Alien Minds Immaculate Bullshit Outstanding Questions

College in the age of ChatGPT.

By Trey Popp

In June 2021, Chris Callison-Burch typed his first query into GPT-3, a natural language processing platform developed by the San Francisco-based company OpenAI. Callison-Burch, an associate professor of computer and information science, was hardly new to AI chatbots or the neural networks that power them. He's been at the forefront of machine translation since the early 2000s, and at Penn he teaches courses in computational linguistics and artificial intelligence. Besides, digital assistants like Siri and Alexa had already woven NLPs into the fabric of everyday life. But the jaw-dropping fluency of OpenAI's new model pitched him into a "career existential crisis."

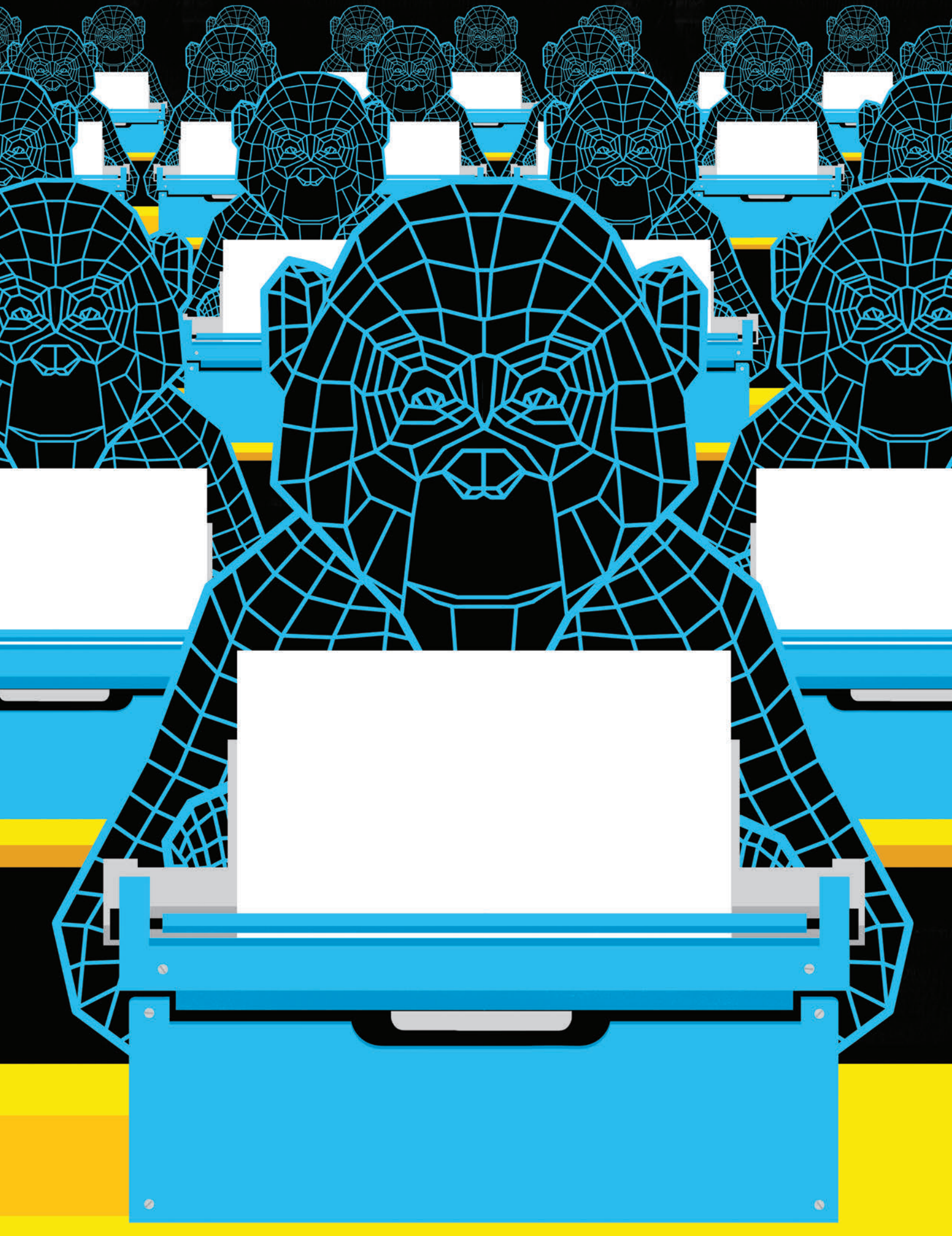
It could respond to prompts with cogent, grammatically impeccable prose. It could turn plain language into Python code. It could expand bullet-point outlines into five-paragraph essays—or theatrical dialogues.

"I was like, 'Is there anything left for me to do? Should I just drop out of computer science and become a poet?'" he later recollected. "But then I trained the model to write better poetry than me."

On November 30, 2022, OpenAI publicly released a refined version called ChatGPT. Its shock-and-awe debut quickly gave Callison-Burch plenty of company on campus. On February 1 he went to a meeting convened by Penn's Center for Teaching & Learning (CTL) to address the anxiety and excitement racing through faculty lounges—especially after the bot had passed a Wharton operations management exam administered to it by Christian Terwiesch, the Andrew M. Heller Professor. "It was probably the best-attended CTL meeting ever," Callison-Burch recalled, with a wry chuckle, a couple weeks later. So many people came that CTL director Bruce Lenthall split them into three sessions—two comprising social sciences and humanities faculty and one that blended professors of math, engineering, and physical sciences with counterparts from the University's health schools.

They'd come for varied reasons. "Some people were just alarmed," Lenthall said. Having ingested vast swathes of internet text, ChatGPT and other so-called generative AI tools are exquisitely adapted to serve as "sophisticated plagiarism machines," in the words of Eric Orts, the Guardsmark Professor in Wharton's department of legal studies and business ethics, who'd experimented with ChatGPT in an MBA course and discussed it within the Faculty Senate executive committee. Other attendees had yet to engage with the tools at all and simply wanted to learn about them. A third group sensed a chance to get in on the ground floor of a revolutionary change. "They suggested that this could be exciting and open up possibilities," Lenthall recalled, "but they didn't really have a good idea of what those might be."

ILLUSTRATON BY CHRIS GASH



Most participants fell somewhere in the middle—worried about the threats ChatGPT posed to established modes of teaching and evaluation, but curious about its potential to advance the scope or pace of instruction. "What was the most gratifying to me," said Lenthall, who is also an adjunct associate professor of history, "was that all the faculty came to the conclusion that they really needed to think through the question: What is it most critical that my students learn to do on their own? And when is it most appropriate for them to do something with another tool?"

Their search for answers gave the spring semester a hothouse atmosphere of probing and experimentation.

Wharton associate professor Ethan Mollick, a Ralph J. Roberts Distinguished Faculty Scholar and academic director of Wharton Interactive, not only permitted but in some cases *required* students in his innovation and entrepreneurship courses to use generative AI platforms, which he likened to "analytic engines." Meanwhile, on the other end of campus, astronomy professor Masao Sako and his analytical mechanics students asked ChatGPT to solve homework problems. "It returns answers and explanations that sound plausible," Sako said, but "failed on every single one." Given the confident authority with which ChatGPT announced its defective solutions, Sako concluded that the tool might indeed have some utility in the realm of upper-level physics. "I've told my students to continue using it to get some practice on identifying errors, which is a useful skill."

Penn Integrates Knowledge (PIK) University Professor Konrad Kording, who teaches psychology and neuroscience with a focus on neural networks and machine learning, emerged as a pithy generative AI maximalist. "It's just a mistaken opportunity for any student to not use ChatGPT for any possible project they're working on," he declared at a late-February panel discussion sponsored by the School of Arts & Sciences' Data Driven Discovery Initiative (DDDI). "It's just incompetent. We should give them bad grades for *not* using ChatGPT." Yet Eric Orts was finding that when he let his MBA students use it for some assignments, it tended to lead them toward bad grades—in the form of "deadening" prose—all on its own. "I'm convinced that there are positive uses emerging for this in the real world," he told me. "But in general I was not impressed by the answers I got from students using it."

Neither was Karen Rile C'80, a fiction writing teacher in the English department whose experimentation with chatbots goes back to a primitive model developed by AOL at the turn of the century. "What I value in writing is specificity, sharpness, clarity—and it fails on every level," she reflected. "It's like a bad student writer who writes in a way that's very generic, with lots of vague cliches and phrases. It feels blurry. I think that it'll probably get sharper and better, but I can't imagine it's ever going to do anything that's literary quality. It'll be very formulaic."

Yet Rile kicked off her fiction seminar this spring by assigning her students a piece by a writer who'd used GPT-3 as a kind of a coauthor. "I wanted to get ahead of it at the beginning of the semester," she told me. Then, in mid-March, she brought in Callison-Burch and his PhD student Liam Dugan EAS'20 GEng'20, who focuses on natural language processing, to give a guest lecture about generative AI and creative writing.

All 10 professors I interviewed, plus another four who participated in that DDDI panel discussion and several with whom I spoke informally, expressed a similarly open attitude. Sako's dim view of ChatGPT's analytical chops didn't keep him from seeing its potential to boost the conceptual sophistication of his mid-level coding class. Mollick mixed breathless boosterism with a running list of warnings about its boundless propensity to deceive users. Skeptics were on the lookout for positive use cases, and enthusiasts frequently offered insights about generative AI's limitations.

I spent the first half of the semester trying to learn from all of them. Bouncing between epiphanies and provocations, I rollercoastered through dizzying loops of intellectual whiplash. Gobsmacked amazement would curdle into stomach-churning dread, veer back into excitement, only to fizz out into underwhelmed deflation.

My crash course led me to several tentative conclusions, one of which may be useful to state at the outset.

A year and a half after Chris Callison-Burch's AI poetry experiment, his quasi-demoralizing success struck me as deriving from one fact above all others: The reason GPT-3 bested him in rhyming verse is that Callison-Burch is not a poet and doesn't really wish to become one.

Therein lies a key to thinking about two dynamics that on the surface may seem opposed: the stubborn mediocrity of most text-based generative AI; and its massive potential to change the nature of work, social relations, and many other aspects of contemporary life—including higher education.

Here is the story of my journey.

BABE IN THE WOODS

This is not one of those articles whose second section reveals that the first was written by ChatGPT. But it wasn't for lack of trying. I just couldn't get the tool to produce prose that didn't make me cringe. (Since every tool serves some purposes better than others, let me specify that mine were limited to the journalistic enterprise. That didn't include coding, for instance. But insofar as journalism shares higher education's fundamental aim—seeking and conveying useful truths and insights—my experience may illuminate some of the challenges and opportunities this technology poses for institutions like Penn.)

ChatGPT "wrote" grammatically flawless but flaccid copy. It served up enough bogus search results to undermine my faith in those that seemed sound at first glance. It regurgitated bargain-bin spec-

ulations about the future of artificial intelligence. When prodded to probe controversial topics or claims, it typically either reproduced a familiar center-left bias, split the difference with mealy-mouthed pablum, or shrank from engaging at all (which became more common as OpenAI reinforced its guardrails). It offered second-rate and sometimes nonsensical editorial suggestions for article drafts that hit my inbox needing cures that I would have loved to outsource.

It excelled at generating productivity-sapping amusement. ChatGPT spun happily-ever-after tales about mischievous dwarf monkeys and praised the benefits of wearing your socks over top of your shoes. It dished up relationship advice as sensible as the innumerable advice columns it had doubtless digested in its training—and dinner recipes whose surface resemblance to *Bon Appetit* masked defects that would ruin your Tuesday night and possibly your cookware. And it lied. Oh, did it lie. Brashly and exuberantly, with a devilish knack for sprinkling in just enough truth to fool a person even about their own past. It's an eerie feeling when a machine lists articles you've written, with titles and topics so plausible that it takes checking your own archive to reveal them as utter fabrications.

These traits arise from the way large language models (LLM) like ChatGPT are designed. You can think of them as supercharged versions of your smartphone's autocomplete feature. Consider a barebones version designed to operate at the level of individual letters after being trained on an English dictionary. If you asked it for a one-syllable word beginning with the letter *q*, it would infer that the next letter would almost certainly be *u*. Its subsequent calculation would be less clear-cut, but would exclude *b*, *c*, *d*, and most other consonants. Purposefully allowing for a degree of probabilistic variance, it might pass over the most common next vowel, *a*, to extend the string to *qui* on the way to its final output: *quid*. That's essentially how ChatGPT works,

"There's a desire not to be replaced that gets in our way of using it well."

only at the level of words, and with a training regimen so intensive that it can infer that what probably comes next is *pro quo*—unless contextual clues lead it instead to place a period at the end of a sentence about a Manchester United fan wagering against Liverpool to win £10. It has also been trained to respond to questions and commands, using the same basic autocomplete methodology.

Crucially, it doesn't matter if there was a *quid pro quo*, or if Liverpool actually crushed Man U—only that a plausible sentence can be written about it. Hence my frustration about ChatGPT's propensity to "hallucinate" facts. It kept trapping me in a maddening double-bind: The only viable way to judge its output was to possess enough expertise (to recognize a recipe flaw that would elude a novice cook; to know that the University of Pennsylvania did *not*, despite ChatGPT's stubborn insistence—"Yes, I am sure"—racially segregate its undergraduate programs until the 1960s) that there wasn't much point in asking the question to begin with.

Beyond minting false facts, Eric Orts fretted that AIs trained to predict words based on an all-you-can-eat internet buffet might wreak a subtler sort of social harm. "One of the words that really comes to mind for me is *genuineness*," the Wharton professor reflected. "There's a lot of bullshit in the world, as [the philosopher] Harry Frankfurt has said. There's a lot of stuff online that nobody believes—and there's no fact-checking, and there's no responsibility for saying something that you really mean. You're just throwing stuff at the wall and seeing what sticks. I worry a little bit that this is another technological phenomenon that's whittling away that sort of genuineness—that sense that we really trust one another, when we're talking to each other, to be saying what we really believe."

My frustration with ChatGPT's formulaic output did not surprise Lyle Ungar, a computational linguist in Penn's department of computer and information science. He'd experienced the same thing while coauthoring a January *Los Angeles Times* opinion piece with Angela Duckworth G'03 Gr'06, the Rosa Lee and Egbert Chang Professor of psychology, urging educators to figure out ways to "use tools like GPT to catalyze, not cannibalize, deeper thinking."

"I tried using it to write my *LA Times* article," Ungar said in March. "I tried really hard to have it find interesting metaphors. [And it] failed. … It was just giving me trite stuff. These things are statistical models that produce something that tries to capture the average—the most likely words. By definition they will be formulaic."

That also explained the main instances of creativity I succeeded in coaxing from the tool, which involved mashing one brand of formulaic prose into another. As long as there's enough of something on the internet, those possibilities

were endless. So even though ChatGPT was lousy at aping the style of Kurt Vonnegut or William Faulkner—possibly because copyright protections may have shielded their works from OpenAI's training-data vacuum—it could nail a Robert Parker Barolo review right down to the whiff of pipe tobacco. But why waste a minute on that when I could squander 10 on tasting notes styled after the King James Bible ("And lo, the Elio Grasso Runcot Barolo 2015 did pour forth from the bottle, a deep and lustrous garnet…"), *Pulp Fiction* ("Yo, this Elio Grasso Runcot Barolo 2015 is one bad motherf***er"), or a Bollywood musical (melding ChatGPT's ken for cheesy rhyming couplets with choreography cues for "rhythmic dancing").

ChatGPT's output seemed most interesting precisely when it was least useful. Yet even these formula mashups soon grew tiringly formulaic, exhausting their amusement value. Meanwhile I was stuck on square one in the main game. What was I doing wrong?

Virtually everything, according to Ethan Mollick.

GURUS AND SKEPTICS

In mid-February Wharton's most prominent AI enthusiast suggested that I'd managed to hit on ChatGPT's signature weaknesses. "It's not a good lookup engine," Mollick said. "It doesn't understand food, so it's just making up things that look like recipes—so it's going to be garbage." It doesn't "understand style" in the manner of AI image generators that can mimic Cézanne or Seurat. And it's a useless guide to current events, since it lacks up-to-date information and also "has guardrails slammed into place, because without them it would be happy to generate conspiracy theories, or harassing letters, or violent threats—because it doesn't care."

What I needed to learn was the art of prompting the tool to deliver outputs that would actually help me. "There's a desire to not be replaced that gets in our way of using it well," Mollick mused. "Because we're kind of happy when it doesn't work—and then we move on. But you're leaving a lot of value on the table.

"The problem," he said, "is that most people don't try to incorporate it into their workflow. They bounce off it because it's not as good as them. But you can train it to be better at doing your stuff."

ChatGPT, which doesn't learn from user interactions, can't actually be trained. But users can be. Students in his entrepreneurship class were now using generative AI to do "three times more" than he'd previously expected of them. "They're writing code—and often they don't know how to code. They're doing product designs and posters—and I wouldn't have expected them to do graphic design before this. They are writing ad copy. I wouldn't have expected them to."

He cast AI tools as equalizers. "People who aren't very good at generating ideas, this generates ideas for you." ChatGPT will gladly serve up 40 ideas for a new kind of toothbrush, as Mollick showed in a Twitter post. Even if 39 of them stink, one might spark a half-decent concept you can try to refine—perhaps by asking the tool for cost-cutting advice, using your brain to evaluate it, and moving to the next step.

"I expect the ideas to be of higher quality, because they're using these tools to actually do work," he continued. "I've had students talk to me about how they weren't good writers, and as a result they weren't taking that seriously. Maybe English wasn't their first language, or maybe it was another reason. And now they're good writers. They write emails and letters and they're much better quality and they get more reactions."

ChatGPT's formulaic output currently suits it best for "low-stakes stuff" like performance reviews and other bureaucratic banes, he conceded. And its untrustworthiness means users need to doublecheck absolutely everything. But it's a mistake to fixate on those weaknesses, he said. "You're coworking with an alien mind that has access to all human knowledge, is eager to please, but also lies a lot. If you think about it that way, there's a lot of uses for that.

"When I get stuck on a paragraph, I feed it in and let it finish the paragraph for me," he continued. "Do I keep it? Not really—but the hybrid paragraphs that I cowrite with ChatGPT are often the ones that people quote the most." And as generative AI gets better, it will become even more valuable to anyone seeking to "overcome the inertia associated with staring at a blank page," by producing first drafts that a user can refine.

"Some people are getting it faster than others," he told me. "But I'm a teacher! I have to figure out how to teach people this. In the world that's coming—or the world that just arrived two months ago—not being good at prompt-crafting is going to hurt you."

It's not enough to say *Write an essay explaining why X is more persuasive than Y*. "You need to say: *You are the writer for an academic journal. You care about accuracy and you use interesting word choices. You don't repeat yourself. You don't use cliches. Your goal is to communicate to the audience clearly but using sophisticated writing.* And then you give it what you want it to write, and you'll get very different results."

Mollick described an example of this on his Substack page, *One Useful Thing*. First he asked ChatGPT to "write an essay with the following points: humans are prone to error; most errors are not that important; in complex systems, some errors are catastrophic; catastrophes cannot be avoided." It responded by expanding the bullet points into three cogent but generic paragraphs. Then Mollick appended extra instructions: "Use an academic tone. Use at least one clear example. Make it concise. Write for a well-informed audience. Use a style like the New Yorker. Make it at least 7 paragraphs. Vary the language in each one. End with an ominous note."

He called the six-paragraph result "typical of how generative AI works: you don't

always get what you ask for, but you can push toward something unique and interesting by playing with prompts."

The second output was undeniably more elaborate than the first, and it struck a suitably ominous tone at the end. It deployed a clear example, choosing the 2011 meltdown of Japan's Fukushima nuclear reactor. Yet in other ways it struck me as less impressive, even malign. Using repetitive prose that bore no resemblance to the *New Yorker*, it articulated an argument that was somewhere between wishy-washy and self-contradictory, declaring that the Fukushima "disaster could not be avoided" immediately after having listed several avoidable causes of the accident. After trying to fill in some of my own ignorance about the Fukushima meltdown, I came to wonder whether ChatGPT's training data included a 2012 paper by the Carnegie Endowment for International Peace titled "Why Fukushima Was Preventable."

Given the nature of neural net architecture, there was no way to know. But it seemed like ChatGPT, when asked to serve up a clear example, semi-randomly picked Fukushima out of a black-box hat labeled *interchangeable catastrophes* and used it to short-circuit the entire point of analytical writing—which is to apply reason to carefully examined evidence in order to draw a conclusion, not start with a conclusion and illustrate it with a hastily selected example that might just as easily support a contradictory thesis. There's robust scholarship on the proneness of complex systems to catastrophe. But ChatGPT appeared to have tainted its own "reasoning" partly because its fidelity to the prompt outweighed any other concern.

Loosing MBA students on AI bots to churn out posters, HTML code, ad copy, and emails is one thing. But in this context (if not his classroom), Mollick looked to be getting out over his skis.

Two days after I shared this observation with him in an email, along with another about the nonsensical way ChatGPT had explained its production of a clever micro-story he had elicited from it, Mollick published a Substack post titled "How to Get an AI to Lie to You in Three Simple Steps." It added more no-no's to his list: asking ChatGPT or Microsoft's new Chat Bing more than it 'knows'; assuming it is a person; and asking it to explain itself.

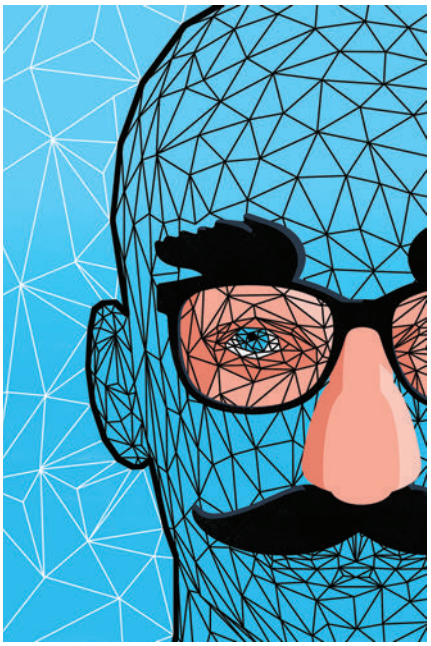
"It can help to think of the AI as trying to optimize many functions when it answers you, one of the most important of which is 'make you happy' by providing an answer you will like. It often is more important than another goal, 'be accurate,'" he wrote. One consequence can be "plausible, and often subtly incorrect, answers that feel very satisfying."

Yet "even knowing all of the above," Mollick confessed, "I keep getting fooled." After all, these tools mold plausible bullshit into authoritative, grammatically perfect declarations by design. They actively promote the illusion of personhood— referring to themselves with the first-person "I"—by design. When asked to explain their answers, they obligingly slather a second opaque coat atop the first, by design.

And if they're slick enough to trick the academic director of Wharton Interactive, where might they lead the rest of us?

Bruce Lenthall wondered the same thing. "What we want our students to be able to do—and humans to be able to do—is to weigh the evidence and figure out what conclusions make the most sense," the CTL director said when reflecting on ChatGPT's Fukushima essay. "And this is removing that." The black-box nature of LLMs compound the problem. "It's not possible for us to go back and say, *let me look at what leads you to this conclusion*. Even if we're trained to do that very thing already, we don't have the capacity." That opacity, he concluded, "is so pernicious because it undermines the kind of thinking we want to teach people to do."

Konrad Kording, a crackling conversation partner with a puckish flair for



"You're coworking with an alien mind that has access to all human knowledge, is eager to please, but also lies a lot. If you think about it that way, there's a lot of uses for that."

devil's advocacy, put a different spin on it. Essayists shouldn't be asking ChatGPT to plug an example into their prose; they should instead ask it to list and elucidate 10 examples from the scholarly literature, use their judgment to determine which one to deploy, and then let the AI thread it in with its trademark grammatical fluidity.

"It draws from a much better set of sources than humans could. But at the same time, it's much worse than humans at evaluating for local logical consistency," Kording said, describing ChatGPT's propensity to cast contradictory facts as being complementary. As a

result, "large language models make it more valuable to think at a high level and less valuable to polish your sentences, and put the comma in the right place, and make sure everything is perfectly grammatical. All of these things are now very automatable—but it just means that, in a way, we get closer to the process of just *thinking* very clearly."

At the DDDI panel he struck an even more provocative note. "In reality you don't actually want to teach your students writing, in my view," he said. "Ultimately, the reason why you want to teach them writing is because there's something about understanding the logic that is necessary to good writing." But words themselves, he suggested, "are just the glue" that binds logical chains of thought together. "The really big thing—the place where students fail—is building proper narratives," he went on. "ChatGPT is very bad at that. So arguably, by allowing students to use those tools, you allow them to do more of what you really want—which is get the logic right, get the narrative right, all those things that are what writing is *really* about—instead of making writing be primarily about words.

"The raw superficial aspect" of word selection, he concluded, "is not a great thing to be grading our students on. In the future, no student ever will not have access to something like ChatGPT. So why do we prepare them for a skill that they will no longer need in their future life?"

Stuck in the comparatively mundane present, I finally found a productivity-boosting LLM power move: automated interview transcription. It was hardly error free, but these models have crossed a threshold I've dreamed about for years. I could now spend 10 minutes correcting what once took me an hour to do unaided.

Yet I continued to hit walls when trying to use AI to either hone my thinking or cast it in engaging prose. ChatGPT could raise awful writing to mediocrity but steered elegant passages in the same direction. (A therapeutic suggestion for aspiring novelists: ask it to rewrite the first paragraph of *One Hundred Years of Solitude*.) When prodded for critical feedback and editorial advice for drafts that needed it, the problem was less its low batting average—throwaway suggestions cost nothing to ignore—than its routine failure to identify the one or two most necessary interventions. Which raised the thorny matter of expertise yet again, especially given how many suggestions would point a less-experienced editor toward formulaic dullness. ChatGPT, in these contexts, worked like bleach: capable of cleansing a snot-stained pillowcase—or sucking the color out of any pattern that's been embroidered or woven with care.

And that's when it occurred to me that I might be exactly the wrong person to judge it.

"The things that you're in the top one percent in the world at doing, it probably won't be as good as you," Mollick had told me. "But there's a lot of work that we all do where we're not the top one percent, or we don't need to do top-one-percent work." Wherever I rank as a writer and editor, I've been doing both for 25 years. The disruptive power of generative AI may lie at the other end of the spectrum. (That's why it holds particular promise for anyone trying to navigate work or life in a second language.) It can be a generic scribe for someone who can't write, a middling coder for someone who can't code, a generator of ideas for someone who doesn't have any. None of those things is very flattering to its users. But it's also a timesaver for people who don't have enough of it. And that's the biggest market there is.

Does that portend a future clotted with illimitable slime wads of insipid text? Perhaps. But that's not exactly a new dynamic. The last quarter-century has familiarized us with the substitution of cheap, middling-quality goods for better but costlier ones; that's why it's so easy to buy a chair that disintegrates in five years than one you can pass down to your children. And maybe that's not the right way to look at it. Graphite tennis rackets fixed a million Sunday serves without diminishing the sheer awe that Steffi Graf wrung out of them.

The question for colleges is twofold: How to guide students toward mastery of generative AI, and how to prevent the tools from hobbling students' intellectual growth in other fields.

Chris Callison-Burch, who cheerfully describes himself as "the most amateur of amateur writers," likened working with generative AI to learning to play a musical instrument. "You want to play the guitar, well, you've got to practice. It's the same way here. You can make it produce super clunky, terrible prose straight out of the box. That's easy. But to make it sing for you, you have to train yourself.

"There's an exciting thing happening right now," he added. "If you think about it, we've trained ourselves over the years to do the boringest web searches imaginable. I have fun examples of people trying to search the web from 1999, when I was college, and they would ask awesome things—complete English sentences—it was really great. And those never really worked, so we learned to just give a couple of keywords. And that's really sad, because we tuned ourselves to what the system could do, and kind of lost our creativity in coming up with questions. And now we have a totally new modality where you can ask it super interesting, detailed things and it'll generate stuff. So we can retrain ourselves to think about how we can interact with knowledge on the web."

Computer science professor Michael Kearns, the founding director of Penn's Warren Center for Network and Data Sciences and coauthor of *The Ethical Algorithm* ["Gazetteer," Nov|Dec 2020], observed that the context matters. "I think people are most impressed by [generative AI] in settings in which there's not a right answer and the expectations are low," he said during the DDDI panel. "The higher the standard you're

holding it to—and the more specific your use case is, and the greater extent to which there's a factually correct answer—these models are quite far from being very helpful in those domains."

But beyond their helpfulness—which may well improve—Kearns questioned the wisdom of using AI bots to generate substantive text at all. "To me, writing isn't some means to an end, or a final artifact," he said. "Writing clearly and creatively reflects thinking creatively and clearly. And I personally don't know of any substitute for writing to force myself to have that clarity of thought. So I don't think that we should be encouraging our students to use ChatGPT as much as possible. I think that will do a disservice to them in many, many ways."

STUDENTS AND TEACHERS

The spring semester brought a bumper crop of ideas about how colleges might use generative AI to boost teaching and learning. The most prolific source was Ethan Mollick, whose rapid-fire brainstorms befitted a professor in a field whose mantras include "fail fast and iterate." The strength of his ideas varied, in my view, but he was doubtless doing other teachers a service by exposing so many of them to public scrutiny. Mollick takes pedagogy seriously, as do many of his Penn colleagues.

The ideas I encountered tended to fall into one of three categories: direct student engagement with generative AI; teacher-mediated engagement geared toward saving professors' time; and a time-intensive, blended style of engagement that I found most intriguing. I came to think of them largely in terms of how vulnerable they are to the tools' factual unreliability and self-explanatory opacity.

That calculus may vary according to the academic context. "It's worth remembering that different disciplines teach different kinds of things," said CTL director Bruce Lenthall. "ChatGPT's ability to help me code things might allow me to ask really complicated questions in chemistry. But if I'm teaching computer science, it might be undermining the skill I want students to learn."

Masao Sako concurred, adding that the instructional level matters, too. When it comes to coding, "I definitely do think it could be a problem in basic intro classes," he said. "But at the same time, I think it's actually quite useful for upper-level classes." Next year he's considering telling advanced students: "Go ahead and use ChatGPT, but know that the questions that I'm going to be asking you are going to be much more complicated than what ChatGPT can simply tell you."

Lyle Ungar added that AI's capabilities, and attitudes about them, are likely to change as the technology matures. "In the short term, it's just a stupid assistant

SIDEBAR

The Coming Economic and Ethical Earthquake

"I've gone through several moments of realization about AI that have transformed my thinking," Chris Callison-Burch said during a February panel discussion. "One is that a computer program could be racist. If you'd told me that in college, I just wouldn't have understood what you were taking about: *It's an algorithm, that's nonsense!* But it is encoded in data and can be biased. The other I've started to change my thinking about is: *What is the obligation we have to people whose data we're training on?*"

Getty Images had recently filed suit against the parent company of AI image generator Stable Diffusion for allegedly copying and processing more than 12 million copyrighted photographs "without permission … or compensation … to train its highly lucrative model." Meanwhile a group of visual artists brought a separate class action seeking compensation for damages and an injunction to prevent future harms. "If Stable Diffusion and similar products are allowed to continue to operate as they do now, the foreseeable result is they will replace the very artists whose stolen works power these AI products with whom they are competing," their legal representative asserted. "AI image products are not just an infringement of artists' rights; whether they aim to or not, these products will eliminate 'artist' as a viable career path."

"This is really important, because it's the first case we can think of," said Konrad Kording. Even if the impact of any particular artist or photographer's work is technically trivial to a generative AI model's training or output, "if we view them as a group, then we have a million people worldwide who made a pretty decent living," Kording observed, "and now the things they created make computer scientists rich. Is that a fair deal?"

No panelist argued in the affirmative, but none thought that these lawsuits would prevail, either—partly because of how novel generative AI is to existing frameworks of intellectual property law.

"I'm pretty certain," said Michael Kearns, "that in the next decade massive bodies of law will be rewritten to compensate people who generate content that affects trained models. I think it will take at least 10 years. And this is not special to generative models," he added. For instance, the European Union's General Data Protection Regulation protects an individual's right to be forgotten: "You can ask to have your data deleted from storage. But what if I trained a predictive model using your data? Do I need to remove your data and retrain the entire model [at an exorbitant cost]? And do I have to do that every single time somebody asks to have the data removed? Or is there some kind of argument that any particular individual's contribution to that model is sufficiently infinitesimal that you don't have to do anything about it?"

Ethical and legal issues around training AI models may just be the first tremors in a series of escalating economic earthquakes. "There's a lot of other things that Penn graduates are doing," said Kording, "that this technology will be coming for as well."

that helps," he said. "But there's a forecasting rule: people always overestimate change in the short run and underestimate change in the long run. Is ChatGPT going to change your life in the next few years? No. It's going to help make life a little more efficient. And it might be embarrassing not to use it, the same way it's embarrassing not to use Google. But in the long run, I think it really will start to change the way people think and teach—the same way that Mathematica has," he said, referencing a software whose interactive visualizations have become a classroom staple. "Mathematica hasn't really changed math. But it is a core tool that I couldn't imagine teaching an intro math course without."

For these reasons, Lenthall does not foresee the University issuing a blanket academic policy regarding the use of generative AI. "It seems like that's antithetical to the way Penn does things," he said. "Because if I am teaching a class, I define what materials you may legitimately bring into the class. If I have an exam, I can tell you that it's open-book or not. It's not that accessing the book inherently is cheating, right? But I define the rules, because it depends on my teaching aims."

The most common suggestions I heard for how students could profit from self-directed use of AI chatbots involved soliciting straightforward explanations of concepts or summaries of text. "What are the current beliefs about tetrachromacy in humans?" Ungar offered by way of example, referring to the perception of color by retinal cells. "That's a reasonable question to ask in a freshman-level cognitive science course. And you could use Google: you could find five or 10 different articles, and then summarize the data and start to form some opinion as to whether there is in fact documented tetrachromacy in humans. [But] you can probably do it 10 times as fast if you use something like ChatGPT."

And if along the way you get confused by some concept, "you can ask it to explain it to you like you're 10 years old," Mollick pointed out. "It's not always perfect, but it's certainly a lot more helpful than not getting it explained."

When I requested simple explanations of topics I knew a fair bit about, I was usually satisfied with the results—though my confidence in ChatGPT's summarizing function was permanently shaken by my first exposure to it. During a video interview, Callison-Burch asked it to summarize a 240-word passage of a *New York Times* article. The result contained a fundamental misattribution error that seemed to arise from the presence of multiple perspectives. In a partial but critical respect, the summary stated the opposite of what the article conveyed; reliance on it would have led to an unpardonable journalistic error. (When I duplicated the attempt three weeks later, ChatGPT offered an error-free summary. But when asked to regenerate it 30 seconds after that, it made the same original mistake—and so did Chat Bing.) Nevertheless, soliciting simple explanations about settled topics seemed relatively low-risk—especially with Chat Bing, which provides hyperlinked source footnotes.

But a more creative form of direct student engagement, proposed by Mollick in a Substack post, underscored the risk that generative AI poses to anyone who lacks the expertise to vet its output. Chat Bing, he suggested, has the ability to "apply general theories to specific, never encountered examples in meaningful ways"—a potentially powerful way to deepen conceptual understanding. As an example, he asked Bing to opine about how John Stuart Mill and Immanuel Kant would have analyzed the ethics of nuclear deterrence via the mutual assured destruction doctrine. Mollick professed to find (pseudo) Kant's argument "particularly interesting." Which it may have been—but not necessarily for its grasp on the German philosopher's thinking. When I shared it with two philosophy professors, both were appalled. "Use of generative AI in this way might well seriously mislead students in philosophy," said one.

Konrad Kording pitched ChatGPT as a potential partner in Socratic dialogue. "Just ask it," he told me, and it would start posing queries rather than merely responding to mine. So I did. I asked it to engage me in a Socratic dialogue about whether citizens should be permitted to cite religious beliefs to justify refusing expression-related commercial services to certain other citizens—as in a case involving a Colorado baker whose refusal to serve a gay couple reached the US Supreme Court in 2018. I chose the topic because it has inspired abundant commentary from multiple perspectives, and I am genuinely of two minds about it. I am skeptical about commercial actors citing religious convictions to gain immunity from generally applicable anti-discrimination statutes; but the conservative commentator David French and Penn political science professor Rogers Smith have articulated two distinct counterarguments that I find compelling. Would ChatGPT hit on one of them—or come up with another—if asked to play the devil's advocate?

After establishing our initial positions we went for three fruitful rounds. Whatever else might be said about it, ChatGPT has an astonishing capacity for fluid, naturalistic conversation. It listened closely and countered sensitively—until suddenly it seemed to listen *too* closely. On our fourth exchange it abruptly capitulated to my position, and iced the proverbial cake with a gloss on *Masterpiece Cakeshop v. Colorado Civil Rights Commission* that got the Supreme Court's decision exactly backwards (despite having correctly characterized it earlier). When I repeated the exercise three weeks later using OpenAI's GPT-4 upgrade (which debuted for $20/month), the bot steered clear of SCOTUS altogether but contradicted its own conceptual argument halfway through our conversation, which devolved into a muddle.

I didn't come out of these dialogues totally emptyhanded. When I pivoted GPT-4 away from Socratic dialogue and toward straightforward explanations of the "strict scrutiny" standard under the 1993 Religious Freedom Restoration Act, it cleared the collegiate bar with ease. But for me, the exercise also underlined how far generative AI remains from offering anything like the intellectual frisson of talking to Konrad Kording himself.

Heather Schneps, a senior neuroscience major with an interest in computer science, emphasized the preciousness of that privilege. Hitting up ChatGPT to explain concepts could be a boon to many students, she told me. "I just hope it doesn't get to a point where that's a substitute for people attending office hours, or that it makes professors feel less like they have to answer questions because kids can just use [AI] instead."

Schneps traced some of her own academic success, as well as an important undergraduate research opportunity, to the office hours of psychology professor Johannes Burge. "I really loved his class," she reflected. "And I enjoyed speaking with him—I felt this rapport that was very meaningful. I felt more connected and motivated when I was in class. And I think that forming those relationships is really important."

Ungar observed that using AI explainers as a first resort doesn't preclude quality facetime with a professor—and might even free up more time for it. "A lot of questions that students have actually involve fairly mundane technical details" that a generative AI can handle, he said. Then, if students want to "talk about career planning, or what should you do if you want to become a deep learning person, come talk to me. Or talk to me about why Stable Diffusion is better than GANs," he added, referring to AI image generators. "Frankly, GPT will probably do some sort of summary, but maybe it's more fun to talk to me: Great! There are things where you want the human interaction, and you want the open-ended discussion."

Ungar predicts that generative AI will finally push interactive digital tutoring over the hump that's thwarted it for decades. The online educational organization Khan Academy, for instance, announced in March that it will pilot a new GPT-4-powered tool as a "virtual tutor for students and a classroom assistant for teachers."

Yet Ungar thinks that such tools will see limited use at Penn. "Because we are, frankly, wealthy, and very expensive. We hire expensive people," he noted. "But for a large part of the world—think of India—you can't afford to hire a Penn professor to go through and read your paper and give comments. You can't go to their office hours and chat with them. I think there's a lot of cases where you're going to see these systems saying: *I've got your first draft. Here's a bunch of comments*."

I spoke with five Penn undergrads, freely offering anonymity to encourage candor, and chatted informally with more. To my surprise, many said they had used ChatGPT very little, if at all. Some were wary of stumbling into an academic-integrity charge. Others feared that using AI as a crutch would undermine the skills, especially in writing, they sought to develop in college. (One initially willing source either got cold feet or became too busy to follow through.) But many of those who'd experimented with the tool had intuited that it was better at boosting efficiency than producing a caliber of work they'd want to turn in.

"It would take you a really long time" to get ChatGPT to produce usable text, said a Wharton junior who'd used it with a professor's blessing. But its single-shot digests of web-based information beat doing a dozen Google searches. "If I can save myself 30 minutes of preliminary research, and then just start getting to work on my own ideas, it's really helpful," she told me. "A lot of times it takes some playing around with it, but it really does give you the nuts and bolts of what you want to talk about. It gives me enough of a base, and maybe some key points, that I'm like, *Alright, I feel confident to now come up with my own articulation of this.*"

She was cognizant, though, of a potential danger. "Sometimes you can fall victim to a little bit of an anchoring bias," she said. "We're all a little lazy. And if I just say, *Oh, that must be all there is to know about it—ChatGPT gave me what was on the internet and I don't really need to look any further*, then I'm kind of anchored to whatever ChatGPT generated in that single response."

Efficiency also lies at the heart of several ideas for teacher-mediated uses of generative AI. Mollick proposed a few in a mid-March Substack post in which he shared carefully worded prompts anyone could use.

One turned ChatGPT (or Bing) into a customizable example-generator for any topic and level a teacher chose. To explain "opportunity cost to college students," for example, ChatGPT came up with four concise, cogently explained examples splendidly tuned to the lives of college students, like "part-time job vs. internship."

Another turned ChatGPT into a "creator of highly diagnostic [and] low-stakes" multiple-choice quizzes. When I used it to create a college-level test on the mid-20th-century South Carolina governor James Byrnes, it passed with flying colors. It took 10 seconds to produce five varied questions that, remarkably, did not stray into Byrnes' more consequential stints as a US Senator and Secretary of State. And the answer key was correct. It was hard to think of an easier way to gauge students' progress with assigned reading. Yet when I solicited a second quiz, on the history of race relations at the University of Pennsylvania, four out of the five answers were wrong and a couple of the questions themselves were so misguided that no correct answer was possible.

Nevertheless, approaches like these mitigate the risks of unreliable output by stationing teachers at the gate. They

have the expertise to jettison bum examples or quiz questions before students can be led astray.

Warp-speed creation of multiple-choice quizzes is hardly a higher-education gamechanger. But a different line of Mollick's pedagogical thinking struck me as genuinely compelling. It flips the notion of using ChatGPT to critique student writing by putting the bot in the pupil's chair instead.

"By acting as a 'student,' the AI can provide essays about a topic for students to critique and improve," he wrote with coauthor (and wife) Lilach Mollick, Wharton Interactive's director of pedagogy, in a white paper. "The goal of this exercise is to have the AI produce an essay based on a prompt and then to 'work with the student' as they steadily improve the essay, by adding new information, clarifying points, adding insight and analysis, and providing evidence. We take advantage of the AI's proneness to simplify complex topics and its lack of insightful analysis as a backdrop for the student to provide evidence of understanding."

"If you put the student in the role of instructor, then they learn," Ethan told me. "It's constructivist learning: You're learning by doing in a very interesting, specific way."

And the coming ubiquity of generative AI, many professors believe, is about to make the skill of critical reading more important than ever.

"There's an asymmetry," Ungar told me. "It has become much cheaper to generate bullshit than to detect it. And that economic shift is going to cause a huge problem.

"It's just going to become a key function that people are going to have to learn," he continued. "When someone gives you something that's beautifully written, with citations and clean grammar and everything looks super impressive, should I believe it or not? You can call it critical thinking. You can call it what you want. But I think it's a huge problem."

COLLEGE IN AN AGE OF IMMACULATE BULLSHIT

Wharton's emergence as an early locus of generative AI experimentation at Penn is not surprising. "I'm a business school professor," Ethan Mollick emphasized in our conversation. Beyond brainstorming pedagogical uses of generative AI, he was ultimately focused on helping students accomplish their practical goals in the world of commerce, to which the tools are coming quickly. But I didn't expect to find the English department on the front foot as well—let alone in a course on John Milton.

Zachary Lesser, the Edward W. Kane Professor of English, was one of the first faculty members to task undergraduates with critiquing AI essays, albeit on an optional basis. Instead of writing a paper using a traditional prompt in his Age of Milton class, they could choose to feed the prompt—altered to suit their purposes—into ChatGPT and pick apart the results. The due date of this "pure experiment" fell after the *Gazette*'s deadline, but Lesser thought it might be an effective way to move students beyond the kind of high-school-level boilerplate that ChatGPT so readily churns out: "flowing, grammatical essays" with "bland, catch-all conclusions" that hew to broad generalities at the expense of "anything concrete."

"My hope is that they'll form their own argument about the same topic," Lesser said, and "develop a more sophisticated understanding"—both about 17th-century English literature and what critical analysis really entails.

For the moment, at least, generative AI struggles in that domain. When it released GPT-4, OpenAI and other researchers demonstrated many impressive capabilities. It could derive a married couple's tax liabilities from a plain-language description of their (uncomplicated) income and deductions combined with a copy-paste of the notoriously convoluted US tax code. It could transform a crude pencil sketch into a primitive website. It could pass the LSAT, the Uniform Bar Exam, the US Medical Licensing Examination, and a raft of AP exams—but not, curiously, AP English, which it flunked by a mile. Perhaps that will change with GPT-5. But insofar as AI amplifies the market value of critical reading, college English departments may have a special role to play. So might philosophy departments. It would be a strange irony if enrollment in the humanities, which has cratered across the country over the last decade, were to be revived by a technology that excels at manufacturing immaculate bullshit.

To become savvy users and analysts of generative AI, students will also need to know more about what goes on under the hood. Chris Callison-Burch's mid-March visit to Karen Rile's fiction writing seminar provided an educational model. Using OpenAI's Application Programming Interface (API), he showed students exactly how GPT's autocomplete inferences work—illustrating not only its propensity to hallucinate, but its tendency to regurgitate the sorts of slander that suffuse the internet.

When asked to complete the phrase *All Trump voters are*, for instance, the AI suggested *bigots*—before its guardrails triggered a statement describing Trump voters as a "diverse group of individuals with a wide range of beliefs" and cautioning against "sweeping generalizations about any group of people based on their political affiliations." GPT triggered the same self-correction every time it was asked to complete the phrase (*racists*, *idiots*, etc.)—just as it later did for me after reflexively maligning Mexican immigrants and Muslim neighborhoods. This is a testament both to OpenAI's content-moderation efforts and the fundamental problem they seek to address.

When Callison-Burch solicited an obituary for "Prof. Karen Rile," for instance, GPT mourned the passing of an accomplished academic with an impressive (if fabricated) CV. But when asked to eulogize "Karen Rile," without the title, GPT produced a paean to a life

filled with baking projects and gendered domesticity. Callison-Burch wondered how far we've really come since the time, some years ago, when typing the beginning of an analogy into an older AI model—*man* is to *computer scientist* as *woman* is to _____—prompted it to answer *homemaker*.

Liam Dugan, his PhD student, observed that as AI chatbots contribute more and more text to the internet, they may create feedback loops that make such problems harder to resolve. "There's a worry that as more machine-generated text proliferates, that it could start to lock in a lot of biases," he said.

A different sort of bias can be expected to arise from how LLMs are trained.

"After OpenAI trained the normal model to complete the next word" based on internet text, Dugan explained to Rile's students, "human annotators were given, say, four or five of the model's responses, and asked to sort which ones they thought were the best and worst. And the model was fed that back, and then retrained to optimize for the ones they thought were best. So there's a lot of human feedback in here, and some of the behaviors the model shows—like the tendency to have this nicely structured five-paragraph essay every time—may be because it's reflecting the preferences of the annotators that they hired. And maybe that's a reflection of what these annotators would like to see, rather than what would be useful to you."

"Theoretically," Callison-Burch added, "you could have a different set of preferences for collegiate writing versus seventh-grade writing."

Or for poetry. One way to get GPT to "sing for you," as he put it, is through a technical process called fine-tuning. It involves feeding a large but focused data set into OpenAI's API—like the 15,000 or so poems Callison-Burch scraped from the Poetry Foundation—to bend the output beyond what well-crafted prompts alone can achieve. If working with generative AI is like learning to play a guitar, fine-tuning is a bit like altering the instrument's shape. Though the verse Callison-Burch's modification elicited was fundamentally derivative, not original, it was also more evocative, edgy, and even haunting than any I'd seen from standard-issue ChatGPT.

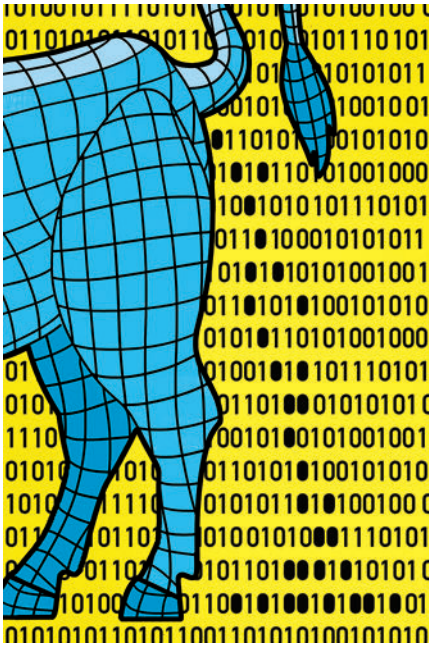
Fine-tuning is already figuring into innumerable start-up companies. Two in the education space include Elicit AI, a literature-review tool that finds and summarizes academic journal articles; and (still in development) Etan Ginsberg EAS'23 W'23 and Shriyash Upadhyay EAS'22's Learn Like a Martian, which aims to sync LLMs with online course management platforms like Canvas to create tailored flashcards and quizzes directly from course materials.

But to run with the musical analogy, fine-tuning can't turn a guitar into a trumpet. The only thing that can do that is massive amounts of money.

"A lot of companies are trying to build general-purpose models that can be used for lots and lots of different tasks," said Daphne Ippolito Gr'22, who studied under Callison-Burch and is currently a senior research associate at Google on her way to a computer science professorship at Carnegie Mellon. "That's because they're super-expensive to train—it's hundreds of thousands to millions of dollars to train the largest models—and so it makes sense to train it to be as general purpose as possible. But sometimes by being general purpose, you make it worse at each individual purpose that you could have it do."

Callison-Burch worries that the prohibitive cost of developing LLMs could concentrate power in the "small handful" of companies that can afford it.

In February he traveled to Arlington, Virginia, to lobby for a $200 million per year federal outlay to create a "national inference engine" infrastructure along the lines of a government/academic/industry/non-profit partnership model. Leaving the development of LLMs solely to a small group of companies, he



"It has become much cheaper to generate bullshit than to detect it. And that economic shift is going to cause a huge problem."

argued, "could ultimately lead to significant economic and social inequality." [See Sidebar, page 29.] Moreover, it would make universities "totally reliant on companies' infrastructure," limiting academic research to "prompt engineering and fine-tuning models, and other limited actions enabled via companies' APIs. This is not science."

So the questions of who controls the underlying AI platforms, how they're monetizing them, and what they're doing with the data users pour into them are further matters that merit scrutiny by students, professors, and university leaders.

They join a daunting but invigorating list.

For ultimately that's what generative AI confronts us with: questions that it can't answer but we can't dodge.