THE EVOLUTION AND REPLACEMENT OF AI IN EDUCATION

Benjamin To

Department of Computer Science, Toronto Metropolitan University

Dr. Sarah Kriger

December 2, 2024

Table Of Contents

[**Table Of Contents 2**](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.riu1mqh6urhr)

[**Summary 3**](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.yp2bsnfwg2u)

[**Introduction 3**](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.e65sk79b0gdp)

[**Discussion 3**](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.yvo3b02bpy87)

[Can AI Tutor Students? Will It Help Them Learn 3](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.wdyu1dj67jcg)

[AI Grading 4](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.jxmjz8u6bq11)

[AI vs Human Experts 4](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.dygadcnqgieq)

[Grade Feedback 5](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.fxdjkwodfdbv)

[AI Privacy, Policies, and Biases 6](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.kkndbonfevsa)

[Privacy 6](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.2ce5fx1ekiye)

[Policies 6](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.c4e1pqtn9htg)

[Biases 7](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.yc5lbsdimrbp)

[Conclusion 8](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.ip9kvc3udsrc)

[Recommendation 8](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.meik4a5pwazb)

[**References 9**](https://docs.google.com/document/d/1qS1owpD6R2NpOUSnMFPT8SJXoOJdK8gsW98Klfty6j0/edit?tab=t.0#heading=h.yfwfamoz87q0)

**Introduction**

The evolution of AI technology in multiple fields from industrial, to daily life, and education has sparked widespread discussions about its potential. Especially at an educational institution like Toronto Metropolitan University, such tasks as auto-grading have been implemented into reality for a long time due to their cost-effectiveness and time-saving. As a result of this, the upcoming intention to replace many GAs (Graduate Assistants) with AI in tutoring students has been recently released. While the benefits include faster processing, personalized tutoring, and 24/7 availability no doubt, the limitations in technology as well as privacy raise questions about fairness, accuracy, and bias. This essay will give an overview of the potential of generative AI to assist in education, highlighting its strengths and weaknesses in tutoring, grading, and feedback generation, as well as its implications for privacy and policy enforcement. Overall, we come to the conclusion that although AI can serve as a valuable supplementary tool, there is still a long way to go before it can fully replace the judgment and adaptability of human GAs.

**Discussion**

Can AI Tutor Students? Will It Help Them Learn

Gen AI has the capability to tutor students effectively, with pretty big examples demonstrating their ability to help students overcome educational gaps, specifically in China AI tutoring has been deployed to help students with personalized learning experiences. The AI systems offer tailored lessons that adapt to individual learning speeds and needs and this results in improved academic performance and greater access to quality education for students who are underprivileged and may lack that ( World Economic Forum,2024).

While AI is very proficient in delivering immediate feedback and is customizable based on students, it has its limits. Research indicates that AI tools are most effective when used to assist regular human tutors rather than remove them 100%. Ai really struggles with tasks that require contextual understanding and creativity.

AI tutoring systems also risk the possibility of generating errors or misleading information when the prompts are outside of the Large language model’s training data. This could impact students negatively if they rely on AI without human oversight (Education Week, 2024). Ai can significantly enhance learning outcomes in structures and data-rich environments, but its role as a stand-alone tutor is still very very limited.

In conclusion, AI can tutor students and contribute to learning but its full potential is realized when it's in connection with human TAs who have the oversight and skill to verify its outputs.

AI Grading

A major portion of graduate assistants' jobs is grading and creating feedback on tests and assignments in the undergraduate classes that they support. This responsibility will fall to the AI if TMU proceeds with the transition from human graduate assistants’ to generative AI. The generative AI would be able to mark tests and provide constructive feedback swiftly compared to the graduate assistants. However, due to generative AI models' deterministic nature and training data, the quality of their grading and feedback would vary. This presents glaring issues in regard to the fairness and caliber of the grade received by students.

AI vs Human Experts

These issues are more prevalent when comparing generative AI to human experts, such as graduate assistants. To start, generative AI would significantly alter the grades given out in courses. In a comparison between human experts and chatGPT, they only agreed on the same marks on “30% of the [total] exams” (Flodén, J., 2024). This lack of agreement raises concerns about the grading metric of generative AI marks and shows the unreliability of AI in marking undergraduate exams and assignments. In fact, generative AI was considered more “friendly” than human experts in terms of grading. In Flodén (2024), it was noted that the generative AI would score the grades clustered around the average while avoiding failing students. Furthermore, as noted in Li et al. (2024) and Lu et al. (2021) the generative AI gave higher marks on average than the human experts. This boosting of grades is bad for the students as the lack of failing grades would cause unprepared students to continue into harder courses while lacking the necessary basics. This will in turn cause the students to have underdeveloped skills that will not help them in future jobs.

Moreover, another major limitation of generative AI when it comes to grading is its lack of ability to understand the context of the questions being asked. If the question contains specific references to content mentioned in class or data that is not contained in its training data, it would have a hard time generating a response. OpenAI, the company behind chatGPT, released a report in which they concluded that the model “ “hallucinates” facts and makes reasoning errors” (Achiam et al., 2024). Although OpenAI has a clear bias for its generative AI model, its findings have been consistent with other reputable sources. Li et al. (2024) stated that chatGPT failed at marking questions for subjects that were not contained in its training data. This limitation of generative AI is not present for the human graduate assistants who have a good understanding of the course material and the specific concepts taught in their courses. In addition, chatGPT's lack of context and understanding of the questions led it to give higher grades than it should have (Lu et al., 2021). If the answers to the questions “contained keywords related to the concepts covered by the question” (Lu et al., 2021) the AI model would still give the mark for the answer even though it may have lacked enough content to receive full marks. The graduate assistants would give higher grades to well-constructed answers rather than those with just keywords. Therefore, in order to have accurate grades that reflect the understanding of students, all grading should be done by graduate assistants.

Grade Feedback

Another major job for graduate assistants is generating feedback for the tests and exams that they have marked. This job also takes up a lot of time and sometimes, the graduate assistants are unable to provide feedback due to deadlines and a lack of time. If the feedback was just generated by generative AI, it would be done faster, however, the quality and accuracy would be lower. Li et al. (2024) stated that the students enjoyed the feedback given by the generative AI for each question. This demonstrates the importance of feedback for the grades given out and how it helps students understand what went wrong and how to do better in the future. However, as mentioned earlier, chatGPT or other generative AI might make up facts and answers when faced with subjects outside of its training data (Achiam et al., 2024). This could potentially result in feedback that is not accurate or nonsensical which can guide students down the wrong path. The best course of action is for the graduate assistants to ask the generative AI to generate feedback on the tests that they marked, then overlook the feedback to make sure it is accurate and helpful to the students (Li et al., 2024). This way graduate assistants can grade the tests or assignments then quickly generate feedback and check their quality while still finishing before their deadlines.

Overall, having the generative AI mark the tests and assignments for undergraduate courses would cause a major shift in grades given out and could negatively impact both the school and the student's understanding of the content. However, generative AI is good at quickly giving back feedback as long as the feedback is checked and the quality and accuracy are fixed by a graduate assistant.

AI Privacy, Policies, and Biases

Privacy

The main topic of replacing Graduate Assistants (GAs) with AI graders and tutors at TMU involves critical considerations regarding privacy, policies, and biases that could significantly impact students, faculty, and the institution’s reputation. As AI systems rely heavily on training and processing a large amount of data in order to work, this can cause protecting and assuring private information becomes a big challenge. Although the core goal is to help utilize the convenience and ease as well as speed up the process, this data often includes sensitive student information, such as personal identifiers, academic performance, and behavioral patterns, which must be securely stored and processed. Mishandling this data could lead to breaches of confidentiality, violating student trust and legal frameworks such as Canada’s Personal Information Protection and Electronic Documents Act (PIPEDA) or even international regulations like the General Data Protection Regulation (GDPR). For example, Marda and Narayan (2021) emphasize the risks of improperly securing educational data, highlighting cases where weak privacy measures led to widespread breaches. Without stringent protections, students could lose faith in the institution’s ability to protect their personal information, hindering the perceived legitimacy of AI in education.

Policies

Strong policies governing the use of AI are critical to ensure its integration is consistent with TMU's commitment to equity and academic integrity, as outlined in Policy 60. One of the major issues of AI is that it may sometimes place efficiency higher than fairness, resulting in oversights that may harm student outcomes. There are obvious reasons why AI is used the most when it comes to auto-grading scenarios: while AI may be the best at identifying surface-level errors or applying standard rubrics, it might struggle to evaluate more subjective aspects of student work, such as creativity or the quality of arguments. For example, AI will be the best grader when it comes to a test that includes only multiple choices, but when grading an essay or a project for a fashion student, then the principles will not be enough, and objective human opinions will be required. Moreover, instructors and GAs frequently adjust and are flexible in their feedback based on the conditions and the scenarios for each of them, a quality that current AI technologies lack. Binns et al. (2018) argue that regulatory frameworks should include mechanisms to audit AI decisions, hold developers accountable for errors, and ensure equity across diverse student populations. Without these safeguards, students may question the credibility of their grades, leading to dissatisfaction and potential disputes.

Biases

Equally concerning is the potential for bias in AI systems, which could aggravate existing inequities within the student body. Although the same algorithm and structure create consensus on a single standard evaluation metric, AI tools are only as unbiased as the data on which they are trained, and training data sets often reflect social biases. For instance, Mehrabi et al. (2021) show that AI grading systems can unintentionally penalize students from underrepresented backgrounds due to biases embedded in the data or the design of the algorithms. A typical example is the 2020 UK grading controversy, where an algorithm used during the COVID-19 pandemic downgraded students from low-income areas, causing widespread backlash. While GAs can adapt their evaluations to account for individual circumstances, AI tools lack this contextual understanding. This rigidity might disadvantage students who communicate their ideas in non-standard ways or come from diverse cultural backgrounds. For example, international students may have different reflective perspectives when the topic is about culture and customs, especially with domestic standards at the location where AI gets the data.

Furthermore, the biases could extend beyond automated AI grading and AI tutoring, which could disproportionately benefit students whose learning styles align with the AI’s programmed responses. AI in some ways helps students research information faster or have an overall understanding of new knowledge. Noble (2018) illustrates how algorithmic systems in education often reinforce systemic inequalities by favoring groups that align with dominant cultural norms. TMU must recognize that these biases may cause students who already face challenges in higher education to feel more struggle. Resulting in the fact that AI tools which can provide many supportive benefits, become less effective than humans in promoting an inclusive learning environment.

To reduce these dangers, TMU should conduct more testing of AI tools before and after deployment, invest in diverse training datasets, and guarantee that monitoring mechanisms allow instructors and staff to appeal AI choices. Additionally, the university must establish transparent policies regarding data use, consent, and accountability, as recommended by legal scholars and AI ethicists. Considering that the increase in costs and complexity will be more costly than the benefit that these measures can bring, it would be more comprehensive to negate some of the financial benefits that prompted the proposal.

In summary, although TMU's goal of maximizing costs and improving the student experience more quickly and promptly through AI is worth considering, the potential disadvantages involved cannot be ignored. to privacy, policy enforcement, and bias. If these issues are not adequately addressed, this change could harm trust, equity, and student outcomes, ultimately weakening the school's reputation.

Conclusion

Generative AI has demonstrated its ability to assist in various educational tasks, such as grading and tutoring. However, its deterministic nature and reliance on training data limit its effectiveness as a standalone solution. Human graduate assistants offer a nuanced understanding of course materials, context, and student needs, ensuring fair grading and meaningful feedback. Replacing graduate assistants entirely with generative AI could lead to a decline in educational quality, as AI struggles with subjective analysis, creativity, and contextual relevance. Furthermore, privacy concerns, biases, and the potential for inaccuracies make it clear that human oversight is essential.

Graduate assistants play an indispensable role in fostering equity and understanding in education. Their ability to adapt to diverse circumstances and provide personalized insights ensures a higher standard of learning. Therefore, while AI can support their work, it should not replace GAs entirely.

The most “optimal approach” as we computer scientists would say is to integrate generative AI as a supplementary tool for graduate assistants rather than as a replacement. AI can expedite grading and feedback generation, allowing TAs to focus on more complex and nuanced tasks. For example, TAs can mark assignments and utilize AI to draft feedback, which they then review and refine to ensure accuracy and relevance.

This collaborative model enhances efficiency while maintaining educational quality. Moreover, rigorous testing of AI tools, robust privacy policies, and efforts to reduce bias in AI systems will ensure these tools align with TMU’s commitment to equity and academic integrity. By using AI as an aid rather than a substitute, TMU can strike a balance between innovation and maintaining the high standards of education that students deserve.

**Recommendation**

Our final recommendation to TMSU is to use AI, not as a replacement for GA’s but in conjunction with them to improve student learning and maintain the high quality education students expect at TMU.

**References**

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024). GPT-4 technical report. *arXiv*.<https://arxiv.org/abs/2303.08774>

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.<https://doi.org/10.1145/3173574.3173951>

Education Week. (2024, October). What happens when an AI assistant helps the tutor instead of the student? *Education Week*.<https://www.edweek.org/technology/what-happens-when-an-ai-assistant-helps-the-tutor-instead-of-the-student/2024/10#:~:text=An%20AI%2Dpowered%20tutoring%20assistant,to%20a%20Stanford%20University%20study%20>

Flodén, J. (2024). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*.<https://doi.org/10.1002/berj.4069>

Li, K., Yang, Q., & Yang, X. (2024). Can autograding of student-generated question quality by ChatGPT match human experts? *IEEE Transactions on Learning Technologies, 17*, 1600–1610.<https://doi.org/10.1109/tlt.2024.3394807>

Lu, O. H. T., Huang, A. Y. Q., Tsai, D. C. L., & Yang, S. J. H. (2021). Expert-authored and machine-generated short-answer questions for assessing students’ learning performance. *Educational Technology & Society, 24*(3), 159–173.<https://doi.org/10.30191/ETS.202107_24(3).001>

Marda, V., & Narayan, A. (2021). Data governance in education: Implications for privacy and equity. *Data & Policy, 3*, e24.<https://doi.org/10.1017/dap.2021.24>

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.<https://doi.org/10.1145/3457607>

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

World Economic Forum. (2024, September 10). This AI tutor could make humans 10 times smarter, its creator says. *World Economic Forum*.<https://www.weforum.org/stories/2024/07/ai-tutor-china-teaching-gaps/>