Stylometric Comparison between ChatGPT and Human Essays

Sheng Wang,[1] **Nello Cristianini,** [2] **Bruce Hood** [3]

[1]School of Computer Science, University of Bristol, UK
[2] Department of Computer Science, University of Bath, UK
[3] School of Psychological Sciences, University of Bristol, UK
sheng.wang@bristol.ac.uk

Abstract

A key priority in web and social media research is to distinguish between human-generated and machine-generated content, especially in the era of widespread use of Large Language Models.

In order to be able to detect Machine-Generated Content, we compared the writing styles of essays written by human subjects and by GPT under the same conditions. Specifically, we analyzed the sentiment, readability and subjectivity of 4895 essays, written by 789 university students in response to 9 specific prompts, with 225 essays of the same length written by GPT-4 in response to the same prompts. We observed that GPT's essays are more difficult to read and less subjective. All of the above quantities depend on the specific prompt, and show a correlation between human subjects and GPT, across the 9 prompts.

Introduction

Social media is a powerful technology for the exchange of opinions and information, but is vulnerable to the risk of misinformation and disinformation(Aïmeur, Amri, and Brassard 2023), which could potentially lead to manipulations of public opinion(Epstein et al. 2022). This risk is increased by the advent of Large Language Models, which significantly eased the creation of content based on specific prompts for social media platforms. OpenAI, the creators of the GPT models, have warned about potentially malicious use of their tools for many years(Partadiredja, Serrano, and Ljubenkov 2020). An important technical question is how to distinguish machine-generated from human-generated text, (Floridi and Chiriatti 2020), for example to warn the readers, or even as a way to filter out artificial content from research studies.

Various approaches have been proposed, including some based on extracting specific textual-features (such as frequency characteristics, fluency, text-based features, etc.) and others based on neural network (such as using existing models for zero-shot classification and fine-tuning pre-trained language models) (Crothers, Japkowicz, and Viktor 2023).

Stylometry, been recognized as a viable tool for authorship attribution for a long time(Holmes and Kardos 2003), has untapped potential in detecting these subtle stylistic differences. Therefore, we propose a novel method based on stylometry aimed at automatically detecting machine-generated content by analyzing and comparing stylistic features of texts. The research question we consider is whether there may be distinguishable differences at the stylistic level between machine-generated and human-generated texts, even though Large Language Models (such as GPT-4) can reach comparable quality to human-generated text.

We compare the writing of GPT-4 with that of human subjects under controlled conditions, so that we can account for topic, and isolate the effects of style.

By style we mean all the choices made by an author, besides the actual content, which can form their voice: word choice (simpler or more complex?), subjective vs objective (do they present facts neutrally, or express personal judgments and impressions?), sentence length, tone (formal vs informal), readability, sentiment.

The effects of these choices on readers has been investigated, for example the presence of emotional words in Facebook posts has been shown to significantly affect the emotion later expressed by the readers, in a controversial experimental study of "emotional contagion" (Kramer, Guillory, and Hancock 2014). Choice of words can also affect understandability, persuasion, and can reveal also author identity.

Establishing how the "voice" (tone and style) of a ChatBot differs from that of humans can have important applications, both for their potential effects on users, for the risk of impersonation, and for their potential as companions, news givers, counsellors, and so on. This information could be useful also to inform the important final alignment phase of any Large Language Model.

Methods

Data Description

Human data description

The participants. A group of 789 participants was assembled, formed by students attending a university course on mental well-being (Hood, Jelbert, and Santos 2021). The students were required to keep a weekly online journal on a given topic (either related to the tasks that they were set or asked students to provide some reflection on their mental well-being over the past week). All the participants were students enrolled on the 'Science of Happiness' (SoH) course

Topic	topic1	topic 2	topic 3	topic 4	topic 5
	topic 6	topic 7	topic 8	topic 9	
Human	653	578	574	527	559
	524	512	497	471	
GPT-4	25	25	25	25	25
	25	25	25	25	

Table 1: The number of human-generated and GPT-4 generated essays for each topic. *(Topic 1: Three Good Things; Topic 2: Optimism; Topic 3: General Reflections 1; Topic 4: Gratitude Letter; Topic 5: Well-being Intervention; Topic 6: Acts of Kindness; Topic 7: Signature Strengths; Topic 8: General Reflections 2; Topic 9: Goal Setting)*

at the University of Bristol in the academic years 2020/21 and 2021/2022, with two teaching blocks each year. The 789 participants were so distributed: (Age: mean = 18.81 years, range 17-31 years, 81 students did not complete this section; Gender: 78.32% female, 19.86% male, 1.82% non-binary/ undisclosed/ prefer not to say, 74 students did not complete this section; Nationality: 78.20% British, 87 students did not complete this section; Ethnicity: 76.08% White, 11.05% Asian, 3.78% Black, 4.62% Multiple ethnic groups, 4.48% Other ethnic groups/ undisclosed/ prefer not to say, 74 students did not complete this section). This research was approved by the University of Bristol School of Psychological Sciences Research Ethics Committee(Approval code: 011020110763). Participants provided informed written consent as part of an online survey during data collection.

The Dataset. Over nine weeks these students were asked to produce a journal entry based on a given topic and prompt (there are missing journal entries, as not all students completed each task). The resulting dataset consists of 4895 valid journal entries. (Topic 1: Three Good Things; Topic 2: Optimism; Topic 3: General Reflections 1; Topic 4: Gratitude Letter; Topic 5: Well-being Intervention; Topic 6: Acts of Kindness; Topic 7: Signature Strengths; Topic 8: General Reflections 2; Topic 9: Goal Setting)

For each journal entry, we computed 1) the Positive emotion and Negative emotion scores based on LIWC2015(Pennebaker et al. 2015); 2) the readability scores based on Dale-Chall formula(Dale and Chall 1948); 3) a linguistics subjectivity score based on the SentiWordNet as described in (Baccianella et al. 2010)(Flaounas et al. 2013). The details of these computations are described below. So for each of the 4895 valid journal entries we have: topic, content, positive emotion, negative emotion, readability scores, subjectivity scores, number of unique words, avaerage word length and average sentence length.

GPT data description

As a comparison, we created a dataset by using ChatGPT 4.0, under the same conditions as the human participants. ChatGPT 4.0 is a Large Language Model created by openAI, which has the capability to generate human-like text in response to a prompt (Achiam et al. 2023). The software was prompted to generate journal entries using the same journal topics and prompts (and asked to make them of the same length as the average entry length in the human data (Topic 1: 301, Topic 2: 213, Topic 3: 210, Topic 4: 229, Topic 5: 245, Topic 6: 204, Topic 7: 176, Topic 8: 208, Topic 9: 152)). We accessed ChatGPT 4.0 through the OpenAI portal at 'https://chat.openai.com/' utilizing the ChatGPT Plus subscription plan.

The GPT dataset. We generated 25 journal entries for each journal prompt/topic, using a new session for each journal entry, in order to avoid interdependencies between essays generated in the same session (an effect known as in-context learning (Brown et al. 2020)). In this way, we obtained 225 journal entries generated by GPT4.0, for each journal entry, we computed the Positive emotion and Negative emotion score based on LIWC2015; Computed readability scores based on Dale-Chall formula and a Linguistics subjectivity score (Flaounas et al. 2013) based on the SentiWordNet. The details of the computations are described below.

LIWC scores extraction

We extracted two psychometric indicators (positive emotion or posemo, negative emotion or negemo) by using the word lists and rules from LIWC2015 (Pennebaker et al. 2015). This produced a numerical score for these two emotions in each journal entry. For each journal entry, we followed the same steps as in (Wang, Lightman, and Cristianini 2021), that is: First lowered uppercase characters and tokenized text into words (alphanumeric strings, referred to as 'tokens'); also, we removed all punctuation. Each word was compared with the rule lists of LIWC2015, incrementing the counter of each indicator when a match occurred. In the end, we normalized all counts by the total number found via tokenization which generated two numerical scores for each journal entries. In this way, we obtained two LIWC scores for each of the 4895 journal entries submitted by the students, as well as for the 225 entries generated by the GPT4.0 model.

Dale-Chall Score

Readability is an important property of writing style and is the result of a writing style that is legible, interesting and comprehensible(Zamanian and Heydari 2012). To compare the writing styles of humans and GPT, we are attempting to quantify the readability of the essays they have written. We accessed this by using the Dale-Chall formula (Dale and Chall 1948), which is a most accurate (Klare 2000) and popular standard formula in scoring text readability(Zamanian and Heydari 2012). The scores range from 0.0 to 10.0, the higher the score, the essay are more difficult to read. We applied this formula to both the humans' and GPT's journal entries and received a numerical score for each journal entry.

Linguistic Subjectivity

Subjective language refers to the use of words that describe impressions, judgments, emotions of the speaker, as opposed to objective descriptions(Quirk and Crystal 2010)(Wiebe and Riloff 2005). We follow the work of (Flaounas et al.

	Human	GPT-4
Unique Word (Mean)	124.4155	154.8222
Unique Word (SEM)	0.9049	1.7425
Average Word Length (Mean)	3.9880	4.4933
Average Word Length (SEM)	0.0040	0.0183
Average Sentence Length (Mean)	27.1644	19.7737
Average Sentence Length (SEM)	0.1305	0.1643

Table 2: Statistical comparison of vocabulary usage and sentence length in human-generated and GPT-4 generated essays

2013), using the percent of adjectives in the text that have positive or negative valence.

The proportion of subjective adjectives (those with a subjectivity score greater than 0.25) relative to the total number of adjectives.

The adjectives were found by using the Spacy library, then their sentiment was assessed via SentiWordNet (Baccianella et al. 2010),

We did this for each journal entry generated by GPT and by human subjects.

Other measures

GPT uses a richer vocabulary, longer words, but shorter sentences. (See Table 2)

Results

Sentiment Analysis

The LIWC scores for both human and GPT-generated essays were compared across nine different topics. For each topic, we computed the average scores for positive emotions (posemo) and negative emotions (negemo) for both human's essays and those generated by GPT. This resulted in two separate series of scores for posemo and negemo, each containing nine data points for both the humans and GPT.

We calculated the correlation coefficient between the score series for GPT and Humans for both posemo and negemo across nine topics, as reported it in Table 3. We can see for both coefficients indicate a significant correlation, with p-value scores for posemo (c(posemo))=0.0023 and for negemo (c(negemo)=0.0004).

The scores for positive emotion ones are slightly larger in GPT than in human (p=0.0274), non significantly larger for negative emotion (p=0.4039).

Figure 2 shows the scores for each of the topics (error bars indicate two standard errors of the mean).

Writing Style

We use the readability and linguistic subjectivity as two indicators of writing style. As described on the method part, we reported the average Dale-Chall scores and linguistic subjectivity score on each topic for humans and GPT on Figure 3 and Figure 4.

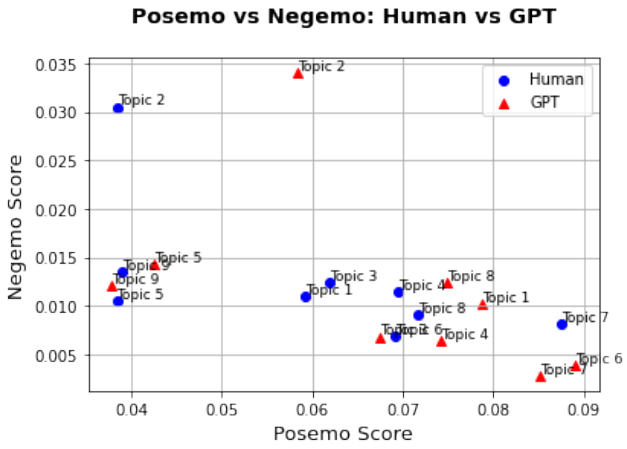

Figure 1: Comparison of topics based on their sentiment (Posemo vs Negemo), for humans (Blue) and GPT (Red). *Topic 1: Three Good Things; Topic 2: Optimism; Topic 3: General Reflections 1; Topic 4: Gratitude Letter; Topic 5: Well-being Intervention; Topic 6: Acts of Kindness; Topic 7: Signature Strengths; Topic 8: General Reflections 2; Topic 9: Goal Setting)*

Scores	Statistic	P-value
LIWC (Posemo) Score	0.868798	0.002364
LIWC (Negemo) Score	0.923459	0.000379
Dale-Chall Score	0.867590	0.002438
Linguistic Subjectivity	0.846760	0.003979

Table 3: Pearson Coefficient Correlation between humans and GPT on scores, topic by topic.

Figure 3 displays the average Dale-Chall scores of essays by humans and GPTs across various topics, with 2SEM error bars indicating whether the average score for each topic can represent the scores of the population under that topic. From this figure, we can observe that for all topic, the average Dale-Chall scores of GPTs are higher than those of humans, suggesting that the essays generated by GPTs are more difficult to read than those written by humans. We also calculated the correlation coefficient between the Dale-Chall score series for GPT and humans. The value reported in Table 3 indicates a significant correlation between these two score series.

For each essay, we also calculated their linguistic subjectivity scores using the defined method on last chapter, as shown in Figure 4. The error bars (2SEM) reveal that, compared to humans, the score range for GPTs in each topic is more dispersed. It is also shown that human's essays score higher in linguistic subjectivity than GPTs' essays, suggesting that human are more likely to use subjective adjectives. Similar to the Dale-Chall score, we calculated the correlation coefficient for these two score series, and the result (reported in Table 3) shows that they are significantly correlated.

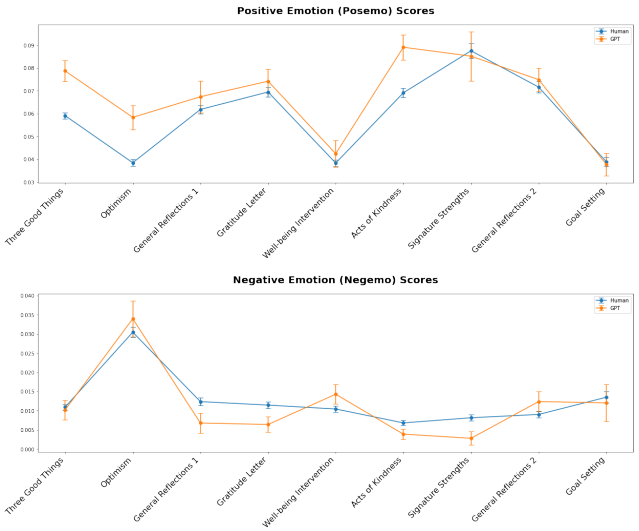

Figure 2: The Average Score of LIWC2015 indicators (Posemo/ Negemo) for each topic, for humans(blue) and GPT (orange). Confidence intervals represent two standard errors of the mean (2SEM).

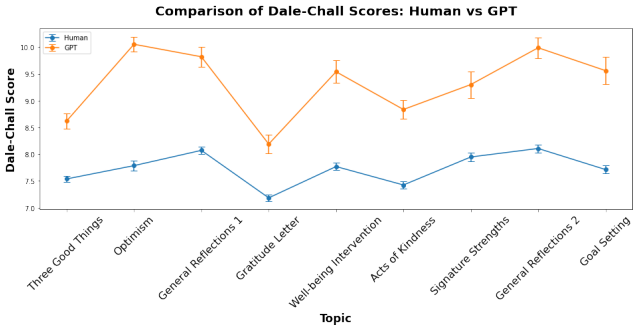

Figure 3: The Average Dale-Chall Scores for each topic, for humans(blue) and GPT (orange). Confidence intervals represent two standard errors of the mean (2SEM).

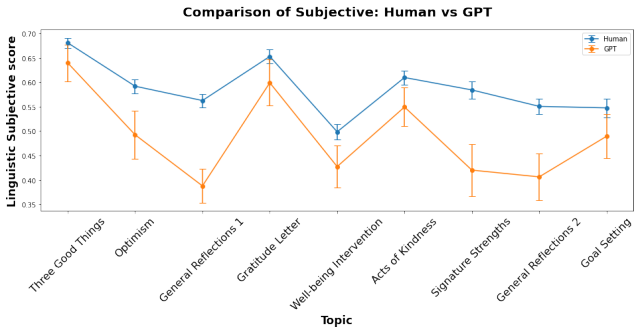

Figure 4: The Average Linguistic Subjective Score for each topic, for Humans(blue) and GPT (orange). Confidence intervals represent two standard errors of the mean (2SEM).

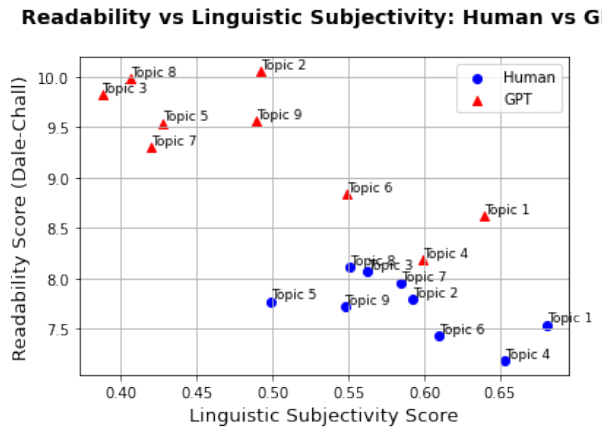

Figure 5: Comparison of writing style, across essays of different topic, based on Readability vs Linguistic Subjectivity, for humans (blue) and GPT (red). *(Topic 1: Three Good Things; Topic 2: Optimism; Topic 3: General Reflections 1; Topic 4: Gratitude Letter; Topic 5: Well-being Intervention; Topic 6: Acts of Kindness; Topic 7: Signature Strengths; Topic 8: General Reflections 2; Topic 9: Goal Setting)*

We plotted the readability and linguistic subjectivity scores by topic on a scatter plot (Figure 5), helping us to visualize different topics on both scales. From this figure, we can observe an interesting fact: when the readability of essays go higher, their scores for subjectivity also tend to increase. Furthermore, this visual representation reinforces our previous findings: essays written by humans exhibit better readability than those generated by GPTs, and they also score higher on subjectivity.

Conclusion

Finding and Contribution

The writing style of GPT-4 was compared with that of human subjects under controlled conditions, and we observed significant stylistic differences: in readability, subjectivity, but not in sentiment. This suggests that more research in stylometric analysis of LLM-produced text might lead to useful automated filters.

We believe that this line of research can lead not only to methods to filter machine-generated content, but also it could play a crucial role in the alignment and auditing of chatbots based on Large Language Models , a step which involves managing risks of impersonation, persuasion, and contagion. This is also a step towards developing a machine-generated text detector by evaluating various text-based feature. Finally this raises the possibility of using chatbots as control groups (to generate baseline data or null hypothesis) for assessing human subjects.

Future Work

Our dataset for this study is based on essays generated by students and those generated by GPT. For each topic, we have created 25 essays written by GPT. We plan to expand the dataset of GPT-generated essays in our future work and will introduce more comparative data, such as tweets writ-

ten by humans and those generated by bots. Incorporating different Large Language Models (LLMs) for comparison is also one of our objectives which could help us to establish a baseline.

Additionally, in the methods section, we will employ more metrics (such as John Burrows' Delta Method) to achieve a more comprehensive system. At the same time, we will also complete the use of psychometric testing of GPT.

Acknowledgements

We appreciate Catherine Hobbs for her significant contributions to the dataset used in this study.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aïmeur, E.; Amri, S.; and Brassard, G. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1): 30.

Baccianella, S.; Esuli, A.; Sebastiani, F.; et al. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, 2200–2204.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Crothers, E.; Japkowicz, N.; and Viktor, H. L. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*.

Dale, E.; and Chall, J. S. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, 37–54.

Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 183–193.

Flaounas, I.; Ali, O.; Lansdall-Welfare, T.; De Bie, T.; Mosdell, N.; Lewis, J.; and Cristianini, N. 2013. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital journalism*, 1(1): 102–116.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Holmes, D. I.; and Kardos, J. 2003. Who was the author? An introduction to stylometry. *Chance*, 16(2): 5–8.

Hood, B.; Jelbert, S.; and Santos, L. R. 2021. Benefits of a psychoeducational happiness course on university student mental well-being both before and during a COVID-19 lockdown. *Health Psychology Open*, 8(1): 2055102921999291.

Klare, G. R. 2000. The measurement of readability: useful information for communicators. *ACM Journal of Computer Documentation (JCD)*, 24(3): 107–121.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National academy of Sciences of the United States of America*, 111(24): 8788.

Partadiredja, R. A.; Serrano, C. E.; and Ljubenkov, D. 2020. AI or human: the socio-ethical implications of AI-generated media content. In *2020 13th CMI Conference on Cybersecurity and Privacy (CMI)-Digital Transformation-Potentials and Challenges (51275)*, 1–6. IEEE.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Technical report.

Quirk, R.; and Crystal, D. 2010. *A comprehensive grammar of the English language*. Pearson Education India.

Wang, S.; Lightman, S.; and Cristianini, N. 2021. Effect of the lockdown on diurnal patterns of emotion expression in Twitter. *Chronobiology International*, 38(11): 1591–1610.

Wiebe, J.; and Riloff, E. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International conference on intelligent text processing and computational linguistics*, 486–497. Springer.

Zamanian, M.; and Heydari, P. 2012. Readability of Texts: State of the Art. *Theory & Practice in Language Studies*, 2(1).