Seeing the Forest from the Trees: Unveiling the Landscape of Generative AI for Education Through Six Evaluation Dimensions

Yael Feldman-Maggor[1](✉), Teresa Cerratto-Pargman[2], and Olga Viberg[3]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
yaelfm@kth.se
[2] Stockholm University, Digital Futures, Stockholm, Sweden
tessy@dsv.su.se
[3] KTH Royal Institute of Technology, Digital Futures, Stockholm, Sweden
oviberg@kth.se

Abstract. Artificial intelligence (AI) holds significant promise as a technology that may improve the quality of educational practices. This includes specialized AI-powered technologies tailored for education and general AI-based technologies, including recently popular generative AI tools that stakeholders are increasingly adapting for teaching and learning. Integrating AI tools into educational settings holds numerous potential pedagogical benefits, such as assisting teachers in planning lessons, promoting personalization, and enhancing student autonomy. However, concerns about bias and discrimination linked to the use of these technologies have rapidly emerged. Today, standardized evaluation criteria to assess the potential contribution of such tools to education and their reliability within the learning and teaching context are lacking. To address this gap, we build on an existing taxonomy for the evaluation of open educational resources (OER) to better suit the unique features of generative AI. The result is a six-dimensional evaluation approach that includes descriptive, pedagogical, representational, communication, scientific content, as well as the ethical and transparency dimension. We then apply this approach to examine the educational potential and ethical concerns around 30 AI tools. The analysis facilitates a critical mapping of the potential and risks of AI-powered technologies in education settings.

Keywords: Generative AI · Algorithm Bias · Open Educational Resource (OER)

1 Introduction

Generative artificial intelligence (GenAI) technologies can enhance learning and teaching opportunities by creating new content, data, and visualizations of scientific phenomena. At the same time, they pose critical ethical challenges such as bias and discrimination [1]. With the rapid development of GenAI tools, governments, international organizations, universities, and researchers have published reports on this topic regarding education [e.g. 2]. While there are various ways for evaluating learning technologies

R. Ferreira Mello et al. (Eds.): EC-TEL 2024, LNCS 15160, pp. 99–105, 2024.
https://doi.org/10.1007/978-3-031-72312-4_12

[3], so far, there is no standardized method for evaluating GenAI tools for learning, teaching, and research purposes. This raises the question of how GenAI tools can be effectively assessed for educational purposes and the need to further develop existing taxonomies for evaluating learning technologies [4, 5] with new categories and adding a new dimension focusing on ethics and potential risks. We do so by evaluating 30 selected, frequently used GenAI tools. Our approach highlights the identified gap and is driven by the motivation to enable end-users to evaluate generative AI tools critically for their specific needs.

The taxonomy adapted for this study was initially developed to assess open education resources (OER) [6]. It has been further developed and updated over the years [4, 5]. It encompassed five dimensions: *descriptive, pedagogical, representational, communication*, and *scientific content*. The *descriptive* dimension presents basic information about the technology, including its creators and relevant technical data. The *pedagogical* dimension aids in evaluating the learning method and its objectives. The *representational* dimension focuses on how information and knowledge about specific topics are presented in text and images. The *communication* dimension assesses the potential for user interaction via the tool. Finally, the *scientific content* dimension includes the concept of "information reliability," noting that books and journal articles undergo peer review by experts before publication, whereas online information is not necessarily subjected to such scrutiny. Consequently, it is up to the user to assess the reliability of the information provided by such tools.

There are additional factors to consider with emerging GenAI technologies that are not part of the adopted taxonomy. GenAI presents unique challenges, including the potential for generating biased information and 'hallucinations' or processing data in inaccurate or biased ways [7]. For example, in examining AI chatbots based on large language models, Sun et al. [8] found two types of hallucinations: intrinsic and extrinsic. Whereas intrinsic hallucinations refer to non-factual statements, such as incorrectly predicting a celebrity's birthday, extrinsic hallucinations are irrelevant or out-of-context responses, such as describing the history of football when the user asks about the number of teams currently in the group. This could introduce bias into educational content. AI systems depend on data for training. Consequently, data is integral to the functionality of these algorithms and systems. When the training data contains biases, the algorithms inherently learn and mirror them in their output [9, 10]. Hence, biases in the data can influence the algorithms that use this data, leading to biased outcomes. These algorithms have the potential to amplify further and perpetuate the biases found in the data [11]. Also, algorithms may exhibit biased behavior due to specific design choices independent of the data's bias. The results of these biased algorithms can harm user decisions and create a bias cycle that affects the data used to train future algorithms [11]. For example, Mehrabi et al. [11] describe three sources of potential bias: data to the algorithm, algorithm to the user, and user to data. Data to the algorithm means that bias in the data can result in biased algorithmic outcomes, affecting different educational stakeholders. The algorithm to the user means biased algorithmic outcomes can bias user behavior. Finally, user-to-data means any inherent biases in users could be reflected in the data they generate.

2 Aim

End-users such as teachers and students cannot necessarily identify all sources of bias. However, they can be aware of these potential biases and take them into consideration when using various GenAI tools in teaching and learning practices. To assist them in this evaluation, this study proposes expanding the existing taxonomy [4–6] to include these critical ethical aspects. Although users cannot examine the data directly, they can assess bias in generated figures, text, and other ethical considerations, including privacy, age restrictions, and copyright issues. The proposed research aims to map the landscape of the selected set of GenAI tools to evaluate their reliability and potential contributions to education. This effort is crucial for providing a deeper understanding of how these emerging technologies can shape educational ecosystems, including learning outcomes, identifying potential risks, and laying the groundwork for informed decision-making for learning and teaching. This research seeks to contribute to the responsible integration of GenAI technologies into educational settings by adopting an evaluative approach that also considers ethics, including various biases.

3 Methods

This qualitative study combines "top-down" and "bottom-up" approaches. First, we followed Holmes and Tuomi to determine whether the selected tool suits students, educators, or institutions or whether there is overlap among these groups [12]. Second, we draw upon an existing taxonomy for evaluating OER [4, 5], which provides a structured approach to assess pedagogical potential in educational settings in a "top-down" way. Further, new categories that emerged through the evaluation were analyzed inductively, following a "bottom-up" approach where themes arise directly from the data [13]. The evaluation was conducted through an experiential engagement with a selected set of 30 GenAI tools frequently used by stakeholders. We review the privacy policy, introduction, training materials, and the questions and answers section on the tool's website. This approach provided an overview of the AI tool, enabled us to identify new criteria, and updated the evaluation taxonomy. The analysis is based on our engagement with the tools and information on their website, such as the use of terms and privacy policy. We first analyzed 10 AI tools that helped us validate and update the taxonomy, and then we evaluated 20 additional AI tools according to the updated taxonomy. The first author conducted the evaluation. Subsequently, the third author evaluated 20% of the AI tools to determine the agreement on the evaluation. All three authors conducted a discussion to reach a full agreement regarding the evaluation criteria.

4 Result and Discussion

Most of the evaluated GenAI tools are not specific to education but can be tailored to the needs of educators, students, and researchers. For instance, educators can use such tools as ChatGPT, Google, and Gemini to prepare lessons, while students can seek hints from the tool on how to solve problems. In our study, we have also evaluated five tools

specifically designed for education (Diffit, MagicSchool AI, Curipod, IllumiDesk, MathGPTPro). Table 1 presents a summary of our evaluation. Each dimension is analyzed, with categories derived from the original taxonomy, and newly emerged categories are indicated in italics. Importantly, we introduce the "Ethics and Transparency" category as a new dimension to the taxonomy. The list of the evaluated tools and the detailed evaluation is available in the provided link.[1]

Although our focus was on GenAI technologies, where users usually can ask follow-up questions, we found that 53% of the tools were limited in enabling dialogical communication (e.g., Gamma), which can be seen as a limitation. The suggested criteria, "information on resources provided" in the "scientific content dimension," can help assess a tool's reliability. With these criteria, we highlight that some tools reveal the resources that serve as the basis for the generated information, highlighting the strengths of these tools and the weaknesses of others that do not disclose their data sources, thus being less transparent. One of the challenges with AI tools is *transparency* [14], as AI tools are often viewed as a "black box" [15]. Despite this, we found that it is possible, to a certain extent, to evaluate the tools from an end-user's perspective. Another issue that can be regarded as a matter of transparency is when AI tools state the types of natural language processing models they use, indicating that the tool is not standalone. In the "Ethics and Transparency" dimension, we highlight this in the suggested criteria "based on another tool" (e.g., tools based on the ChatGPT model).

Less than half of tools (47%) explicitly stated that users are responsible for assessing the information, while others did not. This stresses the importance of critical assessment by all educational stakeholders. In our evaluation, we identified biases not only in textual content but also in other aspects. For example, gender bias was previously identified in ChatGPT and even in non-generative AI technologies [16]. However, we also observed the opposite bias in one tool that exclusively generated female images, skewing perception towards a "female-only" world. We also identified several biases, such as a preference for newer sources to avoid outdated information. This raises the question of how these tools balance acknowledging previous studies and knowledge. Additionally, a lack of critical thinking on the part of the evaluated tools was observed, prompting questions about limitations or future studies based on an uploaded paper. Finally, we also found that AI-generated pictures exhibit bias and inaccurately portray scientific models (e.g., Leonardo). This highlights the importance of integrating bias considerations with scientific content assessment and attention to pedagogy. This comprehensive perspective is vital for ensuring that the use of AI in educational settings supports effective teaching and learning and addresses and mitigates the risk of perpetuating biases. Doing so aims to foster an educational environment that leverages AI technologies to their fullest potential while maintaining a critical awareness of their limitations and ethical implications. AI tools, serving as OER for education, are evolving rapidly daily.

Consequently, these tools are continuously being developed, and evaluating them in a few months could result in changes to some of our findings. To account for this, we have documented the evaluation dates. However, undertaking this evaluation at this stage is

[1] List of tools and detailed evaluation: https://docs.google.com/spreadsheets/d/1_qk-aI91OV910E8Kst_SMh1cau1b6OKc/edit?usp=sharing&ouid=111082985714990261310&rtpof=true&sd=tru

critical. It not only aids in understanding current trends more broadly by mapping 30 tools but also provides valuable feedback to AI tool developers, emphasizing the importance of addressing pedagogic aspects and concerns such as bias. Educators and researchers can also use the evaluation criteria to select GenAI tools for their purposes, ensuring they are informed of their benefits and limitations. In future studies, we aim to involve educators in using these tools for teaching purposes to gather further insights.

Table 1. Summary of the Generative AI Tools Evaluation (N = 30).

Dimension and Category	Frequency	Percentage
Descriptive		
Specific for education	5	17%
Account required	22	73%
Free use option	25	83%
Pedagogical		
Lesson plan assistant	17	57%
Research assistant	20	67%
Content creation	30	100%
Upload data option	27	90%
Exercises feedback option	13	43%
Exercises without feedback	14	47%
Team/Cooperation	4	13%
Data Analysis Assistant	9	30%
Representational		
Text	27	90%
Video	5	17%
Figures	13	43%
Audio	4	13%
Communication		
Search engine	30	100%
Prompt follow-up questions	20	67%
Limited dialog ability	16	53%
Shareability	21	70%
Scientific Content Dimension		
User responsible to evaluate information	30	100%
Statement on reliability	14	47%
Information on resources provided	6	20%

(*continued*)

Table 1. (*continued*)

Dimension and Category	Frequency	Percentage
Ethics and Transparency		
Gender bias	7	23%
Other biases identified	15	50%
Based on another AI-technology	9	30%
Data on users collected	25	83%
Different privacy settings for specific users	16	53%
Copyright	8	27%
Age limit	17	57%

Acknowledgements. This work was supported by the Digital Futures post-doctorate fellowship (Stockholm, Sweden) (and the Israel Science Foundation (ISF) award (Grant 73/24).

References

1. Hwang, G.-J., Chen, N.-S.: Editorial position paper: exploring the potential of generative artificial intelligence in education: applications, challenges, and future research directions. Educ. Technol. Soc. **26**(2), 18 (2023)
2. Holmes, W., et al.: Guidance for generative AI in education and research. UNESCO Publishing (2023)
3. Lai, J.W., Bower, M.: How is the use of technology in education evaluated? A systematic review. Comput. Educ. **133**, 27–42 (2019)
4. Feldman-Maggor, Y., Rom, A., Tuvi-Arad, I.: Integration of open educational resources in undergraduate chemistry teaching–a mapping tool and lecturers' considerations. Chem. Educ. Res. Pract. **17**(2), 282–295 (2016)
5. Nachmias, R., Tuvi-Arad, I.: Taxonomy of scientifically oriented educational websites. J. Sci. Educ. Technol. **10**, 93–104 (2001)
6. Nachmias, R., Mioduser, D., Oren, A., Lahav, O.: Taxonomy of educational websites-a tool for supporting research, development and implementation of web-based learning. Int. J. Educ. Telecommun. **5**(3), 193–210 (1999)
7. El Helou, M., Süsstrunk, S.: Bigprior: toward decoupling learned prior hallucination and data fidelity in image restoration. IEEE Trans. Image Process. **31**, 1628–1640 (2022)
8. Sun, W., Shi, Z., Gao, S., Ren, P., de Rijke, M., Ren, Z.: Contrastive learning reduces hallucination in conversations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 13618–13626 (2023)
9. Tao, Y., Viberg, O., Baker, R.S., Kizilcec, R.F.: Auditing and mitigating cultural bias in llms (2023). arXiv preprint arXiv:2311.14096.
10. Figueras, C., Verhagen, H., Cerratto Pargman, T.: Exploring tensions in Responsible AI in practice. an interview study on AI practices in and for Swedish public organizations. Scand. J. Inf. Syst. **34**(2), 6 (2022)
11. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)

12. Holmes, W., Tuomi, I.: State of the art and practice in AI in education. Eur. J. Educ. **57**(4), 542–570 (2022)
13. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)
14. Nazaretsky, T., Cukurova, M., Alexandron, G.: An instrument for measuring teachers trust in AI-based educational technology. In: LAK22: 12th International Learning Analytics And Knowledge Conference, pp. 56–66 (2022)
15. Khosravi, H., et al.: Explainable artificial intelligence in education. Comput. Educ. Artif. Intell. **3**, 100074 (2022)
16. Gross, N.: What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. Soc. Sci. **12**(8), 435 (2023)