Imitations of Immortality:

learning from human imitative examples in transformer poetry generation

RAY LC
School of Creative Media
City University of Hong Kong
Hong Kong SAR
LC@raylc.org

ABSTRACT

"To me the meanest flower that blows can give
Thoughts that do often lie too deep for tears."
- from *Intimations of Immortality*
by William Wordsworth

Learning to generate poetry in the style of the poet can make models style experts, but humans who create imitative works take a more general approach that incorporates knowledge outside the poet's style. Instead of learning from a large corpus of one poet's works, can machines imitate deep style using only one example of her work? To explore generating poetic variations for a web-based installation art work, I wrote eight poems that imitated the structure of eight poets, and used them to fine tune a transformer model that has seen only one poem by each author. The poems presented show structures borrowing from the human imitation in addition to prompted content of the original, suggesting the model has learned aspects of how humans write variations on content by imitating style. Audience evaluation reveals an ability for machine-generated text to reproduce the nuance of the original text as well as the human variation, despite being less expressive.

CCS CONCEPTS

• Human computer interaction (HCI) • Interaction paradigms • Natural language interfaces

KEYWORDS

machine learning poetry, generative text, machine perception, machine creativity

ACM Reference format:

ARTECH'21, October 2021, Alveiro, Portugal

1 Introduction

Machines can learn to imitate a style by a corpus of examples, but frequently great works of literature have only one salient exemplar. If we want to train a machine to learn to produce Hamlet's soliloquy with the same level of nuance that Shakespeare intended, one approach is to ask humans to produce poetic variations and use them to fine tune the model. Thus the human-written variations serve as examples to point the model towards grammatically, thematically, and structurally relevant state spaces where the particular text generated lives.

Transformer models like GPT-2 have been found to generate poetry indistinguishable from human-written poetry, given enough training examples [5]. Learning to generate particular styles, however, is more subtle, as models often generate free-form streams that don't adhere to poetic forms. Recent approaches to generating based on poetic forms include imposing strict rules on rhyming [7] and using characters in predefined positions [8]. These approaches do not take into account styles of particular poets, which may impose additional structural limits. Work in this area includes modelling poet style explicitly during line-by-line generation [13], and generating poems based on emotions evoked by particular words associated with particular poets [1]. Having such style-specific models allows us to transfer the style in text to other content [4], creating scenarios such as "how would Virginia Woolf write villanelles about loss?"

Given models exposed to styles of different poets, we may want to analyze which instances of generated lines best capture a given style. Attempts to evaluate text quality have included using the cosine similarity score in a doc2vec embedding of the poetry [9] and evaluating semantic correlation using a probabilistic Natural Language Inference model [3]. A meta study found, however, that the different metrics can code well for semantic similarity, but often misses the structural component in measures of syntactic fluency [11]. In consideration of this, I evaluated the text outputs manually, using a standard I intuited based on my own variation exercises on the original poems. The result is shown

interactively on the web as part of an online art exhibition (Figure 1).

Figure 1: Web interface to the poetry text of *Imitations of Immortality*, 2021, interactive.

To evaluate the efficacy of the poetic works generated in regards to capturing the essential nuance of the original poem, I gave naïve online audiences a corresponding selection of text from my own variation and from GPT-2 and asked them to identify whether they were generated by machine or human, to rate the text level of expressiveness and structure, and to rate how they capture the essence of a reference text by an unidentified original author. I found that participants produced similar rates of error to machine and human generated text, indicating that they could not distinguish between the two. Moreover, they were equally likely to indicate human and machine text analogs as most representative of the nuanced style of the original, showing that they served as equally valid variations. While they were perceived to have the same level of structure, the human texts were perceived to be significantly more expressive than the machine texts.

2 Methodology

I chose 8 poems by 8 different authors that represent a cross section of styles ranging from free verse, villanelle, quatrains, haiku, blank verse, rhymed and unrhymed. I composed 8 poems on my own that utilize aspects of each of the originals in terms of form but narrated my own content. For example, here's the Elizabeth Bishop poem and my own:

> "I lost two cities, lovely ones. And, vaster,
> some realms I owned, two rivers, a continent.
> I miss them, but it wasn't a disaster.
> --Even losing you (the joking voice, a gesture
> I love) I shan't have lied. It's evident
> the art of losing's not too hard to master
> though it may look like (Write it!) like disaster."
>
> - from *One Art*
> by Elizabeth Bishop

> "I blank out on my age, name, and every syllable
> I uttered (Emphatic!) to make my existence remain.
> It moved on before me, but even that's forgivable.
> But when you (whose voice moves me, face bespectacle)
> left me my memories, I give but one refrain:
> the science of forgetting is not inscrutable
> but being forgotten, that is unforgivable."
>
> - from *The Science*
> by the author

Note that I imitated not only the villanelle structure, but also the nuanced structures such as using parentheses for a call to action, non-standard punctuation, evocative images, and not simply repeating the ending for each of the last line of the stanzas but varying them in turn, as Bishop did. I believe these are creative inputs that will be difficult for models to pick up using one-shot learning, especially given their difficulty with syntax as opposed to thematic content.

Next I fine-tuned the GPT-2 355M and 124M models with each of the 8 original poems only (not putting in my own poems) for 5000 epochs (5480 tokens, learning rate 0.0001, average loss 0.01-0.02). Then I prompted these models with the beginning (first stanza or equivalent) of each of the 8 original poems at a range of temperatures from 0.8 to 1.8 to see how the models created new content based on the prefixes. I noticed quite a bit of overfitting, as many runs simply repeated the entire poem verbatim given the first stanza. I failed to find many deviations for the Wallace Steven's poem *Thirteen Ways of Looking at a Blackbird*, possibly because the form is strictly bound by the roman numerals. I also saw no noticeable difference between 355M and 124M, and decided to work thereafter with the 124M model for its smaller number of parameters and greater variation in the text generated. When the models do go off the script (especially at higher temperatures), it was hard to see any consistency in the generated forms, although the content was generalized:

> "I
> Among twenty snowy mountains,
> The only moving thing
> Was the eye of the blackbird.
> II
> I was of three minds,
> Like a tree
> In which there are three blackbirds."
>
> - from *Thirteen Ways of Looking at a Blackbird*
> by Wallace Stevens

> "I
> Among twenty snowy mountains,
> The only moving thing
> Was the eye of the blackbird.
> Q. How did you the blackbird see?
> A. I saw behind the blackbird.
> Q. But the blackbird cannot fly.
> Do you not see the blackbird flying?"
>
> - from GPT-2 pretrained to 8 original poems
> by 8 different poets

Thus I decided to pre-train the 124M model to the 8 original poems together with each of my poems written in variation with the original (5000 epochs, average loss 0.01-0.02). The idea is to provide an example of how variations could respect the original poem and push the model towards learning the way humans write variational forms. It also provides an example of syntactic and structural coherence despite thematic and semantic variability.

Table 1: Poetry and Generation Parameters of web installation *Imitations of Immortality*.

Original Poem	Original Poem Author	Variation by the author	GPT-2 Temp
Do Not Go Gentle Into That Good Night	Dylan Thomas	Abscond the Fabled Glorious Primrose Way	1.3
Howl	Allen Ginsberg	Sex	1.8
Daddy	Sylvia Plath	Mommy	1.6
Thirteen Ways of Looking at a Blackbird	Wallace Stevens	Twelve Ways of Listening to Maria	1.3
Variations of the Word Love	Margaret Atwood	Variations on the Rain	1.2
Valediction Forbidding Mourning	John Donne	Eulogy Contra Severance	1.4
One Art	Elizabeth Bishop	The Science	0.8
This is Just to Say	William C. Williams	聞きたかっただけ	1.0

3 Outcomes

3.1 Text Generation

The website compiles the poems generated in an interactive format (Table 1). While the results still need to be curated to a slight extent by the writer, I found much more success with poems that have to fit certain forms, perhaps due to the additional pre-training that constrains the model more towards variations that fit the same style as the original: generalized:

> "Good men, the last wave by, crying how bright
> Their frail deeds might have danced in a green bay,
> Rage, rage against the dying of the light.
> Wild men who caught and sang the sun in flight,
> And learn, too late, they grieved it on its way,
> Do not go gentle into that good night."
> - from *Do Not Go Gentle into That Good Night* by Dylan Thomas

> "Good men, too, were wilt thou find it so?
> Lose what is lost, and reap what is lost?
> Tread thorns to die and live whence they came.
> That awful knowledge it brings or cares,
> Opportunities, vain imaginings, but find believe by heart's desire, Do not go gentle into that good night"
> - from GPT-2 pretrained to *Abscond the Fabled Glorious Primrose Way*, by the author

Note that it incorporated the content in the "thread thorns" line that came from my own variation, but still followed the villanelle form. These examples abound in GPT-2-generated content from Allen Ginsberg's *Howl*. For example:

> "it should be happening soon now I tell my mom, any time within the next one plus or minus two years, in between I gotta find some stuff to do so I don't hang out to dry, so you think I teach those ignorant bombastic know-nothing wanna-be-liberal-but-can't-leave-their-couch- conservative I-fuck-who-my-mom-tells-me-to-fuck I'm-going-to-this-fantastic-party-looking-at-me-I'm-getting-hooked-up I-can-drink-like-a-whale-shoot-like-an-addict little anal cocksuckers for fun?, you think they pay me to figure out who's got a boy back home and a thousand acres and a muel?"
> - from *Sex,* by the author

> "who know how I don't live by the now nothingness that is millennial mean-spirited-conscience care, whoomon-and-patron-waving-makers-through-their-lamb wires-of-ideology-because-I-hate-technology-I-hate-texts-I-hate-music-I-hate-expertise-but-I-hate-your-accent-makes-me-blurry-I-like-it-but-I-hate-your-accent-makes-me-blame-you-on-a-crustacean-of-middle-school-culture I-hate-your-accent-makes-me-blunt-andi-and-don't-you-respect-you-amateur-profanities-you-a-fucking-Bigotry-comedian"
> - from GPT-2 pretrained to *Sex,* by the author

Here, I used the word "it" as paragraph starters in my variation *Sex*, instead of Ginsberg's "who." The *Sex*-pretrained generated poetry similarly managed to switch to "while" in the middle for 4 lines. It also picked up on the use of dashes to join words that I used in *Sex* on line 20. There's a cross-contamination of content from the different poet's works both into my own writing and in GPT-2, since the training utilizes all the original poems. Dylan Thomas's *Do Not Go Gentle Into That Good Night* provided the line "And you, my father, there on the sad height, Curse, bless, me now with your fierce tears, I pray." In my work after John Donne, it became a meditation on our own mortality. In the words of GPT-2, who was trained to the variation by me, and both Dylan Thomas and John Donne poems, the same content generalized:

> "So you, my father, leaves me now,
> Through pains of hemorrhagic nail…
> The death that comes now, my father,
> Is an old trail from which you stray.
> The future of a child's past, before
> We died a little bit each day"
> - from *Eulogy Contra Severance*, by the author

"Silk soaked in oil is no more safe than salt is safe
(The way would be different!).

Still, you father, tears are an art
To inspire a bright future."
- from GPT-2 pretrained to
Eulogy Contra Severance, by the author

This machine-generated line is also notable in that it has established GPT-2 as an original voice, using a parenthetical remark borrowed from Elizabeth Bishop (see next section), and saying things about both "tears" and "father" in its own way, transcending each of the ways Dylan Thomas and myself used those phrases for its own seemingly original intent.

3.2 Content-Structure Trade-off

As I found previously, appropriate content can be obtained from the generative process. Changing the temperature is necessary for each generative process, however, because some poems required a lot of variation in the generation side while others have the variations built into the content itself. Obtaining the right level of variation, however, still requires human supervision. The structural and syntactic aspects, on the other hand, requires careful curation. Often only a small number of lines generated fall into the structural framework of say a villanelle or a rhyming couplet. By curating the structure, we are allowing the variation in content to speak more powerfully in the finished machine-generated text.

This type of trade-off is evident in the text of the GPT-2 version of *One Art*. When prompted by the first stanza of the poem, the first instance of the raw GPT-2 output contained 4, 2, 12, and 6 lines, in sequence, breaking with the structure of the villanelle, which must be in 3-line form. The content however was unique yet evocative of the original:

"Thus close to the truth (grab it, and run!),
how come no one has seen such a sight for years?
The stranger the disaster, the farther a word...
Such people are in themselves no disaster,
but the dead are involved, they are all involved,
and the end of the world is near"
- from GPT-2 pretrained to *The Science,* by the author

Note how it has taken the disaster idea from *One Art* as well as the grammatical structure of the imperative inside parentheses ("write it!" and "the joking voice, a gesture I love" in the original). While the structure fails as a villanelle, the thematic elements appear to diversify. The second instance of the prompted output produced stanzas that were all 4 lines long, much closer to what the strictly 3-line stanzas of the poem is structurally. But it simply repeats the same:

"Such people who are lost surely are reflecting on
their misfortunes,
their intoxication, their intoxication
sure enough, is not bitter still.
Such people who are lost surely are reflecting on
their misfortunes,
their intoxication, their intoxication
sure enough, is not bitter still"
- from GPT-2 pretrained to *The Science,* by the author

Here, the content is not diversified, but the structure is simpler and does not require human curation. The human hand in selecting from the text output is part of the poem creation process. What I did was take the best of the thematic output as casually observed, and put them structurally into stanza forms that makes the poem recognizable to the reader.

3.3 Generating Fragments

In the William Carlos Williams and Wallace Stevens examples, I had to generate short sections of text that corresponded to original works. Here, the curator's hand is more prominent than in generating long text that ramble on, such as Allen Ginsberg and Margaret Atwood examples. Constraints on the structure of words pushed the information content of the poem away from the content in the content-structure tradeoff outlined earlier. I illustrate the curation process using the raw generated text from GPT-2 for Williams's work. The first of the three 4-line stanzas serves as the prefix primer:

"I have eaten
the plums
that were in
the icebox"
- *This Is Just to Say,* by William Carlos Williams

Here, GPT-2 takes a lot of liberties with the source material, having the freedom to say things seemingly unrelated to the premise in much of the generated text. Note that the raw structure of the lines following the prefix are also not constrained. Note also here that the pre-training of the model included my own variation which is in Japanese, but given to GPT-2 in the form of the romanization ("kimi ga ano\\ kokonattsu o\\ hakkou sa sete bin ni\\ irete…"). The thinking here was that the structure of my Japanese poem can provide some structural line-breaks for the generation:

"I have eaten
the plums
that were in
the icebox
And who didn't wish a nostalgic whackle of a ring
In a press of turquoise sparks."
- from GPT-2 pretrained to
聞きたかっただけ, by the author

While the structure did not reflect the original, the content is on the right track: the very next few generations yielded results which were in line with the Williams Carlos Williams content, but without the appropriate line-breaks:

"I dream of Bill.
Of the telephones I have,
And of the hard copy machine I,
Have but the black of a caption."

- from GPT-2 pretrained to
聞きたかっただけ, by the author

The selection process looks at all such appropriate content as well as structurally valid forms. Note that GPT-2 is capable of the content in the Williams poem, but syntactic level constraints are more difficult, and often altered by the human curator. The final selection comes from a long ramble that breaks off at the right time. Here even the syntactic breaks like the beginning of a next sentence on stanza 3 is retained, with noted removal of middle sections by curation.

""I have eaten
the plums
that were in
the icebox

and heard
the ticking of
the box
the first time

It was a lot
like dinner,
and the ticking
stopped."

- from GPT-2 pretrained to
聞きたかっただけ, by the author

3.2 Parameters for Generation

Different poetic styles seem to necessitate different strategies for text generation. I employed the same number of epochs (5000), learning rate (0.0001) and model size 124M for each model, so the diversity in the text generation comes from pretraining via additional poem variation written by me and the temperature setting of the text generation (Table 1).

The temperature determines how unlikely the generated sequences will be given the same model [2]. Turning up the temperature makes the poetry more wild and variable, while lower temperatures can constrain the model to generate more sensible text. In practice, setting the temperature below 1.0 tended to repeat the entire poem used to pretrain the model. The exception to this was the Elizabeth Bishop variation generated with temperature 0.8. It's in general hard to predict what is the optimal temperature to generate variable and yet understandable poetry given its pretraining source, although the desired structure of the output can determine the temperature setting. For example, the variation based on Allen Ginsberg was desired to be chaotic and flowing much like *Howl*, so a temperature of 1.8 was employed in this case.

An example of a more balanced approach is the variation based on Margaret Atwood, which despite using a modest temperature of 1.2, was able to, to some extent, generalize the idea and content of *Variations on the Word Love*, which had revealed the consumer and advertising around our concept of "love."

"This is a word we use to plug
holes with. It's the right size for those warm
blanks in speech, for those red heart-
shaped vacancies on the page that look nothing
like real hearts. Add lace
and you can sell
it. We insert it also in the one empty
space on the printed form."

- from *Variations on the Word Love*
by Margaret Atwood

"Add lace
and you can sell
it. We insert it also in a fake body
and a touch of sexuality to a
gel. It's the size of your hand
and your imagination can do with
lay copulation what text just said"

- from GPT-2 pretrained to *Variations on the Rain*
by the author

3.5 Interactive Presentation

A webpage shows the poetry written and generated interactively (Appendix). Sections are separated into human and GPT-2 variations. The GPT-2-generated text is overlaid on the original text and revealed via hover (Figure 2).

Figure 2: *Imitations of Immortality* (2021). (Upper left) Writing variations. (Upper right) GPT-2-generated variations. (Lower left) Original poem used for pre-training. (Lower right) GPT-2 generated poem with the first stanza highlighted as prefix for the generation.

In order to show the content of the poetry in human-written variations, I used CSS and javascript to make interactive elements in the text (Figure 3). The interactions are matched to the poetry content. For example, the interaction for *The Science* is a slowly disappearing poem that can be brought back into view by using the mouse. The poem itself is about the process of forgetting and how we can bring memories back if needed, but cannot forgive being ignored.

In *Mommy*, based on Sylvia Plath's *Daddy*, the character is described as appearing in movies, so each movie title she appears in is displayed as a neon sign in the web interaction.

In 聞きたかっただけ, the text is written and shown in Japanese. In *Variations on the Rain*, the words of the poem fall from the pages according to the location of the mouse. Falling away of the words periodically washes away the text, producing a raining poem that also illustrates its ending:

> "it brings perpetual
> Spring,
> touches even one who sits lonely upon a
> rock, and when you slap it,
> it turns the other cheek, melting it, not just
> the mortal body. It giveth and it
> washes away, it is part of
> everything we are, and it falls to us
>
> drop by drop."
> - from *Variations on the Rain*, by the author

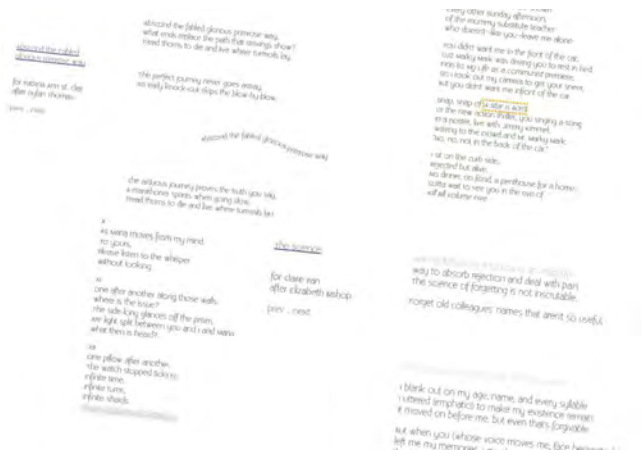

Figure 3: *Imitations of Immortality* (2021). Examples of web-based interaction in the presentation of poetic forms. (Upper left) Using a scroll-based path to illustrate *Abscond the Fabled Glorious Primrose Way*. (Upper right) Using a Neon Lights graphic to illustrate the movies that inspired *Mommy*. (Lower left) Showing what the poem states in the content using typography in each of the variations in *Twelve Ways of Listening to Maria*. (Lower right) Showing the process of forgetting as described in *The Science* of forgetting.

Finally, the interactions are seen most prominently in *Twelve Ways of Listening to Maria*. Here, each mentioning of "Maria" in the text of the poem evokes an interaction that fits the content. For example "a heart pumping from Maria" is shown by a periodically bold-flashing text, "whether Maria is outside, or inside" shows the text leaving the page, "are light split between you and I and Maria" shows the text splitting into opposingly moving colors, etc. This style of showing the meaning behind content suggests the possibility of having a type of interactive meaning in the GPT-2 generated text. A future exercise is to automatically generate CSS-type interactions based on text content [12].

4 Evaluation

To understand how audiences interpret machine-generated and human-written text differently, a survey was given to readers naïve to any of the texts (n=25) asking to guess the identity of the author (machine vs. human).

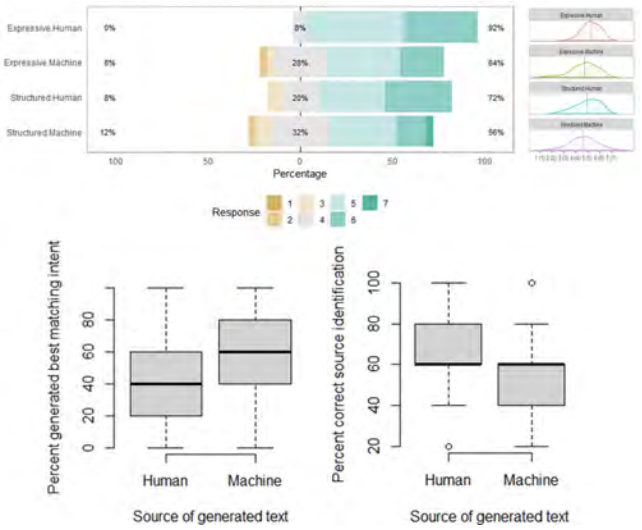

Figure 4: Evaluating the text of *Imitations of Immortality*. (Above) Likert scores of how expressive or structured the sample text is for the audience surveyed (n=25). "How expressive do you feel the following text is?" (1 – not expressive, 7 – very expressive) and "How well structured do you feel the following text is?" (1 – not structured, 7 – very structured) are asked. Significant difference in expressiveness (Wilcoxon Ranked Sum, p=0.03689), no difference in structured (Wilcoxon Ranked Sum, p=0.1985). Density estimation of the Likert scores - right. (Below) Percent of human or machine generated text deemed to be best in matching underlying nuance of the analogous text in the original poem, - left. Percent of human or machine generated text that is identified correctly - right.

The readers were given an online survey titled "Who's Text Is It Anyways" that includes 10 snippets of texts (5 generated by GPT-2, 5 written by the author, given in pairs with no answers given). After being asked to identify whether the text was generated by human or machine, readers were surveyed on how expressive and how structured they feel each snippet was (Figure 4). For each pair of text snippets, participants were asked which one of the two texts best exemplified the underlying nuance of the original poetry text. The entire list of texts presented are found in the Appendix. For example, the following pair (both stanza 5 in their respective texts to match for analogous content) were presented for readers to identify whether they are generated by humans or machines:

"I blank out on my age, name, and every syllable
I uttered (Emphatic!) to make my existence remain.
It moved on before me, but even that's forgivable."

"Such people are in themselves no disaster,
but the dead are involved, they are all involved,
and the end of the world is near."

Readers were then asked which one best matched the underlying nuance of *One Art*'s stanza 5 by Elizabeth Bishop:

"I lost two cities, lovely ones. And, vaster,
some realms I owned, two rivers, a continent.
I miss them, but it wasn't a disaster."

In this case, the text from the left above is by human, and the right is by machine. Each pair of such texts (5 pairs in total) are given to the participant to rate in sequence without revealing their identities. The results were averaged for each participant to give a single percentage correct – machine, percent correct – human, percent best matching – machine, percent best matching – human for each individual. The median expressiveness and structuredness for human and machine generated texts were also calculated for each individual. The results were aggregated across 25 readers.

Results show no significant difference in being able to identify the human-written text (61.6% correct) and the machine-generated text (56.0%) correctly as being written by human and machine, respectively (Wilcoxon Ranked Sum p=0.2177). This suggests that human and machine generated poetry in this context cannot be reliably differentiated from each other, as the previous example given for *Howl* and *Sex* texts suggested. Moreover, there's no significant difference of the percent of best nuance-matching text that are machine-generated from 50% (Wilcoxon Ranked Sum p=0.1343). Overall, 56.8% of the best matching text are by machines, with the remaining 43.2% by human, so if anything, readers believed the machine texts to be as representative of the original author's text as the human-written text, if not more so.

The level of structuredness of the human (median = 5) and machine generated (median = 5) texts is not significantly different (Wilcoxon Ranked Sum p=0.1985), which is not surprising since the curation process was done by humans alike to account for some of the structural uniformity. However, the level of expressiveness of the human-written text (median = 5) was significantly higher than the expressiveness of the machine-generated text (median = 5) (Wilcoxon Ranked Sum p=0.03689). Why this is so can be explained in Appendix Figure 1. Even though the medians are the same, there are significantly more 6 ratings in the expressiveness of human texts. This indicates that although bearing structural similarities, the content of the text by human and machine may exhibit differences in ability to express poetic content.

5 Conclusion

Imitations of Immortality is a creative study on how one-shot machine learning models can begin to learn to produce variations of canonical styles of poetry by pre-training to human-written variants. Our exercise explored the possibilities of machine creativity in the context of producing variations, for in some sense, human creativity itself is related to its ability to variate based on strong precedents followed. For example, *Apocalypse Now* is a variation on *Heart of Darkness*.

Human study of the perception of human-written and machine-generated texts show that both sources of text can equally represent the nuance of the original text, and that people cannot disambiguate human-written text from machine generated text in the poetic context (Figure 4). Interestingly, we found that machine-generated text is found to be significantly less "expressive" even though they may represent the nuance of the original text by the classic authors just as well. Even though operationally the machine-generated texts may represent poetic content just as well, we hypothesize a tangible difference in the way they make audiences feel in how expressive they can be. Further work is needed to distinguish the expressiveness of the text from its operational ability to represent poetic content. One can argue, for instance that the goal of the variation of a poem is to express, and not only to represent.

However this exercise also points to the importance of understanding the release and use of machine language models in terms of equitable access, transparency, authorship, and risks of undull human influence and misinformation [10]. In particular, recent history have suggested that the generation of conspiracy theories and information calculated to influence using falsehoods can lead to severe behavioral changes. Comparison of human-written and machine-generated conspiracy theories show remarkable resemblance [6], suggesting a similar strategy of pretraining used by both humans and machines. Investigating the way text generation occurs given a corpus of creative input text will be crucial to determining its effect on the public.

REFERENCES

[1] Brendan Bena and Jugal Kalita. 2020. Introducing Aspects of Creativity in Automatic Poetry Generation. *arXiv:2002.02511 [cs]* (February 2020). Retrieved October 5, 2020 from http://arxiv.org/abs/2002.02511

[2] Gwern Branwen. 2019. GPT-2 Neural Network Poetry. (March 2019). Retrieved January 19, 2021 from https://www.gwern.net/GPT-2

[3] Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference. *arXiv:2011.10819 [cs]* (November 2020). Retrieved January 14, 2021 from http://arxiv.org/abs/2011.10819

[4] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style Transfer in Text: Exploration and Evaluation. *arXiv:1711.06861 [cs]* (November 2017). Retrieved October 8, 2020 from http://arxiv.org/abs/1711.06861

[5] Nils Kobis and Luca Mossink. 2020. *Artificial Intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry*. arXiv.org. Retrieved October 5, 2020 from https://econpapers.repec.org/paper/arxpapers/2005.09980.htm

[6] Sharon Levy, Michael Saxon, and William Yang Wang. 2021. The Truth is Out There: Investigating Conspiracy Theories in Text Generation. *arXiv:2101.00379 [cs]* (January 2021). Retrieved January 19, 2021 from http://arxiv.org/abs/2101.00379

[7] Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. 2020. Rigid Formats Controlled Text Generation. *arXiv:2004.08022 [cs]* (April 2020). Retrieved October 5, 2020 from http://arxiv.org/abs/2004.08022

[8] Yi Liao, Yasheng Wang, Qun Liu, and Xin Jiang. 2019. GPT-based Generation for Classical Chinese Poetry. *arXiv:1907.00151 [cs]* (September 2019). Retrieved October 5, 2020 from http://arxiv.org/abs/1907.00151

[9] M. C. Santillan and A. P. Azcarraga. 2020. Poem Generation using Transformers and Doc2Vec Embeddings. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. DOI:https://doi.org/10.1109/IJCNN48605.2020.9207442

[10] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. *arXiv:1908.09203 [cs]* (November 2019). Retrieved October 6, 2020 from http://arxiv.org/abs/1908.09203

[11] Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating Evaluation Methods for Generation in the Presence of Variation. In *Computational Linguistics and Intelligent Text Processing* (Lecture Notes in Computer Science), Springer, Berlin, Heidelberg, 341–351. DOI:https://doi.org/10.1007/978-3-540-30586-6_38

[12] Muminov Ibrokhim Botir Ugli. 2020. Will Human Beings Be Superseded by Generative Pre-trained Transformer 3 (GPT-3) in Programming? *IJOT* 2, 10 (October 2020), 141–143. DOI:https://doi.org/10.31149/ijot.v2i10.769

[13] Jia Wei, Qiang Zhou, and Yici Cai. 2018. Poet-based Poetry Generation: Controlling Personal Style with Recurrent Neural Networks. In *2018 International Conference on Computing, Networking and Communications (ICNC)*, 156–160. DOI:https://doi.org/10.1109/ICCNC.2018.8390270

APPENDIX

The original poems in interactive format and text generated by GPT-2 are found at: [link removed for anonymity].

The 10 text snippets used in the survey were (each text from S1-S2, S3-S4, S5-S6, S7-S8, S9-S10 were asked which of the two best matched to the equivalent stanza from *A Valediction Forbidding Mourning, Daddy, Thirteen Ways of Looking at a Blackbird, Variations on the Word Love,* and *One Art*, respectively):

S1: Human - *Eulogy Contra Severance* - stanza 7

S2: Machine - GPT-2 generated from *A Valediction Forbidding Mourning* - stanza 6

S3: Machine - GPT-2 generated from *Daddy* - stanza 15

S4: Human - *Mommy* - stanza 15

S5: Machine - GPT-2 generated from *Thirteen Ways of Looking at a Blackbird* - stanza 10

S6: Human - Twelve Ways of Listening to Maria - stanza 10

S7: Machine - GPT-2 generated from *Variations on the Word Love* - line 7

S8: Human - *Variations on the Rain* - line 27

S9: Human - *The Science* - stanza 5

S10: Machine - GPT-2 generated from *One Art* - stanza 5

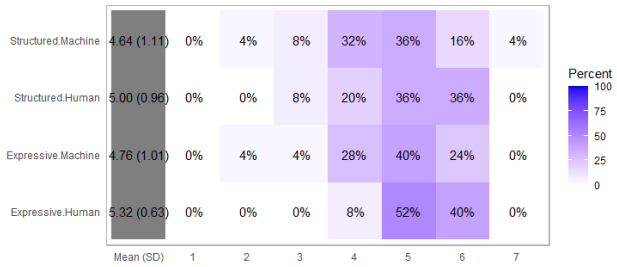

Figure A1: Expressiveness and structured-ness of machine-generated and human-written text as result of survey (1 – not expressive / not structured, 7 – very expressive / very structured) shown in heat-map form.