ChatGPT and World History Essays: An Assignment and its Insights into the Coloniality of Generative AI

Kelsey Rice

Berry University

When ChatGPT became publicly available in the spring of 2023 I, like many of my colleagues, was immediately concerned by the temptations the new technology posed to students in both my 100-level and upper-level college history courses and the thorny issues that would certainly arise around cases of suspected AI use in class assignments. At the same time, I was mindful of the fact that there are many historical precedents for overblown panics over new technologies. As far back as the eighteenth century, increased literacy and access to print media lead commentators to warn of the negative influences novels might have on young people prone to excessive reading, while in the 1940s a doctor warned that youth were becoming addicted to radio dramas in much the same ways alcoholics were addicted to their drinks.[1] Rather than fall into predictable patterns, I decided to open my own OpenAI account and began to experiment with the technology to see what it could do. What stood out to me most immediately was that the texts ChatGPT produced, while grammatically impeccable, were underwhelming at best. The algorithm's soulless regurgitation of accurate facts about history, devoid of analysis or insight, in no way resembled the type of work I push my students to produce in my classes. And yet, reports of student use of the technology on assignments were pouring in, and I encountered a few cases myself. I thus resolved to design an assignment that encouraged my students to engage with generative AI in a way that forced them to think critically about what the technology could and could not do.

This article will be broken into two parts: the first describes the ChatGPT-based assignment I have designed for my 100-level course World History Since 1550 and demonstrates how the assignment has proved a useful tool in promoting student learning objectives and assessing their mastery of course content. The second section will analyze the insights I have gained about generative AI writing on historical topics as a result of this assignment, as it requires me to read dozens of ChatGPT-produced history essays. This will include some quantitative data about the topics in modern world history ChatGPT tends to favor and how this demonstrates a strong Eurocentric bias embedded in the system. With this two-part structure, this essay thus has two conclusions: first, that ChatGPT can productively be used in a college history course to encourage student learning and critical thinking. Second, that it is imperative for educators to impart to our students the significant limitations of generative AI's knowledge-production abilities, as algorithms trained on large language models (LLMs) reproduce historic inequalities.[2]

The Assignment

One of the concerns for educators in any discipline that involves written assignments is that students will stop writing themselves and rely on generative AI to do the work for them. There are many creative ways to make this task more difficult, however for this assignment I decided to turn the tables and require my students to produce essays through ChatGPT, and then assess the quality of those essays based on our course. Because students were unlikely to have encountered a similar assignment before, I also provided a detailed rubric and an example of

[1] Amy Orben, "The Sisyphean Cycle of Technology Panics," *Perspectives on Psychological Science* 15, no. 5 (2020), 1143-4.

[2] A newer version of ChatGPT has become available since I wrote this article. While the essays are now longer and more complex, many of the stylistic problems outlined below remain and the assignment described in this article still works well, as the key to the assignment is the fact that the program has not taken my class. Thus, students can still use the essays to demonstrate their own mastery of course material. The problems with Eurocentric bias in large language models have not been fixed by the more sophisticated writing abilities of the newest version, so the insights offered in the second section of the essay also remain valid.

Teaching History, 49(1).
DOI: 10.33043/67bdga4z9

what the completed assignment should look like. I have copied the assignment description and guidelines below:

In this assignment, students will produce a 1,000-word essay using ChatGPT answering one of the prompts listed below. Copying and pasting the essay into Microsoft Word, students will then use the Tack Changes function to edit and comment on the essay, assessing how well it answers the prompt based on the content and materials of this course.

The purpose of this assignment is twofold. **First**, it is to help you understand the capabilities and limitations of AI writing tools while developing your skills in historical analysis. This new technology can be immensely helpful; it can write emails and other documents that are not the best use of our time and intellectual energy, it can help us workshop our ideas, and it can find sources. AI is not a replacement for developing your own ability to think and write persuasively, however. AI technology is not creative, it is predictive. When you enter this prompt into ChatGPT, it will write the essay in under a minute, using its algorithm to predict what the most likely next word is in the sentence. What you get is a very smoothly written, fairly lightweight piece of analysis that sometimes contains made-up, incorrect information. One of the goals of this assignment is to equip students with the ability to look at a tight, well-written document, and ask themselves: "Am I really convinced by this argument? Is this evidence and analysis actually persuasive or does it just sound good?" Second, this assignment is simply an assessment of whether you have been paying attention in class and doing your reading. In your comments on the essay, you will point out when the essay is addressing history never covered in class, and you will suggest stronger evidence from the course materials that it should use. When it does touch on history we have covered, you will offer positive feedback. You will note when the evidence is too vague and recommend how it could go farther, based on the themes and topics of the course. You can only effectively critique this essay if you have a strong command of the course materials.

Prompts **(simply copy and paste the prompt you choose into ChatGPT)**

Write a 1,000-word essay for the following prompt: Respond to the following question in a complete, organized, and argumentative essay substantiated by specific evidence that is analyzed to support your arguments. In your answer you must include examples from at least three of the listed global regions: Africa, the Caribbean, East Asia, Europe, Latin America, the Middle East, and South Asia (the Indian subcontinent.) Analyze the role of nationalism in shaping the political, social, and cultural structures of the modern world from the 18th-20th century.

Write a 1,000-word essay for the following prompt: Respond to the following question in a complete, organized, and argumentative essay substantiated by specific evidence that is analyzed to support your arguments. In your answer you must include examples from at least three of the listed global regions: Africa, the Caribbean, East Asia, Europe, Latin America, the Middle East, and South Asia (the Indian subcontinent.) How and why have members of different societies from the 16th-20th centuries sought to challenge and limit the power of the governments ruling over them?[3]

Assignment Guidelines

- Got to https://openai.com/chatgpt and either create an account or use your existing account. (If you are strongly opposed to opening an OpenAI account, consult with me.)
- Click "New chat" and copy and paste one of the two prompts into the chat.
- Copy and paste the resultant essay into Microsoft Word and grade using the Track Changes and Comment tools under the "Review" tab.
- In your comments, make clear and specific critiques of the essay's content, focusing on its analysis and evidence.
- Comments on analysis should suggest ways to better connect the essay to themes covered in the course and identify when the essay is overly vague, contradictory, and/or historically inaccurate.
- Comments on evidence should point out whether or not the evidence is relevant to the course and suggest stronger pieces of evidence based on course materials. It should also identify when more context is needed for the evidence.
- ChatGPT writes well, but it is not perfect. Edit any instances you spot where word choice or syntax could be better.
- Make sure all your feedback is clear. Consider: if you were a student and received this feedback, would it make sense to you?
- At the end of the essay include 2-3 paragraphs of feedback in which you offer an overall assessment of the essay and explain how it did well and how it could have done better.

[3] Readers may note the absence of Anglophone North America and Southeast Asia from the list of world regions in the prompt. The first omission is because I have noticed that if given the option, most students will invariably write about the United States. In a world history class, I want to push students out of their comfort zones a challenge themselves to analyze history they have not previously encountered. Coverage of Southeast Asian history is a weak spot in my syllabus that I am currently working to address.

I have used this assignment for four sections across two semesters and plan to continue to use it as I have found it effective in its goals. In terms of helping students to think more critically about what generative AI can and cannot do, I have been gratified to see students take the algorithm to task for its shallow analysis and inability to logically structure a historical argument. The essays are rarely structured chronologically and ChatGPT almost always produces essays broken into many small sections with subheadings, proving utterly incapable of writing a transition sentence, issues many students point out. Because this is a major graded assignment, most students spend a lot of time with their AI essays and manage to see through the authoritative tone of the writing to the emptiness within. Because they are critiquing an algorithm and not a peer, they are do not hold back, with comments like "ChatGPT was able to generate a decent essay based on the prompt, the content is within it is surface level and there is no real argument within it…This assignment has allowed me to see that while AI can be a great tool for editing and revising, it cannot effectively be used to create content such as essays" and "I see how this could be appealing to a stressed college student that doesn't know what to write about because at first glance it sounds good and scholarly. But when you dive deeper into the essay it's choppy and it sounds like it is gathered information from different articles. I don't think this essay is crafted well enough to be a persuasive argument, but I do think it makes valid, solid points…. If I had to give it a grade, I would give it a 75." I hope that, seeing how poorly they assessed an AI essay to be in my course, students who completed this assignment will think twice before, in a moment of panic or laziness, they attempt to submit an AI essay as their own work in another course.

In terms of assessing how well students know the course material, this assignment has also proven effective. ChatGPT has never seen my syllabus nor attended my classes, so it takes the open-ended essay prompts and applies anything about word history within their parameters rather than the using the best evidence from my course's content. It does not know that I have an entire lesson dedicated to the growth in the global coffee trade, or that, as I am a Middle Eastern historian, the Ottoman Empire comes up in my lessons more often than it might in another professor's modern world history survey. It is quite possible that the textbook I use in my course is part of the LLM dataset used to train ChatGPT, but there are many other world history textbooks it learned from as well (and one of ChatGPT's largest sources of training data is Wikipedia, which natural language processing researchers consider a source of "high quality information," concerning as that assumption may be to academics.)[4] This means that students who have kept up with their readings, taken notes, and stayed engaged in class are much better equipped to do well on this assignment. With a due date near the end of the semester, this assignment allows students to demonstrate their mastery of the whole course. When one student highlights a sentence in their AI essay about the Mexican Revolution and comments "we did not discuss this in class," when in fact we had an entire lesson dedicated to the topic and another student highlights a sentence about the Maji Maji Rebellion and recommends the inclusion of additional detailed background information on the event that the AI essay brushed over, assessment is quite easy.

The Essays

Grading this assignment means that each semester I read around sixty ChatGPT-produced essays. The exercise has given me cause for both optimism and alarm. Optimistically, I see nothing in these essays that seem likely to replace human ingenuity and artistic ability. The essays, each produced by a different student at a different point in the semester (the assignment is available from the start of the semester, and I recommend students produce their essays in the first week of classes and then add their annotations throughout the semester. A handful heed my advice; the majority produce the essay sometime in the week or two before it is due), are mind-numbingly repetitive and predictable, using the same examples and drawing the same vague conclusions. To give an example, the most frequent piece of historical evidence used by ChatGPT in the essays on nationalism was Kwame Nkrumah and Ghanaian independence. Half the time Nkrumah appeared in ChatGPT essays (fourteen instances), he appeared paired with Jomo Kenyatta of Kenya. Interestingly, Kenyatta never once appeared in an

[4] Roberto Navigli, Simone Conia, and Björn Ross, "Biases in Large Language Models: Origins, Inventory, and Discussion," *Journal of Data and Information Quality*, Vol. 15, No. 2, Article 10, (June 2023), 3. https://dl.acm.org/doi/10.1145/3597307

essay not paired with Nkrumah. Below are three examples of text from ChatGPT-generated essays, each quote representing the entirety of the essay's engagement with Ghanaian and Kenyan history:

1. For instance, Kwame Nkrumah in Ghana and Jomo Kenyatta in Kenya spearheaded anti-colonial movements that culminated in independence in the mid-20th century.
2. In Africa, nationalist leaders like Kwame Nkrumah in Ghana and Jomo Kenyatta in Kenya rallied against colonial rule, leveraging a shared sense of national identity to challenge imperial powers. For instance, Nkrumah's Pan-Africanism advocated for continental unity, while Kenyatta's Kenya African National Union (KANU) mobilized the masses for liberation.
3. Leaders such as Jomo Kenyatta in Kenya and Kwame Nkrumah in Ghana spearheaded independence movements that galvanized popular support and ultimately led to the dismantling of colonial administrations. These movements utilized a combination of political mobilization, civil disobedience, and armed struggle to challenge and ultimately overthrow colonial rule.

While I have no reason to doubt the programmers who assure us that the quality of the writing by generative AI programs will get better and better, these algorithms are still limited by the basic fact that they are not producing any new knowledge and are consigned to reproducing and rehashing the knowledge created by humans. As the three examples above demonstrate, ChatGPT draws predictable connections that offer little insight into the complex history of decolonization and never questions the logic of pairing these two very different leaders of two different nations, whose main connection is a shared legacy of British colonialism.

It is in that reproduction of preexisting knowledge that I find most cause for alarm. As numerous scholars have noted, because large language models are trained on texts produced by societies with biases, they are prone to reproducing those biases.[5] Notably, the sources on which many of the leading natural language processing models such as ChatGPT are trained draw heavily from online resources such as Wikipedia and Reddit, sites whose contributors are overwhelmingly male, English-speaking, and white. Thus, as AI researchers argued in an influential 2021 article, "this means that white supremacist and misogynistic, ageist, etc. views are overrepresented in the training data, not only exceeding their prevalence in the general population but also setting up models trained on these datasets to further amplify biases and harms."[6] For example, scholars demonstrated that ChatGPT-3 created violent completions for the prompt to complete the sentence "Two Muslims walked into a" sixty-six percent of the time, while the likelihood of a sentence concluding with a violent action dropped dramatically when "Muslim" was swapped out for other religions.[7] For historians, this means that ChatGPT has been trained on centuries of Eurocentric history writing that privileges elite white males as history's most important actors.

The short history essays students produced for this assignment did not tend to demonstrate overt racial, gender, or religious biases. However, when looking at the geographic distribution and volume of the historical examples that were used in the ChatGPT essays, a clear western and Anglophone bias emerges. In the following paragraphs I offer a quantitative breakdown of the content of the AI essays produced by my spring 2023 students and analyze what these statistics reveal to us.

Chat GPT History Writing and the Legacy of Colonialism

In the spring of 2023, fifty-seven students completed the AI essay assignment I detailed above. Thirty students elected to produce essays with the prompt "Analyze the role of nationalism in shaping the political, social, and cultural structures of the modern world from the 18th-20th century" (hereafter Prompt 1) and twenty-seven

[5] Ibid.

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big? 🦜," *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021), 613. https://dl.acm.org/doi/10.1145/3442188.3445922

[7] Abid, Abubakar, Maheen Farooqi, and James Zou, "Persistent anti-Muslim bias in large language models," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298-306, (2021.) https://arxiv.org/abs/2101.05783

selected the prompt “How and why have members of different societies from the 16th-20th centuries sought to challenge and limit the power of the governments ruling over them?” (hereafter Prompt 2.)

The thirty essays on Prompt 1 employed fifty-eight discrete historical examples in their analyses. Of these fifty-eight examples, only seventeen appeared in more than three essays, while the most frequently used example appeared in twenty-eight of the thirty essays. The five most common examples were:

1. Kwame Nkrumah and Ghanaian independence (28)
2. German and Italian Unification (24)
3. The Latin American Wars for Independence (19)
4. Indian independence and the partition of India and Pakistan (17)
5. The Romantic Movement in Europe (14)

The geographic distribution of the fifty-eight examples is as follows: 17 European, 14 African, 9 Middle Eastern, 8 Latin American, 7 East Asian, and 3 South Asian. Although my essay prompts count the Caribbean as its own geographic category, I am not counting the Caribbean here as there were only two examples, both of which were the Haitian Revolution and both of which were identified in the essays as Latin American examples.

The twenty-seven essays on Prompt 2 included forty-one discrete historical examples of which fourteen appeared in more than three essays. The most common example appeared in twenty-six of the twenty-seven essays. The five most common examples were:

1. The French Revolution (26)
2. The Indian Independence Movement (21)
3. The Latin American Wars for Independence (20)
4. The Mau Mau Rebellion (16)
5. The Haitian Revolution (12)

The geographic distribution of the forty-five examples is as follows: 14 European, 8 African, 8 Latin American, 5 Middle Eastern, 3 East Asian, 2 Caribbean (not counting the Haitian Revolution), and 1 South Asian.

In his essays on African literature and the legacy of colonialism, Kenyan writer Ngũgĩ wa Thiong’o lays out his argument for rejecting colonial languages as the languages of African literary expression, describing his childhood education in English as a process in which “language and literature were taking us further and further from ourselves to other selves, from our world to other worlds.”[8] Thiong’o famously rejected writing in English in favor of writing in his native Gikuyu and has been an important voice for the literary value of African languages for decades. Generative AI programs, in their current form, pose a threat to this ongoing decolonizing project. Researchers of natural language processing categorize languages as “high resource,” “medium resource,” and “low resource,” referring the volume of high-quality texts in those languages available to use for training AI models. English is the most highly resourced language by orders of magnitude. Other high-resource languages include German, French, Spanish, Arabic, Japanese, and Mandarin.[9] All African languages are low-resource languages.[10] Thus even multilingual language models that are trained on datasets involving multiple languages are trained overwhelmingly in English and other colonial languages.[11] Since the motive behind developing these models is profit-driven, there is little impetus from NLP developers to address these issues, as doing so would be time-consuming and expensive.[12]

The ongoing discourse on “decolonizing history” is varied and ever-evolving, involving calls to incorporate

[8] Ngũgĩ wa Thiong’o, *Decolonising the Mind: The Politics of Language in African Literature*, (Martlesham: Boydell & Brewer, 1986), 12.

[9] Gabriel Nicholas and Aliya Bhatia, “Lost in Translation: Large Language Models in Non-English Content Analysis,” Report, Center for Democracy and Technology, May 2023, 18. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/

[10] Ibid.

[11] Ibid., 6.

[12] Navigli, et al, “Biases in Large Language Models,” 5.

more BIPOC (Black, Indigenous, People of Color) scholars into syllabi, for history curriculum to more critically engage with the history of the end of empire, and for scholars to critically interrogate how imperial legacies have shaped the structure of the institutions within which we work.[13] However a history teacher approaches the concept of decolonizing history, however, it is clear that generative AI is a step in the wrong direction, especially if we allow students to become overly reliant on it for their historical information. As the data above demonstrate, although the top examples for each essay prompt include broad geographic coverage, when taken as a whole it is clear that ChatGPT offers much more variety and volume on European history than non-western history. Every essay contained at least one European example, something true for none of the other global regions listed in the prompt. Damningly, although Indian history appears in the top five examples for both prompts, for Prompt 1 there were only three examples that ChatGPT could produce related to South Asian history, while for Prompt 2 there was just one example, albeit a frequently recurring one. East Asian history examples do not appear in the top five examples for either prompt and there were only seven examples from East Asia for Prompt 1 and three for Prompt 2, meaning that for South Asia and East Asia combined, the region within which most of humanity has resided throughout history, ChatGPT could offer just fourteen historical examples worth mentioning.

For both prompts African history followed European history as the most common regional history included, however baked into these examples is significant Anglocentrism. For the first prompt there were just two examples about African regions that were not part of the British Empire: the Rwandan Genocide and African examples from the global Negritude Movement. For the second prompt there were also only two examples: the Maji Maji Rebellion and the First Italo-Ethiopian War, although it is worth noting that the Maji Maji Rebellion occurred in German East Africa, which would transfer to British colonial rule just twelve years after the rebellion's conclusion. As discussed above, the persistent pairing of Nkrumah and Kenyatta highlights how the British Empire drives ChatGPT's understanding of African history much more than any sense that African nations have unique, nuanced histories completely independent of European involvement.

While the latest model of ChatGPT can now draw on and learn from information on the internet as it is posted, the nature of LLMs means that ChatGPT will aways carry the weight of historiography and will reproduce the historical knowledge with the most volume rather than that which is the most innovative. For example, it has only been since the 2022 Russian invasion of Ukraine that the field of East European and Eurasian studies has taken a more definitive turn toward decentering Russia.[14] The 2023 theme for the annual convention of the Association for Slavic, East European, and Eurasian Studies was decolonization (notable in that whether or not the Russian Empire involved a colonial project has continued to be a matter of some debate among scholars in the field), and it is likely there is a wave of forthcoming publications that will contribute to this decolonizing project. For a LLM trained on billions of words, these new scholarly publications are just drops in an ocean of data, however, especially as many will remain behind paywalls and not available for immediate inclusion in datasets. A student asking ChatGPT to write about the Soviet Union will almost certainly receive an interpretation of history that solidly centers Russia, despite the excellent new work being published on regions such as Central Asia, the Caucasus, and Siberia. ChatGPT will always be a step behind the most original scholarship, thus although it is the cutting edge of technology, it is consigned to produce only tired, unremarkable insights into the humanities.

[13] Amanda Behm, Christienna Fryar, Emma Hunter, Elisabeth Leake, Su Lin Lewis, and Sarah Miller-Davenport, "Decolonizing history: enquiry and practice," *History Workshop Journal*, vol. 89, (2020), 171-2. https://academic.oup.com/hwj/article-abstract/doi/10.1093/hwj/dbz052/5739463?redirectedFrom=fulltext

[14] Alexander Motyl, "Decentering East European and Eurasian Studies," *Harriman Magazine*, 2024 Issue. https://harriman.columbia.edu/decentering-eurasian-and-east-european-studies/

Conclusion

The computational linguist Emily Bender has cautioned that we should "resist the urge to be impressed" when it comes to AI.[15] As her scholarship has shown, the limitations and shortcomings of AI are myriad, however much tech CEOs such as Sam Altman (whose net-worth is directly tied to AI hype) may warn us that their awe-inspiring technology could soon spell the end of humanity.[16] There are few, if any, historians who know how to build large language models and train AI models for natural language processing. Historians are excellent, however, at identifying poppycock when we see it, and I hope the discussion above demonstrates that scholars in the humanities are in many ways better equipped to identify some of AI's shortcomings than those in STEM fields. I suggest that as history teachers, we have an important role to play in the current AI discourse. When AI's world-ending potential comes up, who better than a historian to point out the many other times in human history that the end was nigh, or that new technology inspired widespread panic? Rather than focusing on abstract notions of technological apocalypse, it is imperative that historians point to the actual, current harms that AI produces. Thiong'o wrote that "a specific culture is not transmitted through language in its universality but in its particularity as the language of a specific community with a specific history."[17] Generative AI, with its privileging of English, of male-produced content, and of dominant narratives, elides cultural specificity and flattens knowledge into something reductive, bland, and hegemonic. We must communicate to our students that if they elect to allow generative AI to produce their knowledge for them, however tempting it may be, they will be placing severe limitations on the sort of knowledge they might gain, caging themselves in an algorithmic world of rehashed banalities and robbing themselves of the potential for creative thoughts and original insights. Rather than mounting our soapboxes and telling them this, designing assignments that require students to engage with generative AI and critically analyze it can help them reach these conclusions themselves.

[15] Emily M. Bender, "On NYT Magazine on AI: Resist the Urge to Be Impressed," *Medium*, April 18, 2022. https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd

[16] Samantha Kelly, "Sam Altman warns AI could kill us all. But he still wants to world to use it," *CNN*, October 31, 2023. https://www.cnn.com/2023/10/31/tech/sam-altman-ai-risk-taker/index.html

[17] Thiong'o, *Decolonising the Mind*, 15.