Xu

Peking University

Zijian Li

Peking University

Naping Chen

Shantou University Medical College

Mingxue Xu

Peking University

Taolin Zhu

Peking University

Linfei Xiao

Peking University

Xinyu Li

Monash University

Oudong Xia

Shantou University Medical College

Changmin Lin

Shantou University Medical Colleg

Danijela Gasevic

The University of Hong Kong https://orcid.org/0000-0001-5976-4011

Lixiang Yan

Tsinghua University

Matthias Stadler

Technical University of Munich

Dragan Gasevic

The University of Hong Kong https://orcid.org/0000-0001-9265-1908

Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11417031/v1

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

Additional Declarations: There is **NO** Competing Interest.

Human-first learning: scaffolding reasoning before AI assistance improves diagnostic accuracy

Anonymised

Abstract

Generative AI could improve physicians' clinical performance, yet its use in medical training raises questions about whether trainees can develop the clinical reasoning needed to use AI effectively and independently. We conducted a longitudinal randomised controlled trial involving 496 medical students (389 analysed). Students completed clinical reasoning tasks with virtual standardised patients and were randomised to six conditions: no intervention, an AI assistant alone, AI scaffolding alone, scaffolding followed by the AI assistant, the AI assistant followed by scaffolding, or both simultaneously. AI scaffolding prompted students to reflect on and evaluate their reasoning without providing direct answers; the assistant provided requested information and answers. The study comprised an AI-free baseline, two intervention phases and a final assessment after all AI support was withdrawn. Frontier large language models (LLMs) completed the same tasks as an additional benchmark. Students initially performed below LLMs but later exceeded them following the interventions. Students who began with AI scaffolding achieved higher diagnostic accuracy than those given access to the AI assistant from the outset both during the subsequent intervention phase (75.2% vs 66.0%; OR 1.73, 95% CI 1.12–2.68) and after withdrawal (24.6% vs 14.4%; OR 2.19, 95% CI 1.15–4.14). After withdrawal, scaffold-first groups were the only groups to outperform both unsupported students (24.5–24.7% vs 9.0%; ORs 3.77–3.87, 95% CIs 1.25–12.00) and LLMs (24.5–24.7% vs 8.6%; ORs 3.48–3.50, 95% CIs 1.22–9.89). Introducing the AI assistant after scaffolding decreased intrinsic cognitive load ($\boldsymbol{t(50) = 3.25}$, $\boldsymbol{p_{BH} = .006}$, Cohen's $\boldsymbol{d_z = 0.46}$, 95% CI [0.17, 0.75]) and increased metacognitive evaluation ($\boldsymbol{t(50) = 3.11}$, $\boldsymbol{p_{BH} = .019}$, Cohen's $\boldsymbol{d_z = 0.44}$, 95% CI [0.15, 0.73]). These findings support a staged and human-first approach to AI use in medical training, in which reasoning-focused scaffolding precedes answer-providing assistance to strengthen diagnostic performance both during AI-supported learning and after AI support is withdrawn.

Main

Large language models (LLMs) can now perform increasingly complex medical reasoning tasks and, in some controlled studies, improve clinicians' diagnostic or management performance when used collaboratively [1–4]. These advances have moved clinical AI beyond examination-style benchmarks, although such benchmarks and other constrained evaluations still constitute a substantial part of the evidence base [5, 6]. Evaluations in more clinically realistic settings reveal important limitations in information gathering, guideline adherence and workflow integration [2], while erroneous or biased AI advice can reduce diagnostic accuracy [7]. Benefits from AI also vary substantially across users and tasks, and combining human and AI capabilities does not consistently outperform the stronger of the two acting alone [8, 9]. Governance and training frameworks have consequently struggled to keep pace with technical development [10, 11]. Recent work on meaningful human oversight has further argued that clinician involvement alone does not ensure effective oversight unless users can evaluate AI outputs and intervene when necessary [12]. These developments create an important tension for medicine. AI holds the potential to extend physicians' capabilities, while performance that approaches or exceeds clinicians on some tasks also raises questions about the continuing contribution of human expertise and how that expertise should be developed for AI-assisted practice [1, 9, 13].

Medical training is a critical setting for determining whether AI can extend future physicians' capabilities while preserving the development of clinical reasoning. Generative AI can improve learners' immediate performance by providing information, explanations and diagnostic assistance, but improved performance while AI is available does not establish that the underlying reasoning has been learned [14–16]. Recent experimental evidence also shows that novice medical students can be adversely affected by plausible but incorrect AI explanations, with reductions in diagnostic accuracy and impaired confidence calibration [7, 17]. These findings have raised concerns about de-skilling, in which previously acquired unaided performance deteriorates after AI exposure [18], and mis-skilling, in which erroneous AI guidance becomes incorporated into clinical judgment [7]. The key developmental concern is never-skilling, in which trainees may fail to acquire foundational reasoning when AI assumes functions before those capabilities have developed [19]. The recent *Nature Medicine* Perspective on never-skilling emphasises that direct evidence for this developmental risk in medical training remains limited and proposes that the educational effects of AI may depend on how and when it is introduced [19]. The unresolved question is therefore broader than whether AI improves students' performance while they use it. Medical education also needs evidence about what students learn through different forms of AI use and what they can subsequently do when AI is unavailable or changes.

The sequence in which learners encounter different forms of AI may influence whether improvements in AI-supported performance translate into gains in clinical reasoning that persist when AI is removed. Evidence outside medicine shows that unrestricted access to generative AI can improve performance during practice yet produce poorer subsequent performance when AI is removed, whereas designs that constrain direct answers and provide structured guidance can reduce this effect [20]. This finding is consistent with a broader distinction between performance achieved

through AI and capabilities that learners develop and internalise themselves [15, 21]. Clinical reasoning is particularly relevant to this distinction because learners need to gather and interpret evidence, revise diagnostic hypotheses, monitor uncertainty and evaluate their conclusions [22, 23]. Prematurely delegating these cognitive and metacognitive processes to AI may reduce learners' engagement in monitoring and evaluating their own reasoning [21, 24]. The metacognitive account of learning and performance in human–AI interaction therefore proposes that learners may benefit from first developing and exercising these capabilities before selectively offloading some task demands to AI [21]. This proposition remains to be established empirically in medical training. One way to test it is to compare the effects of an AI assistant that supplies information and answers with AI scaffolding that withholds direct answers and instead prompts learners to inspect evidence, monitor their reasoning and evaluate their conclusions [11, 23, 25].

Current evidence does not yet establish how different sequences of AI scaffolding and direct AI assistance shape clinical reasoning over time. Studies have increasingly examined what LLMs can accomplish independently, how physicians perform with AI, and how medical students perform while AI assistance is available [3, 6–8, 16]. These studies provide important evidence about immediate human and AI performance, but they do not determine whether the capabilities demonstrated with AI remain available after assistance is withdrawn. Evidence is especially limited on whether beginning with AI that supports learners' own reasoning produces different developmental outcomes from providing direct answers from the outset. The developmental consequences of early AI use have consequently been articulated more clearly conceptually than they have been tested longitudinally in medical training [19, 21]. Proposed responses increasingly emphasise preserving opportunities for independent reasoning and introducing AI in a staged manner [19, 26, 27], but causal evidence for such sequencing remains scarce. The sequence in which AI is introduced may also determine whether learners develop capabilities that complement AI sufficiently for combined human–AI performance to exceed either unsupported human or standalone AI performance, an outcome described as human–AI synergy [9].

We conducted a longitudinal randomised controlled trial to test how different forms and sequences of AI use shape the development of clinical reasoning. The study used challenging clinical reasoning tasks in which medical students gathered incomplete evidence from virtual standardised patients, compared competing diagnostic hypotheses and evaluated conclusions under uncertainty [28, 29]. Neither students nor frontier LLMs performed consistently well alone on these tasks, providing the human and AI benchmarks needed to assess whether subsequent human–AI interaction produced performance beyond either alone (Fig. 1a) [9, 13]. We compared two forms of AI intervention. The AI assistant provided requested information and answers, whereas AI scaffolding withheld direct answers and prompted learners to reflect on and evaluate their own reasoning (Fig. 1b). A total of 496 medical students were randomised to six conditions comprising no intervention (CN), AI assistant only (ASST), AI scaffolding only (SF), scaffolding followed by the AI assistant (SF-ASST), the AI assistant followed by scaffolding (ASST-SF), or simultaneous scaffolding and AI assistant (SF+ASST). Across four phases, students completed an AI-free baseline, an

initial intervention, a subsequent intervention and a final phase in which both interventions were withdrawn (Fig. 1c). Following data-quality screening, 389 participants were included in the analysis. This design allowed us to test whether AI scaffolding and the AI assistant improved clinical reasoning relative to unsupported learning, whether students could progress from performing below frontier LLMs to exceeding them, and whether beginning with scaffolding before introducing the AI assistant produced stronger subsequent and independent performance than access to the assistant from the outset. We examined cognitive load and metacognitive evaluation to assess how these sequences related to changes in learners' cognitive demands and monitoring of their own reasoning. The findings provide evidence that developing clinical reasoning before introducing an AI assistant can support both retained independent performance and subsequent human–AI synergy.

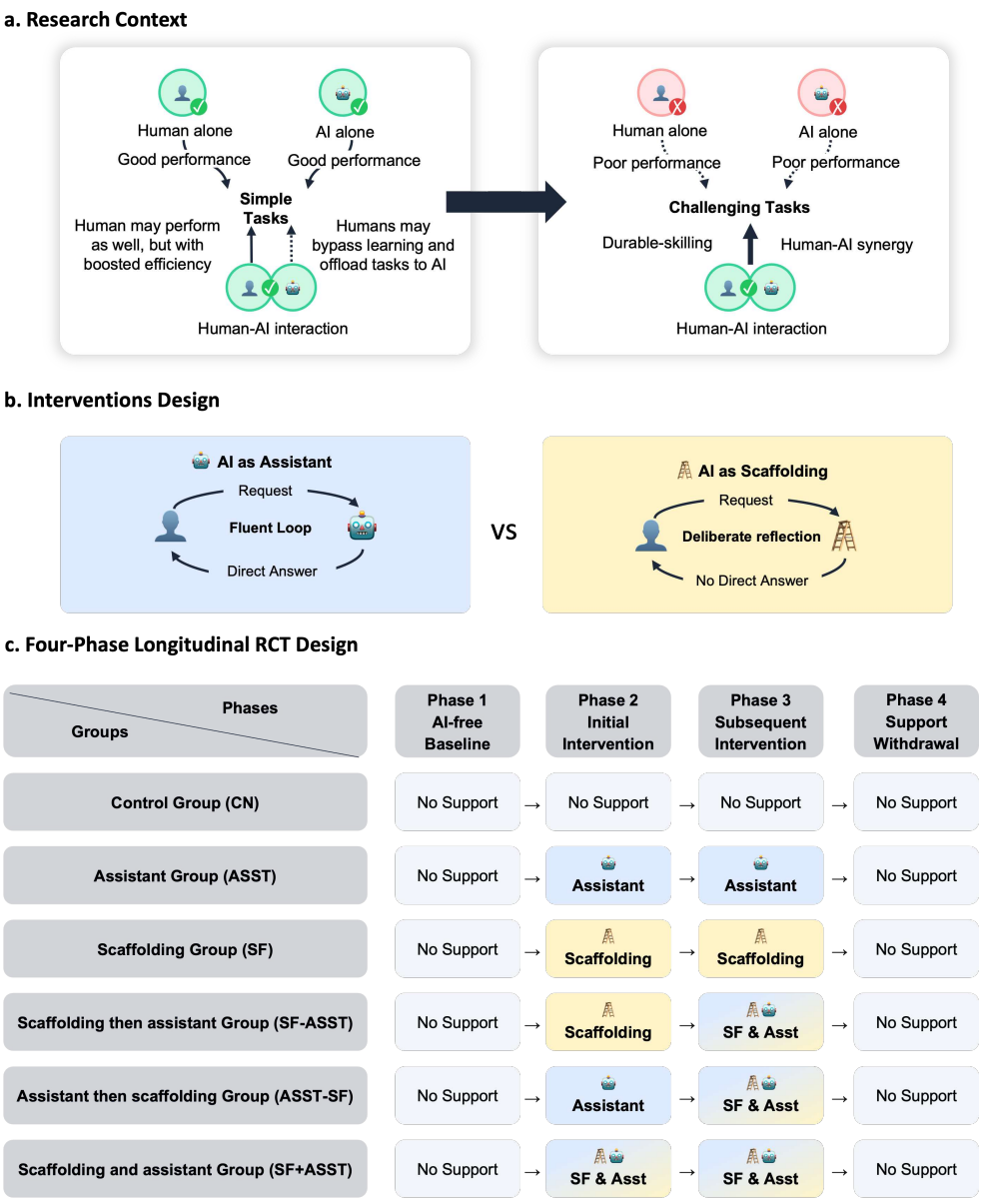


Fig. 1 Research design. a. Research context. Challenging tasks served as the research context of this study, referring to tasks in which neither medical students nor AI can independently achieve high performance. **b. Intervention design.** Two distinct AI-driven interventions were designed: the *AI as assistant* condition provided direct feedback by answering participants' questions. The *AI as scaffolding* condition was also AI-mediated but avoided providing direct answers, instead offering reflective prompts to support participants' metacognition. For example, the AI assistant could provide potential diagnoses or clinical reasoning clues when requested by participants; in contrast, scaffolding prompted participants to reflect on whether they had missed any relevant clinical information and guide their clinical reasoning. **c. Four-phase longitudinal RCT design.** The RCT consisted of four experimental phases, with participants assigned to one of six groups under different experimental conditions.

Results

Diagnostic performance of unsupported students and LLMs

To investigate human–AI synergy, we selected challenging clinical reasoning tasks that neither humans nor AI could solve well independently. These challenging tasks were selected from the Lecturio clinical reasoning case library (i.e., cases the platform marked as hard cases), which was originally developed by *The New England Journal of Medicine*. The cases were further adapted into a high-fidelity interactive environment developed by the authors: the simulation platform featured a virtual patient and two examination reporters, with whom participants engaged in open-ended dialogue to gather evidence and reach a diagnosis. Participants were fourth-year medical students who had completed their basic science curriculum and were in the early phase of clinical training, thus possessing foundational knowledge but limited clinical experience (See Supplementary Table 8 in the Appendix for student information). Case difficulty was confirmed through expert panel review and a preliminary pilot evaluation. Only cases in which both the pilot LLMs and students showed consistently low diagnostic accuracy were retained for the main study. The final LLM benchmark was then established through a separate, independent evaluation using a distinct and expanded set of LLMs that represented the strongest available systems at the time of the study, including GPT-5, GPT-o3, Claude Opus 4.1, Gemini 2.5 Pro, Grok-4, DeepSeek-R1, and Qwen-3-Max. Each model autonomously interacted with the same simulated roles, collected clinical information, and generated diagnoses 10 times (See Section 1.1.3 in Supplementary Information for LLMs benchmarking procedure). Across all selected cases (combined from Phases 1–4), the average diagnostic accuracy was 39.71% for the LLMs and 31.43% for the unaided human group (i.e., CN group). This overall low accuracy for both LLMs and human participants indicates the considerable challenges posed to both human and AI independent judgment, thereby establishing the dual baseline necessary for testing human–AI synergy.

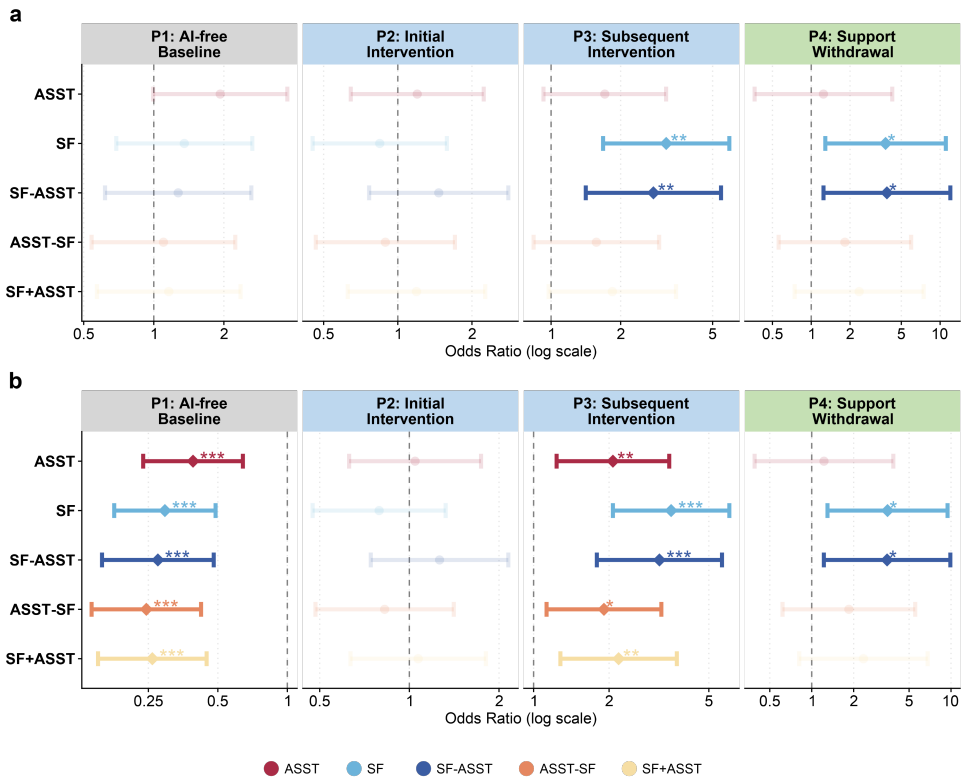


Fig. 2 Intervention group performance relative to control and LLMs benchmarks across four phases. Panel a Odds ratios (ORs) with 95% confidence intervals for each intervention group (ASST, SF, SF+ASST, SF-ASST, ASST-SF) relative to the control group (CN) across four phases (per-protocol sample, $n = 389$ participants). **Panel b** ORs with 95% confidence intervals for each intervention group relative to the LLMs group across four phases. Analyses were based on a per-protocol sample of $n = 316$ participants (ASST: $n = 69$; SF: $n = 73$; SF+ASST: $n = 62$; SF-ASST: $n = 55$; ASST-SF: $n = 63$) and the LLMs benchmark, which comprised 7 models with 10 repeated generations per case (70 observations per case). ORs indicate whether an intervention group had higher or lower odds of responding correctly relative to the control group in panel a, or relative to the LLMs group in panel b; OR > 1 indicates better performance, OR < 1 indicates worse performance, and OR $= 1$ indicates no difference. 95% CIs are unadjusted Wald intervals; p_{BH} denotes Benjamini–Hochberg corrected p-values, adjusted across the four contrasts within each phase. In both panels, solid diamonds indicate statistically significant effects ($p_{\mathrm{BH}} < .05$) and solid circles indicate non-significant effects; horizontal lines represent 95% confidence intervals; the vertical dashed grey line at OR $= 1$ denotes no difference between groups. In panel **a**, scaffolding-led groups (SF and SF-ASST) showed significantly higher ORs than CN at the Subsequent Intervention phase (OR $=$ 2.77–3.15, $p_{\mathrm{BH}} \leq .008$) and the Support Withdrawal phase (OR $=$ 3.77–3.87, $p_{\mathrm{BH}} = .048$). In panel **b**, all intervention groups significantly outperformed the LLMs at the Subsequent Intervention phase ($p_{\mathrm{BH}} \leq .016$); scaffolding-led groups (SF and SF-ASST) significantly outperformed the LLMs at the Support Withdrawal phase (OR $=$ 3.48–3.50, $p_{\mathrm{BH}} = .049$).

AI-free Baseline Performance

No baseline differences detected between intervention groups.

In Phase 1, all six human groups performed the clinical reasoning task without external support. Diagnostic accuracy did not differ significantly among any of the six human groups (intervention versus control: all pBHs $\geq .262$, ORs 1.10–1.93; full pairwise: all pBHs $\geq .680$, ORs 0.52–1.75; all 95% CIs spanned the null value of 1), see Fig.2a, P1.

LLMs outperformed human learners before interventions.

We compared the five intervention groups against the LLMs, using GLMM post-hoc z-tests with Benjamini–Hochberg correction. As shown in Fig.2b, P1, we found that all five intervention groups performed significantly lower than the LLMs (all pBHs $\leq .001$, ORs 0.24–0.39, all 95% CIs excluded the null value of 1), indicating that the LLMs achieved higher task performance than human groups in the AI-free baseline phase (Supplementary Tables 10-12 report full model comparisons, random-effects estimates, and omnibus fixed-effects tests; Supplementary Tables 13-15 report descriptive statistics and full post-hoc comparisons of task performance).

Initial and Subsequent Interventions

Intervention groups' performance converged with LLMs after initial interventions.

In Phase 2, the ASST and ASST-SF groups received AI assistant support, the SF and SF-ASST groups received scaffolding support, and the SF+ASST group received both, whereas the CN group received no support. We used GLMM post-hoc comparisons in this phase to contrast each intervention group with both the CN group and the LLMs group. No intervention group performed significantly differently from the CN group (full pairwise: all pBHs $= .726$, ORs 0.84–1.47; all 95% CIs spanned the null value of 1), see Fig.2a, P2. Likewise, no intervention group differed from the LLMs group (full pairwise: all pBHs $\geq .805$, ORs 0.79–1.26; all 95% CIs spanned the null value of 1), see Fig.2b, P2. These results suggest that, after the initial interventions, intervention groups' performance converged with both the CN group and the LLMs group at Phase 2, with no significant differences detected in either direction.

Human–AI Synergy emerged in the subsequent phase following initial scaffold-only support.

In Phase 3 (the subsequent intervention phase), the ASST group still received AI assistant support, the SF group still received scaffolding support, and the ASST-SF, SF-ASST and SF+ASST groups received both, whereas the CN group still received no support. A Type III Wald χ^2 test revealed a marginally significant Group $\times$ Phase interaction ($\chi^2(15) = 23.38$, $p = .076$; Supplementary Table 12), suggesting an emerging trend toward differing performance trajectories across groups over time. Given this trend and our a priori interest in each group's trajectory, we conducted planned pairwise comparisons within each phase. In this phase, the groups began to diverge. Relative to the CN group, both the SF and SF-ASST groups performed significantly better: the SF group (OR $= 3.15$, 95% CI $[1.68, 5.91]$, $z = 3.57$, $p_{\text{BH}} = .002$) and

the SF-ASST group (OR = 2.77, 95% CI [1.41, 5.44], $z = 2.97$, $p_{BH} = .007$), see Fig.2a, P3. The remaining three groups (ASST, SF+ASST and ASST-SF) did not differ significantly from the CN group after correction (all pBHs $> .09$). All five human groups exceeded the LLMs, see Fig.2b, P3 (all pBHs $\leq .016$, ORs 1.91–3.54; all 95% CIs excluded the null value of 1). These results indicate that, among all intervention groups, only the SF and SF-ASST groups—both of which received scaffolding first—showed the dual advantage over both CN and the LLMs that constitutes evidence of human–AI synergy.

Support Withdrawal Performance

Scaffold-first groups retained a performance advantage one week after AI withdrawal.

At Phase 4, conducted one week after the intervention phase, all support was withdrawn, returning participants to AI-free conditions. The SF and SF-ASST groups showed the strongest evidence of retained performance relative to the CN group. The odds of a correct response were higher for the SF group (OR = 3.77, 95% CI [1.29, 11.08], $z = 2.42$, $p_{BH} = .048$) and the SF-ASST group (OR = 3.87, 95% CI [1.24, 12.00], $z = 2.34$, $p_{BH} = .048$), see Fig.2a, P4. These two groups performed higher than LLMs (OR = 3.50 and 3.48, all 95% CIs excluded the null value of 1, $p_{BH} = .049$), see Fig.2b, P4. The remaining three groups (ASST, SF+ASST and ASST-SF) did not differ significantly from the CN group or the LLMs (all pBHs $\geq .19$).

To test whether providing scaffolding before direct-answer AI support produces better outcomes than providing direct-answer support from the outset, we compared two categories of intervention groups, defined by their initial mode of AI support. The *initial scaffold-only* category comprised the SF and SF-ASST groups, in which participants received AI-supported scaffolding without access to direct answers during the initial intervention phase (Phase 2). The *early direct-answer* category comprised the ASST, SF+ASST, and ASST-SF groups, in which participants had direct-answer AI assistance from the outset. The two categories were compared separately at each phase, with groups within each category weighted equally. The two categories did not differ at the AI-free baseline (Phase 1: $OR = 0.97$, 95% CI [0.62, 1.52], $p_{BH} = .900$) or during the initial intervention phase (Phase 2: $OR = 1.03$, 95% CI [0.68, 1.56], $p_{BH} = .900$). By the subsequent intervention phase (Phase 3), however, the initial scaffold-only groups achieved significantly higher diagnostic accuracy than the early direct-answer groups (75.2% vs. 66.0%, $OR = 1.73$, 95% CI [1.12, 2.68], $p_{BH} = .033$). This advantage persisted after all support was withdrawn in Phase 4 (24.6% vs. 14.4%, $OR = 2.19$, 95% CI [1.15, 4.14], $p_{BH} = .033$). These findings support the potential value of an initial period of reasoning-focused scaffolding for subsequent AI-supported performance and retained independent diagnostic performance.

Cognitive and Metacognitive Trajectories Across Experimental Phases

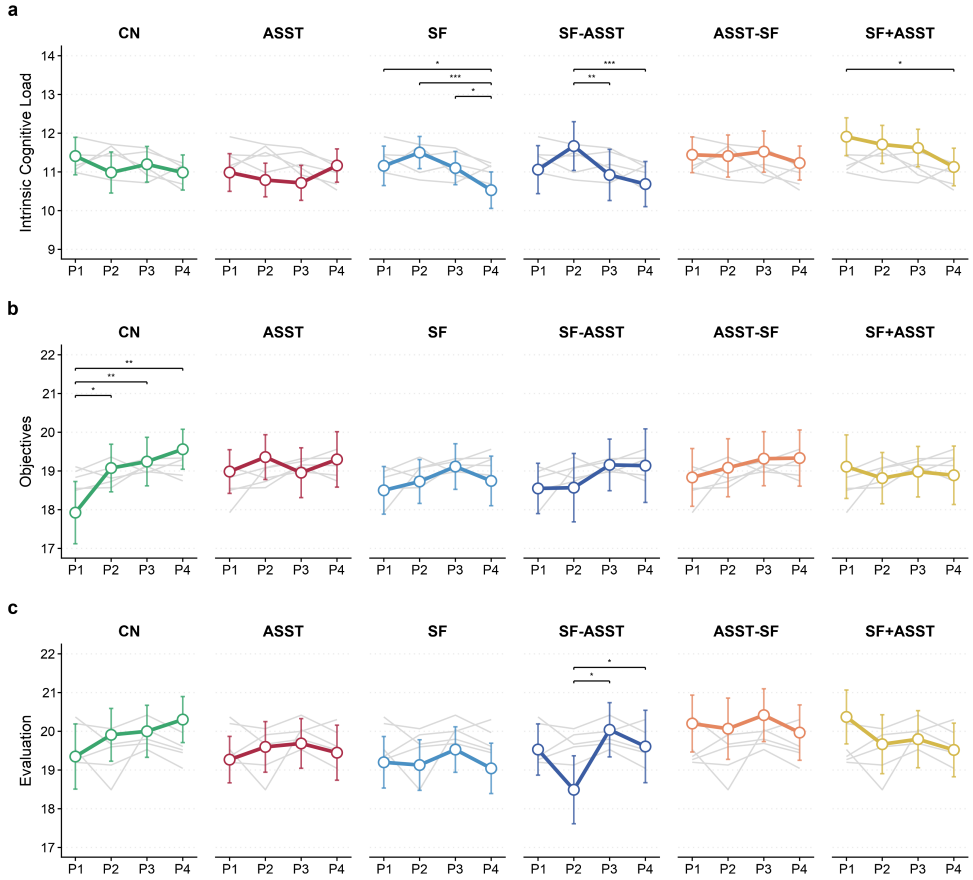


Fig. 3 Trajectories of intrinsic cognitive load, objectives and metacognitive evaluation across four experimental phases. Points indicate group means; error bars represent 95% confidence intervals of the mean. Group differences were assessed using a mixed ANOVA with Greenhouse–Geisser correction and BH-corrected pairwise comparisons. **a. intrinsic cognitive load.** $n = 67$ (ASST), 61 (ASST-SF), 55 (SF+ASST), 66 (CN), 70 (SF), 51 (SF-ASST). A significant group × phase interaction was observed ($F(14.21, 1034.48) = 1.97$, $p = .017$, $\eta_p^2 = .026$); within-group phase effects were significant for the SF, SF-ASST and SF+ASST groups (all $p \leq .040$) but not for the CN, ASST and ASST-SF groups. **b. objectives.** $n = 67$ (ASST), 60 (ASST-SF), 54 (SF+ASST), 66 (CN), 70 (SF), 51 (SF-ASST). A significant group × phase interaction was observed ($p = .042$); within-group phase effects were significant for the CN group (all $p \leq .027$) but not for the intervention groups. **c. metacognitive evaluation.** $n = 67$ (ASST), 60 (ASST-SF), 54 (SF+ASST), 66 (CN), 70 (SF), 51 (SF-ASST). A significant group × phase interaction was observed ($p = .017$); within-group phase effects were significant for the SF-ASST groups (all $p = .019$) but not for the other groups.

In order to examine how different intervention sequences shaped participants' cognitive processing and metacognitive engagement over time, we measured cognitive load along three dimensions: intrinsic cognitive load – the complexity inherent in the task, extraneous cognitive load – the additional demand imposed by the interface and design, and germane cognitive load – the processing devoted to schema construction, as well as metacognitive engagement along five dimensions: Knowledge of Cognition, Objectives, Problem Representation, Monitoring, and Evaluation. Both sets of measures were assessed across four phases (P1–P4), allowing us to trace the longitudinal trajectories of each group throughout the intervention (See Supplementary Tables 16–20 for descriptive statistics, omnibus tests, and follow-up comparisons of cognitive load; See Supplementary Tables 21–25 for descriptive statistics, omnibus tests, and follow-up comparisons of metacognitive engagement).

Control condition revealed growth in goal-directed metacognition.

Among all groups, the CN group was the only one to exhibit a significant phase effect on the Objectives dimension of metacognitive engagement ($F(2.30, 149.34) = 7.596$, $p_{\text{BH}} < .001$, $\eta_p^2 = .105$), with no significant phase effects observed in the remaining five groups (all $p_{\text{BH}} > .05$) (See Figure 3.b). Pairwise comparisons revealed that Objectives scores in Phases 2-4 were each significantly higher than in Phase 1 (Phase 2 vs. Phase 1: mean difference = 1.15, Cohen's $d_z = 0.40$, 95% CI [0.14, 0.65], $p_{\text{BH}} = .027$; Phase 3 vs. Phase 1: mean difference = 1.32, Cohen's $d_z = 0.32$, 95% CI [0.07, 0.56], $p_{\text{BH}} = .006$; Phase 4 vs. Phase 1: mean difference = 1.64, Cohen's $d_z = 0.49$, 95% CI [0.23, 0.75], $p_{\text{BH}} = .001$), with no significant differences among Phases 2-4 (all pBHs $> .05$), suggesting that the rise in Objectives scores occurred primarily at the onset of the intervention and was subsequently maintained at a stable elevated level.

Intrinsic cognitive load decreased and metacognitive evaluation increased after assistant introduction in the scaffold-first group.

Among all groups, only the SF-ASST group showed significant within-group phase effects on both intrinsic cognitive load (See Figure 3.a) and metacognitive evaluation (See Figure 3.c) during this period. Intrinsic cognitive load in the SF-ASST group declined significantly from Phase 2 to Phase 3 (Phase 2 – Phase 3 mean difference = 0.75, Cohen's $d_z = 0.46$, 95% CI [0.17, 0.75], $t(50) = 3.25$, $p_{\text{BH}} = .006$). Concurrently, metacognitive evaluation in the SF-ASST group increased significantly from Phase 2 to Phase 3 (Phase 3 – Phase 2 mean difference = 1.55, Cohen's $d_z = 0.44$, 95% CI [0.15, 0.73], $t(50) = -3.11$, $p_{\text{BH}} = .019$). No significant phase effects on intrinsic cognitive load or metacognitive evaluation were observed in the remaining groups during this period (all pBHs $> .05$).

Reduced intrinsic cognitive load and sustained metacognitive evaluation accompanied retained diagnostic performance.

Following support withdrawal in Phase 4, the SF and SF-ASST groups showed reductions in intrinsic cognitive load relative to earlier phases. In the SF group, intrinsic cognitive load in Phase 4 was significantly lower than in Phase 1 (Phase 1 – Phase 4 mean difference = 0.63, Cohen's $d_z = 0.32$, 95% CI [0.08, 0.56], $t(69) = 2.67$,

$p_{BH} = .029$), Phase 2 (Phase 2 – Phase 4 mean difference = 0.97, Cohen's $d_z = 0.48$, 95% CI [0.23, 0.73], $t(69) = 3.99$, $p_{BH} = .001$), and Phase 3 (Phase 3 – Phase 4 mean difference = 0.57, Cohen's $d_z = 0.29$, 95% CI [0.04, 0.53], $t(69) = 2.38$, $p_{BH} = .040$). In the SF-ASST group, intrinsic cognitive load in Phase 4 was significantly lower than in Phase 2 (Phase 2 – Phase 4 mean difference = 0.98, Cohen's $d_z = 0.61$, 95% CI [0.30, 0.91], $t(50) = 4.29$, $p_{BH} = .001$). Additionally, metacognitive evaluation in the SF-ASST group remained significantly higher in Phase 4 than in Phase 2 (Phase 4 – Phase 2 mean difference = 1.12, Cohen's $d_z = 0.40$, 95% CI [0.11, 0.69], $t(50) = -2.85$, $p_{BH} = .019$), and no significant result between Phase 4 and Phase 3 (Phase 3 – Phase 4 mean difference = 0.43, Cohen's $d_z = 0.13$, 95% CI [-0.15, 0.41], $p_{BH} = .440$), indicating that the elevated evaluative metacognition observed in Phase 3 was sustained after support withdrawal. No significant phase effects were observed for intrinsic cognitive load or metacognitive evaluation in the ASST or ASST-SF groups (all $p_{BH} > .05$). Although the SF+ASST group showed a significant reduction in intrinsic cognitive load from Phase 1 to Phase 4 (Phase 1 – Phase 4 mean difference = 0.78, Cohen's $d_z = 0.43$, 95% CI [0.15, 0.70], $t(54) = 3.15$, $p_{BH} = .016$), no significant changes in metacognitive evaluation were observed in this group, and it did not demonstrate retained diagnostic performance.

Discussion

This longitudinal randomised trial found that medical students who initially received AI scaffolding without direct answers achieved higher diagnostic accuracy than those with direct-answer assistance from the outset, both during subsequent AI-supported training and at an AI-free assessment one week later. In the subsequent intervention phase, all five intervention groups outperformed the LLM benchmark, but only the scaffolding-only and scaffolding-then-assistant groups also outperformed unsupported students. These two groups retained an advantage over unsupported students and LLMs after AI withdrawal. The findings support the educational value of an initial period of reasoning-focused scaffolding, while leaving longer-term retention and transfer to clinical practice to be established.

Previous studies have shown that LLM assistance can improve physicians' and medical students' diagnostic or management performance while AI is available [3, 16]. Our findings extend this evidence by showing that the developmental trajectory of human expertise depends on how AI is introduced and that human capability can develop from initially below standalone AI performance to performance that exceeds it. This longitudinal pattern directly addresses the concern that early AI use may prevent trainees from developing clinical reasoning of their own [19]. The progression from initially underperforming frontier LLMs to subsequently exceeding them also demonstrates that developing human expertise can remain valuable as AI capabilities improve. Human–AI synergy therefore depends on developing and retaining human clinical expertise that supports independent reasoning, evaluation of AI outputs, and effective collaboration with AI [9, 12].

The performance trajectories also reveal different relationships between AI-supported performance and the development of independent clinical reasoning.

Unsupported exposure to difficult cases did not produce consistently strong performance in the control condition. AI scaffolding preserved learners' responsibility for generating and evaluating their own reasoning while providing guidance during challenging tasks, consistent with established accounts of supported learning within learners' developing capabilities [21, 30, 31]. By comparison, students given the AI assistant could improve performance, while direct answers were available without developing an equivalent advantage after the assistant was removed. We refer to this pattern as *contingent-skilling*, in which competence can be expressed effectively with AI support but is not converted into equivalent support-independent performance. Contingent-skilling differs from de-skilling, which concerns deterioration of previously acquired competence [18], and from never-skilling, which concerns failure to develop foundational competence in the first place [19]. The current study was not designed to establish mis-skilling because participants were not systematically exposed to misleading AI, although experimental evidence shows that erroneous AI can impair diagnostic reasoning in novice learners [7, 17]. Contingent-skilling may nevertheless be clinically consequential because dependence on continued AI access leaves performance vulnerable when systems are unavailable, unreliable, or unevenly deployed across healthcare settings [32, 33]. It may also weaken meaningful oversight if clinicians lack sufficient independent capacity to question, override, or interrupt AI-mediated decisions [12].

The strongest causal evidence from the trial concerns the sequence in which different forms of AI were introduced. Students who began with scaffolding alone subsequently achieved higher diagnostic accuracy than students who had direct-answer AI available from the outset, both during later training and after all AI was withdrawn. This result provides experimental support for the metacognitive approach to learning and performance in human–AI interaction proposed by Fan et al. [21]. That approach predicts that learners should first develop and internalise cognitive and metacognitive capabilities before selectively delegating task demands to AI. Previous research has established several components relevant to this prediction. Problem solving before instruction can promote deeper and more transferable learning [31]; unrestricted generative AI can improve supported performance while impairing subsequent unaided performance, whereas guarded designs can mitigate this effect [20]; requiring learners to derive answers from AI-provided suggestions can improve learning [34]; and cognitive-forcing interventions can reduce over-reliance on AI [35]. Recent evidence also indicates that hint-first AI can preserve productive engagement more effectively than unrestricted answer-giving AI [36]. The present study advances this literature by testing the sequencing hypothesis longitudinally in clinical reasoning, comparing different orders of scaffolding and direct-answer assistance, assessing performance after withdrawal, and benchmarking students against frontier LLMs. The SF+ASST condition is also informative because simultaneous access to AI scaffolding and direct answers did not produce the later advantage observed when scaffolding was initially protected from direct-answer access. This pattern supports the interpretation that an immediately available route to answers can reduce the developmental benefit of scaffolding even when reflective support is present.

Changes in cognitive load and metacognitive evaluation provide converging evidence for the proposal that learners can selectively offload some task demands to AI

while retaining the metacognitive monitoring and evaluation needed to regulate their own reasoning [21]. When students in the SF-ASST condition moved from scaffolding alone to scaffolding with the AI assistant, intrinsic cognitive load declined while metacognitive evaluation increased. Elevated metacognitive evaluation was subsequently maintained after AI withdrawal, alongside reduced cognitive load and stronger independent performance. Lower cognitive load alone did not show the same relationship with learning outcomes. The SF+ASST group also showed reduced intrinsic cognitive load by the withdrawal phase without a corresponding increase in metacognitive evaluation or retained advantage in diagnostic accuracy. This pattern suggests that effective use of AI involves more than reducing cognitive demands. Learners may benefit when some task demands are offloaded while monitoring and evaluation of their own reasoning remain active [21, 37]. Experimental evidence that metacognitive calibration improves subsequent offloading decisions provides independent support for this interpretation [38]. Our findings therefore provide empirical support for the predicted combination of reduced cognitive demands, stronger evaluative metacognition, AI-supported performance, and retained independent performance [21]. These self-reported measures do not directly capture moment-to-moment cognitive and metacognitive offloading, so the specific process through which these changes occur remains to be tested with behavioural and process-level measures. Preserved evaluative metacognition may nevertheless be particularly important in medicine because meaningful oversight requires clinicians who can evaluate AI output and intervene when its recommendations are uncertain or incorrect [7, 12]. Human-first learning may consequently provide a developmental route toward human–AI synergy by building the capacities needed to use AI selectively rather than treating synergy as an immediate consequence of combining human and AI performance [39].

The retained advantage of students who began with AI scaffolding after direct AI assistance was withdrawn has immediate implications for medical education. Introducing an AI assistant only after learners have had opportunities to first develop and exercise clinical reasoning may better prepare them to use AI without making their performance dependent on continued access to direct answers. The findings suggest that even a brief initial period of reasoning-focused scaffolding may produce benefits that persist after AI support is withdrawn. By prompting students to draw on existing knowledge and evaluate their own reasoning before accessing direct answers, this approach may encourage active cognitive and metacognitive engagement. Medical education should therefore consider the sequence in which different forms of AI are introduced, alongside the capabilities of the systems themselves [40, 41]. Training on challenging cases may remain especially important because such cases require learners to gather incomplete evidence, manage uncertainty, evaluate competing explanations, and decide when AI output should be trusted or challenged. These capacities are closely related to the contextual judgment, accountability, and integration of patient information that remain central to clinical practice [42–45]. The findings also have implications for assessment. Evaluating performance only while AI is available can obscure whether learners have developed clinical reasoning that remains available independently. Assessment in AI-supported medical education should therefore examine learning trajectories, reasoning processes, appropriate reliance, and performance after

AI withdrawal in addition to immediate diagnostic outcomes [46]. Such approaches may help medical education prepare clinicians who can benefit from increasingly capable AI while developing the capacity to evaluate and challenge its outputs [47]. Whether the same sequencing principle applies to practising clinicians and continuing professional development requires direct investigation.

Several limitations define the scope of these conclusions. Cognitive load and metacognitive engagement were assessed using self-reported measures at the end of each experimental phase. Although the instruments have been validated previously [23, 48], self-report cannot capture the dynamic cognitive and metacognitive processes occurring during individual human–AI interactions. The mechanistic interpretation should therefore be tested using dialogue data, behavioural traces, process measures, and other multimodal indicators. Independent performance was assessed one week after the interventions were withdrawn. The withdrawal phase demonstrated retention beyond immediate AI-supported performance but did not establish durability over longer periods or transfer across different clinical tasks and settings. Case difficulty was screened before phase assignment, yet residual variation across cases may have contributed to fluctuations in absolute accuracy across phases. More stringent calibration or counterbalanced assignment would strengthen future studies. The trial randomised 496 students and included 389 in the final analysis after post-randomisation data-quality exclusions. The resulting estimates therefore apply to the retained analytic sample, and the potential influence of post-randomisation exclusion should be examined through appropriate sensitivity analyses. The study was also conducted with medical students from a single institution in a simulated clinical environment. Generalisation to other medical schools, practising clinicians, real patients, and high-stakes clinical workflows requires further study. Learner characteristics including prior knowledge, AI literacy, and epistemic beliefs may also moderate the effects of different AI sequences and warrant dedicated analysis.

Methods

Ethics, preregistration, and reporting

The study was approved by the Anonymised University Human Research Ethics Committee (Anonymised) and preregistered on the Open Science Framework (OSF) prior to data collection. The study was embedded in the *Clinical Thinking Training* course, and research activities were integrated with regular classroom teaching according to students' learning progress. All participants provided informed consent electronically before beginning the study. Participants logged in using assigned platform accounts, and no directly identifying information, such as names or student IDs, was collected. Research data were anonymised before analysis, and participants could withdraw from the study at any time and request deletion of their data.

Participants, recruitment, and randomization

Participants were recruited from Anonymised University Medical College between October and December 2025. A total of 496 clinical medicine students took part in

the experiment (202 female, 294 male). Participants were assigned to six groups using stratified random allocation generated by R software, stratified by sex and class. To ensure adequate engagement with the experimental tasks and to minimise the potential confounding influence of external support, valid participants were identified based on three criteria: the extent of their interactions with three virtual characters on the platform, whether they had actually utilised the assigned intervention, and whether they had engaged in cheating behaviours by using AI tools outside the platform. Across the three criteria, 57 participants were excluded for insufficient interaction rounds (CN: n = 8, ASST: n = 10, SF: n = 9, SF+ASST: n = 12, SF-ASST: n = 10, ASST-SF: n = 8), and an additional 50 for suspected use of external AI tools, identified via content feature detection (e.g., AI-typical formatting) and behavioural trace analysis (e.g., paste actions without prior on-platform copying) (CN: n = 7, ASST: n = 4, SF: n = 1, SF+ASST: n = 8 SF-ASST: n = 18, ASST-SF: n = 12), resulting in a total of 107 exclusions. The final analytic sample comprised 389 valid participants, distributed across groups as follows: CN group (n = 67), ASST group (n = 69), SF group (n = 73), SF+ASST group (n = 62), SF-ASST group (n = 55), and ASST-SF group (n = 63). The final sample size met the statistical power requirements calculated in the pre-registered study protocol (see Supplementary Information Sections 1.4 and 1.5 for exclusion criteria and detailed demographics).

Experimental procedure

Three types of data were collected in this study, namely task performance, cognitive load, and metacognitive engagement. The task performance data are based on the students' case reports for each clinical reasoning task. Cognitive load and metacognitive engagement were collected through scales.

Clinical reasoning task

Participants completed clinical reasoning tasks on a simulated learning platform that replicated the workflow of real-world clinical practice, including patient consultations and retrieval of examination results (see Supplementary Information Section 3 for screenshots of the platform). To gather relevant information, participants engaged in dialogue with virtual characters on the platform and iteratively refined their diagnoses based on the information obtained (Fig 4). Specifically, the platform featured three AI-powered virtual characters: a virtual standard patient, a physical examination reporter, and a diagnostic tests reporter (Author, 2025). The system prompts for each character were adapted from the selected clinical cases. The virtual patient was instructed to describe symptoms in a naturalistic manner in response to participants' questions, while the reporters were instructed to provide corresponding findings upon request. Upon completing the interaction, participants were required to compose and submit a case report, including their inferred diagnosis and the supporting evidence, which served as the measure of task performance.

Clinical cases were sourced from the Lecturio case library and initially screened for difficulty. An expert panel then reviewed all candidate cases to confirm that the difficulty exceeded the knowledge and skill level of the participant population. Case

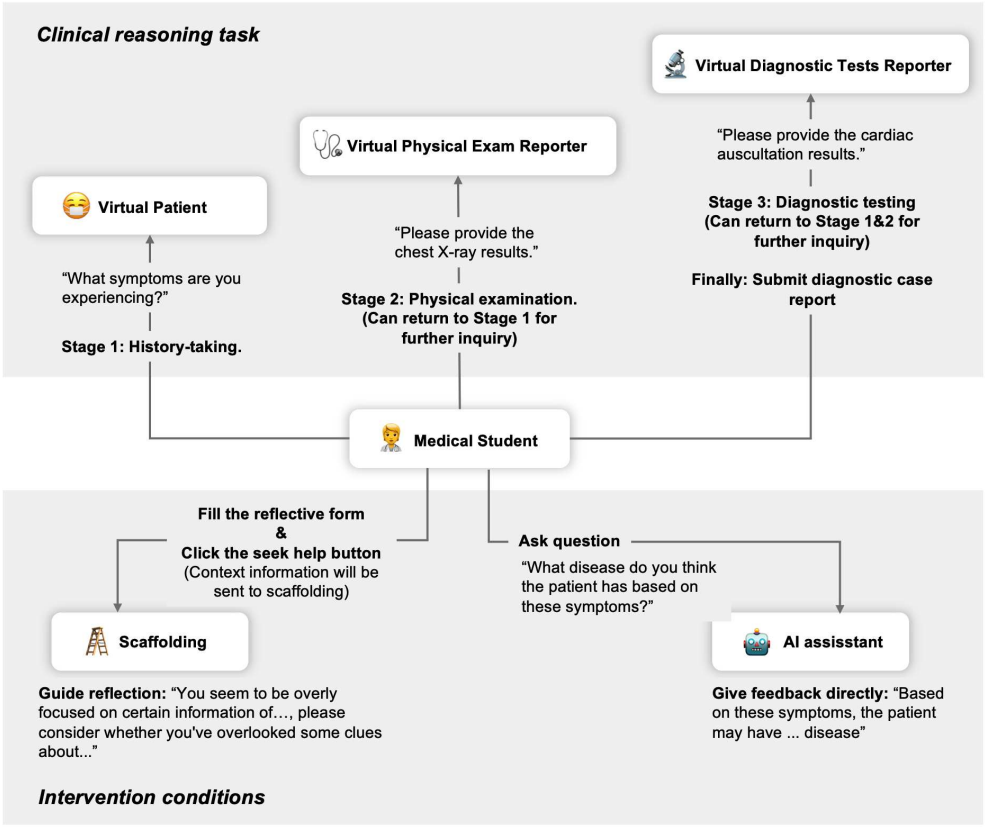


Fig. 4 Interaction between participants and AI-based roles. The clinical reasoning task is the core component of the experiment, in which all participants must interact with a Virtual Patient, a Virtual Physical Exam Reporter, and a Virtual Diagnostic Tests Reporter to gather information and formulate a diagnosis. Different participants engage in conversations with either the Scaffolding or the AI assistant, depending on the intervention assigned to them at each experimental phase.

difficulty was further verified through a preliminary pilot evaluation conducted independently of the main experiment, in which 9 LLMs and 14 medical students each made a diagnosis on the candidate cases; only cases in which both groups showed consistently low diagnostic accuracy were retained for the main study (see Supplementary Information Section 1.1 for details of the pilot evaluation).

To benchmark the diagnostic performance of state-of-the-art LLMs, a distinct and expanded set of frontier LLMs was selected for the formal analysis, including GPT-5, GPT-o3, Claude Opus 4.1, Gemini 2.5 Pro, Grok-4, DeepSeek-R1, and Qwen-3-Max. Each model autonomously interacted with the same three virtual characters following identical procedures as human participants. Each model completed this process 10 times per case to ensure stability of the performance estimate (see Supplementary Information Section 1.1.3 for the full LLM baseline testing protocol). The resulting diagnostic accuracy served as the LLM performance benchmark.

Intervention conditions

Two forms of intervention were implemented: AI assistant (ASST) and scaffolding (SF). Participants who had access to an AI assistant could ask questions during the clinical reasoning task, and the AI assistant provided immediate answers. This condition simulated common uses of AI by offering direct informational support, such as medical knowledge, conceptual clarification, or diagnostic information. Participants who had access to SF needed to monitor and evaluate their own reasoning processes (Author, 2026). Specifically, the SF consists of two components: (1) a Deliberate Reflection Table, which requires participants to record diagnostic reasoning information; and (2) an AI agent called "AI Tutor", which poses reflective questions and feedback based on the content of the table as well as the transcripts of participants' conversations with three virtual characters within the platform. Learners with access to the SF can click the "Ask the Tutor" button in the table to interact with the AI Tutor (Fig 4).

The study used a four-phase design: an AI-free baseline phase, an initial intervention phase, a subsequent intervention phase and a support withdrawal phase. In the AI-free baseline phase (Phase 1), all students completed two tasks on their own, without scaffolding or AI support. In the initial intervention phase (Phase 2), students completed two tasks under their group's assigned support condition. The CN group worked independently; the SF and SF-ASST groups received scaffolding; the AI and ASST-SF groups received AI support; and the SF+ASST group received both scaffolding and AI support. In the subsequent intervention phase (Phase 3), students completed two tasks under their group-specific support condition. The CN group still worked independently; the SF group received scaffolding; the ASST group received AI support; and the SF+ASST, SF-ASST, and ASST-SF groups received both scaffolding and AI support. In the support withdrawal phase (Phase 4), all students again completed a challenging task independently, without AI support or scaffolding, to see whether they could maintain what they had learned after the support was removed.

Outcome measures

Diagnostic accuracy was assessed by comparing participants' most likely diagnosis against a pre-specified expert-derived reference diagnosis. Responses were independently coded as correct (score = 1) when the primary diagnosis matched the expert standard, and incorrect or partially correct (score = 0) when it did not fully correspond to the reference diagnosis.

Cognitive load was measured using the scale developed and validated by Klepsch et al.[48], consisting of 7 items covering intrinsic cognitive load (ICL), extraneous cognitive load (ECL) and germane cognitive load (GCL), rated on a 7-point Likert scale. The metacognitive engagement scale was selected based on domain-specificity considerations. General metacognitive measures are designed for domain-general learning contexts and are considered less appropriate for medical diagnostic problem-solving tasks. Therefore, a task-specific metacognitive scale developed by Wang et al. [23] for medical diagnostic settings was adopted in this study. This scale used a 5-point Likert-type scale and included 21 items across five dimensions: Knowledge of

Cognition, Objectives, Problem Representation, Monitoring and Evaluation. Upon completing the clinical reasoning tasks in each phase, participants were immediately asked to complete both scales online. For the cognitive load scale, reliability analysis showed satisfactory internal consistency across the three dimensions (Cronbach's α = 0.75–0.83). Confirmatory factor analysis supported an acceptable three-factor structure (CFI = 0.893, RMSEA = 0.171, SRMR = 0.089), with standardised factor loadings ranging from 0.62 to 0.86 ($p < .001$). Although CFI (0.893) and SRMR (0.089) approached the margin of acceptable fit, the elevated RMSEA value (0.171) may be partly attributable to the limited degrees of freedom associated with this 7-item scale, as RMSEA is known to be inflated in models with small df even when correctly specified [49]. This is further supported by a large-scale meta-analysis showing a significant positive correlation between the ICL and ECL dimensions ($r = 0.403$), suggesting a discriminant validity limitation that may contribute to poorer model fit [50]. Given that Cronbach's α for all three dimensions in the current study exceeded acceptable thresholds, the elevated RMSEA likely reflects these inherent scale characteristics rather than a systematic measurement problem. For the metacognitive engagement scale, Cronbach's α values ranged from 0.65 to 0.83 across dimensions, indicating good internal consistency. The five-factor model demonstrated good fit (CFI = 0.927, TLI = 0.915, RMSEA = 0.054, SRMR = 0.042), and standardised factor loadings ranged from 0.54 to 0.80 ($p < .001$), supporting construct validity.

Statistical analysis

To examine how the different intervention conditions affected participants' complex problem-solving performance and the development of their competence, we modelled diagnostic accuracy (correct = 1, incorrect = 0) using generalised linear mixed models (GLMMs) with a binomial distribution and a logit link function, fitted with the *lme4* package in R.

For the human–human comparisons, the model included data from all participants, with group (CN, ASST, SF, SF+ASST, SF-ASST and ASST-SF), phase (AI-free baseline, initial intervention, subsequent intervention, support withdrawal) and their interaction as fixed effects, and by-participant and by-case random intercepts. For the human–AI comparisons, the model included data from the five intervention groups and the LLMs group, with group, phase and their interaction as fixed effects and a by-case random intercept. Random-effects structures were selected on the basis of likelihood-ratio tests and AIC comparisons between candidate models, and singular-fit diagnostics were used to confirm model stability. The significance of effects was assessed using Type III Wald χ^2 tests. Post-hoc comparisons were based on estimated marginal means, for which we report odds ratios (ORs), unadjusted 95% Wald confidence intervals, and P values adjusted for multiple comparisons using the Benjamini–Hochberg (BH) procedure.

To directly test whether the primacy of scaffolding, rather than mere exposure to AI assistance, drives clinical reasoning performance, we defined a theory-driven contrast from the human–human GLMM. Specifically, we compared the average performance of the *initial scaffold-only* conditions (SF and SF-ASST, weighted equally) against the average performance of the *early direct-answer* conditions (ASST,

SF+ASST, and ASST-SF, weighted equally), such that each category contributed equally to the comparison regardless of group sample sizes. This contrast was evaluated at each phase separately, with BH correction applied across the four phases.

To examine the cognitive mechanisms through which the intervention conditions shaped human-AI synergy, we analysed each dimension of cognitive load (intrinsic, germane and extraneous) and each dimension of metacognitive engagement (Knowledge of Cognition, Objectives, Problem Representation, Monitoring, and Evaluation) using mixed-design ANOVAs. Each model included group as a between-subjects factor and phase as a within-subjects factor, and we focused on the main effect of group, the main effect of phase and their interaction. Sphericity was assessed using Mauchly's test, and the Greenhouse–Geisser correction was applied where the assumption was violated. Significant interactions were followed up with simple-effects analyses and post-hoc pairwise comparisons. Effect sizes are reported as partial eta-squared (η_p^2) and Cohen's d, and multiple comparisons were corrected using the BH procedure. All analyses were conducted in R (version 4.5.2), with the significance threshold set at $\alpha = .05$.

Data availability

De-identified data supporting this study will be made publicly available after the acceptance of this paper.

References

[1] Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642:442–450. https://doi.org/10.1038/s41586-025-08866-7.

[2] Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine. 2024;30:2613–2622. https://doi.org/10.1038/s41591-024-03097-1.

[3] Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Network Open. 2024;7(10):e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.

[4] Teo ZL, Thirunavukarasu AJ, Elangovan K, Cheng H, Moova P, Soetikno B, et al. Generative artificial intelligence in medicine. Nature medicine. 2025;31(10):3270–3282.

[5] Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA. 2025;333(4):319–328. https://doi.org/10.1001/jama.2024.21700.

[6] Chen SF, Alyakin A, Seas A, Yang E, Choi JJ, Lee JV, et al. LLM-assisted systematic review of large language models in clinical medicine. Nature Medicine. 2026;32:1152–1159. https://doi.org/10.1038/s41591-026-04229-5.

[7] Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. JAMA. 2023;330(23):2275–2284. https://doi.org/10.1001/jama.2023.22295.

[8] Agarwal N, Moehring A, Rajpurkar P, Salz T, Ouyang D, Shih G, et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nature Medicine. 2024;30:837–849. https://doi.org/10.1038/s41591-024-02850-w.

[9] Vaccaro M, Almaatouq A, Malone TW. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour. 2024;8(12):2293–2303. https://doi.org/10.1038/s41562-024-02024-1.

[10] Mello MM, Cohen IG. Regulation of Health and Health Care Artificial Intelligence. JAMA. 2025;333(20):1769–1770. https://doi.org/10.1001/jama.2025.3308.

[11] Abdulnour REE, Gin B, Boscardin CK. Educational Strategies for Clinical Supervision of Artificial Intelligence Use. New England Journal of Medicine. 2025;393(8):786–797. https://doi.org/10.1056/NEJMra2503232.

[12] van de Sande D, Economou-Zavlanos N, van Genderen ME. Meaningful Oversight of Medical AI beyond Human in the Loop. npj Digital Medicine. 2026;9:569. https://doi.org/10.1038/s41746-026-02971-1.

[13] Zöller N, Berger J, Lin I, Fu N, Komarneni J, Barabucci G, et al. Human–AI collectives most accurately diagnose clinical vignettes. Proceedings of the National Academy of Sciences of the United States of America. 2025;122(24):e2426153122. https://doi.org/10.1073/pnas.2426153122.

[14] Yan L, Greiff S, Teuber Z, Gašević D. Promises and challenges of generative artificial intelligence for human learning. Nature Human Behaviour. 2024;8:1839–1850. https://doi.org/10.1038/s41562-024-02004-5.

[15] Yan L, Greiff S, Lodge JM, Gaševi'c D. Distinguishing performance gains from learning when using generative AI. Nature Reviews Psychology. 2025;4(7):435–436. https://doi.org/10.1038/s44159-025-00467-5.

[16] Roemer A, Schlicker N, Kernder A, Albe B, Hack J, Hirsch M, et al. Large language models enhance diagnostic reasoning of medical students in rheumatology: a randomized controlled trial. BMC Medical Education. 2026;26(1):579.

[17] Teng D, Tan L, Cao Q, Xia Y, Zhang N, Li J, et al. Impact of AI Misinformation on Diagnostic Accuracy and Confidence Calibration in Novice Medical Students. npj Digital Medicine. 2026;9:356. https://doi.org/10.1038/s41746-026-02547-z.

[18] Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. The Lancet Gastroenterology & Hepatology. 2025;10(10):896–903. https://doi.org/10.1016/S2468-1253(25)00133-5.

[19] Ke Y, Jin L, Ong JCL, Thirunavukarasu AJ, Car J, Cheung CY, et al. AI-induced never-skilling in medical education. Nature Medicine. 2026;32:1997–2006. https://doi.org/10.1038/s41591-026-04438-y.

[20] Bastani H, Bastani O, Sungu A, Ge H, Kabakcı O, Mariman R. Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences. 2025;122(26):e2422633122. https://doi.org/10.1073/pnas.2422633122.

[21] Fan Y, Yan L, Molenaar I, Greiff S, Gašević D.: A Metacognitive Approach to Learning and Performance in Human-AI Interaction. PsyArXiv preprint. Available from: https://osf.io/preprints/psyarxiv/s7dvk_v1.

[22] Bowen JL. Educational strategies to promote clinical diagnostic reasoning. New England Journal of Medicine. 2006;355(21):2217–2225. https://doi.org/10.1056/NEJMra054782.

[23] Wang CY, Chen S, Huang MY. Exploring medical students' metacognitive and regulatory dimensions of diagnostic problem solving. Medical Education Online. 2023;28(1):2210804.

[24] 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. British Journal of Educational Technology. 2025;56(2):489–530.

[25] Lim L, Bannert M, van der Graaf J, Singh S, Fan Y, Surendrannair S, et al. Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. Computers in Human Behavior. 2023;139:107547. https://doi.org/10.1016/j.chb.2022.107547.

[26] Ong AY, Sui M, Rosen KL, Kvedar JC. Reconciling how clinical reasoning is learned in the age of artificial intelligence. npj Digital Medicine. 2026;9:435. https://doi.org/10.1038/s41746-026-02873-2.

[27] Ong AY, Merle DA, Pollreisz A, Wagner SK, Sevgi M, Keane PA, et al. Flight rules for clinical AI: lessons from aviation for human-AI collaboration in medicine. npj Digital Medicine. 2026;9:201. https://doi.org/10.1038/s41746-026-02410-1.

[28] D"orner D, Funke J. Complex problem solving: What it is and what it is not. Frontiers in Psychology. 2017;8:1153. https://doi.org/10.3389/fpsyg.2017.01153.

[29] Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Academic Medicine. 2020;95(8):1166–1171.

[30] Vygotsky LS. Mind in Society: The Development of Higher Psychological Processes. Cole M, John-Steiner V, Scribner S, Souberman E, editors. Cambridge, MA: Harvard University Press; 1978.

[31] Sinha T, Kapur M. When Problem Solving Followed by Instruction Works: Evidence for Productive Failure. Review of Educational Research. 2021;91(5):761–798. https://doi.org/10.3102/00346543211019105.

[32] Hwang YM, Ng MY, Pillai M, Sahai MP, Hernandez-Boussard T. The Landscape of AI Implementation in US Hospitals. Nature Health. 2026;1:99–112. https://doi.org/10.1038/s44360-025-00016-7.

[33] Car J, Wong TY, Atun R. The Recursive Care Law: Artificial Intelligence Reinforcing Feedback Loops and Health Inequity. The Lancet. 2026;407(10546):2354–2356. https://doi.org/10.1016/S0140-6736(26)00982-7.

[34] Gajos KZ, Mamykina L. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In: Proceedings of the 27th International Conference on Intelligent User Interfaces. Association for Computing Machinery; 2022. p. 794–806.

[35] Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW1):1–21. https://doi.org/10.1145/3449287.

[36] Bassner P, Lenk-Ostendorf B, Beinstingel R, Wasner T, Krusche S. Less Stress, Better Scores, Same Learning: The Dissociation of Performance and Learning in AI-Supported Programming Education. Computers and Education: Artificial Intelligence. 2026;10:100537. https://doi.org/10.1016/j.caeai.2025.100537.

[37] Cash TN, Kelly MO, Macnamara BN, Risko EF. Is AI Making Us Stupid? Trends in Cognitive Sciences. 2026;Online ahead of print. https://doi.org/10.1016/j.tics.2026.06.004.

[38] Ngai C, Gilbert SJ. Metacognitive Training Facilitates Optimal Cognitive Offloading. Cognitive Research: Principles and Implications. 2026;11:21. https://doi.org/10.1186/s41235-026-00714-0.

[39] Vaccaro M, Almaatouq A, Malone T. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. Nature Human Behaviour.

2024;8:2293–2303. https://doi.org/10.1038/s41562-024-02024-1.

[40] Sokol K, Fackler J, Vogt JE. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. npj Digital Medicine. 2025;8(1):345.

[41] Ke Y, Jin L, Ong JCL, Thirunavukarasu AJ, Car J, Cheung CY, et al. AI-induced never-skilling in medical education. Nature medicine. 2026;p. 1–10.

[42] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019;25(1):44–56.

[43] Birch J, Creel KA, Jha AK, Plutynski A. Clinical decisions using AI must consider patient values. Nature medicine. 2022;28(2):229–232.

[44] Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. The Lancet Digital Health. 2020;2(3):e111–e112.

[45] Ning Y, Ong JCL, Cheng H, Wang H, Ting DSW, Tham YC, et al. How can artificial intelligence transform the training of medical students and physicians? The Lancet Digital Health. 2025;7(10).

[46] Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Medical Education. 2025;25(1):129. https://doi.org/10.1186/s12909-025-06719-5.

[47] Collaco BG, Haider SA, Prabha S, Gomez-Cabello CA, Genovese A, Wood NG, et al. The role of agentic artificial intelligence in healthcare: a scoping review. npj Digital Medicine. 2026;9(1):345.

[48] Klepsch M, Schmitz F, Seufert T. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Frontiers in psychology. 2017;8:1997.

[49] Kenny DA, Kaniskan B, McCoach DB. The Performance of RMSEA in Models With Small Degrees of Freedom. Sociological Methods & Research. 2015;44(3):486–507. https://doi.org/10.1177/0049124114543236. https://doi.org/10.1177/0049124114543236.

[50] Krieglstein F, Beege M, Rey GD, Ginns P, Krell M, Schneider S. A Systematic Meta-analysis of the Reliability and Validity of Subjective Cognitive Load Questionnaires in Experimental Multimedia Learning Research. Educational Psychology Review. 2022 Dec;34(4):2485–2541. https://doi.org/10.1007/s10648-022-09683-4.

Supplementary Files

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

- NMClinicalReasoningRCTStudy2026Appendix.pdf