

## THE CHRONICLE OF HIGHER EDUCATION

---

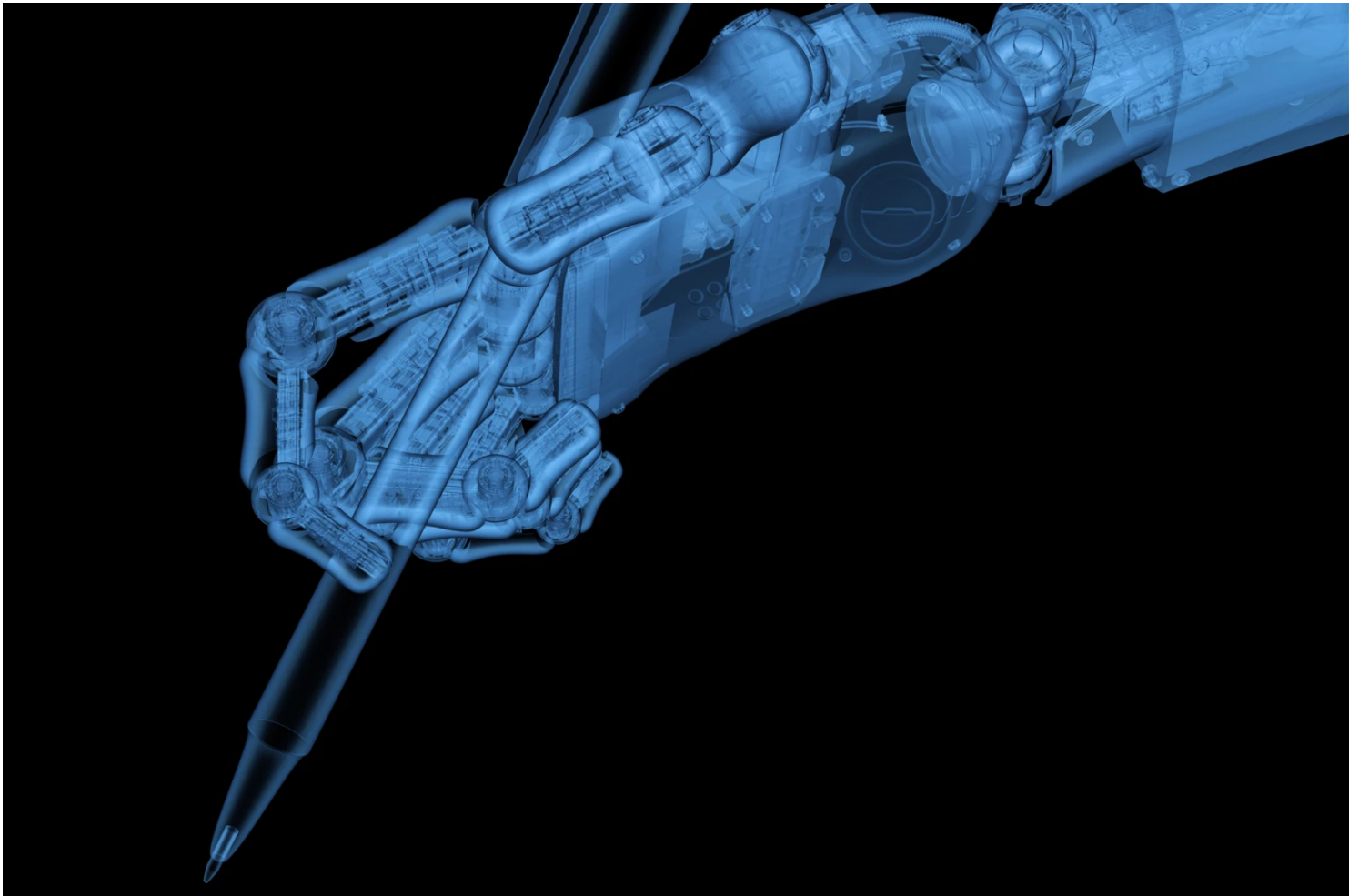
ADVICE

# ChatGPT Just Got Better. What Does That Mean for Our Writing Assignments?

An educator who tested the new GPT-4 before its release offers advice for faculty members on how to respond.

By *Anna R. Mills*

MARCH 23, 2023



GETTY IMAGES

A lot of academics were in the midst of devising policies and testing pedagogical responses to ChatGPT — the chatbot released in November that generates plausible-sounding, original text at a user’s command — when OpenAI this month announced a new version of the underlying software: GPT-4. It’s all been a bit dizzying. But as a writing instructor of 17 years, I was among those who tested the new version over the past six months. So I want to share a few observations on what to expect, how the update should affect our response to ChatGPT, and what this jump in sophistication suggests for the future.

Company officials contacted me back in August to be part of an effort to identify and mitigate risks in their latest and most-advanced model, a process they referred to as “red teaming.” They had noticed me posting on OpenAI’s developer forums and on social media about what text generators might mean for the teaching of writing. Under a nondisclosure agreement, I agreed to test GPT-4’s capacity to generate credible academic prose.

---

#### FROM THE CHRONICLE STORE

---



REPORT

### The Creativity Challenge

Teaching Students to Think Outside the Box

[Visit the Store](#)

---

Initially, as I compared GPT-4’s relatively sophisticated responses to the flatter outputs I had grown accustomed to from its predecessor, GPT-3, I felt amazement and dread. (The cloak-and-dagger associations of working under an NDA for the first time probably added to the frisson.) The prose that GPT-4 produced was more precise and articulate, a bit less boring, and a bit more substantive in the way it showed the connections between ideas.

Here’s an example. I asked both the old and the new models to summarize an article from *The Atlantic*, “[Today’s Masculinity Is Stifling](#).” Both outputs are accurate, but here’s how they differ:

- The GPT-3.5 version is simpler and [less precise](#), with sentences like: “The author recounts her son’s desire to wear dresses to school and the reactions he faced from classmates and parents.”
- By contrast, the [GPT-4 wording](#) gives more information on the same point. It includes a sense of the boy’s self-assurance in the face of likely negative reactions: “The author describes her own experience with her son, who confidently wore dresses to school despite potential backlash.” It correctly describes *The Atlantic* article as framed around the mother’s reactions, not just the son’s choices.
- The GPT-4 summary offers greater variety in sentence structure and word choice, as shown in the contrast between the older model’s simple sentence and the newer model’s complex one.

Despite the increased sophistication of GPT-4, however, my amazement and dread waned after a week or so. It had not gained understanding. This model was better, but it was also the same: Its outputs were still often simplistic and formulaic, and I found myself frequently describing them as “underdeveloped” or “lacking specifics.” Like the older versions, this one sometimes made up facts and sources, and its outputs contained mistakes in reasoning and analysis.

As I settled down to try to evaluate GPT-4's capacities, I was also considering what these changes should mean for the way I teach. I have not found answers to all of the questions that I — and no doubt many other faculty members — are puzzling over. But I have become increasingly convinced of a few basic points.

## We Can't Out-Prompt It

**It's a dead end to focus on designing prompts that AI won't be good at.** Faculty discussions have sometimes centered on how to add requirements to a writing assignment that ChatGPT might not be able to satisfy, such as quotations, discussion of real sources, or analysis of alternate media. Some educators recommended asking about hyper-local issues or about current events that occurred after the software was trained (for GPT-4, the cutoff was September 2021).

But GPT-4 can do some things, like incorporating quotations, that earlier versions weren't great at. In the testing stage, I fed the model with sources, current news, or whatever information it needed to fulfill the requirements (Bing Chat now combines GPT-4 with web search to incorporate current content). I could give it the text of an article and ask it to write an analysis. Even if the assignment drew on more text than would fit in the prompt window, there were workarounds; a user could break that text up into multiple prompts.

Computer scientists don't agree on whether artificial general intelligence is possible or imminent, nor do they agree on what further capacities are likely to emerge in language models. (OpenAI's [mission statement](#) takes it as a given that "artificial general intelligence — AI systems that are generally smarter than humans" — is coming.) However, scientists do agree that [we have not reached the endpoint](#) for these models. It seems futile for faculty members to spend their energies figuring out what a current version can't do.

**Don't depend on multimedia assignments to evade AI for long.** Various commentators, [including myself](#), have proposed assigning the analysis of images and video as a strategy for preventing students from auto-generating essays. That may still work as a deterrent in the short term, but not for long.

GPT-4 can generate textual analysis of images. You can't paste an image into ChatGPT yet, but OpenAI [released samples](#) of GPT-4 image analysis and has granted some developers access, so surely an app is on the way. Various companies are working on AI that can describe video as well.

**Don't depend on genres like personal narrative and metacognitive reflection to prevent students from using AI.** Those genres seem human, but keep in mind that GPT-4 is proficient at mimicking them. For example, it can write a semi-credible, if clichéd, [literacy narrative](#) and a pretty decent [pseudo-reflection](#) on the writing process. (Of course assigning personal writing may still help motivate students to write and, in that way, deter misuse of AI.)

**Don't assume that if AI can do something, it's not worth assigning.** It is tempting to imagine that the kinds of writing a computer can produce are exactly the kinds devoid of interest or learning value. In December, writing teacher John Warner wrote, "If AI can replace what students do, why have students keep doing that?" He [recommended](#) changing "the way we grade so that the fluent but dull prose that ChatGPT can churn out does not actually pass muster." Require students to "demonstrate synthesis," he argued, and "ask them to bring their own unique perspectives and intelligences to the questions we ask them."

His words and teaching advice are inspiring, but let's be clear that GPT-4 can now "pass muster" on those kinds of prompts. It can certainly [compare and contrast texts](#). I haven't seen it produce something truly original, but it can [mimic the results of critical thinking that strives for originality](#). It can make up [text that reads like a "unique perspective"](#) if you ask it, for example, to write "from the perspective of a Latinx single mother" and to throw in some "vivid details." And as Sam Bowman, a professor at New York University working on AI and language, explained at a [recent panel discussion](#), future models are likely to get "voicier."

## We Can't Count on Detecting It

**It may be more and more difficult to distinguish AI prose from student writing.** In the short term, we may be able to recognize that an essay was generated in ChatGPT if it reads too differently from that student's usual style, or if it is too fluent in academese. In testing GPT-4, I noticed that it was not very good at mimicking the styles of developing writers or variants of standard English. I gave GPT-4 a short piece of student writing with errors (used with permission) and asked it to continue in the same style. The result: GPT-4 [spit out grammatically flawless](#) prose. An attentive professor would probably notice such a marked shift in the student's writing. Then I [repeated the experiment](#) and instructed GPT-4 to insert errors. It added a few doozies to its first paragraph only, generating phrases like "she point out the mental health challenges faces by adolescents," whereas the student's first paragraph had only contained minor mistakes, like a missing apostrophe.

However, given the improvements in sentence-syntax sophistication from GPT-3.5 to GPT-4, it seems probable that future versions will get better at copying an individual writer's syntax and error patterns. Students may soon be able to feed a model of their past essays into an AI program and get it to generate new drafts in their style.

Will faculty members be able to use software to help us distinguish auto-generated text from human writing?

Thus far, [not reliably](#). Existing detectors such as OpenAI's classifier and GPTZero have been shown to identify human-written text as "likely AI" a significant portion of the time. Such false positives could lead to false accusations. Besides, where is the legal permission to feed student data to the company that offers the detection software? It's also relatively easy to circumvent the detectors by tweaking the AI text. I still wonder if there can be a positive role for detection software in helping establish a norm of transparency around the use of AI text. But at this point, there is little prospect of a detector that could give us an answer, rather than a probability, as to the origin of an essay.

So much for the things we can't control. What steps, you may be wondering, can we take?

## Focus on Motivation and the Writing Process Itself

I may be pessimistic about out-prompting ChatGPT or detecting AI text, but I am not at all worried that the act of writing has become less valuable to students. Writing practice continues to be intensely rewarding for students and central to intellectual growth in college. If we want to prevent learning loss, why not focus on encouraging students to write? The release of GPT-4 underscores the value of several accepted best practices for college writing assignments:

- **Assign writing that is as interesting and meaningful to students as possible.** Connecting prompts to real-world situations and allowing for student choice and creativity within the bounds of the assignment can help, but those

are only two among myriad possibilities. Why should we ever stop experimenting and reflecting on how students engage with our prompts?

- **Communicate what makes the process of writing valuable.** No one creates writing assignments because the artifact of one more student essay will be useful in the world; we assign them because the process itself is valuable. Through writing, students can learn how to clarify their thoughts and find a voice. If they understand the benefits of struggling to put words together, they are more likely not to resort to a text generator.
- **Support the writing process.** If writing feels less intimidating and overwhelming because students are doing it step by step, they will be less likely to resort to text generators out of desperation. Assigning and giving feedback on prewriting, drafting, revision, and reflection are already best practices — as most writing instructors, writing-studies scholars, and writing-across-the-curriculum proponents well know.
- **Focus on building relationships with students as a way to help them to stay engaged.** As we support the writing process via conferences and check-ins, one-on-one or in small groups, we can show that we respect students as individuals and find their ideas interesting.

Yes, such teaching methods are more labor and time-intensive than simply assigning a paper and grading the final product, with no input along the way. Teachers without training in writing pedagogy may need additional faculty development. And institutions should look to provide structural support for these approaches, like teaching assistants and reduced class size.

Many of us go into education to share our love of learning. We long to encourage intrinsic motivation in students, and doing so may well be the most effective way to prevent misuse of AI.

## Explore the Nature and Risks of AI With Students

**Don't wait until you feel like an expert to discuss AI in your courses.** Learn about it in class alongside your students. AI capacities are changing too quickly underfoot for higher ed to take its usual time to deliberate. Students and faculty members alike are seeking to understand what ChatGPT and other language models can do, and whether or how we want to use them. Why not investigate together? Start by discussing a high-quality reading or video on the topic (see this [Quick Start Guide to AI and Writing](#) for some ideas). Approaching AI in a spirit of open inquiry and listening to students can help build trust and engagement, and prevent misuse.

**Expect emotional responses to AI, and help students prepare.** Humans tend to attribute human qualities to computer systems, even when we theoretically know better. Our impulse to anthropomorphize, known as the “[Eliza effect](#),” has been documented for decades.

Interacting with a sophisticated system like GPT-4 can amplify that tendency in ways that are poorly understood. Kevin Roose, a technology columnist for *The New York Times*, has covered AI extensively, but nonetheless said he felt “[deeply unsettled](#)” by his conversations with Bing Chat (which, it was later revealed, was running GPT-4). A Google engineer, [Blake Lemoine](#), became convinced that the company's LaMDA language tool was sentient.

How might chatting with AI systems affect vulnerable students, including those with depression, anxiety, and other mental-health challenges? That remains to be seen. But we can at least help by demystifying the technology, resisting anthropomorphic characterizations, and introducing discussion of the emotions that may arise for its users.

**Teach students to be on the lookout for authoritative-sounding gibberish, everywhere.** Undergraduates aspire to write at an academic, college level, and so they are especially vulnerable to being taken in by ChatGPT's seeming eloquence. With GPT-4, the outputs are going to seem more authoritative than ever. That's all the more reason to help our students learn how to spot well-written, seemingly well-documented nonsense.

You can do that by showcasing how a confident, entitled, and eloquent academic style can be used in service of an utterly stupid point.

For example, I asked ChatGPT (running GPT-4) to "explain for an academic audience why people who eat worms are more likely to make sound decisions when it comes to choice of life partner." It responded with a brief academic paper that concluded: "While there is no direct causation between worm consumption and sound decision-making in life partner selection, the correlation can be better understood through the examination of underlying traits that are common among individuals who consume worms. Open-mindedness, adaptability, and nonconformity are qualities that contribute to a more discerning approach to personal relationships and partnership."

**Explore AI policy and societal impacts with students.** OpenAI delayed the release of GPT-4 for half a year while it tried to mitigate potential harms. It came nowhere near eliminating them. The company is [quite explicit about dangers](#) still posed by this software. It has released a [sobering, if not outright terrifying, list of possible harms](#), including weapons proliferation, security breaches, personalized phishing, and a flood of disinformation. That document concludes: "While our mitigations and processes alter GPT-4's behavior and prevent certain kinds of misuses, they are limited and remain brittle in some cases. This points to the need for anticipatory planning and governance."

[Critics say](#) the company should not have released the model if it knew about such serious risks, and it should assume responsibility for the results. But OpenAI and its critics seem to agree that guardrails are needed. The college classroom seems like the ideal place to debate these issues and propose guardrails.

After all, we in higher education have a role to play in establishing those guardrails. We shouldn't just scramble to adapt our teaching as the tech morphs before our eyes. A recent essay, "[Now the Humanities Can Disrupt 'AI'](#)," argued that professors are needed in a larger societal conversation about shaping the future of systems like ChatGPT. As a writing teacher and textbook author who found myself testing this startling new software, I want to encourage academics and administrators not to be intimidated by AI, either by its technical nature or by the complexity of the issues it raises for us. There are many unknowns, but there are also simple steps we can take to move forward.

*We welcome your thoughts and questions about this article. Please [email the editors](#) or [submit a letter](#) for publication.*

TEACHING & LEARNING

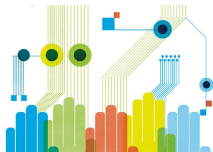
TECHNOLOGY

STUDENT SUCCESS

[Anna R. Mills](#)

Anna R. Mills is an English instructor in the California Community Colleges system and curates the resource area [AI Text Generators and Teaching Writing](#) for the Writing Across the Curriculum Clearinghouse. She is the author of the free and open textbook, [How Arguments Work: A Guide to Writing and Analyzing Texts in College](#).

RECOMMENDED ARTICLES

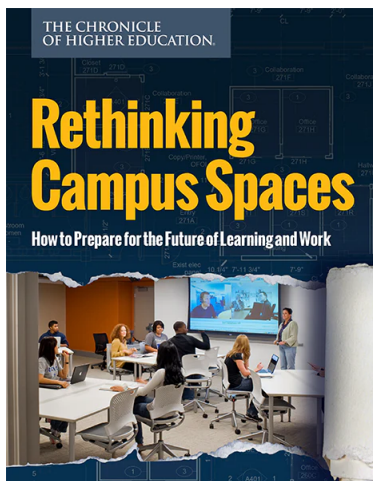


NEWS  
**Cloud Technology Can Lift the Fog Over Higher Education**



NEWS  
**Breaking Down Menus Digitally, Dish by Dish**

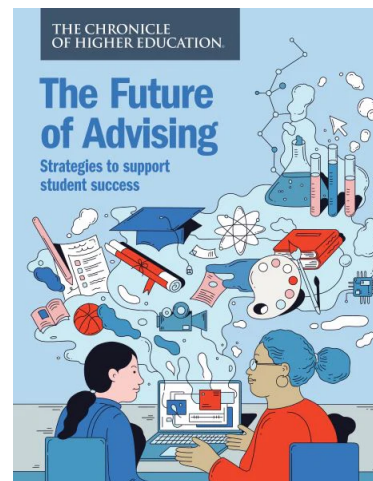
IN THE CHRONICLE STORE



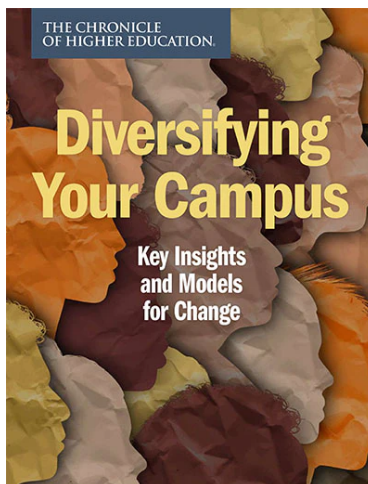
**Rethinking Campus Spaces**



**Building a Faculty That Flourishes**



**The Future of Advising**



**Diversifying Your Campus**



1255 23<sup>rd</sup> Street, N.W. Washington, D.C. 20037  
© 2023 The Chronicle of Higher Education