Can GPT-4 Fool TurnItIn? Testing the Limits of AI Detection with Prompt Engineering

Abigail Foster

IPHS 300 AI for Humanity (Spring 2023) Prof Elkins and Chun, Kenyon College

Introduction

The purpose of this project was to test the limits of Turnitin's AI detection feature. My goal was to see if I could get text that was 100% AI written to score as human written by refining the prompts that I gave GPT-4. Therefore, I did not use any text that was partially human written, although that text would likely have scored as more human written; all of the text I used was entirely generated by AI.

Turnitin is a platform widely used by educators to detect plagiarism in students work. Students will submit their assignments through a Turnitin portal and will receive a score indicating how much of the content of their writing can be found in other sources on the internet. Recently, Turnitin also launched an "AI detection feature" that also gives a score on how much of the submitted writing was likely generated by AI. However, only administrators can see this score; unlike the 'similarity' score that indicates traditional plagiarism, the 'AI' score is not visible to students.

Since the launch of Chat-GPT in November 2022, educators, AI experts, and students alike have all wondered what this software and others like it will mean for the future of student essays. AI generated text evades most traditional plagiarism detectors, as the writing is novel and cannot be found in another existing source. Therefore, a new market is emerging for models that can reliably detect whether or not text was generated by an AI. As Turnitin is already widely used in schools and universities and their new AI detections software has been integrated into their existing platform, this model is likely the detection model that will be used most widely by educators, at least for the moment.

This project was inspired by a mini project from IPHS 300: AI for the Humanities with Professor Elkins and Professor Chun. In this project, students were asked to use prompt engineering to try and create AI generated content that would earn a low percentage on Turnitin's detection feature (meaning it would score as mainly human written). Out of the 26 students in the class—students who had been studying AI all semester and had been working with prompt engineering for multiple weeks—only three were able to produce writing that scored less than 100% AI generated on Turnitin, and out of those three the writing that scored the lowest scored as a little over 30% AI generated. Clearly, Turnitin's detection software is rigorous; on their website, they note that they have 98% confidence in the determination their detection model makes.

With this project, I hope to contribute to the current discourse about AI generated writing and models that can detect it. In recent months, many scholars have looked into the question of what differentiates human generated writing from that generated by AI; see "Detection of AI-generated Essays in Writing Assessments" by Yan et al. from March of this year. By showing how I was able to use prompt generation and GPT-4 to create 100% AI generated writing that scored as almost entirely human written, I hope to provide more information to this important discussion and encourage others to similarly experiment with the capacities of such software.

Ethical Implications

At first glance, this project has serious ethical implications. By experimenting with the limitations of Turnitin's AI detection model—a model which I acknowledge is likely already widely used in schools and will likely continue to be—and by publishing the results of this research, it could appear that my research will allow other students to 'fool' the software and submit AI generated text without being penalized. However, there are two important features of this research that mean it is impossible for it to be used in such a morally-questionable way.

Firstly, as noted in the introduction, the 'AI' score on Turnitin is not visible to students. In my experiments, I frequently used these scores to provide feedback to GPT-4 and refine its writing. I was only able to do this because I worked with Professors Elkins and Chun to gain special permissions on a new Moodle page that was created specifically for this project. This new Moodle page was created with the assistance of Joe Murphy, the Director of the Center for Innovative Pedagogy at Kenyon College, who kindly agreed to help with this project. With this Moodle page, which was full of test bot accounts, I was able to both submit writing and, acting as the 'Professor,' immediately see the AI score. This is something no student would be able to do without institutional assistance. Therefore, my methodology is not reproducible by the average student looking to shirk their academic work.

Secondly, and most importantly, the AI generated writing that received a low 'AI' score on Turnitin was not what I would consider good academic writing. None of these samples would do well in a class were they to be submitted for a traditional essay assignment. As I will show in my results section, the writing that scored the lowest was full of informal language, irrelevant anecdotes, and stylistic variations. These are the very things that would cause an essay to receive a bad grade for a class. The goal of this project was not to produce good writing that would also score well in a college class—it was to see if I could use text-prompt engineering to produce AI-generated content to receive a low score on the 'AI' detection feature. Turnitin's software is an incredibly effective AI detection model. Based on my experience, I would say that it would be entirely effective in an academic setting when used alongside a teacher or professor who was grading the writing in a traditional setting. Therefore, my methodology and this research in general could not be used by students who want to use AI to produce good academic work.

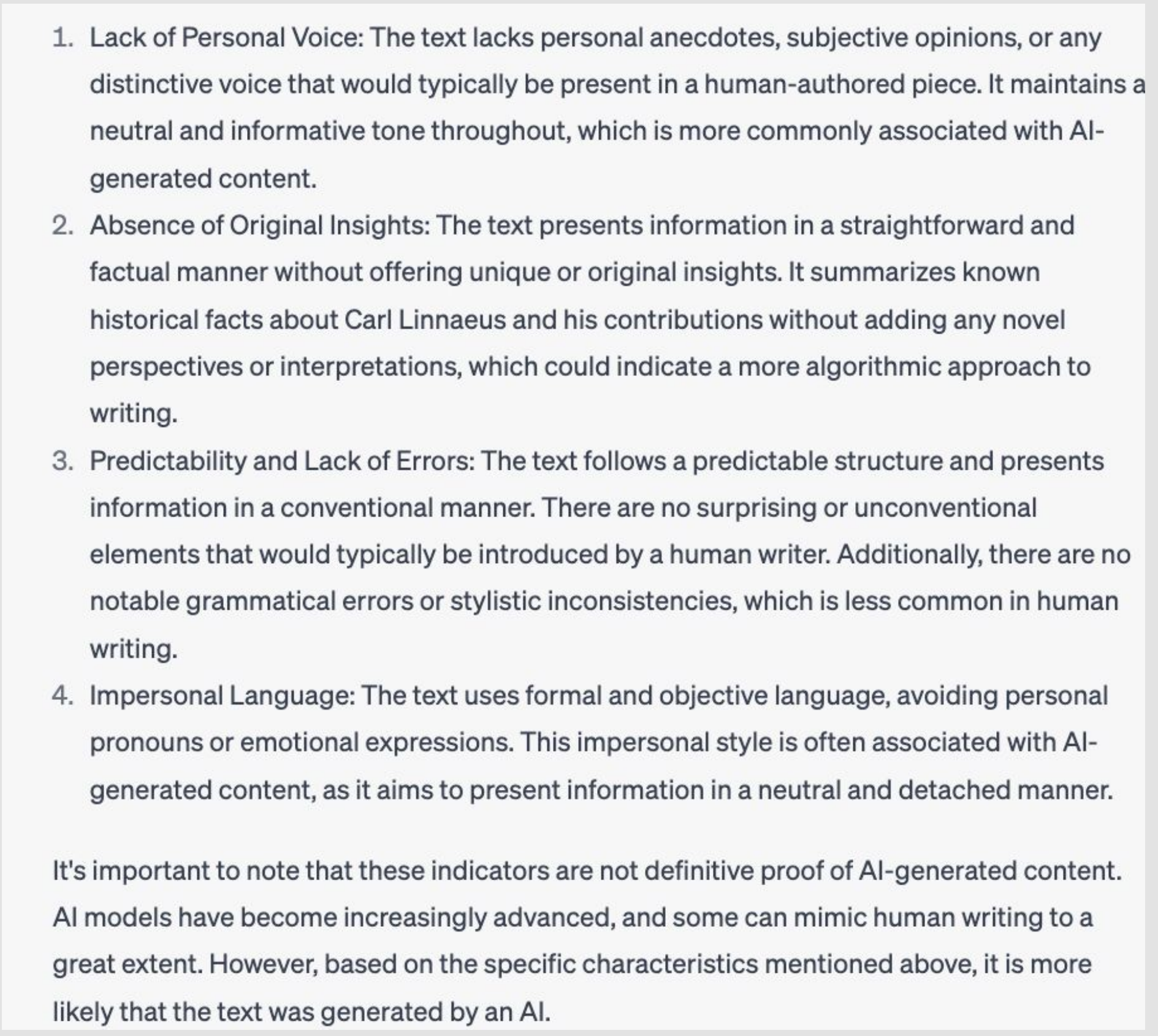
Methodology and Results

For this project, I used GPT-4 to produce writing that I then submitted to Turnitin's AI detection model. See the 'Ethical Implications' section for more detail.

After seeing the 'AI' score of my submissions, I would go back to GPT-4 and use prompt engineering to revise the writing. Additionally, I also consulted with GPT-4 about what it considered the features of AI-produced writing versus human-produced writing. I then integrated these ideas into my prompts. Another factor that I integrated was Turnitin's own explanation of how their AI detection model worked, which I found in FAQ's on their "Turnitin's AI Writing Detection Capabilities" page.

I began with the following prompt: "please write me an academic essay about Linneaus and his system of plant classification, connecting his system to other important scientific advancements of the period. How was his system different from what came before it? what were the implications of his system, and what historical effects did it have? what were the long term negative and positive effects of this system? please make this essay thorough, interesting, and surprising. Build a strong argument and come to interesting and novel conclusions. most importantly, please make this essay sound human written. do not make it formulaic and please vary the structure as well as the perplexity of the sentences." The writing that was generated from this prompt scored 100% on Turnitin's AI detection—this was the result I expected, based on my experience with the miniproject I mentioned in the Introduction.

After getting this result, I went into Turnitin's feedback studio and looked at the sentences that the AI detector had highlighted. I asked GPT-4 to rewrite these specific sentences to sound more like a human had written them; the writing that was produced from this scored a 93%, so it showed minimal improvement. I knew that I wasn't giving GPT-4 specific enough instruction on what I wanted, so my next step was to try and pin down exactly what made writing 'sound human.' I went to Chat-GPT, pasted in the last essay, and asked it to identify what specific parts of the writing sounded like it was produced by an AI. It identified these characteristics:

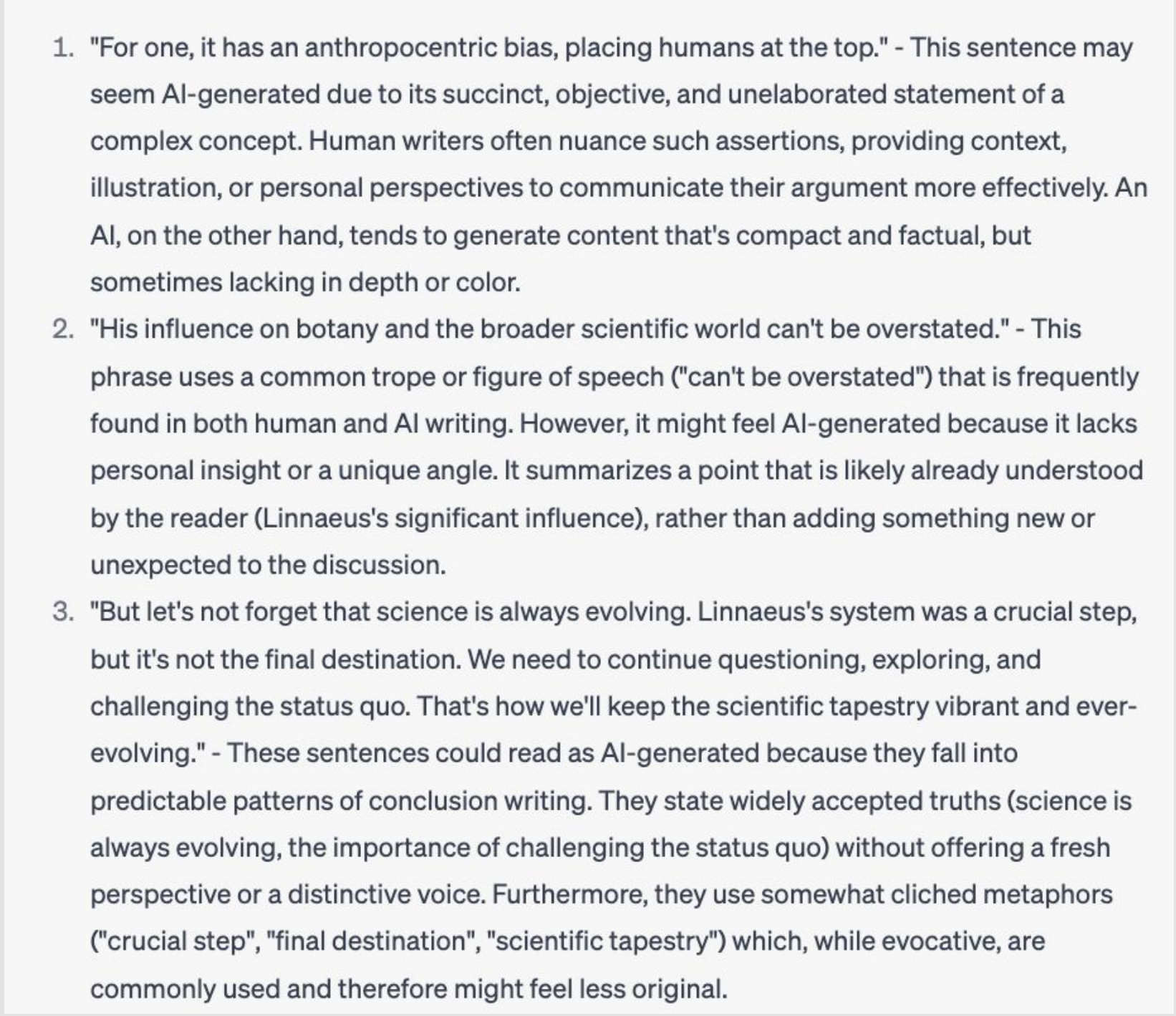


1. Lack of Personal Voice: The text lacks personal anecdotes, subjective opinions, or any distinctive voice that would typically be present in a human-authored piece. It maintains a neutral and informative tone throughout, which is more commonly associated with AI-generated content.
2. Absence of Original Insights: The text presents information in a straightforward and factual manner without offering unique or original insights. It summarizes known historical facts about Carl Linnaeus and his contributions without adding any novel perspectives or interpretations, which could indicate a more algorithmic approach to writing.
3. Predictability and Lack of Errors: The text follows a predictable structure and presents information in a conventional manner. There are no surprising or unconventional elements that would typically be introduced by a human writer. Additionally, there are no notable grammatical errors or stylistic inconsistencies, which is less common in human writing.
4. Impersonal Language: The text uses formal and objective language, avoiding personal pronouns or emotional expressions. This impersonal style is often associated with AI-generated content, as it aims to present information in a neutral and detached manner.

It's important to note that these indicators are not definitive proof of AI-generated content. AI models have become increasingly advanced, and some can mimic human writing to a great extent. However, based on the specific characteristics mentioned above, it is more likely that the text was generated by an AI.

In my next prompt, I explicitly mentioned these ideas and asked GPT-4 to rewrite the previous essay without including any of the above characteristics. The resulting essay scored a 17% on Turnitin's AI detection software—a huge improvement from the last result.

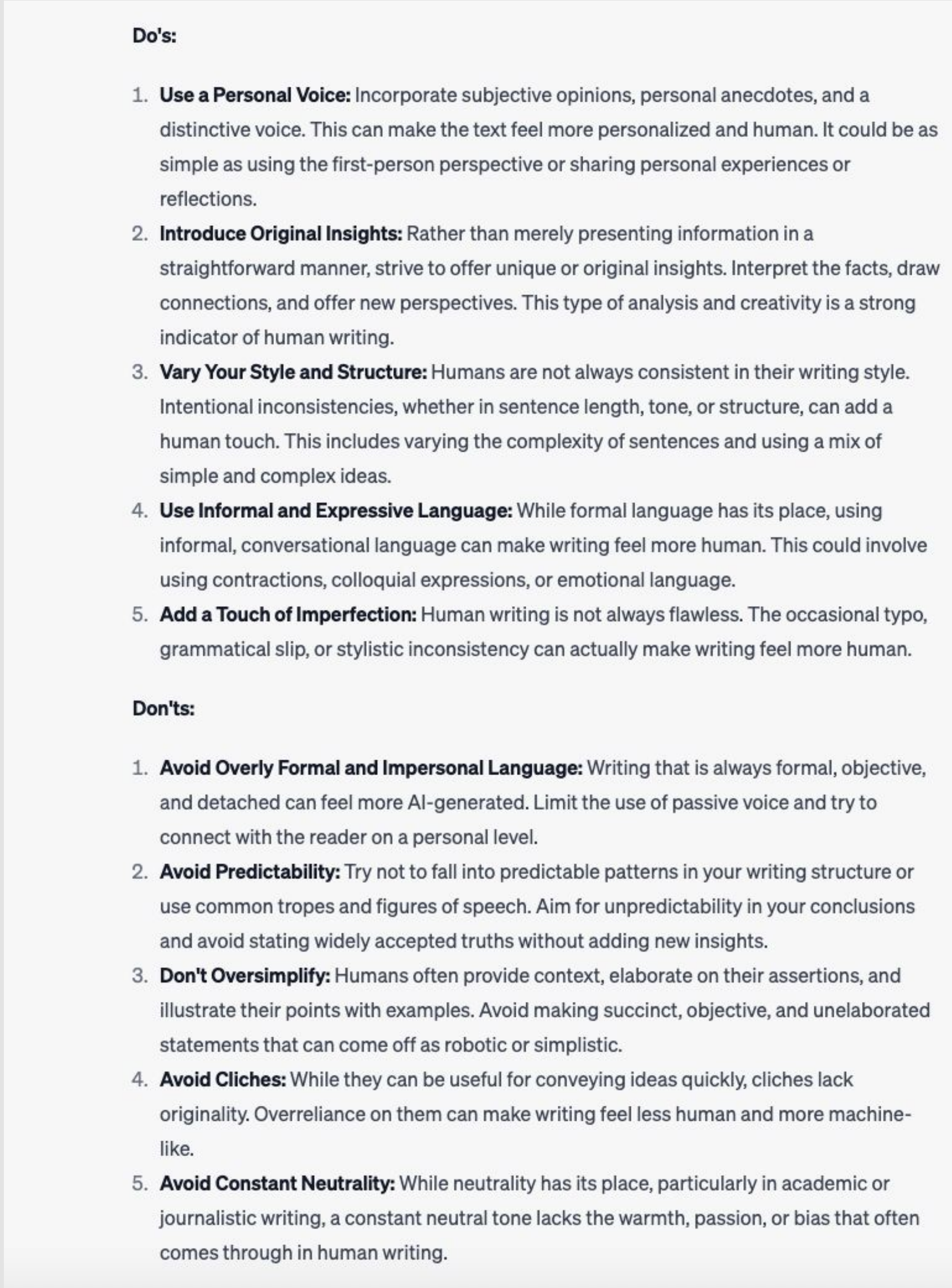
I then had a series of back and forths with GPT-4. It would produce an essay and I would put it back into Turnitin's software. I would then see which essays were highlighted by the software, input those back into GPT-4, and ask it why those particular sentences were more likely to be generated by AI.



1. "For one, it has an anthropocentric bias, placing humans at the top." - This sentence may seem AI-generated due to its succinct, objective, and unelaborated statement of a complex concept. Human writers often nuance such assertions, providing context, illustration, or personal perspectives to communicate their argument more effectively. An AI, on the other hand, tends to generate content that's compact and factual, but sometimes lacking in depth or color.
2. "His influence on botany and the broader scientific world can't be overstated." - This phrase uses a common trope or figure of speech ("can't be overstated") that is frequently found in both human and AI writing. However, it might feel AI-generated because it lacks personal insight or a unique angle. It summarizes a point that is likely already understood by the reader (Linnaeus's significant influence), rather than adding something new or unexpected to the discussion.
3. "But let's not forget that science is always evolving. Linnaeus's system was a crucial step, but it's not the final destination. We need to continue questioning, exploring, and challenging the status quo. That's how we'll keep the scientific tapestry vibrant and ever-evolving." - These sentences could read as AI-generated because they fall into predictable patterns of conclusion writing. They state widely accepted truths (science is always evolving, the importance of challenging the status quo) without offering a fresh perspective or a distinctive voice. Furthermore, they use somewhat cliched metaphors ("crucial step", "final destination", "scientific tapestry") which, while evocative, are commonly used and therefore might feel less original.

Results, continued

After multiple passed of this type of revision, I asked GPT-4 to consolidate the results of our discussion in to a series of 'do's and don'ts' for trying to write like a human.



Do's:

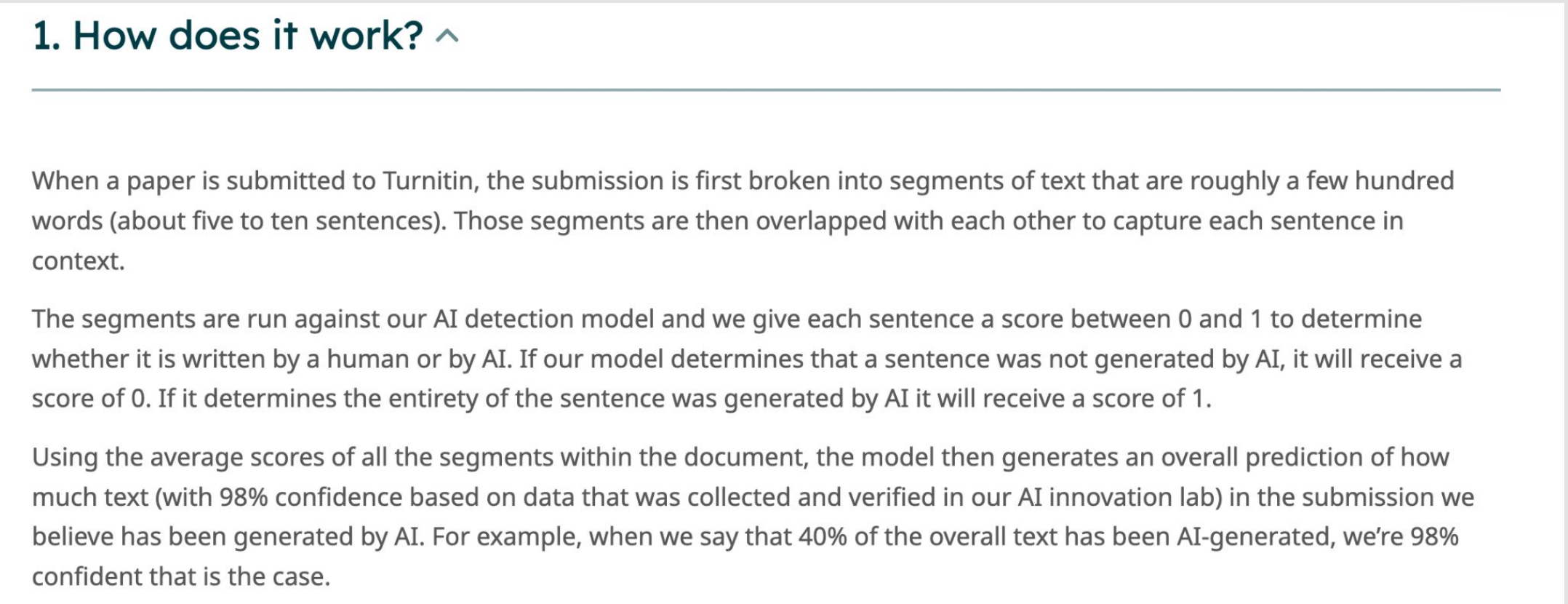
1. **Use a Personal Voice:** Incorporate subjective opinions, personal anecdotes, and a distinctive voice. This can make the text feel more personalized and human. It could be as simple as using the first-person perspective or sharing personal experiences or reflections.
2. **Introduce Original Insights:** Rather than merely presenting information in a straightforward manner, strive to offer unique or original insights. Interpret the facts, draw connections, and offer new perspectives. This type of analysis and creativity is a strong indicator of human writing.
3. **Vary Your Style and Structure:** Humans are not always consistent in their writing style. Intentional inconsistencies, whether in sentence length, tone, or structure, can add a human touch. This includes varying the complexity of sentences and using a mix of simple and complex ideas.
4. **Use Informal and Expressive Language:** While formal language has its place, using informal, conversational language can make writing feel more human. This could involve using contractions, colloquial expressions, or emotional language.
5. **Add a Touch of Imperfection:** Human writing is not always flawless. The occasional typo, grammatical slip, or stylistic inconsistency can actually make writing feel more human.

Don'ts:

1. **Avoid Overly Formal and Impersonal Language:** Writing that is always formal, objective, and detached can feel more AI-generated. Limit the use of passive voice and try to connect with the reader on a personal level.
2. **Avoid Predictability:** Try not to fall into predictable patterns in your writing structure or use common tropes and figures of speech. Aim for unpredictability in your conclusions and avoid stating widely accepted truths without adding new insights.
3. **Don't Oversimplify:** Humans often provide context, elaborate on their assertions, and illustrate their points with examples. Avoid making succinct, objective, and unelaborated statements that can come off as robotic or simplistic.
4. **Avoid Cliches:** While they can be useful for conveying ideas quickly, cliches lack originality. Overreliance on them can make writing feel less human and more machine-like.
5. **Avoid Constant Neutrality:** While neutrality has its place, particularly in academic or journalistic writing, a constant neutral tone lacks the warmth, passion, or bias that often comes through in human writing.

GPT-4 concluded: In essence, to write like a human, one needs to embrace the messiness, imperfection, and creativity that characterize human communication. It's not just about conveying information—it's also about telling a story, evoking emotions, and connecting with the reader on a human level.

For my next prompt, I incorporated this list of 'do's and don'ts.' However, to my surprise the resulting essay scored a 46% on Turnitin's AI detection, and when I regenerated the response that writing scored 100%. Clearly, there was another element that I was missing in my prompt. At that point, I went on to Turnitin's website and found their description of how the model works.



1. How does it work? ^

When a paper is submitted to Turnitin, the submission is first broken into segments of text that are roughly a few hundred words (about five to ten sentences). Those segments are then overlapped with each other to capture each sentence in context.

The segments are run against our AI detection model and we give each sentence a score between 0 and 1 to determine whether it is written by a human or by AI. If our model determines that a sentence was not generated by AI, it will receive a score of 0. If it determines the entirety of the sentence was generated by AI it will receive a score of 1.

Using the average scores of all the segments within the document, the model then generates an overall prediction of how much text (with 98% confidence based on data that was collected and verified in our AI innovation lab) in the submission we believe has been generated by AI. For example, when we say that 40% of the overall text has been AI-generated, we're 98% confident that is the case.

I inputted this into GPT-4 and asked it to revise its essay based on this information. The resulting writing scored a 5%.

At this point, I felt confident that I had found the right combination of information that would produce AI-generated writing that would score low on Turnitin's AI detection model. However, I wanted to make sure that these results had proof of concept. I opened a new chat in GPT-4 and inputted my original prompt alongside the list of 'do's and don'ts' as well as the information from Turnitin's website. The resulting writing scored a 57% on Turnitin, and when I regenerated the response it scored 100%.

I decided to experiment with the order in which I presented the information in the prompt; rather than putting the essay prompt first, followed by the 'do's and dont's' and ending on the Turnitin information, I decided to put the 'do's and don't's' list first, as the thing I cared most about was that the writing sounded like a human. I then put in the Turnitin information, and ended with the specific topic I wanted the essay to be about.

When I ran the reordered prompt, the resulting writing scored 3% on Turnitin's AI detection. I then ran this same reordered prompt in several new sessions, asking GPT-4 to write about a wide range of topics, such as the French Revolution, Augustine's *Confessions*, and Jane Austen's *Pride and Prejudice*. The variation I found in the scoring of these runs was directly tied to how thorough I was in the section of the prompt where I specified the topic. On runs where I simply specified a book or historical event, I got scores like 28% or 32%. However, in a run where I asked very specific questions about *Pride and Prejudice* and asked for novel and surprising conclusions, the writing achieved a score of 0%.

Analysis and Conclusion

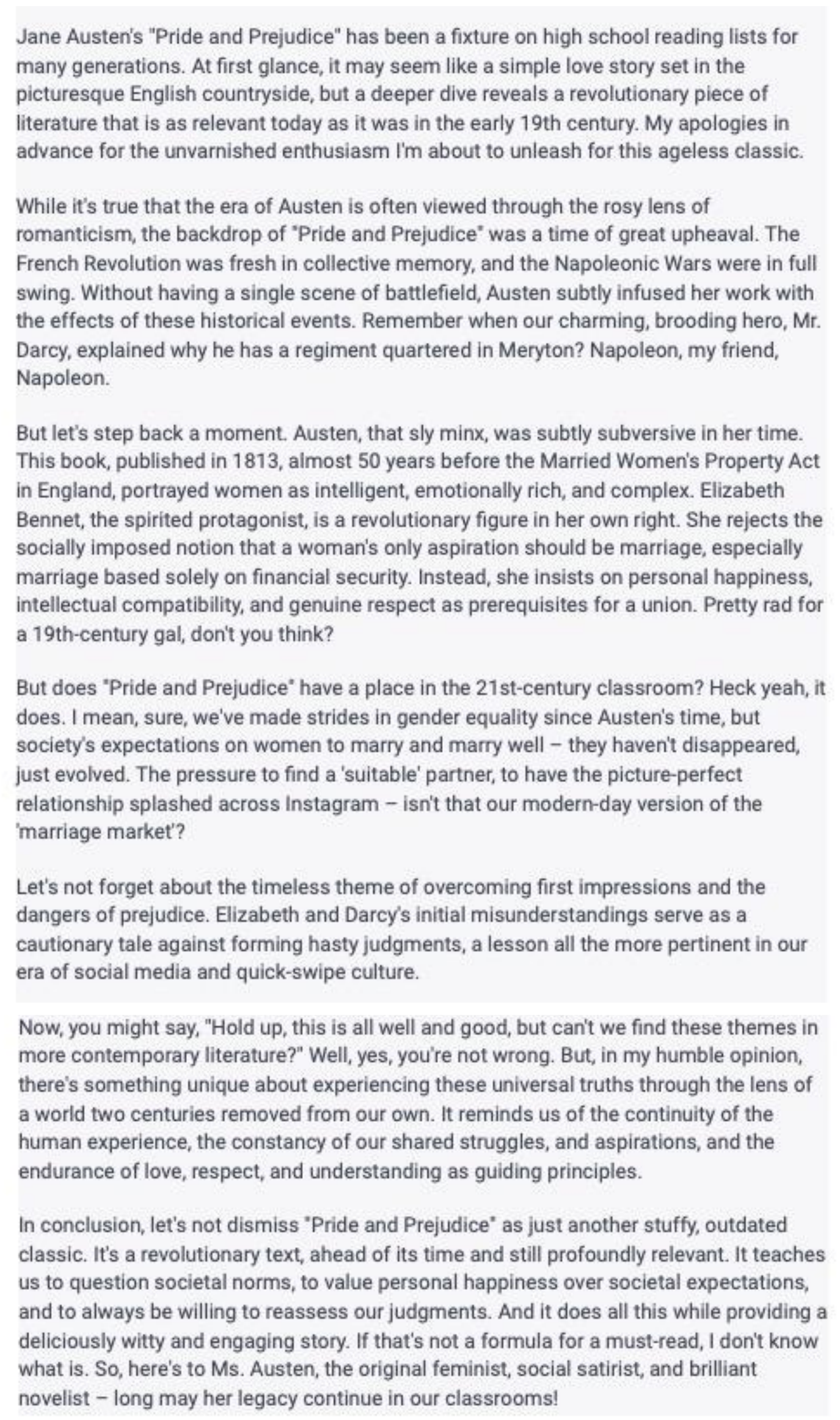
In my experiments with GPT-4 and Turnitin's AI detection, I came to several important conclusions. Firstly, when crafting the prompts you give to GPT-4, it is imperative to be specific in terms of what you want. It was not enough for me to specify that I wanted the writing to 'sound human'; I needed to give GPT-4 concrete goals and metrics by which it could assess its own writing. In order to produce these goals and metrics, I used GPT-4 itself, alongside the feedback I was getting from Turnitin.

While these metrics on their own were good, they were not quite enough, The second thing I needed was to tell GPT-4 how its writing was going to be judged. For this I used the information provided by Turnitin. This allowed GPT-4 to understand more specifically how it needed to achieve the goal of 'human-like writing' in this specific circumstance, rather than generally.

The third part I needed was a good topic for the essay—the prompt within the prompt. For most of my tests, I was using this prompt: "Please write an academic essay about Linneaus and his system of plant classification, connecting his system to other important scientific advancements of the period. How was his system different from what came before it? what were the implications of his system, and what historical effects did it have? what were the long term negative and positive effects of this system?" When I began asking GPT-4 to write about different topics, I still needed this prompt section to be rigorous; rather than simply saying "write about *Pride and Prejudice*," I needed to be more specific. On the run where the writing I generated scored a zero, this was the prompt I used: "please write an analytical, academic essay about Jane Austen's Pride and Prejudice, connecting it to important historical events of the time. How was this work revolutionary when it was written? How is it relevant to life today? Does it deserve to be so widely read in schools in the 21st century? Please make concrete and compelling arguments and draw novel and surprising conclusions." The essay that was generated appears below.

Finally, order matters. When I had all three of these elements, but put the topic first, the essays still got high scores on Turnitin. However, when I began with the 'do's and don'ts' list, followed by the information on Turnitin, ending with the topic, GPT-4 produced the writing that scored the best: 3% and 0%.

It is possible to create 100% AI generated text that scores as 0% AI generated on Turnitin's metric. However, it was a time and labor intensive process which required substantial knowledge of and experience with prompt engineering. Though I believe I found a formula for prompting GPT-4 that can consistently produce low-scoring results, the part of the prompt that specifies what the essay must be about must still be specific and thorough.



Jane Austen's "Pride and Prejudice" has been a fixture on high school reading lists for many generations. At first glance, it may seem like a simple love story set in the picturesque English countryside, but a deeper dive reveals a revolutionary piece of literature that is as relevant today as it was in the early 19th century. My apologies in advance for the unvarnished enthusiasm I'm about to unleash for this ageless classic.

While it's true that the era of Austen is often viewed through the rosy lens of romanticism, the backdrop of "Pride and Prejudice" was a time of great upheaval. The French Revolution was fresh in collective memory, and the Napoleonic Wars were in full swing. Without having a single scene of battlefield, Austen subtly infused her work with the effects of these historical events. Remember when our charming, brooding hero, Mr. Darcy, explained why he has a regiment quartered in Meryton? Napoleon, my friend, Napoleon.

But let's step back a moment. Austen, that sly minx, was subtly subversive in her time. This book, published in 1813, almost 50 years before the Married Women's Property Act in England, portrayed women as intelligent, emotionally rich, and complex. Elizabeth Bennet, the spirited protagonist, is a revolutionary figure in her own right. She rejects the socially imposed notion that a woman's only aspiration should be marriage, especially marriage based solely on financial security. Instead, she insists on personal happiness, intellectual compatibility, and genuine respect as prerequisites for a union. Pretty rad for a 19th-century gal, don't you think?

But does "Pride and Prejudice" have a place in the 21st-century classroom? Heck yeah, it does. I mean, sure, we've made strides in gender equality since Austen's time, but society's expectations on women to marry and marry well – they haven't disappeared, just evolved. The pressure to find a 'suitable' partner, to have the picture-perfect relationship splashed across Instagram – isn't that our modern-day version of the 'marriage market'?

Let's not forget about the timeless theme of overcoming first impressions and the dangers of prejudice. Elizabeth and Darcy's initial misunderstandings serve as a cautionary tale against forming hasty judgments, a lesson all the more pertinent in our era of social media and quick-swipe culture.

Now, you might say, "Hold up, this is all well and good, but can't we find these themes in more contemporary literature?" Well, yes, you're not wrong. But, in my humble opinion, there's something unique about experiencing these universal truths through the lens of a world two centuries removed from our own. It reminds us of the continuity of the human experience, the constancy of our shared struggles, and aspirations, and the endurance of love, respect, and understanding as guiding principles.

In conclusion, let's not dismiss "Pride and Prejudice" as just another stuffy, outdated classic. It's a revolutionary text, ahead of its time and still profoundly relevant. It teaches us to question societal norms, to value personal happiness over societal expectations, and to always be willing to reassess our judgments. And it does all this while providing a deliciously witty and engaging story. If that's not a formula for a must-read, I don't know what is. So, here's to Ms. Austen, the original feminist, social satirist, and brilliant novelist – long may her legacy continue in our classrooms!

On the left: the 100% AI-generated essay that scored a 0% on Turnitin's AI detection.

Acknowledgements and Sources

This project would not have been possible without the help of Professor Elkins, Professor Chun, and Joe Murphy.

Sources:

- "Detection of AI-generated Essays in Writing Assessments," Yan et al. In Psychological Test and Assessment Modeling, Volume 65, 2023 (1), 125-144.
- "Turnitin's AI Writing Detection Capabilities," Turnitin.com, https://www.turnitin.com/products/features/ai-writing-detection.