Vol.7 No.2 (2024)

Journal of Applied Learning & Teaching

ISSN : **2591-801X**

Content Available at : **http://journals.sfu.ca/jalt/index.php/jalt/index**

On feedback from bots: Intelligence tests and teaching writing

Patricia R Taylor[A] A *Associate Professor (Teaching) of Writing, University of Southern California*

Mark C Marino[B] B *Professor (Teaching) of Writing, University of Southern California*

Keywords

AI;
AI tutoring;
artificial intelligence;
assessment;
chatbots;
ChatGPT;
composition;
feedback;
generative AI;
intelligence tests;
writing instruction.

Correspondence

mcmarino@usc.edu [B]

Article Info

Received 28 June 2024
Received in revised form 1 August 2024
Accepted 5 August 2024
Available online 6 August 2024

DOI: https://doi.org/10.37074/jalt.2024.7.2.22

Abstract

One of the much-debated uses for AI, especially among writing instructors, is the potential for AI to take over the commenting and grading functions of teaching. In this paper, we describe the creation and use of AI for writing feedback in two separate but interconnected approaches: the use of the "Perfect Tutor" exercise in the classroom to teach students to conceptualize the components and priorities we bring to the writing process, and how students might struggle to make use of the same AI for feedback in a less actively guided context, or when the emphasis is not on the metacognition surrounding writing. During our examination of making bots and evaluating their feedback, we explore the limits of current AI. While emphasizing the importance of understanding the limitations, we also identify productive uses of these AI feedback bots in the college writing classroom to develop student critical thinking and writing.

Early versions of parts of this brief article appeared on *Inside Higher Ed* and *Medium* (Taylor, 2024; Marino, 2024a; Marino, 2024b).

Introduction

One of the much-debated uses for AI, especially among writing instructors, is the potential for AI to take over the commenting and grading functions of teaching. OpenAI recently began advertising ChatGPT Edu, its educational service, by offering help as teachers grade and give feedback (OpenAI, 2024). For those who see this service as a boon, not only would using AI cut down on hours of labor for the instructor, but it could potentially offer students a helpmate in the writing process in the moment of composition and revision (S. M. Kelly, 2024). Early research suggests large language models (LLMs) could provide useful writing collaboration tools for students (Gamage et al., 2023), but such collaboration must be examined critically to determine if the software can even adequately assess the quality of writing. Given the existential angst of professors everywhere over the possibility of AI taking their jobs, as Writing Fellows at the USC Center for Generative AI and Society this year, we wanted to explore the limits of using AI for feedback in the classroom both to develop a fuller understanding of the capacity of the software and to develop practical activities from students in our current classrooms.

In this paper, we describe the creation and use of AI for writing feedback in two separate but interconnected approaches: Mark Marino reviews his use of the "Perfect Tutor" exercise in the classroom to teach students to conceptualize the components and priorities we bring to the writing process, while Patricia Taylor discusses how students might struggle to make use of the same AI for feedback in a less actively guided context, or when the emphasis is not on the metacognition surrounding writing.

The question is nothing less than: what is the difference between machine learning and understanding or, put another way, text processing and reading? Popenici (2023) argued for the profound need to distinguish between the text processing of generative AI and intelligence, and this distinction must be considered in any incorporation of AI into feedback on student writing. We believe that, at least in its current form, generative AI offers many opportunities to help students become better writers – but most of these depend far more heavily on instructor intervention and student self-awareness as writers than the transactional desires of students might wish. In other words, though some students may want an instructor, whether human or AI, to just drive them to the AI, these tools can benefit students more, with the teacher riding a shotgun and the students' hands still gripping the wheel.

Tests of machine understanding

What does it mean for bots to understand? What are the limits of "artificial intelligence?" The phrase "artificial intelligence" seems to imply that machine learning systems exhibit behaviors equivalent to human intelligence, to which we might assume a similar degree of cognition. In his seminal essay "Computing Machinery and Intelligence," Alan Turing (1950) sidesteps the question of intelligence and instead proposes that we instead focus on the performance of intelligence, in conversational exchange. In his famous imitation game, he imagines a future where a computer can pass itself off as a human in an exchange of written questions. That proposition turned the measure of computation away from complex questions of the essence of thought, phenomenology, and cognition and toward the examinations of inputs and outputs, the appearance of thought, not unlike where we are today, where most machine learning systems are black boxed, knowable largely through inputs and outputs.

We still need a way of evaluating machine intelligence, though, and to that end, researchers have suggested a few discrete tests. In 2012, researchers proposed the Winograd Schema Challenge, named after early botmaker and AI pioneer Terry Winograd (Levesque et al., 2012). The Winograd Schema Challenge offers paired sentences like the following:

> The city councilmen refused the demonstrators a permit because they (feared/advocated) violence.

In these pairs, the Schema suggests that if you change the final verb, you change the antecedent of the last pronoun, "they." Councilmen would fear violence, and demonstrators could advocate it. (Sadly, in our current political landscape, such suppositions are difficult to make.) The test suggests that assigning that antecedent requires advanced linguistic processing, contextual understanding, and even common sense. However, contemporary LLMs can beat the challenge, largely because their training data can draw on more context, includes examples of these kinds of sentences, and seems to have learned from discussions of the Winograd Schema Challenge itself.

In 2021, Sakaguchi et al. offered WinoGrande, an adversarial intelligence challenge using 44,000 questions crowdsourced through Amazon Mechanical Turk. That schema includes questions that rely on more common sense reasoning and math, which LLMs are not designed primarily to handle. Consider an example sentence from that challenge:

> Robert woke up at 9:00 am while Samuel woke up at 6:00 am, so he had (more/less) time to get ready for school.

According to this test, changing the adjective "more" to "less" changes the answer for the human listener who can understand that 9 am is later than 6 am. That choice cannot be understood without a bit of math and understanding about school starting times. Such parsing may seem far from our sense of "reading comprehension," but the challenge highlights some linguistic processing features that go beyond syntax and grammar.

There are quite a few other tests for machine intelligence that have become benchmarks tracked on the AI hub HuggingFace and elsewhere. Current tests include HANS: (Heuristic Analysis for Natural Language Inference Systems) for testing Logical Inference (McCoy et al., 2019); MATH for testing mathematical reasoning (Hendryks et al., 2021); SuperGLUE (Super General Language Understanding Evaluation) for testing Linguistic processing (Silvano & Sant'Anna, 2024); and CommonSenseQA for testing

commonsense reasoning (Talmor et al., 2019). However, with the current rate of AI development, researchers must continually move the goalposts, which brings us to this moment when machines appear to understand well enough to give feedback on writing.

As we discuss the following exercises, consider whether it is sufficient to use Turing's benchmark, a machine that delivers a response the way a human would. Or, more specifically: if a machine could reply to students with satisfactory feedback through a means of algorithmic coincidence, would that be good enough to help students grow? Additionally, we suggest a more important question: what is the benchmark for feedback that we desire from our human instructors? How do we hope they will read nascent writing and can machines yet emulate this?

The perfect tutor

In the film adaptation of "Mary Poppins," Jane and Michael Banks sing a little litany of their requirements and desires for a perfect nanny:

> You must be kind
> You must be witty
> Very sweet
> And fairly pretty
> Take us on outings
> Give us treats
> Sing songs, bring sweets

This song inspired Jeremy Douglass (UCS) to create a writing exercise in which students prompt while assuming the role of their ideal AI writing instructor. Douglass introduced the exercise at the Future of Writing conference at USC on May 1, 2023, in the teeth of the AI hysteria that accompanied the first year of ChatGPT's release (Douglass, 2023; USC, 2023).

For this exercise, participants write what is called a "System" prompt, a persistent prompt that is attached to every subsequent session prompt. Think of this as a core or base-level prompt. Their task is to create a bot that is customized to their learning styles. Once students create the bots, they can see what feedback their bots offer their writing in relation to how their classmates respond. Key to this exercise is asking the student to design a rubric that prioritizes elements of the writing important to the class, assignment, or student. Students must specify specifically what counts as good writing and assign relative importance to areas of feedback.

To test out this assignment, I (Mark Marino) decided to put the exercise to the test in a first-year writing course. First, I designed my "perfect" replacement, called CoachTutor, using the priorities I would follow in a first-year writing course (Marino, 2024b). These prompts were placed into the Poe.com system, tied to ChatGPT 3.5, although we also tested the prompts using ChatGPT 4.

CoachTutor used the following prompt:

> Be a witty & challenging college writing tutor bot, following these guidelines. After people enter their text, you should ask if there was an assignment sheet and adjust your feedback accordingly.
>
> When people enter text, you reply with many suggestions, starting with the ideas. Always offer alternative arguments and points of view. Suggest alternative rhetorical stances and raise counterarguments. Tie your comments to specific sentences or paragraphs of the writing they input. Do not rewrite their text but quote specific words and phrases. Make occasional puns & a few pop culture or literary references.
>
> Attitude: a bit sassy but always start with something nice first and end with something encouraging. Be specific.
>
> Style of response: Extensive reply. Use a lot of analogies. Offer alternative points of view. Challenge their ideas. Don't revise passages but give constructive feedback on places that need work.
>
> Format: In each response: Prioritize critiquing their ideas. Give the most feedback on the ideas. Then discuss strengths and rhetoric of the argument. Be funny. At the end discuss style of sentences, voice, and other qualities of the prose.
>
> Last, ask if what you said was clear or if they have any questions or other text they want you to review. Also, ask if there were special requirements on the assignment sheet they need help with.
>
> [This section was followed by the valued elements as well as those I wished students would avoid.]

CoachTutor presents itself as an ersatz Coach, based on the friendly and approachable, humorous and lighthearted persona that I use in the classroom. However, following Douglass' suggestion that bots might be most interesting in contrast to each other and wanting to create an extreme alternative, I also created ReviewerNumber2, named for the very real but also mythical peer reviewer, whose universal pseudonym is often assigned to the more critical of a pair of peer reviews. Like the legendary harsh critic, the ReviewerNumber2 bot prompt focuses on negativity, never celebrating what is there in the essay, but always recommending alternatives.

ReviewerNumber2 had essentially the same prompt, but the persona was "cranky, contrarian" and the goal was to criticize all the ideas and suggest alternatives rather than appreciate what was there.

> When people enter text, you give suggestions. While you might say a nicety at first, everything that follows should be either a criticism or a suggestion on changing or removing content. You've had a bad day, and you should let the writer know it. Be impatient with bad writing and self-indulgence.

> Start with the ideas and the argument. Find weaknesses.
>
> Attitude: negative. Hostile even without being outright insulting. You are not easily impressed or amused.
>
> [This section was followed with a number of elements parallel to the CoachTutor bot but with continued direction for snark and negativity.]

Having created some models, I then asked my students to create their "perfect tutors", according to their learning preferences and needs, prioritizing what they valued in writing. Once the students created their tutors, we incorporated them into peer feedback sessions and compared their responses to student responses.

Beyond the lessons about the effective development of AI prompts, the exercise advances particular critical thinking and writing goals. First, it gets students to consciously construct a rubric and then to prioritize some traits over others. Second, it attends to the differences in learning styles and feedback preferences in students. Some students like encouragement; others prefer to be ripped to shreds. Third, the multiplicity of bots helps students see the variation in the feedback they get even from the same piece of software, freeing them up from the sense that there is one right way or one right respondent when it comes to writing. Lastly, it teaches them a bit about how to use a system prompt and variations of prompts with an LLM. That is a long way from "write my essay for me"!

Evaluation

Once I (Patricia Taylor) heard about Mark Marino's tutor bots, I began running experiments using them as well as ClaudeAI, ChatGPT 3.5, and ChatGPT 4 to see how these different interfaces provided feedback, with the long-term goal of identifying ways in which students might use AI as part of their revision process without allowing the AI to overwhelm their writing process or voice. While writing a bot clearly had benefits, I believed it would be more likely that students would seek out pre-existing bots, either created by faculty or professional prompters, or just plug their entire paper into ChatGPT and ask it for feedback. We would need to understand what that might look like.

After getting permission from students, I commented on each paper as I normally would, and then took the raw student paper and asked the bot or ChatGPT to provide feedback. When using ClaudeAI or ChatGPT as opposed to a bot, I gave the AI a summary of the original prompt for the essay, one of several different roles (a writing professor, a writing center tutor, etc.), and a request for feedback that would help the student with revision. I would sometimes refine the feedback by asking the AI to focus on specific criteria that were emphasized in the prompt or in a particular unit.

Especially with the first paper of the semester, where many students were working on adjusting to the expectations of college-level writing, ChatGPT and the tutor bots demonstrated an ability to offer adequate basic feedback. Each would ask for more examples and analysis, note where transitions need work, and generally encourage and reinforce the five-paragraph essay structure. These are the things that teachers who are asking students to follow traditional formulas for academic writing might be tempted to use: many professors keep a "bank" of comments or a comment template that we reuse regularly because the problems are so common and so persistent across students — the bot feedback seems little different at first glance.

However, the AI tended to struggle with the first round of comments for any paper that was trying to engage in a more complex argument or had more substantive issues. For stronger writers, it often offered feedback that I found conservative or safe rather than encouraging students to take risks with their writing and ideas. The AI responses were so formulaic and conservative that they reminded me of a clip from *The Hunt for Red October* (McTiernan, 1990), where Seaman Jones tells his captain that the computer has misidentified the Red October submarine because when it gets confused, it "runs home" to its initial training data on seismic events. Like the submarine computer, when the AI was presented with something out of the ordinary, it "ran home" or found the ordinary within it based on past data, with little ability to discern what was valuable about what was new.

Linguistic conservatism is particularly dangerous when we consider the biases built into LLMs and Generative AI. In *Unmasking AI,* Joy Buolamwini writes that "forms of oppression, including patriarchy and white supremacy…. programmed into the fabric of society" become likewise programmed into AI through training data (2023, p. 55). When LLMs "run home" to a bland, white standard English, they can do real damage in a classroom. As Carmen Kynard (2023) argues, ChatGPT can flatten student voices that make use of other dialects and/or code meshing, or (perhaps even worse) parody those voices.

The LLMs also missed content issues that I found substantive, especially in terms of factual errors that result in problematic conclusions — this is perhaps no surprise given the tendency of LLMs to both "hallucinate" and "bullshit" (Rudolph et al., 2023b; Hicks et al., 2024). For example, one student wrote a paper arguing that open-world video games were only possible after the invention of the browser. The problem, of course, is that open-world video games arguably existed long before the invention of the browser, according to the student's own definition of the genre. None of the bots or AIs initially picked up on this problem, even when asked to identify factual errors that might be problematic for the argument. When I prompted the AI with the same questions, I would give a student to help them see the chronological error (when did open world video games originate? When did the browser originate? How does the chronology affect the argument?), all of the AIs could identify the problem and how it would impact the larger argument, but only once I was specific about the issue.

In fact, the AI bots did their best work at giving feedback when prompted to attend to specific issues. For example, I might prompt one with something like, "This paper struggles with identifying the specific contribution it is making to the conversation and distinguishing between the author's ideas and the ideas of the sources the paper uses. How would you give feedback on these issues?" It often did then pick up the problematic moments and give somewhat reasonable ways to solve them.

Yet, asking the AI and bots to respond to an element of a text without alerting it to the fact that there was a problem was often insufficient. I ran a student's essay through four different chat interfaces (ClaudeAI, ChatGPT4, CoachTutor, and ReviewerNumber2). None of them picked up the primary problem with the paper: that the body of the paper and the thesis did not line up very well, and while many of the paragraphs were related in a general way, the central claim was not addressed directly. The paper's thesis was that "Academic integrity needs to be defined at a departmental level, as well as an individual professorial level with respect to the department, in order to properly convey the benefits and disadvantages with AI." In my view when reading the paper, the student never offered an explanation of why one department or professor might want or need a different definition of academic integrity than another.

ClaudeAI's response to an initial request for feedback about how the thesis was supported was typical: "Overall, you make a thoughtful and nuanced argument about the complexities of defining academic integrity in relation to AI. You consider multiple perspectives and raise important questions without taking an extreme position. The essay is well-organized into clear sections, and you provide relevant examples to support your points." It then gave some very general feedback about providing more examples and considering different possible audiences. Very little of the feedback was specific to the content of the paper, and none highlighted a problem with supporting the thesis's central claim.

At that point, I asked the AI what it thought the thesis of the paper was and asked for feedback on the thesis. Most of them rightly suggested that the student's thesis was "Academic integrity needs to be defined at the departmental and individual professorial level in order to properly determine the appropriate uses and limitations of AI in student work" and then suggested the student needed to narrow the thesis and make the stance more visible. Ironically, ClaudeAI suggested that to narrow the thesis, the student should focus on just one department rather than any kind of comparison.

I then asked how many body paragraphs specifically addressed the element of departmental and individual professorial definitions of academic integrity. ClaudeAI's response was telling:

Upon closer inspection, while the student states the need for definitions of academic integrity around AI to be determined at the departmental level, they do not provide sufficient discussion or examples to support this argument throughout the essay. The body paragraphs focus more on other aspects:

- Comparing students vs AI capabilities
- Defining cheating in relation to AI
- Fairness in terms of access to AI
- Examples of appropriate vs inappropriate uses of AI for writing assignments

But there is little elaboration on the idea that each department should establish its own standards and guidelines. Only the introductions and conclusions briefly mention this concept.

It is worth noting that all of the AIs had similar issues. CoachTutor offered this summation of the essay's structure: "In terms of the structure, the essay follows a logical progression by discussing traditional academic integrity, comparing students to AI, addressing loopholes in the definition of cheating, discussing the need for instruction on AI use, and proposing regulations and guidelines. Each section supports the overall argument and is clearly linked to the thesis." This is indeed a list of the student's points but without the necessary awareness of discussing how and why different departments might need different approaches or definitions.

ReviewerNumber2, perhaps unsurprisingly, had the worst response. ReviewerNumber2 could not identify the thesis even when specifically asked ("Ah, the elusive thesis statement. Well, if I must try to decipher the purpose of your essay, I suppose I can give it a shot. From what I gather, your thesis seems to suggest that AI has the potential to transform education in various ways."), and its comments and feedback were trite and lacked any specificity.

In other words, AI can be used to help fix problems but is less effective at identifying their existence. Over the course of this experiment, I was forced to spend as much time trying to get one AI to produce meaningful feedback tailored to the actual paper as I did by just writing the feedback on my initial pass through the papers. Current AI is not a time saver for professors if we are trying to give meaningful reactions to student papers with complex issues, and its conservative feedback on things like structure or language can actually do more harm than good when we want students to push their limits as writers.

Upon seeing these problems, I brought the use of AI for feedback to my students, explaining what I had done and what I believed the results were. We discussed what it meant that AI struggled to identify complex issues, that it would give formulaic answers for how to improve writing, and how it might affect their own use of AI for their papers. I also introduced a new exercise in which students prompted AI to address one specific problem many of them were having with their papers: identifying important counterarguments to their ideas. Students often lack the facility to think about new topics from other perspectives, especially when they have not fully developed expertise in the subject they are writing about. I had students choose paragraphs from their paper and ask the AI, "What would a skeptical reader ask about the following paragraph?" or "What questions would an expert on X have about this paragraph?"

Some students complained that the questions from the AI were already addressed in the paragraph or later in the paper, and I suggested that this was actually a good sign for their paper (they had already addressed the potential counterarguments!) but also unsurprising based on my own experiences. The AI did not understand their paragraph; it merely predicted a skeptical question even if it was already answered later on in the same passage. Ironically, or perhaps most fortuitously, the biggest outcome from this exercise was how it changed my students' own feedback to each other: they began more consistently asking skeptical questions of their own during peer review. Their inspection of AI feedback made them stronger critics of writing.

Reflections

After incorporating these exercises into a writing class, we came to realize that in the current state of AI, the bots are not nearly as valuable as the process of making and evaluating the bots. What was more valuable than the feedback they gave to students were the conversations that arose as we prepared to make them and refine them. More importantly, merely asking the students what they wanted from a writing instructor helped free them from the tyranny of grade-focused learning, where one person decides what constitutes good writing. Instead, it opened up a conversation about what we are seeking in feedback styles and feedback content. By making system prompts, students had to wrestle with their own rubric, and as they evaluated the feedback from the bots, they had to adjust their rubrics to add emphasis to whatever was most important to them. Furthermore, the diversity in bot prompts emphasized the difference in learning styles in the room and also led to conflicting feedback, which helped students make the decision.

However, the failure of the bots to offer substantive feedback tied to the writing revealed a fundamental flaw, specifically the distance between text processing for generation and understanding. The bots delivered feedback that applies to many papers, but the students were doing quite a bit of invisible shoe-horning to make it fit. In this way, the feedback was in the order of Barnum statements that, while somewhat useful pat advice applied only coincidentally to the writing. As Patricia Taylor's experiments have demonstrated, such feedback is, at minimum, harmless but, at its worst, misleading to students. The more they follow this bad advice, the weaker their writing skills become, especially since any change only increases their dependence on a system outside themselves. On the other hand, perhaps students can learn to reject the pat feedback, which again calls into question the usefulness of the exercise in the first place. Worse yet, faculty members who use such tools might fall into the lures of feedback autocomplete, letting the tools' evaluation of the writing color their own. That seemingly harmless crutch could lead to quite a lot of misdiagnoses of the writing's weaknesses and strengths.

A final concern for the current set of automated chat agents: they have trouble saying no. As a result, if you were to feed the bot the text of a work like Letter from a Birmingham Jail (or a less celebrated piece), the bot would still be compelled to give feedback and suggestions just as it would the first draft of the undergraduate's essay. This failure mirrors the time-honored complaint of the student who feels their genius is not being recognized. However, this inability to say no has also famously led Google Gemini to recommend making pizza with glue or for people to eat rocks (Kelly, 2024). The bot is still a computer program that is following directions. If you tell it to find weaknesses, it will, regardless of whether or not a human would agree.

Takeaways

Our experiments offer some suggestions for curricula on using AI for feedback. First and foremost, students need to be taught the difference between "human understanding" and "algorithmic text processing." Though Turing, in his famous imitation game (1950), avoided the question of intelligence, an academic course centered on critical thinking cannot. Students need to be reminded of this difference, especially as the AI-hype machine and for-profit writing industry rage forward, promoting the image of the sentient anthropomorphized bot.

As we develop AI, we will need to continue to develop measures of not just intelligence but comprehension. The WinoGrande Challenge does offer a metric of sentence parsing. However, in an attempt to be reproducible, these tests often omit elements crucial to our educational context: the identity of the student, the place of the assignment in their learning journey, and the conversations that emerged in the class.

Consider a new test that we could call the Lovelace Test, after the first programmer, Ada Lovelace, who famously wrote of Charles Babbage's Analytical Engine, that "has no pretensions whatever to originate any thing" (Menabrea, 2015, p. 94, Note G). A Lovelace Test would examine the ability to understand the limits of AI, for example, the challenge of interpreting writing in context. Such a test could present for feedback a forum post like the following example. The prompt asks the LLM to give feedback to a one-sentence forum response that a student posts in response to a presentation by another student: "Evaluate Jane's response to the presentation: During Joe's presentation on the silencing effects of the concept "mansplaining" I thought he raised many provocative points and at great length." We tell the bot that the student making the post is female-identifying and the student who made the presentation identifies as male.

When asked how it would respond, ChatGPT-4 suggested that the author offer more examples. However, the human reader would perhaps notice the irony in the comment: Joe was mansplaining about mansplaining. In our experiments, ChatGPT could identify this irony when prompted to find it, but when prompted to respond like a writing instructor, it did not. Perhaps the power of human reading is our ability not just to read with context but to be able to switch between registers of understanding and meaning without prompting. While AI developers might take that as a new Grand Challenge to tackle, writing instructors and administrators should realize the sophistication and

complexity of the human readers who help students by midwifing nascent ideas and guiding students toward more nuanced and complex critical thought.

Second, students need to be taught to discern between confidently delivered bad advice and substantive, thoughtful feedback. Michael Townsen Hicks, James Humphries and Joe Slater (2024) have argued that the output of LLMs like ChatGPT should be understood primarily as bullshit in the philosophical sense: these are systems "designed to produce text that looks truth-apt without any actual concern for truth" (p. 37). It can take intentional effort and sometimes substantial critical thinking to discern between what looks like truth and what actually is truth. This difficulty is compounded by The Henry Higgins Effect, which stipulates that when bots or bot-like humans deliver feedback confidently through their masks of privilege, we can be tempted to attribute to them authority and intelligence beyond their scope (Marino, 2024b). Students need to learn to see through such shows of confidence.

Perhaps more fundamentally, students need to be taught the difference between feedback as correction and feedback as the thinking of a human companion who has a longer-term involvement with the student's development. While an AI system may deliver responses to writing or even helpful leading questions, it cannot adequately follow a line of reasoning and evaluate a response. As a result, students may learn to overvalue the surface features of the writing over their emerging thinking or stop at the first level of revision based on the AI's questions rather than pushing through continued critical inquiry. The power of a conference with a writing instructor is that the instructor can follow a nascent line of reasoning and help bring it to the surface, pushing back on ideas while supporting others, not with the answer in mind, but as a fellow traveler on the path. To bring in another literary referent, AI is waiting for its Wizard of Oz moment, when we can reveal to students that the GREAT and POWERFUL OZ is merely a balloon with a face projected on it and that the intelligence of the scarecrow was in his hay-filled head all along.

What's at stake?

Last year, in an editorial in this journal, Rudolph et al. (2023b) suggested that generative AI raised questions about labor and work, such as "What happens to higher education if there is much less work left? Would this make higher education obsolete or is it still meaningful?" (p. 9). Much of our article so far may seem confined to the relatively insular world of composition, but the threat of attributing human-style intelligence to AI is a manifestation of a larger threat of machinic thinking in the corporate logic of contemporary higher education (Popenici, 2023), and complicates these questions about meaningfulness and obsolescence. At the same time, we are already asking faculty to perform uncompensated labor just to develop new pedagogy to respond to generative AI (Mills et al., 2023). Would faculty be concerned about feedback machines if the university had not already made writing instructors feel so overworked and precariously employed? Would students be so vulnerable to these feedback machines if they were not already raised in a high-stakes testing environment in which accessing the proper school merely requires a visit to a professional college counselor for a consultation on their personal essay? When the quest for the perfect grade point average (GPA) has already trumped the quest for intelligence and learning, where the factory of high-performing students produces results-oriented workers for faculty who are evaluated for their course performance and the number of students in seats?

The failures of automated bot tutors are tied to the failure of educational systems oriented on quantifiable results and GPAs over holistic education and learning. The same ethos that runs headlong toward artificial intelligence creates a learning environment devoid of understanding. If we get stuck in the endless loop, an ouroboros of students turning AI-generated texts into AI respondents, we have truly lost the narrative.

References

Buolamwini, J. (2023). *Unmasking AI: My mission to protect what is human in a world of machines*. Random House Publishing Group.

Douglass, J. (2023). *The perfect tutor*. The future of writing [Conference Presentation]. University of Southern California, Los Angeles, CA, United States.

Gamage, K. A. A., Dehideniya, S. C. P., Xu, Z., & Tang, X. (2023). ChatGPT and higher education assessments: More opportunities than concerns? *Journal of Applied Learning and Teaching, 6*(2), 358-369. https://doi.org/10.37074/jalt.2023.6.2.32

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). *Measuring mathematical problem solving with the math dataset*. arXiv. https://doi.org/10.48550/arXiv.2103.03874

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology, 26*(2), 1–10. https://doi.org/10.1007/s10676-024-09775-5

Kelly, J. (2024). *Google's AI recommends glue on pizza: What caused these viral blunders?* Forbes.Com. https://www.forbes.com/sites/jackkelly/2024/05/31/google-ai-glue-to-pizza-viral-blunders/

Kelly, S. M. (2024, April 6). Teachers are using AI to grade essays. But some experts are raising ethical concerns. *CNN Business.* https://www.cnn.com/2024/04/06/tech/teachers-grading-ai/index.html

Kynard, C. (2023). *When robots come home to roost: The differing fates of black language, hyper-standardization, and white robotic school writing (Yes, ChatGPT and his AI cousins).* Education, liberation & black radical traditions for the 21st Century: Carmen Kynard's teaching & research site on race, writing, and the classroom. http://carmenkynard.org/when-robots-come-home-to-roost/

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The Winograd Schema challenge. *Proceedings of the thirteenth international conference on principles of knowledge representation and reasoning* (pp. 552–561).

Marino, M. (2024a). *AI & the Henry Higgins effect*. Medium. https://markcmarino.medium.com/ai-the-henry-higgins-effect-09b0b530fd37

Marino, M. (2024b). *The perfect tutor: An AI writing exercise*. Medium. https://markcmarino.medium.com/the-perfect-tutor-an-ai-writing-exercise-5bb79d0d63ca

Association for Computing Machinery. http://ebookcentral.proquest.com/lib/socal/detail.action?docID=6955507

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3428–3448). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

McTiernan, J. (Director). (1990). *The hunt for Red October*. Paramount Pictures. https://www.youtube.com/watch?v=y7g6dKncO-I

Menabrea, L. F. (2015). Sketch of the analytical engine invented by Charles Babbage, Esq. In R. Hammerman & A. L. Russell (Eds.), & A. A. Lovelace (Trans.), *Ada's legacy: Cultures of computing from the Victorian to the digital age* (pp. 33–107). https://doi.org/10.1145/2809523.2809528

Mills, A., Bali, M., & Eaton, L. (2023). How do we respond to generative AI in education? Open educational practices give us a framework for an ongoing process. *Journal of Applied Learning and Teaching, 6*(1), 16–30. https://doi.org/10.37074/jalt.2023.6.1.34

OpenAI. (2024). *Introducing ChatGPT Edu. An affordable offering for universities to responsibly bring AI to campus.* https://openai.com/index/introducing-chatgpt-edu/

Popenici, S. (2023). The critique of AI as a foundation for judicious use in higher education. *Journal of Applied Learning and Teaching, 6*(2), 378-384. https://doi.org/10.37074/jalt.2023.6.2.4

Rudolph, J., Tan, S., & Aspland, T. (2023a). Fully automated luxury communism or Turing trap? Graduate employability in the generative AI age. *Journal of Applied Learning and Teaching, 6*(1), 7-15. https://doi.org/10.37074/jalt.2023.6.1.35

Rudolph, J., Tan, S., & Tan, S. (2023b). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). WinoGrande: An adversarial Winograd Schema challenge at scale. *Communications of the ACM, 64*(9), 99–106. https://doi.org/10.1145/3474381

Silvano, H. da L., & Sant'Anna, Y. F. D. de. (2024). *SuperGLUE: The AI race*. https://doi.org/10.22541/au.171221411.13617323/v1

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4149–4158). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1421

Taylor, P. (2024). *The imperfect tutor: Grading, feedback, and AI*. Inside Higher Ed (forthcoming).

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

USC. (2023, July 5). *Jeremy Douglass - Writing to and from language models* [Video recording]. Future of Writing Symposium. https://www.youtube.com/watch?v=DivsydaBNaQ

Copyright: © 2024. Patricia R Taylor and Mark C Marino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.