Bits of Grass: Does GPT already know how to write like Whitman?

Piotr Sawicki[1], Marek Grześ[1], Fabricio Goes[3], Dan Brown[2], Max Peeperkorn[1], Aisha Khatun[2]
[1] School of Computing, University of Kent, Canterbury, UK
[2] Cheriton School of Computer Science, University of Waterloo, Canada
[3] Computing and Mathematical Sciences Department, University of Leicester, UK
P.Sawicki@kent.ac.uk, M.Grzes@kent.ac.uk, Fabricio.Goes@leicester.ac.uk,
Dan.Brown@uwaterloo.ca, M.Peeperkorn@kent.ac.uk, Aisha.Khatun@uwaterloo.ca

Abstract

This study examines the ability of GPT-3.5, GPT-3.5-turbo (ChatGPT) and GPT-4 models to generate poems in the style of specific authors using zero-shot and many-shot prompts (which use the maximum context length of 8192 tokens). We assess the performance of models that are not fine-tuned for generating poetry in the style of specific authors, via automated evaluation. Our findings indicate that without fine-tuning, even when provided with the maximum number of 17 poem examples (8192 tokens) in the prompt, these models do not generate poetry in the desired style.

Introduction

The recently introduced GPT-3.5 and GPT-4 models represent significant progress over the previous versions of GPT, achieving human-like performance on many tasks that were so far unattainable to Large Language Models (LLMs) (OpenAI 2023; Bubeck et al. 2023). Among creative tasks, GPT models can write poetry (Gwern Branwen 2022). In this study, however, we are concerned with generating poetry in the styles of specific authors. In our companion paper (Sawicki et al. 2023), we have examined the same challenge of generating poetry in the style of specific authors through fine-tuning GPT-3, and the results were successful. We have also found that poetry generated from GPT-3.5 (text-davinci-003) through prompt engineering only (i.e. without fine-tuning) does not follow the style of the requested author. In here, our aim is to investigate this finding further and also to check whether GPT-3.5-turbo (ChatGPT) or GPT-4 can achieve this task through prompting only. To facilitate comparison with the above-mentioned work, we attempt to generate poetry in the style of Walt Whitman without prior fine-tuning of the GPT models, and we evaluate these poems against the original works of Whitman using the automated evaluation workflows presented in our previous works (Sawicki et al. 2022; 2023).

As a main contribution of this paper, we demonstrate that generating poetry in the style of a specific author through prompting alone (whether with zero-shot or many-shot) from GPT-3.5, GPT-3.5-turbo (ChatGPT) and GPT-4 does not produce good outcome, and therefore fine-tuning is still the recommended approach.

In the next section, we describe our experimental setup, which includes three experiments to address our research question. Then, we summarize the findings of this paper and suggest the directions for future work.

Method

In this section, we describe the methodology used in this paper. First, we visually compare the difference between poems generated through the same prompt from consecutive GPT models. Then, we examine whether GPT is able to retrieve the original poems by Whitman. After that, we describe the data used for further experiments, the evaluation process, and our core results.

Three Different Models And One Prompt

While experimenting with poetry generation from consecutive versions of GPT, we have observed that the models produce poems of increasing level of complexity and length; however, the requested style is clearly not preserved. For example, Walt Whitman's poetry does not follow the 'four lines in a stanza' structure, and does not use rhyming (Bohan 1995). The majority of poems that we generated 'in the style of Walt Whitman' do follow the 'four lines in a stanza' structure and use rhyming. This, in fact, applies to most poetry generated from GPT models (including GPT-4). Only rarely will GPT deviate from this specific structure, and even then, the style does not match that of the requested author. This applies both to zero-shot prompting (where the prompt contains only the instruction to write a poem in the style of the specific author) and few-shot prompting (where in the prompt, apart from the instruction, we provide as examples a few poems by the original author). For that matter, even in a multi-step conversation with ChatGPT (GPT-3.5-turbo) and GPT-4, when the prompt highlights that the generated poems have been in 4-line stanzas with rhyme, and that the desired output should not have this structure, the model, for the most of time, still generates 4-line stanzas with rhyme.

Table 1 shows examples of the first stanzas of the poems generated from the same prompt by all three GPT models we are examining, followed by a fragment of a poem generated from the same summary using GPT-3 Curie fine-tuned for 4 epochs on the poetry of Walt Withman. The prompting procedure for the fine-tuned model is different from the prompt presented, and is fully explained in our companion paper (Sawicki et al. 2023). The results show that there is a obvious contrast between the style of writing of the fine-tuned model and the default models.

PROMPT:

Write a poem in the style of Walt Whitman.
The poem should match the following summary:

This is a poem about having courage and strength in the face of difficulties and hardships. The speaker is confident in their own determination and power, regardless of what life throws their way. They do not fear the darkness or horrors of life, knowing that they have the strength to handle what comes their way.

GPT-3.5:

Oh how I ride and stride,
In the darkest night and night's surprise.
What trouble can come my way,
Will not break me, no matter how I dread.

GPT-3.5-turbo:

Oh, the trials and tribulations of life,
How they test our will and might,
But fear not, my dear friend,
For courage and strength are within sight.

GPT-4:

I sing the song of the sailor, the wanderer of the seas,
The one who beholds the world with eyes wide and free,
In the port of Charleston, where the ships lay anchored,
I saw a sight that stirred my soul, a vision that lingered.

Fine-tuned (4 epochs) GPT-3 Curie:

OUT of the night that covers me like a hearse,
Ambient and sorrowful as the murmurous verse of those I love best,
(Not the unreplying march of Death, whose cold speech is wholly chang'd,
But the reply, the matin song of all that is, and ever shall be, in spire, in choir,
Out of the sung response, arriere glissade, in the ranks of the living,)
A vibration, blue and tremulous as a cloud of wakeful birds,
A departure and a procession stealing away into the night,
(...)

Table 1: Example of the first stanzas of the poems generated by GPT-3.5, GPT-3.5-turbo and GPT-4 from the prompt shown, followed by the fragment of a poem generated from the same summary from the fine-tuned GPT-3 Curie. The summary provided in the prompt is of the poem 'Out Of The Night That Covers Me' by William Ernest Henley. For the fine-tuned model, the prompt did not include the verbal instruction.

The style of writing is expected to change over the course of authors life. Did the young Whitman write in the same way as late Whitman? Could it be that the same author can write in a few different 'styles'? Can GPT mimic those separate 'sub-styles' with precision? Such questions are left for future research, and in here we consider the 'style' only as a very general feature, distinguishing one author's writing from another. This said, the fine-tuning workflow that we presented in (Sawicki et al. 2023) may be able to capture those more fine-grained styles, but a further analysis would be required to verify this.

Does GPT Know Whitman's Poems?

Before proceeding to poetry generation and evaluation, we first wanted to examine whether GPT is acquainted with Whitman's poetry. For that, we have run a simple experiment to check the GPT's ability to provide the complete text of requested poems.

In a sense, we are attempting to use the GPT model as a search engine here, and we are aware that, while LLMs are increasingly being used as search engines, they are notoriously unreliable at this task. Their search results are often incorrect and require verification using reliable sources (Liu, Zhang, and Liang 2023). In here, we want to accentuate the distinction between the ability to cite the text of the poems and the ability to create new poems in a requested style. The retrieved poems are compared against the ground truth, and the accuracy of the retrieved content is quantified in Table 2. These quantification can in fact support the result of (Liu, Zhang, and Liang 2023) that current GPT may return factually incorrect outputs.

This experiment is motivating the subsequent one, and our way of reasoning is as follows. The fact that a person is able to recite certain poems from memory does not imply that they are able to write in the style of that poet. For that, an average person would have to study literature, attend workshops, practice writing, etc. Our other paper (Sawicki et al. 2023) shows that fine-tuning GPT models on the works of a specific poet leads to successful acquisition of the style, similar to human studying. However, since the current GPT-4 models can generate realistic text documents in various styles that were included in its training data, a natural research question is to ask if GPT without fine-tuning has mastered the style of poets whose poems it has seen in its training data. The experiments on generating poetry without fine-tuning in the next section can in fact be seen as measuring the 'no studying' approach to style acquisition, and the fine-tuning workflow (Sawicki et al. 2023) is the 'studying' approach. In other words, if GPT-4 knows the poems of the poet in question (i.e. it has seen them in its training data and it can retrieve them when prompted), then we could expect that our experiment of generating poetry without fine-tuning would succeed in preserving the style. However, later in this paper we will show that this is not the case.

For this experiment, we have randomly selected 10 poems by Walt Whitman, and asked each of the tested GPT models to retrieve the text of the poems using the following prompt:

```
Give me the text of a poem
{TITLE OF THE POEM} by Walt Whitman.
```

Unlike in the previous versions of GPT, in GPT-3.5-turbo and GPT-4, setting the temperature parameter to 0 does not

Retrieving complete text of Whitman's poems			
Poem title	GPT-3.5	GPT-3.5-turbo	GPT-4
Spirit Whose Work Is Done	24.60%	96.05%	20.68%
Aboard At A Ship's Helm	26.43%	91.96%	94.79%
Who Learns My Lesson Complete?	21.21%	16.09%	49.59%
The World Below the Brine	28.06%	98.53%	98.53%
As At Thy Portals Also Death	27.16%	99.47%	99.47%
Eidólons	15.19%	13.82%	94.42%
I was Looking a Long While	27.60%	98.02%	98.14%
Italian Music in Dakota	24.34%	**0.0%**	82.28%
Miracles	22.81%	45.31%	67.18%
By Broad Potomac's Shore	25.05%	24.34%	23.66%
Avg. Result	**24.25%**	**58.36%**	**72.87%**

Table 2: Results of retrieving the complete text of the poems by our chosen author. The average Levenshtein distance, calculated over five trials, is utilized to quantify the similarity between the retrieved text and the original poems.

guarantee repeatability. For this reason, the process was repeated 5 times for every poem and the results were averaged. The averaged results are shown in Table 2. The similarity score reported is Levenshtein distance (Levenshtein 1966) between the original poem and the poem retrieved by the model. The Levenshtein distance is an efficient and versatile method for measuring string similarity, as it determines the minimal number of single-character edits needed to convert one string into another.

The results above 90% indicate correctly retrieved poems, with some minor differences in layout. This is acceptable, since these kind of differences are found even between different websites presenting the same poem. The lower results on GPT-3.5-turbo and GPT-4 almost always indicate that the models started to retrieve the poem correctly, but than deviated from the original text. However, the GPT-3.5 model has never correctly retrieved even a fragment of a requested poem, although these results could be different for retrieving poems by other authors. We can speculate that in the case of this model the results are always around 20% because of similar vocabulary used. It is interesting to note that in the case of "Italian Music in Dakota", GPT-3.5-turbo in all five attempts have responded: *'I'm sorry, but Walt Whitman did not write a poem titled "Italian Music in Dakota. It is possible that you are thinking of a different poet or a different poem title.'*. Therefore, we have entered 0.0% for this poem.

We can speculate that GPT's ability to retrieve the text of the poems is influenced by the number of times the poem appeared in the training dataset. Regardless, GPT-3.5-turbo and GPT-4 are, in many cases, able to retrieve the requested poems, and therefore, we can assume that those models are acquainted with the style of this poet, but as we will show later in this paper, this does not mean that they can write in the style of the requested poet, and for that—at least with the current versions of GPT models—the fine-tuning process is necessary.

Model	Version
GPT-3.5	text-davinci-003
ChatGPT	gpt-3.5-turbo (v. 2023.04.08)
GPT-4	gpt-4 (v. 2023.04.08)

Table 3: GPT versions used for poetry generation.

Experimental Setup

The principal focus of this paper is on evaluating the poetry generated through zero-shot prompts. In Reynolds and McDonell (2021) and in Kojima *et al.* (2022), it is argued that few-shot prompting is in many cases unnecessary. For example, in translation: it is not reasonable to assume that the language models can learn to translate from language A to language B just from the few examples provided in the few-shot prompt. Those works argue that the LLM already possesses the skill of (for example) translating between the two given languages, and the only purpose of the prompt is to 'invoke' that particular skill. We can speculate that this argument could extend to poetry generation using LLMs.

We were, however, intrigued by the possibility of using 8192 token-long prompts in the current version of GPT-4, which was launched 7 weeks before the submission deadline for this paper. Therefore, we also include a preliminary evaluation of poems generated from maximum-length many-shot prompts.

Data Preparation

The original author we have chosen for this work is Walt Whitman (American, 1819–1892). We use the dataset of his works created for our companion paper (Sawicki et al. 2023), which is available on our GitHub repository[1], which contains 300 poems for seven different authors (including Whitman). Since we are examining all three of the top GPT models: GPT-3.5, GPT-3.5-turbo and GPT-4 (Table 3) with zero-shot prompting, and additionally we are examining GPT-4 with many-shot prompting, we have prepared four datasets to be used in this experiment. To match the 300 samples of the original author's works, we generate 300 samples from each of the GPT models examined. For the zero-shot poetry generation, we use the following prompt for all three models (GPT-3.5, GPT-3.5-turbo and GPT-4):

```
Write a poem in the style of {AUTHOR}.
The poem should match the following summary:
{SUMMARY OF THE POEM}
```

We experimented with different ways of structuring the zero-shot prompts, but have found no meaningful differences in output quality between them.

In the case of many-shot prompting of GPT-4, we generated 300 samples with the maximum possible prompt length (8192 tokens), where, apart from the instruction to generate the poem, we provided as examples 17 poems by

[1] https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation

Whitman accompanied by their summaries. The poems included in the 17-shot (i.e. 17-poem) prompt are the following: '1861', 'A Woman Waits For Me', 'Spain 1873-'74', 'Sparkles From The Wheel', 'Spirit Whose Work Is Done', 'States!', 'Tears', 'That Music Always Round Me', 'The Artilleryman's Vision', 'The Base Of All Metaphysics', 'The City Dead-House', 'The Indications', 'Aboard At A Ship's Helm', 'The Ox tamer', 'The World Below The Brine', 'These, I, Singing In Spring', and 'Think Of The Soul'. In this case the structure of the prompt is different than the one used above, to accommodate for poem examples included in the prompt:

```
These are the examples of prompts and
completions. Prompt contains the summary
of the poem, completions contains the poem
based on this summary. Write the last
completion from the prompt preceeding it,
following the examples given.
PROMPT:
{SUMMARY OF POEM 1}
COMPLETION:
{BODY OF POEM 1}
......
PROMPT:
{SUMMARY OF POEM 17}
COMPLETION:
{BODY OF POEM 17}
PROMPT:
{SUMMARY OF THE POEM TO BE GENERATED,
FROM HENLEY AND ROSETTI DATASET}
COMPLETION:
```

As before, we experimented with various ways of structuring this prompt, but found no significant differences in the output quality. One of the approaches we tried was to provide the 17-poem prompt shown above, but without the verbal instruction preceding it, thus attempting to simulate the fine-tuning process, but that did not improve the output quality.

The summaries we use for our poem generation (both zero-shot and many-shot) are taken from the dataset published in our companion paper (Sawicki et al. 2023), and these are the same summaries that were used by us for poetry generation from the fine-tuned models. These summaries were generated for poems by William Ernest Henley (1849–1903) and Christina Rossetti (1830–1894). There are 150 summaries for each author, giving 300 summaries in total. Overall, we obtain four datasets, each containing 300 poems generated from a specific GPT model as label 0, and 300 poems by the original author as label 1. Each dataset is split into training/validation subsets, with 200/100 samples per label, respectively. This two-label setup is necessary for evaluation with binary classifiers described in the Evaluation section.

When examining the dataset generated from the 17-poem prompts, we have observed that only about 25% of generated poems have deviated from the structured/rhymed style and on the surface have resembled Whitman's poetry. We can speculate that the model produces 'higher quality' outputs when prompted with a summary which is related to the subject that Whitman was writing about, and fails when we request a poem on the subject that is not present in Whitman's works, but that would require detailed analysis by the expert in English literature.

We have to stress that few-shot and many-shot prompting of GPT-4 requires a dedicated study, and in here it was treated only as a preliminary experiment.

Evaluation

Having prepared the datasets, we are fine-tuning GPT-3 models for binary classification, following the automated evaluation methodology presented in our companion paper (Sawicki et al. 2023), where evaluation is done in the following way: binary classifiers are trained on two labels, label 0 being the GPT output, and label 1 the works of the original author. If the classifier cannot distinguish between those two classes, it means that the generated poems have preserved the style/quality of the original author. On the contrary, if the classifier can distinguish between the two classes, it means that generated poems do not match the style/quality of the original author. Achieving a 50% score would mean that both labels are indistinguishable to our classifiers, which is the desired outcome.

This approach, however, comes with a caveat because it can be argued that when the evaluation results are approaching 50%, instead of indicating the successful replication of the desired style, it may simply mean that the classifier is of poor quality. For that reason, in our other paper (Sawicki et al. 2023), we have conducted a number of experiments to establish the accuracy of fine-tuned GPT-3 models as classifiers. We found them to achieve a nearly 100% accuracy, regardless of whether the two classes represented very dissimilar texts, like Whitman's poetry vs. fragments from the book on machine learning, or more similar texts, like Whitman's poetry vs. Rudyard Kipling's poetry. In there, we have also found that of the four GPT-3 models that are available for fine-tuning (the default versions of: Ada, Babbage, Curie and Davinci), the highest performing one was GPT-3 Babbage, and therefore this model was chosen as a basis for fine-tuning the classifiers in this work.

The results of classification on all four generated datasets are shown in Table 4. The table additionally includes the results from the best performing fine-tuned GPT-3 model for Whitman's poetry (FT-GPT-3 Curie 4 epochs) from (Sawicki et al. 2023). We can compare our fine-tuned models' results with the current results because of the matching setup, i.e., we used the same dataset of Whitman's works, our evaluation setup contained the same amount of samples per label, the training/evaluation split was the same (200/100), and the poems were generated from the same set of summaries.

The results show that the classifiers were able to distinguish the GPT-generated poems from the original authors' works with almost 100% accuracy. This shows that the poems generated through prompting only do not match the style/quality of writing of the original authors, while the poems generated from the fine-tuned GPT-3 models (Sawicki et al. 2023) are approaching the style/quality of the original authors' works.

These results should be interpreted with caution in the light of the fact that the binary classifiers used are entirely black-box systems, i.e. we do not know how the classification was performed. Further research is needed to address this problem and to decipher the features that lead to high

GPT-x vs Walt Whitman original			
Model	Correct	Incorrect	Accuracy
GPT-3.5	200	0	100%
GPT-3.5-turbo	200	0	100%
GPT-4	200	0	100%
GPT-4 17-poem prompt	199	1	99.5%
FT-GPT-3 Curie 4e	**123**	**77**	**61.5%**

Table 4: Results of our experiments where GPT-generated poetry is compared against the Walt Whitman's original works. Entries in the first column indicate which GPT model's output was evaluated against the Whitman's works.

similarity of the poems that have the same style according to the classifier. However, knowing that fine-tuned GPT-3 models are reliable as binary classifiers, we can, to some extent, rely on these results. Further investigation, especially including human evaluations, is necessary to thoroughly determine the quality of the GPT-generated poetry.

Future work

In the future work, we plan to analyze GPT's ability to write poetry in the 'style' of other poets, especially those who use a structured and rhymed way of writing, as this is closer to the default style of GPT-generated poetry and may yield better results.

We are also intrigued by the question of which styles of writing can be reproduced from prompt engineering alone, and at which point the fine-tuning process becomes necessary.

Now, that the models with very large context window become available (8192 tokens for current version of GPT-4, and 32K for the upcoming version), we should investigate in detail to what extent the 'few-shot' prompt engineering can improve the models' ability to generate poetry in a requested style

Conclusion

In this study, we have examined the poetry generation ability of GPT-3.5, GPT-3.5-turbo and GPT-4 when used with prompting only. We have found that the generated poems do not match the style/quality of the works of the original author, whereas the fine-tuned model can consistently reproduce the complex style of an author like Whitman. It remains to be seen whether later versions of GPT will render the fine-tuning process obsolete (for the purpose of generating poetry in the style of a specific author), but as of now, using prompting of default GPT models does not produce good results, and fine-tuning is a recommended approach.

Acknowledgments

We thank the anonymous reviewers for their feedback and suggested references.

The work of DB is supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada. MP is supported by the University of Kent GTA Studentship Award, Prins Bernhard Cultuurfonds, Hendrik Mullerfonds, and Vreedefonds.

Author contributions

Experimental design: PS with MG, FG, DB, MP; experimental implementation: PS; writing: PS with MG, DB, FG, editing: MG, DB, FG, MP, AK.

References

Bohan, L. R. 1995. Whitman and the poetic form. In Greenspan, E., ed., *The Cambridge Companion to Walt Whitman*. Cambridge University Press. 166–193.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Gwern Branwen. 2022. GPT-3 creative fiction. https://gwern.net/gpt-3.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, 707–710. Soviet Union.

Liu, N. F.; Zhang, T.; and Liang, P. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

OpenAI. 2023. GPT-4 technical report.

Reynolds, L., and McDonell, K. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.

Sawicki, P.; Grzes, M.; Jordanous, A.; Brown, D.; and Peeperkorn, M. 2022. Training GPT-2 to represent two romantic-era authors: challenges, evaluations and pitfalls. In *Proc. of ICCC*. Association for Computational Creativity (ACC).

Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; Khatun, A.; and Paraskevopoulou, S. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proc. of ICCC*. Association for Computational Creativity (ACC). https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation/.