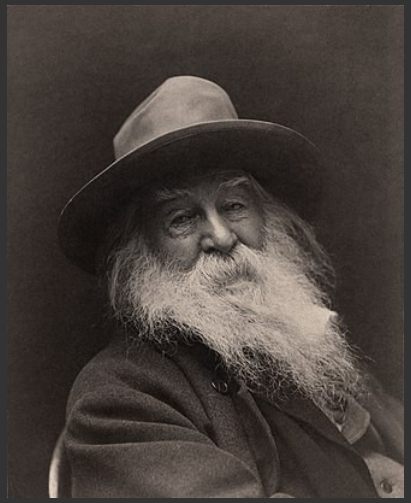
Walt Whitman



Rudyard Kipling

Piotr Sawicki
Marek Grześ
Fabricio Goes
Dan Brown
Max Peeperkorn
Aisha Khatun
Simona Paraskevopoulou

On the power of special-purpose GPT models to create and evaluate new poetry in old styles

ICCC 2023, Waterloo, Canada

Motivation

- Our objective: generate poetry in the style of a specific author and with a specific narrative.

- GPT-3.5 (text-davinci-003), GPT-3.5-turbo (ChatGPT) and GPT-4 cannot do it. See our companion paper: “Bits of Grass: Does GPT already know how to write like Whitman?” during the poster session.

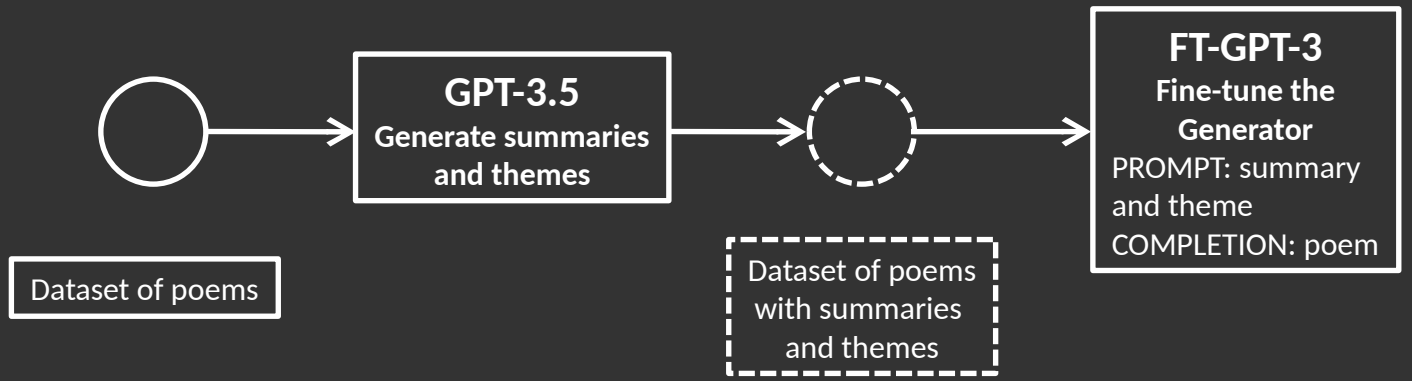
- If prompt engineering is insufficient to accomplish the task, we must resort to fine-tuning.

- We start with automated evaluation to test the method before we engage with human experts.

Presentation Structure

- Poetry Generation
- Evaluation of Generated Poetry
- Conclusion
- Future work

PART 1: GENERATION



Dataset for Poetry Generation

- “LLMs prefer large datasets for fine-tuning” - but most of poets don’t write very much.
- We are looking for English language poets who wrote at least 300 works that are between 100 and 500 words in length, and who passed away more than 75 years ago, because of the copyright laws in the UK.
- We have selected 7 of them:
  - Ella Wheeler Wilcox (American, 1850–1919),
  - Rudyard Kipling (English, 1865–1936),
  - Emily Dickinson (American, 1830–1886),
  - Lord Byron (English, 1788–1824),
  - William Wordsworth (English, 1770–1850),
  - Walt Whitman (American, 1819–1892),
  - Thomas Hardy (English, 1840–1928).

- This gives 300 poems for each author, 2100 poems in total
- We are fine-tuning the models on the works of a single poet (Walt Whitman and Rudyard Kipling), but also on a combined dataset of all 7 poets. This could also allow to mix the styles, for example, “Rudyard Whitman.”

Prompts for Poetry Generation (1)

- The available datasets of poetry do not include summaries. Luckily, GPT-3.5 (and later) are very good at summarizing. We summarized all 2100 poems.

  This is the poem:" + BODY_OF_THE_POEM + "This is the poem's summary:"

- Adding the main theme of the poem, as an additional way of controlling the output.

  "These are the categories: Mysticism, Childhood, God, Love, Life, Art, Poetry, Sadness, Despair, Depression, Death, Religion, Nature, Beauty, Aging, Desire, Travel, Dreams, Birth, War, Failure, Immortality, Fantasy. Choosing from these categories, select one that best describes this poem:" + BODY_OF_THE_POEM"

- **GPT-2 can be fine-tuned on any text, but GPT-3 requires the data to be in the format:**

  {"prompt":"BODY_OF_PROMPT", "completion":"BODY_OF_COMPLETION"}

  If completion is the poem, what goes into a prompt? A summary of the poem!

Prompts for Poetry Generation (2)

- **{"prompt":"BODY_OF_PROMPT", "completion":"BODY_OF_COMPLETION"}**

- **PROMPT:**

- <|startofauthor|>Walt Whitman<|endofauthor|>
- <|startofdates|>1819–1892<|endofdates|>
- <|startofcountry|>United States<|endofcountry|>
- <|startoftitle|>Year Of Meteors, 1859 '60<|endoftitle|>
- <|startofthemes|>Fantasy<|endofthemes|>
- <|startofsummary|>
- BODY OF THE SUMMARY
- <|endofsummary|>
- <|startofpoem|>

- **COMPLETION**:

- BODY OF THE POEM
- <|endofpoem|>

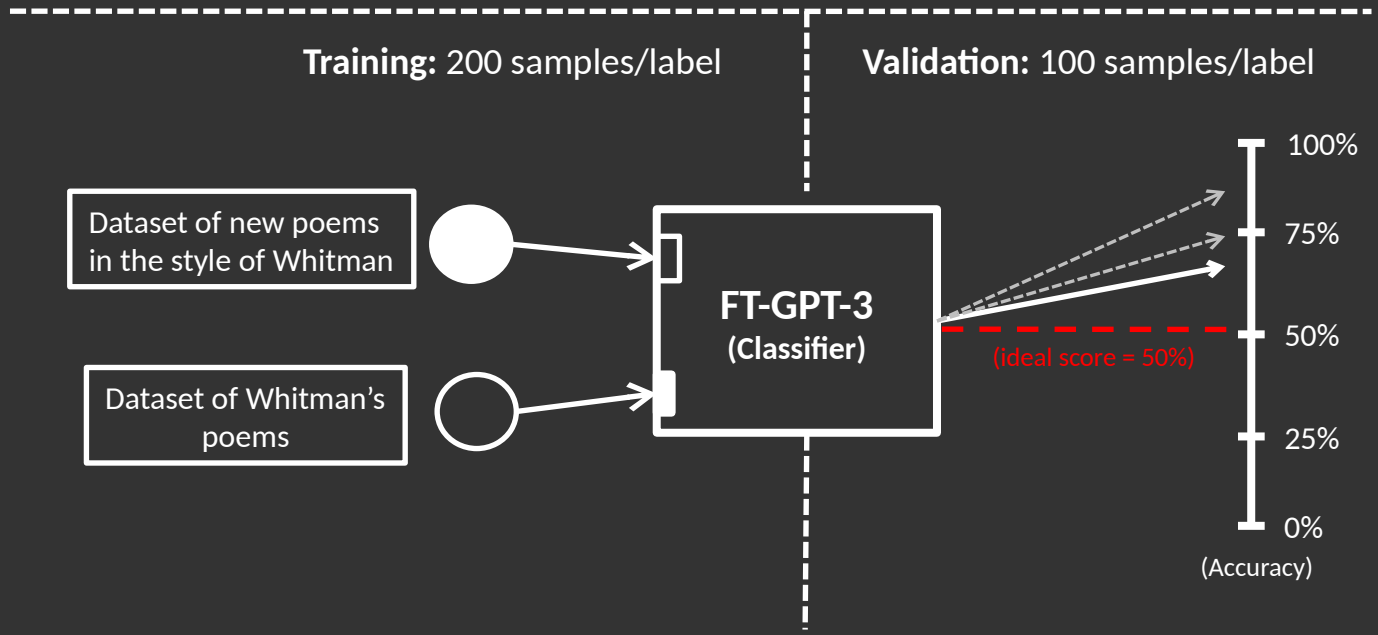
Models for Poetry Generation

- Four GPT-3 models are available for fine-tuning - the base versions of Ada, Babbage, Curie and Davinci. We fine-tune all four models for each dataset. We fine-tune all four models for 4 epochs, but Curie and Davinci we also fine-tuned for 1 epoch.

- From each model fine-tuned on a single author's works we generate 300 poems to be used for evaluation, and from the models fine-tuned on 7 authors' works we generate 300 poems in the style of Whitman and Kipling.

- The summaries we use for generation are taken from two additional poets whose works we summarized, William Ernest Henley and Christina Rossetti (150 summaries each). These summaries were not present in the fine-tuning dataset for any model.

PART 2: EVALUATION



Evaluation Strategy

- Human evaluation? Expensive and difficult. Also, we cannot present human evaluators with 1000s of poems.

- We use the automated evaluation method presented in our previous paper on style preservation through fine-tuning GPT-2. (Training GPT-2 to represent two Romantic-era authors: challenges, evaluations and pitfalls, ICCC 2022)

- GPT-based binary classifiers, where one label contains the works of the original author, the other label contains the poems produced by GPT. If the classifier can distinguish the two classes, then the outcome is not satisfactory. If it cannot, i.e. the accuracy of evaluation approaches 50%, it means that GPT-generated poems are “good”.

- Four datasets, with 200 samples per label for training and 100 samples per label for validation. They range from texts that are very dissimilar to very similar.

Validation of the Method

- 50% could also mean that the classifier is of poor quality. We must check if GPT can be fine-tuned for classification with good effect, and which of the base models available for fine-tuning is the best?

- The datasets are:

  - Walt Whitman vs. fragments from the book on machine learning (by Richard Sutton)
  - Walt Whitman vs. fragments of the prose by Leo Tolstoy
  - fragments of the prose by Leo Tolstoy vs. fragments from the book on machine learning
  - Walt Whitman vs. Rudyard Kipling

- Fragments from prose by Tolstoy and from the book on machine learning have length matching those of the poems, i.e. random length between 100 and 500 words.

Walt Whitman vs book on machine learning			
	Correct	Incorrect	Accuracy
Ada	199	1	99.5%
Babbage	**200**	**0**	**100%**
Curie	**200**	**0**	**100%**
Davinci	199	1	99.5%
Walt Whitman vs Leo Tolstoy			
	Correct	Incorrect	Accuracy
Ada	**200**	**0**	**100%**
Babbage	199	1	99.5%
Curie	**200**	**0**	**100%**
Davinci	196	4	98%
Leo Tolstoy vs book on machine learning			
	Correct	Incorrect	Accuracy
Ada	196	4	98%
Babbage	**200**	**0**	**100%**
Curie	189	11	94.5%
Davinci	180	20	90%
Walt Whitman vs Rudyard Kipling			
	Correct	Incorrect	Accuracy
Ada	196	4	98%
Babbage	**200**	**0**	**100%**
Curie	197	3	98.5%
Davinci	199	1	99.5%

Validation of the Evaluation Method

… accuracy in this test is almost 100%.

Davinci didn’t score the highest in any category, while it is the most expensive model to use (50x the cost of Ada, 25x the cost of Babbage)

Walt Whitman GPT-3 vs Walt Whitman original			
Model	Correct	Incorrect	Accuracy
Ada 4e	127	73	63.5%
Ada 7A 4e	140	60	70%
Babbage 4e	131	69	65.5%
Babbage 7A 4e	134	66	67%
Curie 1e	150	50	75%
Curie 4e	**123**	**77**	**61.5%**
Curie 7A 4e	131	69	65.5%
Davinci 1e	144	56	72%
Davinci 4e	174	26	87%
Davinci 7A 4e	137	63	68.5%

Rudyard Kipling GPT-3 vs Rudyard Kipling original			
Model	Correct	Incorrect	Accuracy
Ada 4e	170	30	85%
Ada 7A 4e	147	53	73.5%
Babbage 4e	**134**	**66**	**67%**
Babbage 7A 4e	142	58	71%
Curie 1e	173	27	86.5%
Curie 4e	160	40	80%
Curie 7A 4e	150	50	75%
Davinci 1e	175	25	87.5%
Davinci 4e	161	39	80.5%
Davinci 7A 4e	163	37	81.5%

Evaluation Results -

Classification of GPT outputs vs. the works of the original author.

Poems generated from fine-tuned Davinci did not score the highest in any case.

The combined dataset of 2100 poems did not improve the quality over the smaller datasets of 300 poems. 300 poems are enough to fine-tune the model for poetry generation.

We did not reach the desired 50%, but 61.5% for Whitman and 67% for Kipling is still pretty good.

- We did not experiment with adjusting hyperparameters!

Limitations of Evaluation

- This way of evaluation is completely blackbox. We have no idea how it was performed. That requires a dedicated study.

- These results only have meaning if they are treated cumulatively, because the classifiers are not useful for evaluating single poems.

How to use our system?

- Fine-tune your own model with any of the datasets we provide.
- Generate poems - just provide the summary of the poem you want, set the title and theme and press enter.
- Dataset and source code is available on our Github: https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation
- After the summary you can provide a fragment of the previously generated poem and the generation will continue from there. You can change the summary and theme too.
- **Our system is not an autonomous poetry generator yet, but it can be used as “poet’s assistant”.**

Contributions

- We present a workflow that allows for generation of poems with a specific narrative and in a specific author's style through fine-tuning GPT-3 models on a dataset of poems accompanied by their summaries. This approach could be extended to other than poetry categories of text, where prompt engineering alone does not give desired results.

- We demonstrate that GPT-3 models fine-tuned for classification are highly accurate as text classifiers.

- We provide a dataset of 2100 out-of-copyright poems (7 authors and 300 poems per author) where each poem is accompanied by a summary and a theme. This dataset can be used for further research on poetry generation.

- We show that the smaller models (Ada and Babbage) produce results comparable to larger models (Curie and Davinci), thus considerably reducing the costs of fine-tuning GPT-3 for poetry generation and evaluation. This indicates that some tasks, like poetry creation, do not require the use of largest models. This opens the possibility of using smaller, open-source models that can be run on a customer grade hardware.

Future Work

- Fine-tuning GPT-3 alternatives: OPT, Llama, GPT-J, GPT-NeoX20B, etc.
- Different than summary ways of encoding the poems
- We need to find a more effective way of automated evaluation of poetry
- Finally, human evaluation.

Thank you!