SHORT COMMUNICATION

Safeguarding Academic Integrity from AI-Assisted Cheating in Online Medical Examinations- Short Communication

Hamid Saboorifar[1,*], Afarin Rabbani[1,2]
[1]PhD Student, Faculty of Nursing and Midwifery, Mashhad University of Medical Sciences, Mashhad, Iran
[2]Department of Nursing, Torbat Heydariyeh University of Medical Sciences, Torbat Heydariyeh, Iran
*Faculty of Nursing and Midwifery, Azadi Sq. Mashhad, 9137913199 Iran
Tel: +989156513677
Email: hmydsbwryfr@gmail.com

Background: The rise of generative artificial intelligence (AI) has created unprecedented challenges for maintaining academic integrity in online medical examinations. AI tools such as large language models can produce human-like, contextually relevant responses that evade traditional plagiarism detection, undermining the validity of assessment results. Traditional assessments are increasingly vulnerable to AI-assisted cheating, threatening the reliability of evaluation outcomes.

Method: This short communication reviews recent literature on assessment redesign strategies and synthesizes key principles, including authentic clinical scenarios, localized data interpretation tasks, process-tracing methods, multimodal assessment, and open-book application-focused formats, to mitigate AI-related risks.

Results: Redesigned assessments emphasizing higher-order cognitive skills, context-bound reasoning, and real-time performance can limit AI's effectiveness in generating undetectable responses. Practical implementation guidelines are presented for both high-resource and resource-limited settings, with particular relevance to countries facing infrastructure limitations, such as Iran.

Conclusion: By re-engineering assessment tasks rather than relying solely on surveillance technologies, medical schools can safeguard examination validity, preserve public trust in medical qualifications, and ensure graduates possess the competencies essential for safe clinical practice. Future research should evaluate the reliability, feasibility, and scalability of redesigned assessment models across diverse medical education settings.

Key Words: Academic integrity, Artificial intelligence, Examination cheating, Online assessment, Medical education

حفاظت از سلامت آکادمیک در برابر تقلب با کمک هوش مصنوعی در آزمون‌های پزشکی آنلاین - گزارش کوتاه

زمینه و هدف: رشد سریع هوش مصنوعی مولد چالش‌های بی‌سابقه‌ای را در حفظ صداقت علمی در آزمون‌های آنلاین علوم پزشکی ایجاد کرده است. ابزارهایی مانند مدل‌های زبانی بزرگ قادرند پاسخ‌هایی مشابه انسان و متناسب با متن تولید کنند که می‌تواند از سامانه‌های متداول تشخیص سرقت علمی عبور کرده و اعتبار نتایج ارزشیابی را خدشه‌دار سازد.

روش: این نوشتار کوتاه بر «بازطراحی نظام ارزیابی» به‌عنوان مؤثرترین و مقیاس‌پذیرترین راهکار برای مقابله با تقلب مبتنی بر هوش مصنوعی تمرکز دارد. با تکیه بر مطالعات اخیر، اصول کلیدی این رویکرد شامل استفاده از سناریوهای بالینی واقعی، تکالیف تحلیلی مبتنی بر داده‌های بومی، روش‌های ردیابی فرایند ارزشیابی چندوجهی و قالب‌های باز و کاربردمحور معرفی می‌شود.

یافته‌ها: ارزیابی‌های بازطراحی‌شده که بر مهارت‌های شناختی سطح بالا، استدلال مبتنی بر زمینه و عملکرد در زمان واقعی تأکید دارند، می‌توانند اثربخشی هوش مصنوعی را در ایجاد پاسخ‌های غیرقابل تشخیص محدود کنند. دستورالعمل‌های اجرایی عملی برای هر دو محیط با منابع بالا و منابع محدود، با ارتباط ویژه با کشورهایی که با محدودیت‌های زیرساختی مانند ایران مواجه هستند، ارائه شده است.

نتیجه گیری: با مهندسی مجدد وظایف ارزیابی به جای تکیه صرف بر فناوری‌های نظارتی، دانشکده‌های پزشکی می‌توانند اعتبار آزمون را حفظ کنند، اعتماد عمومی به مدارک پزشکی را حفظ کنند و اطمینان حاصل کنند که فارغ‌التحصیلان از شایستگی‌های لازم برای عملکرد بالینی ایمن برخوردارند. پژوهش‌های آینده باید قابلیت اعتماد، امکان‌پذیری و مقیاس‌پذیری این الگوهای بازطراحی‌شده را در زمینه‌های متنوع آموزش پزشکی مورد ارزیابی قرار دهند.

واژه های کلیدی: صداقت تحصیلی، هوش مصنوعی، تقلب در آزمون، ارزیابی آنلاین، آموزش پزشکی

حماية النزاهة الأكاديمية من الغش بمساعدة الذكاء الاصطناعي في الامتحانات الطبية عبر الإنترنت - رسالة قصيرة

الخلفية: خلق صعود الذكاء الاصطناعي التوليدي تحديات غير مسبوقة للحفاظ على النزاهة الأكاديمية في الفحوصات الطبية عبر الإنترنت. يمكن لأدوات الذكاء الاصطناعي، مثل نماذج اللغة الكبيرة، إنتاج استجابات شبيهة بالإنسان، وذات صلة بالسياق، تتجنب الكشف التقليدي عن الانتحال، مما يُقوّض صحة نتائج التقييم. أصبحت التقييمات التقليدية أكثر عرضة للغش بمساعدة الذكاء الاصطناعي، مما يهدد موثوقية نتائج التقييم.

الطريقة: تستعرض هذه الرسالة الموجزة الأدبيات الحديثة حول استراتيجيات إعادة تصميم التقييمات، وتُلخص المبادئ الرئيسية، بما في ذلك السيناريوهات السريرية الأصيلة، ومهام تفسير البيانات المحلية، وطرق تتبع العمليات، والتقييم متعدد الوسائط، وصيغ الكتاب المفتوح التي تركز على التطبيقات، للتخفيف من المخاطر المتعلقة بالذكاء الاصطناعي.

النتائج: يمكن للتقييمات المُعاد تصميمها، التي تُركز على المهارات المعرفية العليا، والتفكير المرتبط بالسياق، والأداء الفوري، أن تحد من فعالية الذكاء الاصطناعي في توليد استجابات غير قابلة للكشف. تُقدم إرشادات تطبيقية عملية لكل من البيئات عالية الموارد والمحدودة الموارد، مع أهمية خاصة للدول التي تواجه قيوداً في البنية التحتية، مثل إيران. الخلاصة: من خلال إعادة هندسة مهام التقييم بدلاً من الاعتماد كلياً على تقنيات المراقبة، يمكن لكليات الطب حماية صحة الامتحانات، والحفاظ على ثقة الجمهور بالمؤهلات الطبية، وضمان امتلاك الخريجين للكفاءات الأساسية للممارسة السريرية الآمنة. ينبغي أن تُقيّم الأبحاث المستقبلية موثوقية نماذج التقييم المُعاد تصميمها وجدواها وقابليتها للتوسع في مختلف بيئات التعليم الطبي.

الخلاصة: يُعد استخدام الذكاء الاصطناعي التوليدي استراتيجية فعّالة لتحسين الكفاءة التشخيصية في الإرشاد السريري لطلاب الطب.

الكلمات المفتاحية: النزاهة الأكاديمية، الذكاء الاصطناعي، الغش في الامتحانات، التقييم عبر الإنترنت، التعليم الطبي

آن لائن طبی امتحانات میں AI کی مدد سے دھوکہ دہی سے تعلیمی سلمیت کی حفاظت کرنا- مختصر مواصلات

پس منظر: تخلیقی مصنوعی ذہانت (AI) کے عروج نے آن لائن طبی امتحانات میں تعلیمی سالمیت کو برقرار رکھنے کے لیے بے مثال چیلنجز پیدا کیے ہیں۔ اے آئی ٹولز جیسے کہ بڑے لینگویج ماڈلز انسانی جیسا، سیاق و سباق سے متعلقہ ردعمل پیدا کر سکتے ہیں جو روایتی سرقہ کی کھوج سے بچتے ہیں، تشخیص کے نتائج کی صداقت کو کم کرتے ہیں۔ روایتی جائزے AI کی مدد سے ہونے والی دھوکہ دہی کے لیے تیزی سے خطرے سے دوچار ہیں، جس سے تشخیص کے نتائج کی وشوسنییتا کو خطرہ ہے۔

طریقہ: یہ مختصر مواصلت تشخیص کو دوبارہ ڈیزائن کرنے کی حکمت عملیوں پر حالیہ لٹریچر کا جائزہ لیتی ہے اور AI سے متعلقہ خطرات کو کم کرنے کے لیے کلیدی اصولوں کی ترکیب کرتی ہے، بشمول مستند طبی منظرنامے، مقامی ڈیٹا کی تشریح کے کام، عمل کا پتہ لگانے کے طریقے، ملٹی موڈل اسیسمنٹ، اوراوپن بک ایپلیکیشن فوکسڈ فارمیٹس۔

نتائج: اعلیٰ ترتیب کی علمی مہارتوں، سیاق و سباق کے پابند استدلال، اور حقیقی وقت کی کارکردگی پر زور دینے والے دوبارہ ڈیزائن کیے گئے جائزے ناقابل شناخت ردعمل پیدا کرنے میں AI کی تاثیر کو محدود کر سکتے ہیں۔ عملی نفاذ کے رہنما خطوط اعلی وسائل اور وسائل کی محدود ترتیبات دونوں کے لیے پیش کیے گئے ہیں، خاص طورپر ان ممالک کے لیے جو بنیادی ڈھانچے کی حدود کا سامنا کر رہے ہیں، جیسے کہ ایران۔

نتیجہ: مکمل طور پر سرویلنس ٹیکنالوجیز پر انحصار کرنے کے بجائے دوبارہ انجینئرنگ کی تشخیص کے کاموں سے، میڈیکل اسکول امتحان کی درستگی کی حفاظت کر سکتے ہیں، طبی قابلیت پر عوام کے اعتماد کو برقرار رکھ سکتے ہیں، اور اس بات کو یقینی بنا سکتے ہیں کہ فارغ التحصیل افراد محفوظ طبی مشق کے لیے ضروری صلاحیتوں کے حامل ہوں۔ مستقبل کی تحقیق کو طبی تعلیم کی متنوع ترتیبات میں دوبارہ ڈیزائن کردہ تشخیصی ماڈلز کی وشوسنییتا، فزیبلٹی، اور اسکیل ایبلٹی کا جائزہ لینا چاہیے۔

کلیدی الفاظ: تعلیمی سالمیت، مصنوعی ذہانت، امتحان میں دھوکہ دہی، آن لائن تشخیص، طبی تعلیم

INTRODUCTION

Academic integrity is a foundational requirement for medical education because the authenticity of assessment underpins the competence and trustworthiness of future healthcare professionals and, ultimately, public safety. Recent analyses emphasize that integrity in medical training influences not only scholarly standards but also downstream clinical performance and patient outcomes (1,2). Maintaining credible assessment practices has therefore become a priority for medical schools worldwide.

The rapid expansion of online assessment modalities following the COVID-19 pandemic, together with the rise of powerful generative AI systems, has created novel and urgent threats to assessment validity. Large language models and related generative tools can produce human-like, contextually coherent answers to a wide range of clinical and scientific prompts, thereby challenging the ability of conventional online examinations and plagiarism detectors to distinguish original student work from AI-assisted output (3,4). Concurrently, evidence from recent literature reviews indicates that existing anti-cheating infrastructures—remote proctoring and traditional similarity detection—are often insufficient to detect sophisticated, AI-mediated forms of misconduct (5).

Given these pressures, this short communication narrows its scope to a single, immediately actionable claim: deliberate redesign of assessment format and content is the most effective and scalable strategy to preserve academic integrity in online medical examinations in the era of generative AI. We briefly review the rationale, summarize supporting evidence, and provide concise, practical recommendations for implementation—with attention to applicability in resource-limited or infrastructure-constrained settings.

METHODS

This short communication is based on a targeted narrative review of recent literature (2020–2025) on academic integrity and assessment redesign in medical education. Relevant publications were identified through Google Scholar and PubMed using keywords such as "Academic integrity", "Artificial intelligence", "Examination cheating", "Online assessment", "Medical education". Key concepts and practical strategies were synthesized to propose a framework for mitigating AI-related risks in both high- and low-resource contexts.

Focused discussion — assessment redesign as the primary strategy

Rationale: why format and task design matter

Generative AI undermines assessments principally when evaluation tasks reward outputs that AI can produce reliably (e.g., short factual answers, predictable essay prompts, or decontextualized problem solving). Conversely, assessments that require situated reasoning, real-time demonstration of skills, integration of tacit clinical judgement, or personalized reflection create cognitive demands that are more difficult for off-the-shelf AI systems to replicate convincingly. Recent syntheses and guidance documents highlight that transforming what is assessed—moving from decontextualized recall to authentic, applied tasks—reduces the information advantage that AI provides and increases the signal-to-noise ratio for authentic student performance (6,7).

Evidence from systematic and narrative reviews supports the proposition that task design reduces opportunities for undetectable misconduct. Chiang et al. (2022) concluded that online assessment integrity improves when tasks are authentic and application-oriented rather than rote or easily automatable (8). Singh and colleagues (2024) summarized institutional strategies and found that assessment redesign—particularly adoption of scenario-based, open-resource, and multimodal evaluations—consistently appeared in successful integrity programs across diverse higher-education contexts (9). These findings align with broader observations that proctoring and similarity detection alone cannot recover validity once assessment tasks remain unchanged (5,6).

Concrete redesign principles and examples

Below are practical principles derived from the reviewed literature and mapped to implementable formats that reduce vulnerability to AI-assisted cheating:

1. Prioritize authenticity and clinical context.

Replace isolated recall questions with brief, realistic clinical vignettes that require interpretation of incomplete data, prioritization of differential diagnoses, or justification of management choices. Authentic scenarios force candidates to integrate knowledge with clinical reasoning and to state specific, context-bound decisions that are harder for generic AI prompts to mimic convincingly (8,9).

2. Emphasize data-driven interpretation and localized materials.

Provide bespoke datasets (laboratory output, localized epidemiologic figures, anonymized charts) or images created for the exam session and ask candidates to interpret or act on that material. These items reduce the utility of generic AI responses because the stimulus is unique to the assessment and cannot be retrieved from public corpora (8).

3. Use structured free-response with process tracing.

Ask examinees to show intermediate steps, reasoning pathways, or to record brief time-stamped annotations explaining how they reached an answer. Process-tracing increases detectability of externally generated answers because AI outputs

tend to be concise final products lacking idiosyncratic intermediate reasoning (9).

4. Integrate multimodal assessment components.

Combine written tasks with synchronous oral examinations, short practical stations, or workplace-based assessments (mini-CEX, directly observed procedural skills). Where feasible, include a short viva or oral defense of written answers; in many settings, a focused oral component is disproportionately effective at revealing authentic competence (9).

5. Favor open-book, application-focused formats over closed-book recall.

When the emphasis is on clinical problem solving rather than memorization, open-resource examinations neutralize the advantage of having access to information and instead reward skills in retrieval, synthesis, and judgement—domains where AI assistance may be less decisive when tasks are well-designed (9,8).

6. Blueprint assessment to higher-order cognitive levels.

Align items to Bloom-type objectives emphasizing analysis, synthesis, and evaluation. Well-blueprinted exams make it harder to substitute AI answers for genuine student ability because prompts require individualized application, not generic statements (9).

Operational considerations for implementation

Implementation of assessment redesign must be pragmatic and sensitive to resource constraints, particularly in contexts with limited proctoring infrastructure or bandwidth. The following operational recommendations aim to balance feasibility with integrity:

• Pilot and scale incrementally. Run small, discipline-specific pilots (e.g., a clinical reasoning section within an existing exam) to evaluate logistics, faculty workload, and psychometric properties before broad adoption. Use pilot data to refine rubrics and examiner training.

• Invest in faculty development. Redesign requires faculty who can craft authentic scenarios, construct reliable rubrics, and conduct high-quality oral or workplace assessments. Targeted training workshops and sample item banks accelerate capacity building (10).

• Leverage existing clinical placements. Workplace-based assessments and direct observations in clinical settings are low-technology but high-value alternatives to remote proctoring for core clinical competencies. When clinical placements exist, structured low-stakes observations can complement summative assessments.

• Use short-duration synchronous tasks. Time-limited, supervised short tasks (e.g., 20–30 minute case analyses with live submission) reduce opportunities for external assistance and are manageable for examiners and IT systems.

• Combine technological supports judiciously. Proctoring tools and AI-detection algorithms can supplement redesign but should not be relied upon as primary defenses; they are best applied to monitor process variables or flag anomalies for targeted human review rather than as sole arbiters of misconduct (5).

Addressing likely challenges

Resistance to change, perceived faculty burden, and concerns about standardization are predictable barriers. The literature indicates several mitigation strategies: use shared item templates to reduce preparation time, adopt structured scoring rubrics to enhance inter-rater reliability, and phase implementation so faculty and students adapt progressively (9). Where scale is an issue, a mixed approach—retaining computerized, auto-scored items for foundational knowledge while shifting assessment of reasoning and clinical application to human-rated authentic tasks—optimizes resource use and preserves validity.

Contextualizing for resource-limited settings (example: Iran)

In contexts where internet reliability, bandwidth, and indigenous detection tools are limited, assessment redesign is particularly attractive because it relies primarily on pedagogic engineering rather than heavyweight surveillance technology. Localized items, workplace-based assessments, and brief oral defenses are feasible even under infrastructural constraints and can be implemented with modest investment in faculty training and secure scheduling. Empirical educational interventions in regionally comparable settings have shown that targeted workshops and virtual training can improve students' attitudes and knowledge around academic honesty—an important complement to task redesign (10). Importantly, redesign efforts should be embedded within institutional policy revisions that clarify expectations, specify consequences, and promote a culture of integrity.

Measuring success and next steps

Institutions should evaluate redesign interventions using measurable outcomes: changes in exam reliability and validity metrics, rates of flagged misconduct, student and faculty perceptions of fairness, and downstream indicators such as clinical performance in objective structured formats. Mixed-methods evaluation (quantitative psychometrics plus qualitative feedback) will identify unintended consequences and guide refinements. Where possible, sharing item banks, rubrics, and implementation experiences across institutions can accelerate best practice uptake while preserving confidentiality of assessment content.

CONCLUSION

Generative AI poses a structural challenge to the validity of many existing online medical examinations because it can produce superficially plausible responses to prompts that reward decontextualized output. The most robust and scalable countermeasure is to redesign what is assessed: shift emphasis from recall to authentic, context-bound tasks that require process-based reasoning, clinical judgement, and demonstration of applied skills. Assessment redesign reduces the advantage conferred by AI, increases the detectability of inauthentic work, and produces richer information about learners' readiness for clinical practice. To operationalize this strategy institutions should pilot scenario-based items, integrate short synchronous oral or workplace assessments, invest in faculty development, and evaluate outcomes with clear psychometric and qualitative metrics. In resource-constrained settings, pedagogically focused redesign offers a feasible path to preserving integrity without reliance on expensive surveillance technologies. Future research must document the comparative effectiveness, reliability, and scalability of redesigned assessment models and develop consensus standards for their use in high-stakes medical examinations.

Ethical Considerations

Ethical issues including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc. have been completely observed by the authors.

ACKNOWLEDGMENT

The authors gratefully acknowledge the researchers whose published work informed this focused communication.

Financial Support: None.

Conflict of Interest: The authors declare that there are no conflicts of interest in the present study.

REFERENCES

1. Alsharefeen R, Sayari N. Examining academic integrity policy and practice in the era of AI: a case study of faculty perspectives. Front Educ. 2025;10.
2. Moya Figueroa B, Eaton S, Pethrick H, Hayden A, Brennan R, Wiens J, McDermott B. Academic Integrity and Artificial Intelligence in Higher Education (HE) Contexts: A Rapid Scoping Review. Canadian Perspectives on Academic Integrity. 2024;7.
3. Ekaterina K, Ana M, Maia Z. Academic Integrity Within the Medical Curriculum in the Age of Generative Artificial Intelligence. Health Sci Rep. 2025;8(2):e70489.
4. Jaleel A, Aziz R, Farid G, Bashir MZ. The impact of ChatGPT on academic integrity in medical education: a developing nation perspective. Front Educ. 2025;Volume 10 - 2025.
5. Sozon M, Mohammad Alkharabsheh OH, Fong PW, Chuan SB. Cheating and plagiarism in higher education institutions (HEIs): A literature review. F1000Res. 2024;13:788.
6. Malekar O, Karale A, Pawara P, Gharat R. Real-Time Exam Monitoring. International Journal of Advanced Research in Science, Communication and Technology. 2025:288-94.
7. Singh J, Mishra N, Mishra K. Promoting Academic Integrity: Strategies and Challenges in Higher Education. 2024. p. 191-224.
8. Chiang F, Zhu D, Yu W. A systematic review of academic dishonesty in online learning environments. J Comput Assist Learn. 2022;38:907-28.
9. Na SJ, Yoon BY, Yeo S. The change of the COVID-19 pandemic on medical education in Korea: a national survey of medical schools. Korean J Med Educ. 2023;35(4):349-61.
10. Nikjo P, Vizeshfar F, Zarifsanayee N. Comparison of the effect of training academic honesty using two workshop and virtual training methods on the knowledge and attitude of M.S. students. J Educ Health Promot. 2021;11(4):1-7.