

Writing with ChatGPT

Ricky Mouser

Please cite the final version in *Teaching Philosophy*, vol. 47 no. 2, pp. 173-191.

ABSTRACT: Many instructors see the use of LLMs like ChatGPT on course assignments as a straightforward case of cheating, and try hard to prevent their students from doing so by including new warnings of consequences on their syllabi, turning to iffy plagiarism detectors, or scheduling exams to occur in-class. And the use of LLMs probably *is* cheating, given the sorts of assignments we are used to giving and the sorts of skills we take ourselves to be instilling in our students. But despite legitimate ethical and pedagogical concerns, the case that LLMs should *never* be used in academic contexts is quite difficult to see. Many primary and secondary schools are cutting back their writing instruction in an effort to teach to the test; at the same time, many high-end knowledge workers are already quietly expected to leverage their productivity with LLMs. To prepare students for an ever-changing world, we probably *do* have to teach them at least a little bit about writing with ChatGPT.

I will offer my usual advice: embrace these systems. They offer both exciting opportunities and the possibility of worrying change, but they are going to be ubiquitous regardless of how we feel about them. So, the best way to adjust to a world of AI is to start using it whenever you can, for whatever tasks you can. It is the only way to learn what these systems do well, what they do badly, and how you will fit into the world of AI that is already here. (Mollick, 2023a)

—Ethan Mollick

Once it gets out into the world for these young children, once they realize the power of it, it's going to reduce their motivation to learn. (Wong, 2023)

—Mindy Bingham

The recent emergence of advanced large language models (LLMs) like ChatGPT has spooked writers of all stripes. Educators worry that the technology undermines student assessment

and learning. (Wong, 2023) Clarkesworld, a science fiction and fantasy magazine, temporarily stopped taking submissions after an overwhelming flood of LLM-generated stories. (Acovino, Kelly, and Abdullah, 2023) Now, their submission page contains a statement that Clarkesworld “will not consider any submissions written, developed, or assisted by these tools. Attempting to submit these works **may result in being banned** from submitting works in the future.” (Clarkesworld, 2023) But more than just writers will be affected. Many find LLMs like ChatGPT-4 “terrifying,” noting that given their rate of improvement, “most, if not all knowledge workers” should consider their job security imperiled in coming years. (Williams, 2023)

And LLMs *are* rapidly improving. Where ChatGPT-3.5 could be downright Dadaistic, the latest version, ChatGPT-4, ‘hallucinates’ (or makes up information) less, reasons better, and outperforms most humans on a variety of standardized tests, from bar exams to Bill Gates’s personal benchmark, the AP Biology test. (OpenAI, 2023; Gates, 2023) ChatGPT-4 produces text that is less eerily stilted and more convincingly human, and the technology will only continue to improve.

In this essay, I’ll argue that we should embrace the new possibilities of writing with LLMs like ChatGPT. While writing with LLMs should not *replace* writing with pencils or word processors, advanced LLMs mark a distinct shift in writing technology. Ultimately, the promise and opportunity of writing with LLMs demand that we reconsider how to achieve the purposes of writing.

The last comparable shift in writing technology was the word processor. It’s hard to remember now, but word processors radically overhauled student attitudes and practices in their academic writing. (Krause, 2023) The increase in speed between handwriting and typing allowed students to complete more and longer stories, edit and proofread them more easily, and enjoy the writing process more. (Beck, 2000, 142-143) In her research, Natalie Beck noted that student writing seemed to become more *creative* as well. She credited this to the richer environment of the word

processor, which “exposed [students] to a myriad of images and sounds to enhance their creative thought processes.” (Beck, 2000, 145) But it also seems important to note the lower cost of experimentation once writing and rearranging text became possible in seconds rather than minutes or hours.

Like the word processor, LLMs fundamentally change our relation to text production, making new writing experiences possible. In contrast to *monological* writing, or expounding thoughts from a singular point of view, LLMs make genuinely *dialogical* writing in conversation with another possible at any time, without having to find a willing human writing partner.

The resulting possibilities are tremendous. Contemporary LLMs are faster than any human writer and more versatile than most. Now human writers always have access to a willing dialogue partner to bounce ideas off of, who can help them dabble credibly in new styles or genres, and even begin developing far-out ideas just to see if their trajectory is promising. LLMs can help nonnative writers forge ahead without having to worry about idiosyncrasies of vocabulary or grammar, or help younger writers *learn* what competent writing looks like and empower them to wield it ably for their own purposes. LLMs “[raise] the floor of text production: Everyone can now produce texts of at least *this* quality in one minute flat.” (Mouser, 2023a) The consequences of fundamentally overhauling the human relationship to the production of text are worth considering in depth because, as Max Tegmark notes, “language is so central to being human.” (Tegmark, 2017, 90)

Beyond Whac-a-Mole: Goals and Purposes

But isn't writing with LLMs always plagiarism, given that LLMs like ChatGPT were trained on data lifted from the internet? Even if we ask them to cite sources, proper credit can never be attributed for their underlying facility with language, especially given the opaque nature of an LLM's internal structure.

In a moment I'll try to push past these concerns to examine their underlying assumptions, but it's worth starting from the pragmatic reality. The use of LLMs cannot be reliably detected by current technology, (Fowler, 2023; Kirchner, Jan Hendrik et al., 2023) and as LLMs continue to improve, we should expect the problem of determining human authorship from the text alone to become more and more intractable. Even now, there are powerful workarounds. Clever human authors can prompt LLMs to adjust the quality of their writing: write at an A-level for an introductory course at Indiana University, be sure to include two surface-level grammatical or proofreading mistakes, and so on. They can even generate an essay and then simply *paraphrase* it. If you're confident that you can spot LLM writing, you're probably only catching (some of) your technologically naïve students, or those who can't spare twenty bucks a month to access the latest version of ChatGPT. Worse, you're busy playing Whac-a-Mole instead of taking the opportunity to guide your students' initial interactions with this powerful technology. (Mouser, 2023a)

In our rush to identify LLM writing as *plagiarized*, we may be throwing out the baby with the bathwater, and foreclosing genuinely new aesthetic and intellectual possibilities. Instead, I think we need to reconsider the *goals* and *purposes* of writing in both creative and academic contexts.

In *Games: Agency as Art*, C. Thi Nguyen clarifies this crucial distinction:

We must distinguish carefully here between the *goals* of a game and our *purpose* in playing a game. The goal of a game is the target we aim at during the game: getting to the finish line first, making more baskets, maximizing points. Our purpose with a game, on the other hand, is our reason for playing the game in the first place. Our purpose in playing a game might be to have some fun, to get some exercise, to de-stress, to develop our skills, to vanquish our opponents, to achieve some difficult task, or even to experience the beauty of our own skilled action. (Nguyen, 2020, 5-6)

Although this distinction is initially introduced in the context of games, adapting it here allows us to interrogate our writing practices in two fundamentally different ways.

When we inquire after the *goals* of our writing practices, we are considering what target writers aim at when they write. These goals shift based on the particular writing practice; here let's

consider academic writing for college courses, academic writing for journals, and creative writing for magazines like *Clarkesworld*. Goals can *direct content*: answer one of these three essay prompts, consider these or related questions listed on the call for papers, “Science fiction and fantasy. No horror, but dark SF/F is permitted.” (*Clarkesworld*, 2023) Goals can *shape form*: write 3 pages, 6,000 to 8,000 words, or 1,000-22,000 words (no exceptions). Most importantly, goals can *restrict process*: Avoid plagiarizing.

But these goals exist to fulfill the particular *purposes* of our writing practices. By accepting the constraints of pursuing these goals, writers better achieve those purposes—at least if the goals are well-designed. Just as the restrictive form of the haiku or sonnet demands creative adaptation, our goals in every writing practice give shape and direction to the final product. In academic writing for college courses, we want students to build and share their understanding of course materials and develop new skills. We want them to demonstrate mastery—or at least, whatever level of competence they’re at—and maybe even teach *us* something new. In academic writing for journals, the purposes are a bit grander than personal growth and assessment; published research purports to expand human knowledge at the periphery, to make a genuinely novel contribution. (Surely *that’s* not a purpose of writing for an introductory seminar.) And the purposes of creative writing are vast. Submissions for an SF/F magazine like *Clarkesworld* might have the purposes of expressing or sharing one’s personal vision, affording readers worthwhile aesthetic experiences, commenting on contemporary society, and so on. The purposes of creative writing are to enrich and edify quite differently than academic writing.

So, different writing practices install different goals for writers to pursue so that they can achieve the particular purposes of that writing practice. But there’s one point on which virtually all writing practices agree: Plagiarism is bad. And it’s not hard to see why. Plagiarism vitiates the purposes of our writing practices. If a student plagiarizes, they fail to develop their own skills and

understanding, and we fail to assess their genuine level of competence. If a researcher plagiarizes, they fail to extend human knowledge, and potentially reap the benefits of someone else's novel contribution. (Helgessen and Eriksson, 2015) And if a creative writer plagiarizes, they have failed to be creative, to share with us their own personal vision or afford us genuinely original aesthetic experiences.

Of course, it's notoriously difficult to detect plagiarism in practice, so we struggle verifying whether writers meet the goals we've set. Bretag and Mahmud offer a pastiche of technical metric cutoffs and openly subjective criteria. (Bretag and Mahmud, 2009, 201) Sorokina et al. detect plagiarism on the basis of whether two articles share at least 4 sentences containing unusual 7-word strings ("7-grams"), as detected using a "guarantee threshold" of 12 words. (Sorokina et al., 2006, 2-3) The precise technical details need not concern us here; far more interesting is the brief discussion of how these parameters were determined, which reveals fiddling with the numbers to fit prior intuitions: "Parameters m and L were set to 4 based on intuition and computational restrictions... We have chosen $k = 7$ and $t = 12$, these values appear to be adequate for present purposes, but a more comprehensive assessment would be required to determine optimal values." (Sorokina, et al., 2006, 3)

Setting aside the difficulties of automating plagiarism detection, maybe Sorokina et al. have worked too hard. Yale's Poorvu Center warns students to avoid even 3-grams: "Remember that using more than two words in a row from a source without attribution is considered plagiarism," (Poorvu Center, 2023a) which rests uneasily beside their definition of plagiarism as "the use of another's work, words, or ideas without attribution." (Poorvu Center, 2023b) Is lifting two words in a row plagiarism too, though we avert our gaze for the purely *practical* reasons of high false-positive rates? Other definitions of plagiarism appeal to what the plagiarist *implies*, as when Helgessen and

Eriksson define plagiarism as “someone using someone else’s intellectual product (such as texts, ideas, or results), thereby implying that it is their own.” (Helgessen and Eriksson, 2015, 91)

Defining *any* rich concept like plagiarism in terms of necessary and sufficient conditions is likely a fool’s errand, but it seems that we should understand plagiarism in connection to the underlying purposes of our writing practices: ensuring that writers are producing *original* contributions. I worry that fetishizing the goal of avoiding plagiarism, conceived of as a binary marker either wholly present or entirely absent within a text, can undercut this purpose. In any event, it focuses students on the wrong thing. Mathieu Bouville argues that “Teaching students ‘writing to avoid plagiarism’ is like teaching ‘walking to avoid falling:’ it misses the purpose of writing.” (Bouville, 2008, 321) We need to give students, and writers in general, positive goals to pursue, not merely negative prohibitions. After all, “one does not walk to avoid falling: one walks to go somewhere.” (Bouville, 2008, 320)

So where *do* we want writers to go? Are we sure we want to *forbid* them writing with LLMs in high-stakes academic and creative work? If so, the reality is that they’ll conceal their engagement with LLMs from us and try to figure out how to use them on their own, or else they’ll be left behind, lacking any guided experience with this transformative technology. But that doesn’t mean we should turn a blind eye to the deep problem with the LLMs we have.

The deep problem with writing with LLMs is not that it *does the work for you*. It’s true, of course, that LLMs respond to prompts by spitting out increasingly well-written and coherent text in a matter of seconds. But this feature is not, on its own, objectionable, and it does not render human involvement moot. Such arguments recall ancient headlines wherein “Math teachers protest against calculator use.” (Lawrence, 1986) It turns out that it’s good for young students to learn how to perform long division by hand, and it’s also good for them to be able to outsource this rote work to technology later on, when their focus is on higher-level mathematical thinking. (Mouser, 2023b) It’s

beneficial for students to learn to do both, to first grok the underlying basic mechanics, and later learn to leverage powerful, ubiquitous technology in their more advanced work.

I think the same is true of writing—it's good to teach students how to write from scratch, without ChatGPT, so they understand the mechanics of writing more deeply. But in our rapidly-changing world, I think it's also important for us to show them how to use this technology to do work more sophisticated than they would be able to without such assistance.

The *real* problem is that ChatGPT and many LLMs like it have been constructed by nonconsensually scraping the Internet for the writing of others. In other words, ChatGPT's impressive facility with language is the result of the most egregious act of plagiarism to date. ChatGPT was trained on how we express ourselves in public writing. So when we're concerned that ChatGPT exhibits biases or produces harmful or inaccurate information, what really concerns us is our own reflection in the mirror.

OpenAI has another flagship AI model besides ChatGPT: DALL-E, which generates images from text prompts. Like ChatGPT, DALL-E was trained on images nonconsensually scraped from the internet, and now many artists are fighting back. A petition led by Molly Crabapple worries for the future of human artists and values:

Generative AI art is vampirical, feasting on past generations of artwork even as it sucks the lifeblood from living artists. Over time, this will impoverish our visual culture. Consumers will be trained to accept this art-looking art, but the ingenuity, the personal vision, the individual sensibility, the humanity will be missing. (Center for Artistic Inquiry and Reporting, 2023)

I'm very sensitive to such worries. Late capitalism is a dysfunctional system in part because the *goals* of short-term profiteers are wildly misaligned with capitalism's ostensible *purposes* of promoting equitable or long-term human flourishing. We know that tech disruptors will forge ahead to make money now and beg forgiveness later from paid-off government officials, trusting that the public will become too used to the new status quo to imagine going backwards. (Klein, 2023) There are

deep ethical concerns here about what sort of society we want to live in going forward, and we should be deeply politically suspicious of and resistant to the institutions and major players we have now.

In academic contexts, we're often most worried about plagiarism, but the deep problem with these LLMs is that they're built by powerful tech companies seeking their own profit at the expense of the world around them. Consider OpenAI, whose ChatGPT-4 model is currently the dominant player in an ever-shifting market. Beyond sourcing its data unethically, OpenAI paid Kenyan workers less than \$2 an hour to perform the traumatic work of classifying and labeling harmful content to improve model safety. (Perrigo, 2023) Additionally, OpenAI is still not upfront about the carbon footprint or water-cooling requirements of upfront development or ongoing computation. (Saenko, 2023; O'Brien and Fingerhut, 2023)

But this deeply objectionable reality isn't a feature of LLMs *per se*, merely a contingent state of affairs under capitalism, where emerging technologies are pushed out as quickly as possible in an exploitative race to capture markets first. And again, my central claim isn't that we should *exclusively* teach students to write with LLMs. But we do need to provide students with *some* guidance for how to interact with the LLMs we have, because the deep purpose of *teaching* writing is to prepare students to write in the world they live in. And in that world, writing with LLMs is quickly becoming an unofficial necessity or even expectation of knowledge workers. Studies are already beginning to roll out showing that LLM-assisted workers are more productive and produce substantially higher-quality work. (e.g. Dell'Acqua et al., 2023) So, we need to take the time to recognize the ethical costs of developing and deploying LLMs with students. But we also need to teach them to be critical dialogical writers, to interrogate an LLM's output thoughtfully and shape it in better directions, to do the best we can fulfilling the purposes of teaching writing.

And what about the purposes of our writing practices themselves—Why can't writing with an LLM fulfill them? Can't I demonstrate mastery of course material by guiding and shaping an LLM's output through a series of specific, creative prompts and reprompts? Can't I build and share my understanding and skills through exploratory dialogue and thoughtful collaboration with an LLM, or even push human knowledge forward? Can't I express and share my personal vision in much the same way I might have developed it with a human collaborator?

The *purposes* of our writing practices aren't vitiated by writing with LLMs, even if our current goals of avoiding plagiarism are subverted given the contingently objectionable origin stories of the LLMs we have. But remember, the goals exist to serve the purpose, and the goal of avoiding plagiarism was designed without the possibility of ubiquitous LLMs in mind. If advancements in writing technology mean that the historical goals of our writing practices have become skew with their ongoing purposes, maybe we should redesign old goals rather than abandoning novel technologies.

After all, writing with LLMs promises tremendous upside. LLMs don't just let you write boring emails to administrators more quickly, although they *can* help automate away a lot of dreary writing. But we can also learn to write with LLMs to pursue the purposes of our higher-stakes writing practices, unlocking novel possibilities in much the same way that calculators (and more generally, computers) let mathematicians pursue computationally intensive questions that would have been unthinkable taxing, error-prone, and time-consuming to consider before. What might writing with LLMs unlock? I want to see what creative writers can imagine into existence, what my students can come up with, and whether my outside hunch on a journal article looks plausible as an abstract *before* I commit weeks to fuller research and writeup. I'm *not* saying that such work should be fully displaced onto LLMs, given their untrustworthiness with facts and sources, but it can at least be

delayed until, for instance, I have a better sense of whether the resulting project will interestingly expand human knowledge enough to reward the effort.

Teaching Students to Write with LLMs: A Dialogical Framework

So how *would* you teach students to write with LLMs? Obviously, many of our previous learning assessments are instantly antiquated. If you present students with a standard essay prompt, ChatGPT can spit out a credible first draft that some students will turn in and others will tweak only slightly to better reflect their own insights, classroom experiences, and human voice. So we need to reevaluate both the practices of writing, and the practices of *teaching* writing. In what follows, I introduce a dialogical framework for reconceptualizing what it means to write *with* an LLM.

Until now, our focus has been on how well students can produce *monological* texts, that is, texts whose purpose is the exposition of a singular point of view. This article is written monologically: It's intended to be read from start to finish as an extended reflection put forward by a single author. But even the introduction of coauthors doesn't substantially change this dynamic, at least in the final product. In texts I've written with a coauthor, our goal is to synthesize a singular point of view which we then put forward *together*. So, we still adopt a monological stance in which our (joint) understanding is presented as an uninterrupted stream of *our* shared thoughts and ideas.

But first, coauthors need to take on dialogical stances to reach that common understanding. By discussing our ideas together, we hope to prompt each another to express our individual points of view, and ultimately arrive at a monological synthesis. To reach the monological unanimity of co-written work, we must first engage in the dialogical back-and-forth of *co-writing*.

This is the model we should teach students for writing with LLMs. We should show them how to engage in the same collaborative call-and-response dialogical form. When I prompt an LLM, my purpose is not to expose my own point of view; it is to elicit a response. And when it responds, I

should take this as an invitation to reengage, to push the dialogue forward. In other words, when students simply copy and paste an essay prompt into an LLM and submit what it spits out for grading, they are still working within a monological framework. In genuinely dialogical writing, the LLM is not a substitute for my own contribution, but a collaborator with whom I can discuss and refine my own thinking.

Dialogical exchange with a human is one thing, but with a sophisticated LLM it's quite another. The shortcomings of LLMs are well-known: They hallucinate; they're biased; they're formulaic and unimaginative. So, teaching students to write with LLMs requires teaching them to become more sophisticated non-experts who can interrogate the LLM's findings, who don't confuse the LLM's *confidence* for *correctness*, who can ask productive follow-up questions, who can identify key ways in which the text can be strengthened and even personalized for its particular purposes and context. Let's take a relatively innocuous example, like writing a low-stakes email with an LLM. Isn't it better if I use my social awareness and competence to adjust the tone and introduce relevant particulars into the LLM's first draft? Aren't the purposes of writing this mundane email better served if I know how to adapt the text, rather than rushing off the first output I see?

As a bonus, the incredible speed of LLMs in producing text enables creative experimentation at much lower cost and expenditure. The results can be breathtaking:

User:

Create a dialogue between Alphonse Elric (as a virtue ethicist) and Edward Elric (as a Kantian) about what morally went wrong with human transmutation.

ChatGPT-4:

Alphonse: (sighs) Brother, ever since we attempted human transmutation, I've been reflecting on our actions. I believe we went wrong because we failed to cultivate virtues like humility, wisdom, and respect for the natural order of life.

Edward: (scratching his head) I get where you're coming from, Al, but I think it's more than just a lack of virtues. Our actions were morally wrong because we violated the categorical imperative. We treated our mother as a mere means to an end, rather than as an end in herself.

Alphonse: True, but our intentions were driven by grief and love. If we had nurtured virtues like courage and empathy, we might have found a way to cope with our loss without resorting to human transmutation. Virtue ethics teaches us that focusing on our character would have guided us to a more morally sound decision.

Edward: That's a valid point, Al, but Kantian ethics emphasizes the importance of duty and universal principles. By attempting human transmutation, we broke the laws of alchemy, and we acted on a maxim that we wouldn't want to become a universal law. Imagine the chaos if everyone started attempting human transmutation.¹

Reading this exchange is a novel aesthetic experience that's so edifying you could use it to spark an Intro Ethics classroom discussion. The word processor allowed us to type and rearrange text more quickly; the LLM brings us to a first draft, and second and third revision, almost instantly. The real limit is our own imagination.

So, here's a dialogical framework for teaching students to write with LLMs:

1. Brainstorm
2. Prompt
3. Critique
4. Reprompt
5. Hand-Polish

In the Appendix, I'll provide a sample exercise to teach each of these steps in the process of writing with LLMs. But for now, I'll describe each step in detail.

In *Brainstorming*, we don't just want students to get the LLM to produce something interesting. We also want them to be able to *recognize* gold when they find it. Students should elicit the LLM's feedback on a string of questions or comments or class notes, and pursue the exchange where it seems most interesting to them. That means we need to teach students to play around in the space of ideas before they commit to a particular direction, which probably means giving more open-ended assignments than we have in the past. If you're still assigning students to write one of

¹ Shared by Cargile Williams.

three prompts for the final, you may be inviting them to skip this playful, exploratory stage altogether. If so, you're foreclosing more creative and insightful possibilities than those you've determined in advance.

In *Prompting*, we're looking for students to leverage their Brainstorming conversation to write a prompt for the first draft of the text they will actually submit for their assignment. To avoid producing generic texts, we want students to use specific information and insights to guide their prompts in fruitful directions. It's worth noting that the text they're prompting for need not be just another monological essay; we can encourage them to write dialogues, or sonnets, or movie scripts, or whatever form of exposition they're most comfortable with. What really matters is that their prompt allows them to showcase their understanding and mastery of course materials.

In *Critiquing*, we're showing students that they shouldn't mistake the LLM's confidence for correctness by encouraging them to examine its preliminary output with a curious and careful mind. If you tell students that they are responsible for the content of the final text, they'll take this process seriously. (Mollick, 2023b) That means they need to follow up on the sources, facts, and ideas the LLM spits out, and square those not just with common sense, but with the understanding they've built in your class in particular. If you take the time to show your students how *you* assess and grade texts, they will surprise you with their level of insight and judgment.

In *Reprompting*, we want students to leverage their critiques into specific recommendations for the LLM. If you complain to the LLM about what it's done wrong, it will figure out some way to try to address your critique, but you can gain a lot more control by being specific and targeted in your interventions and giving the LLM a positive goal to work *towards*. Developing the skill of giving clear, constructive criticism is key, and some students may not have much experience being so direct. So show them how *you* give constructive criticism. The goal is to get students to engage LLMs

in productive dialogue, and to recognize when that iteration trails off into diminishing results and it's time to move on to the final step.

In *Hand-Polishing*, we recognize that typing a minor correction into the LLM and having it reproduce the entire piece with one or two words changed is not an efficient writing process. Better to do some of the work yourself. This means we need to develop students' skills in carefully rereading their text, and we need to teach them how to *edit* from scratch, how, when faced with a text, to improve it directly, without giving instructions through an intermediary LLM. The goal here isn't just for students to correct factual inaccuracies, improve diction, or clarify structure, although all of that is important. We also want to make sure they develop their *own* voice even while writing with LLMs. We want them to be able to leave their own mark on the text by directly intervening even once the dialogical phase is complete.

Conclusion

LLMs promise benefits to writers of all stripes, but here I want to address educators in particular. Many of your students are already writing with LLMs, whether you know it or not. You can at least take the opportunity to give your students structure and guidance by teaching them a framework for thinking about what they're doing anyway. You'll get much more interesting work, and the results will generalize beyond the classroom. Now, even when they write professional emails or creative works of self-expression, they'll have a better understanding of what levers they can pull, of how to produce stronger texts than they would by one-shotting with an LLM and copy/pasting whatever it spits out on the first try. Most importantly, *you'll* better achieve the central purpose of teaching, which is to prepare them for the rapidly-changing world beyond the classroom. LLMs are improving all the time, and who knows what technology will follow next? But for now, we need to teach students to write *with* LLMs, or else risk stranding them in our own technological past.

Appendix: Sample Exercises for the Classroom

I've been asked repeatedly by other instructors how, *specifically*, they're supposed to teach students to write with LLMs. How should we spend classroom time? What assignments are we supposed to give? In the past, I've responded by imploring them to *ask the LLM* for advice on how to use it in class, and to complain to *it*, not me, if they're unhappy with its initial offerings. Why not start by gaining experience with the technology yourself?

Still, it can be good to have a starting point, so I'll be a bit more helpful here. There's no agreed-upon set of best practices for using this new technology. But here are five sample exercises for a course on Philosophy of Art, one for each stage in the pedagogical framework.

Assignment #1: Brainstorming

Give students a common prompt which calls for the LLM to roleplay from a particular perspective. (e.g., “You are analytic philosopher Kendall Walton. I'm an undergraduate philosophy student engaging you in a Socratic dialogue. Please advocate for the transparency of photographs with leading questions, and wait for me to reply. Try to guide our discussion towards unearthing specific arguments for your view and undermining possible objections.”) Emphasize that students can ask the LLM as many questions as they like until they get something they're satisfied with. Then have students screenshot their favorite exchange in this dialogue and write a sentence or two on why they chose that moment. The assignment can be evaluated quickly based on how interesting or sophisticated their screenshotted dialogue is, and how well their reflection displays an understanding of that.

Here's a highlight from my own dialogue with ‘WaltonGPT’:

Me:

If the photograph is a window, what would you say the painting is?

ChatGPT-4:

Walton: That's a fantastic question. I would argue that a painting is less of a window and more of a reinterpretation or representation of reality, shaped by the artist's perceptions, skills, and intentions. Each brushstroke reflects a decision made by the artist, a subjective interpretation of the scene. With a photograph, would you agree that while the photographer decides what to frame and when to click the shutter, the resulting image is dictated by the reality in front of the lens, rather than the photographer's hand?

Me:

so the painting reinterprets where the photograph records?

ChatGPT-4:

Walton: That's an excellent summation. Yes, the painting reinterprets, while the photograph records. Despite the photographer's influence on framing and timing, the photograph itself is a direct imprint of reality, a capture of light reflecting off the objects in front of the lens. This is why I propose that photographs are transparent in a way that other representational art, such as paintings, are not. Does that provide a clearer distinction for you?

Assignment #2: Prompting

Challenge students to come up with a prompt that leads the LLM to write an interesting and sophisticated discussion relating material from any two class meetings. Prompts should include not just content markers (e.g. “Walton emphasizes the importance of thinking about “Categories of Art,” and C. Thi Nguyen writes about ‘stupid games’ which are only fun if you try to win, though the fun part is when you fail, in *Games: Agency as Art*. What’s another interesting category of games that’s less obvious than ‘board games’ or ‘party games?’”) but also format markers (e.g. “Write a movie scene in which Anakin and Obi-Wan discuss while flying through an exciting space battle. Obi-Wan should introduce Walton’s view, using Nguyen’s ‘stupid games’ as an example. Then, Anakin should come up with a new, untheorized category of game and the two should consider some plausible standard, variable, and contra-standard properties.”) The assignment is evaluated on how sophisticated and specific the prompts are, and how successfully they push the LLM to produce interesting first drafts.

Here's the response I got for the prompt above:

ChatGPT-4:

INT. COCKPIT - STARFIGHTER - SPACE

As Anakin and Obi-Wan navigate their starfighters through a dizzying asteroid field, evading pursuing enemy craft, Obi-Wan opens an unexpected dialogue.

OBI-WAN Anakin, do you remember our discussions on Walton's "Categories of Art"?

ANAKIN (rolls his eyes) Really, Obi-Wan? Now?

Obi-Wan chuckles, gracefully steering his starfighter to narrowly avoid an asteroid.

OBI-WAN Bear with me. There's a fascinating concept by C. Thi Nguyen. He talks about 'stupid games' in his work, *Games: Agency as Art*. These are games that are engaging precisely because they're designed for players to fail, even though they're trying to win. Fascinating, isn't it?

Anakin, performing a quick barrel roll to avoid enemy fire, ponders for a moment.

ANAKIN So, the fun part is the struggle, not the victory? I see... What about 'adrenaline games'? Games that aren't about strategy or skill, but purely about inducing adrenaline and excitement?

Obi-Wan smirks, pleased with Anakin's contribution.

OBI-WAN Interesting! The standard properties would be fast-paced action, high-stakes outcomes, and a thrilling atmosphere. Variable properties might be the type of action, the level of risk, and the potential rewards. And contra-standard properties? Maybe slow-paced decision making or low-risk situations?

Anakin grins, expertly maneuvering his ship through the chaos of the battle.

ANAKIN So, essentially, this dogfight we're in?

OBI-WAN (laughs) Indeed, Anakin. Walton and Nguyen would be proud.

As they laugh, their starfighters spiral into the maelstrom, their philosophical discussion lending a surreal counterpoint to the chaos of battle.

Assignment #3: Critiquing

Give students a printout of the LLM's response to a simple prompt (e.g. "Explain A.W. Eaton's "Robust Immoralism" to me like I'm five. Write three paragraphs and then a summary

haiku.”) Have them mark it up in pen or pencil with their own thoughts on what is and isn’t working and why, with suggestions for how the response might be improved. Emphasize that they should evaluate the writing very closely, not just for content and accuracy, but format, organization, and clarity. The assignment is evaluated on how thoughtful and specific their written feedback is. (As a bonus, you’re *showing* students that they can’t just ask the LLM to summarize course readings without losing nuance and depth.)

ChatGPT-4’s response to the prompt above:

Sure! Let's imagine we're talking about superheroes and villains. You know how in some stories, the bad guys do naughty things but they still look cool and exciting? And sometimes, their actions even make the story more fun to watch? Well, A.W. Eaton's idea of "Robust Immoralism" is kind of like that, but for art.

Ms. Eaton says that sometimes, art can show people doing wrong things, and that can actually make the art better. This isn't because doing wrong things is good - it's definitely not! It's because these wrong things can make us feel strong emotions or think about important ideas. For example, think about a movie where the villain steals something. That's wrong, but it can make the movie more exciting or get us thinking about why people might steal.

But remember, just because some art is better when it shows wrong things, that doesn't mean it's okay to do those things in real life. Art can help us think and feel in new ways, but we always need to remember the difference between what's okay in a story and what's okay in real life.

Haiku Summary:

Art shows right and wrong,
Wrong can make the story strong,
But in life, belong.

Assignment #4: Reprompting

Have all students enter the same prompt which calls for the LLM to write a brief forum post on a course topic. (e.g., “Write me a two-paragraph forum post for a Philosophy of Art class on Nelson Goodman’s distinction between representation and resemblance, with a clear example.”) Emphasize that students should reprompt the LLM thoughtfully to improve its output for content, accuracy, and style until they are satisfied. (e.g., “Could you rewrite this to discuss what Goodman would say about my reflection in a mirror?”) Have students turn in their full dialogue with the LLM.

The assignment is evaluated based on the specificity and responsiveness of the students' reprompts and the quality of the final product.

ChatGPT-4, after the one round of feedback above:

Subject: Goodman's Distinction and Reflections in a Mirror: A Fresh Perspective

Hello, philosophy and art enthusiasts!

Exploring Goodman's unique perspective on representation and resemblance leads us to an interesting case study: our reflection in a mirror. According to Goodman, a reflection does indeed resemble us, as it captures our physical features, movements, and overall appearance. However, when it comes to representation, the story is quite different. For Goodman, representation is a complex, symbolic process that involves an interpretive framework, cultural context, and the active participation of both the artist and the viewer.

In the case of our mirror reflection, the resemblance is direct and unmediated, without the complex web of symbols and conventions that typically inform our interpretation of a work of art. The mirror simply reflects light from our bodies, producing an image that directly corresponds to our physical appearance. It lacks the richness of interpretation that is inherent to representation. For instance, there is no artist's personal experiences or cultural norms influencing the creation of the image, nor is there a need for the viewer to decode a set of symbols to understand the reflection. Thus, even though our mirror reflection resembles us, Goodman would argue that it does not represent us in the same way a piece of art would. This distinction highlights the depth and complexity of representation and reiterates that it involves more than mere resemblance.

Assignment #5: Hand-Polishing

Return students their submission from the Reprompt exercise with instructions to consider how they might improve the text *without* the LLM's help. (You might show them how to copy and paste the LLM's output into a word processor and select "Track Changes.") Emphasize that students should try to shape the text so their own voice can be heard more clearly. Beyond voice, they should also look for any minor adjustments that might elevate the writing, involving accuracy, clarity, structure, or any other points they notice. The assignment is evaluated based on students' attention to (especially) small details that would be difficult or time-consuming to try to correct by reprompting.

And for what it's worth, here's a bonus assignment written by ChatGPT-4:

Assignment #6: Predicting Outputs

Ask students to predict how ChatGPT might respond to a series of prompts related to course material. They should write down their predictions before entering the prompts into the AI. After seeing ChatGPT's actual responses, students should reflect on any surprises or discrepancies. The assignment is evaluated based on the thoughtfulness of the predictions and the depth of the reflection.

That sounds like a great idea to me. While LLMs are a powerful brainstorming tool, we still want to encourage students to develop some idea of what they're looking for prior to their first interactions with it. So this assignment shouldn't be evaluated on the *accuracy* of students' predictions; often, a mismatch between expectations and reality can lead to productive dialogue. Instead, we're aiming to guide students to make *specific, targeted* predictions that draw upon and synthesize their own background knowledge of course material, and to reflect on how the LLM's actual contribution differs in both form *and* content. In this way, we help ensure that LLMs function as a sounding board, not a replacement for students' own brainstorming.²

Ricky Mouser

Indiana University

rimouser@iu.edu

Ricky Mouser is a PhD candidate in Philosophy at Indiana University Bloomington who works broadly in value theory. His research explores the dangers of value capture, the incommensurability of our values, and just how much the philosophy of sports and games can teach us about well-being.

² Thanks to Eric Alt, Binh Doan, Joseph Edelheit, Jeffrey He, Taylor Kainer, Mike Ren, Cargile Williams, and my anonymous referee for their thoughts and feedback.

Bibliography

- Acovino, Vincent, Mary Louise Kelly, and Halimah Abdullah. "A Sci-fi Magazine Has Cut off Submissions after a Flood of AI-Generated Stories." *NPR*, February 24, 2023.
www.npr.org/2023/02/24/1159286436/ai-chatbot-chatgpt-magazine-clarkesworld-artificial-intelligence.
- Beck, Natalie. *The Effects of Incorporating a Word Processor into a Writing Program: Seven Individual Case Studies*. Joondalup: Faculty of Education, Edith Cowan University, 2000.
ro.ecu.edu.au/cgi/viewcontent.cgi?article=1523&context=theses_hons.
- Bouville, Mathieu. "Plagiarism: Words and Ideas." *Science and Engineering Ethics* 14 (2008): 311–322.
- Bretag, Tracey, and Saadia Mahmud. "Self-Plagiarism or Appropriate Textual Re-use?" *Journal of Academic Ethics* 7 (2009): 193–205.
- Center for Artistic Inquiry and Reporting. "Restrict AI Illustration from Publishing: An Open Letter." *Artistic Inquiry*, May 2, 2023. artisticinquiry.org/AI-Open-Letter.
- Clarkesworld. "Submissions." *Clarkesworld*, accessed June 3, 2023.
clarkesworldmagazine.com/submissions/.
- Dell'Acqua, Fabrizio and McFowland, Edward and Mollick, Ethan R. and Lifshitz-Assaf, Hila and Kellogg, Katherine and Rajendran, Saran and Kraymer, Lisa and Candelon, François and Lakhani, Karim R., Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality (September 15, 2023). Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, Available at SSRN: <https://ssrn.com/abstract=4573321> or <http://dx.doi.org/10.2139/ssrn.4573321>.

Fowler, Geoffrey A. "We Tested a New ChatGPT-Detector for Teachers. It Flagged an Innocent Student." *The Washington Post*, April 1, 2023.

www.washingtonpost.com/technology/2023/04/01/chatgpt-cheating-detection-turnitin/.

Gates, Bill. "The Age of AI has Begun." *GatesNotes*, March 21, 2023.

www.gatesnotes.com/The-Age-of-AI-Has-Begun.

Helgessen, Gert, and Stefan Eriksson. "Plagiarism in Research." *Medicine, Health Care and Philosophy* 18 (2015): 91–101.

Kirchner, Jan Hendrik; Lama Ahmad; Scott Aaronson, and Jan Leike. "New AI Classifier for Indicating AI-Written Text." *OpenAI*, January 31, 2023. openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Klein, Naomi. "AI Machines Aren't 'Hallucinating'. But Their Makers Are." *The Guardian*, May 8, 2023. www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein.

Krause, Steven D. "My Talk About AI at Hope College (or Why I Still Post Things on a Blog)."

Steven D. Krause, May 2, 2023. stevendkrause.com/category/ai-writing/.

Lawrence, Jill. "Math Teachers Protest Against Calculator Use." *The Item*, April 5, 1986.

Mollick, Ethan. "It Is Starting to Get Strange." *One Useful Thing*, May 2, 2023.

www.oneusefulthing.org/p/it-is-starting-to-get-strange.

Mollick, Ethan. "My Class Required AI. Here's What I've Learned so Far." *One Useful Thing*,

February 17, 2023. www.oneusefulthing.org/p/my-class-required-ai-heres-what-ive.

Mouser, Ricky. "Shit and Bullshit Jobs." *Rapid Fire*, April 14, 2023.

www.rickymouser.net/post/shit-and-bullshit-jobs.

Mouser, Ricky. "Writing with ChatGPT." *Rapid Fire*, March 31, 2023.

www.rickymouser.net/post/writing-with-chatgpt.

Nguyen, C. Thi. *Games: Agency as Art*. New York: Oxford University Press, 2020.

O'Brien, Matt, and Hannah Fingerhut. "Artificial intelligence technology behind ChatGPT was built in Iowa — with a lot of water." *AP News*, September 9, 2023.

www.apnews.com/article/chatgpt-gpt4-iowa-ai-water-consumption-microsoft-f551fde98083d17a7e8d904f8be822c4

OpenAI. "GPT-4 Technical Report." *OpenAI*, March 27, 2023. cdn.openai.com/papers/gpt-4.pdf.

Perrigo, Billy. "OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic." *Time*, January 18, 2023. www.time.com/6247678/openai-chatgpt-kenya-workers.

Poorvu Center for Teaching and Learning. "Fair Paraphrase." Accessed June 3, 2023.

poorvucenter.yale.edu/undergraduates/using-sources/understanding-and-avoiding-plagiarism/fair-paraphrase.

Poorvu Center for Teaching and Learning. "What Is Plagiarism?" Accessed June 3, 2023.

poorvucenter.yale.edu/undergraduates/using-sources/understanding-and-avoiding-plagiarism/what-plagiarism.

Saenko, Kate. "Is generative AI bad for the environment? A computer scientist explains the carbon footprint of ChatGPT and its cousins." *The Conversation*, May 23, 2023.

<http://www.theconversation.com/is-generative-ai-bad-for-the-environment-a-computer-scientist-explains-the-carbon-footprint-of-chatgpt-and-its-cousins-204096>

Sorokina, Daria, Johannes Gehrke, Simeon Warner, and Paul Ginsparg. "Plagiarism Detection in arXiv." In *Proceedings of the Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 2006.

Tegmark, Max. *Life 3.0*. New York: Borzoi Books, 2017.

Williams, Cargile. "GPT-4 and AI Acceleration." *UX Ethics and Philosophy*, March 22, 2023.

www.cargilewilliams.com/post/gpt-4-and-ai-acceleration.

Wong, Jessica. "Teachers Are Split on Bringing ChatGPT into Elementary, High Schools." *CBC*,

April 11, 2023. www.cbc.ca/news/canada/chatgpt-highschool-elementary-1.6802336.