Dan Bailey

ENG 381

Final Project

Is ChatGPT Really Language?

While there are some misunderstandings about ChatGPT, there is no denying that it is a world-changing technology. Contrary to what the media says, ChatGPT is *not* AI, at least not in the general sense of the term. ChatGPT is classified as a large language model (LLM), the function of which I will explain in this paper. And while what it produces looks and reads like language, is it truly producing something that can be identified as language? Fortunately, there’s a couple of mathematical tools that can tell us whether ChatGPT bears the same results as written English. In this paper, in addition to explaining how an LLM works, I will discuss two methods using formulae by Zipf and Shannon to perform signal analysis on the output of ChatGPT versus a control group of written English language produced by human authors and compare both to a collection of random words generated by an online API.

How ChatGPT Works

Ultimately, the best way to describe ChatGPT is as a very complex statistical model of language, built on doing an analysis of the language found on the Internet. Explaining how the model is trained and the methodologies used for that is a lengthy topic and outside the scope of this section. The short form would be to say that a large language model (LLM) is quite literally trained to determine which word is most likely to appear next in a sequence of words. (Romponi, 2023) For example, given a fragment of a sentence, such as

The dog chased the

The next word might have a high probability of being *squirrel, cat, mailman,* or *car,* with each word having a different probability. At the opposite end of the spectrum, the next word is unlikely to be *asteroid, book,* or *Lego.* ChatGPT looks backwards several words as it decides which word comes next — effectively looking at “what word comes next in this chain?” (ByteByteGo, 2023.)

Whether this constitutes actual “language” or mere mimicry thereof will be discussed in the afterword section. Let’s get to the testing!

Zipf’s Law: What Is It? Why does it matter here?

Zpif’s Law is a mathematical property of distributions of information. From the Wikipedia article on the article, “holds, approximately, when a list of measured values is sorted in decreasing order. It states that the value of the *n*th entry is [inversely proportional](https://en.wikipedia.org/wiki/Inversely_proportional) to *n*.” (Zipf’s law, 2023.) In plain English, if we assume a list of words in human language, the second-most common word will appear half as many times as the most-common word, the third-most common would appear one-third as often, and so on. A better way to look at this would be via the table below:

|  |  |
| --- | --- |
| **Rank** | **Frequency** |
| 1 |  |
| 2 | 1/2 |
| 3 | 1/3 |
| 4 | 1/4 |
| 5 | 1/5 |
| … | … |
| 10 | 1/10 |

Zipf’s Law has been shown to hold for most natural languages, and even non-natural cases like Esperanto (*Zipf’s law,* 2023.) Based on this information, I decided I would look at the top ten most common words in English when compared to the ChatGPT dataset, the human author dataset, and a random words dataset.

Shannon: What is Entropy and Why Does It Matter?

Claude Shannon published his paper, A Mathematical Theory of Communication in 1948, and it said some interesting stuff. The crux of it is: if someone tells you something that you already know, they haven’t communicated with you.

Wait. What? Let’s try using someone else’s words: “it provides a rigorous mathematical framework for quantifying the amount of information needed to accurately send and receive a message, as determined by the degree of uncertainty around what the intended message could be saying.” (Hartnett, 2022.).

That’s not much more clear is it? Okay, let’s go for a coin flip. Except let’s say you have a defective coin that escaped the US Treasury with George Washington’s face on both sides. If I flip it twice, it takes no information to communicate the results because we can be assured that the results will be the same, no matter how many times we flip it.

Okay, let’s try again with a normal coin. And let’s use bits to represent the faces of the coin. George Washington is a 1, and the eagle on the back is a 0. If we flip the coin twice and record the results, there are four possible combinations. Those are: 00, 11, 01, and 10.

In our first scenario with the defective coin, if you try to guess the outcome, you’re 100% likely to get it right. In our second scenario with the normal coin, trying to guess the possible outcome lands you a 25% chance of getting it right. To transmit the results, you need to send two bits of information to your receiver for them to know with certainty what the outcome was. (There are complications here with missing bits and knowing the position of the missing bit and helping clear up ambiguity, but we don’t need to go down that rabbit hole.)

Basically, what Claude Shannon did with this paper is put forth a formula that can tell you the minimum number of bits required to communicate a message. This is entropy – it’s a measure of disorder. And it’s an idea borrowed from physics.

So let’s look at the physical world for a minute. Let’s compare a boulder and a pile of gravel. The gravel has higher entropy than the tree because there’s more ways to push it around and rearrange it compared to the boulder.

In language, if I send you a bunch of random letters at you, that’s high entropy – it can be rearranged however you like. But in the case of letters obeying strict patterns, that’s lower entropy. The low entropy is the signal.

For the English language, Shannon calculated that the average word length was 4.5 letters, and that the entropy was 2.62 bits per letter. Thus, the entropy for English per word is 11.62 bits. (Shannon, 1948.)

There’s a strong signal in the noise.

The Data Sets for This Experiment

The data sets were targeted to be in the vicinity of 100,000 words each. There was some variability due to numbers and malformed text causing some issues. All punctuation was stripped out and all text was converted to lowercase for the purposes of making it easier to count words (a computer views STOP, stop, and Stop as three different words). There were, ultimately three different sets of data.

The first was approximately 96,000 words of responses from ChatGPT to a variety of requests for information about a variety of topics. This information has been submitted prior to the writing of this paper but will be included in Appendix A of this paper. The second set was approximately 100,000 words of material copied-and-pasted from a variety of e-books, the list of which will be provided in Appendix B. Lastly, was a set of 100,000 random words, generated by an API query (API-Ninja, 2023). Those I will put in a GitHub repository for download, the link to which will be in Appendix C.

The datasets were used to run both the Zipf distribution analysis, and the Shannon entropy analysis, with varying results.

Results

As it was the first method discussed, we will visit the Zipf analysis first. The following table shows the results for the top ten words from each data set and the number of occurences of each word in parentheses after the word.

|  |  |  |
| --- | --- | --- |
| Human Authors | ChatGPT | Random Words |
| the (5823) | the (6228) | cut (7) |
| of (2880) | and (4797) | head (7) |
| to (2879) | of (3356) | address (6) |
| and (2716) | a (2574) | better (6) |
| a (2428) | to (2259) | white (6) |
| in (1681) | in (1873) | carry (6) |
| that (1483) | for (899) | stroke (6) |
| i (1400) | is (886) | devotion (5) |
| is (1054) | as (702) | right (5) |
| It (1028) | with (695) | ground (5) |

Knowing prior to this paper that the two most common words in the English language are “the” and “of”, the ChatGPT results list stuck out like a sore thumb. The randomly-generated words shook out about like I’d expected to – when randomly pulling 100,000 words from a list of approximately 171,000 there are likely to be some duplicates, but not many.

What also jumped out at me, was the word frequency. If we look at column one (the human author generated data) in the above table, the most common word (“the”) appears 5823 times. As such, we’d expect the second-most common word to appear half that amount: 2911 times. “Of,” “To,” “And,” and “A” all are well in the range of being the number two word. The third-most common word should appear approximately 1941 times, which “in” is closest to. “That,” and “I” are pretty close to the fourth most common value, which should be 1320 occurences. The fifth most common word should appear 1056 times, and “is” and “it” are both close (“is” particularly so). So what this looks like is that the word distribution is “clumpy”, but in-line with expectations.

Repeating this same exercise with column two in the table (ChatGPT’s text) shows some serious deviation from Zipf’s Law – the most common word (“the”) appears 6228 times, which means that the second most-common word should appear 3114 times. But it doesn’t. The second-most common word, “and,” appears 4797 times! Almost 50% more than it should! What’s going on here? Our third-most common word, “of,” should appear 2076 times, but appears 3356 – more than 50% more than it should! ChatGPT’s output *does not conform to* *Zipf’s Law!* (At least where this data set is concerned.)

Exploring the data a bit more, I chased down English language word frequency and found that there was a great deal of good information derived from the Google Web Trillion Word Corpus. As observed in the online article “English Word Frequency”, there’s a different set of top ten. (Tatman, 2017.) I compared their data with mine in the table below and highlighted in green the words that appear in all three lists. As you can see, while there is some agreeance, it's not entirely accurate. I would expect that a large language model that was trained off the internet would most-closely conform to the Google Web Trillion Word Corpus. It’s close, to be sure, but not exactly. Given that OpenAI, the creators of ChatGPT, and Google are two separate entities, it is highly probable that there were different rules in-place when the information that contributed to their individual projects was harvested, that resulted in slightly different results.

|  |  |  |
| --- | --- | --- |
| Human Authors | ChatGPT | Google Trillion Word |
| the | the | the |
| of | and | of |
| to | of | and |
| and | a | to |
| a | to | a |
| in | in | in |
| that | for | for |
| i | is | is |
| is | as | on |
| it | with | that |

Zipf, incidentally, has an additional law called the Law of Abbreviation, that states that the more a word is used, the shorter it is. So, there was an interesting opportunity to see if this was the case with both the human authors and the ChatGPT data sets.

I took the entire lists of words and what I did was calculate the length of each word on the list and multiplied it by the number of times it appeared in the data set. Then I totalled those values up and divided by the number of words in the data set. This gave me the average length of a word in the data set.

|  |  |
| --- | --- |
| Data Set | Average Word Length |
| English Language | 4.5 (Shannon, 1948.) |
| Human Authors | 4.5 |
| ChatGPT | 5.5 |

Another case where ChatGPT and Zipf don’t get along. I’ll speculate on that in the next major section (“Conclusions and Afterword”).

Transtioning to the Shannon Entropy Analysis, I’m pulling information from both the section on Zipf and from the earlier section explaining Shannon Entropy to provide these values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | Word Length | Bits/Character | Bits/Word |
| Shannon, 1948. | 4.5 | 2.62 | 11.82 |
| Human Authors | 4.5 | 4.20 | 18.90 |
| ChatGPT | 5.5 | 4.16 | 22.88 |

So the interesting part here is that the human authors and ChatGPT data sets have a higher number of bits per character than Shannon’s original 8000-word data set. The question, of course, is why? Without having access to that data set, there’s no way of knowing for certain. However, it is interesting to note that the Human Authors and ChatGPT both have very similar bits/character rating. The word-length difference means that there’s more entropy/uncertainty with ChatGPT on a per-word basis than there is with human authors.

Conclusions and Afterword

Whether ChatGPT’s output can be considered language is a question that can be addressed from two different angles – philosophically and computationally. Putting aside the philosophical aspect, and looking at the data we’ve crunched here, ChatGPT’s output does not conform to the Zipf – distribution is definitely an issue, and abbreviation may be at-issue as well, given that the average word length in ChatGPT-generated text is nearly a full letter longer than natural English.

ChatGPT and the human authors received extremely similar scores in Shannon Entropy – so similar, in fact, that I’d judge the .04 difference to be statistical noise.

When both harvesting and generating data, ChatGPT is using functions similar to Shannon, which would explain why their entropy score is so much closer to the human side. But why the average word length difference? It’s hard to say. Speculation is easy – programmatic error, bias in data fed to the LLM, and any other number of things could be influencing the output.

Personally, I’m of the opinion that yes, from a computational standpoint, it’s possible to sniff out ChatGPT data. As for the human “smell test”, something stinks. The more data I generated and the more time I spent reading ChatGPT output, the more I felt I could identify it. It felt repetitive at a structural level and less like a human-created work. Something was off.

All that said, I do have ideas for a second round of this. Ideas that would entail new (and different) data sets. This is because I feel like the output is tailored around my interests and preferences and that results in some weirdness – the hundreds of queries (Appendix A) were largely based off of my own interests and a lot of late-night caffeine. So we have hundreds of short articles created by a single “author” (ChatGPT), which we’re comparing to huge chunks of about 25 books out of my personal library, each by a different author.

I think for the next time around, I’d throw out the random words data set I created for this exercise. It was useful for looking at Zipf, but useless for looking at Shannon Entropy (for that, in retrospect, I should have generated 450,000 random characters and analyzed those). I’d also like to do a human-vs.-ChatGPT situation in which a prompt is given to a human and a prompt is given to ChatGPT, for them each to produce a similar word count. This would eliminate the issues with different numbers of sources and different word lengths from each source.

What I’ve discovered here is that, yes, there’s some degree of watermarking in ChatGPT’s output, but that from a computational linguistics perspective, there’s so much more to dig into. (Like every other subject we’ve covered this semester, really.)

Sources

Shonihei. (2017, May 1). *Shonihei/zipfs-law: Confirming that Zipf’s law is real and amazing*.  
GitHub. <https://github.com/shonihei/zipfs-law>

Ramponi, M. (2023, August 4). How CHATGPT actually works. AssemblyAI. <https://www.assemblyai.com/blog/how-chatgpt-actually-works/>

ByteByteGo. (2023, April 24). *How chatgpt works technically | chatgpt architecture*. YouTube. <https://www.youtube.com/watch?v=bSvTVREwSNw>

Wikimedia Foundation. (2023, November 20). *Zipf’s Law*. Wikipedia. <https://en.wikipedia.org/wiki/Zipf%27s_law>

Wikimedia Foundation. (2023, November 20). *Entropy (information theory)*. Wikipedia. <https://en.wikipedia.org/wiki/Entropy_(information_theory)>

Tatman, R. (2017, September 6). *English Word Frequency*. Kaggle. <https://www.kaggle.com/datasets/rtatman/english-word-frequency>

Hartnett, Kevin. (2022, September 6). *How Shannon Entropy Imposes Fundamental Limits on Communication*. Quanta Magazine. <https://www.quantamagazine.org/how-claude-shannons-concept-of-entropy-quantifies-information-20220906/>

Shannon, C. (1948, October). *A Mathematical Theory of Communication*. Harvard Mathematics Website. <https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>

Appendix A: The ChatGPT Queries

Generating the ChatGPT queries required me to use the standard online ChatGPT interface to request information and then to copy and paste results into a ridiculously large text file. (Those files are available for download in my GitHub repository, a link to which is provided in Appendix C.) The queries used are:

* Can you give me a 2000-word summary on string theory?
* Can you give me a 2000-word summary on the history of the Israeli-Palestinean conflict?
* Can you give me a 2000-word summary on the current state of oncology?
* Can you summarize the movie "The Rock" (1996) as a Shakespearean sonnet?
* Can you summarize the Freudian elements present in "Return of the Jedi"?
* Can you provide a 2000-word summary on the Trump administration's response to COVID-19?
* Please provide 2000 words comparing and contrasting Microsoft's Azure cloud services and those of AWS.
* Please provide an analysis of the movie "The Blues Brothers" (1980), including symbology and themes, as well as discussing what influences from literature are present in the film.
* Discuss the validity of the "Tommy Westphal Universe" and whether the shows are interconnected or writers are just lazy.
* Please write 500 words of fiction in the style of Stephen King about a barn haunted by the ghost of a soldier who died in World War 1.
* Write 500 words of fiction in the style of Ernest Hemingway about an Irish Immigrant who finds his way to California in the late 1800's.
* Write a Shakespearean sonnet about the Chicago Cubs finally winning the World Series.
* I need a 2000-word summary of the history of the C programming language.
* Can you give me a 3000-word summary of the Obama presidency?
* Can you provide a 4000-word summary of the major tourist attractions in England?
* Why is doping such a problem in pro cycling?
* Please talk about the strengths and weaknesses of Garrett Lisi's E8 theory.
* Can you give me a 1200-word summary about the current thinking around microinteractions in UX/UI design?
* Can you provide 3000 words on the current state of CO2 emissions in the United States?
* Can you write me a series of 2 or 3 limericks on any subjects you choose?
* Can you provide me with a highly detailed description of the process of building a bicycle frame out of steel?
* What are the primary exports of Zimbabwe?
* Please give me a detailed summary of Proteus syndrome.
* I need a 5000 word summary about the history of the Louvre.
* Please provide a summary of Blaxplotation films from the 1970s including insights about their long-term contribution to both the art of cinema, their impact on racism in society, and whether the artistic influences are worth any negative societal effects.
* The 1970's seems to have been peak time for the number of serial killers that were operating -- please compare and contrast the styles, personalities, and victims of Ted Bundy and Elmund Kemper, and comment on how they were both influenced by the 70's zeitgeist.
* Please discuss the problems with the book "Freakonomics"...
* Please discuss the notable findings of behavioral economics and their implications.
* Discuss the success of the video game "Animal Crossing" -- what contributes to its success?
* Why does the Big Lebowski hate the fuckin' Eagles, man?
* Who would you rather have flying your airplane: Donald Trump or a chimpanzee?
* How much wood would a woodchuck chuck if a woodchuck could chuck wood?
* I need a 1000-word summary and review of the movie "Iron Man" (2008).
* Which band is better Pearl Jam or Soundgarden? Why?
* Why did Seattle become a hotbed of music in the early 1990's?
* Please write a sonnet about Joe Biden's accomplishments during his Presidency.
* Please compare and contrast the worldwide political and social environments of the years 1939 and 2019.
* Can you provide 1000 words on the comic book character Peter Parker? Please include information that includes a psychological analysis and how his character has evolved over time to fit with the cultural zeitgeist.
* Please compare and contrast ChatGPT to other large language models.
* Please compare and contrast the current candidates for a Grand Unified Theory.
* Please discuss the theories about the identity of Jack the Ripper.
* Please discuss the evolution of the rules set of Dungeons and Dragons, starting from the first edition through the fifth edition.
* Who invented the mountain bike?
* Are there any amendments to the Constitution that are currently being considered by legislators?
* Why is Elon Musk such a douchebag?
* Why is Bill Gates venerated?
* Please provide a comparison between the Saturn V rocket and SpaceX's new Superheavy/Starship booster.
* Please provide a 2000-word summary of Mars exploration to this point.
* Please provide a summary of Lego's business model.
* Can you provide a 2000-word summary of the Standard Model of physics?
* Can you provide an analysis of the Star Trek: Deep Space Nine series?
* Can you talk about the themes present in the reboot of Battlestar Galactica?
* Please provide a set of instructions for accessing Amazon S3 Glacier and getting an inventory of a vault.
* Please provide a 2000-word summary on the reign of King Louis XIV of France.
* Please provide a 2000-word timeline, overview, and analysis of the Battle of the Bulge.
* Please provide a 2000-word history of video games.
* Please provide a detailed summary of the Clinton presidency.
* Please provide a 3000-word history of the Apollo program.
* Why did the Nixon Administration cancel the Apollo program?
* Explain the waning popularity of baseball, particularly compared to the rise in popularity of American football.
* Please provide instructions for replacing the struts in a 2005 Subaru Impreza WRX.
* Please provide a recipe for piroshki.
* Provide an analysis of the effects of Sesame Street on American society.
* Compare and contrast Mr. Rogers and the Buddha.
* Provide 2000 words on the subject of your choice.
* What can Tetris tell us about the human need to organize?
* Please examine and provide data about any correlation between the stock market and the sunspot cycle.
* Are biorhythms just pseudoscience or is there any evidence to support that crap?
* Please fabricate the lyrics to a new song in the style of Rage Against the Machine.
* Please give me a 2000-word explanation as to why Vladimir Putin hasn't faced a coup during his ongoing war with Ukraine.
* What does the comic book character Venom symbolize?
* What are some of the unseen external influences on the stock market?
* Can you provide an overview as to why foreign aid to Africa has largely been unsuccessful and how it relates to the end of the colonial era?
* Please provide a summary of the Mars trilogy by Kim Stanley Robinson.
* Why was the initial response to the AIDS crisis so badly mismanaged and how does it compare to the COVID response?
* Please provide a 1000-word summary of the Panic of 1907.
* Please provide a 2000-word history of the bicycle.
* Please provide a basic tutorial on TailwindCSS.
* Please provide a 2000-word story, written in the style of Neal Stephenson's "Snow Crash" about a time-traveling cat with a bad attitude problem.
* Please provide a 2000-word summary on current research into treatment for ADHD.
* Write me seven different haiku about cats.
* Please talk about the Millennium Prizes and explain the questions and their implications in plain English.
* What are the shortcomings and problems with the PHP programming language?
* Why is Wordpress so damn popular?
* Please provide a 3000-word summary on the phenomenon of "ghost malls" -- including causes, remedies, and possible uses for those spaces.
* That wasn't 3000 words, please try again
* Please provide a 1000-word summary of the Ford Nucleon car.
* Please write the lyrics to a Motown-style song called "Stop Being a Butthole"...
* If you were going to name a tuxedo cat, what would you call it?
* Which is better: the lions Voltron or the vehicles Voltron?
* Please provide a 3000-word explanation of Einstein's General Theory of Relativity
* What is the best way to understand trigonometry?
* What are the pros and cons of titanium as a material for bicycle frames?
* Summarize the unification of Germany from 1989 to present.
* Why was the Arc de Triomphe built?
* Please summarize the findings of the Large Hadron Collider in the form of a sonnet.
* Please identify and summarize any similarities between "Happy Gilmore" and "Talladega Nights"
* How can humanity save the environment?
* Please explain, in-detail, the influences the Beatles had on modern music, particularly in bands that were formed after the year 2000.
* please write a biography of Greg LeMond
* What band(s) is/are the logical successor to Queen?
* Please describe the eventual Heat Death of the universe.
* Assuming a Big Crunch scenario, describe the future of the universe.
* Please provide a 2000-word summary of the themes in the Great Gatsby.
* Can you summarize Tolstoy’s “War and Peace” in 3000 words?
* Please summarize the history of the unification of Italy in 2000 words or more.
* How did the invasion of Afghanistan in 1979 shape the downfall of the Soviet Union?
* Please provide a summary of theories of quark substructure.
* How was Apple able to come back from the brink of bankruptcy?
* Explain the current scientific consensus on the mind-body connection.
* Please analyze and describe the themes in the song "Blinded by Rainbows" by The Rolling Stones.
* Please write a review of Stephen King's book "The Stand"
* Please write 3000 words on the subject of your choice.
* Please provide another 3000 words on the subject of your choice.
* What would have happened if Patton had been allowed to use nuclear weapons in the Korean War?
* What are the keys to a successful marriage?
* What the key commonalities in a failed marriage?
* Is there an easier way to generate huge blocks of ChatGPT 3.5 responses for the purposes of research?
* Is it possible that I can function better late at night while I am trying to generate data for my research project?
* Should Minnesota secede from the Union and join Canada?
* Can we kick Texas out of the Union?
* Please provide 3000 words on the subject of your choice.
* I need a 4000-word description of the physics around a black hole.
* What are the top sights a tourist should see in Rome?
* What are the top sights a tourist should see in Milan?
* Talk a bit about using Zipf functions to analyze language.
* Talk a bit about using Shannon functions to analyze language.
* Please provide, in plain English, a description of how quantum computing works.
* Tell me more about Dashboard Confessional.
* Please discuss the global impacts of an eruption of the Yellowstone super volcano. For extra credit, explain what would happen in Minnesota, specifically.
* What are the implications of MOND theory?
* Is personal pain required to be a successful comedian?
* How to write a novel?
* Please discuss the unifying themes of the album "August and Everything After..." by Counting Crows.
* How did the Gregorian calendar come to be?
* Please provide 3000 words on the subject of your choice.
* Please talk about the idea of "hidden chords" in music.
* Provide 2000 words on the history of the NHL.
* Provide a summary of the AI-invented sport, Speedgate.
* Explain the offsides rule in soccer.
* Rewrite the plot of the movie “Love, Actually” if it were directed by Quentin Tarantino
* Rewrite the plot of the movie “Pulp Fiction” if it were directed by Wes Anderson.
* What are the top sights for tourists on the island of Hokkaido?
* Why was it so difficult to legalize gay marriage?
* Explain how blacksmiths make a sword.
* Who are the five most controversial SNL hosts?
* What are the five best scary movies?
* What is the worst Halloween candy and why is it candy corn?
* How does Picasso fit into art history? What are his influences and who has he gone on to influence?
* Please discuss in-detail recurring themes in the movies of Kevin Smith.
* Is there any scientific basis for phrenology?
* Is chess a sport?
* Should artificial intelligences have the same rights as people?
* Please explain, in plain English, how a CVT works.
* Where am I going wrong in life?
* Is "Die Hard" a Christmas movie?
* Please write a short story about a farmer living on haunted land.
* Write a short story, retelling the tale of Beowulf, set in a modern call center.
* Rewrite the plot of "Star Wars" if it were directed by Kevin Smith.
* Please provide a 3000-word essay on the topic of the cyberpunk movement in science fiction during the 1980's.
* Please write a 4000-word essay on the topic of your choice.
* Write a history of file sharing, starting with the 1970's BBS systems and proceeding through the platforms of the 2000's, all the way to the BitTorrent sites of the 2020's.
* Please write a short story retelling "The Fifth Element" but set in Middle Earth.
* Write your own take on Romeo & Juliet, written in the style of Cormac McCarthy's "The Road"...
* Write a review for a box of wood screws.
* What is the future of data storage?
* Please write 3000 words on the topic of cheating in the Olympics.
* Please explain the recurring themes on the album "Hot Fuss" by The Killers.
* Write a short synopsis of the entire body of Star Wars movies and shows.
* Please provide a 2000 word essay detailing the layout, character, and history of the city of your choice.
* Tell me everything you know about Amur tigers.
* Who were the belligerents in the Battle of the Bulge?
* Why do Canadians say "aboot"?
* Please provide a 2000-word essay on the subject of your choice.
* Provide a 1000-word essay on the topic of crime and punishment as it exists in 2020.
* At what point does global warming pass the point of no return?

Appendix B: The Sources of Written Text Data

* *Holly* by Stephen King
* *Elon Musk* by Walter Isaacson
* *Blood Meridian* by Cormac McCarthy
* HBR’s 10 Must Reads on Mental Toughness by Harvard Business Review
* Sure, I’ll Join Your Cult by Maria Bamford
* *Diaspora* by Greg Egan
* *Today Matters* by John C. Maxwell
* Zen in the Art of Archery by Eugen Herrigel
* *The Bell Jar* by Sylvia Plath
* The Android’s Dream by John Scalzi
* Bayesian Statistics the Fun Way by Will Kurt
* *Inhibitor Phase* by Alastair Reynolds
* The Anthropocene Reviewed by John Green
* Sure, I’ll Be Your Black Friend by Ben Philippe
* *Rage* by Bob Woodward
* A Memory Called Empire by Arkady Martine
* *The Hurt Artist* by Shane Niemeyer
* I’ll Be Gone in the Dark by Michelle McNamara

Appendix C: The GitHub Repository

The data files and scripts used to run this analysis, and other assorted files, including a copy of this paper, can be found at the following URL:

<https://github.com/dan-bailey/ENG381>