COMMENTARY

Automated Essay Writing: An AIED Opinion

Mike Sharples[1] ⓘ

Abstract

This opinion piece emerged from research for the book, *Story Machines: How Computers Have Become Creative Writers*, by Mike Sharples and Rafael Pérez y Pérez, published by Routledge. While thinking of ways to promote the book, I realised that students could employ AI story generators to write essays. That led me to research automated essay writing, write a Twitter thread that has garnered 43,000 engagements, and author a blog article (Sharples, 2022). The current piece is a revised and expanded version of that article.

Keywords Transformer AI · Plagiarism · Assessment · Text generation · Generative AI

Essay Writing with Transformer AI Systems

Essays have formed part of academic assessment since the early 19th century, when students in some European universities were required to write scholarly papers for discussion in seminars (Kruse, 2006). Academic essays continue to be a mainstay of assessment in schools, colleges and universities because they are easy to set, assess depth of understanding, and train students in how to express an argument. In recent years, the scope of written assignments has expanded to include narrative, argumentative, reflective, expressive, responsive and analytical papers.

The widespread practice of setting written assignments as coursework has been criticised as laborious and unfair (Race, 2018). It is also subject to contract cheating through "essay mills" that sell written assignments to order, at fees of up to £400 for

✉ Mike Sharples
mike.sharples@open.ac.uk

[1] Institute of Educational Technology, The Open University, Milton Keynes, UK

writing a 20-page essay[1]. A study by Newton (2018) found that 15.7% of students surveyed admitted to paying someone else to write an assignment. Now, an unintended consequence of generative "Transformer" AI systems such as GPT-3 is that they democratise cheating. A student can generate an entire essay in seconds, at a cost of around 50 US cents[2]. Equally worrying, would-be academic researchers can call on these systems to generate articles for submission to conferences and journals.

How do such Transformer AI systems work? How can we detect whether an academic essay has been written by an AI? Could any good come from widespread use of such technology? What does this mean for the AIED community?

How Does a Transformer AI Work?

Take time to read this brief undergraduate student essay on "A Critique of Learning Styles". If this were your student, how would you assess the written work? What mark would you give? What are its strengths and weaknesses?

> The construct of "learning styles" is problematic because it fails to account for the processes through which learning styles are shaped. Some students might develop a particular learning style because they have had particular experiences. Others might develop a particular learning style by trying to accommodate to a learning environment that was not well suited to their learning needs. Ultimately, we need to understand the interactions among learning styles and environmental and personal factors, and how these shape how we learn and the kinds of learning we experience.
>
> **Learning Styles and Learning Modes**
>
> In the 1970s, learning styles became an important feature of educational practice. In the 1990s, "learning modalities" became a major focus of learning style research. This is the idea that we process information in different ways. For example, some individuals may be auditory learners, benefiting from listening to lectures and using recorded materials. Others may be visual learners, who gain from using visual materials such as textbooks, handouts, and computer graphics. Still others may be kinesthetic learners, who benefit from learning by movement and doing. Numerous experiments have been conducted to test this idea, and a number of different classifications of learning modalities have been proposed.
>
> In a review of this research, Dunn and Dunn (1997) found that the evidence does not support the idea that people have a particular learning mode that is dominant or that they process information in a particular way. In their study, the authors found that when students were given a particular instructional treatment, they could be classified in a particular learning modality category. But when they were given a different instructional treatment, they were classified in

---

[1] https://www.essaymills.co.uk/pricing.

[2] https://openai.com/api/pricing/. The cost to access the most powerful GPT-3 model is US$0.06 for 1000 tokens (or approximately 750 output words).

a different category. In a follow-up study in which they gave students a number of instructional treatments, Dunn and Dunn found that students could be placed in a particular category in some treatments, but not others.

Dunn and Dunn also found that the results varied by age. Older students generally fit the results of more than one modality, whereas younger students were more likely to fit the results of just one modality. This led them to conclude that style preferences are learned over the course of development. This is consistent with the idea that people are flexible in responding to the way they are taught.

References

Dunn, R., & Dunn, K. (1997). Learning styles: Research and practice. Journal of Research in Education, 7(2), 139–151.

The entire essay was written by GPT-3, a Transformer AI program. It was given the prompt "The construct of 'learning styles' is problematic because" and it generated the rest, including heading and citation. The text shown here is the first attempt. The reference at the end was generated by adding the prompt "References" to the end of the text and resubmitting.

Gaining access to GPT-3 is straightforward. Anyone with internet can sign up to the OpenAI website[3], gain an account, click the "Playground" tab, type a prompt such as the title of an essay, set the maximum length of output (up to 4000 language "tokens", or approximately 3000 words) and click the Submit button. A few seconds later, the system produces a typed and formatted text. Some companies are already promoting AI-based essay writing services[4].

GPT-3 works like a highly trained text completer of the kind found on mobile phones and email interfaces. Instead of looking back at the last few characters and using these to predict the next word or two, it can attend to the previous 750 words it has written, to continue an entire short story, blog, or student essay. The same program can also summarise a scientific article in simpler language, write a review, translate languages, and answer general questions. In short, a Transformer AI is a general-purpose language machine.

How Can We Detect Whether An Essay is Written by a Transformer AI System?

Plagiarism software will not detect essays written by Transformers, because the text is generated, not copied. A Google search shows that each sentence in the essay above is original. Two plagiarism detectors each indicated over 95% original text[5].

What better way to discover whether the essay is written by human or machine than to ask GPT-3? Here is its response when presented with the essay above and the prompt "Was that essay written by a human or a computer?".

---

[3] https://openai.com/api/.

[4] https://www.nichepursuits.com/best-ai-essay-writers/.

[5] https://plagiarismdetector.net/, https://smallseotools.com/plagiarism-checker/. The example essay has now been published online, so plagiarism detectors no longer indicate it as original.

It is difficult to say for certain, but it seems more likely that it was written by a computer. The language is very technical and does not seem to flow smoothly. In addition, the ideas are presented in a very concise, bullet-point style, which is often seen in computer-generated texts.

Impressive, yet it gave a similar response when presented with the introduction to a highly-cited research paper written by Sharples (Sharples, 2010), claiming it "does exhibit some characteristics that are often associated with AI-generated text, such as a lack of clear authorial voice and a somewhat stilted, awkward style of writing". Rather than attempt further embarrassing comparisons, we note that any sufficiently powerful program to determine whether a piece is written by human or machine could frequently be outwitted by an equally powerful AI text generator, in a futile computational arms race.

Humans fare no better than machines at detecting AI-generated essays. In a small study by EduRef.net[6], college professors were asked to grade essays produced by human writers and by GPT-3, without being informed that any piece was generated by machine. For a Research Methods topic, the machine-written essay was given a grade of C, while the human essays were graded B and D. For US History, machine and human were given similar grades. For a Law essay, GPT-3 was graded B-, while the human essays ranged from A- to F. For one topic, Creative Writing, the machine essay was failed, while human-written essays were graded from A- to D+. The professors gave similar written feedback to the machine productions as to human writers.

A comprehensive study of state-of-the-art methods to determine if a piece of extended text is human-written or machine-generated concludes that "humans detect machine-generated texts at chance level" and that for AI-based detection "overall, the community needs to research and develop better solutions for mission-critical applications" (Uchendu et al., 2021). Students employ AI to write assignments. Teachers use AI to assess and review them (Lu & Cutumisu, 2021; Lagakis & Demetriadis, 2021). Nobody learns, nobody gains.

On the surface, our sample text appears to be a mediocre to good (though very short) student essay. It is correctly spelled, with good sentence construction. It begins with an appropriate claim and presents a coherent argument in support, backed up by evidence of a cited research study. The essay ends with a re-statement of the claim that learning styles are flexible and change with environment.

But look more closely and the paper falls apart. It references "Dunn, R., & Dunn, K. (1997). Learning styles: Research and practice. Journal of Research in Education, 7(2), 139–151." There is a journal named *Research in Education*, but no issue 7(2) in 1997. Dunn & Dunn did publish research on learning styles, but not in that journal. GPT-3 has fashioned a plausible-looking but fake reference. The program also appears to have invented the research study it cites. We can find no research study by Dunn and Dunn which claims that learning styles are flexible, not fixed.

To understand why a Transformer AI program should write plausible text, yet invent references and research studies, we turn to the seminal paper written by the developers of GPT-3. In a discussion of its limitations, the authors write: "large pre-

---

[6] https://best-universities.net/features/what-grades-can-ai-get-in-college/.

trained language models are not grounded in other domains of experience, such as video or real-world physical interaction, and thus lack a large amount of context about the world" (Brown, et al. 2020, p.34). Transformers are models of language not experiential knowledge. They are not designed to be scholarly – to check academic references and ensure that evidence is grounded in fact. In human terms, they are essentially inexperienced, unthinking and amoral. They have no ability to reflect on what they have written, to judge whether it is accurate and decent.

OpenAI has provided an add-on to GPT-3 that filters bad language. However, it is unlikely that the company will produce tools to check for accuracy. Its focus is on artificial general intelligence not education. Other companies could, in the future, provide tools to check generated references for accuracy or add genuine references to an article. But these would not overcome the fundamental limitation of Transformer language models such as GPT-3: that they have no internal inspectable model of how the world works to provide a basis for the system to reflect on the accuracy and scholarship of its generated work. Research is in progress to develop explainable neural AI (Gunning et al., 2019) and hybrid neural/symbolic AI systems (Garcez & Lamb, 2020) that might address this problem.

Could Any Good Come from Widespread Use of Such Technology?

Transformer AI systems belong to an alternative history of educational technology, where students have appropriated emerging devices – pocket calculators, mobile phones, machine translation software, and now AI essay generators – to make their lives easier. The response from teachers and institutions is a predictable sequence of ignore, resist, then belatedly accommodate.

It will be hard to ignore the growing number of students who submit assignments written by AI. Turnitin, the leading plagiarism checking company, admits that "we're already seeing the beginnings of the oncoming AI wave … when students can push a button and the computer writes their paper" (Turnitin, 2020). As we have already indicated, resisting AI-generated assignments by deploying software to detect which ones are written by machine is likely to be a futile exercise. How, then, can we accommodate these new tools?

Teachers could restrict essay assignments to invigilated exams, but these are formal and time consuming. Alternatively, they could set reflective and contextualised written assignments that could not be generated by AI. For example, a teacher could set each student an independent research project, then ask for a written report on that specific project, give the student feedback on the report, then ask for the student to write a critical reflection on the feedback and issues raised by the project.

An imaginative way to incorporate AI-generated text into teaching could be for the teacher to employ Transformer AI to generate a set of alternative essays on a topic, then ask students to critique these and write their own better versions. Or set a complex question then ask each student to generate AI responses to the question and for the student to evaluate these responses in relation to the marking criteria. The

student would then write an integrative essay drawing on the AI answers to address the original question.[7]

Transformer AI can be a tool for creative writing. For example, the student writes a first paragraph, the AI continues with the second paragraph, and so on. The AI writing partner helps maintain a flow of words and also takes the story in unexpected directions, to which the student must respond. Generating a few alternative continuations to a story may help a student writer see creative writing not as a linear progression, but an exploration of a space of possibilities.

AI assisted writing exercises could focus on skills of critical reading, accuracy, argumentation and structure. Assignments where AI is not allowed could be assessed for style, expression, voice and personal reflection.

Additionally, teachers could explore with students the ethics and limits of generative AI. How does it feel to interact with an expert wordsmith that has no intrinsic morals and no experience of the world? Is writing with AI tantamount to plagiarism (Fyfe, 2022).

What Does This Mean for the AIED Community?

Reviewers for IJAIED will not be able to avoid the challenge of assessing whether a submitted article has been written with the aid of an AI system. As an exercise, we generated an entire short research paper. First, we chose at random the title of a real paper published in IJAIED: "Domain-Specific Modeling Languages in Computer-Based Learning Environments: a Systematic Approach to Support Science Learning through Computational Modeling" (Hutchins et al., 2020). Then we used GPT-3 to generate three alternative Abstracts, just from the title. We chose a generated abstract for a "review paper". Then, we added to the abstract the heading "Introduction" and requested GPT-3 to generate the paper. We followed with a prompt of "Discussion" and then "References", which GPT-3 added in neat APA format. Finally, we presented GPT-3 with just the newly generated Abstract and requested it to generate a new title for the paper. It responded with: "Using Domain-Specific Modeling Languages to Support Science Learning: A Review of the Literature".

The result is an original 2,200 word "academic paper" produced in under five minutes. The output can be read at https://t.co/RTxogLRlJT. It will probably not pass a first Editor's review but is a harbinger of a flood of papers generated with the aid of AI. IJAIED will not be the only journal to receive papers produced wholly or partly by AI, but we are well placed to lead a debate on how to detect and deal with them.

A related issue is how to respond to genuine papers in the general area of AI Transformer systems in education. Should we publish papers on new tools to automate essay writing, or software to detect AI-generated essays? Where is the dividing line between promoting competition between AI generators and detectors, and enabling new forms of academic writing assisted by generative AI?

---

[7] These suggestions are based on responses by @jennicarr8 to a Twitter discussion on rethinking assessment in an era of generative AI. https://twitter.com/sharplm/status/1534051131978047494.

This could be a pivotal time for education, as students equip themselves with powerful new AI tools that substitute for what some perceive as the drudgery of assessment. These will not just write essays for students but answer complex questions and generate computer code (Bombazine et al., 2021). An education system that depends on summative written assessment to grade student abilities may have reached its apotheosis.

Every new educational technology arrives with affordances and limitations. AI Transformer technology is a powerful general-purpose language model that is already becoming embedded in education through chatbots, text summarisers, language translators, and now essay generators and tools for creative writing. The AIED community is well placed not only to debate the application of these systems to education, but to design new generative AI tools for writing, reasoning and conversation for learning.

Declarations

Conflict of Interest There is no conflict of interest.

References

Bombazine, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S. … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P. … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901

Fyfe, P. (2022). How to cheat on your final paper: Assigning AI for student writing. *AI and Society*, 1–11

Garcez, A. D. A., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. *arXiv preprint arXiv:2012.05876*

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120

Hutchins, N. M., Biswas, G., Zhang, N., Snyder, C., Lédeczi, Á., & Maróti, M. (2020). Domain-specific modeling languages in computer-based learning environments: A systematic approach to support science learning through computational modeling. *International Journal of Artificial Intelligence in Education*, *30*(4), 537–580

Kruse, O. (2006). The origins of writing in the disciplines: Traditions of seminar writing and the Humboldtian ideal of the research university. *Written Communication*, *23*(3), 331–352

Lu, C., & Cutumisu, M. (2021). Integrating Deep Learning into an Automated Feedback Generation System for Automated Essay Scoring. *International Educational Data Mining Society, 2021*

Newton, P. M. (2018, August). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Frontiers in Education*, *3*(67), doi: https://doi.org/10.3389/feduc.2018.00067

Race, P. (2018). Is the 'time of the assessed essay' over? *University of Sussex blog article*, November 14, 2018. https://blogs.sussex.ac.uk/business-school-teaching/2018/11/14/is-the-time-of-the-assessed-essay-over/

Lagakis, P., & Demetriadis, S. (2021, November). Automated essay scoring: A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)* (pp.1–6). IEEE

Sharples, M., Taylor, J., & Vavoula, G. (2010). A Theory of Learning for the Mobile Age. In B. Bachmair (Ed.), *Medienbildung in neuen Kulturräumen. Stuttgart* (pp. 87–99). Kohlhammer Verlag

Sharples, M. (2022). New AI tools that can write student essays require educators to rethink teaching and assessment. *LSE blog article*, May 17, 2022. https://blogs.lse.ac.uk/impactofsocialsciences/2022/05/17/new-ai-tools-that-can-write-student-essays-require-educators-to-rethink-teaching-and-assessment/

Turnitin (2020). How Teachers Can Prepare for AI-Based Writing. *Turnitin blog article*, May 21, 2020. https://www.turnitin.com/blog/how-teachers-can-prepare-for-ai-based-writing

Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2001–2016). Punta Cana, Dominican Republic, November. Association for Computational Linguistics

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.