# Legal aspects of generative artificial intelligence and large language models in examinations and theses


## Abstract

The high performance of generative artificial intelligence (AI) and large language models (LLM) in examination contexts has triggered an intense debate about their applications, effects and risks. What legal aspects need to be considered when using LLM in teaching and assessment? What possibilities do language models offer?

Statutes and laws are used to assess the use of LLM:

– University statutes, state higher education laws, licensing regulations for doctors
– Copyright Act (UrhG)
– General Data Protection Regulation (DGPR)
– AI Regulation (EU AI Act)

LLM and AI offer opportunities but require clear university frameworks. These should define legitimate uses and areas where use is prohibited. Cheating and plagiarism violate good scientific practice and copyright laws. Cheating is difficult to detect. Plagiarism by AI is possible. Users of the products are responsible.

LLM are effective tools for generating exam questions. Nevertheless, careful review is necessary as even apparently high-quality products may contain errors. However, the risk of copyright infringement with AI-generated exam questions is low, as copyright law allows up to 15% of protected works to be used for teaching and exams.

The grading of exam content is subject to higher education laws and regulations and the GDPR. Exclusively computer-based assessment without human review is not permitted. For high-risk applications in education, the EU's AI Regulation will apply in the future.

When dealing with LLM in assessments, evaluation criteria for existing assessments can be adapted, as can assessment programmes, e.g. to reduce the motivation to cheat. LLM can also become the subject of the examination themselves. Teachers should undergo further training in AI and consider LLM as an addition.


**Keywords:** assessment, AI, large language models, legal framework


**Maren März**[1]
**Monika Himmelbauer**[2]
**Kevin Boldt**[3]
**Alexander Oksche**[4,5]

1 Charité – University Medicine Berlin, AG Progress Test Medicine, Teaching Division, Berlin, Germany

2 Medical University of Vienna, Teaching Centre, Vienna, Austria

3 The State Commissioner for Data Protection and Freedom of Information Rhineland-Palatinate, Mainz, Germany

4 Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), Mainz, Germany

5 Justus Liebig University Giessen, Rudolf Buchheim Institute for Pharmacology, Giessen, Germany


## Introduction

Artificial Intelligence (AI) is one of the key technologies of the fourth industrial revolution, which has the potential to fundamentally change industries and societies through global networking, digitalisation and the merging of the physical, digital and biological worlds [1].

Generative Artificial Intelligence (GAI) such as Large Language Models (LLM) is reaching a level of maturity that will impact healthcare. It could soon contribute to medical practice and empower patients to systematically shape their healthcare [2], [3], [4], [5], [6], [7], [8]. The rapid development, adoption and use of AI technologies in healthcare requires healthcare professionals to master experimental techniques, even if they are not yet recognised as standard [9].

GAI uses deep learning for content creation. LLM process natural language. They generate human-like text based on statistical principles that calculate the probability of a word or character depending on the context [10], [11], [12], [13]. Models such as ChatGPT are optimised for dialogue using reinforcement learning with human feedback (RLHF) [14], [15], [16]. LLM are used for translation and content production, automating literature reviews, identifying relevant studies, extracting key findings [17], [18], facilitating information retrieval and knowledge discovery, and providing decision support [19], [20].

LLM achieve remarkable exam results: ChatGPT passed the United States Medical Licensing Examination [21] and outperformed most students on the German progress test medizin [22]. LLM outperformed first and second year students on free-text clinical reasoning exams [23], scored 75% on the open-ended ENT and head and neck surgery specialist exam [24], 83% on a simulated 500-question written neurosurgery exam [25], and around 60% on the European core cardiology exam [26]. GPT-4

significantly outperformed previous models such as GPT-3 and GPT-3.5 in all areas analysed, demonstrating the rapid evolution of LLM [23], [24], [25], [27], [28]. GPT-3 was in the bottom 10% of US uniform bar examination graduates, while GPT-4 was in the top 10% [15], [27].
Additionally, there are challenges and limitations. The quality of the underlying training data can lead to discriminatory, unfair, and inaccurate content [29]. Training data should be accurate, complete, up-to-date, representative, and free from historical bias; however, these characteristics are often not fully known and therefore difficult to assess [29], [30]. In rapidly developing areas, data may also have limited public availability. LLM then generate plausible-sounding but incorrect answers, which are known as "hallucination" [31]. Previous measures such as retrieval LLM (RAG) reduce erroneous results, but do not completely prevent them [17], [32], [33]. Therefore, it is essential to subject the generated content to careful scrutiny [20], [34], [35]. Another weakness is the lack of transparency in LLM decision-making processes. These limitations have prompted a comprehensive debate about the applications, effects, and risks associated with these technologies [23], [36], [37], [38], [39].
The issue of examinations is particularly prominent, especially where examination systems are centred on written forms [40], [41]. With the rise of online examinations, there is growing concern about academic misuse, fuelled by anonymity, lack of supervision and access to electronic texts [42], [43], and LLM exacerbates existing challenges [40], [41], particularly for written work such as assignments, bachelor or master theses and dissertations. This is a complex issue, not only in terms of content, but also from a legal perspective. The following aspects need to be considered:

## 1. University statutes, higher education laws of the federal states, German Research Foundation guidelines, licensing regulations for doctors (AO)

Universities regulate examination requirements and procedures in study and/or examination regulations. They contain provisions on failures, breaches of regulations, performance assessment and grading, and can regulate the use of aids and define the use of unauthorised aids as cheating [44]. The AO (2002) leaves the decision on the consequences of violations of regulations or attempts to cheat in examinations to the discretion of the relevant state examination office (cf. §§ 14 para. 5, 15 para. 6).
The DFG guidelines for *safeguarding* good scientific practice apply to all researchers engaged in projects funded by the German Research Foundation (DFG). Furthermore, these guidelines are intended for implementation by universities and research institutions in Germany, which are expected to incorporate them into their own regulations [45].

## 2. Copyright act (UrhG)

The legal framework governing copyright is based on an EU directive that has been transposed into national legislation by each member state. It protects personal intellectual creations, as set forth in Section 2 (2) of the German Copyright Act (UrhG). An author is always a natural person, that is to say, a human being. This confers upon them the exclusive right to use their work. The extent to which AI-generated output is protected by copyright is contingent upon the degree to which the individual utilises the computer as a technical aid. (cf. Dreier/Schulze/Schulze UrhG Section 2 para. 8) [41], [42].

## 3. General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) is directly applicable in all EU member states. It regulates the protection of personal data and the free movement of data, as well as protecting the fundamental rights and freedoms of natural persons. Data processing must be legally justified (Art. 1, Art. 5 para. 1 lit. A, 6 para. 1 GDPR). Individuals whose data is processed have certain rights, including the controller's obligation to provide information and the right to access that information (Art. 13, 14, 15 GDPR). Furthermore, the regulation applies to companies outside the EU that process the data of EU citizens, in accordance with the "marketplace principle" (Art. 3 para. 2 GDPR).

## 4. AI regulation (EU AI Act)

The AI Regulation establishes a legal framework for trustworthy AI. Its objectives include ensuring security, transparency, traceability, non-discrimination, and environmental friendliness. It was adopted by the European Parliament on 13 March 2024 and will apply in all EU member states from 2026. AI systems are categorised into four risk classes: unacceptable risk (prohibited), high, low, and minimal risk. "General-purpose AI systems" (GPAI), which in principle include many LLM, are initially classified as limited risk and must fulfil transparency and documentation obligations and a copyright policy (Art. 52, 52c AI Regulation) [46]. GPAI with systemic risk are subject to additional requirements and need to be registered [47], [48]. In addition, high-risk AI systems must implement measures such as supervision, quality and risk management, extensive documentation, and rigorous data quality and system security standards [49].

# Consideration of the legal aspects in detail

## Significance for examination candidates

The DFG guidelines ensure academic integrity in teaching and research [45]. University statutes address academic offences such as plagiarism and cheating.

Plagiarism is an offence against good scientific practice and copyright law when works are used without appropriate attribution [38], [44], [50]. Plagiarism occurs when copyrighted texts are included in the product [40], [41], [51]. This can occur, for example, with "shake & paste" plagiarism, in which text passages from different sources are combined [38]. The use of AI-generated text certainly harbours the risk of plagiarism. The responsibility for these offences (i.e. the plagiarism) lies with the persons who adopt such texts without attribution [52]. The providers of the LLM are held accountable for any infringements of copyright law (ongoing legal proceedings in the USA against OpenAI and Google) [53].

Cheating is defined as presenting a work produced with unauthorised resources as one's own. This is against good academic practice and the study or examination regulations [38].

In the case of digital or paper-based examinations supervised (without aids) in the presence of an examiner, or in the case of decentralised digital examinations with proctoring and (in some cases) the use of secure browsers, the risk of cheating through the use of LLM is reduced [40], [44]. Oral and application-based examinations, such as OSCE, are less susceptible to the use of LLM. The design of such examinations is crucial to ensure that other forms of assessment error, such as subjectivity, do not become sources of error [40], [41], [54], [55], [56]. LLM are particularly problematic for written assessments that are completed independently and without supervision, such as homework [40], [41]. The reliability of synthetic text recognition remains variable, although it has improved significantly [57], [58], [59], [60], [61]. In addition, human judgement can be crucial. One university successfully rejected a Master's application because the essay was of unexpectedly high quality for a Bachelor's graduate and appeared to be AI-generated [62].

## Significance for universities and State Examination Office

### Creation of examination content and tasks

The use of LLM to create examination questions has been demonstrated [63], [64]. For example, ChatGPT has the potential to generate MCQ of comparable quality for final medical examinations in a short time [64], [65], [66]. However, it has been observed that questions that query higher levels of learning objectives show certain limitations [63]. In general, it is worth experimenting with different prompts. Prompt engineering represents a systematic approach to effective communication with LLM, exerting a significant influence on the resulting output [15], [66], [67]. Tasks created by AI must undergo a review process, as even linguistically well-constructed and plausible products, such as examination questions, can be erroneous [66]. In contrast, the risk of copyright infringement when utilising AI to create examination tasks is relatively low. This is due to the fact that, in accordance with Section 60a of the German Copyright Act (UrhG), up to 15% of a work may be made available for non-commercial purposes for illustration in class and for examinations.

### Evaluation of examination content and tasks

The criteria for examinations are generally laid down in higher education legislation and are specified in examination and study regulations. In the event that independent assessment by the examiner is envisaged, the examiner must assess the examination performance independently. Furthermore, a language model can only be used for support purposes. It should be noted that there is also no independent assessment if the assessment is not adopted exactly but is based solely on the AI-generated result [41]. From a data protection perspective, such an assessment is generally in violation of the ban on automated decision-making (Art. 22 GDPR). If a performance assessment is carried out by an AI, it is necessary to assume that a high-risk AI system is being used in accordance with the AI Regulation [49].

# Options for action

LLM represent a challenge, but also an opportunity for examination procedures. Faculties can react to AI developments without necessarily reverting to restrictive formats [41]. A general ban on AI applications is difficult to implement and hardly relevant; algorithms are already integrated into existing systems such as browsers or word processing programmes [41], [68]. It may therefore make sense to specifically authorise or integrate AI applications.

## Adaptation of regulations

Framework conditions should define under which conditions or for which purposes the use is legitimate and authorised and in which areas the use is prohibited [41]. Even if monitoring may be difficult, a clear explanation increases the binding nature and clarifies the consequences of rule violations, including in declarations of independence. Without these restrictions, no violation of the examination rules can be assumed [28], [41], [44], [52], [62]. Students should be aware that they bear responsibility for errors such as copyright infringements [28], [41], [52] (see attachment 1 for sources of sample texts and checklists).

## Adaptation of assessments

Evaluation criteria can be adapted for existing assessments. The critical use of sources and positioning in the specialised discourse could be weighted more heavily, linguistic correctness and expression could lose importance [41], [52]. LLM still produce incorrect or unweighted source references [17], [69]. A review would therefore appear to be useful.

## Adaptation of assessment programmes

The motivation to cheat can be reduced by changing the assessment programmes. Intrinsically motivated, performance-orientated students cheat less often than extrinsically motivated students who primarily want to pass [43], [70], [71] or are stressed [69]. One approach could be to reduce the stakes of individual examinations and at the same time increase the relevance of the content. Assessments should relate to real experiences and knowledge and therefore be authentic. Formats such as MC questions can certainly be used [41], [52], [72], [73]. A "moral anchor" and moral awareness can reduce cheating, promoted by exemplary teachers and training in self-awareness, ethics and decision-making [43], [68].

## LLM as an examination subject, LLM as a learning aid

In the spirit of "AI literacy", teachers should view LLM as a supplement and undergo continuous further training. In addition to knowledge about AI, the application, such as prompt generation, and the critical evaluation or validation of AI-generated texts could also be tested [18], [67], [74], [75].
LLM can identify knowledge gaps in formative testing environments through thematic text analysis of assessment data and provide individualised, timely and continuous feedback, similar to a constantly available tutor [74], [76].

# Limitations

The field of GAI is dynamic and developing rapidly. The performance capabilities and limitations mentioned could soon become obsolete. However, all developments must be in line with the applicable legal framework, including UrhG, DS-GVO and KI-VO. In the DACH region, this applies to Germany and Austria. For Switzerland, the legal framework is not known to the authors and must be checked before using GAIs.

# Acknowledgements

We would like to thank Daniel Bauer, Daniel Tolks and Katharina von der Wense for their critical reading and expert advice.

# Notes

## Translation

DeepL was used for English translation. For editing, DeepL, ChatGPT and Copilot were used (prompts: "Summarise the following text" and "Improve the following text").

## Authors' ORCIDs


- Maren März: [0000-0002-2661-5076]
- Monika Himmelbauer: [0000-0001-5516-1993]
- Alexander Oksche: [0000-0003-4592-1770]


# Competing interests

The authors declare that they have no competing interests.

# Attachments

Available from https://doi.org/10.3205/zma001702

1. Attachment_1.pdf (135 KB)
   Model guidelines, link collections

# References


1. Majumdar D, Banerji PK, Chakrabarti S. Technology Analysis & Strategic Management Disruptive technology and disruptive innovation?: ignore at your peril?! Technol Anal Strateg Manag. 2018;7325(11):1247-1255. DOI: 10.1080/09537325.2018.1523384
2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CM, Schwarzkopf SC, Unger M, Veldjuizen GP, Wagner SJ, Kather JN. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3(1):141. DOI: 10.1038/s43856-023-00370-1
3. Webster P. Six ways large language models are changing healthcare. Nat Med. 2023;29(12):2969-2971. DOI: 10.1038/s41591-023-02700-1
4. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Arcas BA, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Emturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. DOI: 10.1038/s41586-023-06291-2
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. DOI: 10.1038/s41591-023-02448-8
6. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, Zhang M, Long J, Ng AY, Rajpurkar P, Sinha SR. Development and Validation of an Artificial Intelligence System to Optimize Clinician Review of Patient Records. JAMA Netw Open. 2021;4(7):e2117391. DOI: 10.1001/jamanetworkopen.2021.17391

7. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic Reasoning Prompts Reveal the Potential for Large Language Model Interpretability in Medicine. NPJ Digt Med. 2024;7(1):20. DOI: 10.1038/s41746-024-01010-1

8. Li R, Kumar A, Chen JH. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? JAMA Intern Med. 2023;183(6):596-597. DOI: 10.1001/jamainternmed.2023.1835

9. Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. Lancet Digit Heal. 2020;2(3):e111-e122. DOI: 10.1016/S2589-7500(20)30023-6

10. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: FAccT 2021: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. March 2021. p.610-623. DOI: 10.1145/3442188.3445922

11. Bhavya B, Xiong J, Zhai C. Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT. arXiv. 2022. DOI: 10.48550/arXiv.2210.04186

12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv. 2020. DOI: 10.48550/arXiv.2005.14165

13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017:5999-6009. DOI: 10.48550/arXiv.1706.03762

14. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Conerly T, El-Showk S, Elhage N, Hatfield-Doods Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nanda N, Olsson C, Amodei D, Brown T, Clark J, McCandish S, Olah C, Mann B, Kaplan J. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv. 2022. DOI: 10.48550/arXiv.2204.05862

15. Open AI, Achiam J, Adler S, Agarwal S, et al. GPT-4 Technical Report. arXiv. 2023;4:1-100. DOI: 10.48550/arXiv.2303.08774

16. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. Training language models to follow instructions with human feedback. arXiv. 2022. DOI: 10.48550/arXiv.2203.02155

17. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 20239;11(6):887. DOI: 10.3390/healthcare11060887

18. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, Bright TJ, Tatonetti N, Won KJ, Gonzalez-Hernandez G, Moore JH. ChatGPT and large language models in academia: opportunities and challenges. BioData Min. 2023;16(1):20. DOI: 10.1186/s13040-023-00339-9

19. Thapa S, Adhikari S. ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. Ann Biomed Eng. 2023;51(12):2647-2651. DOI: 10.1007/s10439-023-03284-0

20. Watkins R. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. AI Ethics. 2023. DOI: 10.1007/s43681-023-00294-5

21. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Heal. 2023;2(2):e0000198. DOI: 10.1371/journal.pdig.0000198

22. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing?? Med Educ Online. 2023;28(1):2220920. DOI: 10.1080/10872981.2023.2220920

23. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, Chen JH. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. JAMA Intern Med. 2023;183(9):1028-1030. DOI: 10.1001/jamainternmed.2023.2909

24. Long C, Lowe K, Santos A dos, Zhang J, Alanazi A, O'Brien D, Wright E, Cote D. Evaluating ChatGPT-4 in Otolaryngology–Head and Neck Surgery Board Examination using the CVSA Model. medRxiv. 2023. DOI: 10.1101/2023.05.30.23290758

25. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, Asaad WF, Cielo D, Oyeles AA, Doberstein CE, Gokaslan ZL, Telfeian AE. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. Neurosurgery. 2023;93(6):1353-1365. DOI: 10.1227/neu.0000000000002632

26. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Hear J Digit Heal. 2023;4(3):279-281. DOI: 10.1093/ehjdh/ztad029

27. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the Bar Exam. SSRN Electron J. 2023:1-35. DOI: 10.2139/ssrn.4389233

28. FernUni. A teacher's guide to ChatGPT and remote assessments. FernUni.ch. 2023.

29. Neutatz F, Abedjan Z. Whats is "Good" Training Data. In: Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, Rostalski F, editors. Künstliche Intelligenz Wie gelingt eine vertrauenswürdige Verwendung. Tübingen: Mohr Siebeck; 2022. DOI: 10.1628/978-3-16-161299-2

30. Stoyanovich J, Howe B, Jagadish HV. Responsible data management. Proc VLDB Endow. 2020 ;13(12):3474-3488. DOI: 10.14778/3415478.3415570

31. Heaven WD. Geoffrey Hinton tells us why he's now scared of the tech he helped build. MIT Technol Rev. 2023. Zugänglich unter/available from: https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/

32. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, Berkowitz ST, Finn AP, Jahangir E, Scoville EA, Reese TX, Friedmann DL, Bastarache JA, van der Heijden YF, Wrigth JJ, Ye F, Carter N, Alexander MR, Choe JH, Chastain CA, Zic JA, Horst SN, Turker I, Agarwal R, Osmundson E, Idrees K, Kiernan CM, Padmanabhan C, Bailey CE, Schlegel CE, Chabless LB, Gibson MK, Osterman TJ, Wheless LE, Johnson DB. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023;6(10):e2336483. DOI: 10.1001/jamanetworkopen.2023.36483

33. George C, Stuhlmüller A. Factored Verification: Detecting and Reducing Hallucination in Summaries of Academic Papers. arXiv. 2023. DOI: https://doi.org/10.48550/arXiv.2310.10627

34. Huang J, Chen X, Mishra S, Zheng HS, Yu AW, Song X, Zhou D. Large Language Models Cannot Self-Correct Reasoning Yet. arXiv. 2023;1:1-19. DOI: 10.48550/arXiv.2310.01798

35. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023;307(2):e230163. DOI: 10.1148/radiol.230163

36. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. DOI: 10.1038/s41591-018-0300-7

37. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int. 2023;00(00):1-12. DOI: 10.1080/14703297.2023.2190148

38. Georg-August-Universität Göttingen. Handreichung zum Umgang mit Plagiaten für Lehrende an der Sozialwissenschaftlichen Fakultät der Georg-August-Universität Göttingen. Göttingen: Georg-August-Universität Göttingen; 2021. p.1-2.

39. Hochschulforum Digitalisierung. ChatGPT im Hochschulkontext – eine kommentierte Linksammlung. Essen: Hochschulforum Digitalisierung; 2023. Zugänglich unter/available from: https://hochschulforumdigitalisierung.de/chatgpt-im-hochschulkontext-eine-kommentierte-linksammlung/

40. Moritz S, Romeike B, Stosch C, Tolks D. Generative AI (gAI) in medical education: Chat-GPT DQG and co. GMS J Med Educ. 2023;40(4):Doc54. DOI: 10.3205/zma001636

41. Salden P, Leschke J. Didaktische und rechtliche Perspektiven auf KI-gestütztes Schreiben in der Hochschulbildung. Bochum: RUB; 2023. p.41. Zugänglich unter/available from: https://hss-opus.ub.ruhr-uni-bochum.de/opus4/frontdoor/deliver/index/docId/9734/file/2023_03_06_Didaktik_Recht_KI_Hochschulbildung.pdf

42. Stabsstelle IT-Recht der bayerischen staatlichen Universitäten und Hochschulen. Prüfungsrechtliche Fragen zu ChatGPT. Würzburg: Universität Würzburg; 2022. Zugänglich unter/available from: https://www.rz.uni-wuerzburg.de/fileadmin/42010000/2023/ChatGPT_und_Pruefungsrecht.pdf

43. Simkin MG, McLeod A. Why do college students cheat? J Bus Ethics. 2010;94(3):441-453. DOI: 10.1007/s10551-009-0275-x

44. Radcke A. Der Einsatz von KI in Hochschulprüfungen und dessen prüfungsrechtlichen Auswirkungen. Potsdam: Universität Potsdam; 2023. Zugänglich unter/available from: https://www.uni-potsdam.de/fileadmin/projects/zfq/Lehre_und_Medien/E-Assessment/Auswirkung_KI_auf_Pruefungen_20230524.pdf

45. Deutsche Forschungsgemeinschaft. Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Bonn: DFG; 2019. DOI: 10.5281/zenodo.3923601

46. European Commission. Artificial Intelligence – Questions and Answers. Brussels: European Commision; 2023. Zugänglich unter/available from: https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683

47. Europäisches Parlament. KI-Gesetz: erste Regulierung der künstlichen Intelligenz. Brüssel: Europäisches Parlament; 2023. Zugänglich unter/available from: https://www.europarl.europa.eu/news/de/headlines/society/20230601ST093804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz

48. Maximilian Borkowsky. Der EU AI Act: Was bedeutet er für künstliche Intelligenz in Unternehmen? 2023. Zugänglich unter/available from: https://www.melibo.de/blog/der-eu-ai-act-was-bedeutet-er-fur-kunstliche-intelligenz-in-unternehmen

49. Madiega TA. Artificial intelligence act. Brussels: Europäisches Parlemant; 2024. Zugänglich unter/available from: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792

50. Fishman T. “We know it when we see it ” is not good enough?: toward a standard definition of plagiarism that transcends theft , fraud , and copyright. Wollongong (NSW, Aust): University of Wollongong; 2009. p.28-30.

51. Khalil M, Er E. Will ChatGPT get you caught? Rethinking of Plagiarism Detection. arXiv. 2023. DOI:

52. Gimpel H, Ruiner C, Schoch M, Schoop M, Hall K, Eymann T, Röglinger M, Vandirk S, Lämmermann L, Urbach N, Mädche A, Decker S. Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education. Hohenheim: Universität Hohenheim; 2023. p.47.

53. Daniel J Sokolov. Large Language Models: US-Autoren verklagen OpenAI wegen Copyright-Verletzung. 2023. Zugänglich unter/available from: https://www.heise.de/news/Large-Language-Models-US-Autoren-verklagen-OpenAI-wegen-Copyright-Verletzung-9301736.html

54. Susnjak T. ChatGPT: The End of Online Exam Integrity? arXiv. 2022. DOI: 10.48550/arXiv.2212.09292

55. Davis MH, Karunathilake I. The place of the oral examination in today’s assessment systems. Med Teach. 2005;27(4):294-297. DOI: 10.1080/01421590500126437

56. Wass V, Van Der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945-949. DOI: 10.1016/S0140-6736(00)04221-5

57. Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S. Can AI-Generated Text be Reliably Detected? arXiv. 2023. DOI: 10.48550/arXiv.2303.11156

58. Krishna K, Song Y, Karpinska M, Wieting J, Iyyer M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. arXiv. 2023. DOI: 10.48550/arXiv.2303.13408

59. Wu J, Yang S, Zhan R, Yuan Y, Wong DF, Chao LS. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. arXiv. 2023. DOI: 10.48550/arXiv.2310.14724

60. Mo Y, Qin H, Dong Y, Zhu Z, Li Z. Large Language Model (LLM) AI Text Generation Detection based on Transformer Deep Learning Algorithm. Int J Eng Manag Res. 2024;14(2):154-159. DOI: 10.5281/zenodo.11124440

61. Nguyen TT, Hatua A, Sung AH. How to Detect AI-Generated Texts? In: 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE; 2023. p.0464–0471. DOI: 10.1109/UEMCON59035.2023.10316132

62. Zenthöfer J. Erstes Urteil zu ChatGPT an Hochschulen. Frankfurter Allgemeine. 2024. Zugänglich unter/available from: https://www.faz.net/aktuell/karriere-hochschule/erstes-urteil-zu-chatgpt-an-hochschulen-student-benutzte-ki-fuer-bewerbung-19564795.html

63. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. 2023;15(6):e40977. DOI: 10.7759/cureus.40977

64. Cheung BH, Lau GK, Wong GT, Lee EY, Kulkarni D, Seow CS, Wong R, Co MT. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S. A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023;18(8):e0290691. DOI: 10.1371/journal.pone.0290691

65. Klang E, Portugez S, Gross R, Kassif Lerner R, Brenner A, Gilboa M, Ortal T, Ron S, Robinzon V, Meiri H, Segal G. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ. 2023;23(1):772. DOI: 10.1186/s12909-023-04752-w

66. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions. Acad Med. 2024;99(5):508-512. DOI: 10.1097/ACM.0000000000005626

67. Heston TF, Khun C. Prompt Engineering in Medical Education. Int Med Educ. 2023;2(3):198-205. DOI: 10.3390/ime2030019

68. Crawford J, Cowling M, Allen KA. Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). J Univ Teach Learn Pract. 2023;20(3). DOI: 10.53761/1.20.3.02

69. Goodman RS, Patrinely JR, Osterman T, Wheless L, Johnson DB. On the cusp: Considering the impact of artificial intelligence language models in healthcare. Med. 2023;4(3):139-140. DOI: 10.1016/j.medj.2023.02.008

70. Anderman EM, Koenka AC. The Relation Between Academic Motivation and Cheating. Theory Pract. 2017;56(2):95-102. DOI: 10.1080/00405841.2017.1308172

71. Hsiao YP, Klijn N, Chiu MS. Developing a framework to re-design writing assignment assessment for the era of Large Language Models. Learn Res Pract. 2023;9(2):148-158. DOI: 10.1080/23735082.2023.2257234

72. Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment? J Learn Dev High Educ. 2023;27. DOI: 10.47408/jldhe.vi27.1009

73. Schuwirth LW, Van Der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011;33(6):478-485. DOI: 10.3109/0142159X.2011.565828

74. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. Acad Med. 2023;99(1):22-27. DOI: 10.1097/acm.0000000000005439

75. Busse B, Kleiber I, Eickhoff FC, Andree K. Hinweise zu textgenerierenden KI- Systemen im Kontext von Lehre und Lernen. Köln: Universität zu Köln; 2023. Zugänglich unter/available from: https://uni-koeln.sciebo.de/s/7uwYRX5a92eznVI#pdfviewer DOI: 10.13140/RG.2.2.35392.61449/1

76. Cardona MA, Rodríguez RJ, Ishmael K. Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations. US Dep Educ Off Educ Technol. 2023;(1):1-71. Zugänglich unter/available from: https://www2.ed.gov/documents/ai-report/ai-report.pdf

**Corresponding author:**
Maren März
Charité – University Medicine Berlin, AG Progress Test Medicine, Teaching Division, Charitéplatz 1, D-10117 Berlin, Germany, Phone: +49 (0)30/450-576047
maren.maerz@charite.de

***Please cite as***
*März Maren, Himmelbauer M, Boldt K, Oksche A. Legal aspects of generative artificial intelligence and large language models in examinations and theses. GMS J Med Educ. 2024;41(4):Doc47. DOI: 10.3205/zma001702, URN: urn:nbn:de:0183-zma0017020*

***This article is freely available from***
*https://doi.org/10.3205/zma001702*

***Received:*** *2024-04-11*
***Revised:*** *2024-07-01*
***Accepted:*** *2024-07-09*
***Published:*** *2024-09-16*

# Rechtliche Aspekte zu generativer Künstlicher Intelligenz und Large-Language-Modellen in Prüfungen und Abschlussarbeiten

**Maren März**[1]
**Monika Himmelbauer**[2]
**Kevin Boldt**[3]
**Alexander Oksche**[4,5]

1 Charité – Universitätsmedizin Berlin, AG Progress Test Medizin, Geschäftsbereich Lehre, Berlin, Deutschland

2 Medizinische Universität Wien, Teaching Center, Wien, Österreich

3 Der Landesbeauftragte für den Datenschutz und die Informationsfreiheit Rheinland-Pfalz, Mainz, Deutschland

4 Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), Mainz, Deutschland

5 Justus-Liebig-Universität Giessen, Rudolf-Buchheim-Institut für Pharmakologie, Giessen, Deutschland

## Zusammenfassung

Die hohe Leistungsfähigkeit von generativer Künstlicher Intelligenz (KI) und großen Sprachmodellen (LLM) in Prüfungskontexten hat eine intensive Debatte über ihre Anwendungen, Auswirkungen und Risiken ausgelöst. Welche rechtlichen Aspekte sind beim Einsatz von LLM in Lehre und Prüfungen zu berücksichtigen? Welche Chancen bieten Sprachmodelle?
Für die rechtliche Bewertung des Einsatzes von LLM finden Satzungen und Gesetze Anwendung:
– Universitäre Satzungen, Hochschulgesetze der Länder, Approbationsordnung für Ärzte
– Urheberrechtsgesetz (UrhG)
– Datenschutz-Grundverordnung (DS-GVO)
– KI-Verordnung (KI-VO)
LLM und KI bieten Chancen, erfordern aber klare universitäre Rahmenbedingungen. Diese sollten den legitimen Einsatz und die Bereiche, in denen die Nutzung untersagt ist, definieren. Täuschungen und Plagiate verstoßen gegen die wissenschaftliche Praxis und das UrhG. Eine Täuschung ist schwer nachzuweisen. Plagiate durch KI sind möglich. Nutzer*innen der Produkte sind in der Verantwortung.
LLM sind effektive Tools zur Generierung von Prüfungsfragen. Dennoch ist ein sorgfältiges Review notwendig, da selbst qualitativ hochwertig scheinende Produkte Fehler enthalten können. Das Risiko von Urheberrechtsverletzungen bei KI-generierten Prüfungsaufgaben ist hingegen gering, da das Urheberrecht den Einsatz geschützter Werke für Lehre und Prüfungen bis zu 15% erlaubt.
Die Bewertung von Prüfungsinhalten unterliegt Hochschulgesetzen und -ordnungen und der DSGVO. Eine ausschließlich computergestützte Bewertung ohne menschliche Überprüfung ist nicht zulässig. Für Hochrisiko-Anwendungen in der beruflichen Lehre findet künftig die KI-VO der EU Anwendung.
Im Umgang mit LLM in Prüfungen können Bewertungskriterien bestehender Prüfungen angepasst werden, aber auch Prüfungsprogramme, z.B. um die Täuschungsmotivation zu reduzieren. LLM können zudem selbst Gegenstand der Prüfung werden. Lehrende sollten sich in KI weiterbilden und LLM als Ergänzung betrachten.

**Schlüsselwörter:** Prüfungen, KI, große Sprachmodelle, rechtliche Rahmenbedingungen

## Einführung

Künstliche Intelligenz (KI) ist eine der Schlüsseltechnologien der vierten industriellen Revolution, die zur globalen Vernetzung, Digitalisierung und durch Verschmelzung der physischen, digitalen und biologischen Welt das Potenzial hat, Industrien und Gesellschaften grundlegend zu verändern [1].

Generative Künstliche Intelligenz (GKI) wie Large Language Models (LLM) erreicht einen Reifegrad, der das Gesundheitswesen beeinflussen wird. Sie könnte bald zur medizinischen Praxis beitragen und Patient*innen befähigen, ihre Gesundheitsversorgung systematisch mitzugestalten [2], [3], [4], [5], [6], [7], [8]. Die rasche Entwicklung, Einführung und Nutzung von KI-Technologien im Gesundheitswesen erfordert, dass das Gesundheitspersonal experimentelle Techniken beherrscht, auch wenn diese noch nicht als Standard anerkannt sind [9].

GKI nutzt Deep Learning zur Inhaltserstellung. LLM verarbeiten die natürliche Sprache. Sie erzeugen menschenähnliche Texte basierend auf statistischen Prinzipien, die die Wahrscheinlichkeit eines Wortes oder Zeichens in Abhängigkeit vom Kontext berechnen [10], [11], [12], [13]. Modelle wie ChatGPT werden durch Reinforcement Learning mit menschlichem Feedback (RLHF) für den Dialog optimiert [14], [15], [16]. LLM werden für Übersetzung und Inhaltsproduktion eingesetzt, automatisieren Literaturübersichten, identifizieren relevante Studien, extrahieren Schlüsselergebnisse [17], [18], fördern Informationsbeschaffung, Wissensentdeckung und bieten Entscheidungsunterstützung [19], [20].

LLM erzielen beachtliche Prüfungsergebnisse: ChatGPT bestand die United States Medical Licensing Examination [21] und übertraf die meisten Studierenden im deutschsprachigen Progress Test Medizin [22]. LLM übertrafen die Leistungen der Studierenden im ersten und zweiten Studienjahr in Freitextprüfungen im klinischen Denken [23], erreichten in der Fachärzt*innenprüfung für HNO und Kopf-Hals-Chirurgie mit offenen Fragen 75% [24], in einer simulierten schriftlichen neurochirurgischen Prüfung mit 500 Fragen 83% [25] und im Europäischen Examen in Core Cardiology ca. 60% [26]. GPT-4 übertraf frühere Modelle wie GPT-3 und GPT-3.5 in allen untersuchten Bereichen deutlich und zeigt die rasante Entwicklung von LLM [23], [24], [25], [27], [28]. GPT-3 lag in den unteren 10% der Absolvent*innen der Uniform Bar Examination in den USA, GPT-4 hingegen in den oberen 10% [15], [27].

Es gibt auch Herausforderungen und Einschränkungen: Umfang und Qualität der zugrunde liegenden Trainingsdaten können zu diskriminierenden, unfairen und falschen Inhalten führen [29]. Die Trainingsdaten sollten korrekt, vollständig, aktuell, repräsentativ und frei von historischen Verzerrungen sein, sind aber oft nicht vollständig bekannt und daher schwer zu beurteilen [29], [30]. In sich schnell entwickelnden Bereichen können Daten zudem nur begrenzt öffentlich verfügbar sein. LLM erzeugen dann plausibel klingende, aber inhaltlich falsche Antworten („Halluzination“) [31]. Bisherige Maßnahmen wie abfragende LLM (RAG) reduzieren fehlerhafte Ergebnisse, verhindern sie jedoch nicht vollständig [17], [32], [33]. Eine sorgfältige Prüfung der generierten Inhalte bleibt daher unverzichtbar [20], [34], [35]. Eine weitere Schwäche liegt in der mangelnden Transparenz der Entscheidungsprozesse von LLM. Diese Einschränkungen haben eine breite Debatte über Anwendungen, Auswirkungen und Risiken ausgelöst [23], [36], [37], [38], [39].

Das Thema Prüfungen nimmt einen besonders prominenten Platz ein, insbesondere wenn die Prüfungssysteme auf schriftliche Formen ausgerichtet sind [40], [41]. Mit der Zunahme von Online-Prüfungen wächst die Sorge vor akademischem Missbrauch, der durch Anonymität, mangelnde Aufsicht und Zugang zu elektronischen Texten begünstigt wird [42], [43] und LLM verstärken die bereits bestehenden Herausforderungen [40], [41], insbesondere bei schriftlichen Arbeiten wie Hausarbeiten, Bachelor- oder Masterarbeiten sowie Dissertationen. Dies ist nicht nur inhaltlich, sondern auch rechtlich komplex. Folgende Aspekte sind zu berücksichtigen:

## 1. Universitäre Satzungen, Hochschulgesetze der Länder, Leitlinien der Deutschen Forschungsgemeinschaft, Approbationsordnung für Ärzte (AO)

Hochschulen regeln Prüfungsanforderungen und -verfahren in Studien- und/oder Prüfungsordnungen. Sie enthalten Bestimmungen über Versäumnisse, Ordnungsverstöße, Leistungsbewertung und Notenbildung und können die Verwendung von Hilfsmitteln regeln und die Benutzung nicht zugelassener Hilfsmittel als Täuschung definieren [44]. Die AO (2002) legt die Entscheidung über Konsequenzen von Ordnungsverstößen oder Betrugsversuchen bei Prüfungen in das Ermessen des jeweils zuständigen Landesprüfungsamtes (vgl. §§ 14 Abs. 5, 15 Abs. 6).

Die DFG-Leitlinien zur Sicherung guter wissenschaftlicher Praxis gelten für alle Wissenschaftler*innen, die an von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekten beteiligt sind. Sie richten sich auch an Hochschulen und Forschungseinrichtungen in Deutschland, die diese Leitlinien in ihren eigenen Regelwerken umsetzen sollen [45].

## 2. Urheberrechtsgesetz (UrhG)

Das Urheberrecht basiert auf einer EU-Richtlinie, die von den Mitgliedsstaaten in nationales Recht umgesetzt wurde. Es schützt persönliche geistige Schöpfungen (vgl. § 2 Abs. 2 UrhG DE). Urheber*in ist grundsätzlich immer eine natürliche Person, also ein Mensch. Die Person hat das ausschließliche Recht zur Verwendung ihres Werkes. Ob KI-generierter Output urheberrechtlichen Schutz genießt, hängt im Wesentlichen davon ab, inwieweit sich der Mensch des Computers als technisches Hilfsmittel bedient (vgl. Dreier/Schulze/Schulze UrhG § 2 Rn. 8) [41], [42].

## 3. Datenschutz-Grundverordnung (DS-GVO)

Die Datenschutz-Grundverordnung (DSGVO) gilt unmittelbar in allen EU-Mitgliedsstaaten. Sie regelt den Schutz personenbezogener Daten und den freien Datenverkehr, schützt die Grundrechte und Grundfreiheiten natürlicher Personen. Die Datenverarbeitung muss rechtlich begründet sein (Art. 1, Art 5 Abs. 1 lit. A, 6 Abs. 1 DS-GVO). Betroffene haben Rechte wie Informationspflichten der Verantwortlichen und das Recht auf Auskunft (Art. 13, 14, 15 DS-GVO). Die Verordnung gilt auch für Unternehmen außerhalb der EU, die Daten von EU-Bürgern verarbeiten (Marktortprinzip - Art. 3 Abs. 2 DS-GVO).

## 4. KI-Verordnung (KI-VO)

Die KI-Verordnung schafft einen Rechtsrahmen für vertrauenswürdige KI. Sie soll Sicherheit, Transparenz, Nachvollziehbarkeit, Nicht-Diskriminierung und Umweltfreundlichkeit gewährleisten. Sie wurde am 13. März 2024 vom Europäische Parlament verabschiedet und gilt ab 2026 in allen EU-Mitgliedstaaten. KI-Systeme werden in vier Risikoklassen eingeteilt: unannehmbares Risiko (verboten), hohes, geringes und minimales Risiko. „General-purpose AI systems“ (GPAI), zu denen grundsätzlich viele LLM gehören, werden zunächst als begrenztes Risiko eingestuft und müssen Transparenz-, Dokumentationspflichten und eine Urheberrechtspolitik erfüllen (Art. 52, 52c KI-VO) [46]. GPAI mit systemischem Risiko unterliegen zusätzlichen Anforderungen und werden registriert [47], [48]. Hochrisiko-KI-Systeme erfordern zudem Maßnahmen wie Aufsicht, Qualitäts- und Risikomanagement, umfangreiche Dokumentation und hohe Anforderungen an Datenqualität und Systemsicherheit [49].

# Betrachtung der rechtlichen Aspekte im Einzelnen

## Bedeutung für Prüfungskandidat*innen

Die DFG-Leitlinien sichern die wissenschaftliche Integrität in Lehre und Forschung [45], universitäre Satzungen adressieren akademische Verstöße wie Plagiat und Täuschung.

Plagiate sind Verstöße gegen die gute wissenschaftliche Praxis und das Urheberrecht, wenn Werke ohne angemessene Zuordnung verwendet werden [38], [44], [50].

Plagiatserkennungssoftware hat Schwierigkeiten bei der Identifizierung von KI-generierten Texten, denn generative Modelle erzeugen auch bei identischen Eingabeaufforderungen unterschiedliche Texte [40], [41], [51]. Ein Plagiat liegt vor, wenn urheberrechtlich geschützte Texte im Produkt enthalten sind. Dies kann zum Beispiel bei „Shake & Paste“-Plagiaten auftreten, bei denen Textpassagen aus verschiedenen Quellen kombiniert werden [38]. Die Nutzung von KI-generiertem Text birgt durchaus das Risiko des Plagiierens. Die Verantwortung für diese Verstöße (i.e. die Plagiate) liegt bei den Personen, die solche Texte ohne Zuordnung übernehmen [52]. Für Verstöße gegen Urheberrechte sind die Anbieter der LLM verantwortlich (laufende Verfahren in den USA gegen OpenAI und Google) [53].

Eine Täuschung liegt vor, wenn jemand eine mit unerlaubten Hilfsmitteln erstellte Leistung als eigenständig vorgibt. Dies verstößt gegen die gute wissenschaftliche Praxis und Studien- oder Prüfungsordnungen [38].

Bei digitalen oder papierbasierten Prüfungen unter Aufsicht (ohne Hilfsmittel) in Präsenz, oder bei dezentralen digitalen Prüfungen mit Proctoring und (teilweise) dem Einsatz sicherer Browser, ist die Gefahr einer Täuschung durch den Einsatz von LLM reduziert [40], [44]. Mündliche und anwendungsorientierte Prüfungen, wie OSCEs, reduzieren die Möglichkeit der LLM-Nutzung. Wichtig ist hier die Konzeption, damit nicht andere Beurteilungsfehler wie Subjektivität zu Fehlerquellen führen [40], [41], [54], [55], [56]. Problematisch sind LLM vor allem für schriftliche Prüfungen, die selbstständig und ohne Aufsicht absolviert werden, wie beispielsweise Hausarbeiten [40], [41]. Die Erkennung von synthetischem Text ist noch unterschiedlich zuverlässig, hat sich aber bereits deutlich verbessert [57], [58], [59], [60], [61]. Auch menschliche Einschätzungen können ausschlaggebend sein. Eine Universität lehnte erfolgreich eine Masterbewerbung ab, da der Essay unerwartet hochwertig für einen Bachelorabsolventen erschien und augenscheinlich KI-generiert war [62].

## Bedeutung für Hochschulen und LPÄ

### Erstellung von Prüfungsinhalten und -aufgaben

LLM wurden zum Erstellen von Prüfungsfragen eingesetzt [63], [64]. Für Wissensfragen hat z.B. ChatGPT das Potenzial, in kurzer Zeit MCQ von vergleichbarer Qualität für medizinische Abschlussprüfungen zu generieren [64], [65], [66]. Fragen, die höhere Lernzielebenen abfragen, zeigen dagegen gewisse Einschränkungen [63]. Grundsätzlich lohnt sich das Experimentieren mit unterschiedlichen Eingabeaufforderungen („Prompts“). Prompt-Engineering ist ein systematischer Ansatz zur effektiven Kommunikation mit LLM mit großem Einfluss auf das Ergebnis [15], [66], [67]. Durch KI erstellte Aufgaben müssen in einem Reviewprozess überprüft werden, denn auch sprachlich gut und plausibel formulierte Produkte, z.B. Prüfungsfragen, können fehlerhaft sein [66]. Das Risiko einer Verletzung von Urheberrechten beim Einsatz von KI zur Erstellung von Prüfungsaufgaben ist dagegen eher gering, denn nach § 60a UrhG dürfen bis zu 15% eines Werkes für nicht-kommerzielle Zwecke zur Veranschaulichung im Unterricht und für Prüfungen zugänglich gemacht werden.

### Bewertung von Prüfungsinhalten und -aufgaben

Bewertungskriterien für Prüfungsleistungen sind grundsätzlich in Hochschulgesetzen festgelegt und werden in Prüfungs- und Studienordnungen konkretisiert. Ist eine eigenständige Bewertung durch die prüfende Person vorgesehen, muss diese die Prüfungsleistung eigenständig würdigen. Ein Sprachmodell kann nur unterstützend verwendet werden. Eine eigenständige Bewertung fehlt auch dann, wenn die Bewertung nicht exakt übernommen, aber allein auf Grundlage des KI-generierten Ergebnisses erfolgt [41]. Aus datenschutzrechtlicher Sicht verstößt eine solche Bewertung grundsätzlich gegen das Verbot der automatisierten Entscheidungsfindung (Art. 22 DSGVO). Erfolgt eine Leistungsbeurteilung durch eine KI, ist gemäß KI-VO zudem von einem Hochrisiko-KI-System auszugehen [49].

# Handlungsmöglichkeiten

LLM stellen eine Herausforderung, aber auch eine Chance für Prüfungsabläufe dar. Fakultäten können auf KI-Entwicklungen reagieren, ohne notwendigerweise zu restriktiven Formaten zurückzukehren [41]. Ein generelles Verbot von KI-Anwendungen ist schwer umsetzbar und kaum sachdienlich, Algorithmen sind bereits jetzt in bestehende Systeme, wie Browser oder Textverarbeitungsprogramme integriert [41], [68]. Somit kann es sinnvoll sein, KI-Anwendungen gezielt zu erlauben oder zu integrieren.

## Anpassung von Ordnungen

Rahmenbedingungen sollten definieren, unter welchen Bedingungen oder für welche Zwecke der Einsatz legitim und freigegeben und in welchen Bereichen der Nutzen untersagt ist [41]. Auch wenn die Kontrolle unter Umständen schwierig ist, erhöht eine klare Erläuterung die Verbindlichkeit und verdeutlicht die Konsequenzen bei Regelverstößen, auch in Eigenständigkeitserklärungen. Ohne diese Einschränkungen kann kein Verstoß gegen die Prüfungsregeln angenommen werden [28], [41], [44], [52], [62]. Studierenden sollte bewusst sein, dass sie die Verantwortung für Fehler, wie z.B. Urheberrechtsverletzungen tragen [28], [41], [52] (siehe Anhang 1 für Quellen zu Mustertexten und Checklisten).

## Anpassung von Prüfungen

Bei bestehenden Prüfungen können Bewertungskriterien angepasst werden. Der kritische Umgang mit Quellen und die Positionierung im Fachdiskurs könnten stärker gewichtet werden, sprachliche Korrektheit und Ausdruck an Bedeutung verlieren [41], [52]. Noch generieren LLM auch fehlerhafte oder ungewichtete Quellenangaben [17], [69], eine Überprüfung scheint daher sinnvoll.

## Anpassung von Prüfungsprogrammen

Durch eine Veränderung der Prüfungen kann die Motivation zum Täuschen gesenkt werden. Intrinsisch motivierte, leistungsorientierte Studierende täuschen seltener als extrinsisch motivierte, die vor allem bestehen wollen [43], [70], [71] oder gestresst sind [69]. Als Ansatz könnten die Anforderungen der Einzelprüfungen reduziert und gleichzeitig die inhaltliche Relevanz gesteigert werden. Prüfungen sollten einen Bezug zu realen Erfahrungen und Erkenntnissen aufweisen und somit authentisch sein. Formate, wie MC-Fragen können durchaus eingesetzt werden [41], [52], [72], [73]. Ein „moralischer Anker“ und ein moralisches Bewusstsein können Täuschungen verringern, gefördert durch vorbildliche Lehrende und Trainings in Selbstbewusstsein, Ethik und Entscheidungsfindung [43], [69].

## LLM als Prüfungsgegenstand, LLM als Lernhilfe

Im Sinne der „AI-Literacy“ sollten Lehrende LLM als Ergänzung betrachten und sich kontinuierlich weiterbilden. Neben Wissen über KI könnte auch die Anwendung, wie z.B. die Prompt-Generierung, und die kritische Bewertung oder Validierung von KI-generierten Texten geprüft werden [18], [67], [74], [75].
LLM können in formativen Testumgebungen Wissenslücken durch eine thematische Textanalyse von Bewertungsdaten identifizieren und individuelles, zeitnahes und kontinuierliches Feedback bereitstellen, ähnlich einem ständig verfügbaren Tutor [74], [76].

# Limitationen

Der Bereich der GKI ist dynamisch und entwickelt sich rasant. Angesprochene Leistungsfähigkeiten und Limitationen könnten bald veraltet sein. Alle Entwicklungen müssen jedoch im Einklang mit dem geltenden Rechtsrahmen stehen, einschließlich UrhG, DS-GVO und KI-VO. Dies gilt in der DACH-Region für Deutschland und Österreich. Für die Schweiz sind die rechtlichen Rahmenbedingungen den Autor*innen nicht bekannt und müssen vor dem Einsatz von GKI geprüft werden.

# Danksagung

Wir bedanken uns bei Daniel Bauer, Daniel Tolks und Katharina von der Wense für das kritische Durchlesen und die fachlichen Hinweise.

# Anmerkungen

## Übersetzung

Für die englische Übersetzung wurde DeepL genutzt. Zur Editierung wurden DeepL, ChatGPT und Copilot genutzt (Prompts: „Fasse folgenden Text zusammen“ und „Verbessere folgenden Text“).

## ORCIDs der Autor*innen


- Maren März: [0000-0002-2661-5076]
- Monika Himmelbauer: [0000-0001-5516-1993]
- Alexander Oksche: [0000-0003-4592-1770]


# Interessenkonflikt

Die Autor*innen erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

# Anhänge

Verfügbar unter https://doi.org/10.3205/zma001702

1. Anhang_1.pdf (137 KB)
   Musterrichtlinien, Linksammlungen

# Literatur


1. Majumdar D, Banerji PK, Chakrabarti S. Technology Analysis & Strategic Management Disruptive technology and disruptive innovation?: ignore at your peril?! Technol Anal Strateg Manag. 2018;7325(11):1247-1255. DOI: 10.1080/09537325.2018.1523384
2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CM, Schwarzkopf SC, Unger M, Veldjuizen GP, Wagner SJ, Kather JN. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3(1):141. DOI: 10.1038/s43856-023-00370-1
3. Webster P. Six ways large language models are changing healthcare. Nat Med. 2023;29(12):2969-2971. DOI: 10.1038/s41591-023-02700-1
4. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Arcas BA, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Emturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. DOI: 10.1038/s41586-023-06291-2
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. DOI: 10.1038/s41591-023-02448-8
6. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, Zhang M, Long J, Ng AY, Rajpurkar P, Sinha SR. Development and Validation of an Artificial Intelligence System to Optimize Clinician Review of Patient Records. JAMA Netw Open. 2021;4(7):e2117391. DOI: 10.1001/jamanetworkopen.2021.17391
7. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic Reasoning Prompts Reveal the Potential for Large Language Model Interpretability in Medicine. NPJ Digt Med. 2024;7(1):20. DOI: 10.1038/s41746-024-01010-1
8. Li R, Kumar A, Chen JH. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? JAMA Intern Med. 2023;183(6):596-597. DOI: 10.1001/jamainternmed.2023.1835
9. Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. Lancet Digit Heal. 2020;2(3):e111-e122. DOI: 10.1016/S2589-7500(20)30023-6
10. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: FAccT 2021: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. March 2021. p.610-623. DOI: 10.1145/3442188.3445922
11. Bhavya B, Xiong J, Zhai C. Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT. arXiv. 2022. DOI: 10.48550/arXiv.2210.04186
12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv. 2020. DOI: 10.48550/arXiv.2005.14165
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017:5999-6009. DOI: 10.48550/arXiv.1706.03762
14. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Conerly T, El-Showk S, Elhage N, Hatfield-Doods Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nanda N, Olsson C, Amodei D, Brown T, Clark J, McCandish S, Olah C, Mann B, Kaplan J. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv. 2022. DOI: 10.48550/arXiv.2204.05862
15. Open AI, Achiam J, Adler S, Agarwal S, et al. GPT-4 Technical Report. arXiv. 2023;4:1-100. DOI: 10.48550/arXiv.2303.08774
16. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. Training language models to follow instructions with human feedback. arXiv. 2022. DOI: 10.48550/arXiv.2203.02155
17. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 20239;11(6):887. DOI: 10.3390/healthcare11060887
18. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, Bright TJ, Tatonetti N, Won KJ, Gonzalez-Hernandez G, Moore JH. ChatGPT and large language models in academia: opportunities and challenges. BioData Min. 2023;16(1):20. DOI: 10.1186/s13040-023-00339-9
19. Thapa S, Adhikari S. ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. Ann Biomed Eng. 2023;51(12):2647-2651. DOI: 10.1007/s10439-023-03284-0
20. Watkins R. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. AI Ethics. 2023. DOI: 10.1007/s43681-023-00294-5
21. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Heal. 2023;2(2):e0000198. DOI: 10.1371/journal.pdig.0000198
22. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing?? Med Educ Online. 2023;28(1):2220920. DOI: 10.1080/10872981.2023.2220920
23. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, Chen JH. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. JAMA Intern Med. 2023;183(9):1028-1030. DOI: 10.1001/jamainternmed.2023.2909
24. Long C, Lowe K, Santos A dos, Zhang J, Alanazi A, O'Brien D, Wright E, Cote D. Evaluating ChatGPT-4 in Otolaryngology–Head and Neck Surgery Board Examination using the CVSA Model. medRxiv. 2023. DOI: 10.1101/2023.05.30.23290758
25. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, Asaad WF, Cielo D, Oyeles AA, Doberstein CE, Gokaslan ZL, Telfeian AE. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. Neurosurgery. 2023;93(6):1353-1365. DOI: 10.1227/neu.0000000000002632

26. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Hear J Digit Heal. 2023;4(3):279-281. DOI: 10.1093/ehjdh/ztad029

27. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the Bar Exam. SSRN Electron J. 2023:1-35. DOI: 10.2139/ssrn.4389233

28. FernUni. A teacher's guide to ChatGPT and remote assessments. FernUni.ch. 2023.

29. Neutatz F, Abedjan Z. Whats is "Good" Training Data. In: Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, Rostalski F, editors. Künstliche Intelligenz Wie gelingt eine vertrauenswürdige Verwendung. Tübingen: Mohr Siebeck; 2022. DOI: 10.1628/978-3-16-161299-2

30. Stoyanovich J, Howe B, Jagadish HV. Responsible data management. Proc VLDB Endow. 2020 ;13(12):3474-3488. DOI: 10.14778/3415478.3415570

31. Heaven WD. Geoffrey Hinton tells us why he's now scared of the tech he helped build. MIT Technol Rev. 2023. Zugänglich unter/available from: https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/

32. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, Berkowitz ST, Finn AP, Jahangir E, Scoville EA, Reese TX, Friedmann DL, Bastarache JA, van der Heijden YF, Wrigth JJ, Ye F, Carter N, Alexander MR, Choe JH, Chastain CA, Zic JA, Horst SN, Turker I, Agarwal R, Osmundson E, Idrees K, Kiernan CM, Padmanabhan C, Bailey CE, Schlegel CE, Chabless LB, Gibson MK, Osterman TJ, Wheless LE, Johnson DB. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023;6(10):e2336483. DOI: 10.1001/jamanetworkopen.2023.36483

33. George C, Stuhlmüller A. Factored Verification: Detecting and Reducing Hallucination in Summaries of Academic Papers. arXiv. 2023. DOI: https://doi.org/10.48550/arXiv.2310.10627

34. Huang J, Chen X, Mishra S, Zheng HS, Yu AW, Song X, Zhou D. Large Language Models Cannot Self-Correct Reasoning Yet. arXiv. 2023;1:1-19. DOI: 10.48550/arXiv.2310.01798

35. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023;307(2):e230163. DOI: 10.1148/radiol.230163

36. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. DOI: 10.1038/s41591-018-0300-7

37. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int. 2023;00(00):1-12. DOI: 10.1080/14703297.2023.2190148

38. Georg-August-Universität Göttingen. Handreichung zum Umgang mit Plagiaten für Lehrende an der Sozialwissenschaftlichen Fakultät der Georg-August-Universität Göttingen. Göttingen: Georg-August-Universität Göttingen; 2021. p.1-2.

39. Hochschulforum Digitalisierung. ChatGPT im Hochschulkontext – eine kommentierte Linksammlung. Essen: Hochschulforum Digitalisierung; 2023. Zugänglich unter/available from: https://hochschulforumdigitalisierung.de/chatgpt-im-hochschulkontext-eine-kommentierte-linksammlung/

40. Moritz S, Romeike B, Stosch C, Tolks D. Generative AI (gAI) in medical education: Chat-GPT DQG and co. GMS J Med Educ. 2023;40(4):Doc54. DOI: 10.3205/zma001636

41. Salden P, Leschke J. Didaktische und rechtliche Persprektiven auf KI-gestütztes Schreiben in der Hochschulbildung. Bochum: RUB; 2023. p.41. Zugänglich unter/available from: https://hss-opus.ub.ruhr-uni-bochum.de/opus4/frontdoor/deliver/index/docId/9734/file/2023_03_06_Didaktik_Recht_KI_Hochschulbildung.pdf

42. Stabsstelle IT-Recht der bayerischen staatlichen Universitäten und Hochschulen. Prüfungsrechtliche Fragen zu ChatGPT. Würzburg: Universität Würzburg; 2022. Zugänglich unter/available from: https://www.rz.uni-wuerzburg.de/fileadmin/42010000/2023/ChatGPT_und_Pruefungsrecht.pdf

43. Simkin MG, McLeod A. Why do college students cheat? J Bus Ethics. 2010;94(3):441-453. DOI: 10.1007/s10551-009-0275-x

44. Radcke A. Der Einsatz von KI in Hochschulprüfungen und dessen prüfungsrechtlichen Auswirkungen. Potsdam: Universität Potsdam; 2023. Zugänglich unter/available from: https://www.uni-potsdam.de/fileadmin/projects/zfq/Lehre_und_Medien/E-Assessment/Auswirkung_KI_auf_Pruefungen_20230524.pdf

45. Deutsche Forschungsgemeinschaft. Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Bonn: DFG; 2019. DOI: 10.5281/zenodo.3923601

46. European Commission. Artificial Intelligence – Questions and Answers. Brussels: European Commision; 2023. Zugänglich unter/available from: https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683

47. Europäisches Parlament. KI-Gesetz: erste Regulierung der künstlichen Intelligenz. Brüssel: Europäisches Parlament; 2023. Zugänglich unter/available from: https://www.europarl.europa.eu/news/de/headlines/society/20230601ST093804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz

48. Maximilian Borkowsky. Der EU AI Act: Was bedeutet er für künstliche Intelligenz in Unternehmen? 2023. Zugänglich unter/available from: https://www.melibo.de/blog/der-eu-ai-act-was-bedeutet-er-fur-kunstliche-intelligenz-in-unternehmen

49. Madiega TA. Artificial intelligence act. Brussels: Europäisches Parlemant; 2024. Zugänglich unter/available from: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792

50. Fishman T. "We know it when we see it " is not good enough?: toward a standard definition of plagiarism that transcends theft , fraud , and copyright. Wollongong (NSW, Aust): University of Wollongong; 2009. p.28-30.

51. Khalil M, Er E. Will ChatGPT get you caught? Rethinking of Plagiarism Detection. arXiv. 2023. DOI:

52. Gimpel H, Ruiner C, Schoch M, Schoop M, Hall K, Eymann T, Röglinger M, Vandirk S, Lämmermann L, Urbach N, Mädche A, Decker S. Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education. Hohenheim: Universität Hohenheim; 2023. p.47.

53. Daniel J Sokolov. Large Language Models: US-Autoren verklagen OpenAI wegen Copyright-Verletzung. 2023. Zugänglich unter/available from: https://www.heise.de/news/Large-Language-Models-US-Autoren-verklagen-OpenAI-wegen-Copyright-Verletzung-9301736.html

54. Susnjak T. ChatGPT: The End of Online Exam Integrity? arXiv. 2022. DOI: 10.48550/arXiv.2212.09292

55. Davis MH, Karunathilake I. The place of the oral examination in today's assessment systems. Med Teach. 2005;27(4):294-297. DOI: 10.1080/01421590500126437

56. Wass V, Van Der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945-949. DOI: 10.1016/S0140-6736(00)04221-5

57. Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S. Can AI-Generated Text be Reliably Detected? arXiv. 2023. DOI: 10.48550/arXiv.2303.11156

58. Krishna K, Song Y, Karpinska M, Wieting J, Iyyer M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. arXiv. 2023. DOI: 10.48550/arXiv.2303.13408

59. Wu J, Yang S, Zhan R, Yuan Y, Wong DF, Chao LS. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. arXiv. 2023. DOI: 10.48550/arXiv.2310.14724

60. Mo Y, Qin H, Dong Y, Zhu Z, Li Z. Large Language Model (LLM) AI Text Generation Detection based on Transformer Deep Learning Algorithm. Int J Eng Manag Res. 2024;14(2):154-159. DOI: 10.5281/zenodo.11124440

61. Nguyen TT, Hatua A, Sung AH. How to Detect AI-Generated Texts? In: 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE; 2023. p.0464–0471. DOI: 10.1109/UEMCON59035.2023.10316132

62. Zenthöfer J. Erstes Urteil zu ChatGPT an Hochschulen. Frankfurter Allgemeine. 2024. Zugänglich unter/available from: https://www.faz.net/aktuell/karriere-hochschule/erstes-urteil-zu-chatgpt-an-hochschulen-student-benutzte-ki-fuer-bewerbung-19564795.html

63. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. 2023;15(6):e40977. DOI: 10.7759/cureus.40977

64. Cheung BH, Lau GK, Wong GT, Lee EY, Kulkarni D, Seow CS, Wong R, Co MT. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S. A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023;18(8):e0290691. DOI: 10.1371/journal.pone.0290691

65. Klang E, Portugez S, Gross R, Kassif Lerner R, Brenner A, Gilboa M, Ortal T, Ron S, Robinzon V, Meiri H, Segal G. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ. 2023;23(1):772. DOI: 10.1186/s12909-023-04752-w

66. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions. Acad Med. 2024;99(5):508-512. DOI: 10.1097/ACM.0000000000005626

67. Heston TF, Khun C. Prompt Engineering in Medical Education. Int Med Educ. 2023;2(3):198-205. DOI: 10.3390/ime2030019

68. Crawford J, Cowling M, Allen KA. Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). J Univ Teach Learn Pract. 2023;20(3). DOI: 10.53761/1.20.3.02

69. Goodman RS, Patrinely JR, Osterman T, Wheless L, Johnson DB. On the cusp: Considering the impact of artificial intelligence language models in healthcare. Med. 2023;4(3):139-140. DOI: 10.1016/j.medj.2023.02.008

70. Anderman EM, Koenka AC. The Relation Between Academic Motivation and Cheating. Theory Pract. 2017;56(2):95-102. DOI: 10.1080/00405841.2017.1308172

71. Hsiao YP, Klijn N, Chiu MS. Developing a framework to re-design writing assignment assessment for the era of Large Language Models. Learn Res Pract. 2023;9(2):148-158. DOI: 10.1080/23735082.2023.2257234

72. Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment? J Learn Dev High Educ. 2023;27. DOI: 10.47408/jldhe.vi27.1009

73. Schuwirth LW, Van Der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011;33(6):478-485. DOI: 10.3109/0142159X.2011.565828

74. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. Acad Med. 2023;99(1):22-27. DOI: 10.1097/acm.0000000000005439

75. Busse B, Kleiber I, Eickhoff FC, Andree K. Hinweise zu textgenerierenden KI- Systemen im Kontext von Lehre und Lernen. Köln: Universität zu Köln; 2023. Zugänglich unter/available from: https://uni-koeln.sciebo.de/s/7uwYRX5a92eznVl#pdfviewer DOI: 10.13140/RG.2.2.35392.61449/1

76. Cardona MA, Rodríguez RJ, Ishmael K. Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations. US Dep Educ Off Educ Technol. 2023;(1):1-71. Zugänglich unter/available from: https://www2.ed.gov/documents/ai-report/ai-report.pdf



**Korrespondenzadresse:**
Maren März
Charité – Universitätsmedizin Berlin, AG Progress Test Medizin, Geschäftsbereich Lehre, Charitéplatz 1, 10117 Berlin, Deutschland, Tel.: +49 (0)30/450-576047
maren.maerz@charite.de



***Bitte zitieren als***
*März Maren, Himmelbauer M, Boldt K, Oksche A. Legal aspects of generative artificial intelligence and large language models in examinations and theses. GMS J Med Educ. 2024;41(4):Doc47. DOI: 10.3205/zma001702, URN: urn:nbn:de:0183-zma0017020*

***Artikel online frei zugänglich unter***
*https://doi.org/10.3205/zma001702*

***Eingereicht:*** *11.04.2024*
***Überarbeitet:*** *01.07.2024*
***Angenommen:*** *09.07.2024*
***Veröffentlicht:*** *16.09.2024*