The Poet Who Couldn't Know It: How ChatGPT's Imitation of Poetry Differs From the Real Thing

Gross, Max

McGill University Montreal, QC

Contents

1 Theory

Masahiro Mori's 1970 paper "The Uncanny Valley" defined a term now ubiquitous in discussions about robotics, specifically those autonomous inventions designed to operate like humans. His graph, one based in subjective data mind you, explores the idea that "familiarity" with an apparatus increases with how "human-like" it is, up until the point right before the apparatus is an actual, healthy human. When it is incredibly human-like, but not human itself (like a corpse or a zombie in fiction), there is a dip in the increase of this graph, and ergo the "uncanny valley" is shown.[5]

This paper seeks to demonstrate one example of this phenomenon, the "uncanny valley" of ChatGPT's human-like text generation. Although so close to seeming authentic – to feeling familiar – it does not push far enough to seem exactly human and thus plummets into the valley. However, where Mori's work is subjective, the textual analysis tools of our modern age allow us to apply more quantifiable tests to our data, and thus we can make use of modern developments to offer further evidence of a theory proposed half a century ago.

One of ChatGPT's novel abilities is to produce creative writing, i.e. fictional works which are not as grounded in reality as, say, a math proof or error-detection in a computer program. This is where such an "uncanny valley" effect can truly take hold, as the model must insert its own "voice" into the piece, and it cannot simply rely on regurgitating formulas it has seen online. And because ChatGPT is mostly used for shorter demonstrations of text generation, this study will focus on how the program generates poetry, specifically poetry in the style of notable poets for comparison's sake.

ChatGPT is a technology that exists in a field where "competition is forcing them [A.I. companies] to go too fast and cut too many corners"[3], rapidly developing with a future-facing attitude that does now allow for reflection or time for understanding of what is transpiring. However, as researchers, it is important to study a development even if it is still developing – especially if it is able to be swayed off a dangerous course. Questions considered in this paper include: is ChatGPT more accurate with male poets over female poets? Will ChatGPT express strong political messages if it's replicating ideological poets? When does ChatGPT change the original source material, and when does it copy directly? With the lack of literature on the subject, it being so new, this paper seeks to understand poem generation, this one, perhaps microcosmic, aspect of ChatGPT. It aims to be a stepping stone in the uphill climb in defining it, and – eventually – regulating it so that these biases which creep in, those which create an "uncanny valley" effect, can be minimized. To do so, we must answer the general question of how ChatGPT generated poems diverge from their "real" counterparts before narrowing down to the specifics.

2 Conceptualization

What is empirically being studied in this paper, and why

2.1 Data

The aim of this project is to analyze ChatGPT generated poetry in comparison to real poetry. However, "poetry" here is a broad categorization and one that is only broadening as technology allows art to transcend the printed page into what scholar Ramero-López describes as hypertextual, ecphrastic, and serendipitous literature[2]. In order to be comprehensive and representative while still concise and focused, this paper will focus on five poets who represent different poetic movements, each coming from different backgrounds, sexual orientations, genders, and racial groups. However, of the five, four are men, three are white men and two are (presumably) straight white men. This is simply a reflection of whose poetry has entered the canon, and marginalized groups were not afforded the same notability as majority groups in poetry until very recently; this limits the study, but is unavoidable if time is also to be studied, which it is. The five poets are:

- William Shakespeare, representing Elizabethan sonnetry
- William Butler Yeats, representing the Modernist movement
- Langston Hughes, representing the Harlem Renaissance
- Sylvia Plath, representing the Confessional movement
- Allen Ginsberg, representing the Beat movement

Through analyzing ChatGPT's ability to generate poetry in the style of these writers, each definitely with a unique perspective and quality to their work, this paper will be able to discuss not only if ChatGPT is wrong or right, but where it is wrong and where it is right. Presumably, writing a Shakespearean sonnet is trivial (we get 9$^{\text{th}}$ graders to do it all the time) while understanding the deep depressive aspects of Plath's work requires more humanity and less binary computation. Each group will thus be internally compared to its GPT-generated counterpart of course, but diversifying across five movements allows external comparison as well.

2.2 Semantics

The first metric for which these poems will be analyzed is their semantic relatedness to the source material. Put more simply, each poet has a wheelhouse of topics that their movement and time demands, and ChatGPT is either accurate in its assessment about what these authors would have written or woefully wrong and thinks Shakespeare wrote about the iPhone and Plath compared thee to a summer's day (well, thou art more lovely and more temperate, but that's beside the point).

Sadly, one of the limitations of this project is the sheer volume of documents one would need to run machine learning algorithms, so topic modelling would be a challenge out of the scope of this paper[1]. However, in its place, it is possible to use outside information to gauge roughly what each poet is notable for discussing in their work and then seeing how common words of this theme are when assessing the top words (removing stop words and the like) from the real and GPT data.

For example, if we see that Ginsberg's top words revolve around urban, pacifist, and psychedelic themes (for example, using a word list with "city", "peace", "drug", etc.) and that ChatGPT is more general and themeless, this is empirical evidence to suggest that ChatGPT's ability to replicate semantics is faulty.

The only worry here is that a simple word list is not wholly indicative of the theme of a piece, which is almost certainly true. However, a trend across all authors would seem to suggest some misstep between ChatGPT and the real poems, even more so if GPT is great for more "simple" authors like Shakespeare and struggles to understand more complex or ironic Confessionalist themes, for instance.

2.3 Emotion

Second, poetry is a necessarily emotional medium. When reading a poem, one experiences its *qualia* and is not necessarily focused on the actual content as much as how it makes them feel. Reading Langston Hughes can be empowering and unifying, and reading Coleridge can be mystic and take you back to the natural world. Assessing the quality of emotion thus emands the question: what does reading a ChatGPT generated poem feel like? Then, how does this feeling compare to the authentic experience?

The best way, and the way done in this paper, to answer such a question is to use emotional analysis vectors to determine the emotions being used in the poems. What will be used here to convert from text to emotion is the NRC Word-Emotion Association Lexicon (or EmoLex) from Mohammad and Turney[4], which will allow us to compare the non-GPT to the GPT generated poems.

Although subjectively, there is much more emotional depth behind real poetry, it would make sense that this more abstract emotion will not register as highly with EmoLex which tags by more common emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). Going the other way, it would make sense for ChatGPT to overplay simple emotions at the expense of more complex ones, and thus have a higher score with EmoLex.

Note that EmoLex used here is a comparative tool, not an absolute one. This means that on its own, a corpus analysed under this lens does not hold much water. It is only when it is compared to another corpus that the disparity in emotion is noticed, giving us a result that is significant statistically to indicate if ChatGPT can replicate emotion.

3 Data Collection

Specifications on the data being used and how it was acquired

3.1 "Real" Poems

The real poems, those by Shakespeare, Coleridge, Yeats, Hughes, Plath, and Ginsberg were graciously loaned from the .txtlab at McGill University. They were chosen as representative of the aforementioned poetic movements and because of their varying backgrounds, but also because they were so prolific and therefore there is a lot of data to analyse. Here is a poem and word count for all of the following authors:

- William Shakespeare, 154 sonnets, 18319 words

- William Butler Yeats, 360 poems, 68638 words
- Langston Hughes, 629 poems, 93371 words
- Sylvia Plath, 315 poems, 74664 words
- Allen Ginsberg, 476 poems, 181079 words

Each author has their own file which is a collection of .txt files containing their poems. Note that for Shakespeare, the data was taken in as one large .txt file of all of his sonnets and then a program was created by ChatGPT to separate that large file into 154 different ones to match the other authors.

3.2 "Fake" Poems

The fake poems, those written by ChatGPT (specifically GPT 3.5 Turbo) were generated using 100 queries to its API using the prompt "Write me a poem in the style of [AUTHOR'S NAME]". This was fully automated using code inspired by ChatGPT. Thus, for each real poet, there are 100 pseudo-poems written in their image. Because ChatGPT is non-deterministic, the same prompt should generate a slightly different poem every time[6], giving us 100 unique poems if all goes well. This paper is necessarily comparative, so it would have been better to have generated about the same amount of words by ChatGPT as for the real poets; however, given the cost of querying the API, this was decided against.

The poem and the word count from ChatGPT is as follows:

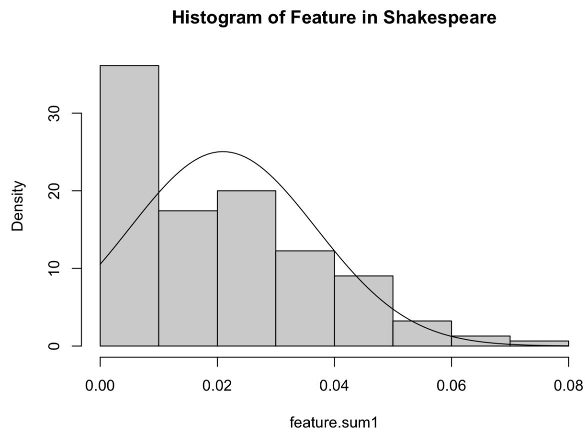
- Pseudo-William Shakespeare, 100 sonnets, 11692 words
- Pseudo-William Butler Yeats, 100 poems, 12139 words
- Pseudo-Langston Hughes, 100 poems, 12461 words
- Pseudo-Sylvia Plath, 100 poems, 12316 words
- Pseudo-Allen Ginsberg, 476 poems, 12435 words

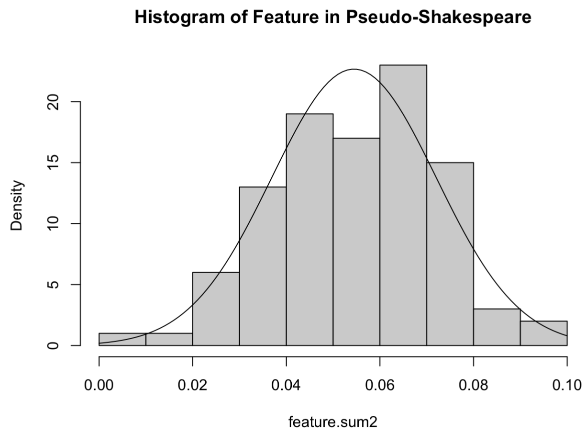
4 Measurements

What the data says

4.1 Semantics, cont'd.

Shakespeare versus Pseudo-Shakespeare
Since this dataset is made up entirely of sonnets, it makes sense to test the semantic differences between the Bard (not Google's version) and ChatGPT using words of love. Here, the list selected was as follows: "beauty", "love", "like", "feel", "want", "desire", "crave", "hair", "eyes", "smile", "whisper", "youth", and "lady". Below are the two graphs representing the feature selection for real Shakespeare and Fakespeare.



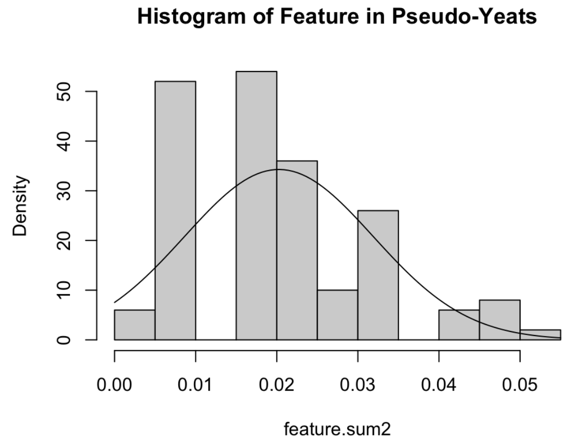


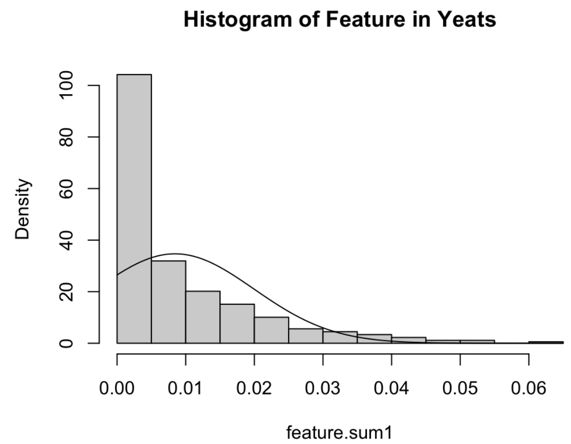
According to a Wilcoxon rank-sum test with continuity correction, GPT-generated sonnets use significantly more “love” words than real sonnets in our sample of poetry (p = 0.01754386). We found that the median value for ChatGPT (0.05882353) was 236% higher than the median value for Shakespeare (0.01754386) resulting in an overall increase of 0.4127967 or roughly 6 “love” words per sonnet (assuming a sonnet is around 150 words).

Clearly, GPT overestimates the amount of love in Shakespearean sonnets, filling them to the scenes with romantic language even if the original source material was much less inclined to do so.

Yeats versus Pseudo-Yeats Yeats here acts as a representative of the modernist movement, which sought to emphasize the natural world and the supernatural which lies beneath it. In addition, he was a nationalist Irishman, which factored heavily into his work. Thus, his word list is: “ireland”, “myth”, “love”, “dream”, “spirit”, “heart”, “imagine”, “world”, “vision”.

We can see from the following graphs that ChatGPT used these words much more frequently, directly referencing the modernist philosophy. According to a Wilcoxon rank-sum test with continuity correction, GPT-generated poems use significantly more modernist words than real poems in our sample of poetry (p-value ¡ 0.00000000000000022). We found that the median value for ChatGPT (0.0171677) was 338% higher than the median value for Yeats (0.01754386) resulting in an overall increase of 0.01327665. or roughly 2 more modernist words per poem (assuming a poem is around 150 words).

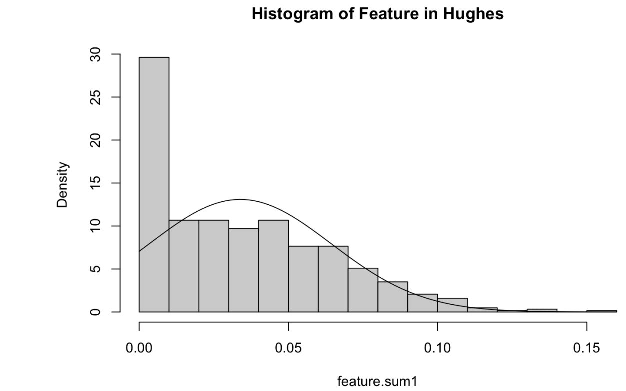


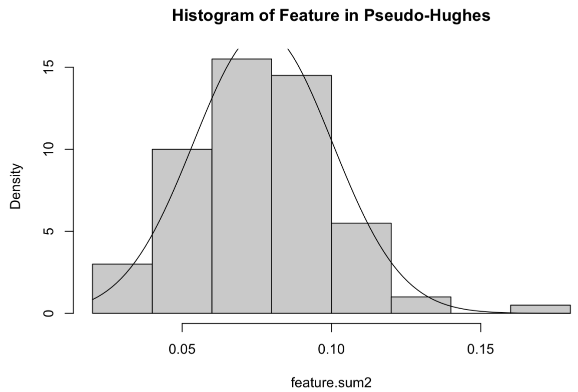


Hughes versus Pseudo-Hughes Hughes is the most notable member of the Harlem Renaissance, a mid-twentieth century rebirth of poetry among racialised groups in New York City. Here, the focus of their poetry was about Black struggle against oppression and the glimmer of a new day in America almost shining through the clouds. They emphasized unity and organisation as well, and thus the word list is as follows:

“black”, “white”, “jazz”, “harlem”, “america”, “jim”, “crow”, “race”, “negro”, “color”, “colored”, “city”, “jazz”, “rhythym”, “i”, “we”, “us”

Here, the idea is that Hughes focused around race, urban life, jazz music, and solidarity (the "I", "we", and "us" in the list). Below are the two graphs representing the feature selection for Hughes and fake Hughes.





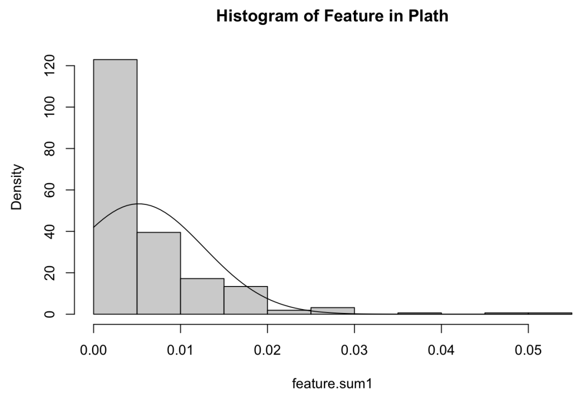
According to a Wilcoxon rank-sum test with continuity correction, GPT uses significantly more of the listed words than the real Langston Hughes in our sample of poems (p < 0.00000000000000022). We found that the median value for GPT (0.07874016) was 176% higher than the median value for the authentic poetry (0.02851336) resulting in an overall increase of 0.0502268 or roughly 8 listed words per 150 word poem.

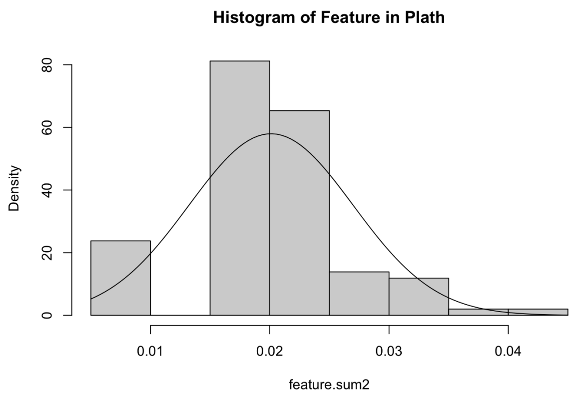
Keeping the trend, ChatGPT vastly overestimates how much Langston Hughes and other Harlem Renaissance authors discuss explicit themes about race in their texts, opting to do so more directly and less abstractly as the real Hughes would have done.

Plath versus Pseudo-Plath Sylvia Plath is a Confessional poet who wrote very deeply about her struggles with mental illness, relationships/sexuality, and death, all major themes across her work. She is the only female poet on this list, and thus occupies a uniquely gendered niche. As such, the word-list chosen is as follows:

"drugs", "pain", "dead", "death", "ill", "sex", "she", "mind"

Her attempts at curing her own mental illnesses with electroconvulsive therapy would prove futile, and she ended up committing suicide in 1963. The words were chosen to express this, as well as her sexuality with "sex" and "she".





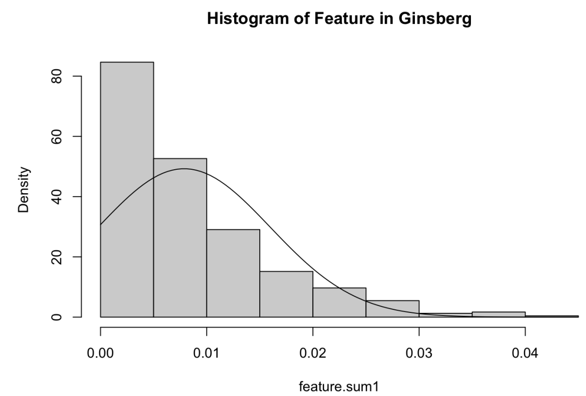
According to a Wilcoxon rank-sum test with continuity correction, GPT uses significantly more of the listed words than the real Plath in our sample of poems (p-value < 0.00000000000000022). We found that the median value for GPT (0.01818182) was 723% higher than the median value for the authentic poetry (0.00220291) resulting in an overall increase of 0.01597891 or roughly 2.5 listed words per 150 word poem.

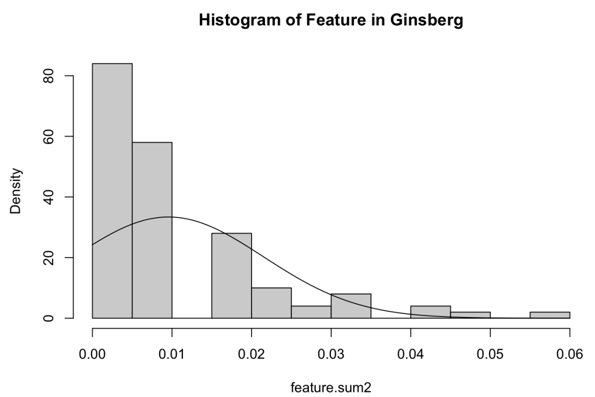
In this case too, ChatGPT exaggerates, quite aggressively how obvious the poems are.

Ginsberg versus Pseudo-Ginsberg Allen Ginsberg is the picture of the sixties counter-cultural movement, anti-war, pro-drug, and openly gay. The Beat generation, as his contemporaries have been called, write about their lived experience in New York City and discuss these themes directly in their texts, at least more directly than any poets previously. The word list to discuss the Beats is as follows:

"drugs", "mind", "city", "war", "peace", "violence", "high", "street", "light", "gay", "violence", "police"

They were psychedelic and heavily political, and these general statements in regards to both aspects should cover what the Beat generation was through Ginsberg's work. As we can see in the following graphs and data analysis, ChatGPT was more open about these basic terms like "war" and "gay", but not as much so as in the previous poets mentioned.



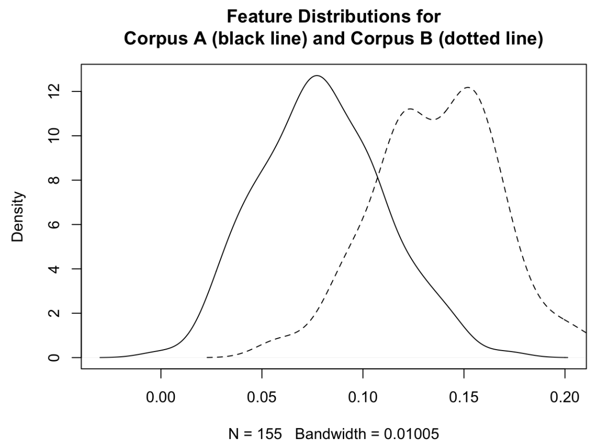


According to a Wilcoxon rank-sum test with continuity correction, GPT uses moderately more of the listed words than the real Ginsberg in our sample of poems (p-value = 0.9522). We found that the median value for GPT (0.007843258) was 26% higher than the median value for the authentic poetry (0.006285355) resulting in an overall increase of 0.001557903 or roughly a third of a listed word per 150 word poem.

Ergo, while GPT is relatively accurate in terms of content, Ginsberg's clear political messaging also allows the large language model to grasp on to easy concepts to replicate and it does so with only slight exaggeration.

4.2 Emotion, cont'd.

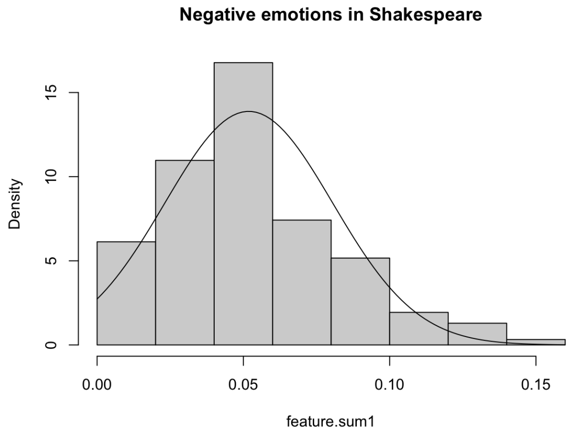
Shakespeare versus Pseudo-Shakespeare For positive emotions, as in the following graph, Fakespeare is about doubly as likely to be explicitly positive in its sonnets than the real Shakespeare.

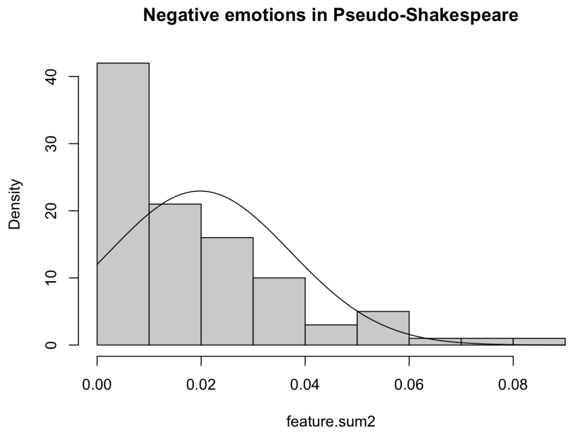


Since the data was normally distributed for Shakespeare (p=0.6999) and ChatGPT (0.8983), the means can be compared. ChatGPT's mean (0.13708403) was about 74% greater than Shakespeare's mean (0.07958572).

This shows that, for positive emotions, GPT overcompensated and went too far.

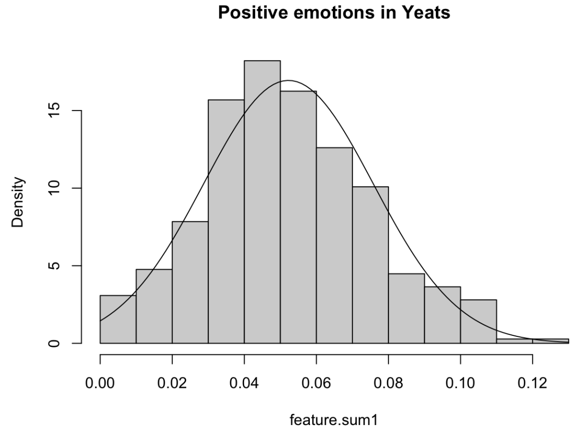
For negative emotions, a similar but opposite trend held. Here, the data was not normally distributed, but a Wilcoxon rank sum test was performed which indicated that Shakespeare uses significantly more negative emotions than ChatGPT in sonnets (p-value ¡ 2.2e-16). Here, the median for Shakespeare (0.04615385) was about 172% higher than the median for ChatGPT (0.01694915), demonstrating that ChatGPT saw sonnets as purely positive expressions but Shakespeare's actual corpus was much more nuanced.

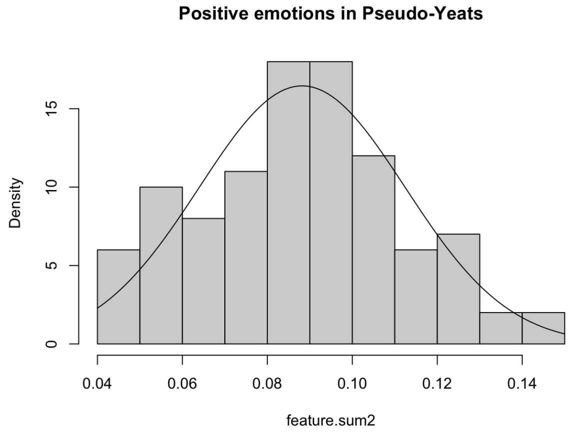




Yeats versus Pseudo-Yeats For Yeats, positive emotions are also over represented in ChatGPT's impersonation, more than half as often. While GPT's data was normally distributed (p=0.3449), Yeats' was not (p=0.01416).

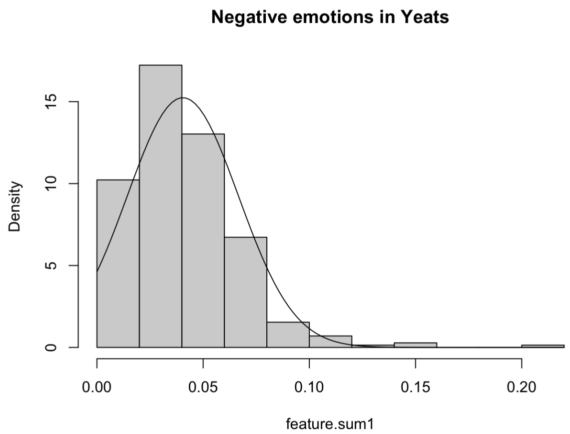
Thus, the median values should be taken. According to a Wilcoxon rank-sum test with continuity correction, GPT uses moderately more positive emotions than the real Yeats in our sample of poems (p-value ¡ 2.2e-16). We found that the median value for GPT (0.08870968) was 56% higher than the median value for the authentic Yeats (0.5699694).

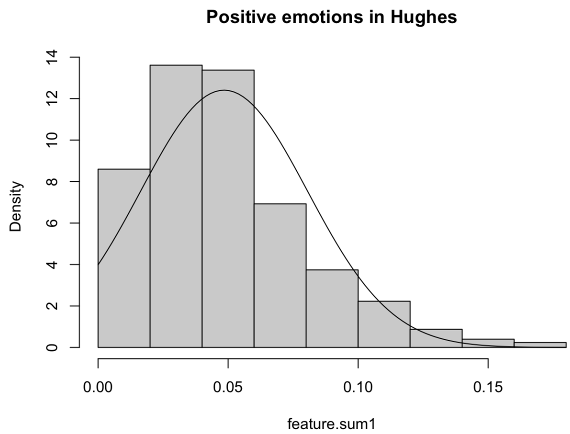


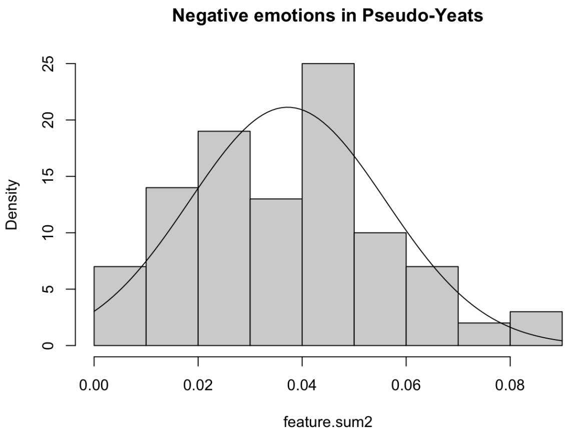


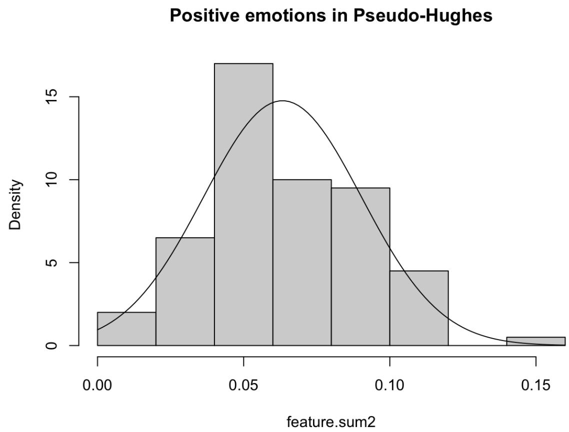
ChatGPT was more accurate when it came to negative emotions, however, but it was still not negative enough. The data was not normally distributed. According to a Wilcoxon rank-sum test with continuity correction, GPT uses only slightly less negative emotions than the real Yeats in our sample of poems (p-value = 4.396e-14). We found that the median value for Yeats (0.03773585) only was 12% higher than the median value for Yeats-GPT (0.0.03376068).

This would not translate to much more negativity than positivity in one actual poem, but it certainly makes sense given how much more positive the poems are that Yeats's work would be slightly more negative even if he was not much of a negative writer.



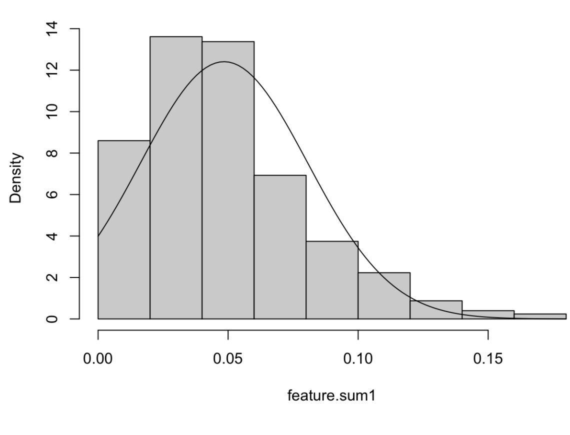


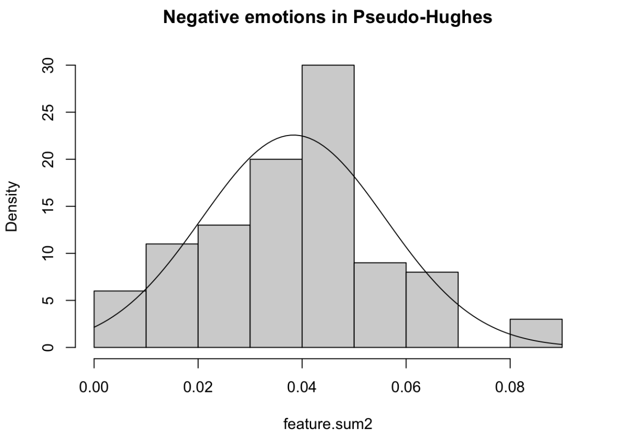




Hughes versus Pseudo-Hughes Positive emotions in GPT's Hughes are similarly enhanced, although not by so wide of a margin. Here, the data is not normally distributed. According to a Wilcoxon rank-sum test with continuity correction, GPT uses moderately more positive emotions than the real Hughes in our sample of poems (p-value = 2.017e-07). We found that the median value for GPT (0.05882768) was 22% higher than the median value for the authentic Hughes (0.04427579).

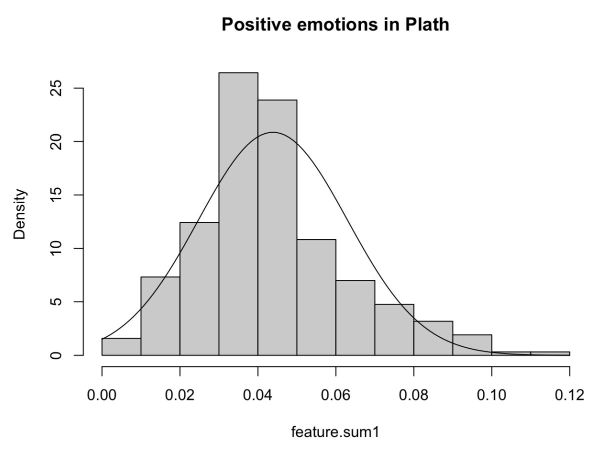
This is also true for negative emotions in Hughes, giving that Langston Hughes is much less negative in his work than ChatGPT. According to a Wilcoxon rank-sum test with continuity correction, GPT uses signifigantly more negative emotions than the real Hughes in our sample of poems (0.004133). We found that the median value for GPT (0.04016129) was 43% higher than the median value for the authentic Hughes (0.02807778).

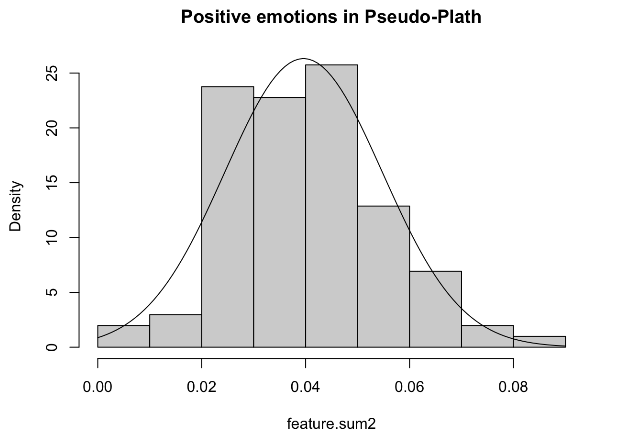




Overall, this paints the picture that ChatGPT here expresses emotions more clearly in both the positive and negative direction, here prioritising the negative perhaps in reference to themese divulged in Harlem Renaissance poetry (racism, police brutality, poverty, etc); yet, Hughes was not as obvious at all.

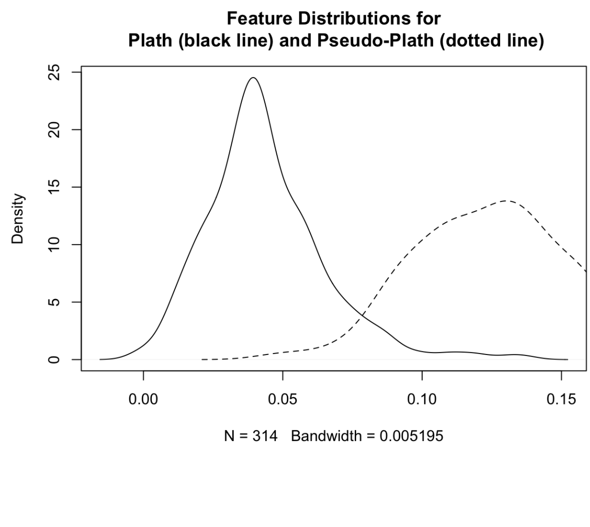
Plath versus Pseudo-Plath For positive emotions, Plath and ChatGPT were neck-in neck with Plath's median at 0.04137931 and GPT's at 0.03937008, Plath was about 5% more positive in her poems than ChatGPT would have predicted. Still, these are small amounts as Plath was not known for her happy-go-lucky attitude and kinder spirit.





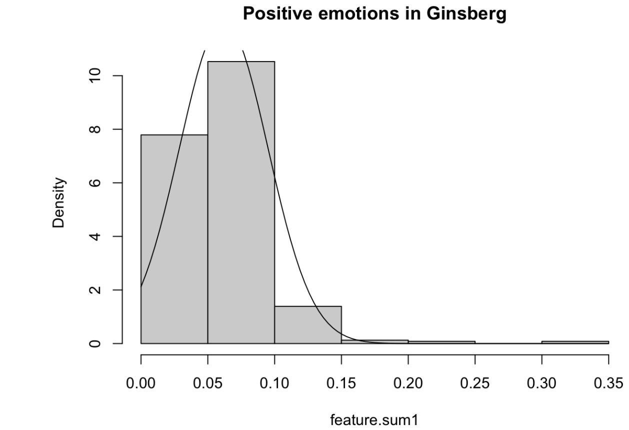
What is much more important here is how GPT handles Plath's negative emotions. Here, the expected happens. According to a Wilcoxon rank-sum test with continuity correction, GPT uses ultra-signifigantly more negative emotions than the real Plath in our sample of poems (p-value ¡ 2.2e-16). We found that the median value for GPT (0.1269841) was 207% higher than the median value for the authentic Plath (0.04079304).

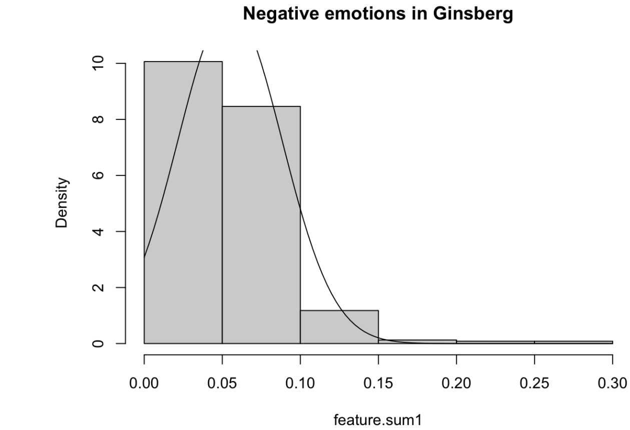
Seemingly because GPT deems Plath to be a depressing (and depressed) poet, it greatly embellishes how sad her poems would actually be, here by a factor of three. Although the following graph is not technically fair, as the data was not normally distributed, regard how vast the difference between Plath and Pseudo-Plath is here for negative emotions (further right is more negative).

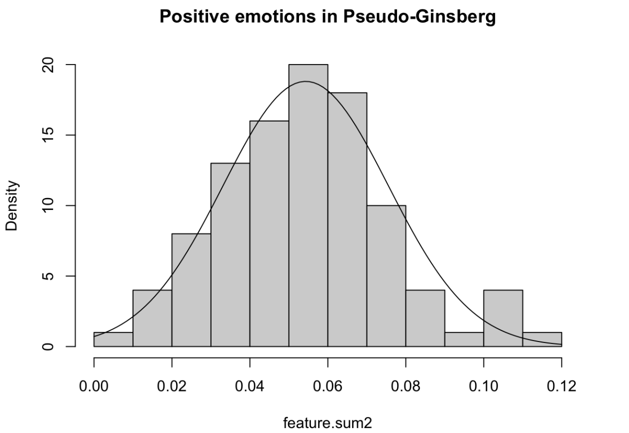


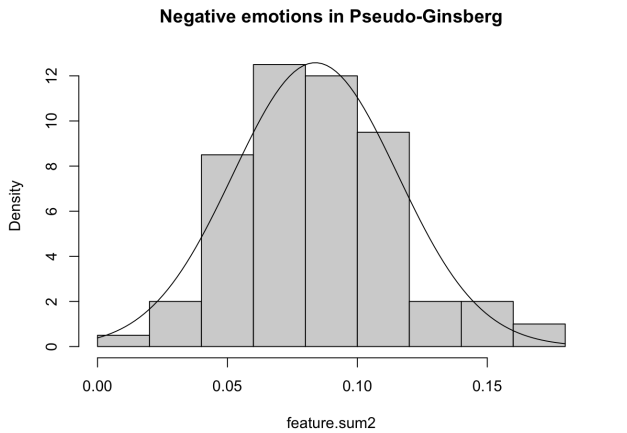
Ginsberg vs. Pseudo-Ginsberg Similarly to Plath, there is not much of a difference between Ginsberg and GinsbergGPT when it comes to positive emotions, with both the real and the faux being only lightly positive. Here, the Wilcoxon gives a p-value of 0.06857, which indicates a smaller difference which is true as there is a 4% median increase from ChatGPT (0.0546875) to Allen Ginsberg (0.05660377).

On the negative side, there is a much greater difference, indicating that ChatGPT sees Ginsberg as a very negative poet when the real man himself was not so. Here, p-value ¡ 2.2e-16 and the median of ChatGPT (0.08265027) is 60% higher than should be expected (0.05).









5 Results and Discussion

From the results, it is very clear that in terms of semantics and emotion, ChatGPT's poems differ greatly from their "real" counterparts, offering a glimpse at what lays beneath the "uncanny valley".

The running trend here is explicitness; simply put, ChatGPT is clear and direct with its messaging in poems, while the actual poets themselves would never be so obvious as to state outright their goals. Take the semantics of Shakespeare for instance. Here, we see that ChatGPT would add 6 more love words per sonnet – and remember, a sonnet is a short poem – than Shakespeare. While the real Bard would aim to use flowery language (sometimes literally) to seduce his audience, ChatGPT comes right out and says "love" or "beauty" or "lady". There is no subtelty.

This holds for all of the poets here, with HughesGPT being unrivaled in its ability to discuss racial politics outwardly and without recourse. Langston Hughes used 8 less racial words per poem, but ChatGPT cares more about replicating the cultural image of Hughes as a radical thinker than actually replicating the man himself. Plath would not write "drugs", "death", "sex", etc., but disguise these difficult concepts under metaphor and poetic device; here, ChatGPT has no such ability.

Only for Allen Ginsberg do the semantics match, although there is still disparity. It makes sense that the biggest loudmouth of an incredibly vocal and wordy movement would almost be as clear as ChatGPT, but he was still 26% less fierce.

And for emotion, of course, ChatGPT too does not understand the concept of subtlety. When the poets are lovey and happy, ChatGPT is veritably euphoric as it is with Shakespeare, becoming incredibly positive on the EmoLex scores. And when the poets are depressed and alone, as is true with Plath, it is as if ChatGPT is a baby who has lost his sucker.

The baby comparison here is apt, because ChatGPT is basically operating how an infant acquires language. It mimics, true, but it responds to stimuli in the most aggressive way. If Langston Hughes talked about racial issues and was upset by them, then ChatGPT must talk even more about racial issues and be even more (here, 43%) upset. At every single data point, across every single genre and era, this held true.

Yet, when one subjectively reads a poem by Plath, they feel deeper than when reading one by ChatGPT. They feel sadder, even if EmoLex is telling us that GPT's negative emotions are through the roof. This is the most important conclusion of this paper, that real poetry should be cherished because it is not reducible to a computation, that the intense metaphor and verbiage used is not meant to be a clear-cut message but a journey that takes reading and re-reading to truly understand.

The "uncanny" valley is a result of this effect, that one is looking at ChatGPT generated "poetry" as poetry, i.e. that it has this deeper meaning and complex emotional sphere. Really, the opposite is true. It offers not much more than a Wikipedia entry for the poet in a rhyme scheme with some lines ripped right from the source material to fill in the gaps. It has the right look to it, the right syntax, but everything else is offputting and wrong.

This is not to say that the study is conclusive. One point of error is the word-list generated, as that was done subjectively and based off of *a priori* knowledge of the source material. In a way, I was doing what ChatGPT is doing, by just throwing existing information at the problem to solve it. In a way, the high levels of association between my word list and ChatGPT's poems strengthens the argument that ChatGPT copies more from what it knows about the author than the content of their poems. However, it is still subjective and other researchers could make their own, perhaps larger, word lists to test for.

Also, in general, the poems ChatGPT generated were all at once, 100 times each. However, these models are not built for 100 such prompts and even though they are non-deterministic, it is inevitable that the same poem (maybe with a word changed) would crop up time and time again. This lowers the sample size considerably, and compares hundreds of real poems to only a handful of faux ones.

Also, despite the poor results, choosing these specific poets could have been an advantage for ChatGPT. They were chosen because they had immense data, and they had immense data because they are some of the best known ever. Could ChatGPT parrot a poem from lesser known figures, or more contemporary authors who are still writing? This question is left unanswered in this study.

Lastly, and importantly, this study avoids studying the syntax of these poems. Although it was a goal of the project, it was not realised and – if given another semester – it should be included in the analysis. Poems are necessarily syntactical, look at e.e. cummings and malfunction trying to get a computer read "the Grasshopper", but it was out of the scope of this paper. Analysing the rhyme scheme or the part-of-speech variation could be a small piece in this puzzle, and I think that I will continue with it even as the semester has come to an end. From an eye-test angle, ChatGPT tended to rhyme ABAB often, no matter the poet-to-mimic. Having a computer proove this would have been nice, and its exclusion is thus a limitation.

Overall, the results were strong to suggest that ChatGPT knows poets and cranks up their qualities to one hundred, embellishing their talking points and demeanor. While there is definitely more to investigate broadly, this first dive into A.I. generated poetry shows with confidence many of its flaws. I think that these are meaningful results, especially as educators will soon see ChatGPT generated poems fly into their gradebooks.

For knowledge, I would definitely give GPT an A+. For creativity and style, that's going to be around a C+.

References

[1] Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3, 2020.

[2] Francisco Chico Rico. *Ciberliteratura i comparatisme = Ciberliteratura y comparatismo. Prólogo.* Alicante : Biblioteca Virtual Miguel de Cervantes, 2013, 2013.

[3] Ezra Klein. The surprising thing a.i. engineers will tell you if you let them. *New York Times*, 2023.

[4] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[5] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[6] Michael Reiss. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark, Apr 2023.