How Well Can GPT-4 Really Write a College Essay? Combining Text Prompt Engineering And Empirical Metrics

Abigail Foster

IPHS 484 Senior Seminar (Spring 2023) Prof Elkins and Chun, Kenyon College

Introduction

When ChatGPT became publicly available in November 2022, AI experts and non-experts alike began speculating about the implications for classrooms around the globe. This project seeks to understand the impact that text generative AI models will have on history writing at colleges like Kenyon, using text prompt engineering combined with an empirical and methodical approach. First, I established the metrics by which I would evaluate the performance of these models. Secondly, I performed a total of 27 trials to provide data which I then evaluated based on my metrics. Finally, I considered my results and placed them within the larger context of the Kenyon community.

For the purposes of this project, I only looked at GPT 3.5 and GPT 4. When I began this project in February, only GPT 3.5 was publicly available, but when GPT 4 launched in March it became clear that I would need to incorporate it into my metrics. Therefore, part of my project also became about comparing the strengths and weaknesses of the two models. Another element of my project was prompt engineering. As I continued to collect data, I discovered how my prompts changed the results I got, and I was able to draw concrete conclusions about the elements a prompt must include to achieve superior results.

Part of my project also involved meeting with members of the Kenyon community—specifically those whose jobs would be most impacted by this technology—and discussing their opinions about and experiences with it so far. Throughout that process I was also able to test out theories or ideas from these stakeholders, and I incorporated many of those suggestions into my final methodology.

Though my project was specific to historical writing, it has implications across disciplines. The most important is the creation of concrete metrics through which to evaluate the capability of a model like GPT 3.5 or 4. Since the release of ChatGPT in November, many people have experimented with the capabilities of this model; fewer have turned those experiments into quantitative data and systematically recorded and evaluated them in the way that this project has. I think the model I present is a valuable addition to discussions about the capacity of text generative models and how they may impact higher education.

Methodology

This project was based off of my background as a History major at Kenyon and as an employee of the Kenyon College Writing Center. In my time at the Writing Center, I have been the course liaison for several history classes, where I've been exposed to a substantial amount of undergraduate history writing. Using these experiences, as well as the feedback I have gotten from history professors here about assignments I have turned in, I created a rubric to function as a comprehensive guide for the qualities that good undergraduate history writing must posses. I came up with five categories: clarity of organization, quality of historical argument, use of quotations and evidence, complexity of grammatical structures and vocabulary, and logic and coherence. In my experience most students who come to the Writing Center for help with their historical writing are looking for help in one of these areas, and I believe that a good historical essay must excel in all of them.

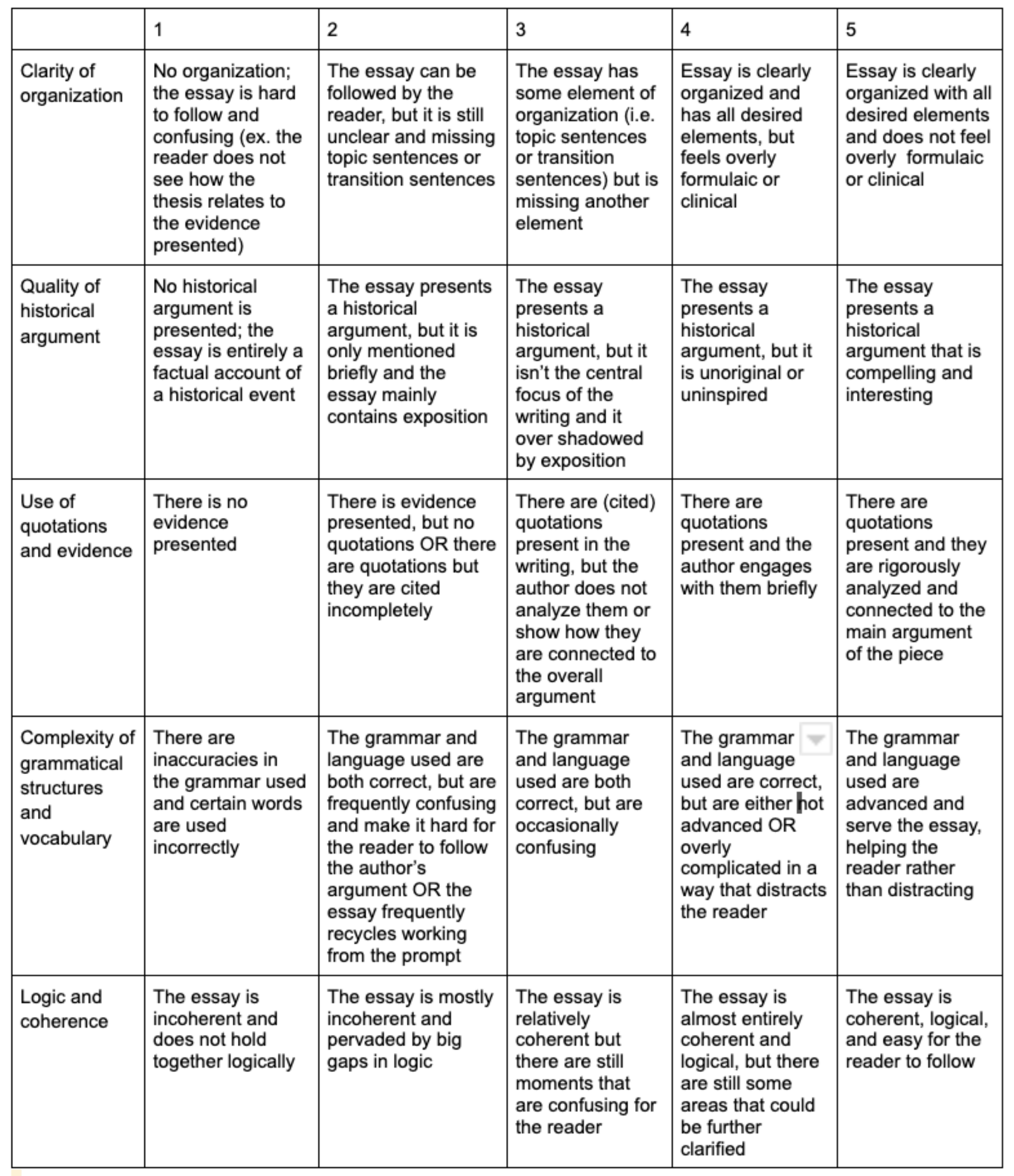
Throughout the creation of this rubric, I made sure to contextualize each part within my own experience reading the writing of other Kenyon students. Essays that score mostly 1s are very poor; I would be surprised to see this type of essay at Kenyon, and if it were to be submitted for a class it would likely receive a very poor grade and the teacher would probably want to meet with the student. An essay that scores mostly 2s needs some serious revision and if it were to be submitted in a college level class would likely get a C range grade or lower. However, this type of essay is realistic in terms of what is often brought in to the Writing Center as a first draft. An essay that scores mostly 3s is an essay that could benefit from revision, but could conceivably be submitted to a college level class and would likely receive around a B or B-. An essay that scores mostly 4s is quite good; it has all the elements that a good essay needs. I would imagine that essays of this level are frequently given A-range grades in college classes. Finally, an essay that scores mostly 5s goes above and beyond, and excels in all five categories. If the AI is presenting writing with all of these elements I would argue that it is as good as or better than a substantial amount of college history writing.

Alongside the 5 categories outlined in the rubric, all of which can be graded on a scale of 1 (poor) through 5 (excellent), there are also three other characteristics I was looking for: factual correctness, usage of the past tense, and inclusion of an element of surprise. These qualities were either present or not present rather than being graded on a scale.

I also evaluated the generated responses I produced using two tools that claim to identify AI-produced text: GPTZero and OpenAI's own AI Text Classifier. GPTZero's two metrics are perplexity, which measures the perplexity of the sentences, and burstiness, which measures the variation in perplexity of the sentences. There are already many services claiming to detect AI-written content, and there will certainly be more emerging. The most prominent ones at the time of writing are GPTZero, OpenAI's text classifier, Originality AI, and TurnItIn's AI detection feature. For this project, I was only able to access the first two, as Originality AI has no free version and TurnItIn requires you to be logged in as the instructor of a course to see the 'AI' score of a piece of writing. However, I will be exploring TurnItIn's AI detection in my final project for IPHS 300, AI for the Humanities.

Acknowledgements

This project would not have been possible without the support of Professor Elkins and Professor Chun, or the feedback from my fellow IPHS concentrations. I would also like to thank Professor Anna Scanlon, Dean Thomas Hawks, and Professor Wendy Singer for their willingness to speak with me about the impact of this technology at Kenyon.



	1	2	3	4	5
Clarity of organization	No organization; the essay is hard to follow and confusing (ex. the reader does not see how the thesis relates to the evidence presented)	The essay can be followed by the reader, but it is still unclear and missing topic sentences or transition sentences	The essay has some element of organization (i.e. topic sentences or transition sentences) but is missing another element	Essay is clearly organized and has all desired elements, but feels overly formulaic or clinical	Essay is clearly organized with all desired elements and does not feel overly formulaic or clinical
Quality of historical argument	No historical argument is presented; the essay is entirely a factual account of a historical event	The essay presents a historical argument, but it is only mentioned briefly and the essay mainly contains exposition	The essay presents a historical argument, but it isn't the central focus of the writing and it over shadowed by exposition	The essay presents a historical argument, but it is unoriginal or uninspired	The essay presents a historical argument that is compelling and interesting
Use of quotations and evidence	There is no evidence presented	There is evidence presented, but no quotations OR there are quotations but they are cited incompletely	There are (cited) quotations present in the writing, but the author does not analyze them or show how they are connected to the overall argument	There are quotations present and the author engages with them briefly	There are quotations present and they are rigorously analyzed and connected to the main argument of the piece
Complexity of grammatical structures and vocabulary	There are inaccuracies in the grammar used and certain words are used incorrectly	The grammar and language used are both correct, but are frequently confusing and make it hard for the reader to follow the author's argument OR the essay frequently recycles working from the prompt	The grammar and language used are both correct, but are occasionally confusing	The grammar and language used are correct, but are either not advanced OR overly complicated in a way that distracts the reader	The grammar and language used are advanced and serve the essay, helping the reader rather than distracting
Logic and coherence	The essay is incoherent and does not hold together logically	The essay is mostly incoherent and pervaded by big gaps in logic	The essay is relatively coherent but there are still moments that are confusing for the reader	The essay is almost entirely coherent and logical, but there are still some areas that could be further clarified	The essay is coherent, logical, and easy for the reader to follow

Results

My results for this project came in several forms. Firstly, using the metrics I created, I was able to compare the performance of GPT 3.5 and GPT 4. This, however, is not necessarily novel; since GPT 4 was released, many comparisons between the two models have been made. The more interesting results came when I differentiated between the 'first pass' responses, responses where I used the 'regenerate response' feature after the first pass, and responses I got after editing the first pass response.

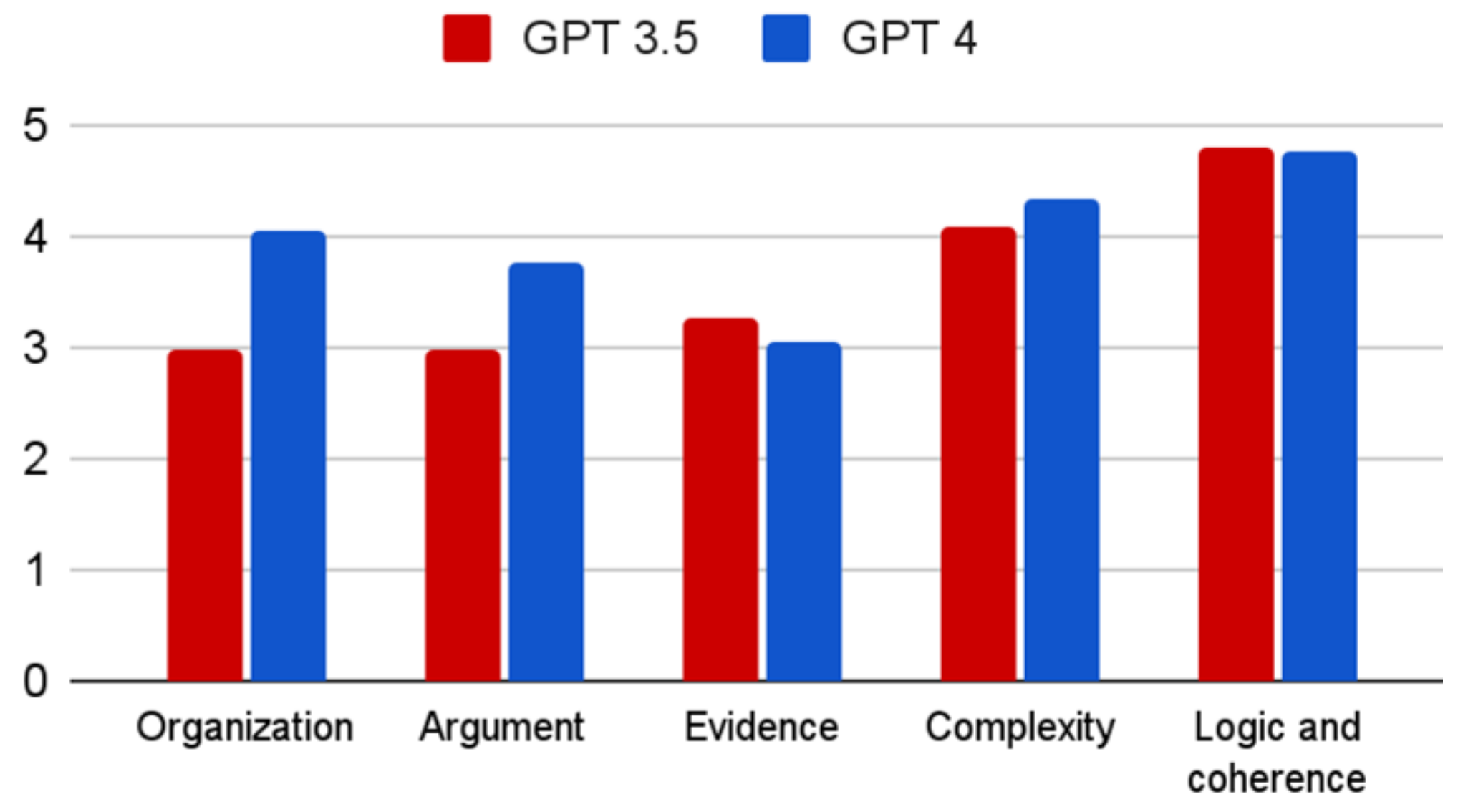
Overall, I performed 27 trials; 14 of the trials were for GPT 3.5, while 13 of them were GPT 4. GPT 4 performed, on average, 1.07 points better than 3.5 in 'clarity of organization,' 0.79 points better in 'quality of historical argument,' in 0.37 'complexity of grammatical structures and vocabulary,' but performed 0.20 points worse in 'use of quotations and evidence' and 0.03 points worse in 'logic coherence.' In ChatZero's 'Perplexity' metric, it scored on average 20 points higher, and on the 'Burstiness' metric it scored on average 72.5 points better. I then divided the data into four categories: the overall average, the average of the 'first pass' results, the average of the regenerated results, and the average of the 'edited' results (or results where I had provided feedback beyond the initial prompt). I did this for both GPT 3.5 and GPT 4. In GPT 3.5, the average of the 'edited' results consistently outperformed the 'first pass results,' while the 'regeneration' response was, on average, worse than the 'first pass.' On the other hand, in GPT 4, the 'regeneration' responses outperformed the 'first pass' responses except in 'logic and coherence.' on the other hand, the 'edited' results were the same as the 'first pass' responses, slightly worse in 'quality of historical argument,' but better in 'use of quotations and evidence.' This is likely because most of the edits I made after the first pass responses were to ask for more primary source quotes; frequently the first pass responses wouldn't have any, although they were specified in the prompt. However, if I had had more time I would have liked to collect more data to see if this trend continued to be true.

In terms of the three categories not included in the metric—factual correctness, use of past tense, and element of surprise—both GPT 3.5 and GPT 4 consistently failed on the first two, though GPT 4 performed better than 3.5 in factual correctness. However, I was able to adjust my prompts so that the essays produced did have what I would consider 'an element of surprise,' something that I think any compelling history essay needs.

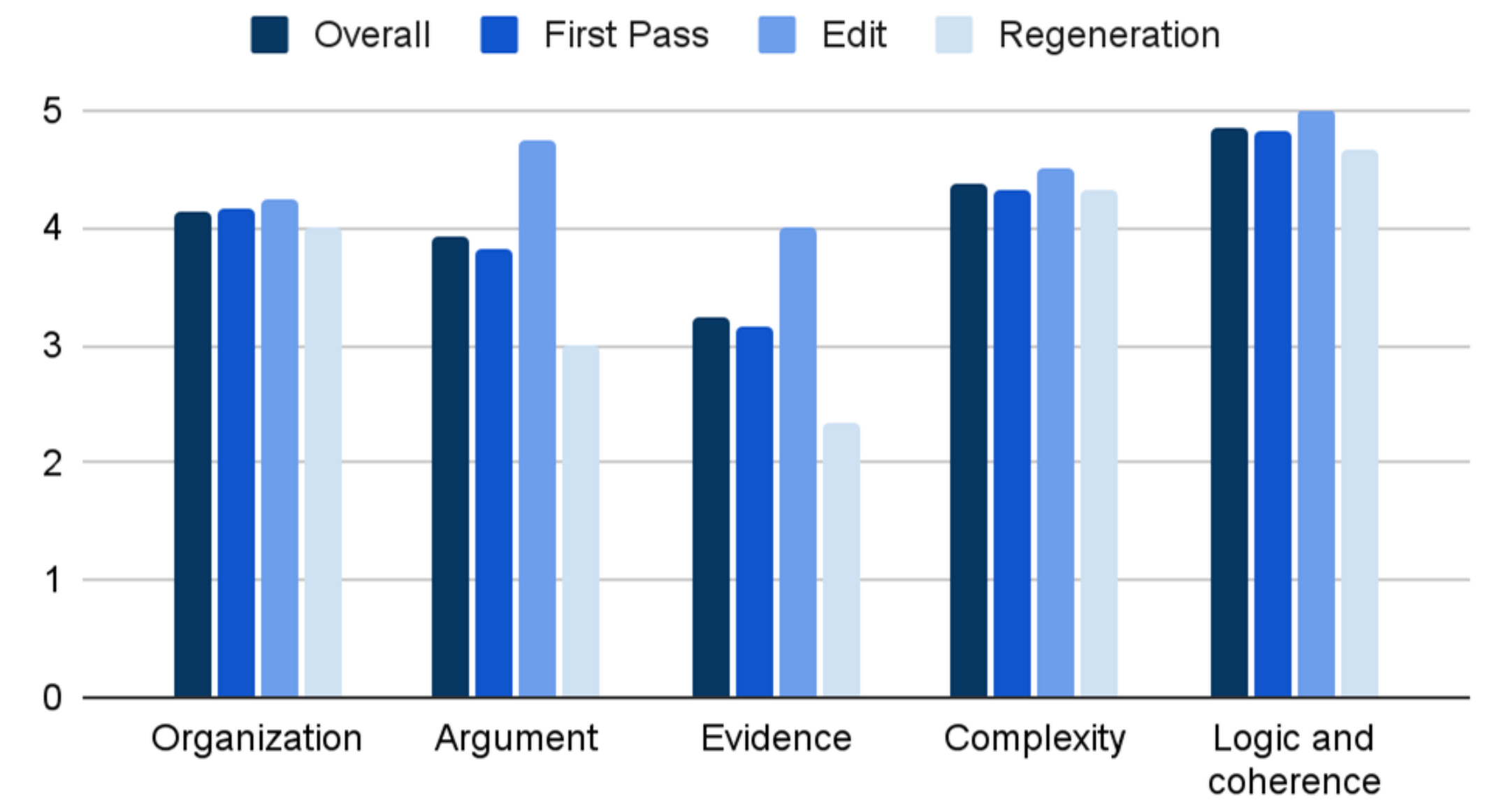
Regardless of the model, the writing that did better (was deemed more human) on the AI classifiers was either a regeneration or based on an edit, not a first pass. For GPT 3.5, GPTZero said most trials were "likely to be written entirely by AI," while two trials (one a regeneration and one an edit) were deemed "may include parts written by AI." OpenAI's text classifier considered the trials to be "likely AI-generated," while three trials (all either edits or regenerations) were "unclear if it is AI-generated." In GPT 4, 6 trials were deemed by GPTZero "likely to be written entirely by AI," 2 were "mostly likely human written," while 5 were "likely to be written entirely by a human." OpenAI's text classifier said that 2 of the trials were "possibly AI-generated," 6 were "unclear if it is AI-generated," 2 were "unlikely AI-generated," and 3 were "very unlikely AI-generated." The samples that scored as the most human on both models were responses that had come from multiple rounds of regenerations and edits.

The biggest takeaway from these results are that the writing samples that will produce the best quality historical writing and read as the most human will be produced by GPT 4 rather than 3.5 and will have gone through multiple rounds of not only feedback/edits, but also multiple regenerations. Overall, the text generated by both models performed quite well on the metrics I created; 3.5 scored an average of 18/25 points, while 4 averaged 20/25 points. As I have demonstrated, there was significant variation depending on whether or not the result had been edited or regenerated. However, the both models averaged what I would consider very strong results, results that according to my metrics would likely score a B+ or A- in a college level history class. With additional text prompt engineering, the results were even better.

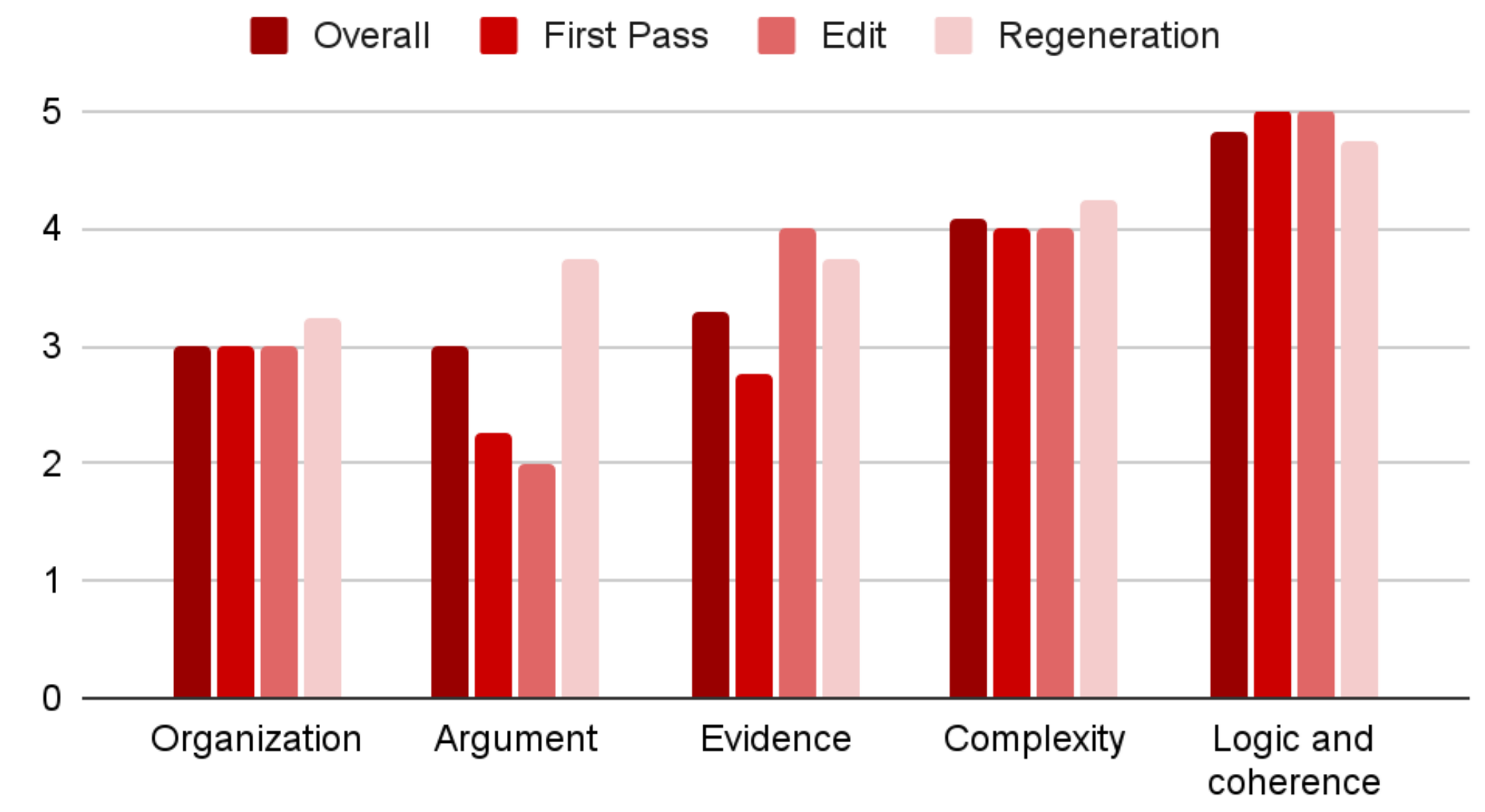
Comparing GPT 3.5 and GPT 4 Performance on Metrics for Historical Analysis



GPT 4 Performance on Metrics for Historical Analysis



GPT 3.5 Performance on Metrics for Historical Analysis



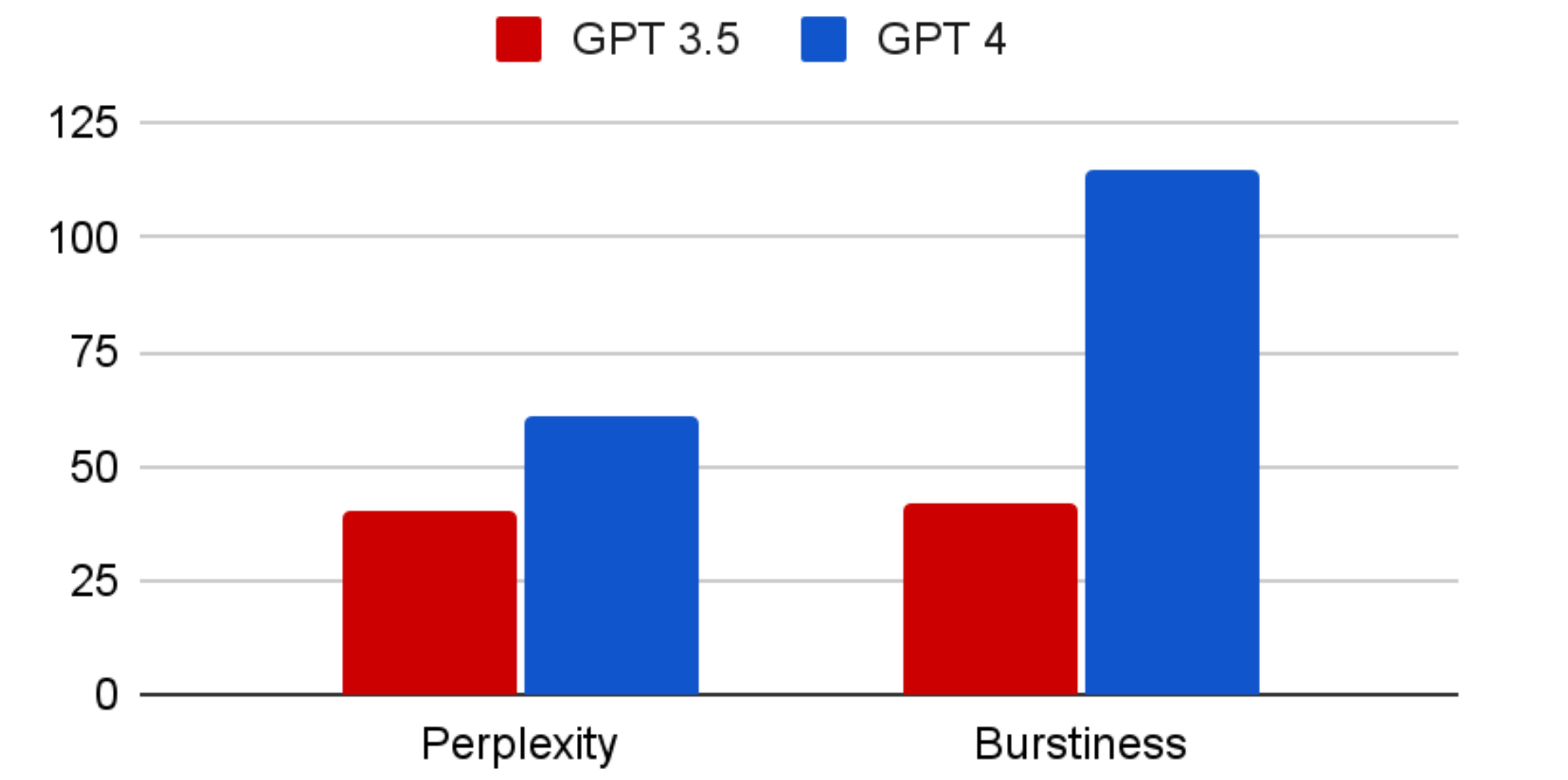
Text Prompt Engineering

As I worked on collecting the data for this project, I became familiar with the emerging field of prompt engineering. I refined my prompts so that I could more quickly get the results that would score highly on my rubric. The first thing I learned was that I got better results when I asked for what I wanted; instead of hoping that the writing I got should include a element of surprise, I should be explicit and ask for that in the prompt. My first task, therefore, was to figure out how to incorporate my rubric into the prompt itself. After trying a variety of methods, I settled on outlining the five categories and saying that I was looking for an essay that would score highly on all of them. The only downside to this was that the model would sometimes interpret my desire for clear organization by giving each paragraph its own title or heading, which I did not want.

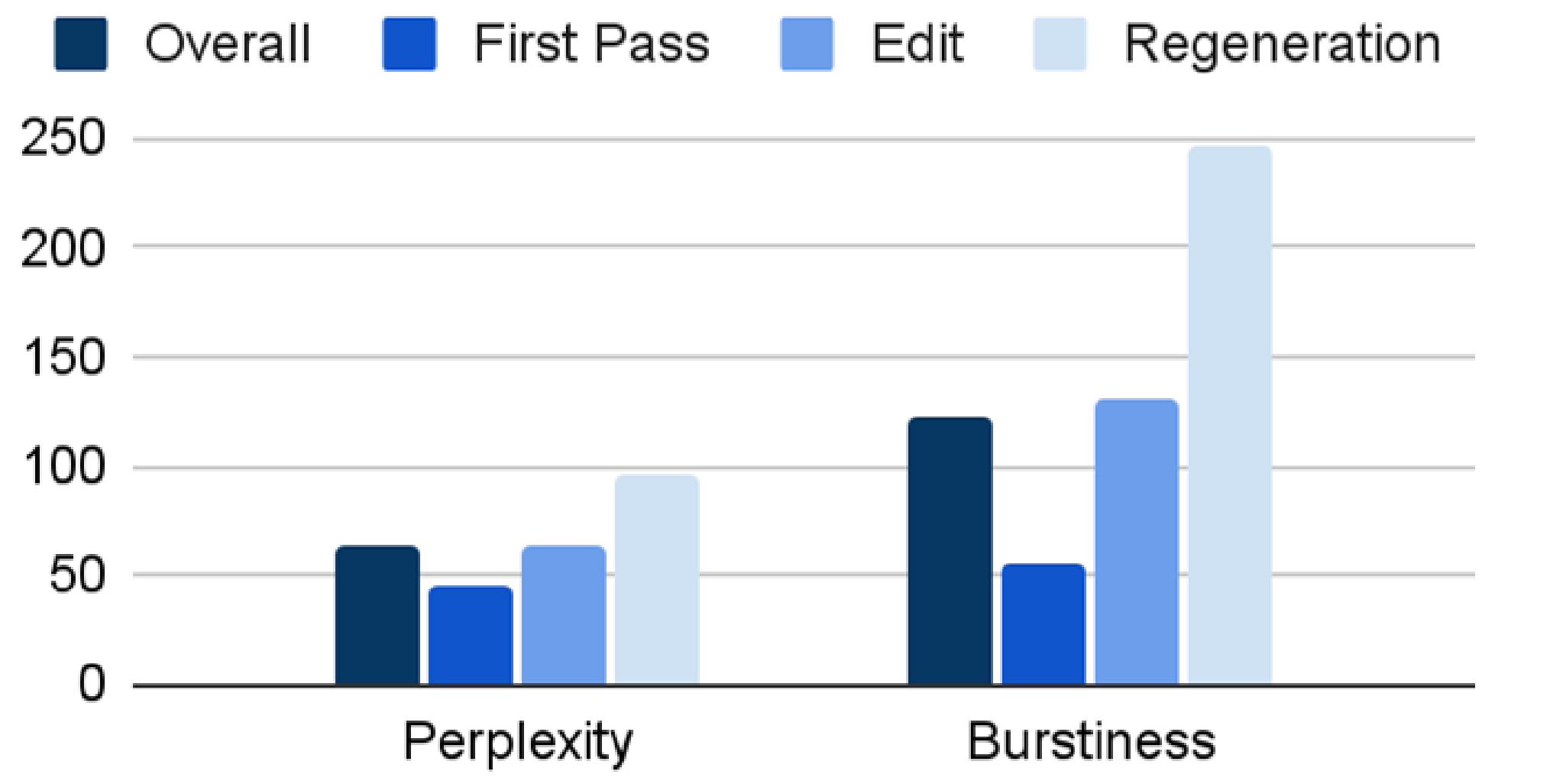
The next task I faced was to decide how open-ended I wanted the prompt to be. I began by giving very specific prompts which specified a primary source and a stance I wanted the writing to take. However, I discovered that the model produced the most interesting and high-scoring writing when I left the prompts more open ended, specifying only a historical event and asking it to use sources to construct an argument. This type of prompt led to some fascinating responses, where the model created connections I wouldn't have thought of. Finally, based on a comment from Professor Wendy Singer, who noted that the best historians don't need a prompt to write a good historical essay, I simply inputted the syllabus for a hypothetical history class and asked the model to write an essay based off of those readings. However, I found that this type of prompt did not yield as high-scoring of a result as when I asked a more clear but still open ended question.

As of now, I believe that the text prompt engineering required to get the publicly available GPT 3.5 to produce really strong results requires a set of skills that are currently unfamiliar to most students. However, I also think that some students have likely learned many of these skills through their academic pursuits in other disciplines. Students like myself who have worked at writing centers or as peer tutors will already be equipped with an understanding of what qualities good writing should possess and what to change about a text in order to improve it. These skills are the ones that will become the most valuable in the coming years, as prompt engineering becomes a more sought-after ability.

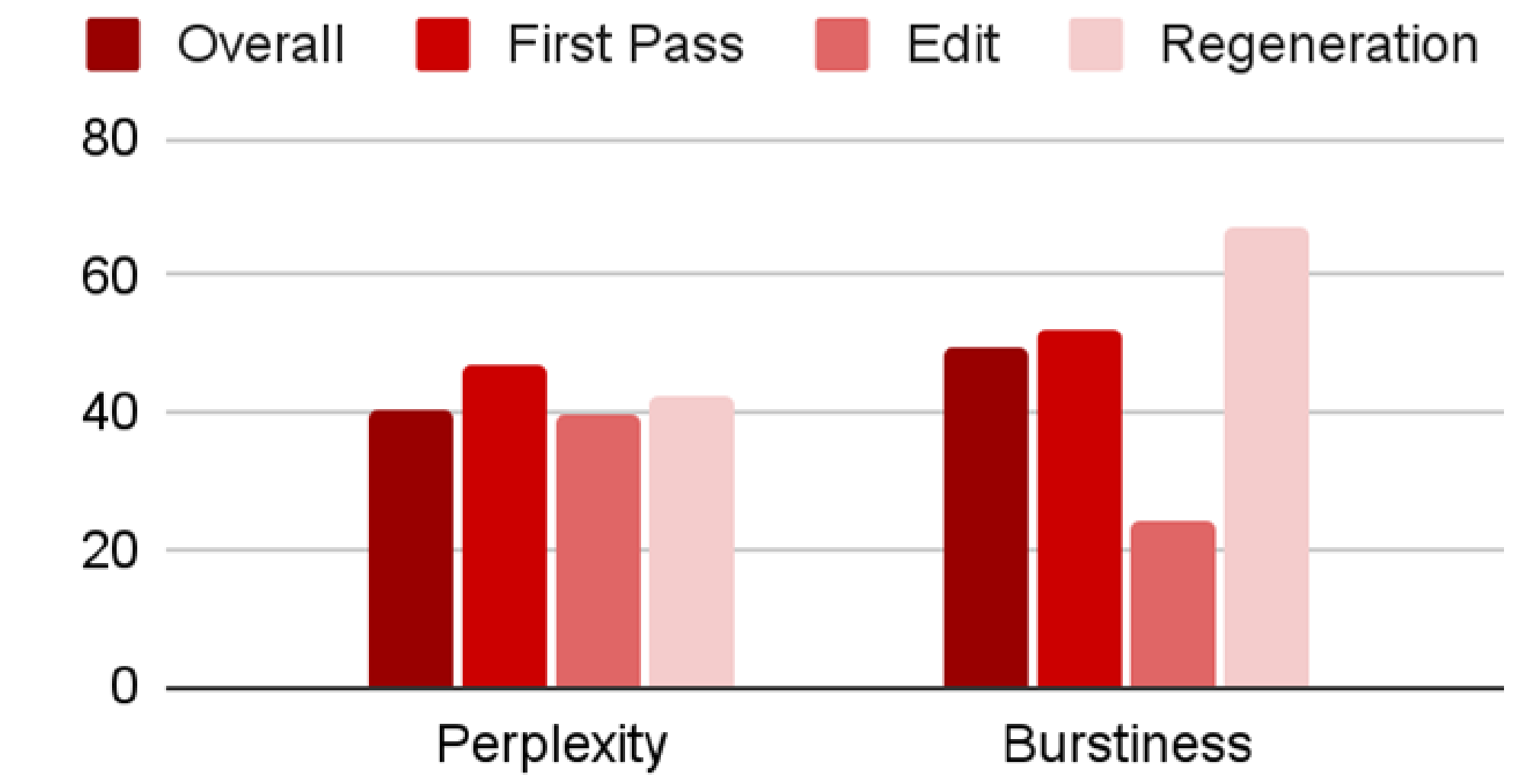
Comparing GPT 3.5 and GPT 4 Performance on GPTZero Metrics



GPT 4 Performance on GPTZero Metrics



GPT 3.5 on GPTZero Metrics



Impacts at Kenyon: Is the Essay Doomed?

Since the public launch of ChatGPT, many people have started asking questions about what this technology means for the future of higher education or jobs related to writing. Part of my goal for this project was to better understand what text generative models would mean for a place like Kenyon, which prides itself so highly on the quality of the writing that is produced here. After working with this technology all semester and speaking with members of the community about it, two things are clear to me. Firstly, this technology is already widely used among the student body. Secondly, Kenyon will need to find a way to shift the academic integrity policy to account for its use; however, how exactly this should be done remains generally unclear.

Of the people in the Kenyon community I spoke to about this project, three were what I would consider 'stakeholders': people whose jobs will likely be most impacted by this technology. Those people were Professor Anna Scanlon, director of the Kenyon College Writing Center; Dean Thoman Hawks, Dean of Academic Advising and former English teacher; and Professor Wendy Singer, Associate Provost and History Professor. All three of these people were excited to speak with me about this technology, and had already been engaged in dialogue about it with their coworkers, friends, or family. They all considered ways that this technology could be integrated into a Kenyon education. In my conversation with Dean Hawks, he made what I thought was a very compelling point: we must consider how a Kenyon education prepares students for the world they will live in after their graduation. If AI is going to be widely used in the professional world to produce writing, shouldn't that be part of your Kenyon education? In the coming years, this technology will continue to improve, and will continue to challenge the ways we think about writing. I encourage everyone in the Kenyon community to see it not as something that should be feared and distrusted, but rather as an exciting new tool that can enhance our lives and educational experiences. Several of the people I spoke to predicted that many schools would move back to having students hand write essays in class, although this system poses issues for accessibility. Others suggested encouraging students to use AI, but emphasising that they must cite its usage. Overall, I predict that in the coming years text generative technology will become integrated into traditional academic writing assignments. I also think that people will begin to think more critically about the role of these assignments in college, and that some places—though not places like Kenyon—will decide that essays are not an effective way to demonstrate learning for a particular discipline. However, I personally think there is immense value in the process of constructing and defending a historical argument, and I do not think that those skills will ever be fully automated at a school like Kenyon.