npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02022-1

A generative AI teaching assistant for personalized learning in medical education

Thomas Thesen[1,2] ✉ & Soo Hwan Park[3]

Medical education faces a scalability crisis, where rising class sizes strain individualized instruction, while students increasingly adopt unvalidated Generative AI (GenAI) tools for individualized learning support. This study investigated how medical students integrate constrained GenAI systems into their self-directed learning practices using Retrieval-Augmented Generation (RAG), which limits large language model responses to instructor-curated materials, thereby reducing hallucinations while maintaining pedagogical utility. We deployed a RAG-based teaching assistant in a medical school basic science course across two consecutive cohorts, examining usage patterns, conversation content, and student feedback to understand adoption and learning behaviors. Students demonstrated strategic, context-dependent usage, with engagement intensifying during high-stakes assessment periods and substantial after-hours utilization. Users primarily sought clarification on foundational concepts and valued the system's continuous availability and source-grounded responses. However, knowledge-base constraints that ensured accuracy also limited broader inquiries, creating tension between reliability and comprehensiveness that shaped how students incorporated the tool into their study routines. These findings provide empirical evidence of how medical students navigate constrained AI tools for self-directed learning, informing institutional strategies for integrating these technologies into pedagogical frameworks.

Generative AI has the potential to provide personalized learning opportunities in medical education. Rising class sizes in medical schools, coupled with increasingly diverse student learning needs, create environments where individualized attention becomes increasingly difficult to provide. At the same time, didactic courses struggle to accommodate the varied pacing and comprehension levels of students, particularly in content-heavy subjects like organ-system=based preclerkship courses. This issue is further complicated by the limited availability of faculty for individual clarification and support outside standard business hours, a time when many students study[1,2].

AI-driven platforms have the potential to tailor instruction to individual students' weaknesses and provide immediate custom feedback. The educational benefit of personalized instruction is well-established. Studies show that reducing student-to-teacher ratios significantly improves learning outcomes[3,4], and one-on-one tutoring can substantially enhance student performance[5]. While peer-tutoring programs address some of these concerns, providing each medical student with individual human tutoring across all medical school courses remains impractical at scale. And today's medical students, as digital natives, increasingly turn to online resources for immediate, interactive learning support[6]. In fact, with the wide commercialization of generative artificial intelligence (GenAI), surveys indicate that approximately half of medical students use large language model (LLM) chatbots, such as ChatGPT, during their studies, with many engaging weekly or more for learning and writing assistance[7]. Notably, students often prefer asking LLMs questions over consulting textbooks or instructors, driven by the appeal of instant, personalized responses. This represents a fundamental shift toward digital, on-demand learning tools in medical education[8,9].

However, this technological adoption comes with specific challenges. LLMs, while powerful and eager in generating helpful explanations, are prone to producing incorrect or fabricated information, so-called "hallucinations"[10]. This poses major challenges for their practical application in medical training, where content accuracy and alignment with curriculum standards and current best-practices in medicine are critically important. Retrieval-augmented generation (RAG) offers a promising solution to address these accuracy concerns. RAG is a hybrid approach that combines the generative capabilities of LLMs with a retrieval mechanism

[1]Department of Medical Education, Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [2]Department of Computer Science, Dartmouth College, Hanover, NH, USA. [3]Department of Medicine, California Pacific Medical Center, Stanford School of Medicine, Palo Alto, CA, USA. ✉e-mail: thomas.thesen@dartmouth.edu

that searches through a curated knowledge base. When a user asks a question, the system first retrieves relevant information from a specific database (in this case, course materials), then provides this retrieved content as context to the LLM, which generates a response grounded in these authoritative sources rather than relying solely on its pre-trained knowledge[11]. By constraining LLM responses to instructor-curated, course-specific materials rather than allowing unconstrained generation from general training data that may contain unverified or outdated medical information from the internet or old textbooks, RAG significantly reduces the risk of hallucinations while maintaining the conversational flexibility that makes LLMs valuable as educational tools. This approach has shown success in varied business and educational settings for reducing hallucinations and increasing response relevance and accuracy[12,13] (See Fig. 1).

Despite RAG's promise in delivery of accurate information, it is unclear how well this technology can be integrated into medical education given that medical students represent a distinct adult-learner population with specialized learning needs. The cognitive demands of processing large amounts of complex information in compressed timeframes, coupled with medical students' established habits of utilizing multiple learning resources create unique implementation and adoption challenges for AI learning tools in medical education[14].

The Technology Acceptance Model (TAM)[15,16] posits that perceived usefulness and ease of use adoption determine system usage, and numerous studies have investigated student adoption of learning technologies using this framework. Students tend to embrace tools they perceive as both beneficial to their learning goals and straightforward to integrate into existing study practices[17]. Following this framework, we developed the following research questions: (1) How would medical students integrate an AI teaching assistant into their self-regulated learning processes? (2) Would the RAG-based design enhance perceived usefulness by increasing trust in AI-generated answers?, and (3) Would the conversational interface and 24/7 availability improve perceived ease of use compared to traditional resources? Drawing from TAM, we hypothesized that students would adopt NeuroBot TA if they perceived it as enhancing their learning efficiency, i.e., usefulness, while requiring minimal effort to learn and operate, i.e., ease of use, and that usage patterns would correspond to periods of high information need when the added value of the system was most evident. We expected this strategic, exam-focused adoption pattern based on cognitive load theory that predict students preferentially seek external support when cognitive demands peak and when time constraints make efficient clarification of uncertainties most valuable[18].

While there has been a substantial interest in deploying LLMs for medical education, to our knowledge, this is the first study that reports on the deployment and evaluation of usage patterns and student attitudes towards a RAG-based LLM platform in medical education. The only deployed, corpus-grounded system has been used in a graduate medical education journal club that embedded assigned articles in a vector database and was evaluated qualitatively with residents and faculty[19]. By contrast, most work to date has relied on base LLMs without automated and constrained retrieval that range from clerkship feedback tools and ward-based case studies to co-designed tutoring concepts, but not employing provenance-grounded retrieval[20–23].

Results

Usage and engagement statistics

Across two academic years, students initiated 360 unique conversations with NeuroBot TA, generating a total of 2946 individual messages. See Fig. 2 for usage analytics and engagement patterns. Conversation length ranged from 1 to 47 turns (with 1 turn being defined as one new message by either a student or the assistant within a conversation) with a mean of 3.6 turns per

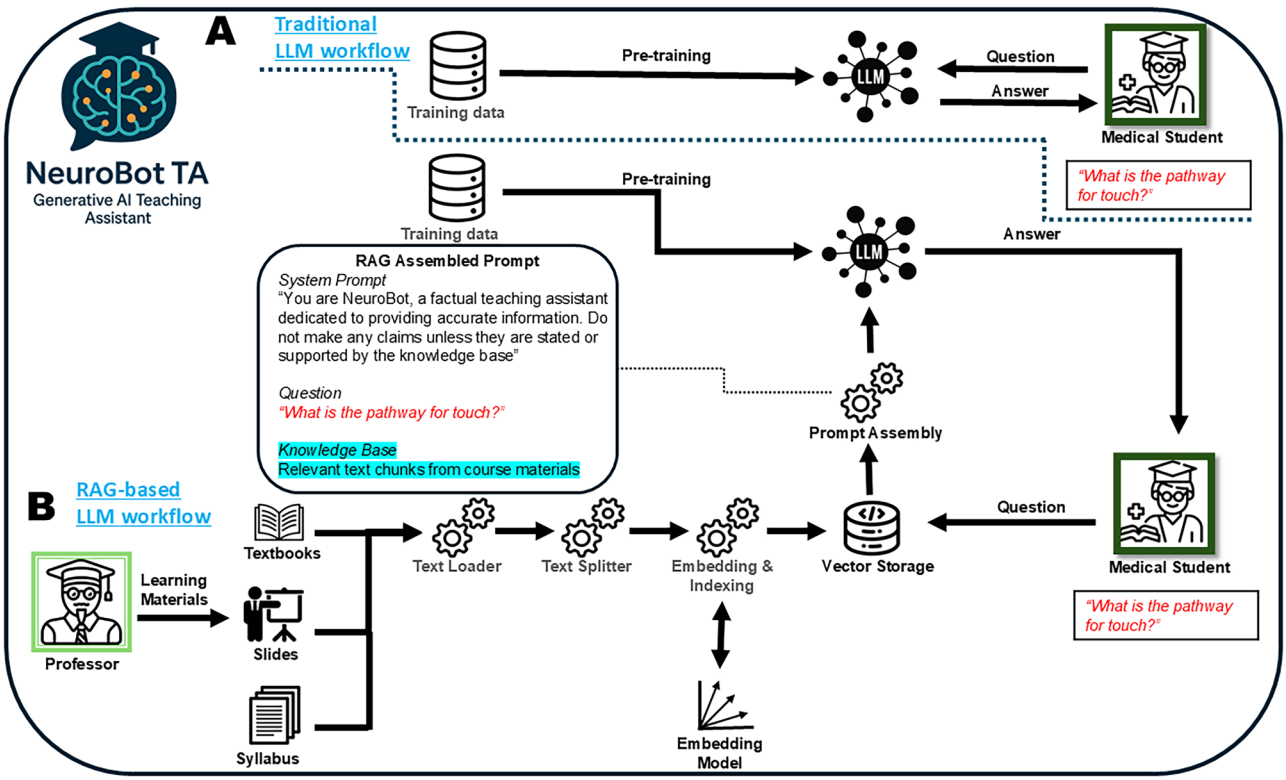

Fig. 1 | NeuroBot TA System Architecture and Information Flow. Schematic diagram contrasting standard LLM implementation with retrieval-augmented generation (RAG) pipeline used in NeuroBot TA. **A**Traditional LLM workflow: the model is pre-trained on general training data and directly responds to student questions without course-specific context. **B** Rag LLM Workflow Implemented In Neurobot Ta: course materials provided by the instructor undergo processing through text loading, splitting, and embedding before storage in a vector database. When a student submits a question about course content, the system retrieves relevant text chunks from the vectorized knowledgebase, assembles them with the system prompt into a context-enriched query to the LLM, which in turn provides answers specifically grounded in course-related materials rather than purely based on general knowledge from pre-training of the LLM.

conversation (Median = 2). The distribution of conversation turns is shown in Fig. 2A. Student message length ranged from 1 to 272 words (median: seven words, SD = 20.65) while bot responses ranged from 2 to 1338 words (median: 82 words, SD = 130).

Temporal analysis of weekly interaction patterns showed that usage peaked mid-week, with lowest engagement on Fridays and Saturdays (Fig. 2B). Between academic years, total number of conversations decreased from 256 to 104 (−59.4%) and number of messages from 1922 to 1024 (−46.7%).

Time-of-day analysis

Daily patterns showed highest use during the day but also substantial after-hours utilization (i.e., after 5 pm) (Fig. 2A). Timeseries plots of conversation frequency from both cohorts showed large increases in the days leading up to exams (Fig. 3B, D). While students in cohort 1 engaged with NeuroBot TA consistently throughout the course, usage by students in cohort two dropped significantly throughout the course. Across both cohorts, independent samples t-tests comparing conversation volume during pre-exam periods (defined as the three days leading up to each exam, $M = 5.11$, SEM = 1.00, $n = 18$) with regular course periods ($M = 1.14$, SEM = 0.14, $n = 236$) showed a significant increase during pre-exam periods compared to regular course periods ($t(252) = 3.95$, $p < 0.001$, Cohen's $d = 1.74$), overall representing a 329.6% increase in conversation volume during the three days leading up to exams. Sub-analysis of the 2023–2024 academic year showed that conversation frequency increased during pre-exam periods (7.33 ± 1.22 conversations/day) compared to average frequency across the whole period (1.58 ± 0.23 conversations/day), representing a 363.2% increase ($t = 4.616$, $p = 0.0014$, Cohen's $d = 2.232$). For 2024–2025, a similar pattern emerged with pre-exam periods averaging 2.89 ± 1.22 conversations/day versus 0.67 ± 0.13 during regular periods (329.6% increase), though this difference did not reach statistical significance ($t = 1.809$, $p = 0.1073$, Cohen's $d = 1.344$). The results of the differences in conversation volumes are shown in Fig. 3C. Two-way ANOVA confirmed significant main effects for both conditions (pre-exam vs. regular, $p < 0.001$) and academic year ($p < 0.001$), with a significant interaction effect ($p = 0.0010$), indicating that the first cohort engaged more with the TA bot for exam preparation.

Conversation content analysis

Thematic analysis of student messages to the chatbot identified eight primary content domains (Fig. 4A). Neuroanatomy and physiology constituted the most frequent theme (65.9%, 216 conversations, examples: *tell me about the vermis of the cerebellum; what do the mamillary bodies do?)*, followed by clinical syndromes and disorders (53.7%, 176 conversations, examples: *what is Brown-Sequard syndrome?; What causes ALS?*). Others included educational methods and resources (31.4%, 103 conversations, examples: *what is the course grade composed of?; Give me some high yield information for benchmark #2*), Neural Pathways and Tracts (29.6%, 97 conversations, examples: *What is the pathway for touch?; Is the dorsal column medial lemniscus pathway myelinated?*), and Course and Exam Information (28.4%, 93 conversations). Less frequent topics included Pharmacology and Treatment (14.9%, 49 conversations, examples: *What do I need to know about trihexyphenidyl?; What are the side effects of carbidopa?*), Clinical Case Discussions (14.3%, 47 conversations, examples: *What are the differential diagnosis for a patient with pain in the left eye?; A 23-year-old-woman is brought to the emergency room after being involved in a car crash…*), and Imaging and Diagnostic Techniques (7.0%, 23 conversations, examples: *what would parkinsons show on mri?; what is the 'hot cross bun' sign on mri?*). Percentages do not add to 100% because a conversation could include multiple themes. On the independently human-coded validation subset, overall LLM–human percent agreement was 78.4%. Weighted Cohen's κ for the ordered 0–3 intensity ratings was 0.64 (95% bootstrap CI 0.58–0.70). When collapsing ratings to present vs absent, overall macro-F1 was 0.76, together showing moderate to good alignment with human coding.

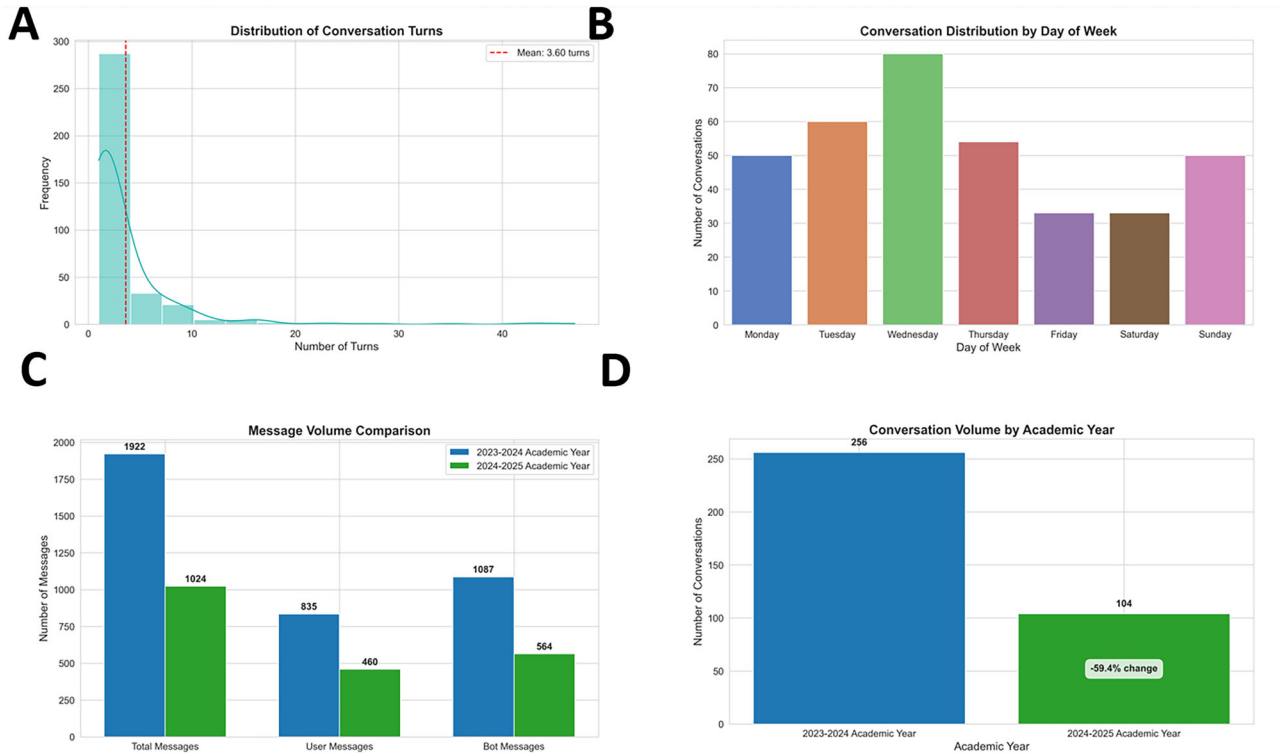

Fig. 2 | Usage Analytics and Engagement Patterns. A Distribution of conversation turns showing frequency of interactions by number of turns. Histogram with kernel density estimation overlay. Mean conversation depth: 3.6 turns (vertical red line). **B** Conversation distribution by day of week showing total number of interactions across all days of the week. **C** Message volume comparison between academic years 2023–2024 (blue) and 2024–2025 (green). Data segregated by message type: total messages, user messages, and bot messages, with counts labeled on each bar. **D** Total conversation volume comparison between academic years.

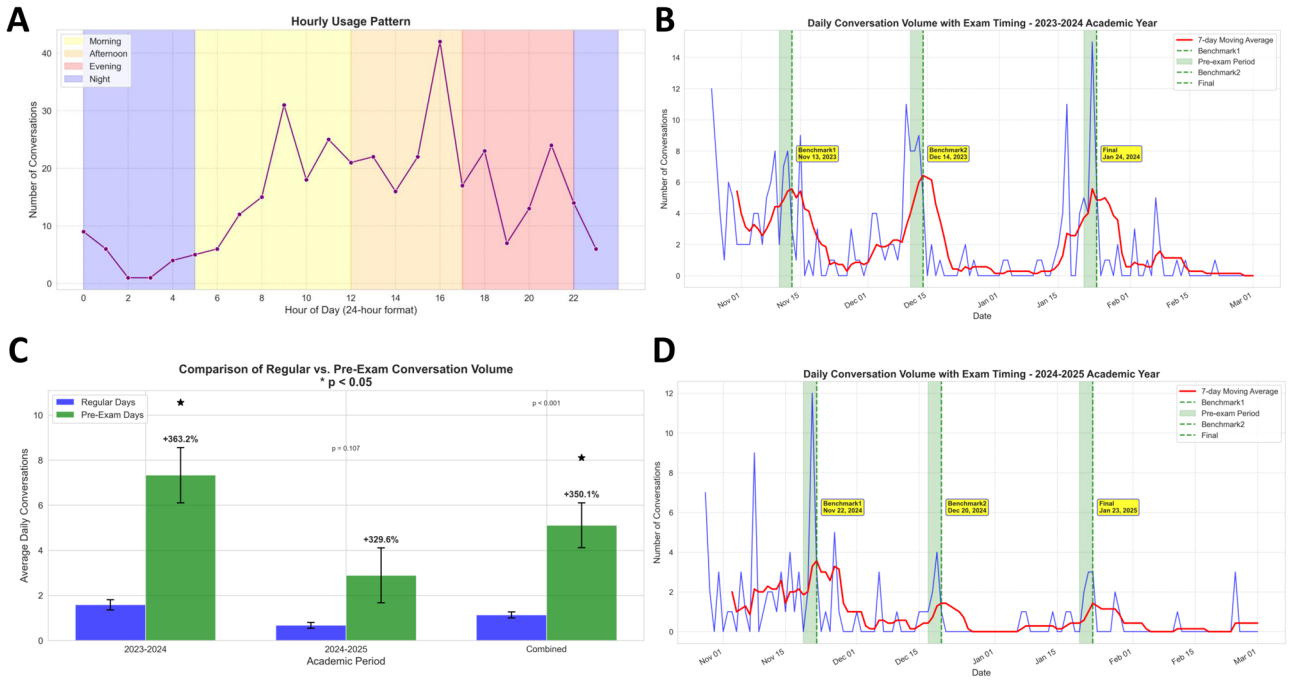

Fig. 3 | Usage Patterns and Exam Timing Analysis. A Average hourly distribution of NeuroBot TA usage throughout the day. Color-coded time periods indicate Morning (5:00-12:00, yellow), Afternoon (12:00-17:00, orange), Evening (17:00-22:00, red), and Night (22:00-5:00, blue). **B** Daily conversation volume during the 2023–2024 academic year relative to exam dates. Green vertical bands indicate pre-exam periods (3 days before each assessment). Blue line represents daily conversation counts while red line shows 7-day moving average. Yellow labels indicate exam dates. **C** Average daily conversation volume comparison between regular course days (blue) and pre-exam days (green) across academic periods. Error bars represent standard error of the mean. **D** Daily conversation volume during the 2024–2025 academic year relative to exam dates. Format follows (**B**).

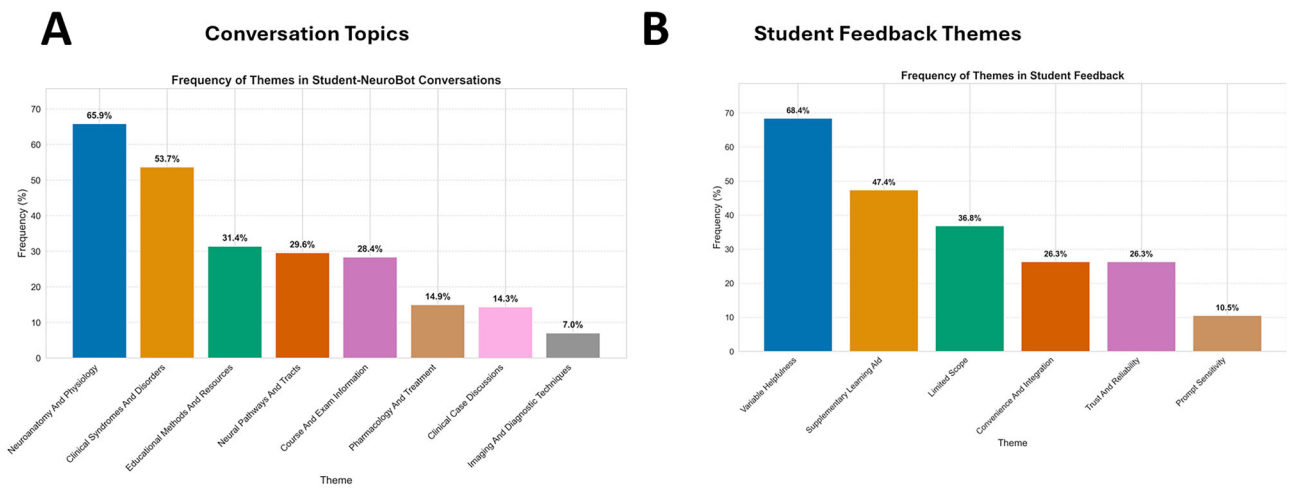

Fig. 4 | Distribution of Primary Themes in Conversations. A Proportional distribution of conversation volume across eight major topics. **B** Distribution of primary themes in student feedback.

Student feedback

In the first cohort (2023–2024), 39.3% (22/56) of students who answered the end-of-course survey reported using the AI assistant at least once during the course, while usefulness was judged at 2.8/5 (SD = 1.4). In the second cohort (2024–25), 26.4% (23/87) of students who completed the survey reported using NeuroBot TA at least once, with usefulness judged as 3.3/5 (SD = 1.0). Overall usage across both cohorts was 31.4%.

Figure 4B shows eight primary themes and their frequency identified in student feedback comments. The 'Variable Helpfulness' theme was the most prevalent theme (68.4%, 13 comments, example: *"I tried to have the Bot come up with a study guide off of the red starred slides and it wasn't able to, but it could give me feedback on grading breakdowns."*), followed by the 'Supplementary Learning Aid' theme (47.4%, 9 comments, example: *"Was extremely useful to be able to quickly ask the AI if a question came up while I was preparing for an exam."*), and the 'Limited Scope' theme (36.8%, 7 comments, example: *"wasnt helpful. I use other AI tools to help create comparative/summary tabels and this was very helpful, but Neurobot ta wasn't able to answer a lot of questions and wasn't that helpful"*). Less frequent themes included 'Convenience and Integration' (26.3%, 5 comments, example: "*is pretty good at summarizing info instead of going through lots of slides for a specific question*"), 'Trust and Reliability' (26.3%, 5 comments, example: *"I don't trust AI yet to give me learning materials, especially after having tried Chat GPT with research articles. I'm aware that the NeuroBot TA only pulls from class materials, which is great."*), and Prompt

Sensitivity (10.5%, 2 comments, example: *"The NeuroBot would often give very long answers for relatively simple questions. I don't blame the bot for this though, I was probably using inefficient prompts and am someone who is still in the habit of searching the web for information rather than using a chatbot."*).

Discussion

This study examined the implementation of a RAG-based AI teaching assistant in medical education across two consecutive academic cohorts and demonstrates a scalable, always-available AI support tailored to curricular contexts and students needs. The results identified both opportunities and challenges in deploying RAG-based AI teaching assistants in medical education programs.

The data from participating students demonstrate how some medical students integrated NeuroBot TA into their self-regulated learning processes. Students exhibited strategic, context-dependent usage patterns rather than continuous engagement. The 329% surge in usage during pre-exam periods and substantial after-hours utilization (post-5 pm) indicate that students primarily leveraged the system as a just-in-time learning resource during intensive study sessions that often extended beyond the hours of instructor availability. The average conversation length of 3.6 turns suggests focused, targeted queries rather than extended tutoring sessions. This pattern indicates that medical students use the chatbot as a targeted reference tool during self-directed study, asking specific questions to clarify concepts rather than seeking extended tutoring or comprehensive instruction.

The RAG-based design showed mixed results in enhancing perceived usefulness through increased trust. Positive indicators included students' appreciation for source citations (26.3% of feedback comments mentioned trust and reliability themes), with one student noting that knowing "the NeuroBot TA only pulls from class materials" increased confidence. The predominance of curriculum-aligned queries (66% neuroanatomy, 54% clinical disorders) suggests students trusted the system for course-specific content. However, the modest usefulness ratings (2.8/5 and 3.3/5 across cohorts) and the "limited scope" frustration theme (36.8% of comments) indicate that while restricting responses to content in the course materials may enhanced accuracy and trust, it simultaneously reduced perceived utility of the tool by restricting response breadth, thereby creating a tension between response reliability and comprehensiveness.

The conversational interface and 24/7 availability did improve perceived ease of use, as evidenced by usage patterns and feedback. Some users valued the "convenience and integration" (26.3% of feedback), with comments highlighting quick access to answers during exam preparation. The temporal distribution showing after-hours usage and mid-week engagement peaks suggest that users found the system accessible when needed. However, some students noted challenges with prompt engineering (10.5% mentioned prompt sensitivity), suggesting that while the interface was accessible, optimal utilization may have required AI skills students were still developing.

These findings collectively support our TAM-based hypothesis partially: students adopted NeuroBot TA when perceived benefits were most salient (pre-exam periods), but overall adoption remained moderate (31.4%) due to the tension between the system's reliability constraints and students' expectations for comprehensive support.

Content analysis of the chat transcripts confirmed that students primarily used NeuroBot TA to reinforce core course knowledge. The vast majority of queries centered on the course's fundamental content and neuroanatomy and neurophysiology concepts were the single largest topic, appearing in about 66% of conversations, and clinical neurological disorders appeared in ~54%. Students also frequently asked about study resources or clarifications of course logistics (appearing in roughly 28–31% of chats, e.g., exam information or "high-yield" review tips). Other academic topics like neural pathways, tract anatomy, and pharmacology were moderately represented, whereas more complex clinical case discussions or neuroimaging interpretation were relatively infrequent. This distribution suggests that the AI assistant was mostly used for fact-based clarification and review of taught material, rather than for open-ended clinical reasoning practice. It may also reflect the constraints of its knowledgebase as NeuroBot TA was intentionally limited to instructor-curated course materials, which ensured answers stayed aligned with the curriculum but inherently capped the scope of questions it could address. The content pattern therefore illustrates a trade-off where the bot excelled at fielding questions on covered topics, but students may have recognized that queries outside the provided content, or requiring extensive synthesis, were beyond its scope.

Student feedback on NeuroBot TA highlighted both the promise of AI support as well as opportunities for improvement. Qualitative comments showed that while many appreciated the bot as a convenient supplemental resource, its performance was inconsistently helpful. The most prevalent theme in feedback was "variable helpfulness" and students noted the AI could answer certain questions well (for example, providing quick explanations or grading clarifications) but failed at other tasks (such as generating comprehensive study guides). Similarly, nearly half of commenting students viewed NeuroBot as a "supplementary learning aid", valuing the ability to get instant answers during self-study (e.g., while preparing for exams). On the other hand, many users were frustrated by the system's "limited scope" and about 37% of comments noted that the bot could not address numerous questions, especially those requiring information beyond the uploaded course slides. This limitation sometimes led students to revert to other AI tools or resources for comparison. Additional feedback themes underscored practical considerations. For example, some praised the convenience and integration of the 24/7 chat format within the learning platform, noting it saved time searching through slides. However, trust and reliability emerged as a concern (26% of comments) where a subset of students expressed reluctance to fully trust any AI-generated answers without verification. Encouragingly, students acknowledged that constraining the bot to official class materials through RAG improved their trust. Lastly, some users mentioned that the quality of answers depended on how questions were asked, with the bot sometimes giving overly lengthy answers to simple questions. This indicates that students were still learning how to interact optimally with LLM chatbots, which is an expected challenge as both users and technology co-evolve.

Usage declined notably in cohort 2, potentially reflecting the rapidly evolving GenAI landscape. When deployed with cohort 1 in Fall 2023, NeuroBot TA represented novel technology less than 10 months after ChatGPT's initial release. Internal survey data indicated limited prior AI experience among cohort 1 students, making the Neurobot TA system novel and distinctive. Conversely, cohort 2 in Fall 2024 entered during a period of more widespread AI adoption in higher education. Medical student utilization of general AI platforms increased dramatically during this period, with up to 89% reporting regular use[24]. While the present study design does not allow us to draw clear inferences, the growing availability of commercial alternative models with reasoning capabilities that are demonstrating improved performance on medical knowledge benchmarks may lead students to use these systems more frequently[25,26]. The usage patterns observed with NeuroBot TA seem to reflect this demand for quick, accessible support where students tended to use the assistant as a just-in-time tutor during intensive study periods (notably mid-week and pre-exam) when instant clarification of concepts was most valued.

The finding that conversation volume increased substantially during pre-exam periods demonstrates how assessment events can drive the use of self-directed learning resources, including AI chatbots. This behavior indicates that students viewed the system primarily as an optional review tool rather than a continuous learning partner and suggests a strategic but yet limited integration into their study practices. The course where the chatbot was deployed was the last organ-system course at the end of Year 2 in the preclerkship curriculum. At this point students will have already found their preferred study method and tools and are therefore less likely to engage in strong shifts in how they study and take the risk of adapting a new, untested study tool. This may have affected the moderate adoption rate despite the societal hype of Generative AI at the time. In the future,

implementation of RAG-based AI tutor system may early in the medical school curriculum when students have not yet solidified their study approach may lead to higher adoption rates, and should include best prompting techniques for interacting effectively with AI-teaching assistants.

There was a clear hierarchy of student learning priorities through chatbot interactions, with core biomedical content dominating conversation frequency, followed by clinical disorders, a finding that matches the content and focus of the preclerkship course. Students also asked a significant number of questions about course organization and exams, highlighting the utility of RAG-supported LLMs in answering questions specific to an individual course based on information that was not part of the original LLM training data. To that end, students primarily leveraged the AI assistant for clarifying course content and reviewing concepts, treating it as an on-demand tutor for course-related content.

Overall, students demonstrated moderate willingness to engage with the course-constrained AI tool, and those who did valued its instant and around-the-clock access to verified information. This availability complements the contemporary nature of how medical students study, as they often study at odd hours, already use digital resources like question banks and online tools, and value self-paced studying. NeuroBot TA provided an additional tool that fit naturally into these eclectic patterns as it was accessible on the same devices they already use and at any time in any location. Additionally, NeuroBot TA's retrieval-based design provides highly targeted, on-demand explanations that align with students' individual study goals (e.g., lecture-specific clarification before exams). Overall, the principles of personalization that LLM-based RAG chatbots offer are consistent with the emerging framework of *Precision Medical Education* that advocates for tailoring educational interventions to each learner's specific needs and context[27,28].

Our findings also align with TAM predictions and demonstrate how perceived usefulness and ease of use shaped NeuroBot TA adoption. The increase in usage during pre-exam periods reflects TAM's principle that perceived usefulness drives adoption when benefits are most salient. Neuroanatomy-focused conversations and positive responses to source citations demonstrate that curriculum alignment and verifiability enhanced perceived usefulness. Substantial after-hours utilization suggests that students found the system accessible when most convenient, satisfying TAM's ease of use dimension. Finally, frustration regarding knowledge scope limitations highlights TAM's expectation alignment principle. This principle predicts that when system capabilities do not match student expectations, student satisfaction decreases, underscoring the importance of clearly communicating system capabilities and boundaries in educational AI implementations.

Notably, NeuroBot TA's intentional constraint to its knowledge base led to student frustration when it refused to address queries beyond the scope of the knowledge base. This compares directly to all-purpose commercial chatbots who typically provide plausible-sounding answers regardless of factual accuracy. Student comments help illustrate this point in the context of the evolving GenAI landscape and the frustration with restricted answer space. A *cohort 1* student, for example, noted, "The chatbot is interesting to play around with but if I have a question, I tend to just pull up Google because it is convenient and is what I have done my whole life,", highlighting the student's unfamiliarity with Generative AI chatbots and preference towards established strategies to access knowledge. In contrast, a *cohort 2* student stated, " I use other AI tools to help create comparative/summary tabels (sic) and this was very helpful, but Neurobot ta wasn't able to answer a lot of questions." A year later, the cohort 2 student had already incorporated GenAI technologies into their study practices, but valued unconstrained answers even if the likelihood of factual incorrectness may be higher.

Pedagogically, educators can address this through explicit instructions in Generative AI uses and misuses and providing students with the knowledge how to responsibly navigate these new and ubiquitous tools for their learning. Technically, future work may focus on developing RAG-based systems that are able to generate flexible study schemes that matches a student's learning habits (e.g., automatic generation of lecture-specific tables or flash cards) and showing students how to most effectively initiate conversations with specific aims.

With only a subsection of students using NeuroBot TA and survey responses from a subset of students who reported usage, our findings do not represent the entire medical student population. Statements about student preferences and strategic adoption should be interpreted within this self-selected sample in mind. In addition, the per-conversation analysis could not distinguish whether usage patterns reflected typical behavior or were driven by a small number of "super users," as we did not track individual user engagement. This limits the ability to generalize directly to the average student experience. Despite theoretical and empirical works demonstrating reduced hallucinations and increased relevance of RAG-constrained responses, the current study did not systematically assess response accuracy. While periodic informal review by the course director identified no critical issues requiring intervention, we cannot quantify accuracy improvements or definitively confirm that RAG constraints eliminated all hallucinations. Finally, the content analysis relied primarily on GPT-4o for thematic coding with human validation of only 15% of conversations for theme coherence rather than comprehensive accuracy assessment. This approach, while efficient, may have missed errors that the LLM could not identify. Lastly, the deployment of the tool at only a single medical school and restricts the study's generalizability across diverse institutional contexts with varying curricula, teaching methodologies, and student demographics.

RAG-constrained LLM chatbots show potential as adjunct study tools for some medical students in self-directed learning contexts. However, they need to be deployed thoughtfully in order to engage students meaningfully and contribute to their learning. We recommend that educators introduce the new technology early in the medical school curriculum, preferably in the first course when students are trying different study techniques that work for them in medical school. Educators who are considering broader implementation should focus foremost on carefully curated content that is submitted to the vector database to improve response relevance. Specific documents that contain useful information about the course resources, assessments and effective study strategies should be included to allow the bot to answer questions in this area, as we found that students asked these questions frequently. Bot responses should also enable a deeper dive into the material by highlighting the text sources that were retrieved and to link directly to the document of origin and the place of quotation. Furthermore, educators need to decide whether to restrict the bot to answer questions solely based on course material at the risk of student frustration or to allow increasingly sophisticated models to answer questions beyond immediate course content. Educators also need to communicate system capabilities and limitations clearly to help students select and leverage these tools effectively in their courses. Lastly, medical programs should ensure students gain fundamental knowledge of GenAI, including prompt engineering, to effectively select and use suitable AI learning tools, whether they are provided by the school or through commercial sources.

Several approaches could address the tension between response accuracy and comprehensiveness identified in this study, while also following best pedagogical practices for long-term learning. For example, a hybrid system could clearly mark responses derived from the course-specific RAG database as highly reliable while flagging answers requiring external knowledge with accuracy warnings, which would allow students to assess information trustworthiness. Beyond RAG, knowledge graph architectures could enable more sophisticated cross-topic synthesis while maintaining accuracy through formal ontological constraints that explicitly map relationships between medical concepts[29]. Furthermore, incorporating Socratic tutoring methods, where the AI guides students through problems with targeted questions rather than direct answers, could transform the system from a passive answer service into an active learning partner that promotes deeper understanding and long-term retention[30]. Such systems could also adapt their approach based on context, providing direct answers during time-sensitive exam preparation while employing Socratic dialog during regular study sessions to develop critical thinking skills.

Methods

This study was conducted in two consecutive cohorts (Cohort 1: $n = 92$, Cohort 2: $n = 98$) in a second-year organ-system course focusing on Neuroscience & Neurology at Geisel School of Medicine at Dartmouth that runs over 14 weeks. Classes for cohort 1 ran from October 2023 to January 2024 and for cohort 2 from October 2024 to January 2025. The course covered foundational neuroanatomy, neurophysiology, and clinical neurology topics. The instructional format was a blend of lectures, active learning sessions, case discussions, and laboratory sessions. The placement of the course was at the end of the basic science preclerkship phase and immediately before the dedicated USMLE Step 1 study phase. All course materials were made available to students via the learning management system (LMS), including lecture slide decks, assigned textbook readings, pre-class preparatory materials, a detailed syllabus, course and session learning objectives, and video recordings of all lectures.

NeuroBot TA system

NeuroBot TA was developed using a commercial AI platform (getcody.ai) that supports RAG with the latest available OpenAI GPT model (GPT-4 for cohort 1 and GPT-4o for cohort 2). We assembled a comprehensive knowledge base comprising of 145 documents in the English language comprising all documents available to students through the course's learning management system, including lecture slides, excerpts from textbook chapters, prework materials, instructor handouts, and the course syllabus. To create a RAG-compliant database documents were split into smaller "chunks" of text (~ 200–300 words each) which were then converted into vector embeddings (representing the native space of LLMs) and stored in a vector database for subsequent retrieval. This enabled vector-based similarity searches, like performed by Google search, which prioritize semantic similarity. When a student asks a question in the chat interface, NeuroBot TA's backend retrieves the most relevant content chunks from the course materials in vector space based on semantic similarity to the query. The retrieved text, along with source identifiers, is then appended to the prompt given to the LLM. The system message of the Neurobot TA was set to instruct the LLM as a helpful teaching assistant and to use the retrieved course content as context to answer the question, cite the source of the information (see Fig. 1B for a visual representation of the RAG-based approach). Importantly, NeuroBot TA was restricted to answering only questions grounded in its curated knowledge base. It also did not have access to the open internet, tools, or any data beyond the curated course documents. This was to ensure both accuracy and relevance, and to prevent the chatbot from generating any inappropriate content. The system was thus effectively an open-book exam of the course content for the AI. Questions outside the scope of the materials in the database elicited a response indicating the information was unavailable. See Supplementary Materials for screenshot examples.

Before initial deployment, pilot testing of the platform was conducted with faculty and a few volunteer students (not part of the study cohorts) to verify appropriate handling of questions, which led to prompt tuning for more refined answers. For example, we adjusted the assistant's tone to be friendly and encouraging and ensured consistent source citation to bolster trust and help students locate the referenced material for further reading.

Deployment and student access

For both cohorts, NeuroBot TA was introduced to students during the course orientation session as an optional study aid, with a live demonstration showing how to access the bot and ask questions. Access to NeuroBot TA was available 24/7 throughout the course through the LMS and via a link to the chat interface. Although no real-time moderation occurred, the course director (T.T.) had access to the anonymous chat transcripts and periodically reviewed them to monitor answer quality with a plan to correct the knowledge base or tune the prompts further should any critical inaccuracies be found. No major intervention was needed during either deployment.

Usage data collection and engagement analysis

We evaluated the TA bot's impact and student attitudes using a mixed-methods approach, collecting both quantitative usage data and qualitative feedback. Analyses were performed in Python 3.6. This quality improvement project received exemption from the Dartmouth College IRB. All data collection was anonymous, participation was voluntary, and students were informed that aggregated usage patterns might be analyzed for educational improvement purposes. Participation in the end-of-course survey was voluntary and had no effect on grades. The chat conversation logs collected over two academic years (2023–2024 & 2024–2025) were analyzed for descriptive engagement metrics (e.g., conversation count and message volume). Temporal analyses compared usage during pre-exam and regular periods using paired t-tests and ANOVA as appropriate to identify students' 24/7 interaction patterns. At the end of each course, an anonymous standardized course evaluation survey was administered that included specific questions about NeuroBot TA. We asked students a yes/no question whether they have used the bot during the course and the statement "NeuroBot TA was a helpful resource for this course", both rated on a 5-point Likert scale. In addition, an open-ended text box invited any comments on their experience with the AI platform.

Content Analysis

We conducted systematic content analyses of both student messages sent to the bot and student feedback comments. The content analyses were done in several stages by combining GPT-4o LLM processing and human-in-the-loop verification adapted from Braun & Clarke's framework[31]. First, conversations were preprocessed to isolate student messages from TA bot responses. Conversations with fewer than 50 characters of student input were excluded to ensure sufficient material for meaningful semantic content analysis. Initial themes were identified by randomly sampling 100 conversations/20 feedback comments and processing with GPT-4o for generation of candidate themes, which were reviewed by a human expert for coherence and relevance. Themes were then refined using a second sample ($n = 50/n = 15$), again with human verification of the refinement process to ensure accurate representation of data. Results were validated through co-occurrence analysis to establish distinctiveness of themes, resulting in a total of 8 message themes and 6 feedback themes. All valid conversations underwent systematic coding, where each was rated by GPT-4o for theme presence using a standardized scale (0 = not present, 1 = slightly present, 2 = moderately present, 3 = strongly present).

The computational coding was validated through manual review of a subset of conversations (15%). This human–LLM hybrid approach follows LLM-assisted content analysis workflows that have achieved moderate to substantial agreement with human coders[32–34]. LLM-human agreement was analyzed with weighted ordinal Cohen's k with 95% bootstrap CIs, and, on a present/absent collapse, overall macro-F1.

Data availability

All data and analysis code are publicly available at https://doi.org/10.6084/m9.figshare.30068977[35].

Received: 22 July 2025; Accepted: 20 September 2025;
Published online: 04 November 2025

References

1. Pincavage, A. T. et al. A national survey of undergraduate clinical education in internal medicine. *J. Gen. Intern. Med.* **34**, 699–704 (2019).
2. Pendergrast, T. R. & Walter, J. M. Use of an asynchronous discussion platform during the pre-clerkship curriculum: a multiyear retrospective study. *Med. Sci. Educator* **34**, 397–403 (2024).
3. Ten Cate, O. & Durning, S. Peer teaching in medical education: twelve reasons to move from theory to practice. *Med. Teach.* **29**, 591–599 (2007).
4. ten Cate, O., Gruppen, L. D., Kogan, J. R., Lingard, L. A. & Teunissen, P. W. Time-variable training in medicine: theoretical considerations. *Acad. Med.* **93**, S6 (2018).

5. Chebrolu, S. & Potti, R. Impact of 'tutorial classes' on learning outcomes, among medical students: a systematic review and meta-analysis. *Al-Azhar Assiut Med. J.* **22**, 105–109 (2024).
6. Ryan, L., Sheehan, K., Marion, M. I. & Harbison, J. Online resources used by medical students, a literature review. *MedEdPublish* **9**, 136 (2020).
7. Zhang, P. & Kamel Boulos, M. N. Generative AI in medicine and healthcare: promises, opportunities and challenges. *Future Internet* **15**, 286 (2023).
8. Chokkakula, S. et al. Quantum leap in medical mentorship: exploring ChatGPT's transition from textbooks to terabytes. *Front. Med.* **12**, 1517981 (2025).
9. Ganjavi, C. et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of US medical students. *PLOS Digital Health* **3**, e0000596 (2024).
10. Rawte, V., Sheth, A. & Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023).
11. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
12. Swacha, J. & Gracel, M. Retrieval-augmented generation (RAG) chatbots for education: a survey of applications. *Appl. Sci.* **15**, 4234 (2025).
13. Vijayasekaran, G. et al. Personalized Learning Platform using Artificial Intelligence. in 1883–1890 (IEEE, 2024).
14. Alrashed, F. A. et al. Incorporating technology adoption in medical education: a qualitative study of medical students' perspectives. *Adva. Med. Educ. Pract.* **15**, 615–625 (2024).
15. Davis, E. L. Marta: a stimulant to Atlanta development? *Transp. Plan. Technol.* **10**, 241–256 (1986).
16. Venkatesh, V. & Davis, F. D. A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manag. Sci.* **46**, 186–204 (2000).
17. Scherer, R., Siddiq, F. & Tondeur, J. The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Comp. Educ.* **128**, 13–35 (2019).
18. Leppink, J. & Van den Heuvel, A. The evolution of cognitive load theory and its application to medical education. *Perspect. Med. Educ.* **4**, 119–127 (2015).
19. Umer, F., Naved, N., Naseem, A., Mansoor, A. & Kazmi, S. M. R. Transforming education: tackling the two sigma problem with AI in journal clubs–a proof of concept. *BDJ Open* **11**, 46 (2025).
20. Bany Abdelnabi, A. A., Soykan, B., Bhatti, D. & Rabadi, G. Usefulness of large language models (LLMs) for student feedback on H&P during clerkship: Artificial intelligence for personalized learning. *ACM Transactions on Computing for Healthcare* (ACM, 2025).
21. Skryd, A. & Lawrence, K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. *JMIR Formative Res.* **8**, e51346 (2024).
22. Thesen, T., Tuan, R. L., Blumer, J. & Lee, M. W. LLM-based generation of USMLE-style questions with ASPET/AMSPC knowledge objectives: all RAGs and no riches. *Brit. J. Clin. Pharmacol.* (2025).
23. Wang, A. et al. Generative AI for medical education: insights from a case study with medical students and an AI tutor for clinical reasoning. *Cuerus* **17**, 1–8 (2025).
24. Zhang, J. S., Yoon, C., Williams, D. K. A. & Pinkas, A. Exploring the usage of ChatGPT among medical students in the United States. *J. Med. Educ. Curric. Dev.* **11**, 23821205241264695 (2024).
25. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
26. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
27. Burk-Rafel, J. & Triola, M. M. Precision medical education: institutional strategies for successful implementation. *Acad. Med.* **100**, 655–660 (2025).
28. Triola, M. M. & Burk-Rafel, J. Precision medical education. *Acad. Med.* **98**, 775–781 (2023).
29. Abu-Salih, B. & Alotaibi, S. A systematic literature review of knowledge graph construction and application in education. *Heliyon* **10**, e25383 (2024).
30. Bastani, H. et al. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proc. Natl. Acad. Sci. USA* **122**, e2422633122 (2025).
31. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**, 77–101 (2006).
32. Balt, E. et al. Deductively coding psychosocial autopsy interview data using a few-shot learning large language model. *Front. Public Health* **13**, 1512537 (2025).
33. Zhang, H. et al. Exploring inductive and deductive qualitative coding with AI: investigating inter-rater reliability between large language model and human coders.*AHFE Open Access* **195**.
34. Long, Y., Luo, H. & Zhang, Y. Evaluating large language models in analysing classroom dialogue. *npj Sci. Learn.* **9**, 60 (2024).
35. Thesen, T. & Park, S. A Generative AI Teaching Assistant for Personalized Learning in Medical Education. https://doi.org/10.6084/m9.figshare.30068977.

Acknowledgements

We thank Robert A. Shumsky for inspiration and discussion.

Author contributions

T.T. planned and implemented the study, analyzed the data and wrote the manuscript. S.P. contributed to data analysis and interpretation and wrote the manuscript.

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/s41746-025-02022-1.

Correspondence and requests for materials should be addressed to Thomas Thesen.