Original Paper

Curriculum-Aligned Conversational AI to Support Pathology Self-Directed Learning in Preclinical Medical Students: Crossover Randomized Pilot Study

Kasie Roark[1,2,3], MS; Yumin Choi[2], BS; Elizabeth Hoy[2], BS; Noel Tufts[2], MPH; Robert Amevor[4], MS, PhD; Rahul Ghosal[4], PhD; Brian Keisler[5], MD; Leonardo Bonilha[2,3,6], MD, PhD; Ioulia Chatzistamou[1], MD, PhD

[1]Department of Pathology, Microbiology and Immunology, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States
[2]AI in Medicine Extracurricular Track, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States
[3]Office of Research, University of South Carolina, Columbia, SC, United States
[4]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, United States
[5]Department of Family and Preventive Medicine, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States
[6]Department of Neurology, Floyd School of Medicine, University of South Carolina, Columbia, SC, United States

Corresponding Author:
Ioulia Chatzistamou, MD, PhD
Department of Pathology, Microbiology and Immunology
Floyd School of Medicine
University of South Carolina
6439 Garners Ferry Rd, Bldg #1, Room B33
Columbia, SC, 29209
United States
Phone: 1 803 2163467
Email: chatzistamou@uscmed.sc.edu

Abstract

Background: AI is reshaping the landscape of medical education, yet the widespread adoption of general-purpose AI tools in student learning raises substantial concerns about curricular alignment, information accuracy, and the absence of pedagogical scaffolding.

Objective: This crossover pilot study evaluated a curriculum-aligned custom GPT that was developed to support AI-scaffolded self-directed learning (SDL), preparation that remains student-initiated but is supported by structured Socratic prompting, compared with fully unguided traditional preparation in second-year medical students at the University of South Carolina, Floyd School of Medicine (USC Floyd SoM).

Methods: Forty-four second-year medical students voluntarily participated in the 2 faculty-led pathology small group discussions (Cushing disease and prostate cancer), using the custom GPT for one case and conventional resources for the other. Faculty assessed SDL documentation forms and small group discussion performance using standardized rubrics.

Results: Students using the curriculum-aligned custom GPT achieved significantly higher SDL rubric scores (10.3 vs 9.4 out of 12; mean difference 0.9, 95% CI 0.5-1.3; $P<.001$; Cohen $d=0.71$) and small group discussion scores (14.9 vs 13.7 out of 16; mean difference 1.2, 95% CI 0.7-1.7; $P<.001$; Cohen $d=0.78$), with moderate-to-large effect sizes. Delayed retention (end-of-block examination items) did not differ significantly (Wilcoxon signed-rank $P=.052$). Students also reported significantly greater gains in understanding and pathophysiological reasoning. A total of 34 out of 44 (77%) reported improved reasoning clarity, and 31 out of 44 (70%) improved conceptual understanding. Assessment of 5 prespecified AI-related concern domains did not identify major concerns.

Conclusions: In this crossover pilot study, use of a curriculum-aligned custom GPT was associated with higher faculty-assessed SDL documentation and small group discussion performance. Delayed knowledge retention did not differ significantly in the primary analysis. These findings support further investigation of curriculum-aligned AI tools in larger, multisite studies.

(JMIR Med Educ 2026;12:e102908) doi: 10.2196/102908

XSL•FO
RenderX

KEYWORDS

artificial intelligence; medical education; self-directed learning; large language models; pathology; clinical competence

Introduction

The integration of AI into medical education has grown rapidly over the last few years. From adaptive question banks to AI-generated explanations and clinical decision support tools, medical schools are increasingly asking a key question: how can AI be used in ways that support the learning environment and enhance, rather than hinder, the development of clinical reasoning, self-directed learning (SDL), and professional judgment? This is not only a pedagogical issue, but it also has implications for accreditation, particularly with respect to Liaison Committee on Medical Education (LCME) Standard 6.3, which emphasizes that medical curricula should prepare students for lifelong SDL [1].

With the rise of general-purpose large language models such as ChatGPT (OpenAI), students now have on-demand access to medical information at an extraordinary scale [2,3]. However, while this access can be beneficial, it also introduces important challenges. These tools are not designed around a specific curriculum, learning objectives, or case structure [4,5]. They often provide direct answers rather than guiding students through the reasoning process, are not calibrated to the learner's level, and lack the intentional design that distinguishes active learning from passive information retrieval [6,7]. In a survey of undergraduate medical students in the United Kingdom, respondents reported that AI integration enhanced personalized learning but raised concerns about critical thinking and overreliance [8]. In practice, when students use a general-purpose AI to prepare for a case discussion, their learning is no longer aligned with the course's learning objectives or faculty expectations. In many ways, this is like studying from an unstructured, unvetted source rather than engaging with material that is intentionally designed to support learning. Ultimately, the issue is not just about accuracy; it is about alignment. Without clear boundaries around scope, depth, and pedagogical intent, general-purpose AI tools may not consistently align with the goals of competency-based medical education when used without curricular or pedagogical constraints.

Curriculum-aligned custom GPTs—user-configured versions of ChatGPT that follow author-defined instructions and draw on author-supplied reference materials—offer a more structured and pedagogically grounded way to approach this challenge. Instead of providing broad general medical information, a custom GPT can be intentionally designed around specific learning objectives, a case-specific knowledge base, and a Socratic interaction style that prompts students to explain, synthesize, and defend their reasoning rather than passively receive answers [6,7]. This design reconceptualizes AI from a passive information source into an active learning partner—serving as a "study buddy" that helps identify knowledge gaps, fosters deeper thinking and reasoning, and guides students through the clinical and pathophysiological logic embedded in the curriculum. Critically, this approach is designed to address many of the quality and alignment concerns associated with general AI use. By constraining the tool to curriculum-aligned content and clearly defined educational goals, it creates a more controlled and reliable learning environment.

Despite the growing interest in AI in medical education, there is still limited rigorous evidence supporting curriculum-aligned AI tools in SDL, particularly in relation to accreditation standards. Most existing studies relied on self-reported data, lacked appropriate control groups, or focused on general-purpose AI rather than purpose-built educational tools [9-12]. To our knowledge, no prior study has evaluated a curriculum-aligned AI chatbot using a crossover design that includes faculty assessment of both SDL documentation quality and small group discussion (SGD) performance, alongside delayed retention outcomes.

This study aims to address that gap. We present findings from a crossover pilot study (N=44) evaluating a curriculum-aligned custom GPT versus traditional resources for SDL preparation for pathology SGDs in the second-year Endocrine and Reproductive System block at the University of South Carolina, Floyd School of Medicine (USC Floyd SoM). We use the term AI-scaffolded SDL to describe the GPT condition: preparation that is student-initiated but guided by structured Socratic prompting aligned with the curriculum. The traditional condition reflects fully unguided, self-directed preparation using conventional resources at the student's discretion. We hypothesized that use of the custom GPT would lead to higher-quality SDL documentation, stronger SGD performance, and improved knowledge retention compared with traditional preparation methods, without raising major concerns in 5 prespecified AI-related domains.

Methods

Study Design and Setting

We conducted a randomized crossover pilot study during the spring 2026 semester in the Endocrine and Reproductive System block of the integrated second-year curriculum at USC Floyd SoM. This block includes 2 faculty-led pathology SGDs: one focused on Cushing disease and one on prostate cancer. Both cases required students to complete SDL preparation using a standardized SDL form consistent with LCME Standard 6.3 SDL competency domains. The complete study design is summarized in Multimedia Appendix 1. This study is reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Multimedia Appendix 2).

Participants

Second-year medical students were invited to participate voluntarily. Forty-four students enrolled and were randomly assigned to 2 groups of 22 using a computer-generated random sequence; allocation was concealed from faculty evaluators until group assignments were finalized. Group 1 used the custom

XSL•FO
RenderX

GPT for the Cushing disease preparation and traditional resources for prostate cancer; Group 2 completed the reverse sequence. Each student served as their own control to enable pairwise testing. Participation was voluntary and did not affect grades. The 2 SGD sessions were separated by approximately 13 days. Students were instructed not to discuss their preparation method or custom GPT access with peers during the study period, consistent with the academic honor code; they were also informed that all enrolled students would receive access to the custom GPT tool at the conclusion of the study, removing any incentive to share access prematurely. As a pilot study, the sample size was determined by voluntary enrollment within the block rather than a formal a priori power calculation; the observed effect sizes (Cohen d=0.71-0.78) provide a basis for powering a future, larger trial. The study was granted an exemption by the University of South Carolina Institutional Review Board (IRB; see the "Ethical Considerations" section). Because FERPA (Family Educational Rights and Privacy Act) regulations precluded access to student academic records from prior blocks, randomization was unstratified. End-of-block final grades were comparable between groups (Group 1: mean 87.1%, SD 6.4%; Group 2: mean 86.8%, SD 6.7%); age, sex, and prior AI experience were not collected due to these access restrictions.

Custom GPT

We developed 2 curriculum-aligned custom GPTs, one for each case, using the GPTs feature of ChatGPT. The system prompts were shaped with input from faculty to ensure alignment with the learning goals, while still allowing flexibility in how students engaged with the material. Faculty also reviewed the case flow and embedded resources to ensure alignment with the depth and reasoning expectations of the pathology curriculum. They also refined the Socratic prompts to encourage clinical and pathophysiological reasoning rather than simple answer retrieval. The goal was to create a structured but interactive tool that could guide students through the cases in a way that supports reasoning rather than simply providing answers. The design focused on several key elements: (1) keeping the interaction tightly aligned with each case's learning objectives; (2) using a Socratic approach that required students to explain and justify their clinical and pathophysiological reasoning; (3) guiding students through the case stages, aligned with the structure of the SDL forms; (4) incorporating case-specific clinical vignettes, imaging descriptions, and laboratory data provided in the study guide students received before the discussion sessions; and (5) directing students to open-access, peer-reviewed resources. The tools were reviewed by pathology faculty with content expertise and piloted and refined with input from students participating in the AI in Medicine Extracurricular Track at USC Floyd SoM to ensure usability and alignment with the learner experience and needs. Students assigned to the GPT condition were instructed to use the custom GPT as part of their SDL preparation. Objective usage metrics were not available to independently verify the extent or duration of engagement. Students accessed each custom GPT through their institutional USC ChatGPT account via the link provided in an email at no cost, as part of required SDL preparation for each case; no separate account creation or payment was required. Students were not given a specific time or interaction quota and used the tool ad libitum before each SGD. No automated prompts or reminders were sent to encourage use beyond standard course scheduling communications. Students did not receive formal orientation or training on use of the custom GPT before the study, although the discussion identifies a brief orientation as a potential improvement for future implementations. The content and configuration of both custom GPTs were frozen for the duration of the spring 2026 data collection period, with no functional or content changes made once data collection began.

Traditional Resources

Students in the traditional condition completed SDL preparation using conventional resources: course lecture notes, pathology textbooks, peer-reviewed literature via PubMed, and other clinical resources at their discretion. Access to the custom GPT tool was not provided to students in this condition, as the tool link was distributed exclusively to students in the GPT condition.

Assessment Instruments

All students completed a standardized SDL form documenting learning goals, identified knowledge gaps, information sources, clinical reasoning synthesis, and self-assessment. SDL forms were scored by 1 of the 3 pathology small group facilitators using a grading rubric (range 6-12), assessing quality of gap identification, information synthesis, pathophysiological reasoning, and resource use. SGD performance was assessed by the faculty pathology facilitator using a grading rubric scored out of 16, evaluating diagnostic depth, clinical-pathological integration, communication, and professionalism. Both scoring rubrics are provided in Multimedia Appendices 3 and 4. Each SDL form and SGD performance was scored by a single facilitator; interrater reliability was not assessed. All faculty evaluators (IC and 2 additional pathology faculty facilitators) were blinded to study participation: they evaluated their entire small group cohort under routine conditions and were unaware of which students had enrolled in the study or which preparation method any student had used. The faculty facilitators rotated across all 9 small groups on a fixed schedule established at the start of the academic year by the pathology content leader; this schedule was followed for all pathology small groups throughout the year. The group assignment process was computer-generated and was not influenced by the principal investigator (IC), who was aware of the overall study enrollment but was blinded to each student's preparation condition for the case being assessed on that date. Every participant was assessed by a different facilitator for each of the 2 cases, and both study conditions were represented among each facilitator's groups in both sessions. Short-term retention was assessed using student performance on 2 in-house multiple-choice questions per quiz administered 2-4 days after each SGD, while delayed knowledge retention was assessed using 2 multiple-choice questions per case from the end-of-block examination, a customized National Board of Medical Examiners (NBME) examination (NBME Customized Assessment Services). The SDL documentation form was completed and submitted immediately before each SGD as a required deliverable, capturing the student's self-identified learning needs, resources used, information synthesis, and self-assessment of gaps; it preceded and directly

XSL•FO
RenderX

informed the discussion. Representative examples of student interactions with the custom GPT tool for each case, provided by consenting study participants, are available in Multimedia Appendices 5 and 6. Students completed surveys immediately after preparation, via a REDCap (Vanderbilt University) questionnaire accessed through a link distributed by email by KR, including 5-point Likert items on pre- and postpreparation understanding, confidence, knowledge impact, mechanism understanding improvement, and resource helpfulness, along with open-text reflections. Potential AI-related concerns were assessed across 5 prespecified domains informed by previous reports on AI use in educational settings: academic integrity, time burden, overreliance, content accuracy, and fairness across performance tiers. Other potential concerns, including hallucinations, data privacy, and inappropriate clinical guidance, were not formally evaluated.

Statistical Analysis

We analyzed the quantitative data using paired *t* tests for within-student (crossover) comparisons and Mann-Whitney *U* tests for between-group survey responses (GPT condition: n=44 survey observations; traditional condition: n=42 survey observations). For the delayed retention score (range 0-2, based on 2 binary items per case), the Wilcoxon signed-rank test was used as the primary analysis because of the discrete nature of the outcome. For the SDL rubric and SGD scores, paired *t* tests were used because these outcomes were treated as approximately continuous, given the number of possible score values and the within-subject crossover design. We acknowledge that treating rubric-based scores as continuous represents an analytical assumption and therefore a potential limitation [13]. Chi-squared tests were used for categorical variables, and effect sizes were reported using Cohen *d* for paired samples, based on within-subject differences, with 95% CIs for the mean differences (*t* distribution) and for Cohen *d* (noncentral *t* distribution). All tests were 2-tailed, with statistical significance set at α=.05. Faculty-assessed SDL and SGD rubric scores were complete for all 44 enrolled students (N=44 paired observations). Two students did not complete the postpreparation survey in the traditional condition, yielding n=42 survey observations for that condition; these were treated as missing data under a complete-case approach, and no imputation was performed.

We also conducted exploratory post hoc analyses by grouping students into mutually exclusive tertiles based on end-of-block academic performance (low: final grade <84.5%, n=15; middle: 84.5%-90.3%, n=15; high: >90.3%, n=14) to explore whether the impact of the GPT differed by academic performance tier. Because final block grade was calculated after the intervention period, it is a concurrent rather than preintervention measure. These analyses were not prespecified, were not powered for subgroup comparisons, and should be considered hypothesis-generating. Sequence effects were examined by comparing within-student GPT-minus-traditional differences between the 2 crossover groups using a Mann-Whitney *U* test.

Ethical Considerations

This study was reviewed by the University of South Carolina IRB and determined to be exempt from IRB review under Category 1 (research conducted in established or commonly accepted educational settings) on August 12, 2025 (study ID: STUDY00000007). As an exempt protocol, formal documented informed consent was not required by the IRB; all participants were informed of the study's purpose, voluntary nature, and that participation would not affect grades before enrolling.

This pilot study was not registered in a public trial registry. As a minimal-risk, exempt educational pilot study conducted within a single course block, formal registration was not pursued; the authors note this as a limitation and recommend prospective registration for future, larger-scale studies of this intervention.

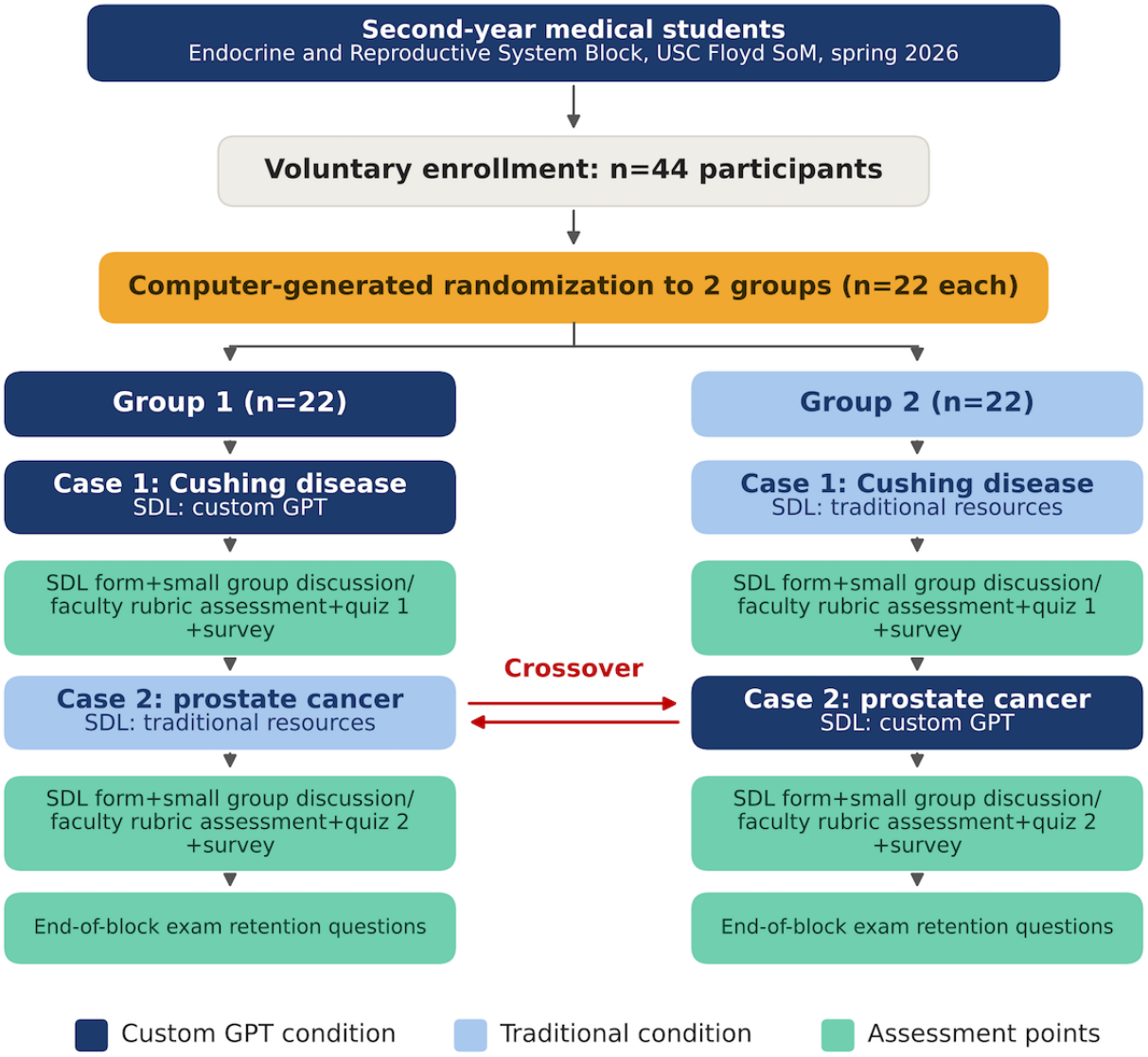
Results

Overview

All 44 enrolled students were randomly assigned (22 per group), completed both cases, and were included in all analyses of faculty-assessed outcomes and delayed retention; 2 students in the traditional condition did not complete the postpreparation survey (Figure 1).

XSL•FO
RenderX

Figure 1. Study design: crossover randomized pilot study. SDL: self-directed learning; USC Floyd SoM: University of South Carolina, Floyd School of Medicine.



SDL Documentation and SGD Performance

Students who used the custom GPT performed significantly better on both faculty-assessed outcomes, with moderate-to-large effect sizes (Table 1; Figure 2). SDL rubric scores were higher in the GPT group (mean 10.3, SD 1.4 vs mean 9.4, SD 1.5 out of 12; mean difference 0.9, 95% CI 0.5-1.3; P<.001; Cohen d=0.71, 95% CI 0.38-1.04). Small group discussion performance showed a similar pattern (mean 14.9, SD 0.8 vs mean 13.7, SD 1.5 out of 16; mean difference 1.2, 95% CI 0.7-1.7; P<.001; Cohen d=0.78, 95% CI 0.44-1.11). **Scores were higher in the GPT condition for both cases.** A significant sequence effect was observed for SGD scores (P=.007), with a larger GPT-associated difference for the prostate cancer case (+1.8 points) than for the Cushing disease case (+0.6 points). No significant sequence effects were identified for SDL scores (P=.09) or delayed retention (P=.98).

XSL•FO
RenderX

Table 1. Summary of quantitative outcomes: GPT versus traditional condition.

Outcome measure	GPT, mean (SD)	Traditional, mean (SD)	Δ (95% CI)[a]	*P* value	Cohen *d* (95% CI)[b]
SDL[c] rubric score (out of 12)[d]	10.3 (1.4)	9.4 (1.5)	+0.9 (0.5-1.3)	<.001	0.71 (0.38-1.04)
SGD[e] score (out of 16)[d]	14.9 (0.8)	13.7 (1.5)	+1.2 (0.7-1.7)	<.001	0.78 (0.44-1.11)
Delayed retention score (out of 2)[f]	1.7	1.5	+0.2	.052	—[g]
Postpreparation understanding (1-5)[h]	4.3	3.9	+0.4	<.001	—
Mechanism understanding improvement (1-5)[h]	3.6 (1.0)	3.1 (0.8)	+0.5	.003	—
In-case quiz score (out of 2)	1.8	1.9	−0.1	.32	—

[a]Mean difference (GPT minus traditional); 95% CIs are reported for the paired rubric outcomes.

[b]Calculated for the paired rubric outcomes only.

[c]SDL: self-directed learning.

[d]Within-student comparison (n=44 pairs); paired *t* test.

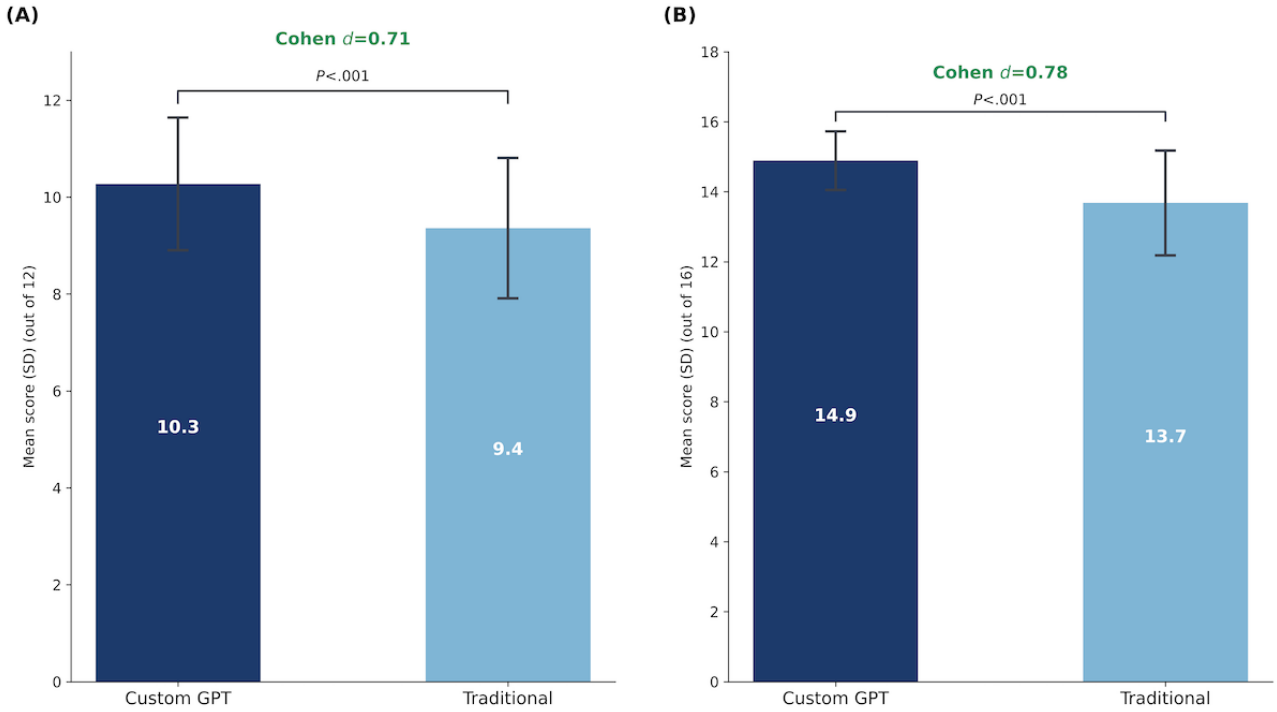
[e]SGD: small group discussion.

[f]Within-student comparison (n=44 pairs); Wilcoxon signed-rank test.

[g]Not applicable.

[h]Between-condition comparison of survey responses (GPT condition: n=44; traditional condition: n=42); Mann-Whitney *U* test.

Figure 2. Faculty-assessed performance scores: custom GPT versus traditional condition. (A) Mean (SD) self-directed learning (SDL) documentation rubric scores (out of 12). (B) Mean (SD) small group discussion (SGD) rubric scores (out of 16). Both comparisons used within-subject crossover data (n=44 paired observations; paired *t* test). Error bars represent ±1 SD. The custom GPT condition produced significantly higher scores on both measures (SDL: 10.3 vs 9.4, *P*<.001, Cohen *d*=0.71; SGD: 14.9 vs 13.7, *P*<.001, Cohen *d*=0.78), with moderate-to-large effect sizes.



Delayed Knowledge Retention

No significant differences were observed between groups in quiz performance 2-4 days after the SGD (1.8 vs 1.9 out of 2; *P*=.32). The primary analysis of delayed knowledge retention did not reach the conventional significance threshold (Wilcoxon signed-rank *P*=.052; Figure 3B). Mean scores were 1.7 versus 1.5 out of 2. Among the 44 paired observations, scores were identical for 26 out of 44 (59%) students, higher under the GPT condition for 13 out of 44 (30%), and lower for 5 out of 44 (11%). In an exploratory post hoc analysis, we divided students into 3 groups based on final block performance (low <84.5%, n=15; middle 84.5%-90.3%, n=15; high >90.3%, n=14). The difference was +0.4 in the lower-performing group, +0.1 in the middle group, and +0.1 in the higher-performing group (Figure 4A). These subgroup findings are exploratory and based on small numbers.

XSL•FO
RenderX

Figure 3. Mechanism understanding improvement and delayed knowledge retention. (A) Distribution of self-reported mechanism understanding improvement. A total of 28 out of 44 (64%) of GPT users reported significant or extreme improvement in pathophysiological reasoning versus 12 out of 42 (29%) traditional users (Mann-Whitney U, *P*=.003). (B) Delayed retention (end-of-block examination items), pooled across both cases. Overall, 84% versus 75% of answers were correct (1.7 vs 1.5 out of 2); the difference was not statistically significant (Wilcoxon signed-rank *P*=.052).

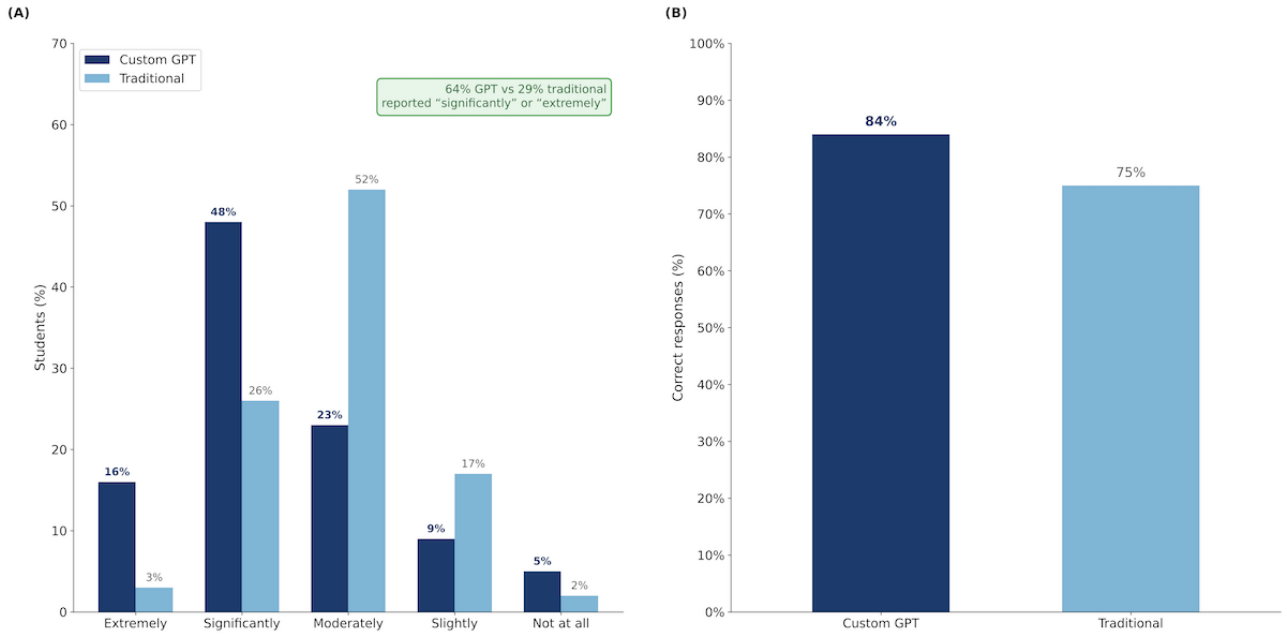
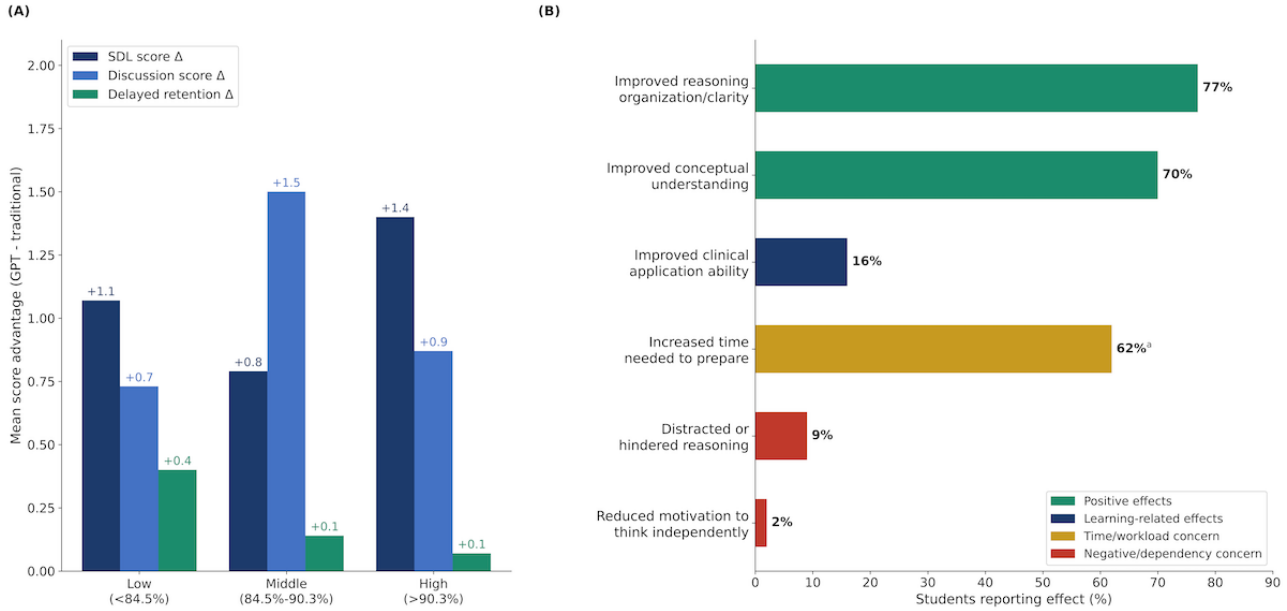


Figure 4. GPT benefits by academic performance tier and student-reported learning effects. (A) GPT score advantage (GPT minus traditional) on self-directed learning (SDL) documentation, small group discussion (SGD), and delayed retention, by post hoc academic performance tertile (n≈14-15 per tier). Low performers (<84.5%) showed the largest delayed retention difference (Δ=+0.4), while middle and high performers showed larger discussion score differences. These subgroup comparisons are exploratory. (B) Student-reported effects of the custom GPT (multiple-select, n=44). A total of 34 out of 44 (77%) reported improved reasoning clarity; 31 out of 44 (70%) improved conceptual understanding. The most frequently cited concern was a perceived increase in the time needed to prepare (16 out of 26 commenting students, 62%). [a]: Proportion of the 26 students who provided optional open-text comments.



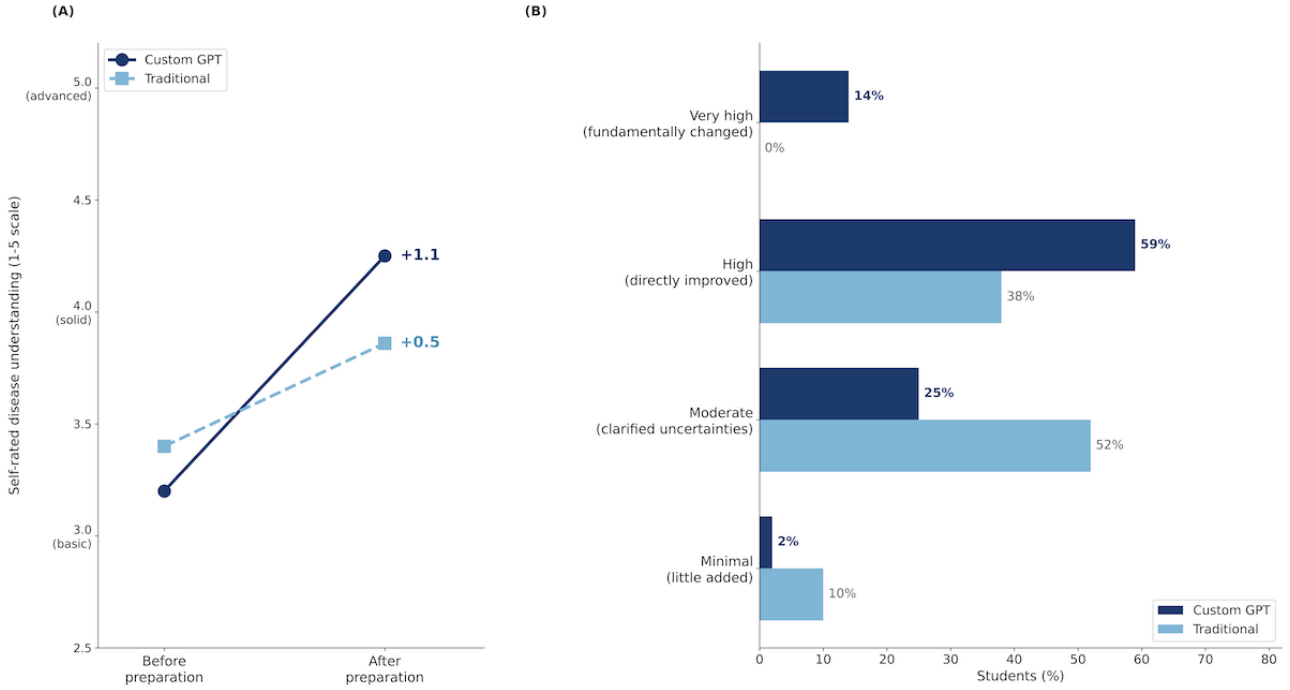
Self-Reported Learning Outcomes

Students using the GPT reported larger gains in their understanding of the disease from pre- to postpreparation (+1.1 vs +0.5; *P*<.001). While both groups improved significantly within-group (*P*<.001), the GPT group ended at a higher overall level of understanding (4.3 vs 3.9; Figure 5A). Perceived impact on knowledge also differed between groups (χ^2_1=9.1; *P*=.003; Figure 5B): 32 out of 44 (73%) GPT users reported a high or very high impact, compared with 16 out of 42 (38%) students using traditional resources. No students in the traditional group reported a "very high" impact (ie, a fundamental change in understanding), whereas 6 out of 44 (14%) GPT users did. A similar pattern was seen for mechanism understanding: 28 out of 44 (64%) GPT users reported significant or extreme

XSL•FO
RenderX

improvement, compared with 12 out of 42 (29%) in the traditional group (P=.003; Figure 3A).

Figure 5. Self-reported learning gains: pre-to-post understanding and knowledge impact. (A) Slope graph of self-rated disease understanding (1=none to 5=advanced) before and after preparation. GPT users improved by +1.1 points versus +0.5, finishing at a significantly higher postpreparation level (4.3 vs 3.9; Mann-Whitney U, P<.001). Within-group changes were significant (Wilcoxon signed-rank, P<.001). (B) Knowledge impact distribution. GPT users were significantly more likely to report high or very high impact (32 out of 44, 73% vs 16 out of 42, 38%; 4-category distribution: χ^2_3=13.8; P=.003). No traditional-condition student reported very high knowledge impact.



Patterns Consistent With LCME Standard 6.3 SDL Competency Domains

When we examined SDL form content in relation to the competency domains described in LCME Standard 6.3, differences emerged in how students approached SDL (Table 2). Students using the custom GPT were more likely to focus on disease mechanisms and pathophysiology as their main learning need (33/44, 75% vs 20/42, 48%), while students using traditional resources more often emphasized treatment and management algorithms (12/42, 29% vs 4/44, 9%). AI chatbot users also reported identifying knowledge gaps they were not previously aware of—a metacognitive skill closely aligned with the competency domains described in LCME Standard 6.3.

These differences were reflected in the SDL rubric scores as well, particularly in goal setting and the quality of learning documentation, which showed a moderate-to-large effect size (Cohen d=0.71) and were consistent across both cases.

Table 2. Type and depth of learning gaps identified: GPT versus traditional condition.

What students said they needed to learn[a]	GPT group (n=44), n (%)	Traditional group (n=42), n (%)	Difference, percentage points
Disease mechanism/pathophysiology	33 (75)	20 (48)	+27
Diagnostic criteria/tests	15 (34)	8 (19)	+15
Treatment/management	4 (9)	12 (29)	−20
Prognosis/outcomes	8 (18)	11 (26)	−8

[a]Students could select more than one category.

Student Satisfaction and Concerns

Overall satisfaction with the GPT was high (Figure 4B). Most students reported meaningful benefits, with 34 out of 44 (77%) noting improved clarity in their reasoning and 31 out of 44 (70%) reporting better conceptual understanding. Only 7 out of 44 (16%) students reported improved clinical application ability. Additionally, student reflections highlighted the value of the coaching-style interaction. One student shared that “this custom AI engine is extremely impressive; having it in a coaching mindset was phenomenal in terms of making sure I truly grasped the material,” while another noted that “the customized ChatGPT was extremely helpful in mastering lecture content and breaking down complex pathophysiology into manageable chunks.” The most common worry was the time required to use the tool, mentioned by 16 out of 26 (62%) GPT-condition students who provided optional comments. Students also suggested a few practical improvements, including fewer simultaneous questions

XSL•FO
RenderX

per exchange, clearer signals for when they had reached the end of a task, and occasional support for recalling foundational facts when needed. Findings across the 5 prespecified AI-related concern domains are summarized in Table 3.

Table 3. Assessment of prespecified AI-related concerns.

Concern domain[a]	What the data show	Interpretation
Academic integrity	Only 1 out of 44 (2%) students felt the tool reduced their independent thinking. Faculty scores were higher in the GPT group, suggesting students still did their own analytical work	No major concern identified in assessed measures
Time and effort	16 out of 26 (62%) students who provided comments said the tool took more time than usual. However, this extra effort was accompanied by higher faculty-assessed scores	Student-reported time burden identified
Overreliance on the tool	Students who used the GPT for one case performed similarly independently on the subsequent case; no evidence of increased reliance on the tool	No signal identified; assessment limited
Accuracy of content	2 out of 44 (5%) students expressed trust concerns; however, because the tool was built from course-specific material, it may be less prone to the inaccuracies seen in general-purpose AI tools	Not formally evaluated
Fairness across students	In an exploratory post hoc analysis, the largest retention difference was observed in students with lower academic performance (Δ=+0.4)	Exploratory finding; requires confirmation

[a]Potential AI-related concerns were reviewed across 5 prespecified domains using student surveys, faculty rubric scores, and written student reflections.

Discussion

Principal Findings

This crossover pilot study suggests that a curriculum-aligned custom GPT can improve second-year medical students' SDL preparation and small group discussion performance. The improvements observed across 2 different cases, in both faculty-assessed and student-reported outcomes, suggest that thoughtfully designed AI tools may have meaningful educational value in undergraduate medical education.

It is also important to distinguish between general-purpose AI tools and curriculum-aligned custom GPTs. General-purpose large language models are not specifically designed around individual medical curricula or the development of SDL competencies. While they may provide accurate information, that information is not necessarily aligned with course learning objectives, may not address specific reasoning gaps, and often lacks the structured, iterative engagement that supports deeper learning [4,5]. In contrast, the curriculum-aligned custom GPTs in this study were intentionally built around specific cases and learning objectives. They asked students to explain and justify their reasoning while directing them toward relevant, curriculum-aligned resources. The improvements observed, particularly in pathophysiological reasoning and SDL documentation quality, may reflect this more intentional and guided learning approach. These findings are consistent with recent controlled studies showing that structured, curriculum-aligned AI tools improve SDL outcomes compared with unassisted preparation [11,12] and add faculty-assessed objective measures that were absent from prior work.

The patterns observed in SDL content are especially interesting in the context of LCME Standard 6.3. Students using the GPT were more likely to identify disease mechanisms and pathophysiology as learning needs, a pattern consistent with the intervention's intended emphasis on mechanistic reasoning. Students also reported identifying previously unrecognized knowledge gaps, suggesting potential support for metacognitive aspects of SDL. These findings should be interpreted as patterns consistent with SDL-related competencies rather than evidence that LCME Standard 6.3 competencies were achieved. The qualitative finding that students were able to identify knowledge gaps they were not previously aware of is particularly noteworthy, as this level of metacognitive awareness can be difficult to achieve with unguided SDL and is essential for lifelong learning [14]. By contrast, only 7 out of 44 (16%) students reported improved clinical application ability, which may reflect the preclerkship context, in which students have limited clinical experience against which to judge such changes; this could be examined in clerkship-stage learners. In the post hoc analysis, the largest retention difference was observed in lower-performing students (Δ=+0.4), and GPT users in the lowest academic tier reached SGD scores of 14.7, compared with 13.8 among higher-performing students using traditional preparation. While these subgroup findings are exploratory and based on relatively small numbers (n≈14-15 per group), they suggest that the GPT may offer helpful structure for students who struggle more with self-directed learning. This potential equity-related benefit is worth exploring further [15].

The primary analysis of delayed knowledge retention did not reach statistical significance (Wilcoxon signed-rank P=.052). Interpretation was limited by the outcome measure, which was based on only 2 end-of-block examination items per case. This result is therefore inconclusive and warrants confirmation in larger studies using more comprehensive, multi-item measures of knowledge retention.

A main limitation was related to the time required to use the tool. Students perceived preparation with the GPT as taking longer than more traditional, reading-based approaches, as reported by 16 out of 26 (62%) students who provided comments. Another factor to consider is timing. The study was conducted in the middle of the spring semester, at a point when students had already established their approaches to preparing

XSL•FO
RenderX

for pathology small group discussions. Adopting a new method at that stage may have added to the perceived time burden, as students were not only learning the content but also adjusting to a different way of studying. Some targeted design changes, such as asking fewer questions at a time, providing clearer end points, and offering a brief orientation, could help reduce the time burden without compromising the educational value. Part of the observed educational benefit may reflect increased time-on-task rather than the GPT's design alone, as 16 out of 26 (62%) GPT students who provided comments reported spending more time than usual on preparation; objective time-on-task was not measured. Future studies should measure preparation time and engagement metrics to disentangle these effects.

Several limitations should be considered when interpreting these findings. First, this pilot study was conducted at a single institution using 2 pathology cases from one preclinical block, which limits the generalizability of the results. In addition, the study was not powered by a formal a priori sample-size calculation; the enrolled sample (n=44) reflected voluntary participation within the course. Although moderate-to-large effect sizes were observed (Cohen *d*=0.71-0.78), these should not be interpreted as evidence that the study was adequately powered, particularly given the width of their 95% CIs. The observed effect sizes may, however, provide a useful basis for sample-size calculations in future multisite studies. Subgroup (tier-based) analyses were exploratory and should be interpreted cautiously. The delayed knowledge retention outcome was also limited by using only 2 end-of-block examination items per case, reducing the precision and sensitivity of the measure to detect modest treatment effects. In the traditional preparation condition, 16 out of 42 (38%) survey respondents reported using general-purpose AI tools. Consequently, the comparison was not between AI and no AI, but rather between a curriculum-aligned custom GPT and usual self-directed preparation, during which some students independently used general AI tools. The direction and magnitude of any resulting bias cannot be determined. It is important to note that this study did not include a condition in which students used general-purpose AI tools without curriculum alignment. Accordingly, the findings do not permit conclusions about whether a curriculum-aligned custom GPT is more effective than unrestricted general-purpose AI use. The advantage demonstrated here is relative to unguided traditional preparation, a comparison that also involves some degree of general AI use in the control condition. Future studies using 3-arm designs (traditional preparation, general-purpose AI, and curriculum-aligned AI) would be needed to test the incremental benefit of curriculum alignment directly. Additionally, objective usage metrics for the custom GPT (eg, number of logins, message counts, or time-on-task) were not collected. Engagement with the custom GPT was inferred from completion of the required SDL form rather than measured directly, limiting our ability to distinguish the effects of the tool's design from the effects of increased time-on-task. Baseline participant characteristics from prior coursework were unavailable because of FERPA restrictions. Although end-of-block performance was comparable between the randomized groups (Group 1: mean 87.1%, SD 6.4%; Group 2: mean 86.8%, SD 6.7%), prior academic performance and prior AI experience could not be fully characterized. However, all students had institutional access to the ChatGPT platform and AI training resources provided by the University of South Carolina. A significant sequence effect was observed for SGD performance, reflecting a larger GPT-associated benefit for the prostate cancer case than for the Cushing disease case. Because case and intervention order are linked in the crossover design, case-specific and order effects cannot be fully distinguished. Notably, sequence effects were not significant for SDL documentation or delayed retention. Each SDL form and SGD performance was scored by a single facilitator, and interrater reliability was not assessed, although facilitators were blinded to preparation condition. Finally, participants could not be blinded to their assigned preparation method because of the nature of the intervention. In addition, multiple outcomes were evaluated without formal adjustment for multiplicity, increasing the possibility of a type I error. Accordingly, the findings should be interpreted in the context of the overall pattern of results rather than any single statistical test.

Conclusions

In this single-institution crossover pilot study of 44 second-year medical students, access to a curriculum-aligned custom GPT was associated with significantly higher faculty-assessed SDL documentation and small group discussion performance compared with unguided traditional preparation (Cohen *d*=0.71 and 0.78, respectively). These findings were consistent across both pathology cases. The primary analysis of delayed knowledge retention did not reach statistical significance. Assessment across 5 prespecified AI-related concern domains did not identify major concerns (Table 3). The main drawback was perceived increased preparation time, reported by 16 out of 26 students who provided comments (62%). These findings support further investigation of curriculum-aligned AI tools in undergraduate medical education in larger, multisite, prospectively registered studies.

Acknowledgments

The authors thank the second-year medical students in the class of 2028 at the University of South Carolina, Floyd School of Medicine (USC Floyd SoM) for their participation in this study. We also thank Drs Christine Wheeler and Avery Dingle for validating the tools and for their blinded assessment of self-directed learning documentation and small group discussion performance. We thank Dr James Catroppo for his contributions to the integrated curriculum of USC Floyd SoM, his role in organizing the pathology small group sessions, and for developing the student and faculty study guides.

XSL•FO
RenderX

Funding

This project was supported by an internal Medical Education Research Innovation (MEDRI) grant from the University of South Carolina School of Medicine, Greenville. The funder had no role in study design, data collection, analysis, interpretation, or the decision to submit this work for publication. RA and RG were supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM155896 through the Center for Clinical & Translational Research at USC (ClinTrUSC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

Datasets are available from the corresponding author upon reasonable request. The study protocol and the system prompts/configuration of the custom GPTs are likewise available from the corresponding author upon reasonable request.

Generative AI Use

ChatGPT (OpenAI) was used to assist with grammar and language refinement and figure preparation. All content was reviewed and verified by the authors.

Authors' Contributions

Conceptualization: IC, BK, LB
Data collection: IC, YC, EH, NT, BK
Formal analysis: IC, KR, YC, EH, NT, RA, RG
Methodology: IC, KR, BK, LB
Writing – original draft: IC
Writing – review & editing: IC, KR, YC, EH, NT, RA, RG, BK, LB

Conflicts of Interest

The study authors, including IC, developed, validated, and tested the custom GPTs evaluated in this study; no commercial or proprietary interest is held in the OpenAI custom GPT platform or in the specific tools beyond their role as study investigators and developers.

Multimedia Appendix 1

Study design summary.
[PDF File (Adobe PDF File), 32 KB-Multimedia Appendix 1]

Multimedia Appendix 2

CONSORT-EHEALTH checklist.
[PDF File (Adobe PDF File), 165 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Self-directed learning documentation rubric.
[PDF File (Adobe PDF File), 55 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Small group discussion rubric.
[PDF File (Adobe PDF File), 37 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Example of interaction with custom GPT 1.
[PDF File (Adobe PDF File), 5 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Example of interaction with custom GPT 2.
[PDF File (Adobe PDF File), 7 KB-Multimedia Appendix 6]

XSL•FO
RenderX

References

1. Liaison Committee on Medical Education. Functions and structure of a medical school: standards for accreditation of medical education programs leading to the MD degree. Liaison Committee on Medical Education. 2024. URL: https://lcme.org/publications/ [accessed 2026-10-31]
2. Triola MM, Rodman A. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies. Acad Med. 2025;100(4):413-418. [FREE Full text] [doi: 10.1097/ACM.0000000000005963] [Medline: 39705530]
3. Rodman A, Mark NM, Artino AR, Lessing JN. Using generative artificial intelligence in medical education. Acad Med. 2025;100(2):250. [FREE Full text] [doi: 10.1097/ACM.0000000000005937] [Medline: 39622003]
4. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of chatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33. [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Hwang GJ, Chang CY. A review of opportunities and challenges of chatbots in education. Interact Learn Environ. 2021;31(7):4099-4112. [FREE Full text] [doi: 10.1080/10494820.2021.1952615]
7. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. [FREE Full text] [doi: 10.1016/j.lindif.2023.102274]
8. Sunmboye K, Strafford H, Noorestani S, Wilison-Pirie M. Exploring the influence of artificial intelligence integration on personalized learning: a cross-sectional study of undergraduate medical students in the United Kingdom. BMC Med Educ. 2025;25(1):570. [FREE Full text] [doi: 10.1186/s12909-025-07084-z] [Medline: 40247306]
9. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. Med Sci Educ. 2025;35(1):555-567. [doi: 10.1007/s40670-024-02206-6] [Medline: 40144083]
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
11. Shalong W, Yi Z, Bin Z, Ganglei L, Jinyu Z, Yanwen Z, et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach. 2025;47(7):1126-1133. [doi: 10.1080/0142159X.2024.2413023] [Medline: 39425996]
12. Kalam KA, Masoud FD, Muntaser A, Ranga R, Geng X, Goyal M. ChatGPT as a learning tool for medical students: results from a randomized controlled trial. Cureus. 2025;17(6):e85767. [doi: 10.7759/cureus.85767] [Medline: 40656250]
13. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013;4:863. [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
14. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Into Practice. 2002;41(2):64-70. [FREE Full text] [doi: 10.1207/s15430421tip4102_2]
15. Ricotta DN, Richards JB, Atkins KM, Hayes MM, McOwen K, Soffler MI, et al. (on behalf of Millennium Conference 2019 writing group). Self-directed learning in medical education: training for a lifetime of discovery. Teach Learn Med. 2022;34(5):530-540. [doi: 10.1080/10401334.2021.1938074] [Medline: 34279167]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
FERPA: Family Educational Rights and Privacy Act
IRB: Institutional Review Board
LCME: Liaison Committee on Medical Education
NBME: National Board of Medical Examiners
SDL: self-directed learning
SGD: small group discussion
USC Floyd SoM: University of South Carolina, Floyd School of Medicine

XSL•FO
RenderX

Edited by P Kanzow; submitted 29.May.2026; peer-reviewed by J Wilkinson, L Corral-Gudino; comments to author 03.Aug.2026; revised version received 09.Oct.2026; accepted 13.Oct.2026; published 06.Nov.2026

Please cite as:
Roark K, Choi Y, Hoy E, Tufts N, Amevor R, Ghosal R, Keisler B, Bonilha L, Chatzistamou I
Curriculum-Aligned Conversational AI to Support Pathology Self-Directed Learning in Preclinical Medical Students: Crossover Randomized Pilot Study
JMIR Med Educ 2026;12:e102908
URL: https://mededu.jmir.org/2026/1/e102908
doi: 10.2196/102908
PMID:

©Kasie Roark, Yumin Choi, Elizabeth Hoy, Noel Tufts, Robert Amevor, Rahul Ghosal, Brian Keisler, Leonardo Bonilha, Ioulia Chatzistamou. Originally published in JMIR Medical Education (https://mededu.jmir.org), 06.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX