npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

Article in Press

https://doi.org/10.1038/s41746-026-03372-0

From simulation to pedagogy: multi-model and randomized evaluation of structured AI standardized patients

Check for updates

Ping Wu[1], Yu Han[2], Jing Zhang[1], Yunqi Li[1], Mengna Jiang[1], Xinyu Lu[1], Haibin Zhang[1,3], Danyang Xu[1,4], Hao Ming[1,5], Lihong Wang[1,6] & Qingping Wen[1]

Standardized patients (SPs) are central to clinical communication training but remain constrained by cost, scalability, and reliance on trained actors. We present AI standardized patients (AI-SPs), large language model–driven patient simulators built on a three-layer information architecture that adaptively modulates information disclosure according to learner skill. In Phase 1, blinded expert assessment of 350 simulated consultations generated by five frontier LLMs showed that, under a common architecture, learner skill level accounted for more of the observed performance variation than model choice ($\eta^2 = 0.31$ vs. 0.06); whether this reflects the architecture itself cannot be determined without comparing models with and without it. In Phase 1b, analysis of 155 live student consultations identified systematic failures to elicit safety-critical information, enabling automated curriculum diagnostics without expert observation. In a Phase 2 three-arm pilot randomized study ($n = 58$), OSCE performance gains with AI-SP training did not differ significantly from those with human SP training; the trial was designed to estimate effect sizes for a future confirmatory study, not to demonstrate equivalence. Self-efficacy gains were numerically largest with AI-SP training, but this exploratory difference did not survive correction for multiple comparisons. These findings suggest that architecture-driven AI-SPs could provide a scalable, model-portable approach to clinical communication training.

Clinical communication skills are fundamental to patient safety, yet their development requires deliberate practice with realistic patient encounters[1,2]. Standardized patients (SPs) remain the gold standard[3,4], but their dependence on trained actors, scheduling coordination, and faculty supervision creates resource barriers that constrain training frequency, particularly in specialties such as anaesthesiology, where preoperative consultations demand integration of history-taking, risk communication, and emotional support within high-stakes, time-pressured encounters[5,6].

Large language models (LLMs) offer a potential solution. Current models demonstrate clinical knowledge[7,8], conversational fluency[9], and the capacity to adopt complex personas[10], capabilities that surpass prior virtual patient systems built on branching scripts or constrained natural language processing[11,12]. Early work has shown promise for generative AI in personalised medical education[13] and clinical communication[14,15]. Recent studies have advanced AI-based patient simulation: Luo et al.[16] demonstrated LLM-based history-taking training in ophthalmology; Lavigne et al.[17] reported a two-arm RCT showing AI-simulated clinical encounters improved OSCE performance over standard training; and Yu et al.[18] developed an LLM-based simulated patient agent validated with medical students. However, these studies share important limitations: all used a single LLM platform (raising generalisability concerns), none compared AI-SP directly against human SP practice (the clinical gold standard), and none employed a structured pedagogical framework governing information disclosure. Creating educationally effective AI-SPs requires more than conversational fluency; it demands pedagogically controlled information disclosure that rewards skilled interviewing, emotional responses calibrated to communication quality, and clinical consistency across conversational turns[19].

[1]Department of Anesthesiology, The First Affiliated Hospital of Dalian Medical University, Dalian, 116000, China. [2]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxfordshire, UK. [3]Department of Anesthesiology, First Hospital of Shanxi Medical University, Taiyuan, 030001, China. [4]Department of Anesthesiology, Affiliated Hospital of Chifeng University, Chifeng, 024000, China. [5]Department of Anesthesiology, Shandong Second Provincial General Hospital, Jinan, 250000, China. [6]Department of Anesthesiology, Chifeng Obstetrics and Gynecology Hospital, Chifeng, 024000, China.
e-mail: wenqp@dmu.edu.cn

Here, we introduce a *three-layer information architecture* for AI-SP design that addresses this gap. Surface information (Layer 1) is volunteered spontaneously; prompted information (Layer 2) is disclosed only upon direct questioning; and hidden information (Layer 3) is withheld, not through deliberate concealment but because the simulated patient lacks the health literacy to recognise its clinical relevance, requiring targeted, empathetic questioning to elicit (Fig. 1a). This design mirrors documented patient behaviour in which critical clinical details emerge only through skilled communication[20], and operationalises cognitive apprenticeship theory[21] by generating a calibrated gradient between what novice and competent students can discover. Crucially, the hidden-information discovery rate provides an objective, automated metric of clinical communication skill without requiring real-time expert observation, enabling scalable assessment alongside scalable practice. The iceberg metaphor (Supplementary Fig. 1) further visualises this graduated disclosure: Layer 1 sits above the waterline accessible to all students, while Layer 3 resides in deep water, discoverable only through empathetic probing that builds sufficient patient trust.

We evaluate this approach through a three-phase progressive validation pipeline targeting preoperative anaesthesia consultations (Fig. 1d–j). This pipeline follows a deliberate escalation logic: deploying an AI training system in a clinical education setting without first establishing its fidelity and ecological validity would be methodologically premature and ethically questionable. Each phase, therefore, addresses a prerequisite question that must be answered before the next phase can be justified. Phase 1 addresses construct validity: can AI-SPs produce educationally adequate simulations? Seven anaesthesiology educators blindly evaluated dialogues from five frontier LLMs (GPT-4o, Claude 4.5 Sonnet, Gemini 2.5 Flash, Qwen-2.5 Plus, DeepSeek-R1) across five clinical scenarios at two student skill levels (Competent, Novice), yielding 350 assessments. Phase 1b addresses ecological validity: do scripted-dialogue findings generalise to authentic student interactions? Thirty-one students completed 155 real-time consultations to characterise interaction quality and per-item discovery patterns under uncontrolled conditions. Phase 2 addresses training efficacy: does AI-SP practice improve clinical skills comparably to human SP practice? A three-arm pilot RCT (n = 58) compared AI-SP, human SP, and standard curriculum controls on OSCE performance and self-efficacy.

Translating the three-layer architecture into a system that functions reliably across multiple LLMs required a structured prompt engineering process (Fig. 2). Over four iterative development rounds with expert review, a clinician–engineer team identified and resolved recurring failure modes, including role boundary violations, premature information disclosure, affective flattening, and stylistic artefacts, producing a final six-section prompt structure that is identical across all five LLMs tested.

Results

Results are presented following the progressive validation sequence: first establishing that AI-SPs can produce educationally adequate simulations (Phase 1), then confirming these findings generalise to authentic student interactions (Phase 1b), and finally testing whether AI-SP practice translates to measurable training outcomes (Phase 2). Because this journal presents Methods after Results, we summarise the design briefly here for orientation: Phase 1 was a blinded expert evaluation of scripted dialogues from five hosted LLMs (seven evaluators; 350 assessments); Phase 1b was a live single-model study of 31 students (155 consultations); and Phase 2 was a three-arm randomized educational study (n = 58: AI-SP, human SP, standard curriculum) with pre– and post-training OSCE and self-efficacy assessment. Full participant, instrument, and analysis details are given in Methods.

Phase 1: expert validation of AI-SP fidelity

This phase asked whether AI-SPs can produce educationally viable simulations, and whether fidelity depends on the LLM or on the pedagogical architecture. Blinded expert evaluation revealed significant between-model differences ($p < 0.001$), with a two-tier structure (Fig. 3a; Table 1; per-scenario and per-evaluator scores in Supplementary Tables 1 and 2): three models (Qwen-2.5 Plus, Claude 4.5 Sonnet, Gemini 2.5 Flash) exceeded the pre-specified educational viability threshold, while DeepSeek-R1 and GPT-4o fell below it. Post-hoc comparisons confirmed that the top-tier models significantly outperformed the lower tier (Bonferroni-corrected $p < 0.025$), while within-tier differences were non-significant.

The key finding was that student skill level, rather than model choice, drove fidelity variation. Two-way ANOVA confirmed that student level accounted for five-fold more variance than model identity (η^2 = 0.31 vs 0.06), with no significant interaction ($p < 0.001$; Table 2; Fig. 3b). The largest differentiation occurred in educational value and communication naturalness, and the competent–novice gap was consistent across all six dimensions and all five models (Supplementary Fig. 2).

This gradient extended to hidden information disclosure (Table 3): competent students discovered all 26 hidden Layer 3 items, while novice students discovered only three, none of which were safety-critical. Missed items included self-discontinued clopidogrel creating stent thrombosis risk (S3), unreported ephedrine-containing supplements (S4), and saffron capsules posing serotonin syndrome risk with concurrent sertraline (S5).

Fidelity was robust across scenario complexity (Fig. 3c). Top-performing models maintained scores near the viability threshold even for high-complexity scenarios, with decrements of < 1 point from low to high complexity. GPT-4o showed the largest complexity-related decline (change score = 1.2).

Phase 1b: ecological validation through live student interactions

This phase asked whether the Phase 1 findings hold when real students interact freely with AI-SPs, and whether the system can generate actionable curriculum diagnostic data. To assess whether scripted-dialogue findings generalise to authentic interactions, 31 medical students each completed all five scenarios, generating 155 live consultations. Sessions exhibited features absent from scripted dialogues, including spontaneous follow-up questions, unexpected tangents, premature topic transitions, and natural communication errors.

The overall hidden-information discovery rate was 65.6%, with substantial inter-scenario variation (Table 4): S2 (thyroidectomy) achieved the highest rate and S4 (lumbar disc) the lowest, consistent with the difficulty of eliciting information that patients minimise.

Per-item analysis revealed wide variation in discovery rates (Fig. 4). Items related to common clinical concerns (S2) were near-universally discovered, while the hardest items clustered in S4 (where the patient minimised substance use), with concussion history proving the most elusive. Student-level variation was similarly wide, with a clear separation between top- and bottom-quartile performers.

A post-training questionnaire indicated high overall satisfaction (Table 5; Cronbach's α = 0.85). Students rated learning effectiveness and communication skill improvement most highly, while their understanding of the three-layer architecture received the lowest score, suggesting that the pedagogical mechanism operated effectively without being transparent to learners.

Phase 2: randomized controlled trial

This phase asked whether AI-SP training produces skill gains comparable to human SP practice, and whether it confers any additional benefits. Sixty-one students were randomized via sex-stratified block randomization; three withdrew before training, yielding 58 analysed participants across three arms (Fig. 5). Groups were comparable at baseline on all measures (Table 6).

Inter-rater reliability differed markedly by scoring instrument: the binary checklist achieved good reliability (ICC = 0.72) while the global rating scale showed poor reliability (ICC = 0.15; Fig. 6). Accordingly, the checklist score was designated as the primary outcome, with the total score reported as secondary (full ICC values with 95% CIs in Supplementary Tables 3–6).

For the primary outcome, the OSCE checklist score, all three groups improved significantly from baseline, with no significant between-group difference (Table 6; Fig. 6a; per-scenario score changes in Supplementary Table 7 and Supplementary Fig. 3). The AI-SP group showed the numerically largest checklist improvement, but the effect size versus human SP was

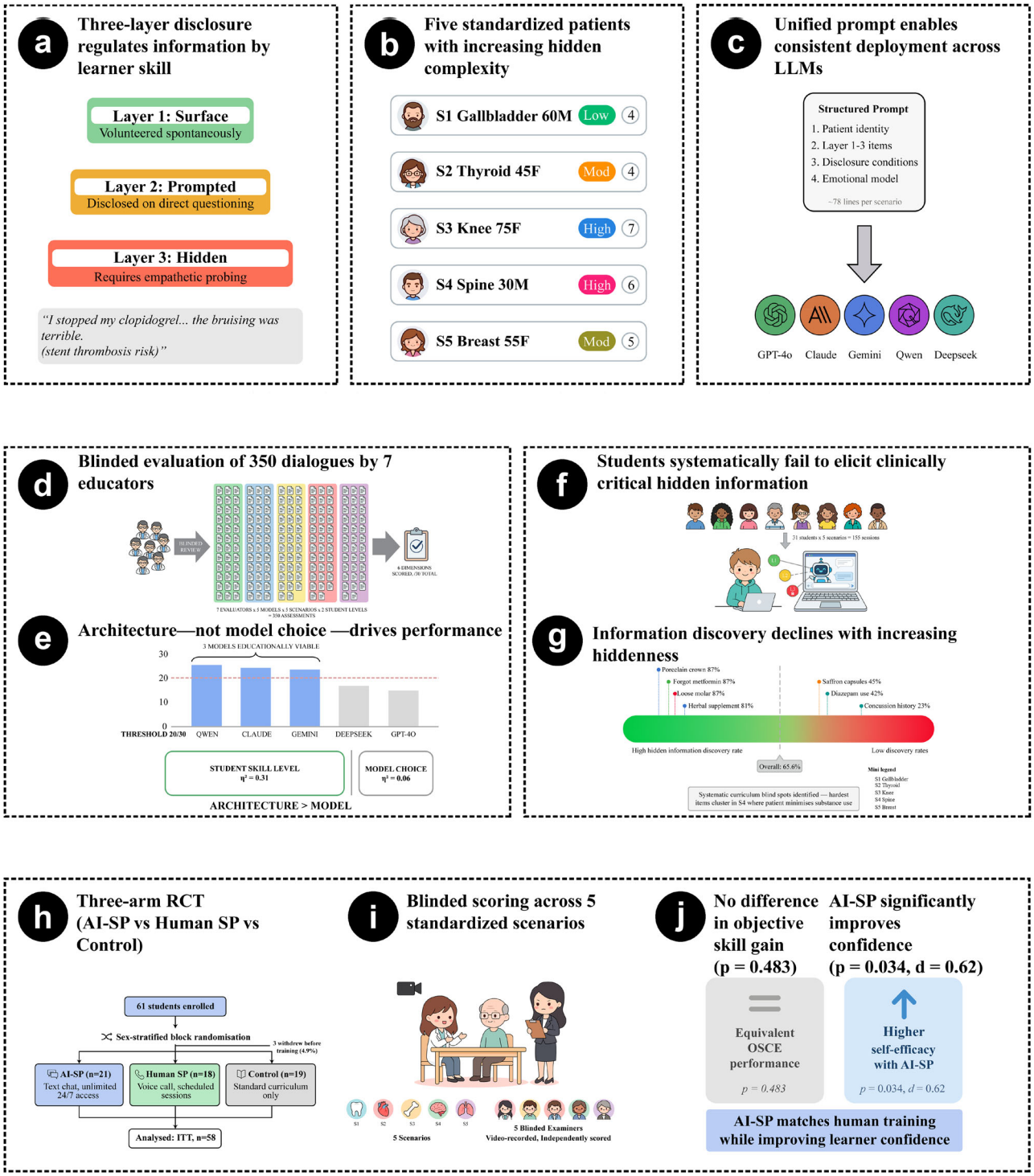


Fig. 1 | AI standardized patient system design and three-phase validation overview. a Three-layer information architecture: Layer 1 (surface) is volunteered spontaneously, Layer 2 (prompted) requires direct questioning, and Layer 3 (hidden) requires empathetic probing to unlock safety-critical information. **b** Five preoperative scenarios spanning low to high complexity, each containing safety-critical hidden items. **c** Unified structured prompt enabling consistent deployment across five large language models. **d** Phase 1 (expert evaluation of AI-SP dialogues): blinded evaluation of 350 dialogues by seven educators. **e** Phase 1: pedagogical architecture, rather than model choice, drives performance (η^2 = 0.31 vs 0.06). **f** Phase 1b (student interaction study): students systematically fail to elicit clinically critical hidden information. **g** Phase 1b: information discovery declines with increasing hiddenness. **h** Phase 2 (randomized controlled trial): three-arm design comparing AI-SP, human SP, and control arms. **i** Phase 2: blinded scoring across five standardized scenarios. **j** Phase 2: no significant difference in objective skill gain (p = 0.483) while AI-SP training significantly improves self-efficacy (p = 0.034, d = 0.62). AI-SP, AI standardized patient; SP, standardized patient; OSCE, objective structured clinical examination.

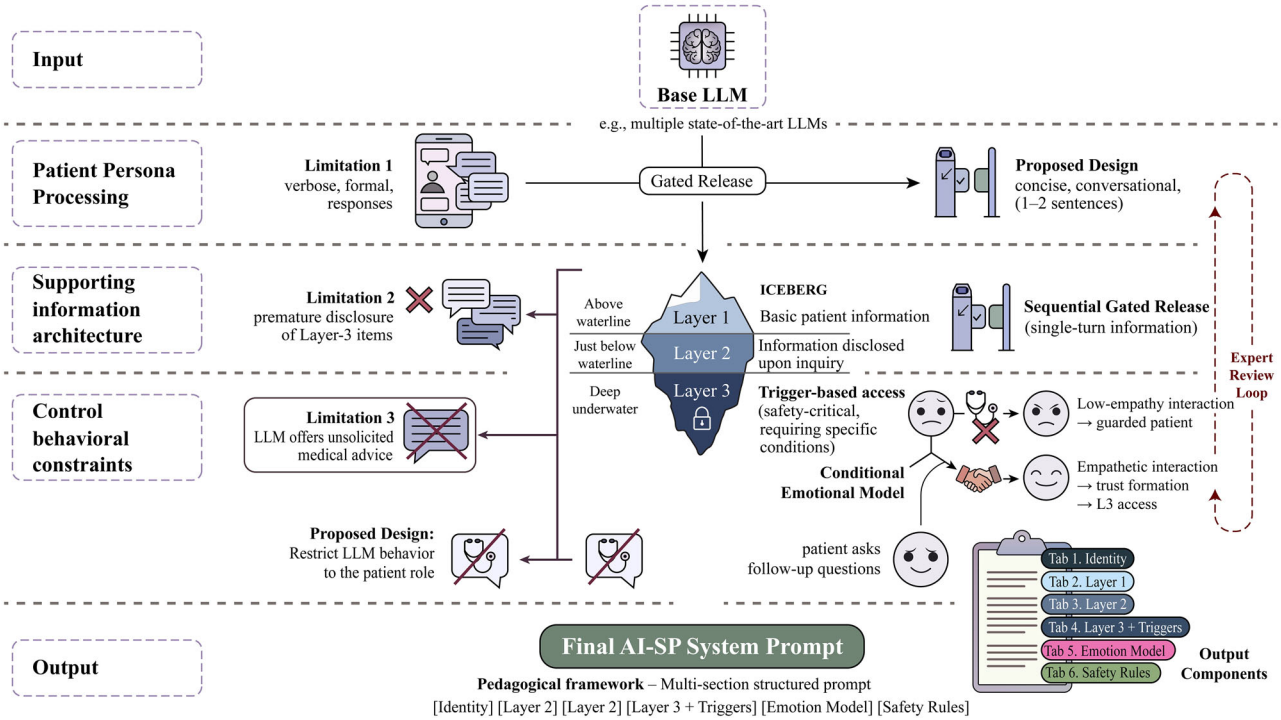


Fig. 2 | AI standardized patient prompt engineering framework. Starting from a base LLM, four iterative development rounds addressed distinct failure modes: verbose and formal patient responses (persona calibration), premature disclosure of safety-critical Layer 3 information (gated release with explicit trigger conditions), flat emotional affect regardless of interviewer quality (conditional emotional model linking empathy to trust and information access), and role boundary violations where the LLM offered unsolicited medical advice (behavioural constraints restricting the model to the patient role). Expert review after each round refined the prompt. The final system prompt comprises six structured sections (identity, Layers 1–3 with trigger conditions, emotional model, and safety rules), averaging 78 lines per scenario, deployed identically across all five LLMs.

small. The control group's comparable improvement suggests a substantial testing effect from repeated OSCE exposure. ANCOVA likewise showed no significant between-group difference after adjusting for baseline scores, which alone explained nearly half of post-training variance, leaving limited residual variance for training-specific effects.

The secondary outcome, the total OSCE score (checklist + global rating), showed a consistent pattern: all groups improved substantially with no significant between-group difference (Fig. 6c; Table 6), paralleling the checklist finding with a medium AI-SP versus human SP effect size.

Self-efficacy improved within all three groups (Wilcoxon signed-rank $p < 0.01$ in each arm). The between-arm difference in self-efficacy change was not significant on the omnibus test (Kruskal–Wallis $p = 0.064$; Table 6). In exploratory pairwise comparisons, the AI-SP arm's gain exceeded that of controls (+10.2 vs +6.1; unadjusted $p = 0.034$, $d = 0.62$) and, less clearly, that of the human SP arm (+10.2 vs +6.8; unadjusted $p = 0.069$, $d = 0.51$); neither difference remained significant after Holm correction (adjusted $p = 0.103$ and 0.138, respectively). The AI-SP arm also started from the lowest baseline (Table 6). Individual trajectories (Fig. 7a) show nearly uniform upward shifts among AI-SP students, whereas the human SP and control groups were more variable, including several students whose self-efficacy declined, partly reflecting differences in baseline. Per-item analysis (Supplementary Table 8; Supplementary Fig. 4) showed the largest numerical AI-SP advantages in managing time pressure and handling special cases, but no item-level difference between arms was significant (all Kruskal-Wallis $p \geq 0.060$, uncorrected).

Training satisfaction did not differ significantly between arms (Table 6). Per-item comparison (Fig. 7b; all eleven satisfaction dimensions in Supplementary Fig. 5) shows a mixed pattern beneath this overall similarity: AI-SP scored numerically higher on time flexibility and emotional support (affordances of asynchronous, judgement-free practice), while human SP scored higher on interaction naturalness, the one dimension where voice-based interaction has an inherent advantage. That the two modalities achieved near-identical satisfaction despite these complementary strengths suggests students valued different aspects of each format. Both arms scored above 3.8/5.0 on all dimensions, with high willingness to continue using AI-SP and recommend it to peers, indicating acceptability for sustained deployment.

Discussion

This study shows that structured AI standardized patients, governed by a three-layer information architecture, can achieve expert-validated fidelity and produce meaningful diagnostic data about student clinical communication skills. When used for training in a pilot trial, they yielded OSCE gains that did not differ significantly from those after human SP practice or test-retest learning alone, together with an exploratory signal of larger self-efficacy gains in the AI-SP arm. Three findings merit particular discussion.

The most consequential finding is that the variance attributable to student skill level ($\eta^2 = 0.31$) was five-fold larger than that attributable to model choice ($\eta^2 = 0.06$). This is consistent with the three-layer information architecture, rather than the underlying LLM, being a primary driver of pedagogical differentiation; we note, however, that because the same architecture was applied to every model, our design cannot fully separate the architecture's contribution from that of prompt quality, persona coherence, or language register (see the limitations discussed below). The practical implication is that institutions need not invest in the most expensive commercial API to achieve educationally adequate AI-SP performance: under the same architecture, several models met the viability threshold, so in this setting the LLM behaved as a largely replaceable engine. Indeed, three models spanning distinct accessibility tiers (a Chinese-developed model, Qwen-2.5 Plus; a premium Western model, Claude 4.5 Sonnet; and a free-tier model, Gemini 2.5 Flash) all exceeded the viability threshold, while the open-weight DeepSeek-R1—accessed via API in this study but deployable

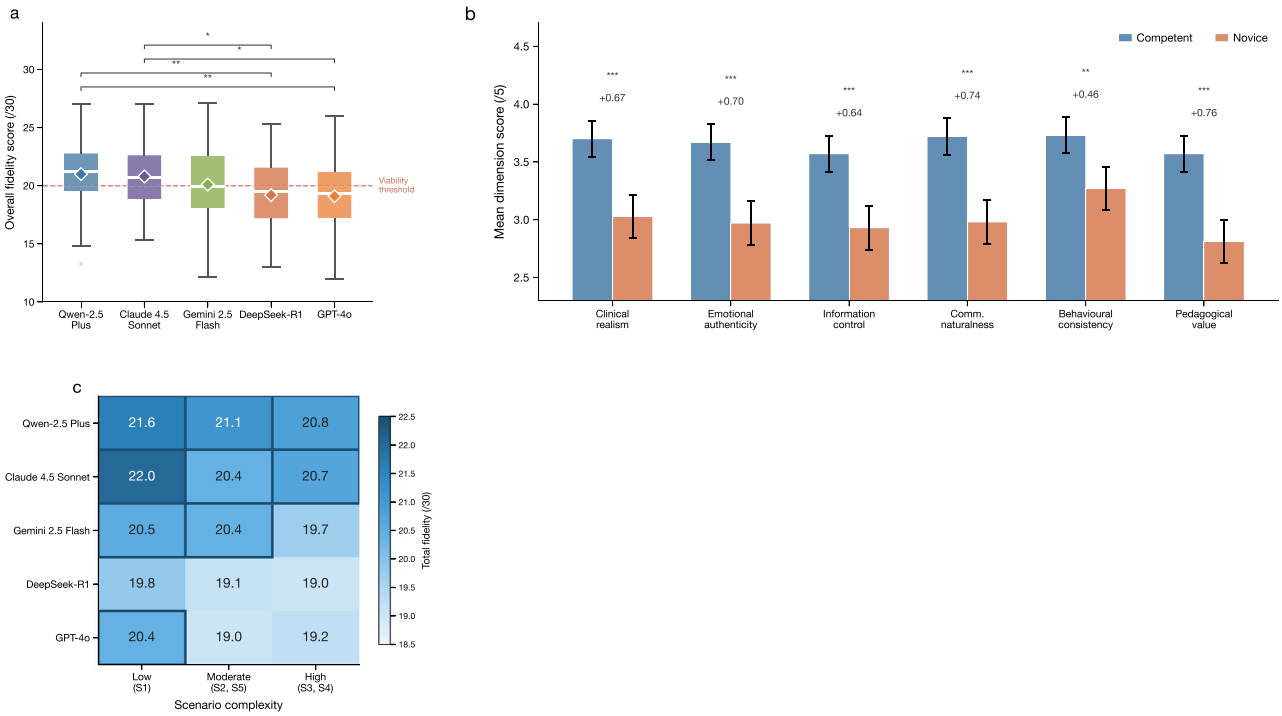


Fig. 3 | Phase 1 expert fidelity evaluation ($n = 7$ evaluators, 350 assessments). **a** Overall fidelity by model. Box plots show median and IQR; diamonds indicate means; dashed line marks the educational viability threshold (≥20/30). Significance brackets: Bonferroni-corrected (*$p < 0.05$; **$p < 0.01$). **b** Fidelity comparison between competent and novice student interactions across six dimensions. Error bars: SEM. **c** Fidelity heatmap across models and scenario complexity levels (Low: S1; Moderate: S2, S5; High: S3, S4). Bold cell borders indicate cells ≥20/30 (educationally viable).

Table 1 | AI-SP fidelity scores across six evaluation dimensions (mean ± SD, scale 1–5). Total scores sum to 30

Model	Clin. real.	Emot. auth.	Info. ctrl.	Comm. nat.	Behav. cons.	Pedag. val.	Total (/30)
Qwen-2.5 Plus	**3.51 ± 0.70**	**3.60 ± 0.75**	**3.34 ± 0.59**	**3.59 ± 0.71**	3.60 ± 0.62	3.39 ± 0.71	**21.0 ± 2.9**
Claude 4.5 Sonnet	3.53 ± 0.65	3.41 ± 0.73	3.26 ± 0.56	3.46 ± 0.77	**3.63 ± 0.57**	**3.53 ± 0.61**	20.8 ± 2.9
Gemini 2.5 Flash	3.53 ± 0.76	3.33 ± 0.77	3.20 ± 0.67	3.41 ± 0.79	3.46 ± 0.72	3.19 ± 0.77	20.1 ± 3.5
DeepSeek-R1	3.23 ± 0.75	3.21 ± 0.66	3.23 ± 0.59	3.20 ± 0.73	3.30 ± 0.64	3.07 ± 0.79	19.2 ± 3.1
GPT-4o	3.10 ± 0.85	3.19 ± 0.75	3.21 ± 0.61	3.13 ± 0.70	3.33 ± 0.61	3.13 ± 0.80	19.1 ± 3.3

Bold indicates highest score per dimension. Educational viability threshold: ≥20/30 (≥3.33 mean per dimension). Three models exceeded this threshold; DeepSeek-R1 and GPT-4o fell slightly below.

Table 2 | AI-SP fidelity scores by student interaction level (total score /30)

Model	Competent	Novice	Change
GPT-4o	21.23	16.94	4.29
Gemini 2.5 Flash	22.17	18.06	4.11
Claude 4.5 Sonnet	22.66	18.97	3.69
Qwen-2.5 Plus	22.57	19.49	3.09
DeepSeek-R1	20.66	17.83	2.83

Change = competent − novice. Higher change indicates greater sensitivity to student skill level. Sorted by differentiation magnitude.

self-hosted—offers a pathway to institution-controlled deployment for settings with data-residency constraints[22]. This landscape means resource-constrained programmes can select from multiple viable options and focus investment on instructional design rather than model selection.

That said, model choice is not irrelevant. In our Chinese-language context, Qwen's native training data conferred advantages in communication naturalness and emotional authenticity, dimensions requiring culturally calibrated patient behaviour including colloquial expressions and age-appropriate speech patterns. GPT-4o's lower fidelity despite strong clinical knowledge[23] highlights that benchmark superiority does not guarantee culturally authentic patient simulation, a finding relevant for non-English medical education contexts worldwide.

The three-layer architecture's 88.5 percentage-point discovery differential between competent and novice students is consistent with its functioning as a pedagogical mechanism that operationalises cognitive apprenticeship principles[21]: just as real patients share sensitive information selectively based on interaction quality[20], AI-SPs withhold Layer 3 items until students demonstrate the requisite communication skills. This mechanism offers three advantages over traditional SP training. First, the hidden-information discovery rate provides an *objective, automated assessment metric* without requiring expert observation. Second, AI-SPs apply disclosure rules deterministically, creating a *consistent assessment standard* that human SPs (whose disclosure behaviour may vary despite training[3]) cannot match. Third, *per-item difficulty profiling* generates granular curriculum diagnostic data that would be difficult to obtain at scale with human SPs.

Table 3 | Hidden information (Layer 3) discovery rates by scenario and student level

Scenario	Hidden items	*n*	Competent	Novice
S1 (Low)	Aspirin, alcohol, snoring, loose molar	4	4/4 (100%)	0/4 (0%)
S2 (Mod.)	Herbal thyroid tea, dental prosthesis, missed metformin, unresolved cold	4	4/4 (100%)	0/4 (0%)
S3 (High)	Self-stopped clopidogrel, palpitations, husband death, OSA, memory decline, dentures, penicillin allergy	7	7/7 (100%)	1/7 (14.3%)
S4 (High)	Daily ibuprofen, fitness supplements (ephedrine), concussion, heavy drinking, benzodiazepine use, e-cigarettes	6	6/6 (100%)	0/6 (0%)
S5 (Mod.)	Saffron capsules (5-HT risk), double-dose lorazepam, extreme diet, advance directive, smoking	5	5/5 (100%)	2/5 (40.0%)
Total		**26**	**26/26 (100%)**	**3/26 (11.5%)**

Novice students failed to discover any safety-critical hidden items across all five scenarios. *n* = number of hidden Layer 3 items per scenario.

Table 4 | Phase 1b live student–AI-SP interaction results by scenario (*n* = 31 students, 155 sessions)

Scenario	Hidden items	Discovery rate	Items found	Messages	Duration
S1 Cholecystectomy	4	65.3% ± 31.4%	2.6/4	45 ± 12	15.4 min
S2 Thyroidectomy	4	85.5% ± 34.0%	3.4/4	34 ± 9	14.9 min
S3 Knee replacement	7	67.3% ± 37.8%	4.7/7	38 ± 13	14.5 min
S4 Lumbar disc	6	51.1% ± 36.2%	3.1/6	37 ± 17	10.7 min
S5 Mastectomy	5	58.7% ± 42.2%	2.9/5	35 ± 13	13.3 min
Overall	**26**	**65.6% ± 37.8%**	—	**38 ± 14**	**13.8 min**

Discovery rate = proportion of hidden Layer 3 items disclosed during the interaction. Values are mean ± SD across 31 students.

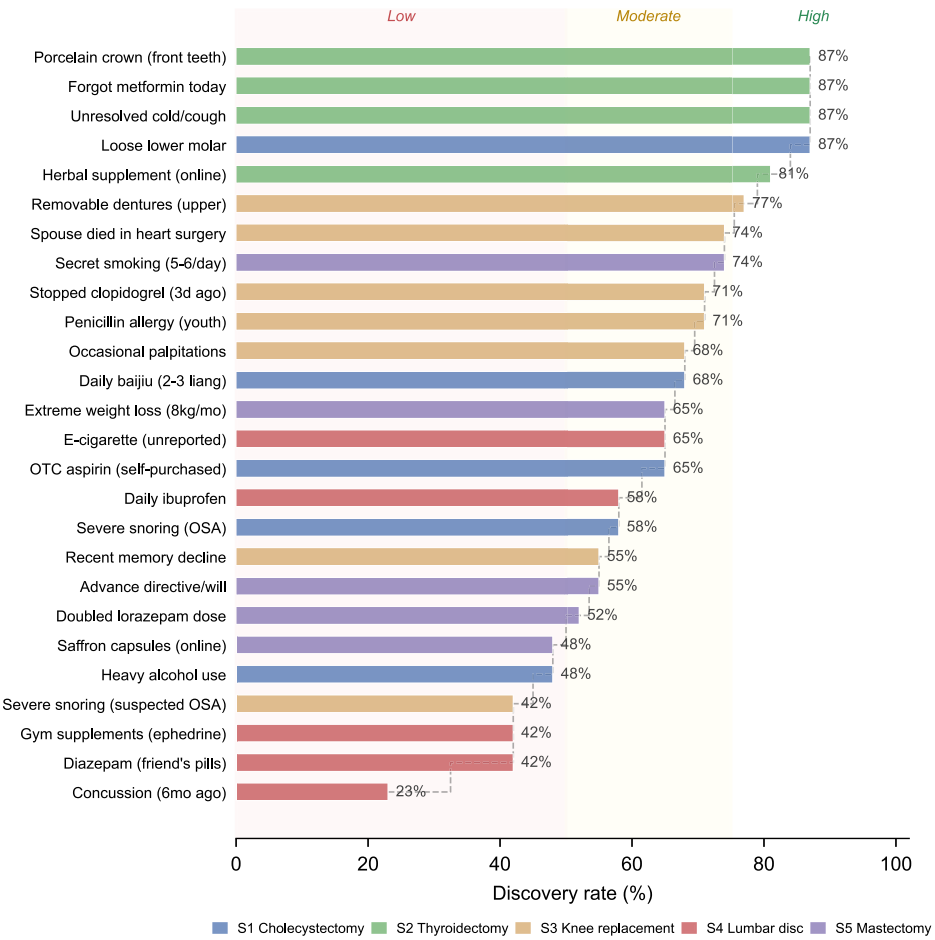


Fig. 4 | Per-item hidden-information discovery rates across five scenarios (*n* = 31 students, 26 items). Items ranked by discovery rate and colour-coded by scenario. S2 items were near-universally discovered (> 80%) while S4 items, particularly concussion history (23%) and diazepam use (42%), proved most challenging, reflecting the difficulty of eliciting information that patients actively minimise.

Table 5 | Phase 1b post-training questionnaire results (1–5 Likert scale, *n* = 21 respondents)

Item	Mean ± SD
Learning effectiveness	4.76 ± 0.44
Communication skill improvement	4.71 ± 0.46
Overall satisfaction	4.52 ± 0.60
Comparable to human SP	4.33 ± 0.73
Understanding of three-layer architecture	3.67 ± 0.97

after correction for multiple comparisons. We consider the two outcomes in turn.

The absence of a statistically significant between-group OSCE difference ($p = 0.483$; checklist $d_{AB} = 0.38$) is consistent both with true equivalence and with a clinically meaningful difference that this pilot, with under 25 per cent power for an effect size of about 0.38, was unable to detect; it does not, by itself, constitute evidence of equivalence. A formal non-inferiority conclusion would require a pre-specified margin, a one-sided confidence interval, and a demonstration that its lower bound excludes that margin, a test this pilot was neither designed nor powered to perform. We therefore

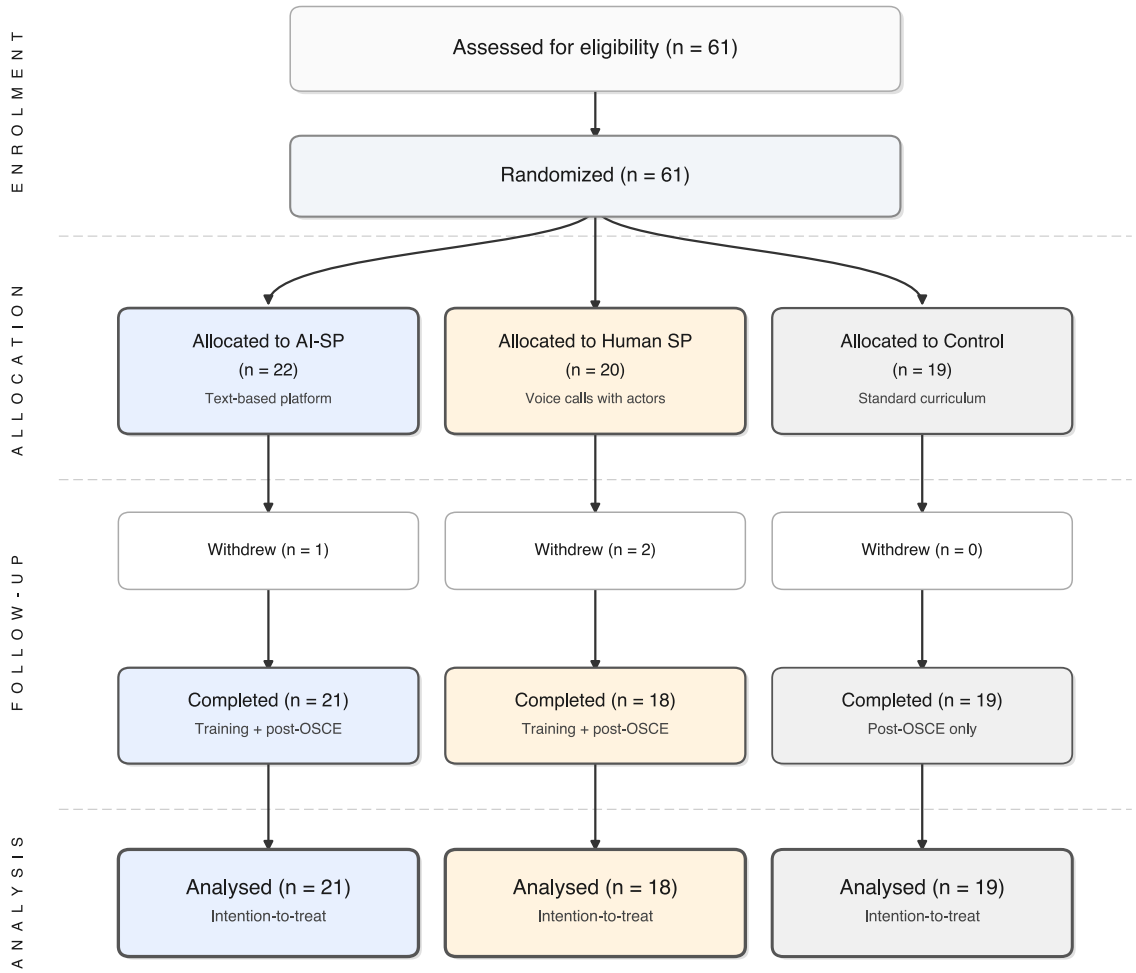


Fig. 5 | CONSORT flow diagram for the Phase 2 pilot RCT. Sixty-one students were randomized into three arms; three withdrew before training commenced (attrition rate: 4.9%). All 58 remaining participants completed the study protocol and were included in intention-to-treat analysis.

The Phase 1b ecological validation demonstrated this diagnostic capacity: per-item discovery rates ranging from 23% (concussion history) to 87% (porcelain crown) across 26 items revealed that students reliably inquire about conditions they associate with anaesthetic risk but consistently miss information that patients minimise or normalise. The concentration of low-discovery items in S4, where a young male patient normalises recreational supplement use and conceals benzodiazepine dependence, identifies a specific curricular gap in training students to recognise and probe patient minimisation behaviours. These findings extend the paradigm demonstrated by Luo et al.[16] in ophthalmology by showing that AI-SP systems can provide not only training but also systematic diagnostic data about educational blind spots[24].

In the Phase 2 pilot, all three groups (AI-SP, human SP, and untrained controls) achieved comparable OSCE skill gains, whereas self-efficacy gains differed numerically between arms without reaching statistical significance

report the observed effect sizes to inform the sample size of a future confirmatory trial rather than as an equivalence claim. We did not conduct a formal economic evaluation; as an order-of-magnitude anchor, the total large language model application-programming-interface cost for the entire study (all phases, more than 500 simulated consultations) was approximately CNY 100 (US$14), reflecting marginal inference cost only and excluding platform development, maintenance, and examiner time. A full cost-effectiveness comparison against human SP programmes remains to be performed. The dominant source of improvement was a testing effect: ANCOVA confirmed that baseline scores alone explained 48.5% of post-training variance, leaving limited variance for training-specific effects to capture[25]. Importantly, the three-arm design allows contextualisation that two-arm studies cannot: the largest published AI-SP RCT (Lavigne et al., $n = 247$ [17]) reported a significant AI advantage but compared AI only against standard training without a human SP arm, a design that cannot distinguish

Table 6 | Phase 2 RCT outcomes: baseline characteristics and pre–post changes across study arms

Outcome	AI-SP (n = 21)	Human SP (n = 18)	Control (n = 19)	p	Effect size
Baseline (pre-training)					
OSCE checklist (/14)	7.5 ± 1.1	7.7 ± 1.3	7.2 ± 1.5	0.715	—
OSCE total (/39)	18.5 ± 1.8	19.1 ± 2.7	18.3 ± 2.6	0.588	—
Self-efficacy (/50)	25.1 ± 9.1	30.6 ± 10.1	27.3 ± 9.5	0.193	—
Primary outcome					
Checklist change (ICC=0.72)	+1.4 ± 1.0	+0.9 ± 1.3	+1.0 ± 1.2	0.483[a]	d_{AB}=0.38
Secondary outcomes					
OSCE total change (ICC=0.40)	+8.6 ± 1.9	+7.3 ± 3.4	+7.6 ± 2.6	0.266[a]	d_{AB}=0.52
Self-efficacy change	+10.2 ± 6.3	+6.8 ± 6.8	+6.1 ± 6.8	0.064[a]	d_{AC}=0.62[c]
Satisfaction (/55)[e]	45.0 ± 7.2	44.4 ± 6.9	—	0.876[b]	d_{AB}=0.09

Values are mean ± SD.
Human SP: n = 18. Holm-adjusted across the three pairwise comparisons, p = 0.103.
[a]Kruskal–Wallis H test across three arms.
[b]Mann–Whitney U test (two arms).
[c]AI-SP vs Control Mann-Whitney U test: unadjusted p = 0.034.
d_{AB}: Cohen's d for AI-SP vs Human SP; d_{AC}: AI-SP vs Control.
[e]Satisfaction collected for intervention arms only (AI-SP: n = 21.

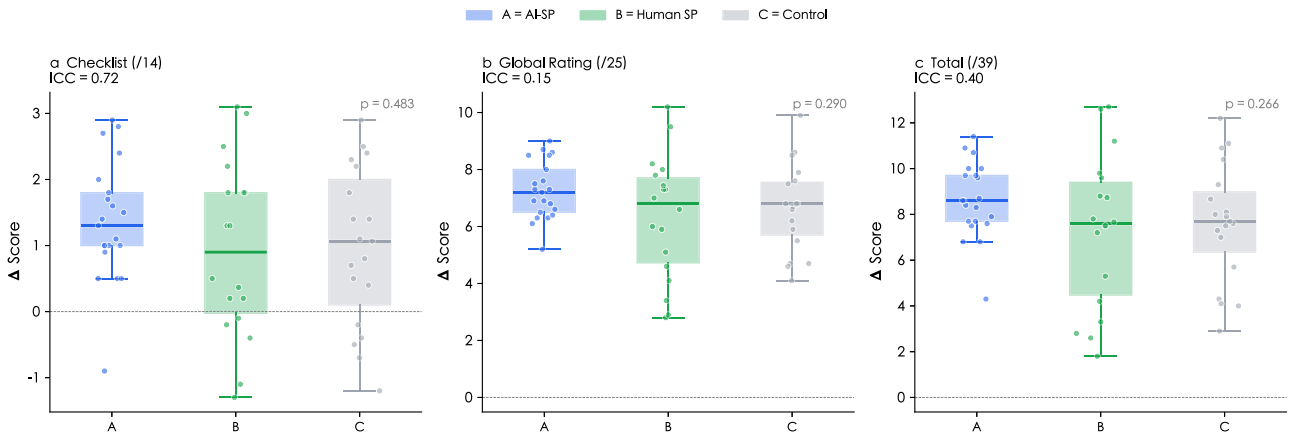


Fig. 6 | Phase 2 score changes by outcome measure. Each panel shows pre–post score change by study arm (box plots with individual data points; diamonds indicate means). Panel letters (**a**–**c**) denote the three outcome measures; within each panel the study arms are labelled A (AI-SP), B (Human SP), and C (Control). **a** Checklist score (/14; ICC = 0.72; primary outcome): all groups improved comparably (p = 0.483). **b** Global rating score (/25; ICC = 0.15): poor inter-rater reliability motivated its exclusion as primary outcome. **c** Total OSCE score (/39; ICC = 0.40): no statistically significant between-arm difference (underpowered pilot) (p = 0.266).

AI efficacy from the absence of an active comparator. Our pilot RCT (n = 58), though smaller, provides a structurally more informative comparison: the inclusion of both human SP and no-intervention arms positions a future adequately powered study to test superiority (vs control) and non-inferiority (vs human SP) simultaneously; the present pilot supplies the effect-size estimates such a test requires. This is a qualitative design advantage, not merely a quantitative limitation. The pilot was designed to estimate effect sizes for a future confirmatory trial[26,27] rather than to detect small differences; post-hoc power analysis (N≥216 needed for 80% power at d = 0.38 [28]) confirms that definitive equivalence testing requires a larger multicentre study. Critically, the three-phase design means the RCT is contextualised by 350 expert assessments and 155 live student sessions; the total evidence base supporting AI-SP validity is substantially larger than the Phase 2 sample alone.

The self-efficacy results should be read as exploratory. AI-SP students gained more confidence than controls (mean change = +10.2 vs +6.1; d = 0.62) and than human SP students (+10.2 vs +6.8; d = 0.51), but the three-arm comparison was not significant (p = 0.064) and neither pairwise difference survived Holm correction (adjusted p = 0.103 and 0.138). The mechanisms below are therefore hypotheses for a confirmatory trial to test, not explanations of an established effect. Bandura's self-efficacy theory[29] identifies four sources of efficacy beliefs: mastery experience, vicarious experience, verbal persuasion, and physiological/affective states. AI-SP training may amplify two of these. First, the psychologically safe environment, where errors carry no interpersonal consequences and practice can be repeated at self-selected times, reduces performance anxiety (affective state), enabling students to accumulate positive mastery experiences without the social evaluation apprehension inherent in human SP encounters. Second, unlimited repetition provides more frequent mastery experiences per unit time than scheduled human SP sessions. The largest per-item AI-SP advantage was in managing time pressure, which is compatible with this account, although that item-level difference was not significant (p = 0.060, uncorrected). Self-efficacy is not merely a subjective outcome; in clinical education, it predicts willingness to attempt challenging patient interactions, persistence when communication breaks down, and ultimately clinical performance in unsupervised settings[4]. Notably, the modality difference (text vs voice) may also contribute: text-based interaction allows more deliberate response composition, potentially reinforcing a sense of communicative competence that transfers to self-efficacy ratings.

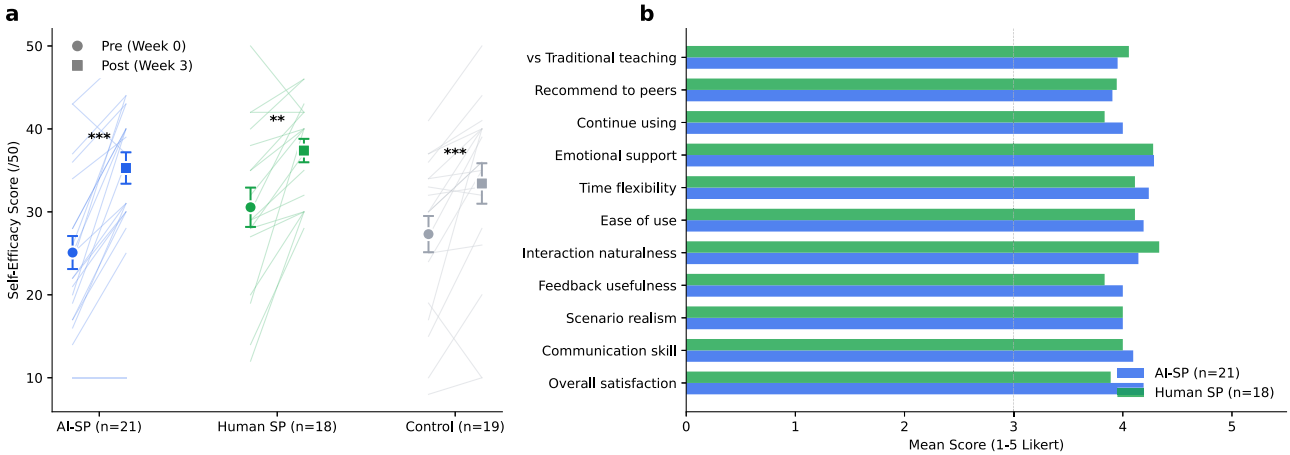


Fig. 7 | Phase 2 secondary outcomes. a Self-efficacy scores (/50) before and after training. Individual student trajectories (thin lines) and group means (± SEM; circles = pre, squares = post). All groups improved significantly (**p < 0.01; ***$p < 0.001$), and the AI-SP group showed the largest mean gain (mean change = +10.2); its difference from controls ($d = 0.62$; unadjusted $p = 0.034$, Holm-adjusted $p = 0.103$) is exploratory, as the three-arm Kruskal–Wallis test was not significant ($p = 0.064$). **b** Per-item training satisfaction comparison between AI-SP and human SP arms (1–5 Likert scale). No item showed significant between-group differences (all $p > 0.27$); interaction naturalness numerically favoured human SP, while time flexibility favoured AI-SP.

Training satisfaction did not differ significantly between AI-SP and human SP arms ($p = 0.876$), with both scoring above 3.8/5.0 on all dimensions. This finding contrasts with early concerns that students would resist AI-mediated clinical training, and suggests that the educational value and convenience of AI-SP practice compensate for any perceived limitations in interaction naturalness.

Notably, the entire Phase 2 protocol (training, assessment, and data collection) was conducted fully remotely. AI-SP training used text-based chat; human SP training used voice calls via videoconferencing software with trained actors following identical scripts; OSCE assessments were video-recorded and scored asynchronously by blinded examiners. This shows that an AI-SP training programme can be delivered without dedicated simulation facilities, on-site scheduling infrastructure, or synchronous examiner coordination, a deployment pathway particularly relevant for programmes serving geographically distributed learners and infeasible with traditional in-person SP programmes. The practical implication is straightforward: if a confirmatory trial upholds these pilot estimates, namely comparable skill gains, a possible self-efficacy advantage, and comparable satisfaction at low marginal cost, then the principal barrier to wider adoption would shift from evidence toward implementation.

Several limitations warrant consideration. First, the AI-SP and human SP arms used different communication modalities (text-based chat vs voice calls), introducing a confound that prevents attributing outcome differences solely to the AI vs human distinction. This was a deliberate design choice reflecting current deployment reality: text-based LLM interaction is the most mature, accessible, and reproducible modality available today, while human SP training is inherently voice-based; forcing it into text would misrepresent the comparator and reduce ecological validity. Matching modalities would improve internal validity but at the cost of testing a comparison that no institution would actually implement. Notably, text-based interaction is conventionally considered a communicative disadvantage relative to voice, as it lacks paralinguistic cues, real-time turn-taking, and tonal modulation that are central to clinical communication[30,31]. That text-based AI-SP reached comparable OSCE outcomes despite text's recognised communicative disadvantage is suggestive, but we emphasise that this reasoning does not control the confound: modality (text vs voice) and agent (AI vs human) remain entangled in our design, and no rhetorical argument substitutes for a factorial comparison. This is the study's most significant unresolved design limitation. A confirmatory study should cross the two factors, at minimum by adding a text-based human SP arm and a voice-based AI-SP arm, so that modality and agent effects can be estimated separately.

Second, the Phase 2 RCT ($n = 58$) was designed and powered as a pilot study; while baseline equivalence was confirmed, the sample was insufficient to detect small between-group OSCE differences (post-hoc power < 25% for observed $d = 0.38$; $N \geq 216$ needed for 80% power). The absence of a significant OSCE difference is therefore compatible with either genuine equivalence or limited statistical power, a distinction that only a larger confirmatory trial can resolve. Secondary outcomes were compared without pre-specified correction for multiple comparisons, so the self-efficacy findings are hypothesis-generating.

Third, within-group variability in OSCE score changes was substantial (checklist change score SD = 1.0–1.3 on a 14-point scale), compounded by a strong testing effect that elevated all groups' scores and compressed between-group differences. We addressed this by: (a) using ANCOVA to adjust for baseline scores, which accounted for 48.5% of post-training variance; (b) pre-specifying the high-reliability checklist (ICC = 0.72) as the primary outcome; and (c) reporting both parametric (ANCOVA) and non-parametric (Kruskal–Wallis, Mann-Whitney) tests. Fourth, using the same five scenarios for both training and assessment introduced scenario-specific familiarity effects. This confound acts in a specific direction: it inflates all post-training scores and compresses between-group differences, making a null between-arm result more likely regardless of true training effects, and, unlike baseline imbalance, it cannot be removed by statistical adjustment. It should therefore be read as an intrinsic limit on this design's ability to detect between-arm differences. Future implementations should employ a larger scenario bank with non-overlapping training and assessment scenarios to isolate genuine communication skill transfer from scenario memorisation.

Fifth, the study assessed immediate post-training outcomes only; longer-term skill retention (4–12 weeks) and transfer to real patient encounters were not evaluated. Sixth, inter-rater reliability for the global rating scale was poor (ICC = 0.15), necessitating reliance on the checklist score as the primary outcome. While binary checklist items are inherently more reliable than subjective Likert scales, this limits assessment to observable behaviours rather than holistic communication quality. As Newble notes, checklists and global ratings suit different stations, and global judgements depend on assessor experience[32], so lower global-rating agreement partly reflects this instrument's known sensitivity to rater calibration. Low global rating reliability is a well-documented challenge in OSCE assessment: Regehr et al.[33] reported global rating ICCs of 0.20–0.40 across clinical tasks, and Gershov et al.[34] recently reported comparable variability in AI-assisted competency assessment. Our decision to report ICC transparently and conduct three layers of robustness analysis (leave-one-out, Supplementary Table 9; sensitivity, Supplementary Table 10; examiner variability,

Supplementary Table 6; Supplementary Note 1, Supplementary Figures 6–8) represents a methodological strength. Enhanced examiner calibration with video-anchored exemplars should be prioritised in future studies.

Additional methodological considerations include: the study was conducted at a single institution with Chinese-language scenarios, limiting cross-linguistic and cross-cultural generalisability. We designed the disclosure rules to operate at the level of communication behaviour rather than language syntax, which we conjecture may support transfer across languages; however, all scenarios and prompt engineering were Chinese-language, and whether Layer 3 trigger conditions transfer across languages and cultures is an untested empirical question, not an established property. The five-model comparison does embed deployment diversity across US and Chinese providers, open- and closed-weight platforms, and three accessibility tiers, but this speaks to model portability, not cross-linguistic generalisability. The remote video-conferenced OSCE format, while validated in post-pandemic literature[35,36], does not assess paralinguistic and non-verbal communication skills. A multicentre, multilingual replication across specialties is planned as the next phase of this research programme.

Several further limitations qualify the interpretation and deployment of these findings. First, the "architecture over model" inference is partially confounded by design: because the identical architecture was applied to every model, the larger variance attributable to student skill than to model identity ($\eta^2 = 0.31$ vs 0.06) cannot by itself separate the specific contribution of the three-layer disclosure mechanism from that of prompt quality, persona coherence, or language register. A stronger test would compare the same models with and without the architecture; we frame the present result as consistent with, but not proof of, an architecture-dominant account. Second, Phase 1 fidelity was assessed on scripted dialogues authored to represent competent and novice students; real novices ramble, misunderstand questions, and pursue tangents in ways scripts do not, which limits the predictive validity of scripted-dialogue fidelity for live performance. Relatedly, Phase 1b was conducted with a single model (Qwen-2.5 Plus); its per-item discovery rates, session durations, and satisfaction data therefore characterise that deployment and should not be generalised to the other four models evaluated in Phase 1. Third, the standardised patient, whether AI or human, is a deliberately bounded model of a person: a genuine patient is more than a set of partly hidden cues and risk factors, and skilled clinicians "read between the lines" in ways that matter most outside the structured pre-anaesthesia encounter (for example, in family medicine); whether current models support this remains uncertain, and our task deliberately does not test it. Fourth, all evaluated models are hosted services under continuous development and subject to silent updates, deprecation, and capability drift, so the published prompts may over time elicit different behaviour from the models they were tuned for; deployments should pin model versions and periodically re-validate. Fifth, in real deployment students aware that the patient is an AI may attempt to elicit hidden information through direct requests or prompt injection rather than skilled interviewing; adversarial prompt hardening was not part of the present system design and is required before high-stakes assessment use. Sixth, student–AI conversation logs are recorded, analysed, and stored, raising data-governance obligations; for the public repository, participant identifiers were replaced with anonymous codes and free-text comments and timestamps were removed prior to release, but a full governance framework (retention, access control, and jurisdiction-specific compliance) should accompany any operational deployment.

Taken together, this study addresses three gaps in the emerging AI-SP literature. Unlike prior single-model studies[16,17,37], the five-model comparison provides multi-model evidence on the portability of the pedagogical architecture. Unlike two-arm designs that compare AI-SP to standard training, our three-arm RCT includes the clinically meaningful human SP comparator alongside a no-intervention control, enabling both superiority (vs control) and, in a future adequately powered study, non-inferiority (vs human SP) inferences. And unlike technical benchmarking studies[18], the three-phase validation pipeline progressively establishes construct validity, ecological validity, and training efficacy. Future research should prioritise adequately powered multicentre equivalence trials with extended follow-up, cross-specialty validation, investigation of optimal AI-SP–human-SP integration ratios in curricula, and development of multimodal AI-SPs incorporating voice and video interaction. The exploratory self-efficacy signal should be tested as a pre-specified outcome, with its psychological mechanisms examined directly. The three-layer information architecture may also be applicable beyond medical education to any professional training context where skilled questioning is central, including journalism, social work, and legal interviewing.

Methods

Study design and ethics

This three-phase study was conducted at the Department of Anaesthesiology, The First Affiliated Hospital of Dalian Medical University, China. Phase 1 validated AI-SP fidelity through blinded expert evaluation of scripted dialogues. Phase 1b extended validation to live student–AI-SP interactions. Phase 2 employed a three-arm, parallel-group randomized controlled trial. The study was reported following DECIDE-AI guidelines[38] for AI system evaluation (completed checklist in Supplementary Table 11) and CONSORT guidelines for the RCT component (completed checklist in Supplementary Table 12). The study was determined to be exempt from formal ethics review by the Ethics Committee of The First Affiliated Hospital of Dalian Medical University (exemption reference: YJ-MS-2025-02, dated 9 October 2025), as it involved comparison of standard educational methods among medical students with no patient contact, no clinical intervention, and no collection of biological specimens. This was an educational evaluation comparing teaching methods rather than a clinical trial of a health intervention; consistent with its educational nature and its exemption from clinical ethics review, it was not registered in a clinical trials registry. We acknowledge the absence of prospective registration as a limitation and address its implications for outcome selection in the Discussion. All student participants provided written informed consent.

Scenario design and AI-SP construction

Five preoperative anaesthesia consultation scenarios were developed through iterative expert consensus, stratified by complexity: S1 (low: laparoscopic cholecystectomy, 60M), S2 (moderate: thyroid nodulectomy, 45F), S3 (high: total knee replacement, 75F), S4 (high: lumbar disc surgery, 30M), and S5 (moderate: mastectomy, 55F). Scenarios covered the spectrum of preoperative communication challenges including medication reconciliation, substance use disclosure, psychosocial assessment, and management of acute anxiety.

Each AI-SP was constructed with a patient persona and the three-layer information architecture described in the Introduction (Fig. 1a). Each scenario contained 4–7 hidden Layer 3 items, including at least one safety-critical item whose non-discovery would represent a clinically significant assessment gap (Table 3). System prompts were structured in four sections: *patient identity* (demographics, personality, speech patterns), *Layer 1 items* (volunteered spontaneously), *Layer 2 items* (disclosed when directly asked), and *Layer 3 items with disclosure conditions* (gated behind specific communication behaviours). Additionally, each prompt specified an emotional response model defining baseline anxiety, triggers for escalation or de-escalation, defensive reactions to insensitive questioning, and gradual trust-building responses to empathetic communication[1,39]. All five LLMs received identical system prompts. Model parameters were standardised across models (temperature = 0.7, top_p = 0.9). The response token limit was 4096 for Phase 1 scripted-dialogue generation; the live student-facing platform (Phases 1b and 2) applied a 2048-token per-turn limit appropriate to interactive chat.

Translating this architecture into effective system prompts required iterative prompt engineering (Fig. 2; Supplementary Table 13). Over four development rounds, a clinician–engineer team tested each draft prompt against all five LLMs across both competent and novice student interaction patterns, identifying and resolving four recurring failure modes: (i) *role*

boundary violations, where the model offered differential diagnoses or treatment suggestions rather than remaining in the patient role; (ii) *premature information disclosure*, where multiple Layer 3 items were revealed in a single turn regardless of conversational context; (iii) *affective flattening*, where the model responded with uniformly cooperative behaviour irrespective of the student's communication quality; and (iv) *stylistic artefacts*, including verbose, formal phrasing inconsistent with the patient persona and bracketed action descriptions (e.g., "[hesitates]") absent from natural patient speech. Each failure mode was addressed through targeted behavioural constraints appended to the prompt: an explicit safety block forbidding medical advice, a "one hidden item per turn" disclosure rule, conditional emotional response rules linking patient cooperation to interviewer empathy, and response length limits enforcing 1–3 sentence turns in colloquial register (representative prompt in Supplementary Table 13). The final prompts averaged 78 lines (~ 1.8 KB) per scenario.

Platform and model selection

The AI-SP system was deployed as a web-based platform (Fig. 1) comprising a real-time chat interface, a scenario management module, an LLM integration layer routing messages to the assigned model, and an automated data collection backend recording all dialogue turns, timestamps, and session metadata. For Phase 2, a server-side whitelist access control mechanism enforced group allocation: upon login, each participant's student ID was verified against the randomization list stored as a server environment variable, automatically assigning the correct study arm without self-selection. Only Arm A participants were granted access to the AI-SP chat interface; Arm B and Arm C participants received group-appropriate instructions but were technically unable to access the AI practice module, preventing cross-arm contamination at the platform level.

Five frontier LLMs were selected to span three accessibility tiers relevant to global deployment: GPT-4o (OpenAI; commercial API)[23], Claude 4.5 Sonnet (Anthropic; commercial API), Gemini 2.5 Flash (Google; free-tier API), Qwen-2.5 Plus (Alibaba; free-tier API), and DeepSeek-R1 (DeepSeek; an open-weight model that can be self-hosted but was accessed here through the official DeepSeek API). This selection enables comparisons across performance tiers (premium, cost-efficient, open-weight), geographic provenance (US, China), and provider ecosystems, representing the spectrum of options available to medical education programmes worldwide. For reproducibility, all models were accessed through their providers' APIs during February–March 2026 using the following identifiers: `gpt-4o`, `claude-sonnet-4-5-20250929`, `gemini-2.5-flash`, `qwen-plus` (Qwen-2.5 Plus), and `deepseek-reasoner` (DeepSeek-R1). Because all five were accessed as hosted services, model weights, parameter counts, and quantization were fixed by each provider and not under our control; we therefore report provider, identifier, and access window in place of local deployment specifications, and note that hosted models are subject to silent provider-side updates (addressed in the Discussion).

Phase 1: dialogue generation and expert evaluation

For each scenario, two scripted student interaction patterns were developed representing distinct skill levels: competent (empathetic, open-ended, systematic) and novice (checklist-style, minimal empathy, jargon-heavy), based on documented communication patterns from institutional OSCE assessments[40]. Each scenario × student level × model combination was generated once using standardised parameters, yielding 50 dialogues (5 scenarios × 2 levels × 5 models).

Seven anaesthesiology faculty served as blinded evaluators (mean clinical experience: 12.4 years, range: 8–18; mean simulation education experience: 7.2 years). Evaluators were blinded to model identity and presentation order was randomised. Each evaluator assessed all 50 dialogues, yielding 350 total assessments. Six dimensions were scored on 1–5 Likert scales with anchored rubrics: Clinical Realism, Emotional Authenticity, Information Control, Communication Naturalness, Behavioural Consistency, and Pedagogical Value. Prior to scoring, evaluators completed a 3-hour calibration workshop. The educational viability threshold (≥20/30, equivalent to ≥3.33 mean per dimension) was established through a two-round modified Delphi process among the seven evaluators: the threshold corresponds to the "adequate for unsupervised student use" anchor point on the rubric, and maps to the institutional OSCE borderline pass standard (60% on a converted percentage scale) that the department applies to human SP encounters. This is a criterion-referenced standard rather than a data-derived cut-point: it was fixed by expert consensus before scores were examined, following the principle that simulation fidelity standards should be anchored to the intended educational use of the tool[4]. We use it descriptively to distinguish models that met the department's "adequate for unsupervised practice" bar, not as an inferential test.

Phase 1b: live student validation

Thirty-one medical students (fourth- and fifth-year) each completed all five scenarios using Qwen-2.5 Plus (selected as the deployment model because it achieved the highest Phase 1 fidelity score, offers a free-tier API with low latency in China, and natively handles Chinese-language interaction without translation artefacts), generating 155 live sessions. The platform recorded all dialogue turns, timestamps, and session duration in real time. Hidden-information discovery was assessed automatically: upon session completion, the dialogue transcript was submitted to a second Qwen-2.5 Plus instance with an analysis prompt listing all Layer 3 items for that scenario, returning a binary classification (discovered/not discovered) for each item. Each hidden item has explicit trigger conditions hard-coded in the system prompt (Supplementary Table 13), and the detection task reduces to verifying whether the patient character's response contained the pre-specified disclosure content, a constrained binary classification with minimal ambiguity compared with open-ended assessment tasks. To validate this automated classification rather than assert its reliability, a random sample of 20 sessions (104 item-level judgements) was independently re-coded by three anaesthesiology educators, each blinded to the automated labels and to one another (Supplementary Table 14). Agreement between the automated classifier and the three human coders was 98.1%, 86.5%, and 95.2% (Cohen's κ = 0.93, 0.63, and 0.84; mean κ = 0.80), while the three coders agreed with one another at a mean κ of 0.79 (pairwise κ = 0.69–0.91). Against the majority judgement of the three coders, the automated classifier achieved 95.2% agreement (Cohen's κ = 0.84; 95% CI 0.71–0.98). The automated classifier therefore agreed with expert human judgement as closely as the experts agreed with one another, supporting the use of the automated discovery rate as the Phase 1b measurement instrument. All disagreements arose from items the classifier scored as disclosed but a coder scored as not disclosed; no item judged disclosed by a coder was missed by the classifier. Following all scenarios, students completed an anonymous 11-item Likert-scale questionnaire (1–5) spanning AI-SP realism, learning value, and adoption willingness, plus three open-ended questions. Internal consistency was assessed using Cronbach's α.

Phase 2: randomized controlled trial

This phase was designed as a pilot RCT; the target sample size of approximately 60 was determined by the single-cohort rotation size available during the study period rather than a formal a priori power calculation, consistent with recommendations for pilot trials aimed at estimating effect sizes and assessing feasibility[26,27]. Sixty-one medical students on anaesthesiology rotation were recruited. Participants were randomized using computer-generated sex-stratified block randomization (seed = 20260308; sex distribution: $\chi^2 = 0.07$, $p = 0.956$) into three arms: AI-SP (n = 22), human SP (n = 20), and control (n = 19). Three students withdrew before training commenced (one AI-SP, two human SP), yielding 58 analysed participants (AI-SP: 21, human SP: 18, control: 19; attrition rate: 4.9%). Students who had participated in Phase 1b were excluded to prevent contamination.

The entire protocol was designed as a fully remote intervention. Arm A (AI-SP) completed 10 practice consultations over one week (five scenarios, each attempted twice) using the text-based AI-SP platform, with automated post-session feedback. Arm B (Human SP) completed an identical schedule of 10 practice consultations via voice calls over videoconferencing software,

with trained graduate student actors following identical scenario scripts and three-layer disclosure rules. The two arms intentionally used different communication modalities: text-based chat for AI-SP reflects the current deployment reality of LLM-driven systems, while voice interaction for human SP preserves the ecological validity of actor-based training[35]. This design prioritises external validity, comparing AI-SP and human SP as they would actually be deployed, over strict internal control of modality. The implications of this modality difference are addressed in the Discussion. Arm C (Control) followed the standard curriculum with no additional practice. To prevent cross-arm contamination, the AI-SP platform enforced server-side whitelist access control: only Arm A student identifiers were authorised for the practice module.

The OSCE comprised two scoring instruments: a 14-item binary checklist assessing observable consultation behaviours (e.g., "verified patient identity", "explained anaesthetic options") and a 5-item global rating scale assessing holistic communication quality on 1–5 Likert scales with descriptive anchors (e.g., 1 = "no attempt", 3 = "adequate but formulaic", 5 = "exemplary, patient-centred"). The pre–post change in the OSCE checklist score (/14) was designated the primary outcome on the basis of its markedly superior inter-rater reliability (ICC(2,1) = 0.72) relative to the global rating scale (ICC = 0.15) and total score (ICC = 0.40). We note explicitly that this designation was made after inter-rater reliability had been estimated rather than pre-registered; the study was an educational evaluation and was not registered in a trials registry (see Study design and ethics). To guard against selective reporting, all three measures, the checklist, the global rating, and the total score, are reported in full (Table 6; Fig. 6), and their reliability is reported transparently. The poor global rating ICC likely reflects the inherently subjective nature of holistic quality judgements in asynchronous video scoring without real-time calibration; leave-one-out and sensitivity analyses (Supplementary Note 1, Supplementary Figures 6–8) confirmed that this was not driven by a single outlier examiner but reflected systematic inter-examiner variability across all five dimensions. OSCEs were conducted at baseline (Week 0) and immediately post-training (Week 1) across five stations (10 minutes each) via remote videoconference. All sessions were video-recorded and scored asynchronously by five blinded examiners, each assigned primary responsibility for one station with cross-scoring of 15 randomly selected recordings per station for ICC estimation (26% dual-rated). This video-recorded OSCE approach is consistent with validated telehealth assessment protocols[35,36]. Secondary outcomes included total OSCE score change (/39), self-efficacy change (10-item questionnaire, /50; assessed at Week 0 and Week 3), and training satisfaction (11-item Likert scale, /55; collected for Arms A and B only, as Arm C received no intervention to evaluate).

Statistical analysis

Phase 1 analyses employed one-way ANOVA (five models) and two-way ANOVA (model × student level; model × complexity) with Bonferroni-corrected post-hoc comparisons. Inter-rater reliability was assessed using ICC (two-way random, absolute agreement). Effect sizes are reported as η^2 (ANOVA) and Cohen's d (pairwise).

Phase 1b used descriptive statistics for per-scenario and per-item discovery rates. Sequential learning effects were assessed using paired t-tests (S1 vs S5). Pearson correlation assessed the engagement–performance relationship.

Phase 2 employed non-parametric tests given the small sample and non-normal distributions. Between-group comparisons used Kruskal–Wallis H tests with pairwise Mann–Whitney U tests. Pairwise comparisons were not pre-specified with a multiplicity correction; for the pairwise self-efficacy comparisons reported in the text we therefore give Holm-adjusted p values (across the three pairwise comparisons) alongside unadjusted values, and all secondary-outcome comparisons, including the ten item-level self-efficacy tests (reported uncorrected), are treated as exploratory. Within-group pre–post comparisons used Wilcoxon signed-rank tests. ANCOVA (post-training score ~ group + baseline score) was used to adjust for baseline differences. Effect sizes are reported as Cohen's d (between-group) and paired d (within-group). Inter-rater reliability was assessed using ICC(2,1) (two-way random, single measures, absolute agreement). All analyses were intention-to-treat. Statistical significance: $\alpha = 0.05$. Analyses used Python 3.9 (SciPy 1.11, statsmodels 0.14).

Data availability

The datasets generated and analysed during the current study (anonymised evaluation scores, AI-SP prompt templates, scenario specifications, scoring rubrics, complete anonymised dialogue transcripts, and the DECIDE-AI and CONSORT reporting checklists) are available in the GitHub repository at https://github.com/RegAItool/ai-standardized-patient, which also contains the source data underlying all figures and tables. The repository is not publicly available during peer review because the manuscript is under consideration; during this period the data are available from the corresponding author on reasonable request. The repository will be made publicly available upon acceptance.

Code availability

Analysis scripts (Python 3.9) for all statistical analyses and figure generation are available in the same GitHub repository at https://github.com/RegAItool/ai-standardized-patient, subject to the same access conditions during peer review.

Received: 28 March 2026; Accepted: 26 September 2026;
Published online: 23 October 2026

References

1. Makoul, G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. *Academic Med.* **76**, 390–393 (2001).
2. Ericsson, K. A. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Med.* **79**, S70–S81 https://doi.org/10.1097/00001888-200410001-00022 (2004).
3. Barrows, H. S. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Med.* **68**, 443–451 (1993).
4. Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L. & Scalese, R. J. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med. Teach.* **27**, 10–28 https://doi.org/10.1080/01421590500046924 (2005).
5. Blitz, J. D. et al. Preoperative evaluation clinic visit is associated with decreased risk of in-hospital postoperative mortality. *Anesthesiology* **125**, 280–294 https://doi.org/10.1097/ALN.0000000000001193 (2016).
6. Bosse, H. M., Nickel, M., Huwendiek, S., Schultz, J. H. & Nikendei, C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. *BMC Med. Educ.* **15**, 183 https://doi.org/10.1186/s12909-015-0468-1 (2015).
7. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 https://doi.org/10.1038/s41586-023-06291-2 (2023).
8. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Med.* **7**, 20 https://doi.org/10.1038/s41746-024-01010-1 (2024).
9. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **388**, 1233–1239 https://doi.org/10.1056/NEJMsr2214184 (2023).
10. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493–498 https://doi.org/10.1038/s41586-023-06647-8 (2023).
11. Kron, F. W. et al. Using a computer simulation for teaching communication skills: a blinded multisite mixed methods randomized

controlled trial. *Patient Educ. Counseling* **100**, 748–759 https://doi.org/10.1016/j.pec.2016.10.024 (2017).
12. Lane, H. C., Hays, M. J., Core, M. G. & Auerbach, D. Learning intercultural communication skills with virtual humans: feedback and fidelity. *J. Educ. Psychol.* **105**, 1026–1035 https://doi.org/10.1037/a0031506 (2013).
13. Thesen, T. & Park, S. H. A generative AI teaching assistant for personalized learning in medical education. *npj Digital Med.* **8**, 627 https://doi.org/10.1038/s41746-025-02022-1 (2025).
14. Mandal, S. et al. Utilization of generative AI-drafted responses for managing patient-provider communication. *npj Digital Med.* **8**, 591 https://doi.org/10.1038/s41746-025-01972-w (2025).
15. Stanceski, K. et al. The quality and safety of using generative AI to produce patient-centred discharge instructions. *npj Digital Med.* **7**, 329 https://doi.org/10.1038/s41746-024-01336-w (2024).
16. Luo, M.-J. et al. A large language model digital patient system enhances ophthalmology history taking skills. *npj Digital Med.* **8**, 502 https://doi.org/10.1038/s41746-025-01841-6 (2025).
17. Lavigne, E. et al. AI-standardized clinical examination training on OSCE performance. *NEJM AI* **2** https://doi.org/10.1056/AIoa2500066 (2025).
18. Yu, H. et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Commun. Med.* **6**, 27 https://doi.org/10.1038/s43856-025-01283-x (2025).
19. Haltaufderheide, J. & Ranisch, R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models. *npj Digital Med.* **7**, 183 https://doi.org/10.1038/s41746-024-01157-x (2024).
20. Nestel, D. & Tierney, T. Role-play for medical students learning about communication: guidelines for maximising benefits. *BMC Med. Educ.* **7**, 3 https://doi.org/10.1186/1472-6920-7-3 (2007).
21. Collins, A., Brown, J. S. & Holum, A. Cognitive apprenticeship: making thinking visible. *Am. Educator* **15**, 6–11 (1991).
22. Zhang, G. et al. Closing the gap between open source and commercial large language models for medical evidence summarization. *npj Digital Med.* **7**, 239 https://doi.org/10.1038/s41746-024-01239-w (2024).
23. Wu, C. Towards evaluating and building versatile large language models for medicine. *npj Digital Med.* **8**, 58 https://doi.org/10.1038/s41746-024-01390-4 (2025).
24. Jin, Q. et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digital Med.* **7**, 190 https://doi.org/10.1038/s41746-024-01185-7 (2024).
25. McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H. & Wayne, D. B. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad. Med.* **86**, 706–711 https://doi.org/10.1097/ACM.0b013e318217e119 (2011).
26. Lancaster, G. A., Dodd, S. & Williamson, P. R. Design and analysis of pilot studies: recommendations for good practice. *J. Evaluat. Clin. Pract.* **10**, 307–312 (2004).
27. Thabane, L. et al. A tutorial on pilot studies: the what, why, and how. *BMC Med. Res. Methodol.* **10**, 1 https://doi.org/10.1186/1471-2288-10-1 (2010).
28. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 https://doi.org/10.3758/BF03193146 (2007).
29. Bandura, A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* **84**, 191–215 https://doi.org/10.1037/0033-295X.84.2.191 (1977).
30. Mehrabian, A. *Silent Messages: Implicit Communication of Emotions and Attitudes* (Wadsworth, 1971).
31. Walther, J. B. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Commun. Res.* **23**, 3–43 https://doi.org/10.1177/009365096023001001 (1996).
32. Newble, D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med. Educ.* **38**, 199–203 (2004).
33. Regehr, G., MacRae, H., Reznick, R. K. & Szalay, D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad. Med.* **73**, 993–997 https://doi.org/10.1097/00001888-199809000-00020 (1998).
34. Gershov, S., Mahameed, F., Raz, A. & Laufer, S. Towards accurate and interpretable competency-based assessment: enhancing clinical competency assessment through multimodal AI and anomaly detection. *npj Digit. Med.* **9**, 219 https://doi.org/10.1038/s41746-025-02299-2 (2026).
35. Boursicot, K. et al. Conducting a high-stakes OSCE in a COVID-19 environment. *MedEdPublish* **9**, 54 https://doi.org/10.15694/mep.2020.000054.1 (2020).
36. Chan, S. M., Choa, G., Kelly, J., Maru, D. & Rashid, M. A. Implementation of virtual OSCE in health professions education: a systematic review. *Med. Educ.* **57**, 833–843 https://doi.org/10.1111/medu.15089 (2023).
37. Brügge, E. et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med. Educ.* **24**, 1391 https://doi.org/10.1186/s12909-024-06399-7 (2024).
38. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
39. Kurtz, S., Silverman, J. & Draper, J. Teaching and learning communication skills in medicine. (2003).
40. Frank, J. R. et al. Competency-based medical education: theory to practice. *Med. Teach.* **32**, 638–645 https://doi.org/10.3109/0142159X.2010.501190 (2010).

Acknowledgements

The authors thank the expert evaluators for their contributions to the fidelity assessment, the graduate student standardised patient actors in Phase 2, and all 92 participating medical students from Dalian Medical University. This work was supported by the Centre for School Development and Planning under the Ministry of Education (Grant No. CSDP26FC5A425) and by the General Project of the Undergraduate Teaching Reform Research Programme of Liaoning Provincial Institutions of Higher Education (Grant No. 2025YBXM0491). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

P.W. conceptualised the study, led the clinical methodology, designed the clinical scenarios and patient personas, curated the data, conducted the investigation, and wrote the original draft. Y.H. designed the study and its three-layer information architecture, developed the cross-model prompt-engineering framework, built the web-based AI-SP platform and automated analytics pipeline, and led the formal analysis, statistical analysis, and visualisation. J.Z., Y.L., and M.J. developed and validated the clinical scenarios, recruited participants, and delivered the training programme. X.L., H.Z., D.X., H.M., and L.W. served as blinded examiners for the OSCE and the expert fidelity evaluation. Q.W. conceptualised and supervised the project, acquired funding, and reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

Funding

The authors received no specific funding for this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-03372-0.

Correspondence and requests for materials should be addressed to Qingping Wen.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.