Teaching with Large Language Models in the History of Philosophy: A Recursive Approach

Michael Otteson

Utah State University

Introduction

The rise of machine learning and specifically large language models (LLMs) has created a watershed moment in higher education. Never before in human history have automated systems been capable of replicating or imitating human language to this degree of plausibility. Written essays, once a staple of college humanities and social science courses, are potentially suspect as meaningful exercises in practice and evaluation for students.[1] The temptation to use these technologies to shortcut the writing process may be substantial for many college students, especially ones who mainly view their education as a credentialing or vocational exercise as opposed to an opportunity for personal growth. If it is unclear to a student why they must take humanities courses unrelated to their major or (ostensibly) their career path, they may not understand what they are missing by forgoing practice in written communication.

This problem could give rise to another, given that students often compete to get into graduate school or well-paying jobs after they leave college. College is a large financial investment on the part of students, and thus they have an interest in earning enough after graduation to pay for it. This could lead to a collective action problem for universities if the LLMs represent a significant competitive advantage on essays for the students who use them. Students who don't use LLMs to complete traditional writing assignments may typically get worse grades than their peers who do, in which case the use of LLMs may become a baseline technique for those who want to avoid competitive disadvantage on job or graduate school applications.[2] Even students who want to do good work and engage with the writing process may feel pressure to cut corners to keep up with their peers. In such an environment, universities have an interest in developing pedagogical practices that will help them maintain the legitimacy of their credentials, and, more deeply, their ability to help students learn the skills and competencies that they believe are essential for intellectual thought and activity.

However, LLMs are not just threats or impediments to education. Indeed, the new technology, like many other technologies before it, represents both danger and opportunity for higher education. There are many legitimate uses of LLMs, and they will increasingly become part of standard, everyday life and education, much like search engines, calculators, and word processors before them.[3] However, as with these other tools, they will

[1] Daniel Herman, "The End of High-School English," *The Atlantic*, December 9, 2022, https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/672412/. There are of course many examples of this kind of article, but this is a representative one.

[2] Russell Hardin and Garrett Cullity, "The Free Rider Problem," *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/entries/free-rider/. As this article notes in Section 1.1, the discussion of the free rider problem is at least as old as Plato (*Republic* II.360b-c, Grube translation), and there has been a long history of the topic in subsequent philosophy, economics, and related fields. While I am not aware of any study that specifically focuses on LLMs and the incentive to cheat in college courses or philosophy classes specifically, the basic concern is straightforward. If Jill and Jane are competing for a spot in medical school, and Jill uses a cheat code to help her get in without getting caught, then Jane also has an incentive to use the same tool even if she wouldn't be inclined to cheat otherwise.

[3] There are legitimate concerns about LLMs and appropriate credit for creative work, but nonetheless these concerns do not seem to have stopped the proliferation of the technology. See the lawsuits filed against the creators of one of these LLM models: Matt O'Brien, "Sarah Silverman and novelists sue ChatGPT-maker OpenAI for ingesting their books," *AP News*, July 12, 2023, https://apnews.com/

Teaching History, 49(1).
DOI: 10.33043/66z64ff68

only benefit students (and society at large) if they learn how to use them as a supplement to creative and critical thinking in as opposed to a substitute for them. Access to generative AI does not make someone a historian or classicist any more than possessing a calculator makes a person a mathematician. In light of this, it is important for university instructors to think about how to help students understand the value and limits of LLMs in their courses and assignments.

In this paper, I outline my own attempts to produce assignment prompts that "lean in" to LLMs but still require students to master the same skills and competencies as traditional essay assignments that have long been a staple of humanities disciplines. I explain how I came to think about this topic based off suggestions from Julia Staffel, a philosopher who was already grappling with this topic. I go over her suggestions for how to deal with the problem of LLMs and explain how I crafted an assignment prompt that built off her insights. The assignment involves student evaluation of AI content that was recursive or circular in nature, and I discuss how students have performed on the assignment prompt. Finally, I review how the results indicate that the assignment can help students develop some of the same critical evaluation skills that are necessary for traditional essay assignments without the same vulnerability that written essays have to LLM shortcuts.

Possible Ways Forward

Since Spring 2023, I have developed assignment prompts that have asked students to use LLMs on larger, take-home projects in my history of philosophy and ethics classes. Like many academics in the winter of 2022-23, I was concerned about how my classes, which had utilized many essay prompts, would no longer be useful in an academic context. However, there were already philosophers focused on this problem who publicly discussed how to deal with LLMs. Julia Staffel published a video that went through various features of ChatGPT and teaching philosophy. Her suggestions were central to the development of my assignment prompt. As I discuss below, she suggests a number of assignments that could potentially avoid abuse by LLMs. The ones I decided to pursue include her suggestion to add a recursive element to the presentation, change assignments to video presentations (instead of written ones), and ask for extensive citations to the philosophical texts that the assignments were based around.[4]

Details of the Assignment

I built an LLM-based prompt for my Ethics class in Spring 2023 that asked students to follow up on this recursive model. In my lower division ethics course, I gave students the same prompt that I gave for traditional essay assignments. In the case of a paper on Aristotle, the prompt is as follows: "Explain what Aristotle says human happiness is and how virtue relates to human happiness." However, instead of asking the students to write an essay, I told them to plug this prompt into an LLM. Once they did this, they were to ask the LLM to produce an answer to the prompt. Once they secured an answer, the prompt then instructed them to put the answer they got from the LLM back into the LLM and ask for a critique of the answer, thus generating a second piece of writing from the LLM. I did this to incorporate the circular element from Staffel's presentation on the topic. Insofar as LLMs are not as adept at criticizing their own work, this would serve as an opportunity for students to carefully consider the original source material that the prompts are drawn from and critically assess writing about these primary texts.

The assignment then asks students to produce a video presentation to explain which of the two responses was superior to the other. I did this for three reasons. First, it added another layer of recursion to the process, which would make it even more difficult to simply plug a prompt into the LLM and get a passable output without any

article/sarah-silverman-suing-chatgpt-openai-ai-8927025139a8151e26053249d1aeec20

[4] Julia Staffel, "Writing Assignments and Chat GPT," YouTube, January 4, 2023, educational video, 20:20-20:30, 15:30-15:40, 15:50-16:00, https://www.youtube.com/watch?v=bkjVkfU9Gro&ab_channel=juliastaffel

additional thinking or understanding of the material the on the part of the student. This follows Staffel's suggestion to have students evaluate different responses produced by LLMs.[5] Second, as Staffel also points out, having the student do a video presentation would make it much harder to read off of a script that an LLM produced given how difficult it is to read that kind of a script and come off as natural and conversational.[6] It would have been noticeable if a student didn't understand what the presentation was about. Third, the assignment itself provides my students with an alternative mode of assessment that allows them to develop verbal communication skills that traditional essay do not offer.

I told my students that they needed to provide evidence from our readings to serve as the basis for their evaluation of the ChatGPT output. They had to quote or reference the text of *Nicomachean Ethics* where they explained whether ChatGPT got anything right or wrong about what Aristotle said in his own work. They also had to do the same thing if they felt that the two responses from the LLM had left anything critical out from the text that related to the prompt. Thus, students still had to read and analyze the text carefully in order to justify their critiques of the LLM responses. Students could skip the readings and lectures from the course at their own peril. As Staffel notes, ChatGPT often comes up with bogus sources that don't exist, or fake quotations from real texts, so requiring sourcing on the part of students is another way to prevent or disincentivize simple copy/paste submissions from students.[7]

In Spring 2024, I also assigned a slightly modified version of the same assignment to my two Ethics classes in order to test a more streamlined version of the assignment. The prompt that students gave to the LLM was the same[8], but instead of asking students to feed the first ChatGPT response back to itself for critique, I had them ask the deep neural network to produce two plausible but different responses to the prompt. They then had to explain which response was better using supporting evidence from the text of *Nicomachean Ethics* analogous to the 2023 version of the assignment. This allowed for less complexity in the assignment design while still retaining elements of recursion (in line with Staffel's suggestions) that would prevent utilizing copy/paste submissions.

Student Performance on the Assignment

I first gave my students the assignment prompt for my class at for the last assignment of the semester in 2023. However, I still let them pick a traditional essay assignment if they wanted.[9] This enabled a relative comparison in the semester right after the release of publicly accessible LLMs.[10] While it is possible that many students could have utilized the technology to write traditional essays, at no point in the future would knowledge (and use) of the technology be as limited as it was then, assuming that familiarity with new technology generally increases over time instead of decreasing. Thus, there would be no better time to run something analogous to an experiment in a classroom setting. Now, this is not a perfect experiment given the lack of randomized selection to the "control"

[5] Staffel, "Chat GPT," 20:20-20:30. I would also like to thank David Ménager of Parallax Research and Ramón Alvarado of the University of Oregon for their help in workshopping the recursive aspect of this assignment.

[6] Staffel, "Chat GPT," 15:30-15:40

[7] Staffel, "Chat GPT," 15:50-16:00

[8] "Explain what Aristotle says human happiness is and how virtue relates to human happiness."

[9] There is also research that suggests that giving students autonomy in their learning is beneficial. See Chris Babits, "A Fun and Different Course: How Gamification Transformed an Online U.S. History Survey," *Teaching History* 48, no. 1 (2023): 65-75, https://openjournals.bsu.edu/teachinghistory/article/view/4273. Babits discusses on pg. 68-69 the positive feedback he got from students when he gave his students some autonomy over their assignment choices in his course. Babits explains this feedback in part by referencing Ryan and Deci's article on Self-Determination Theory (SDT). SDT holds that an intrinsic motivation to perform an activity or task is far more likely to develop when the agents in question have autonomy over their participation in their activities or work. For their discussion, see Richard M. Ryan and Edward L. Deci, "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being," *American Psychologist* 55, no. 1 (2000): 70.

[10] In total, twenty-one students wrote an essay, and eight completed the LLM assignment.

and the “treatment” group, but to do this to students would not be fair to students if both options are available.[11] Furthermore, insofar as a student picked the LLM assignment, this suggests that, all things being equal, the student likely feels more comfortable with the technology, which may in fact skew the comparison in favor of the LLM assignment (though this is ultimately speculation).

When my students turned in their assignments, there were many students who picked each option, though there were less who went with the LLM option. Nonetheless, the final results were telling: the average for both assignments was 88%. Given that this was the last assignment of the semester for each class (and they had had a chance to receive feedback and instruction from me for a semester at that point), it was not surprising that their scores were overall higher than previous versions of the traditional essay assignment earlier in the semester. It should also be noted that almost 63% students received a grade in the “B” range for the LLM option. This means that the assignment was tractable for many students, but it also wasn’t an automatic “A” for them either. In reviewing the specifics of the submissions, it was clear that many of the students could correctly identify certain aspects of the core concepts related to the prompt but did not have a mastery of all of them. Aristotle’s *Nicomachean Ethics* is a difficult text that has admitted of serious philosophical argumentation and disagreement for most of its 2400-year history, and it is not surprising that many students do not score perfectly on an assignment that asks them to interpret it. The assignment thus was able to demonstrate to me as an instructor what they did and did not understand about the course material, which is one of the primary goals of any assignment at the college level.

Furthermore, the fact that there is something of a distribution of scores indicates that the assignment is at an appropriate level for lower division college courses. Insofar as the average score is neither an “A” nor an “F”, the assignment seems to present a set of tasks that are tractable but challenging for students in the class.

The results from Spring 2023 suggested to me that the LLM assignment was viable, so I eliminated the traditional essay option in Spring 2024. The grades distribution was somewhat similar. In one class, the average on the same Aristotle assignment was 87%, and in the second it was 83%. In the first class, 50% of students received a grade in the “B” range and almost 59% were below the “A” range, and in the second class 72% of students fell in the “B” range while almost 89% of students fell below the “A” range. These scores indicate that once again, the assignment is at an appropriate difficulty for college students in lower division classes. It was doable but not trivially easy for students.

Changes for the Future

It is worth noting that the particulars of LLM assignments may have to change along with new methods and procedures for utilizing these tools. It may be the case that the users of the technology develop new techniques for eliciting responses from LLMs that will require different assignment prompts that LLMs can easily complete with one prompt from a student. Going forward, there will have to be a certain amount of trial and error on the part of instructors. Part of this involves using the assignment prompts like the one covered in this paper over the course of multiple semesters and evaluating how students interact with it. I am offering a modified version of this assignment in upper division courses along with other assignment options involving LLMs that students can pick from depending on their preferences. After assigning this LLM centered assignment over multiple semesters, there are some modifications that I will institute going forward.

One of these modifications involves how to frame generative AI. Students often approach the outputs of LLMs with credulity. They seemed primed to believe that LLMs offer substantive and useful responses to the prompts they fed the deep neural networks. It was often the case that students complimented ChatGPT on how well it responded to the prompts. As the instructor, it was clear to me that these systems could pull adequate definitions from notable philosophical texts (found in many places on the internet) but could not adequately explain those definitions. Furthermore, they had a hard time explaining complex arguments found in these texts, and they were not able to synthesize and explain long chains of reasoning and logical moves that are inherent to complex philosophical or analytic projects. LLMs often reverted to vague platitudes that were often restatements

[11] I cannot force them to separate into groups for the sake of an experiment where grades are involved.

of the definitions they provided as opposed to arguments supporting them. It seems likely that they did this because of their basic function as predictive algorithms that generate words based on what words often appear in online discussions of the philosophers in question.

Here is a representative response to my Aristotle assignment prompt that I elicited from ChatGPT:

> Virtue, in Aristotle's view, is excellence of character. It's the mean between two extremes: deficiency and excess. For example, courage is the mean between cowardice and recklessness, while generosity is the mean between stinginess and extravagance. Virtue is developed through habituation and practice, leading to moral and intellectual excellence.[12]

This response is technically correct, but it leaves out the actual definition of virtue broadly (Nicomachean Ethics [NE] Book II, Chapter 6 [II.6] 1106a15-20), the distinction between moral and intellectual virtue and what role they play in the soul for Aristotle (NE I.13 1102a32-1103a10, NE II.1 1103a11-25), or the connection between moral virtue and the passions (NE II.3 1104b13-16).[13] It goes on to say that, "virtue enables individuals to fulfill their potential and achieve their true purpose in life," which arguably gets close to explaining the definition of virtue writ large, but it doesn't ever define what is the substance of human life for Aristotle. This is ChatGPT 3.5, and the upgraded versions offer better responses, but this version demonstrates problems that all of them have in an especially concise form (and is the one students are most likely to use).

I am going to explain to my students before the assignments are due that they should not automatically trust these systems, and in the future, I will assign Harry Frankfurt's On Bullshit, which deals with the difference between language that is meant to get at truth and language that is meant to project an image.[14] The term Harry Frankfurt gave for this second kind of language is "bullshit." While the term is obviously pejorative, deep neural networks don't bullshit intentionally, and thus by themselves don't deserve the same opprobrium as people who bullshit. The tools are, again, predictive algorithms that can help humans sift through large amounts of information if the humans know what they are doing. In other words, the tools have their uses even if they cannot care about the truth in a way that a human should. The point of bringing this up to students in the first place is to emphasize that while the tools themselves could be morally neutral, passing their outputs off as simply your own work is perhaps a paradigmatic case of bullshitting (and is thus wrong). Anyone who puts forward language from ChatGPT after a simply copy/paste as their own does not really care about whether the statements are true; they just want to convince you that they know what they are talking about when they do not. Frankfurt's work will hopefully help students understand how to approach generative AI text in a productive way and recognize that they can at times give information that is not connected to the truth in the way that their own language should be. Adding some discussion of "hallucinations" and LLMs in my classes may also help with this as well.[15]

Another modification entails altering the assignment instructions in regard to citations. The later versions of ChatGPT (and other bots like Perplexity AI) are getting better at accurate citations to the text, but they are still somewhat cursory and lack extended integration of those citations into a larger argument or thesis that would be required for a cohesive project. In light of this, I will tell my students to ask ChatGPT for responses to the prompt that are explicitly adversarial to each other. This will help make them consult the text directly and offer more extended explanations for why one response is superior to the other.

[12] Text generated by ChatGPT 3.5, May 30, 2024, response to "Explain what Aristotle says human happiness is and how virtue relates to human happiness," OpenAI.

[13] All references can be found in the Revised Oxford Translations of Aristotle. Aristotle, *Nicomachean Ethics*, trans. W.D. Ross/revised by J.O. Urmson in *The Complete Works of Aristotle: The Revised Oxford Translation*, ed. by Jonathan Barnes (Princeton University Press, 1984).

[14] Harry Frankfurt, *On Bullshit*, (Princeton: Princeton University Press, 2005), 31-32, 47-48, Apple Books.

[15] Cade Metz, "Chatbots May 'Hallucinate' More Often Than Many Realize," *The New York Times*, November 6, 2023 (updated Nov. 16), https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html

Analysis and Implications

There are varying degrees to which this assignment will deter students from using LLMs in a way that is a substitute for critical analysis and understanding of the text. A determined student may or may not be able to get ChatGPT or an analogous system to produce a complete script for these sorts of assignments. However, no evaluative system will be foolproof against cheating, much in the same way that no safe or security system will be able to deter every conceivable assault. Nonetheless, the assignment does in fact raise the barrier to simple copy/paste cheating significantly. While it might be possible to cause an LLM to produce a functional script for this presentation with proper citations, it would take many hours and prompts from a user to get to the point of a workable prototype. This is in many ways a virtue of this style of assignment in the age of deep neural networks. LLMs will not be able to produce quality material with one prompt, much in the same way that computer programs generally require a lot of careful work from programmers to produce useful outputs. Often, this requires a lot of trial and error that only a skilled programmer will be able to perform, precisely because it involves reconfiguring inputs to produce better results requires an underlying understanding of the goals, methods, and processes that govern the activity.

In the same way, students and LLM users generally must understand complex philosophical ideas and arguments in order to produce quality writing in conjunction with deep neural networks. A student who doesn't understand, for instance, Aristotle's Nicomachean Ethics will not be able to see errors in an LLM's comments on Aristotle's conception of virtue or his definition of happiness. They will be unable to adequately evaluate competing interpretations of the same text without first reading the work in question and having a firm grasp of what the philosopher says. Finally, they will not be able to provide evidence to support their own evaluation of rival interpretations if they do not know what the philosophers argue for in their work. All of this is manifest in my own students' attempts to do this assignment. Some students received high marks on their projects, but many of them did not. Indeed, the average for both the traditional essay assignment and the ChatGPT presentation were not substantially different. Doing this kind of work with LLMs is not a shortcut to an "A" grade in the class, or even a higher grade generally than traditional essay assignments.

In other words, telling students to engage with LLMs as they work through their own understanding of complex writing and concepts will require them to require the same level of engagement with course materials as a traditional essay. With a sufficiently circular assignment prompt, regurgitating the first output from an LLM will not help students earn high grades in their classes or master the skills that college is supposed to impart and test for. Recursive engagement with LLMs thus enables humanities instructors to continue advancing the core goals of their courses and the broader mission they have within the university. The results from my classes suggest that Staffel's suggestions about how to incorporate generative AI into college level classrooms are a fruitful way forward.

Indeed, the nature of LLMs requires certain skills of students that are perfectly in line with learning methods that are legible within existing pedagogical frameworks. In some ways, LLMs are not substantially different from what tools like Google or JSTOR were able to provide students previously. As with these older tools, LLMs can allow students access to a vast body of information that would have been impossible for the philosophers and scholars of centuries past to access. None of this, however, does away with the need to understand the methods and practices of any given field. In fact, information networks of this size require even more from users in order to utilize them effectively. Information without wisdom or knowledge is unintelligible. This means that assignments that require students to use LLMs in the right way would be just as useful even if somehow it was impossible to use them as a crutch and traditional essay assignments were not as vulnerable to them. LLM oriented assignments, properly calibrated, will allow instructors to continue offering assignments that help students to improve their critical analysis and evaluation skills without the worry of widespread shortcuts that make their classes pointless or irrelevant. My hope is that this assignments stresses to students the importance of original thinking and genuine comprehension in a way that helps them understand that they cannot merely rely on these sorts of systems to do their work for them, even if they are an extremely powerful tool in some respects.[16]

Works Cited

Aristotle. *Nicomachean Ethics*. Translated by W.D. Ross/revised by J.O. Urmson. In *The Complete Works of Aristotle: The Revised Oxford Translation*. Edited by Jonathan Barnes. Princeton, NJ: Princeton University Press, 1984.

Babits, Chris. "A Fun and Different Course: How Gamification Transformed an Online U.S. History Survey." *Teaching History* 48, no. 1 (2023): 65-75. https://openjournals.bsu.edu/teachinghistory/article/view/4273

Frankfurt, Harry. *On Bullshit*. Princeton: Princeton University Press, 2005. Apple Books.

Plato. *Republic*. Translated by G.M.A. Grube and C.D.C. Reeve. Indianapolis, IN: Hackett Publishing Company, 1992.

Hardin, Russell and Cullity, Garrett, "The Free Rider Problem," in *The Stanford Encyclopedia of Philosophy* edited by Edward N. Zalta. Stanford: The Metaphysics Research Lab, Winter 2020 Edition. https://plato.stanford.edu/archives/win2020/entries/free-rider/.

Herman, Daniel. "The End of High-School English," *The Atlantic*, December 9, 2022. https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/672412/

Metz, Cade. "Chatbots May 'Hallucinate' More Often Than Many Realize," *The New York Times*, November 6, 2023 (updated Nov. 16). https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html

O'Brien, Matt. "Sarah Silverman and novelists sue ChatGPT-maker OpenAI for ingesting their books." *AP News*, July 12, 2023, https://apnews.com/article/sarah-silverman-suing-chatgpt-openai-ai-8927025139a8151e26053249d1aeec20

Ryan, Richard M., and Deci, Edward L. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55, no. 1 (2000): 68-78.

Staffel, Julia. "Writing Assignments and Chat GPT." Youtube Video, 22:06. January 4, 2023. https://www.youtube.com/watch?v=bkjVkfU9Gro&ab_channel=juliastaffel

[16] If assignments like this can help students see the value of original work, it will also potentially help them understand the importance of giving proper credit to those who produce good writing and content.