Can We Trust AI in the Classroom? Examining Ethical and Privacy Challenges of LLMs – A Literature Review

Ivana Ružić, Igor Balaban
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42 000 Varaždin, Croatia
{ivruzic, igor.balaban}@foi.unizg.hr

***Abstract**. This review paper examines the ethical and data privacy challenges associated with using Large Language Models (LLMs) in education. While LLMs like ChatGPT offer personalized support and expanded access to learning, their implementation raises concerns about bias, academic integrity, and student data protection. Key risks include over-reliance on AI, insufficient transparency, and unclear accountability. The paper highlights the lack of research on teacher perspectives and use in education. It concludes by calling for ethical frameworks, robust privacy policies, and inclusive design to ensure that LLMs enhance educational equity and uphold trust in digital learning environments.*

Keywords. Large Language Models, education, data privacy, ethical consideration

1 Introduction

Large Language Models (LLMs) have rapidly emerged as transformative tools in education, offering new opportunities to enhance learning and teaching processes. Their ability to generate coherent, human-like responses makes them valuable for supporting students across a range of tasks, from studying and brainstorming to complex problem-solving, including coding (Due et al., 2024). Tools such as ChatGPT are increasingly being integrated into classrooms and educational platforms, promoting greater engagement and personalized learning experiences. These capabilities suggest a growing potential to democratize education and reduce disparities by providing accessible, real-time support to learners.

A significant advantage of LLMs is their ability to deliver immediate, context-sensitive assistance. Students encountering challenges with homework or complex concepts can receive clear and simplified explanations, much like having access to a personal tutor (Due et al., 2024). Moreover, LLMs accommodate various learning preferences, some students benefit from reading detailed content, while others prefer interactive, dialog-based support. This versatility makes LLMs particularly useful in under-resourced settings where individual teacher support may be limited, ensuring that learners still receive timely and customized help.

Despite their promise, the integration of LLMs into educational contexts raises serious ethical and data privacy concerns. A central ethical issue involves algorithmic bias and lack of transparency. These models may unintentionally sustain social stereotypes or produce unfair outcomes, especially if not properly monitored and evaluated (Fenu et al., 2022; Wambsganss et al., 2023). Furthermore, the use of LLMs involves processing enormous amounts of data, which raises critical questions about student privacy and data protection (Abbo et al. (2025); Zhang et al. (2022); Andries & Robertson (2023)). Ensuring anonymity, implementing secure data storage, and maintaining transparency about how data is collected and used are essential steps toward ethical deployment (Fenu et al., (2022); McDonald & Pan (2020)).

Wambsganss et al. (2023) note that while LLMs are increasingly used as writing support tools in educational environments, this trend requires ethical oversight to avoid unintended data exposure or improper use. Their study confirms that even when LLMs do not transfer gender bias to students' outputs, robust privacy measures are still necessary. As student writing and behaviour are increasingly subject to algorithmic analysis, educational stakeholders must establish clear ethical guidelines and implement fairness-oriented evaluation frameworks to maintain trust and accountability (Fenu et al., 2022).

The integration of LLMs into education offers new opportunities for personalized learning and data analysis, but it also raises critical ethical and privacy concerns. Compliance with the General Data Protection Regulation (GDPR) is essential, as it mandates strict safeguards for handling personal data within the EU. Educational use of LLMs must ensure data minimization, anonymization, and informed consent (März et al., 2024; Mamalis et al., 2024). GDPR also limits automated decision-making, requiring human oversight in any impactful decisions regarding students (März et al., 2024). Privacy risks, such as unintentional memorization of sensitive data, demand technical solutions like differential privacy and data anonymization (Xiao et al., 2023; Miranda et

al., 2024). Ethical concerns include potential algorithmic bias and misinformation, highlighting the need for transparency and fairness (Dungca, 2023). Techniques such as Privacy Protection Language Models (PPLM) and instruction-based tuning help balance effectiveness and privacy (Xiao et al., 2023). To responsibly utilize LLMs in education, institutions must prioritize privacy, legality, and ethical oversight while maximizing their educational benefits.

2 General Objective and Research Questions

The aim of this paper is to explore and understand the primary ethical concerns associated with the use of LLMs in education. The study focuses on identifying key ethical issues resulting from the integration of these technologies into educational settings, as well as analysing the perceived risks and challenges related to data privacy for students. Additionally, this literature review attempts to provide insights that can guide the responsible and secure implementation of LLMs in education.

The following research questions are at the centre of interest:

RQ1: What are the primary ethical concerns associated with the use of Large Language Models in education?

RQ2: What are the perceived risks and challenges related to data privacy when using Large Language Models in educational settings?

3 Material and Method

The literature review process was divided into four stages: identification, screening, eligibility, and inclusion, according to Boland et al. (2017) and the complete process is illustrated on Fig. 1.

In the first identification stage, a structured search strategy was created for use on the scientific databases Scopus and Web of Science. A search string: *Large Language Model** AND *ethic** AND *data privacy* AND *education*, was used for database search.

For the second stage, screening, 206 results from Web of Science and 107 results from Scopus were identified. The following additional selection criteria were used:

1. Published in English,
2. Published within the time frame 2019-2025,
3. Document type were article and conference paper,
4. Full text was available,
5. The subject areas were Computer and Social Science (Scopus) and Educational Research in Education (Web of Science).

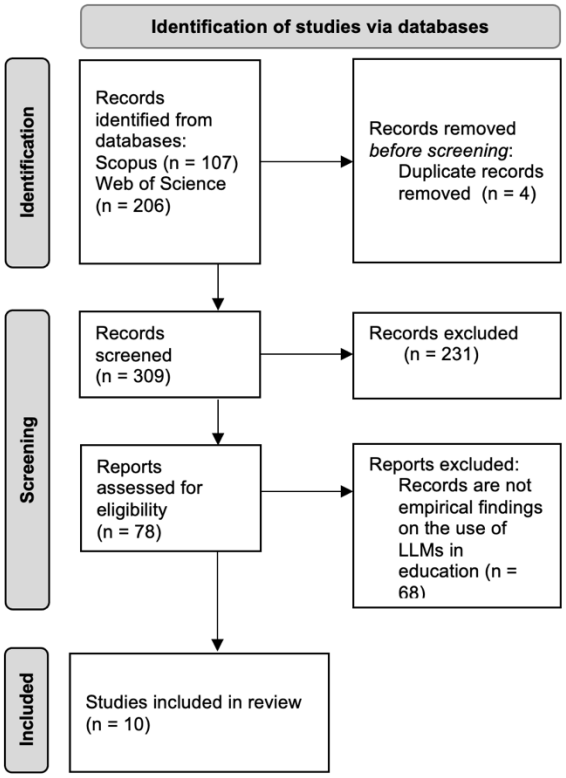

Figure 1. PRISMA flow diagram of the literature review process

As a result, 31 studies were obtained from Scopus and 51 from Web of Science. 4 duplicates were detected and excluded, and 78 studies were reviewed by titles and abstracts.

In the next stage, the final eligibility criterion was applied – studies should have included empirical findings on the use of LLMs in education. From the focus are excluded all studies that examine LLMs in education but are not related to ethical and data privacy issues, as well as studies that explore LLMs without presenting empirical research findings, offering instead only guidelines or conceptual frameworks for their use in education. A total of 68 studies were removed because they did not meet the required criteria. A total of 10 studies were included in the fourth stage of the literature review. The three studies are available in Scopus under the preprints category, reflecting their relevance and emerging importance within the rapidly evolving field of AI in education.

4 Results

The results of the literature review are presented in Table 1, showing the type of AI tool used, the number and role of participants, the educational level, and the duration of the educational activity.

Table 1. Presentation of the research results

Authors	Participants			Duration	Tool
	Role	Level	No		
Abbo et al. (2025)	students	Higher education	21	-	LLM, Robot Misty II
Zhang et al. (2022)	students	Middle education	25	30 hours	Teachable Machine, GANs
Shalevska & Kostadinovska-Stojchevska (2024)	students	Higher education	114	-	ChatGPT
Wambsganss et al. (2023)	students	Middle education	231	min. 15 min	ChatGPT
Sublime & Renna (2024)	students	Middle and Higher education	395	-	ChatGPT
Prather et al. (2023)	students	Higher education	171	-	ChatGPT
Andries & Robertson (2023)	students	Primary education	194	-	Alexa
Vrågård et al. (2024)	educators	Higher education	-	-	ChatGPT
Knowles (2021)	educators	Higher education	35	-	LLM
McDonald & Pan (2020)	students	Higher education	20	-	LLM

The results show that most studies focused on students as the primary participants, reflecting a strong interest in understanding learners' experiences and outcomes when interacting with AI tools. Only two studies (Vrågård et al., 2024; Knowles, 2021) included educators, indicating that research on teachers' perspectives and experiences remains relatively underrepresented despite its importance for effective integration of AI in educational contexts.

The majority of studies were conducted in the context of higher education, followed by middle (secondary) education. Only one study (Andries & Robertson, 2023) targeted primary education, while one study (Sublime & Renna, 2024) included participants from both secondary and tertiary levels. This distribution suggests that AI interventions are primarily explored within more advanced educational settings, potentially due to students' higher digital literacy and autonomy.

The number of participants varied significantly, ranging from small-scale studies with fewer than 30 participants (e.g., Abbo et al., 2025; Zhang et al., 2022; McDonald & Pan, 2020) to large-scale studies involving over 200 individuals (e.g., Sublime & Renna, 2024; Wambsganss et al., 2023; Andries & Robertson, 2023).

Most studies did not report the duration of the intervention, which limits our understanding of the temporal dynamics of AI tool use. Notable exceptions include Zhang et al. (2022), who conducted a 30-hour intervention, and Wambsganss et al. (2023), who reported a minimum usage duration of 15 minutes.

The most commonly used AI tool across studies was ChatGPT, featured in five studies, particularly within higher and secondary education. LLMs more broadly were employed in three studies, while other AI technologies such as Teachable Machine, GANs, Alexa, and the Misty II robot were used more sporadically. This indicates a dominant reliance on text-based conversational agents, with less frequent use of embodied or multimodal AI systems.

Some studies explored more innovative or less conventional tools such as Generative Adversarial Networks (GANs) and Teachable Machine (Zhang et al., 2022) suggesting a growing interest in creative applications of AI, particularly in STEM education. The inclusion of a physical robot (Misty II) in Abbo et al. (2025) represents a promising, though still rare, integration of social robotics into educational contexts.

4.1 Ethical Concerns Associated with the Use of Large Language Models in Education

The integration of LLMs into educational settings has caused significant ethical debate. Recent studies highlight a range of concerns that span data privacy, academic integrity, cognitive development, and equity, underscoring the need for critical oversight and responsible deployment of these technologies.

A dominant ethical concern across the literature is the handling of sensitive data. Abbo et al. (2025), Zhang et al. (2022), and Andries & Robertson (2023) emphasize that the processing of personal information, particularly in primary education, requires strict protections to avoid abuse and data breaches. Wambsganss et al. (2023) further advocate for anonymization practices and transparent data governance. McDonald & Pan (2020) also highlights risks to student privacy and autonomy, highlighting concerns about the extensive data collection by AI systems and possible lack of consent for its use.

The potential of LLMs to spread and increase biases inherent in their training data is well-documented (Zhang et al., 2022; Wambsganss et al., 2023). These biases may manipulated learning materials, reinforce stereotypes, and negatively affect learners' perceptions, especially in formative years. Abbo et al. (2025) and Prather et al. (2023) caution that this can erode fairness and inclusivity in educational outcomes. McDonald & Pan (2020) further differentiate fairness from equality, noting that AI must account for students' diverse backgrounds to avoid discriminatory impacts. Similarly, Knowles (2021) warns that AI systems built on historical data may

reinforce systemic inequities, disadvantaging marginalized groups.

LLMs raise concerns about academic dishonesty. Shalevska & Kostadinovska-Stojchevska (2024) report widespread student use of AI for assessments, potentially undermining educational standards. Prather et al. (2023) and Sublime & Renna (2024) highlight these concerns, noting risks of plagiarism and cheating, which complicate authorship attribution and damage the credibility of academic achievements. Vrågård et al. (2024) similarly emphasize the critical risk that over-reliance on AI-generated content may compromise authenticity, especially in younger students whose understanding must be genuinely reflected in their work.

Several studies (Sublime & Renna, 2024; Andries & Robertson, 2023; Prather et al., 2023) warn against over-reliance on AI-generated responses. This dependency may delay the development of critical thinking, problem-solving, and expressive writing skills. Moreover, Andries & Robertson (2023) raise the issue of anthropomorphism, where young users attribute human-like traits to LLMs, possibly distorting their understanding of AI capabilities and affecting emotional development. Vrågård et al. (2024) further highlight concerns that reliance on AI may reduce creativity and original thinking, potentially creating a dependency that limits natural development in young learners. Knowles (2021) adds that overdue postponement to AI decisions may reduce opportunities for students to engage in moral and reflective reasoning, impacting their capacity for critical ethical judgment.

LLMs can occasionally generate inaccurate or misleading content. Abbo et al. (2025) and Prather et al. (2023) highlight that without critical evaluation, students may accept false information as valid, posing long-term risks to knowledge construction and trust in educational resources. Vrågård et al. (2024) point out that this is especially problematic in primary education, where learners' critical evaluation skills are still developing.

Wambsganss et al. (2023) and Andries & Robertson (2023) underscore the lack of transparency in model training and operation. Clear documentation of data sources and decision-making processes is essential to ensure alignment with pedagogical standards. Accountability mechanisms must be established to oversee ethical deployment and usage within schools. Knowles (2021) and McDonald & Pan (2020) emphasize that unclear responsibility for AI-driven decisions poses risks, as educators and students might be misled to trust AI recommendations without critical oversight, undermining trust and redress possibilities.

Zhang et al. (2022) and Prather et al. (2023) express concern that unequal access to high-quality AI tools may compound existing educational inequalities, disadvantaging students with fewer technological resources. Knowles (2021) and McDonald & Pan (2020) further highlight how AI can sustain systemic inequities if ethical frameworks and equitable policies are not firmly in place.

McDonald & Pan (2020) argue for the integration of ethics training and empathy within AI development and educational curricula to foster systems mindful of users' diverse social contexts and needs. Knowles (2021) stresses the importance of preserving the human aspects of education: empathy, moral reasoning, and judgment, that may be reduced if AI is deployed without strong ethical oversight.

4.2 Perceived Data Privacy Risks and Challenges in Educational Settings

The integration of LLMs into educational settings introduces numerous perceived risks and challenges concerning data privacy, which span technical, ethical, and institutional dimensions.

One of the primary concerns is the potential for unintentional disclosure of sensitive personal data. As Abbo et al. (2025) highlight, LLMs often operate by generating responses based on user-provided input, which can include identifiable or sensitive student information such as learning difficulties, performance records, or behavioural notes. Since these models lack an understanding of contextual confidentiality, they may unintentionally reproduce or process private data inappropriately, particularly when embedded in educational environments that involve dynamic, real-time, or multimodal interactions (Abbo et al., 2025).

The lack of transparency of LLM data processing further complicates privacy assurance. Prather et al. (2023) emphasize the lack of transparency regarding how LLMs collect, store, and utilize user data, leading to uncertainty among users about what information is retained and who has access to it. This concern is increased by the centralized architecture of many LLM platforms, where data is stored on external servers, increasing the risk of large-scale data breaches and unauthorized third-party access. Similar concerns are highlighted by Andries and Robertson (2023), who note that users, particularly children, often lack awareness of how their data is handled, potentially leading to privacy violations, profiling, or the repurposing of educational data for commercial analytics without informed consent.

The legal and regulatory challenges are also prominent. Prather et al. (2023) point out that the use of LLMs hosted on third-party servers may not comply with regulations such as GDPR in the EU or FERPA in the United States, thereby putting institutions at risk of legal breaches. The lack of institutional control over how data is shared or retained on commercial platforms introduces governance concerns, particularly when educators and administrators cannot easily ascertain the limits of data access or deletion protocols (McDonald & Pan, 2020).

Accountability and responsibility for data protection remain unclear. McDonald and Pan (2020)

argue that the integration of LLMs into educational systems blurs the lines of responsibility between technology providers, institutions, and users. This obscurity increases the likelihood of privacy mismanagement, especially in cases where data anonymization is inadequate or when systems permit inference of additional private details from seemingly safe inputs.

While some studies acknowledge privacy concerns implicitly, they do not explore them in depth. For instance, Wambsganss et al. (2023) primarily focus on gender bias in LLM outputs and only briefly address privacy by noting that student data was anonymized. This suggests an awareness of ethical data practices, but the study does not provide a substantive discussion of privacy risks such as data misuse or long-term retention vulnerabilities.

Mitigation strategies are needed but currently underdeveloped. Abbo et al. (2025) highlights the importance of implementing strict technical and regulatory frameworks to monitor and control how LLMs access, store, and disseminate data. Additionally, they promote for user education on safe AI usage and stronger encryption and consent mechanisms. Similarly, Prather et al. (2023) and Andries and Robertson (2023) recommend that robust data governance policies and clearer disclosures about data practices are essential to restoring trust and ensuring safe educational use of LLMs.

Although LLMs offer significant potential for enhancing learning experiences, their deployment in educational contexts must be accompanied by rigorous safeguards to address the multifaceted risks to data privacy. These include the risks of accidental data exposure, insufficient user consent, non-transparent data handling practices, unclear accountability, and the potential misuse of personal information. As the use of LLMs continues to expand, establishing transparent, enforceable, and ethically grounded privacy policies will be critical to protecting the rights and safety of all educational stakeholders.

5 Discussion

The integration of LLMs into education represents a significant technological and pedagogical challenge, offering new ways to enhance learning and teaching processes. The reviewed studies confirm that LLMs, particularly tools such as ChatGPT, are being increasingly adopted in education, primarily due to their capacity to provide immediate, context-sensitive support and personalized learning experiences (Due et al., 2024). However, this enthusiasm is limited by a range of ethical concerns and data privacy challenges that must be addressed to ensure responsible use.

One of the most prominent findings across the literature is obvious ethical conflict between the advantages of LLMs and the dangers they bring. On one hand, their ability to support diverse learning preferences, increase engagement, and provide timely academic assistance, especially in under-resourced educational settings, is widely acknowledged (Due et al., 2024). On the other hand, ethical issues related to algorithmic bias, academic integrity, and the developmental risks of over-reliance on AI are consistently emphasized. Several studies (e.g., Abbo et al., 2025; Prather et al., 2023; Wambsganss et al., 2023) reveal that LLMs may unintentionally maintain social stereotypes, distort learners' perceptions, or slow down the development of critical cognitive and ethical reasoning skills. This duality reinforces the urgent need for ethical oversight frameworks and inclusion of educators in the co-design and governance of AI tools, an area that remains underexplored, as only a minority of studies included teacher perspectives (Vrågård et al., 2024; Knowles, 2021).

In addition to ethical concerns, data privacy risks represent a substantial challenge. Many of the reviewed studies highlight the potential for LLMs to unintentionally process or reveal sensitive student information (Abbo et al., 2025; McDonald & Pan, 2020). This is particularly concerning in educational contexts, where children and adolescents may not fully comprehend the implications of data sharing. Furthermore, the lack of transparency regarding data storage and usage practices, especially on commercial platforms, raises serious questions about compliance with legal and institutional regulations, including data ownership, user consent, and long-term retention. Several authors (Prather et al., 2023; Andries & Robertson, 2023) stress that these issues are added by the centralized nature of most LLM platforms, which limits institutional control and complicates efforts to ensure accountability.

Another significant theme in the literature concerns fairness in access and outcomes. Studies by Knowles (2021) and Zhang et al. (2022) raise important concerns about the potential for AI to reinforce existing educational inequalities. Students with limited access to high-quality digital infrastructure or AI-illiterate educators may be disadvantaged, which runs counter to the democratic promise of LLMs to support inclusive and accessible education. Additionally, the risk of AI tools increasing historical biases, if not carefully monitored and corrected, may further marginalize underrepresented student groups.

Across studies, there is a consensus that the implementation of LLMs in education must be accompanied by clear ethical guidelines, robust data protection policies, and transparency mechanisms. However, existing mitigation strategies remain underdeveloped or inconsistently applied. While some studies propose anonymization techniques, secure data storage, and user education (Abbo et al., 2025; Andries & Robertson, 2023), few offer concrete frameworks for institutional accountability or cross-stakeholder collaboration. The limited duration and scope of many interventions, often with small sample sizes and a focus on higher education, also limit the generalizability of

findings, particularly for younger learners or those in diverse educational contexts.

While LLMs have demonstrated considerable potential to enrich educational practice, their integration must be critically examined through both ethical and technical focus. Future research should prioritize longitudinal studies, include diverse learner and educator perspectives, and explore strategies for safeguarding data privacy and promoting fairness. Only by addressing these challenges comprehensively can we ensure that LLMs serve as responsible partners in the educational process, rather than sources of new forms of unfairness or negative impact.

6 Conclusion

The integration of LLMs in educational settings presents both significant opportunities and risks. On the one hand, LLMs have shown great potential to support personalized learning, foster student engagement, and bridge gaps in educational access. On the other hand, their deployment raises critical ethical concerns related to data privacy, algorithmic bias, and the risk of reducing human agency in the learning process.

This paper highlights the need for a balanced approach, one that embraces the pedagogical benefits of LLMs while rigorously addressing their ethical and technical challenges. Despite the growing interest in LLM applications, the current research remains fragmented, with limited focus on younger learners, teacher involvement, and long-term effects. Future research should adopt a multidisciplinary approach, actively involving educators, learners, policymakers, and developers in the co-design of AI-supported educational tools. In parallel, educational institutions must establish clear ethical guidelines, ensure transparency in data handling, and prioritize equity in access and outcomes.

Only through responsible, inclusive, and ethically informed implementation can LLMs fulfil their promise as transformative tools for education, empowering rather than replacing the human elements of teaching and learning.

References

Abbo, G. A., Desideri, G., Belpaeme, T., & Spitale, M. (2025). " Can you be my mum?": Manipulating Social Robots in the Large Language Models Era. arXiv preprint arXiv:2501.04633.

Andries, V., & Robertson, J. (2023). Alexa doesn't have that many feelings: Children's understanding of AI through interactions with smart speakers in their homes. Computers and Education: Artificial Intelligence, 5, 100176.

Boland, A., Cherry, G., & Dickson, R. (2017). Doing a Systematic Review: A Student's Guide. Sage.

Due, S., Das, S., Andersen, M., López, B. P., Nexø, S. A., & Clemmensen, L. (2024). Evaluation of Large Language Models: STEM education and Gender Stereotypes. arXiv preprint arXiv:2406.10133.

Dungca, P. A. P. (2023). The incorporation of large language models (llms) in the field of education: Ethical possibilities, threats, and opportunities. In Philosophy of Artificial Intelligence and Its Place in Society (pp. 78-97). IGI Global.

Fenu, G., Galici, R., & Marras, M. (2022, July). Experts' view on challenges and needs for fairness in artificial intelligence for education. In International Conference on Artificial Intelligence in Education (pp. 243-255). Cham: Springer International Publishing.

Knowles, M. A. (2021). Five motivating concerns for AI ethics instruction. Proceedings of the Association for Information Science and Technology, 58(1), 472-476.

Mamalis, M., Kalampokis, E., Fitsilis, F., Theodorakopoulosand, G., & Tarabanis, K. (2024). A Large Language Model based legal assistant for governance applications.

März, M., Himmelbauer, M., Boldt, K., & Oksche, A. (2024). Legal aspects of generative artificial intelligence and large language models in examinations and theses. GMS Journal for Medical Education, 41(4), Doc47.

McDonald, N., & Pan, S. (2020). Intersectional AI: A study of how information science students think about ethics and their impact. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-19.

Miranda, M., Ruzzetti, E. S., Santilli, A., Zanzotto, F. M., Bratières, S., & Rodolà, E. (2024). Preserving privacy in large language models: A survey on current threats and solutions. arXiv preprint arXiv:2408.05212.

Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., ... & Savelka, J. (2023). The robots are here: Navigating the generative ai revolution in computing education. In Proceedings of the 2023 working group reports on innovation and technology in computer science education (pp. 108-159).

Shalevska, E., & Kostadinovska-Stojchevska, B. (2024). Ethics in times of advanced ai: investigating students' attitudes towards chatgpt and academic integrity. International journal of Education Teacher, 27, 72-78.

Sublime, J., & Renna, I. (2024). Is ChatGPT Massively Used by Students Nowadays? A Survey on the Use of Large Language Models such as ChatGPT in

Educational Settings. arXiv preprint arXiv:2412.17486.

Vrågård, J., Brorsson, F., & Aghaee, N. (2024, October). Generative AI in Higher Education: Educators' Perspectives on Academic Learning and Integrity. In Proceedings of The 23rd European Conference on e-Learning. Academic Conferences International.

Wambsganss, T., Su, X., Swamy, V., Neshaei, S. P., Rietsche, R., & Käser, T. (2023). Unraveling downstream gender bias from large language models: A study on AI educational writing assistance. arXiv preprint arXiv:2311.03311.

Xiao, Z., Li, T. W., Karahalios, K., & Sundaram, H. (2023, April). Inform the uninformed: improving online informed consent reading with an AI-powered Chatbot. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-17).

Zhang, H., Lee, I., Ali, S., Dipaola, D., Cheng, Y., & Breazeal, C. (2022). Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-022-00293-3