# The ChatGPT Exam: Critiquing Generative AI to Assess Learning

Ashley Black
California State University, Stanislaus

It is difficult to overstate the alarm that generative AI has inspired among university educators in just two short years. So much of the panic we have seen undoubtedly stems from the sudden appearance of this technology in our classrooms and the rapid pace at which it continues to evolve. Keeping up with artificial intelligence is a daunting task, but it is imperative that we, as educators, engage with this technology. As José Bowen and Edward Watson have recently argued, AI literacy will be a vital skill for many of our students as they enter a rapidly evolving workforce, and we must prepare them for this.[1] Despite my own trepidation at this task, my first attempt to incorporate artificial intelligence into my teaching has shown me how much potential this tool has to improve student learning and teaching effectiveness if we use it responsibly and teach our students to do so as well.

In the fall of 2023, as faculty faced the new academic year, it felt as though anxiety around artificial intelligence was reaching a fever pitch. I knew I would soon have to face the reality of what this new technology meant for my teaching, but my strategy to that point had been, quite frankly, to bury my head in the sand. That shifted when I read an article in Perspectives magazine by Jonathan S. Jones, published that September.[2] In the article, Jones describes a classroom activity he designed to address AI: students work in small groups to critique an essay written by ChatGPT related to their course content. Their objective is to evaluate the essay using what they have learned in class. I was intrigued by the idea, and the piece empowered me to engage generative AI head-on.

Immediately, I saw the potential for this assignment, but I wanted to take it beyond an in-class activity to use it for assessment purposes—to replace the traditional exam. When assessing student learning, my primary focus is on higher-order thinking; I am less interested in their ability to recall facts than their comprehension of complex ideas and their capacity to synthesize a wide range of materials. When testing students, I have therefore relied on short answer and essay questions, but I have long been frustrated with the format. Timed exams tend to reward students who are good at taking tests while disadvantaging others.[3] At the same time, I have struggled to design an alternative that achieves the same objectives: to accurately measure learning while incentivizing students to pay close attention in class, take thorough notes, and review the material. With midterms approaching, I decided to implement Jones' ChatGPT assignment as a test.

The design for the exam was simple. First, I plugged the prompt for their midterm essay into ChatGPT. I used the free version, which was then GPT 3.5. While there are more powerful AI tools available, I chose GPT 3.5 for two main reasons: first, the limited capacity seemed more likely to produce the kind of errors that students could identify and, second, I wanted to use the tool most likely to be used by my students, who are predominantly low income and unlikely to pay a premium to access more powerful AI models. I then uploaded the result to Canvas and allowed them an entire day to work through the AI-generated essay independently and asynchronously. Although I briefly considered assigning this as in-class test, I decided that allowing them ample time to complete

[1] José Antonio Bowen and C. Edward Watson, *Teaching with AI: A Practical Guide to a New Era of Human Learning* (Baltimore: Johns Hopkins University Press, 2024), 26.

[2] Jonathan Jones, "Students Critique a ChatGPT Essay," *Perspectives on History*, September 2023, https://www.historians.org/research-and-publications/perspectives-on-history/september-2023/students-critique-a-chatgpt-essay-a-classroom-experiment.

[3] Donald A. Saucier, Noah D. Renken and Ashley A. Schiffer, "Five Reasons to Stop Giving Exams in Class," *Faculty Focus*, Feb. 18 2022, https://www.facultyfocus.com/articles/educational-assessment/five-reasons-to-stop-giving-exams-in-class/.


Teaching History, 49(1).
DOI: 10.33043/gg58bfzzg

the task at home would more likely achieve the objective of having them thoroughly review the class materials. Their instructions were threefold: (1) identify any errors or places where they disagreed with the essay, (2) add citations from course materials where relevant, and (3) build in evidence that was missing from a very surface-level essay using only assigned class materials. Apart from the first task, this exercise is not unlike essays I have assigned in the past, either as in-class or take-home exams: students are required to review the material and consider how everything they have learned connects to a larger theme, drawing connections across a range of topics and sources. Although they are not responsible for correctly answering the essay question, the assignment tests their knowledge by having students identify incorrect information. It achieves this while removing the onus on writing.

Though there were a few problems to address, I was happy with the results overall. I used the ChatGPT exam in two classes that semester, for both midterms and finals, and have continued to use it since. I have solicited feedback from students that has helped me refine the assignment, and I have been impressed with the outcomes. The ChatGPT exam addresses my original objectives to assess learning and incentivize students to engage deeply with the material, but it also does much more. This assessment addresses one of the most significant challenges posed to instructors by AI: it provides a way to educate students about its limitations and responsible use.

The most important outcome, however, was one I had not anticipated: as students identified problems with AI-generated essays, they grew more confident in their abilities. I teach at a Hispanic-Serving Institution with a student body that is predominantly first-generation; most are Pell-eligible, and a significant number speak English as their second language.[4] Many of my students enter college with the belief that they do not belong there. For some, that idea persists until the day they graduate.[5] Disabusing them of that notion means developing their skills and building their self-confidence. After completing the first ChatGPT midterm in September 2023, my students reported that critiquing AI-generated writing helped them realize how much they had learned. This was the exact opposite outcome of what I have observed using traditional exams, which leave many of my students demoralized.

One of the most disheartening experiences I have had in the classroom is watching some of my brightest and most engaged students score terribly on exams. Testing caters to students who are good at taking tests: those who think fastest, write quickly, or have the best recall. Timed exams do not reward careful, methodical thinking, and they often disadvantage second-language learners and students from marginalized backgrounds.[6] For these reasons and more, forgoing timed exams has been identified as an effective practice of inclusive teaching.[7] My alternative to blue book exams has generally been the take-home essay. But this presents its own set of problems. By giving students an extended period to compose a carefully considered essay, the assignment—by its very nature—puts more emphasis on writing than a timed exam. This now disadvantages those who struggle with writing, again including my many second language learners.[8] Furthermore, in the age of generative AI, some students will undoubtedly turn to ChatGPT as they prepare their take-home essays. I need an AI-proof assessment that incentivizes students to review the material, gives them time and space to carefully consider their ideas, and disaggregates the task of synthesis from the task of writing. The ChatGPT exam has achieved all these objectives.

---

[4] Our most recent institutional data, from fall 2023 identifies 64.7% students as under-represented minorities, 70.9% as first generation, and 58.2% as Pell eligible.

[5] Writing about the "imposter phenomenon," Jeff Davis notes that "negative feelings associated with anxiety about academic success are more frequent and more acute for first-generation students." Jeff Davis, *The First Generation Student Experience: Implications for Campus Practice, and Strategies for Improving Persistence and Success* (New York: Routledge, 2010), 49. ProQuest.

[6] Morton Ann Gernsbacher, Raechel N. Soicher and Kathryn A. Becker-Blease, "Four Empirically Based Reasons Not to Administer Time-Limited Tests," *Translational Issues in Psychological Science* 6, no. 2 (2020): 178-181, http://dx.doi.org/10.1037/tps0000232.

[7] Gil Moreu and Markus Brauer, "Inclusive Teaching Practices in Post-Secondary Education: What Instructors Can Do to Reduce the Achievement Gaps at U.S. Colleges," *International Journal of Teaching and Learning in Higher Education* 34, no. 1 (2022): 178. ERIC.

[8] Zakrajsek and Nilson note the importance of using a wide variety of teaching strategies and assignments in the classroom. In a discipline that prioritizes writing, inclusive teaching means providing alternate ways for students to demonstrate their learning. Todd Zakrajsek and Linda Nilson, *Teaching At Its Best*, 5th ed. (Hoboken, NJ: Jossey-Bass, 2023), 109.

Below, I will provide a detailed description of the assignment and an evaluation of what worked, what did not, and how I have revised the assignment in response to student feedback.

## Designing the Assignment

The first step in preparing this assignment is to have AI generate a suitable essay. There has been a significant learning curve in this process, and the first few times I plugged in essay prompts, I received some alarmingly good essays in response. In addition to being clearly written, they touched upon the most important points of the assigned topics and provided specific examples as illustration. Giving these to students would be counterproductive; not only would it be difficult for them to critique such essays, but it could also convey to them the value of using generative AI in place of their own writing. A key step in this process, then, is to familiarize yourself with "prompt engineering": the art of using the right inputs to get your desired output. For me, the solution is to make the prompt as broad as possible while limiting the word count. I focus on expansive, overarching themes that will allow students to draw upon a wide range of class materials. This approach will prompt ChatGPT to produce a shallow essay loaded with arguments but lacking in detail and supporting evidence. This creates an opportunity for students to plug in everything that is missing, drawing on course materials to either back up or refute statements made by AI. An essay of 800 words limits the amount of detail ChatGPT will provide, and it is a reasonable length for my students to work with.

A second useful strategy I have found in generating effective essays is to ask AI to consider a specific reading or case study when composing its response. ChatGPT and other text generators are notoriously bad with sources and have been known to "hallucinate," to invent citations or factual information when it doesn't know the answer, and to state such facts with great certainty. According to an April 2024 article in the F*inancial Times*, when asked to generate scientific abstracts, ChatGPT hallucinated roughly 30% of references.[9] Incorporating a specific book into the prompt will often lead ChatGPT to make mistakes, although the character of those mistakes will depend on how widely the book is known. In fall 2023, for example, I had students in one of my classes read Elizabeth Newman's *Biography of a Hacienda*, an award-winning scholarly monograph in historical archaeology.[10] The hacienda in question caused all kinds of problems for the AI, which identified it by the wrong name that it then included in a falsification of the book's subtitle. It also claimed that the hacienda was part of the territory lost to the United States during the Mexican American War, despite the fact that it was located in the state of Puebla (which, most assuredly, remains a part of Mexico). Alternatively, when I asked ChatGPT to consider John Charles Chasteen's *Born in Blood and Fire*, a popular teaching book now in its fifth edition, AI did considerably better.[11] The essay contained no such egregious errors, and yet it was off in more subtle ways. For example, much of the essay focused on the role of Latin America's rising middle class. While Chasteen does consider this topic, ChatGPT significantly exaggerated its importance to the text. A second problem with the essay is that it consistently lauded liberal policies in Latin America and missed Chasteen's many criticisms of nineteenth century liberalism—criticisms that echoed what students were learning from classroom discussions and lectures. The obvious errors in the first essay were easy for most students to identify, and they quickly revealed who had not read the book. However, the more subtle errors in the second essay did a better job of testing students' familiarity with the materials, for it demanded a deeper grasp of the content.

To administer the exam, I loaded the essay into Canvas with instructions to identify errors or points of disagreement, plug in citations wherever the essay connected to class content, and add missing details by drawing evidence from class materials wherever they supported the claims made by ChatGPT. They were to do all the above using the track changes feature of Microsoft Word (or the edit function of Google Docs), which had the

[9] Henry Mance, "AI keeps going wrong. What if it can't be fixed?" *Financial Times*, April 5, 2024, https://www.ft.com/content/648228e7-11eb-4e1a-b0d5-e65a638e6135.

[10] Elizabeth Newman, *Biography of a Hacienda: Work and Revolution in Rural Mexico* (Tucson: University of Arizona Press, 2014).

[11] John Charles Chasteen, *Born in Blood and Fire: A Concise History of Latin America*, 4th ed. (New York: W.W. Norton & Co., 2016).

added benefit of familiarizing my students with a tool many of them were unaware of. The exam opened at eight in the morning, and they had the entire day to complete the exercise and submit their response.[12]

After the first round of ChatGPT exams, I was excited about the potential for the assignment but saw clear areas for improvement. Despite a good deal of anxiety among my students, for whom this exercise was entirely unfamiliar, many did quite well. Those who performed best were able to clearly demonstrate their understanding and draw connections across a wide range of sources. They inserted evidence from the readings in a meaningful and effective way, supplementing ChatGPT's shallow arguments with an abundance of supporting details. Many of them demonstrated a remarkable level of creativity, connecting ideas and materials in ways I had never considered myself.

Nonetheless, there were two main problems with the first iteration of the assignment. First, I wrote prompts centered around the main readings to nudge ChatGPT to make mistakes. Despite my instructions to consider as many of the course materials as possible, some students interpreted this to mean the essay was entirely about the book and failed to incorporate any other assigned readings or lectures. Second, many students were so focused on the first instruction—to identify errors and points of disagreement—that they barely added any evidence or detail to the essay. This was particularly problematic for the Chasteen essay, whose errors were subtle enough that many students missed them entirely. By getting caught up in their search for errors, they missed the most important objective of the assignment, which was to review the assigned materials and connect them to the arguments made by AI.

Armed with my own sense of what worked with the assignment and what did not, I asked my students for their input. The biggest challenge they cited was unfamiliarity with the task; they had never done anything like this and were anxious about doing it correctly. In short, they needed more structure and more preparation. It was obvious I needed to develop a clear and detailed rubric, which I had not done for the first round because I was not yet sure what the results would look like. I was better prepared to create a rubric for the second round. This would address the problem of structure, but in terms of preparation, the best solution came from a student who suggested that I give them an opportunity to meet in small groups and do a practice essay in class ahead of time. This easily fit into my class structure, which centers around collaboration and peer support. It would also provide crucial scaffolding by allowing the students to perform the task in a low-stakes environment, lowering anxiety and building their confidence.[13]

At the end of the semester, I made several changes to address the problems outlined above. The first thing I did was create a detailed rubric that outlined three main criteria: (1) incorporation of evidence, (2) breadth of sources, and (3) familiarity with materials. The first two categories foregrounded the main objective of the exam and assigned the greatest weight to the use of evidence. To score well, the evidence that they incorporated had to be used effectively; it had to be relevant to the points made in the essay, and it had to be clear that the students understood what their evidence meant. The second criterion gave students specific targets: they had to cite at least five of their weekly lectures and multiple readings (documents, articles, or book chapters) to earn the highest possible score. Finally, "familiarity with materials" included not only success in locating errors and critiquing problems in the essay but also directed students to look for topics not covered in class materials or, alternatively, topics that we had covered but which AI missed entirely. The first two criteria were weighted most heavily, with each counting for forty out of a hundred points, while the third accounted for only twenty. Criteria one and two allowed me to stress the importance of evidence as the main objective, while two and three guided students not to get hung up on either a single source or the pursuit of errors. Expanding the third criterion to include things

[12] Although I considered limiting the amount of time students could work on the essay, I quickly realized that a limited timeframe did not contribute to the main objective of the assignment, which was to have students review and engage with course materials; the more time they had, the more they could review. Giving students the entire day also ensured the greatest amount of flexibility, which is vital for students who work multiple jobs or have children, as many of mine do.

[13] Flower Darby and James M. Lang, *Small Teaching Online: Applying Learning Science in Online Classes* (San Francisco: Jossey-Bass, 2019): 145-147, ProQuest.

missing from the essay or to identify topics not covered in class more thoroughly tested the degree to which they had learned the assigned materials. Using the example of the Chasteen essay cited above, a student could highlight the lengthy passages on the middle class and note that neither the lectures nor the readings had granted so much importance to this topic. This would demonstrate a high level of familiarity with both.

The second major change I made was to incorporate a practice session the week before the exam. I used class time to have students brainstorm major themes that we had been dealing with all semester. This gave them an opportunity to start thinking about big ideas. I then took two of their suggestions and had ChatGPT generate an essay for each; one I used for the exam and the other I handed out in class the week before. We read the practice essay out loud before I put them into groups and instructed them to make note of what sources they might use as evidence for each paragraph and identify any problems they saw with the essay. At the end of the session, they shared their ideas with the class. The practice run allowed them to work through their ideas with their classmates and gave those who better understood the assignment an opportunity to model their approach to groupmates who were less sure. They could then take the practice essay home to further develop their responses.

Performance clearly improved on the second round of exams. Nobody got so hung up looking for errors that they forgot to build in the evidence, more of them drew upon a wider range of lectures and readings, and many used annotations to build in explanatory notes and reflection alongside their evidence. Students benefited from the structure and guidance provided by the rubric, the opportunity to workshop the assignment with their peers, and the familiarity of having done the assignment once before. Although there was still some anxiety, they were much more comfortable with the task the second time.

## Outcomes

The ChatGPT exam has exceeded my expectations in achieving learning outcomes. As noted, the main objective of this assignment is to incentivize students to review and learn the materials. To score well, they must not only cite a wide range of sources, but they must demonstrate that they understand how those sources connect to each other and to the major themes of the class. As I read through the final exams in fall 2023, I was impressed by just how well some of my students were able to demonstrate their comprehension of the materials using this method. They did so by refining language where they felt the AI was too vague, providing missing definitions of complex terms and ideas, and drawing creative and thought-provoking connections between the arguments made in the essay and the materials assigned in class. In the best example I have seen so far, a student cited every lecture, every reading, and every document in a way that was cohesive and highly effective. Alternatively, there were students who merely plugged in direct quotations that were not relevant to the points being made in the essay, which indicated to me that they had reviewed the materials but failed to think critically.

Critiquing the essay has proven particularly useful in evaluating critical thinking. Failure to identify clear factual errors, such as the misnaming of the hacienda at the center of the Newman book, is an effective way to identify those who probably never read the book in the first place. Pointing to AI's overly generous take on nineteenth century Latin American liberalism, on the other hand, demonstrates a much deeper understanding of class content. Such a nuanced critique allows me to distinguish between an A- and a B-level exam. It was also telling when students failed to note the complete absence of Indigenous peoples in one of the essays, despite the centrality of this topic in their class materials. Because the bulk of the grade comes from the incorporation of evidence, students can build enough points to get into the B range by drawing in evidence from a wide range of sources, but it takes familiarity with the materials, a more nuanced category, to elevate them into the highest bracket.

An outcome of this assignment I had not considered when I designed it is that even though students are not required to write in formal prose—though they are not *doing* the writing—they are nonetheless engaging in the writing process. They are developing their ability to support arguments with evidence. As students add details, quotations, and definitions from various lectures and readings, they are synthesizing materials. Many of them critique not only factual errors or missing information, but the writing itself. After completing this assignment,

for example, numerous students have pointed out ChatGPT's tendency toward repetition and empty prose. They have also been quick to note when the essay is missing a clear thesis, or when the AI equivocates instead of taking a clear position. It is easy to be blinded by the sophisticated prose generated by artificial intelligence, but when students engage with it on a deeper level, they begin to see the flaws that are not obvious at first sight.

One of the main objectives of this assignment has been to develop information literacy and prepare students for a world in which AI is a reality of daily life.[14] To this end, the ChatGPT Exam has been highly successful. For the practice session of their final exam, I gave one of my classes an essay that was much better than the one they had worked with for the midterm. We read it together, and, before I put them into groups, I asked them to tell me what grade they would assign to this paper. Almost every student said they would give it an A. An hour later, they had identified important gaps in the essay and major themes that were not even mentioned. Most now gave it a C. It would be naïve to think this assignment will deter students from using AI entirely, nor should it. AI can be a useful tool to aid students in their work. My hope is that this demonstrates to them the danger of simply submitting a ChatGPT essay as their own, to take at face value whatever the AI tells them. My hope is that it teaches them to use this technology in a way that is ethical and responsible.

To close, I would like to return to the idea of empowering students. When I asked my classes for feedback after their first ChatGPT exam, the most memorable comment came from a student who told me it made him feel smart. He went into the midterm unsure of his knowledge, but his ability to find flaws in the work of AI—to critique, correct, and improve upon the essay—revealed to him that he knew the materials far better than he first thought. It gave him authority over his own knowledge. At an institution that is predominantly made up of minoritized students—first generation, non-white, Pell eligible—giving students confidence in their own ability to learn is perhaps the most valuable outcome of all.

[14] Bowen and Watson, 39-41.

## Appendix: ChatGPT Exam Rubric

| Criteria | Ratings | | | Points |
|---|---|---|---|---|
| **Incorporation of Evidence**<br>Provides details to support arguments made in essays. | **40 pts: Full Marks**<br>Adds ample detail in a way that is effective and coherent. All lectures, readings, documents, films, etc. work to support the points made in the essay. Reflects great effort. | **20 pts: Half Marks**<br>Adds materials in a way that may lack coherence. Materials cited may be unrelated to argument, for example. May provide few details. Reflects some effort. | **5 pts: Unsatisfactory**<br>Adds some materials but they have nothing to do with question. May provide very few details. Reflects minimal effort. | **40** |
| **Breadth of Sources**<br>Incorporates a wide range of class materials into essay. | **40 pts: Full Marks**<br>Incorporates a wide range of class materials: at least 5 lectures and multiple documents, readings and/or book chapters. | **20 pts: Half Marks**<br>Incorporates a limited range of assigned materials. May focus mainly on 1 or 2 lectures, or almost entirely on a small portion of the assigned reading. Might miss important sources. | **5 pts: Unsatisfactory**<br>All evidence comes from 1 or 2 sources, for example, a single lecture or book chapter. Does not draw any connections between different sources. | **40** |
| **Familiarity with Materials**<br>Demonstrates knowledge of content and assigned materials. | **20 pts: Full Marks**<br>Exam reflects deep understanding of assigned materials. Finds factual errors or points of disagreement. Makes note of points in the essay that were NOT covered in class. Discussion of materials is correct and reflects understanding of class topics. | **12 pts: Pass**<br>Exam reflects sufficient understanding of materials. Picks up on obvious errors in the essay (if there are any) but may miss more subtle errors. Reflects limited understanding of topics. | **5 pts: Unsatisfactory**<br>Exam reflects minimal understanding of materials. Might miss factual errors or demonstrate lack of engagement with course content. | **20** |
| TOTAL: 100 POINTS | | | | |

## Works Cited

Bowen, José Antonio and C. Edward Watson. *Teaching with AI: A Practical Guide to a New Era of Human Learning.* Baltimore: Johns Hopkins University Press, 2024.

Chasteen, John Charles. *Born in Blood and Fire: A Concise History of Latin America,* 4th edition. New York: W.W. Norton & Co., 2016.

Darby, Flower and James M. Lang. *Small Teaching Online: Applying Learning Science in Online Classes.* San Francisco: Jossey-Bass, 2019. ProQuest.

Davis, Jeff. *The First Generation Student Experience: Implications for Campus Practice, and Strategies for Improving Persistence and Success.* New York: Routledge, 2010. ProQuest.

Gernsbacher, Morton Ann, Raechel N. Soicher, and Kathryn A. Becker-Blease. "Four Empirically Based Reasons Not to Administer Time-Limited Tests." *Translational Issues in Psychological Science* 6, no. 2 (2020): 175-190. http://dx.doi.org/10.1037/tps0000232.

Jones, Jonathan. "Students Critique a ChatGPT Essay." *Perspectives on History,* September 2023. https://www.historians.org/research-and-publications/perspectives-on-history/september-2023/students-critique-a-chatgpt-essay-a-classroom-experiment.

Mance, Henry. "AI keeps going wrong. What if it can't be fixed?" *Financial Times.* April 5, 2024. https://www.ft.com/content/648228e7-11eb-4e1a-b0d5-e65a638e6135.

Moreu, Gil, and Markus Brauer. "Inclusive Teaching Practices in Post-Secondary Education: What Instructors Can Do to Reduce the Achievement Gaps at U.S. Colleges." *International Journal of Teaching and Learning in Higher Education* 34, no. 1 (2022): 170-182. ERIC.

Newman, Elizabeth. *Biography of a Hacienda: Work and Revolution in Rural Mexico.* Tucson: University of Arizona Press, 2014.

Saucier, Donald A., Noah D. Renken, and Ashley A. Schiffer. "Five Reasons to Stop Giving Exams in Class." *Faculty Focus. F*eb. 18 2022. https://www.facultyfocus.com/articles/educational-assessment/five-reasons-to-stop-giving-exams-in-class/.

Zakrajsek, Todd and Linda Nilson. *Teaching At Its Best,* 5th edition. Hoboken, NJ: Jossey-Bass, 2023.