Literary performance of pre-trained large language models: a case study of *Pharmako-AI*

Anonymous ACL submission

Abstract

This study investigates the literary performance of hybrid text co-created by a professional author and a large language model (LLM), exemplified by the novel *Pharmako-AI* by K Allado-Mcdowell (2020). While pre-trained LLMs have gained particular popularity for their literary output such as poetry, recent studies suggest potential shortcomings in LLMs' literary performance. Benchmarking and assessing LLMs are important, because they improve LLMs' interpretability and transparency in various aspects, including literary performance. Computational linguistic assessments for hybrid text are relatively understudied, which motivates this study. From NLP perspective, we aim to characterize and evaluate the collaboration between a human writer and an LLM. Drawing on previous investigations of LLMs' creative capabilities, our study utilizes sentiment and similarity analyses to analyze *Pharmako-AI*'s literary aspects, comparing human and LLMs generated text. Findings reveal nuanced sentiment patterns and higher sophistication in human-generated text. This indicates the distinctive nature of the hybrid text relative to a traditional work of fiction. Our study contributes to understanding the interplay between human and LLM contributions in literary collaboration.

1 Introduction

Pre-trained large language models (LLMs) are progressing rapidly. LLMs' literary performance has been impressive, as indicated by the popularity and prevalent usage of ChatGPT. Users are particularly impressed by LLMs' ability to produce literature. Despite the seemingly creative and well written text, interpretability studies show that under careful scrutiny, LLMs generated text lacks good literary performance (Chakrabarty et al., 2023; Yang et al., 2022). What about hybrid literary text co-created by a professional author and a LLM? As far as our knowledge goes, there is no NLP assessment of such literary text. To bridge the gap, this study explores approaches to characterize and evaluate hybrid text, through the lens of the novel *Pharmako-AI* by Allado-McDowell (2020). The goal is to measure and index literary performance of pre-trained LLMs, and to provide an approach that can potentially tease apart superficial fluency from deep cohesion. By "literary performance", we are referring to the LLMs' production of text that is meant to be literary, and we are evaluating by way of the chosen metrics (sentiment and similarity) to what extent the LLMs text is successful in coming across as literary, for example, emotional and character development, originality, and repetition.

Previous studies provide critical views on the extent to which LLMs generated text is "creative" or "coherent". Yang et al. (2022) suggest absence of long-range coherence in GPT3 generated stories. Chakrabarty et al. (2023) recruited creative writers to assess stories using creativity tests. The stories were written either by professional authors or LLMs. Results indicate that LLMs' written stories are not as fluent or original as they appear to be at first glance, and they are not as creative as those written by professionals.

In addition to NLP researchers assessing literary performance in LLMs, literary scholars, especially digital humanity researchers, utilize NLP methods such as sentiment analysis to analyze literary texts. Sentiment analysis uses computational techniques to analyze the emotional tone expressed in text, categorizing it as positive, negative, or neutral. Sentiment analysis provides a systematic way to explore the emotional dimensions of literature, enhancing traditional qualitative analyses with quantitative insights. Elkins and Chun (2019) applied sentiment analysis to Virginia Woolf's *To the Lighthouse*, a famous modernist novel. Their results suggest that NLP approaches indicated efficacy in accounting for literary analysis. They developed a hybrid model leveraging both NLP tools and close reading. This model leads to innovative findings

that can advance the field of modernist and narrative studies. In this line of investigation, Elkins and Chun (2020) examine the question of whether GPT3 can pass a writer's Turing test. They discuss both the impressive achievements and limitations of GPT3, highlighting the importance of understanding its capabilities and limitations.

Inspired by and built up on Elkins and Chun (2019), we use sentiment and similarity methods, to capture literary performance of *Pharmako-AI*. According to a podcast about the creation process of this book (Allado-McDowell, 2021), *Pharmako-AI* is a collaboration between the professional writer and artist K Allado-Mcdowell and a pre-trained LLM GPT3. Approximately over 60% of the text is generated by GPT3, with occasional instruction and revision by the human writer. The book is completed through K Allado-Mcdowell interacting with GPT3 using an iterative process of writing prompts and probing GPT3 to generate responses. The prompt composed by K Allado-Mcdowell is in a more literary text form than a plain prompt would be. It is a unique book because of the way it is created: human writer and GPT3's texts are integrated, with GPT3 not only responding to a prompt but also influencing the human writer's writing. We hope to provide an approach to analyze such text from NLP perspectives, discovering how the interplay of human and LLMs develops in literary collaboration.

2 Experiments

2.1 Datasets

We accessed the text files of *Pharmako-AI* from a digital library Xenotheka (Roman, 2022) and those of *Frankenstein* from Sweigart (2012). *Pharmako-AI* is a novel exploring philosophical questions, technology, and ecology. We selected seven chapters: chapter five *post-cyberpunk* through chapter eleven *the poison path*. This is because the other chapters involve poetry or poetic forms, where repetitive text is prevalent. This creates confounding variables in similarity computation. Also, we intend to focus on non-poetry form, keeping our comparisons as feasible and controlled as possible. We chose a traditional work of fiction *Frankenstein* as a baseline to establish similarity interpretations. Both the selected *Pharmako-AI* chapters and *Frankenstein* are classified as fiction in terms of literary form, and both belong to the novel genre.

2.2 Sentiment analysis

We used the Syuzhet package in R (Jockers, 2015) for sentiment analysis. We first compared the shape of sentiment trajectory over narrative time in GPT3 and professional author created texts. We represented texts as sentiment vectors, and used the simple_plot() function that applies three smoothing methods: loess, a moving average, and discrete cosine transformation (DCT). According to Jockers (2015), Elkins and Chun (2019), and Chun and Elkins (2023b), smoothing is an approach to extract literary meaningful signals from noisy text data. Loess (LOcal regrESSion) is particularly effective in capturing local variations and trends in the data. Rolling mean, or a moving average, generates a curve through sliding a fixed size window over each sentence's sentiment values, and assigns an average at the moving window's center. DCT is a technique widely used in signal processing to better represent extreme and edge values in the smoothed sentiment vectors. We applied these smoothing methods because when input texts are large, a percentage based measure would water down the emotional valences' maximums and minimums. We hope to reveal the smoothed sentiment development trajectories in human and GPT3 generated texts.

2.3 Similarity analysis

To establish an interpretation baseline, we additionally analyzed a literary text completely created by a professional author: *Frankenstein* by Mary Shelley. We used an OpenAI model *text-embedding-ada-002* to represent *Pharmako-AI* and *Frankenstein* text as embedding vectors, and computed cosine similarity of adjacent sentences vectors. We propose that hybrid text is more likely to have repetition or incoherence than traditional literary text. Reflected in similarity, extremely high and extremely low similarity scores are more likely to be found in a hybrid text than in a human written literary text. Therefore, we hypothesize that relative to similarity in literary texts that are totally created by a human author, hybrid text such as *Pharmako-AI* should show high similarity variance.

3 Results and discussion

3.1 Sentiment analysis findings

Results suggest interesting patterns in GPT3 generated text, professional writer created text, and their interactions. Fig 1. shows two plots stacked, illustrating sentiment trajectory in *Pharmako-AI*

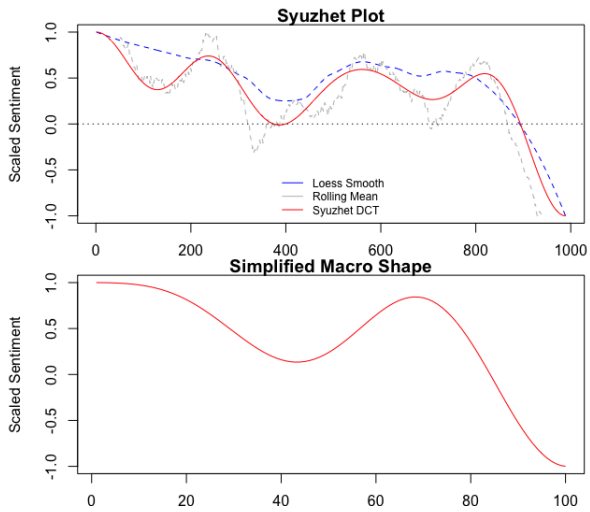

Figure 1: Sentiment plot of *Pharmako-AI* text generated by GPT3

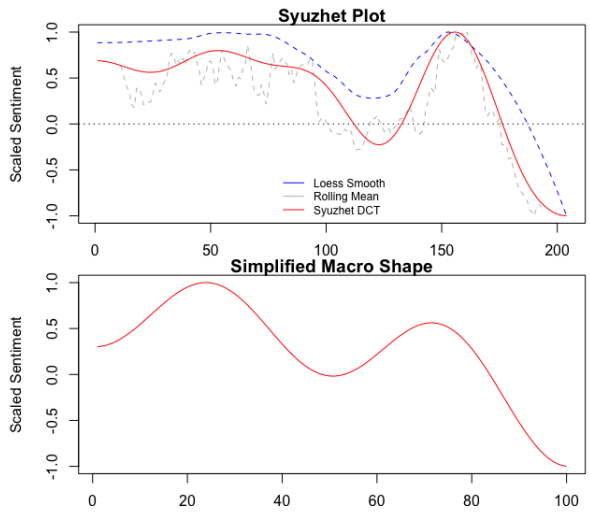

Figure 2: Sentiment plot of *Pharmako-AI* text generated by the professional author

texts generated by GPT3. The x-axis represents "narrative time", where longer time indicates an incrementally larger amount of text. The y-axis represents scaled sentiment. The top graph shows all three smoothing methods on the same graph. The bottom shows only the DCT smoothed line, with a normalized narrative time axis. Fig 2. illustrates the sentiment trajectory over narrative time in *Pharmako-AI* texts generated by the professional author. The same smoothing methods were applied.

First, the top plot in Figure 1,2 shows that the three superimposed lines are mostly consistent in trends, suggesting robust sentiment plot development patterns across different smoothing methods. This provides independent supportive evidence for Elkins and Chun (2019) that NLP methods have potential in informing literary analysis. Second, for the bottom plot, with normalized text length, sentiment shape shows two salient emotion peaks in professional author generated text (Figure 2, narrative time 25 and 70), but one salient peak in GPT3 generated text (Figure 1, narrative time 70). This indicates higher sentiment sophistication and complexity in human professional's text than in GPT3's.

Figure 3 shows a narrative development comparison between texts generated by GPT3 and the professional author. Given the difference in length between human and GPT3 generated texts, we divided each text vector into 31 samples to control the length variable. As a consequence, each line in Figure 3 is derived from equal amount of data (31 sampled points) on the x-axis. There are three observable gaps at narrative time 0.4, 0.6, and 0.8,

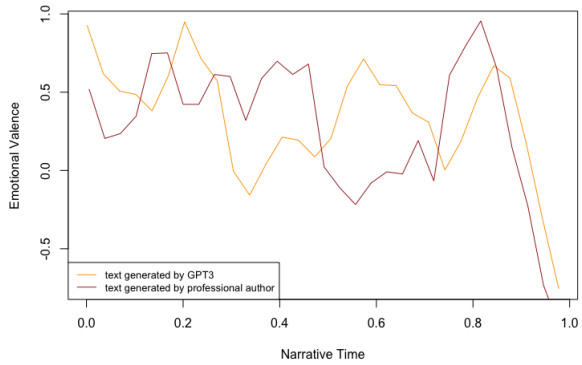

Figure 3: Sentiment development plot of *Pharmako-AI*

in addition to small gaps. This indicates that with respect to emotional valence, there are multiple discrepancies between a pre-trained LLM and a human author. With equal sample size, we constructed a Pearson's correlation model to quantify the comparison. Results suggest that there is no significant correlation in emotional valence between GPT3 generated and professional author written texts ($p = 0.37$, $r = 0.17$, 95%CI [-0.2, 0.49]). This implies great derailment and circumstantiality in pre-trained LLM literary performance.

Our finding aligns with Chakrabarty et al. (2023): the key characteristics that led the experts to believe that a story is written by LLMs include, for example, struggles at narrative ending, inadequate narrative conclusions, and poor character development, where a character would disappear abruptly. We speculate that the aforementioned features are likely to trigger the three noticeable gaps and mis-

Figure 4: Cosine similarity comparison of *Pharmako-AI* and *Frankenstein*

alignment in emotional valence between GPT3 and human written texts.

3.2 Similarity analysis findings

Descriptive statistics showed that standard deviation of similarity in *Frankenstein* is 0.03, and 0.05 in *Pharmako-AI*. Similarity in *Pharmako-AI* showed more variance than that in *Frankenstein*. This is as predicted, although the difference seems marginal. Interestingly, mean similarity in *Frankenstein* is 0.79, and 0.83 in *Pharmako-AI*. Figure 4 visualizes the distribution and the overall trend.

Pharmako-AI showed generally higher similarity than *Frankenstein*. Our interpretation is that higher similarities in *Pharmako-AI* are likely to be attributed to larger amount of perseveration and repetition in *Pharmako-AI*'s GPT3 generated text. This is in line with Breithaupt et al. (2024), which argues that human retelling of stories is notably more creative and less repetitive than a probabilistic LLM. Additionally, our finding can be explained by Chakrabarty et al. (2023), which suggested that LLMs generated text showed lower originality, fluency, and elaboration, compared to text created by professional writers. Reflected in similarity, it is possible that these attributes give rise to higher similarity in *Pharmako-AI* GPT3 generated text.

4 Limitation and future research

In this study, we used sentiment analysis to capture emotion development in *Pharmako-AI* GPT3 generated text, human professional created text, and their interplay. We additionally computed cosine similarities of adjacent sentences in *Pharmako-AI* and *Frankenstein*, assessing similarity of hybrid text, relative to the science fiction classic *Frankenstein*. Our findings suggest that sentiment analysis tools developed in NLP can reveal robust sentiment patterns in literary text. For a hybrid book such as *Pharmako-AI*, human generated text showed higher sophistication than GPT3 created text. *Pharmako-AI* showed larger similarity variance and higher similarity mean than *Frankenstein* did.

We acknowledge that quantifying literary texts is challenging. It becomes even more complex when it comes to hybrid text co-written by human professionals and pre-trained LLMs. With the rapid development of LLMs, we maintain that literary books such as *Pharmako-AI* will be increasingly common. Collaboration between artists and LLMs may become closer and closer. There is a need in both computational linguistics and in digital humanities to benchmark and analyze hybrid text. Our investigation can inform LLMs in terms of where they need improvement. We also hope that our exploration can inspire literary scholars to incorporate NLP and LLMs into literary analyses.

We titled this study as measuring literary "performance" but not "competence" of pre-trained LLMs. We take this preliminary study as an initial step of an incremental project. We do not intend to provide a literary review for the book *Pharmako-AI* or to characterize how competent and enjoyable it is for readers to appreciate as a literary work. We leave that for professional literary scholars and literary reviews. Instead, we focus on assessing LLMs' behavioral performance using well-established NLP methods, such as sentiment analysis and cosine similarity. The intention is to introduce literary meaningful NLP methods to analyze LLMs-generated literary text.

Moreover, we acknowledge that a case study about one particular hybrid text is limited and might not be representative. To expand and generalize sentiment analysis to broader digital humanities research areas, such as narrative generation and diachronic text sentiment analysis, we plan to utilize the greybox ensemble framework developed by Chun and Elkins (2023b). To improve reproducibility and to make the methodology more robust, we also hope to integrate close reading, incorporating professional reviews on hybrid texts. Through comparing professional literary analysis and NLP analysis of literary text, we can then further examine to what extent NLP can assist in revealing the underlying patterns of great literature.

References

K. Allado-McDowell. 2020. *Pharmako-AI*. Ignota.

K Allado-McDowell. 2021. *Podcast: Pharmako-AI: co-writing with an AI and navigating the dark hallways of the mind with K Allado-McDowell (and GPT-3)*.

K Allado-McDowell and Francesco Bentivegna. 2022. Cybernetic animism: Voice and ai in conversation. *Journal of Interdisciplinary Voice Studies*, 7(1):107–118.

Hannes Bajohr. 2022. The paradox of anthroponormative restriction: Artistic artificial intelligence and literary writing. *CounterText*, 8(2):262–282.

Fritz Breithaupt, Ege Otenen, Devin R Wright, John K Kruschke, Ying Li, and Yiyan Tan. 2024. Humans create more novelty than chatgpt when asked to retell a story. *Scientific Reports*, 14(1):875.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

Jon Chun and Katherine Elkins. 2023a. The crisis of artificial intelligence: A new digital humanities curriculum for human-centred ai. *International Journal of Humanities and Arts Computing*, 17(2):147–167.

Jon Chun and Katherine Elkins. 2023b. explainable ai with gpt4 for story analysis and generation: A novel framework for diachronic sentiment analysis. *International Journal of Digital Humanities*, 5(2):507–532.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press.

Katherine Elkins and Jon Chun. 2019. Can sentiment analysis reveal structure in a plotless novel? *arXiv preprint arXiv:1910.01441*.

Katherine Elkins and Jon Chun. 2020. Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 5(2).

Angus Fletcher. 2021. Why computers will never read (or write) literature: a logical proof and a narrative. *Narrative*, 29(1):1–28.

Iona Gilburt. 2023. A machine in the loop: the peculiar intervention of artificial intelligence in writer's block. *New Writing*, pages 1–12.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Matthew L. Jockers. 2015. *Syuzhet: Extract Sentiment and Plot Arcs from Text*.

Nathan Jones. 2022. Experiential literature? comparing the work of ai and human authors. *APRIA Journal*, 5(5):41–57.

Miro Roman. 2022. *Xenotheka: a digital library*.

Mary Wollstonecraft Shelley. 2015. *Frankenstein-The Modern Prometheus: eBook Edition*. Jazzybee Verlag.

Al Sweigart. 2012. *Frankenstein eBook*.

Robert Twomey. 2023. Communing with creative ai. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(2):1–7.

Virginia Woolf. 1992. *To the lighthouse*. Springer.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.