RESEARCH ARTICLE

Received: 13 FEV 2025 | Accepted: 17 FEV 2025 | **DOI:** https://doi.org/10.57849/ulisboa.fm.jscml.0000005.2025

JORNAL DA SOCIEDADE DAS CIÊNCIAS MÉDICAS DE LISBOA

Medical Students' Attitudes, Perceptions, and Usage of Large Language Models in Education: A Questionnaire-based Study

Sara Pereira[1,2], **João Costa**[2], **Thomas Hänscheid**[1,2]

[1] Universidade de Lisboa, Faculdade de Medicina, Instituto de Microbiologia, Avenida Prof. Egas Moniz, 1649-028, Lisboa, Portugal.

[2] Universidade de Lisboa, Faculdade de Medicina, Departamento da Educação Médica, Avenida Prof. Egas Moniz, 1649-028, Lisboa, Portugal.

✉ **Corresponding author:**
Thomas Hänscheid
Universidade de Lisboa, Faculdade de Medicina, Instituto de Microbiologia, Avenida Prof. Egas Moniz, 1649-028, Lisboa, Portugal.
E-mail: t.hanscheid@medicina.ulisboa.pt

ABSTRACT: Objectives: This study examined medical students' attitudes, perceptions, and usage patterns of AI, particularly large language models (LLMs), in medical education. The goal was to explore how these tools are used for academic purposes and their potential integration into medical curricula. **Methods:** A cross-sectional questionnaire was distributed to medical students across six academic years at a Portuguese institution during autumn 2024. Respondents rated their study habits, the relevance of digital resources, frequency of engagement with LLMs, trust in AI-generated content, and opinions on curricular integration. **Results:** A total of 306 students (13.4% response rate) completed the survey. AI was used by 87% of respondents, primarily for resolving theoretical doubts (84%), while its application in complex academic tasks was limited. Freely available models (GPT-3.5) were the most commonly used, whereas only 17% had experience with paid versions such as GPT-4. Trust in AI-generated clinical recommendations was low, with only 16% considering them reliable in a clinical case-based scenario. Limited familiarity (69%) and cost (58%) were identified as key barriers to broader adoption. No substantial evidence suggested widespread use of AI for academic misconduct. Despite scepticism regarding its reliability in clinical contexts, most respondents supported AI integration into the curriculum, with 65% favouring an optional course. **Significance:** Students frequently use AI for theoretical learning but remain sceptical of its reliability in medical decision-making. Addressing concerns through AI literacy and reducing cost barriers may encourage responsible adoption in medical education.

KEY WORDS: Medical education, artificial intelligence, survey, GPT

INTRODUCTION

Artificial intelligence (AI) has transitioned from a niche curiosity into a widely used, transformative force across numerous fields, including medicine and medical education. A key advancement in this domain is the development of Large Language Models (LLMs), exemplified by OpenAI's GPT-3 and GPT-4, though similar models such as Claude 3, Gemini, and DeepSeek are rapidly emerging[1,2]. LLMs process and generate human-like text using deep learning architectures and self-attention mechanisms, enabling tasks such as summari-

zation, translation, and answering medical licensing exam questions with accuracy comparable to human professionals[3,4]. While debates persist about whether these models rely on "simple" statistical patterns (the so-called stochastic parrot, or "glorious autocomplete") rather than genuine understanding[1], their growing influence in medical education, for instance as study aids, simulation tools, and clinical reasoning supports, is evident. Examples include enhancing case-based learning, exam preparation, and clinical documentation. At NYU Grossman School of Medicine, AI tools facilitate interactive learning [5], while studies highlight their role in generating question banks for standardized exams[6,7].

This integration aligns with broader trends in technology-driven learning. Medical students increasingly rely on digital platforms like Osmosis, AMBOSS, and Lecturio, which supplement traditional resources with interactive content[8–10]. Tools like ChatGPT, though not designed for medical education, are widely used to clarify concepts, assist with scientific writing, and provide feedback[11]. Notably, GPT-4 has achieved scores exceeding 86% on medical licensing exams[3], and it has also addressed ethical and professional topics with competence[4], which further increases the interest about its role in exam preparation and clinical reasoning. However, concerns persist about over-reliance, critical thinking shortfalls and bias, and inaccuracies in AI-generated content, particularly when confabulated responses mimic seemingly accurate statements[12].

Despite these developments, medical students' perceptions and usage patterns of AI tools like GPT-3/4 remain understudied. Existing research indicates that students recognize AI's potential but express concerns about ethical limitations and risks of dehumanizing clinical practice[13,14]. While interest in AI is high, foundational understanding of its principles and applications is often lacking[15,16]. Additional debates focus on privacy, trust, and the role of human oversight in AI-augmented workflows[17,18]. As medical curricula adapt to new technologies, understanding student perspectives is key to ensuring AI enhances rather than replaces essential clinical skills.

This study aims to explore medical students' attitudes, beliefs, and usage patterns regarding AI, focusing on LLMs such as GPT-3.5 and GPT-4 at the Faculty of Medicine, University of Lisbon (FMUL), Portugal.

METHODS

Study Design and Population:

This cross-sectional questionnaire-based study targeted medical students (years 1–6, n = 2,278) at the Faculty of Medicine, University of Lisbon (FMUL).

Recruitment and Data Collection:

Students were recruited via institutional emails and WhatsApp groups representing each cohort. Invitations were distributed in three waves: initial emails to year representatives (Week 1), reminders (Week 3), and direct emails to students in Years 2–4 (Week 6). The anonymous questionnaire, administered in Portuguese via Google Forms (Google LLC, Mountain View, CA, USA), was available from November 4 to December 12, 2024. Eligibility was restricted to FMUL students using institutional email addresses, with a one-response-per-account limit to prevent duplicates.

Questionnaire Design:

The 13-item questionnaire (Supplementary file and Table S1 for full questionnaire translated into English) assessed two domains: (I) Study Habits: Relevance of resources (e.g., in-person classes, AI platforms) for semester-long study and exam preparation; (II) AI Engagement: Usage frequency, perceived utility in medical contexts, trust in outputs, and opinions on curricular integration. Questions utilized Likert scales (1–7), multiple-choice, and open-ended formats (Table 1).

TABLE 1. Summary of Key Questionnaire Domains

Question	Focus	Response Type	Key Options/Scale
1	Year of study	Multiple-choice	Years 1–6
2	Study methods (semester)	Likert scale (1–7)	11 resources (e.g., AMBOSS, AI platforms)
5	AI platform usage frequency	Ordinal scale	ChatGPT-3.5, GPT-4, Gemini, Copilot
8	Trust in AI clinical recommendation	Scenario-based	5 options reflecting trust/skepticism
12	AI curriculum integration	Multiple-choice	5 strategies (e.g., mandatory courses)

Academic year (Q1, Years 1–6), study methods (Q2, rating resources like AMBOSS and AI platforms on a Likert scale), AI usage frequency (Q5, use of ChatGPT-3.5, GPT-4, Gemini, and Copilot), trust in AI recommendations (Q8, scenario-based responses), and curriculum integration (Q12, preferences for AI incorporation into FMUL's curriculum). The full questionnaire is available in Supplementary file and summarised in Table S1.

Ethical Considerations:

Data were anonymized and restricted to FMUL-affiliated emails. Ethical approval was granted by the CAML Ethical Committee (Ref. 193/24, September 2024).

Data Analysis:

All responses were exported from the online platform as CSV files, checked for completeness, and prepared for analysis. Sample characteristics were summarized using frequencies and percentages for categorical variables and medians with interquartile ranges (IQRs) for ordinal Likert scale items. Categorical variables were analysed using chi-square tests to assess differences in distribution across groups, with Cramér's V as a measure of effect size. A binomial test was used to determine whether the proportion of sixth-year students differed significantly from the expected value. Likert-scale responses were visualized using boxplots, where medians were marked by red lines, boxes represented IQRs, whiskers extended to 1.5 × IQR from Q1 and Q3, and outliers were displayed as dots. Given the non-normal distribution of Likert-scale data, comparisons between independent groups were performed using the Mann-Whitney U test. Effect sizes were calculated with Cliff's Delta to quantify the magnitude of observed differences. All statistical analyses and visualizations were conducted using Python 3.11.0 on macOS 12.7.6.

RESULTS

Of the 2,278 medical students at FMUL, 306 completed the survey (55% in preclinical and 45% in clinical years), yielding a 13.4% response rate (margin of error: ±5.2%). Response distributions varied across all 6 years, with significant deviations ($p < 0.001$, Cramér's $V = 0.33$, moderate effect; Figure 1). Due to weak effect sizes in preclinical vs. clinical comparisons ($p = 0.02$, Cramér's $V = 0.13$, weak effect), analyses were conducted at the preclinical versus clinical level. Response rates for most individual questions exceeded 97%, except for open-ended items (Supplementary Table S1). These received responses from as few as 3.2% of students and thus were excluded from further analysis.

Several traditional learning methods, e.g. textbooks and theoretical classes lost importance while a shift towards digital media, particularly video platforms, was observed (Supplementary Figures S1, S2). Learning platforms such as AMBOSS and other online resources received moderately positive ratings (Figure 2), though with a wide IQR (2–7). For exam preparation, these resources were pooled into a single composite, which showed a narrowing of the IQR to 4–6 while the median remained unchanged (Figure 2). However, each individual platform was less used during the semester than the composite for examen preparation (both $p < 0.001$; AMBOSS: $\delta = -0.23$, other online study platforms: $\delta = -0.16$, small effect sizes). Doubt-resolution tools (Figure 2), such as Google/Wikipedia or AI platforms, were

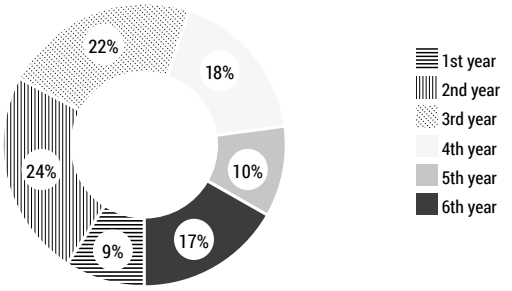

FIGURE 1. Participation in survey by year.

Striped patterns represent preclinical years (1st–3rd, left side), and solid shades represent clinical years (4th–6th, right side). Yearly participation differed significantly from an equal distribution ($p < 0.001$, Cramér's $V = 0.33$, moderate effect size), with 1st (-7.7%, $p < 0.001$) and 5th-year students (-6.7%, $p = 0.002$) underrepresented, while 2nd (+7.3%, $p < 0.001$) and 3rd-year students (+5.3%, $p = 0.012$) were overrepresented. Preclinical students (Years 1–3) outnumbered clinical students (Years 4–6) (56.5% vs. 43.5%, $p = 0.02$, Cramér's $V = 0.13$, weak effect size).

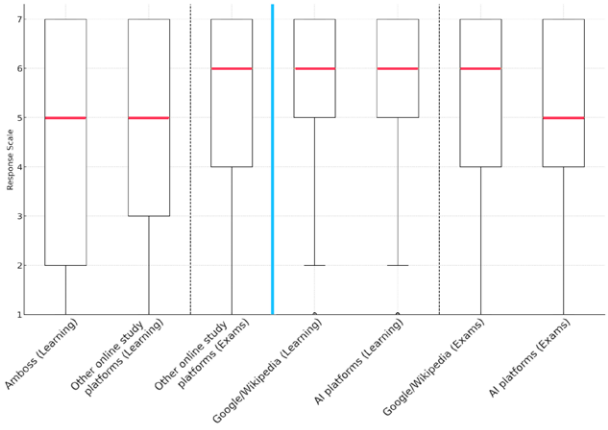

FIGURE 2. Relevance ratings of study methods.

Students rated study methods on a scale from 1 ("Not relevant at all") to 7 ("Extremely relevant"). A blue vertical line separates online study platforms (left) from doubt-resolution tools (right). Dashed grey lines distinguish semester-long study methods (left) from exam preparation tools (right). For exam preparation, semester-long resources are grouped under a single "online platform" category. Red lines indicate medians, boxes represent interquartile ranges (IQR), whiskers extend to 1.5 × IQR from Q1 and Q3, and outliers are shown as dots.

consistently rated highly. AI platforms demonstrated a slight but statistically significant decline in perceived relevance from semester-long learning to exam preparation (median: 6 to 5, IQR: 5–7 to 4–6; $p < 0.001$, $\delta = 0.19$, small effect size). In contrast, ratings for Google/Wikipedia increased slightly during exam preparation; however, the difference compared to AI platforms was not statistically significant ($p = 0.14$). Preclinical-year students preferred practical and theoretical-practical classes, *Sebentas* (study materials prepared by students from previous years), and online videos, whereas clinical years students favoured university-provided video lectures and AMBOSS (Supplementary Table S2). The strongest effect was observed for AMBOSS ($\delta = -0.52$, $p < 0.001$, medium to large effect), indicating a notable increase in its relevance for clinical year students.

Students expressed scepticism about AI's role in more practical tasks, rating it more useful for acquiring general knowledge than for developing practical skills (Supplementary Figure S3). Similarly, AI was perceived as helpful for checking basic medical questions and medical fact-checking but notably less useful during clinical rotations, where more complex clinical uncertainties arise (Supplementary Figure S2). Clinical-year students rated AI slightly higher for answering medicine-related questions ($p = 0.02$, $\delta = 0.15$, small effect size), but its utility during clinical rotations was rated similarly across groups ($p = 0.9$). Specific responses on AI's use in medical education reflected scepticism about its practical applicability (Figure 3, section A). While AI was perceived as useful for generating exam questions, its usefulness for creating practical simulations was rated lower. Trust in AI was limited, as shown in a dual-purpose question assessing both medical knowledge and confidence in AI-generated recommendations. When presented with GPT-4's suggestion of intravenous artesunate as the first-line treatment for severe malaria, only 16% of students trusted the AI's recommendation (details in Supplementary file)[19].

Students perceived AI as similarly useful for text correction and drafting, though drafting was rated slightly higher (Figure 3, section B; Supplementary Figure S4). Correction was viewed as moderately useful, with clinical students rating it slightly higher than preclinical students, though the difference was not statistically significant ($p = 0.120$; Supplementary Figure S5). In contrast, AI's usefulness for drafting was rated higher overall, particularly among clinical students, but again, the difference was not significant ($p = 0.409$; Supplementary Figure S5).

Astonishingly, 13% of students reported lacking both interest in and knowledge of AI. However, this may be an underestimate given the low 13.4% response rate. While 87% expressed interest, 69% had little prior knowledge, and only 18% actively engaged in independent learning. Most AI platforms saw minimal use, with 83%–90% of students reporting "never/almost never" using ChatGPT-4 (paid), Google Gemini, or Microsoft Copilot (Supplementary Figures S6). In contrast, ChatGPT-3.5 (free) was used more frequently with 58% of students using it less than once per day (Figure 4). The majority supported integrating AI into the FMUL curriculum through at least one approach. An optional AI course was the most preferred option (65%), while 30% supported integration into existing courses and 49% favored its inclusion in medical ethics subjects. Additionally, 58% wanted FMUL to fund AI tools, such as ChatGPT Plus subscriptions.

DISCUSSION

Most students in the sample used GPT-3.5 (the free version of ChatGPT) primarily as a general doubt-resolution tool, similar to Google or Wikipedia. Approximately 80% had never used more advanced models such as GPT-4 or alternative (paid) premium models, which may contribute to their scepticism regarding the reliability of AI and its limited recognition of its superior accuracy, as evidenced for example by premium model performance in medical examinations[3,6]. The reliance on free AI tools for quick academic queries suggests that access to more advanced models may be perceived as unnecessary. Concerns regarding AI-generated "hallucinations" (more accurately, confabulations) may also contribute to scepticism, as also reflected in the low acceptance rate of GPT-4's correct clinical case management recommendations in this survey[19]. Previous research indicates that negative experiences with AI can create a persistent bias against subsequent accurate responses[20,21]. Furthermore, AI-generated recommendations are often disregarded in favour of initial human judgments[22], reinforcing scepticism. If AI errors are recalled more readily than its correct outputs, this may further discourage its integration into clinical decision-making.

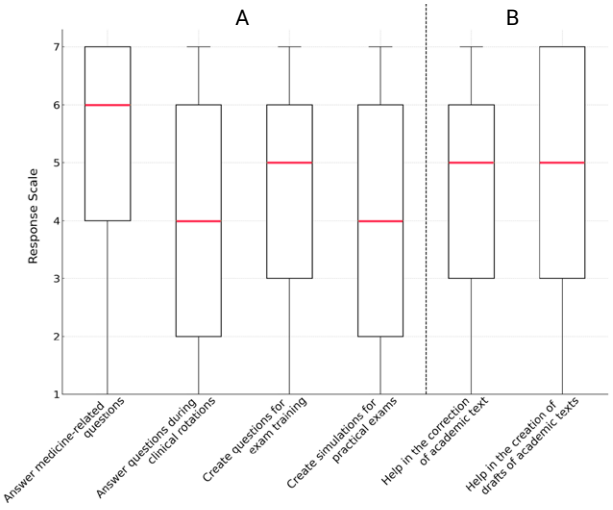

FIGURE 3. Perceived usefulness of AI in medical education.

Tasks related to doubt resolution and exam preparation (A) and academic writing, including assignments (B) are separated by a dashed grey line. Red lines indicate medians, boxes represent interquartile ranges (IQR), whiskers extend to 1.5 × IQR from Q1 and Q3, and outliers are shown as dots.

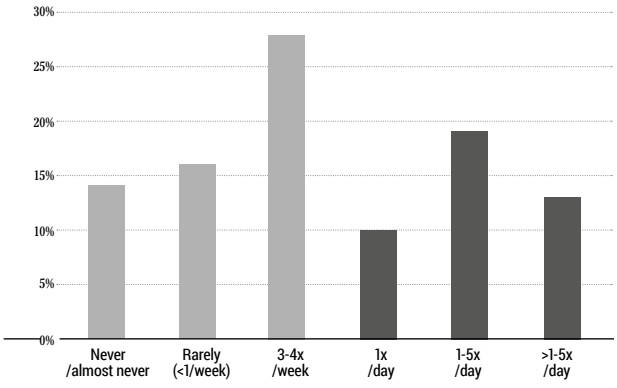

FIGURE 4. ChatGPT-3.5 (free Version) Usage Frequency.

Usage categories include less frequent users (grey bars) and more frequent users (black bars).

Addressing these concerns requires AI literacy programs that provide training on how AI generates outputs, its limitations, and strategies for critically evaluating its recommendations. Educational initiatives should incorporate data on AI error rates, reliability, and comparative performance against human decision-making to facilitate a more evidence-based approach to AI use in medical practice.

The monthly subscription fee (~€20) could also be a barrier for some. Others might see subscription services as binding or costly commitments. Unsurprisingly, 58% of respondents favoured institution-sponsored access, mirroring broader findings that they perceive cost is a major hurdle[23,24]. For instance, at Harvard, 40% of students use AI daily, but only 30% pay for subscriptions; those receiving financial aid are half as likely to do so[25]. Some institutions address this barrier through curricular integration or subsidized partnerships. At FMUL, universal premium access for around 2,300 students at €20 per month each would total €552,000 annually, roughly 2.5% of the university's €22.7 million budget and 13% of its goods-and-services allocation[26]. However, while significant, such an expense could be justified if it may yield equitable access, improved learning outcomes, and alignment with evolving global trends in medical education.

Academic misconduct involving generative AI has raised significant concerns in higher education[27,28]. However, our data suggest that this does not seem to be a major issue at the moment. Students are aware of AI for text creation, and it was rated as moderate, yet considerably lower than the use of traditional peer-created study materials such as *Sebentas* (see Supplementary Data). Interestingly, the exceptionally high rating for *Sebentas* seem to suggest a certain level of honesty in responses, given that they lack formal faculty endorsement. Limited experience with AI and restricted access to advanced models (like GPT-4) may also contribute to these findings. As AI writing tools become more sophisticated and widely available, usage patterns might likely shift, perhaps increasing the risk of misconduct. While institutions may consider AI-detection technologies, these appear to be more unreliable than often assumed[29,30] and risk falsely flagging legitimate student work. A balanced approach, integrating technological solutions with clear guidelines, may ultimately prove more effective in preserving academic integrity.

The low response rate (13.4%) with a ±5.2% margin of error may affect generalizability, particularly given the underrepresentation of certain academic years. However, studies in higher education suggest that response rates as low as 10–20% can still provide reliable estimates if nonresponse bias is minimal[31]. Furthermore, the proportion of students who are nev-

er/almost never users and heavy users (>5x/day) was relatively high and similar, making it less likely that our conclusions are driven by a non-representative sample favoring one site. While broader participation would strengthen representativeness, our findings align with international trends, supporting their relevance. It is possible that students less interested in AI may have been less inclined to participate. Additionally, self-reported data can be influenced by recall bias or social desirability. Despite these constraints, our findings mirror international trends of AI use in medical training[15,16].

The relatively high proportion of students who are either very low/low or high/very high users highlights the need to discuss the benefits of AI and increase foundational knowledge competencies for the very low/low users, as well as to call students' attention to the limitations of AI, potential overreliance, and ethical concerns for the high/very high users. Going forward, structured AI competencies—emphasizing not only technical skills and responsible use but also limitations—could help foster more informed integration of AI into clinical education. Coupled with institutional investments that mitigate cost barriers, such curricula could accelerate the safe and ethical adoption of AI tools among future physicians.

While AI is poised to enhance access to information and support critical thinking and medical decision-making, it is crucial to recognize that medicine inherently involves complexities that technology cannot simplify. As highlighted by Elder, delivering high-quality, patient-centered care requires medical training that is long enough, broad enough, and deep enough[32]. Therefore, it should not be expected that AI tools will make medical education and practice less challenging.

CONCLUSION

In summary, student interest in AI appears high scepticism about clinical reliability, concerns over cost, and limited exposure to more advanced models constrain broader acceptance. Addressing these issues, through dedicated AI-focused curricula, institutional support for premium tools, and ongoing investigations into AI's accuracy and ethical implications, will be pivotal in shaping a medical education landscape where AI enhances rather than undermines clinical expertise.

SUPPLEMENTARY FILE 1

Complete questionnaire (GPT4 - translation to English)

USE OF ARTIFICIAL INTELLIGENCE (AI) BY MEDICAL STUDENTS

My name is Sara Pereira, and I am a 6th-year medical student at FMUL.

Artificial Intelligence, despite its long history, has recently emerged as the leading technology of the moment, with countless applications in everyday life, clinical practice, and medical education.

As part of my Master's Final Project in the field of Medical Education, I am conducting a study on the use of Artificial Intelligence by medical students. Through this study, I aim to gather relevant data on students' perceptions of Artificial Intelligence, with the goal of positively influencing the medical school curriculum at FMUL.

The questionnaire, intended for FMUL medical students from the 1st to the 6th year, should take approximately 5 to 10 minutes to complete. This questionnaire has been approved by the President of the Pedagogical Council and the President of the Department of Medical Education (DEM).

Your participation in this study is completely anonymous and voluntary, and you may withdraw at any time. The confidentiality of your data is guaranteed in accordance with the legislation in force and the guidelines of the National Data Protection Commission (CNPD) (*Deliberação n.º 1704/2015, de 22 de outubro,* and *Decreto-Lei n.º 67/1998, de 18 de março*), as well as the terms required by the General Data Protection Regulation (GDPR; Regulation (EU) 2016/679 of the European Parliament and of the Council, of April 27, 2016).

INFORMED CONSENT – By proceeding with the completion of the questionnaire, I declare that I have been informed of its objectives and authorize the processing of my data exclusively for research purposes.

SECTION 1 – STUDY METHODS

The following questions will address the techniques and resources you use to study throughout the semester and prepare for evaluations.

1. What is your year of study? *(Mandatory question; Select only one option)*

- 1st year
- 2nd year
- 3rd year
- 4th year
- 5th year
- 6th year

2. On a scale of 1 to 7, where 1 means "Not relevant at all" and 7 means "Extremely relevant", rate how you would classify the following resources in terms of their relevance for learning during the semester:

- Attending theoretical classes in person.
- Watching recorded video lectures provided by the university.
- Attending practical and theoretical-practical classes.
- Studying using recommended bibliography textbooks (paper or PDF).
- Using study materials (notes, transcripts, summaries, etc.) prepared by students from previous years.
- Using Amboss for studying.
- Using other online study platforms (e.g., Sketchy, Osmosis, Lecturio, etc.) for studying.
- Using AI platforms (e.g., ChatGPT) for studying.
- Watching online videos (e.g., YouTube) for studying.
- Using search engines (e.g., Google) or online encyclopedias (e.g., Wikipedia) to clarify doubts.
- Using AI platforms (e.g., ChatGPT) to clarify doubts.

3. On a scale of 1 to 7, where 1 means "Not relevant at all" and 7

means "Extremely relevant", rate how you would classify the following resources in terms of their relevance to your evaluations (e.g., written exams, oral exams, TEM, OSCE, tests, quizzes):

- Watching recorded video lectures provided by the university.
- Consulting recommended bibliography textbooks.
- Using study materials prepared by students from previous years.
- Using online study platforms (e.g., Amboss, Sketchy, etc.).
- Watching online videos (e.g., YouTube).
- Using search engines (e.g., Google) or online encyclopedias (e.g., Wikipedia).
- Using AI platforms (e.g., ChatGPT).

4. If you use another method not mentioned or have any comments, please write them here: *(Open-ended, long answer question)*

SECTION 2 – USE OF ARTIFICIAL INTELLIGENCE

In this section, the goal is to better understand how you integrate Artificial Intelligence (AI) into your daily life and how you envision its utility in the future.

5. How often do you use the following AI platforms? *(>5x/day; 1-5x/day; 1x/day; a few times per week; rarely; never/almost never)*

- ChatGPT 3.5 (free version)
- ChatGPT 4 and 4o (paid version)
- Google Gemini
- Microsoft Copilot
- Other AI platform

6. If you selected "Other AI platform," please indicate which one. *(Open-ended short answer question).*

7. On a scale of 1 to 7, where 1 means "Not useful at all" and 7 means "Extremely useful," indicate how relevant you find the use of AI platforms (e.g., ChatGPT) in the following areas:

- Searching for general knowledge information.
- Learning practical skills unrelated to medicine (e.g., new languages, recipes, etc.).
- Clarifying medicine-related doubts.
- Clarifying doubts during clinical rotations.
- Assisting with the correction of academic text (e.g., assignments, medical histories, theses).
- Assisting with drafting academic texts (e.g., assignments, medical histories, theses).
- Creating exam training questions (multiple choice, open-ended, oral).
- Simulating clinical scenarios (e.g., OSCE training).

8. The following scenario aims to investigate how you would use information provided by GPT-4 and assess the extent to which you trust its accuracy in a clinical context:

A 35-year-old patient returns from Thailand with severe malaria. Immediate intravenous treatment must be urgently initiated. After consulting GPT-4 for guidance on treatment, it suggests that both intravenous artesunate and quinine are valid options for treating severe malaria, stating that artesunate is generally superior to quinine. However, GPT-4 also mentions that "artemisinin resistance" has been widely reported in Southeast Asia, citing a recent publication ("Time to contain artemisinin resistance," The Lancet, link). Despite this, GPT-4 recommends starting treatment with intravenous artesunate as the first-line therapy.

Based on the scenario above, and considering GPT-4's response, how would you proceed with the patient's treatment?

- Given the urgency of the situation and the information about artemisinin resistance, you base your decision on GPT-4's response and therefore initiate intravenous quinine as an alternative.
- You find GPT-4's response about artesunate being clinically superior interesting but consider starting quinine because "artemisinin resistance" has been reported in the region.
- Despite the urgency, you verify the information about "artemisinin resistance" in the region using other sources, and if confirmed, switch to quinine.
- Given the urgency of the treatment and the fact that GPT-4 mentions the superiority of artesunate, you follow the recommendation to initiate treatment with intravenous artesunate despite the mention of "artemisinin resistance," as it states that artesunate remains the recommended and effective first-line option.
- You ignore GPT-4's information about the continued usefulness of artesunate when "artemisinin resistance" is reported and consider it a "confabulation" or "hallucination" by the AI, deeming it unreliable.

9. On a scale of 1 to 7, where 1 means "Strongly disagree" and 7 means "Strongly agree," indicate how much you agree with the following statements about the use of Artificial Intelligence in Medicine:

- AI can be helpful in complex clinical situations (e.g., absence of clear clinical signs and symptoms, multimorbidity, deprescription of medications, etc.).
- AI can be helpful in non-clinical specialties (e.g., radiology, neuroradiology, clinical pathology, etc.).
- AI can assist with administrative tasks (e.g., writing clinical records, discharge summaries, etc.).
- AI can serve as a tool to reduce healthcare errors.
- AI could make some non-clinical specialties irrelevant.
- AI can substantially change clinical practice.
- AI can increase healthcare errors.
- AI raises ethical concerns regarding data protection and patient privacy.

10. To what extent do the following statements about AI knowledge apply to you? *(Select only one option)*

- I am interested in AI and am learning about it independently.
- I am interested in AI and am taking a course on the subject outside of university.
- I am interested in AI but do not have much knowledge.
- I am not interested in the subject and have no knowledge about it.

11. Would you like FMUL to promote the integration of Artificial Intelligence into the curriculum? *(Select only one option)*

- I do not believe AI knowledge is relevant for a medical student.
- Yes, I would like FMUL to promote the integration of AI into the curriculum.

12. If you answered yes to the previous question, how would you like FMUL to promote the integration of AI into the medical curriculum? *(Select all that apply)*

- I would like FMUL to integrate knowledge about AI into various subjects in the mandatory curriculum.
- I would like FMUL to create a mandatory subject about AI.
- I would like FMUL to teach about AI in the optional curriculum.
- I would like FMUL to include questions about AI in medical ethics subjects.
- I would like FMUL to promote the frequent use of AI in learning (e.g., through funding subscriptions to ChatGPT Plus for all students).

13. If you have another opinion not mentioned or any comments, please write them here *(open-ended long answer question).*

SUPPLEMENTARY FILE 2

Description of dual-purpose question

Question 8 served two main purposes. Firstly, it was designed to assess whether students understood the concept of partial resistance to a drug, rather than complete resistance. The correct answer (option d) would only be selected by students who were aware that "artemisinin resistance" refers

to delayed parasite clearance, not complete resistance, as described in the literature. Because most students did not choose the correct answer, this revealed a widespread misunderstanding of the term. These findings were the basis for a separate publication (Pereira SM, Grobusch MP, Hänscheid T. How a GPT-aided survey reveals a medical student's misunderstanding of the term 'artemisinin resistance'. New Microbes New Infect. 2024 Dec 5;63:101552. doi: 10.1016/j.nmni.2024.101552. PMID: 39759404; PMCID: PMC11699336.).

At the same time, the question also tested how students trusted information provided by GPT-4. The scenario described a clinical case of severe malaria, with the GPT-4 chatbot indicating that "artemisinin resistance" had been documented yet still maintaining that intravenous artesunate was the recommended first-line therapy. Only 16% of students selected the correct answer by following the chatbot's guidance (figure), while most of the remaining participants chose to switch to quinine or verify the suggested resistance through external sources before deciding on a treatment. Some found the chatbot's recommendation merely "interesting" but did not adhere to it, and a small subset dismissed the AI-based guidance entirely. These patterns show the challenges in establishing trust in AI systems for clinical decision-making, especially when the concepts involved—in this case, partial versus complete drug resistance—are subject to misunderstanding.

Original questions with *[comments in square brackets]***:**

The following scenario aims to investigate how you would use information provided by GPT-4 and assess the extent to which you trust its accuracy in a clinical context:

A 35-year-old patient returns from Thailand with severe malaria. Immediate intravenous treatment must be urgently initiated. After consulting GPT-4 for guidance on treatment, it suggests that both intravenous artesunate and quinine are valid options for treating severe malaria, stating that artesunate is generally superior to quinine. However, GPT-4 also mentions that "artemisinin resistance" has been widely reported in Southeast Asia, citing a recent publication ("Time to contain artemisinin resistance," The Lancet, link). Despite this, GPT-4 recommends starting treatment with intravenous artesunate as the first-line therapy.

Based on the scenario above, and considering GPT-4's response, how would you proceed with the patient's treatment?

a) Given the urgency of the situation and the information about artemisinin resistance, you base your decision on GPT-4's response and therefore initiate intravenous quinine as an alternative.
[3% – This indicates partial trust in GPT-4 (acknowledging its mention of resistance) but ultimately opting against its recommended first-line therapy.]

b) You find GPT-4's response about artesunate being clinically superior interesting but consider starting quinine because "artemisinin resistance" has been reported in the region.
[15% – These respondents acknowledge GPT-4's superiority claim for artesunate but choose quinine, reflecting uncertainty or scepticism about the AI's recommendation.]

c) Despite the urgency, you verify the information about "artemisinin resistance" in the region using other sources, and if confirmed, switch to quinine.
[62% – This represents the majority, who prefer caution by verifying GPT-4's statement externally before potentially changing the recommended treatment.]

d) Given the urgency of the treatment and the fact that GPT-4 mentions the superiority of artesunate, you follow the recommendation to initiate treatment with intravenous artesunate despite the mention of "artemisinin resistance," as it states that artesunate remains the recommended and effective first-line option.
[16% – This is the correct option, indicating trust in GPT-4's guidance despite the mention of resistance.]

e) You ignore GPT-4's information about the continued usefulness of artesunate when "artemisinin resistance" is reported and consider it a "confabulation" or "hallucination" by the AI, deeming it unreliable.
[4% – These respondents fully reject GPT-4's assertion that artesunate remains effective, dismissing the AI's input as untrustworthy.]

TABLE S1 Number of responses per question

Question number	Responses (n)	Responses (%)
1 *(mandatory)*	*306*	*100%*
2	*306*	*100%*
3	*305*	*99.6%*
4 *(open)*	*10*	*3.2%*
5	*306*	*100%*
6 *(open)*	*22*	*7.1%*
7	*305*	*99.6%*
8	*285*	*93.1%*
9	*301*	*98.3%*
10	*303*	*99%*
11	*297*	*97%*
12	*250*	*81.6%*
13 *(open)*	*16*	*5.2%*

The table summarizes the number of responses per survey question. Question 1 was the only mandatory question, ensuring a response from all participants (306), while questions 4, 6 and 13 were open questions where participants could enter text.

TABLE S2 Comparison of Learning Methods Between Preclinical and Clinical Students

Learning Method	p-value	Effect Size (d)	Effect Size Interpretation	Preference
Classes	< 0.001	0.242	Small to Medium	Preclinical
Sebentas	<0.01	0.180	Small	Preclinical
Online Videos	< 0.001	0.315	Small to Medium	Preclinical
Video Lectures	< 0.001	-0.274	Small to Medium*	Clinical
Amboss	< 0.001	-0.517	Medium to Large*	Clinical

"Classes" refer to practical and theoretical-practical classes, and "Video Lectures" to faculty-recorded lectures. Effect sizes (Cliff's δ) reflect differences in ratings: positive values favour preclinical students; negative values favour clinical students. The strongest effect was for Amboss showing a marked increase in its relevance among clinical students. *Sebentas:* study materials prepared by students from previous years.

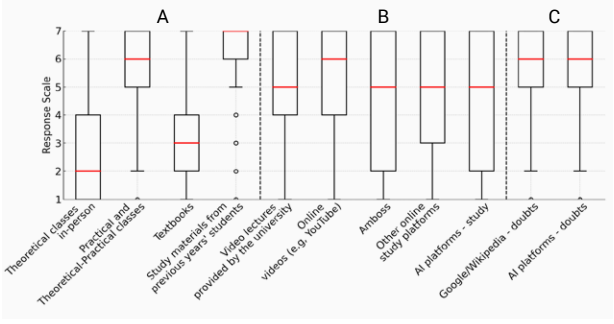

FIGURE S1 Methods used for learning during the semester

Study methods used by students to learn throughout the semester, rated on a scale from 1 ("Not relevant at all") to 7 ("Extremely relevant"). Red horizontal lines indicate the medians. The first vertical bar separates traditional methods (A: theoretical classes, practical/theoretical-practical classes, textbooks, and *"Sebentas"*: study materials prepared by students from previous years) from online resources. The second bar divides study-focused platforms (B: video lectures, Amboss, and others) from doubt-resolution tools (C: Google/Wikipedia and AI platforms).

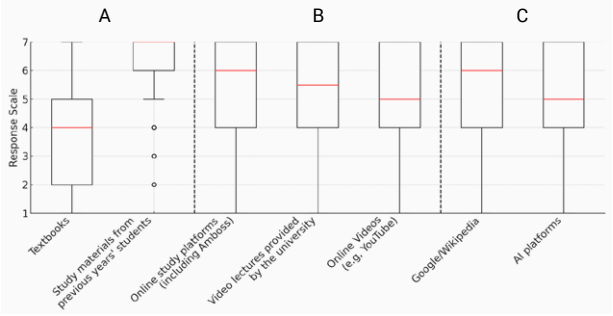

FIGURE S2 Methods used for exam preparation

Boxplot showing students' relevance ratings (1 = "Not relevant at all" to 7 = "Extremely relevant"). Red lines indicate medians. The first vertical bar separates traditional resources (A: textbooks and *"Sebentas"*: study materials prepared by students from previous years) from online resources. The second bar divides general online platforms (B: Amboss, video lectures, and YouTube) from tools primarily used for resolving doubts (C: Google/Wikipedia and AI platforms). Theoretical and practical classes were excluded, as they are unavailable during exams. Amboss was categorized under "Online study platforms," and AI-related options were combined into a single category, "AI platforms."

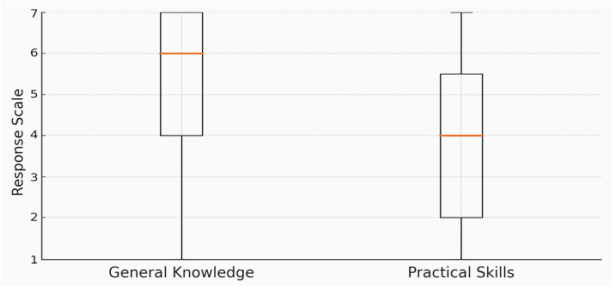

FIGURE S3 AI's utility for general knowledge vs. practical skills

Comparison of AI's usefulness in acquiring general knowledge and developing practical skills (scale: 1 = "Not useful at all" to 7 = "Extremely useful"). Red lines indicate median values.

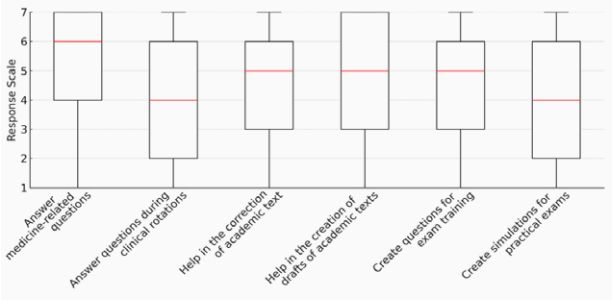

FIGURE S4 Perceived AI utility in medical education

Boxplot showing ratings for AI's usefulness in medical education (scale: 1 = "Not useful at all" to 7 = "Extremely useful"). Red lines indicate median values.

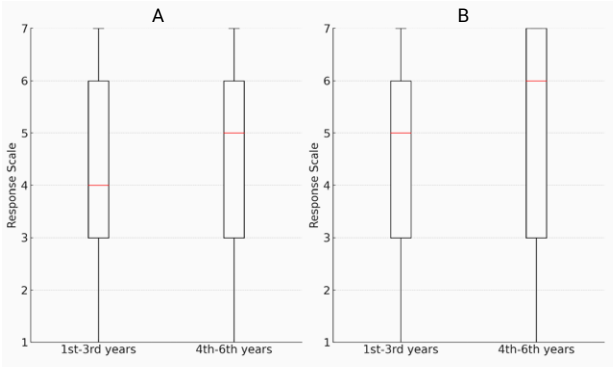

FIGURE S5 Students' perceived utility of AI in academic text tasks

(a) AI was rated moderately to highly useful for correcting academic texts, with slightly higher median ratings among clinical (4th–6th years) compared to preclinical (1st–3rd years) students.
(b) Clinical students also rated AI higher for drafting texts, though their responses showed greater variability than those of preclinical students.

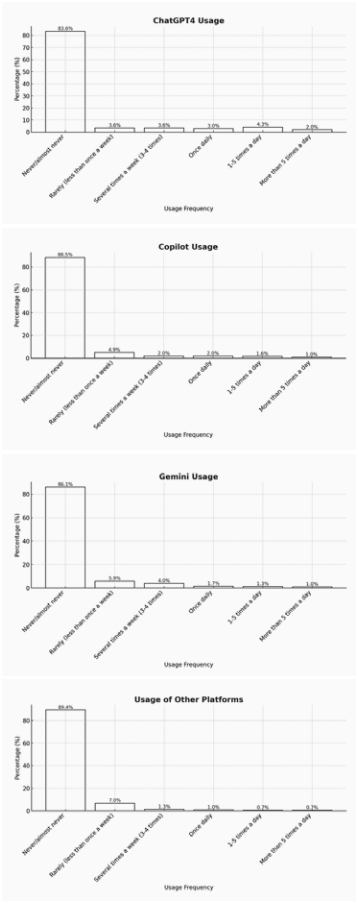

FIGURE S6 Frequency of AI use of different platforms

This figure presents the reported usage frequency of four AI platforms among students: ChatGPT 4 and 4o, Google Gemini, Microsoft Copilot, and Other AI platforms. These platforms typically require a paid subscription or have restricted access. The x-axis categorizes usage into six levels: 1-5 times per day, 1 time per day, more than 5 times per day, several times per week (3-4 times per week), rarely (less than once per week), and never/almost never. The y-axis represents the percentage of respondents for each category. The data indicate that almost 90% of students either never or very rarely use these platforms, highlighting their limited adoption among the surveyed population.

ACKNOWLEDGMENTS: *We thank all students who participated in this study for their time and contributions. The authors used artificial intelligence (AI) tools, including GPT-4 and DeepSeek (version/date of use: January 2025), strictly for language editing and clarity improvement. No AI-generated content was used for original analysis, interpretation of results, or drafting of scientific conclusions, except where explicitly described in the Methods section. No AI-generated creative or substantive text was included. Writing—Review & Editing: GPT-4 and DeepSeek were used at the sentence level for grammar and clarity improvements throughout the manuscript, without generating original arguments, analyses, or substantive content. Privacy and Security: no identifiable or sensitive data were shared with GPT-4 or DeepSeek during the editing process, and usage was conducted in accordance with institutional privacy and security guidelines. The final manuscript was reviewed and approved by all authors to ensure accuracy and adherence to ethical research and publication standards.*

DISCLOSURES: *All authors declare no conflicts of interest. No funding was received for this study.*

ETHICAL COMPLIANCE: *Ethical approval was granted by the CAML Ethical Committee (Ref. 193/24, September 2024).*

REFERENCES

1. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM; 2021. p. 610–23.
2. Rothman J. Why the Godfather of A.I. fears what he's built [Internet]. The New Yorker. 2023 [cited 2024 Dec 2]. Available from: https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai
3. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. 2023 Mar 20;
4. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023 Oct 1;13(1):16492.
5. NYU Grossman School of Medicine. AI in Medical Education: A Case Study. American Medical Association (AMA). 2023.
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023 Feb 9;2(2):e0000198.
7. 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 Feb 8;9:e45312.
8. Amboss. Powerful learning and clinical tools combined into one platform [Internet]. Amboss. 2023 [cited 2024 Dec 3]. Available from: https://www.amboss.com/int?_gl=1*1asi3jf*_up*MQ..*_gs*MQ..&gclid=CjwKCAiA9bq6BhAKEiwAH6bqoKuf4H9U1c1xMOoVph-Vayqc_4xNrFrVJr5JTWO1OzavDMm9FB7dDRoChOkQAvD_BwE
9. Lecturio. Achieve Mastery of Medical Concepts [Internet]. Lecturio GmbH. 2024 [cited 2024 Dec 3]. Available from: https://www.lecturio.com/medical/gateway/
10. Osmosis. Learn Visually with Osmosis [Internet]. Elsevier. 2024 [cited 2024 Dec 3]. Available from: https://www.osmosis.org/
11. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Communications Medicine. 2023 Oct 10;3(1):141.
12. Ramsøy TZ. The Misunderstood Musings of AI: Confabulation, Not Hallucination [Internet]. Thomas Ramsoy. 2024 [cited 2024 Nov 26]. Available from: https://thomasramsoy.com/index.php/2024/03/12/the-misunderstood-musings-of-ai-confabulation-not-hallucination/
13. Pedro AR, Dias MB, Laranjo L, Cunha AS, Cordeiro J V. Artificial intelligence in medicine: A comprehensive survey of medical doctor's perspectives in Portugal. PLoS One. 2023 Sep 7;18(9):e0290613.
14. Jackson P, Ponath Sukumaran G, Babu C, Tony MC, Jack DS, Reshma VR, et al. Artificial intelligence in medical education - perception among medical students. BMC Med Educ. 2024 Dec 1;24(1).
15. Stewart J, Lu J, Gahungu N, Goudie A, Fegan PG, Bennamoun M, et al. Western Australian medical students' attitudes towards artificial intelligence in healthcare. PLoS One. 2023 Aug 31;18(8):e0290642.
16. Weidener L, Fischer M. Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects. JMIR Med Educ. 2024 Jan 5;10:e51247.
17. Amiri H, Peiravi S, rezazadeh shojaee S sara, Rouhparvarzamin M, Nateghi MN, Etemadi MH, et al. Medical, dental, and nursing students' attitudes and knowledge towards artificial intelligence: a systematic review and meta-analysis. BMC Med Educ. 2024 Apr 15;24(1):412.
18. Kimmerle J, Timm J, Festl-Wietek T, Cress U, Herrmann-Werner A. Medical Students' Attitudes Toward AI in Medicine and their Expectations for Medical Education. J Med Educ Curric Dev. 2023 Jan 6;10.
19. Pereira SM, Grobusch MP, Hänscheid T. How a GPT-aided survey reveals a medical student's misunderstanding of the term "artemisinin resistance." New Microbes New Infect [Internet]. 2025 Feb [cited 2025 Jan 13];63:101552. Available from: https://www.sciencedirect.com/science/article/pii/S2052297524003366?via%3Dihub
20. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large Language Model Influence on Diagnostic Reasoning. JAMA Netw Open. 2024 Oct 28;7(10):e2440969.
21. Marsh EJ, Fazio LK. Learning errors from fiction: Difficulties in reducing reliance on fictional stories. Mem Cognit. 2006 Jul;34(5):1140–9.
22. Topol E, Rajpurkar P. When Doctors With A.I. Are Outperformed by A.I. Alone. The New York Times. 2025;
23. Ganjavi C, Eppler M, O'Brien D, Ramacciotti LS, Ghauri MS, Anderson I, et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLOS Digital Health. 2024 Sep 5;3(9):e0000596.
24. Tao W, Yang J, Qu X. Utilization of, Perceptions on, and Intention to Use AI Chatbots Among Medical Students in China: National Cross-Sectional Study. JMIR Med Educ. 2024 Oct 28;10:e57132–e57132.
25. Hirabayashi S, Jain R, Jurković N, Wu G. Harvard Undergraduate Survey on Generative AI. 2024 Jun 2;
26. FMUL. Plano de Atividades 2025 [Internet]. Direção de Serviços de Planeamento Estratégico, Qualidade e Inovação (DSPEQI), Direção de Serviços de Comunicação e Relações Públicas (DSCOM); 2024 [cited 2025 Feb 11]. Available from: https://www.medicina.ulisboa.pt/sites/default/files/2024-12/plano-de-atividades-fmul-2025-final.pdf
27. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International. 2024 Mar 3;61(2):228–39.
28. Susnjak T, McIntosh T. ChatGPT: The End of Online Exam Integrity? Educ Sci (Basel). 2024 Jun 17;14(6):656.
29. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting Training Data from Large Language Models. In: Proceedings of the 30th USENIX Security Symposium. USENIX Association; 2023. p. 2633–50.
30. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity. 2023 Sep 1;19(1):17.
31. Fosnacht K, Sarraf S, Howe E, Peck LK. How Important are High Response Rates for College Surveys? Rev High Ed. 2017;40(2):245–65.
32. Elder A. Medicine is difficult—there are no shortcuts. BMJ. 2024 Oct 3;q2163.