Understanding human-machine phraseologies through an AI-driven analysis of Henry James

Quentin Clark, ENG/CAS BU, [qtcc@bu.edu](mailto:qtcc@bu.edu)

Introduction

Artificial intelligence, while long recognized for its stunning ability to unveil patterns in data and perform tasks previously relegated to people, has recently begun being recognized in the public consciousness for its remarkable abilities to replicate and produce art of various kinds: a task previously thought to be in some way uniquely human. Art generation models like Stable Diffusion and DALL.E, music generation models like Jukebox and SOUNDRAW, and text generation models like GPT-3 have proven to produce end material near-unrecognizable from works produced by human beings. Inherent to this shift in power once exclusively shared by humans now shared with our familiar yet conspicuously alien computers[[1]](#footnote-1) is a cultural and philosophical re-evaluation of our understanding of art itself and the nature of the artistic process.

In this work I use existing methods to create artificial samples replicating the style of American/British author Henry James, who employed a distinct free form yet structured prose in his many works exploring the nature of desire, love, and life in the rapidly changing turn of the century world. Borrowing on contemporary writers concerned with technology, I will analyze what potential use such a method might have in helping us understand James, if it has any at all.

Technical Methods

To generate a machine learning model capable of mimicking the style of Henry James’ writing, I employed models from the Natural Language Processing (NLP) field, where text generation has been one of many long established goal for aspiring models. All of these models (called Large Language Models, or LLMs, due to the size of the network architecture) are in some way based on Attention models, which distilled to its simplest form take in series of text and attempt to learn probabilistic relationships between the words in a sequence. For example, for the sentence “An apple a day keeps the - ”, any person would expect the word “doctor” to be the next word in the sentence with high probability, given the frequent use of that idiom in the English language. But if the sentence was “an apple a week keeps the – “ there would be greater uncertainty due to the previous works having a less probabilistically rigorous connection to any particular following word in common usage.[[2]](#footnote-2)

First, before one can consider what model one wishes to use in mimicking a writing style, one must generate a robust corpus (or collection of writing) in that style. This was quite easy for Henry James, as he was an incredibly prolific writer: Project Gutenberg has around a hundred full-length written works by James, encompassing novels, philosophical pieces, and compilations of his other writing like letters (some of these have his full-length novels split into multiple volumes). I compiled the majority of these from Gutenberg, removing formatting text like the copyright notice (or attempting to)[[3]](#footnote-3), and trying to only include the longer length works and not short essays or compilations including works from other authors. This resulted in a training corpus of around 5 million words. From now on I will refer to this by the HJC (short for Henry James Corpus).

With an attention model, one has a couple of potential options to train the model. One is to simply train it directly over the input data one wishes to mimic. This was my approach. I trained an existing implementation of a model similar to GPT-1 on the HJC, except with 10 self-attention layers/heads instead of 12 to save computational time. The basic intuition behind how GPT-1’s architecture works is that it contains several “blocks” that attempt to find probabilistic relationships between words in a text as described above. This model, and indeed all machine models, essentially train via some form of loss reduction. Some function is defined to give a numerical value representing the “closeness” of a model’s guess to a true value, and the model it updated via taking the derivative of that loss function with respect to the parameters of the model which are updated along that gradient. This derivative method is called “backpropagation”, and the exact formula or method for updating the parameters of the model are called optimization methods. It was trained with a batch size of 254 (how many samples it examines at once), with the AdamW optimizer with L2 weight regularization (a particular optimization function incorporating several clever tricks for faster optimization), over around 30 epochs (passes through the whole training data set) with a rough, manually set decreasing learning rate schedule from around 5e-2 to 5e-6 (the model updated more quickly early in training, and more slowly later). This training took a couple dozen hours on a consumer graphics unit. Another approach would have been to fine-tune the existing GPT-1 model (trained over a more general corpus of text) on the HJC, which empirically has strong results for a variety of cases.

There is potentially an approach that *wouldn’t* require specialized training: using a model capable of understanding *instructions*. This is on the cutting-edge of NLP research, and open source code for the newest models in this area (like OpenAI’s GPT-3)[[4]](#footnote-4) aren’t available to the public, but there are publicly accessible interfaces to use the end result model such as OpenAI’s own Playground. There is something problematic and worrying about groups like OpenAI releasing models with such capabilities into the public without documenting details about how their model was created, but they did eventually release a paper documenting their method. The essence is that existing data models were turned into instructional learning tasks by creating instructions prompts for existing text (ie, feeding it instructions like “Write a sonnet” to accompany a dataset of sonnets) and training over that to gain an understanding of how instructions correlate to end products. In theory, an advanced enough model should be able to understand that when asked to write a story in Henry James’ style, it should write in a way mimicking his style of writing. GPT-3 shows remarkable progress towards this but does not quite succeed. When given the prompt “Write a story in the style of Henry James beginning with “Today on a walk I saw a dog”” it gives a very plain ordinary English story that does not reflect his writing style. Inputting the same prompt with other authors (like French post-modernist Jean Baudrillard or Austrian logician Ludwig Wittenstein, who also have distinctive writing styles) the model does much the same, generating another cookie-cutter story. But when given more direct prompts to imitate writers GPT-3 performs much better, such as amusingly parroting random lines of Wittgenstein’s TLP with incorrect numbering when asked to imitate the TLP. It even performs quite well in writing a short and quite amusing TLP like passage about a dog.[[5]](#footnote-5) But that same success did not translate to Henry James’ writing: both attempts to have GPT-3 imitate his writing directly and of a more specific work (like *Washington Square*) produced writing without his distinctive style. For that latter prompt, it produced a generic story but with somewhat similar themes to *Washington Square* about love and indecisiveness, which shows that while GPT-3 at least for Henry James picks up some of the content of his major works but struggles to reproduce his writing style. I can say with certainty that with fine-tuning on James’ work specifically it would perform much better at the task I trained the much more primitive GPT-1 to do, but such a project is outside of my scope of knowledge or resources. GPT-3 in the proper settings is capable of simply extraordinary results: it can understand word relations along long bodies of text, inferentially answer questions to a depth many humans would struggle with, and perform many tasks it was not even trained to do (such as translation).

Experimentations and Implications

The purpose of this project was not merely to demonstrate the power of LLMs (which is well understood and recognized) but to try to generate some insight into how humans and machines use language differently, and see if that insight can tell us anything about Henry James as an artist. This following section will discuss an experiment aided by the model previously trained, along with human participants.[[6]](#footnote-6)

Inspired by Dr. Juliet Floyd’s draft paper “Revisiting the Turing Test: Humans, Machines, and Phraseology”, I devised a test bearing resemblance to Alan Turing’s Turing Test. In its original formulation, a human is meant to talk to another human and a computer behind a obfuscatory veil where both the human and computer claim to be human; afterwards, the human participant lifts the veil. This has often been claimed as a test for *machine consciousness*, in line with an functionalist view of the nature of consciousness (if it can produce human-like outputs, it must have human-equivalent internal essence) but this is not something Turing explicitly defended himself or that I am interested in examining here. Instead, I want to investigate the sociological aspect of the Turing Test that Floyd elaborates on, where one takes or administers the test and then has a conversation with the other participant, to reveal the sociological implications behind such an encounter mediated by technology.[[7]](#footnote-7) Floyd even specifically argues that the use of chat bots and AI in common discourse mean there is always an “infinitely evolving picture” of Turing Tests all the time throughout social media and, increasingly, ordinary life.[[8]](#footnote-8) I suggest a reformulation of this evolution into one of *phenomenon* and *experiment*: if social interactions mediated by technology is the sociological phenomenon (or a group of them) we can think of the Turing Test as a classic scientific experiment isolating particular elements and seeing how they interact in a vacuum.

Having trained the smaller version of GPT-1 on the HJC, I generated samples of its writing around 3-4 sentences long, cleaned up the grammar and intelligibility of the writing as little as possible while making it grammatically legible, and took samples from real Henry James works to generate a blind A/B test of four comparisons, each of which is presented to the user as a choice between a Henry James original work and an AI produced piece. In reality, the third comparison was both artificial and the fourth comparison was both James. This intentional (but hopefully innocuous) deception was then revealed to the participant after the game to gauge their reaction and glean some insight into their decision making process, which should hopefully elucidate the differences between human and machine phraseology (as they currently exist). It also offered sociological insight: how did people feel about being deceived by AI produced work presented as human? It is worth questioning the validity of this “Turing Test” I have set up with regards to how it isolates social factors. There is something socially distinct about a clinical laboratory setting the Turing Test is usually administered in, and a different atmosphere is formed comparing the friendly relationships I had with the participants versus interacting with a distant, functionally anonymous stranger on the internet. Regardless I found some of the insights from the test amusing.

At the top level, were people able to tell the imitation James from the real deal? Yes, mostly. In the first two comparisons (each having a sample of each) no one I tested failed to distinguish between the first two after spending a minute or two reflecting at the material. People *did* mostly fail to recognize the deception in the third and fourth samples. When informed of that, people were mostly surprised, but realized in hindsight that they had greater struggles making a decision on the third and fourth comparisons due to their uniform quality. At least one person became skeptical of the initial problem formulation, asking if the comparisons could contain samples from the same source, which when it was revealed they quickly swept the challenge (this was also the person most familiar with Henry James).

The next relevant question is what this taught us (if anything) about James’ writing style. I thank Caroline Wall for the most insightful comment, which was that when reading James it is easy for one to fall into a “lazy reading” pattern where you skim the sentence for key nouns and adjectives, but don’t attempt to parse the sentence structure in your head. Especially with James’ more complex works (In the Cage and his essay on death come to mind as examples) his sentences can be near-unreadably complex for a naïve, untrained reader like myself. I would rather read the *Tractatus* again before I attempted to read either of those works again. This was initially a boom for the AI model but quicky began to hurt it, as it was incapable of incapable of actually reproducing Jamesian sentence structure due to its (relatively) low level of model complexity and training time, instead randomly throwing out semicolons and long adjectives with reckless abandon. This meant while at first blush or upon a lazy readying both samples appeared similar, a deeper inspection that required the reader to grasp the meaning of the sentence made the difference clear. Another problem was that while James always had some form of *intentionality* while writing (i.e., he was always writing about some subject or topic in particular) the model has no concept of “subject” and so even when semantically correct, will frequently talk about nonsense with no particular direction even when prompted on a particular issue. Many of these issues have been improved with more recent models, but this general behavior bears some crude parallels to how even the newer LLMs can produce very convincing or authoritative looking outputs on first blush that are often incredibly wrong.[[9]](#footnote-9) Other participants noted that the real James’ passages appeared focused on a single topic in a way the bot’s responses were not, tying back into the intentionality question.

To turn the question away from the technology and back onto people, what meaningful social interactions came as a result of the game? Not many. My testing subjects were people already amenable to me and to the nature of the experiment and general, and no one seemed truly upset I deceived them. I think it was simply amusing for them, which is a perfectly reasonable reaction.

One interesting thing I noted was that when briefly presenting this work for a class on Henry James and New Media I read a short passage generated by the bot in real time from a user prompt out loud, and the bot’s work received a small applause. I find that idea of applauding a machine, a uniquely human activity meant to show appreciation for work, fascinating. Applause has many dimensions of customs and different cultural norms surrounding it, but it nearly always contains *some* element of trying to convey appreciation for an artists or performers work.[[10]](#footnote-10) There was no artist or performer here in the sense that the model is probably not conscious, capable of artistic intent, or capable of understanding (or appreciative of) others’ reactions to its work. It may have learned from Henry James, but in a way quite dissimilar to humans (I know that when trying to imitate writers I do not perform copious amounts of matrix multiplication). Surely the audience was not applauding Henry James, but they also must have intuitively known the AI was not capable of receiving their applause. Perhaps it is a place where people simply adapt human customs out of reflex or intuition: upon hearing human-sounding work, they respond as they would to humans out of cultural instinct. It will be interesting to see how those customs adapt and change to AI’s specialties in the coming future.

Conclusions

For the lack of a better term, where are we meant to be left from all of this? It seems to me from this experience that natural language processing methods, despite their undoubtably impressive performance at many tasks[[11]](#footnote-11), have a ways to go before human writing or language retention concepts can be meaningfully grafted onto machines. A lack of current commercial or practical applications for this clearly powerful tech also point to this: in an optimization driven economy, why aren’t professional writers using these tools more often? Partially due to latent mistrust of AI, but also because the tools simply are not good enough yet to trust without external validation other than in small use cases. There has been great recent “moral panic” over the use of these models in academia[[12]](#footnote-12) and how it might redefine the writing or learning process for students. If a student need not learn how to write, or a teacher how to read, because of AI assisted tools, that would imply a grim future of both sides of the classroom spitting machine-generated and filtered garbage at each other in some sort of value-deprecating ouroboros. I suspect that while academia will have growing pains adapting to how people utilize these models, it will not cause such a dystopic vision for precisely the reasons the imitation Henry James model failed to trick people: its lack of intentionality in writing or understanding of textual meaning. Even at its full capacity, LLMs cannot write college-length essays like this one,[[13]](#footnote-13) and still struggle with the kinds of long relationships across text needed for human writing capabilities. Their nature of being essentially advanced parrots of existing works also means they often struggle with producing novel, creative solutions to problems. There is a reason one of these models has not independently performed scientific research or written a published novel. However, do not think we will not one day get there. The small model I trained for this assignment was state of the art a brief 5 years ago, and even in just the language field AI has advanced considerably. The meaningful task ahead will be to manage how people view and use AI socially as it “grows up” from a child, innocently copying what its parents say and do, to a reckless adolescent capable of independent thought but without the established cultural customs human children are bounded by and flourish through.

References and Acknowledgements

Thanks to Professor Juliet Floyd for a copy of her draft paper on Turing and enlightening conversations. Thanks to Professor Ashok Cutkosky for the codebase for the GPT implementation.

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

Floyd, Juliet. “Revisiting the Turing Test: Humans, Machines, and Phraseology”. For Nudging Choices Through Media – Ethical and philosophical implications for humanity, Eds. J.E. Katz, K. Schiepers and J. Floyd (Springer Nature Switzerland) Draft Version of 10/6/22.

Appendix

**GPT-3 Prompting Results** (original text highlighted, rest is in plain text)

Wittgenstein: (note how many of the lines are exact copies of real lines from the Ogden translation, but with incorrect enumeration!)

Write text imitating Wittgenstein's Tractatus Logico-Philosophicus.

4.1212 The world is everything that is the case.

4.221 What is the case—a fact—is the existence of atomic facts.

4.222 The logical picture of the facts is the thought.

4.23 The thought is the significant proposition.

4.24 The propositions of logic are tautologies.

4.241 The tautologies express the logical form of reality. They are the analytic propositions.

4.25 The propositions of logic determine the structure of reality.

4.251 Reality is the world.

4.26 The structure of the world is the totality of facts, not of things.

4.27 The facts in logical space are the world.

4.271 The world is the totality of facts, not of things.

Write text imitating Wittgenstein's Tractatus Logico-Philosophicus about a dog.

2.01 A dog is a mammal.

2.02 A mammal is an animal.

2.03 An animal is a living thing.

2.04 A living thing is a thing that is alive.

2.05 A thing that is alive is a thing that has life.

2.06 A thing that has life is a thing that is alive.

**James Turing Test**

Comparison 1:

Sample 1: The thing is that the stranger had has no success of his own in his situation, and has already, for the present, been in a way in which he is a complicative statement, of consequences, of a series of struggling lights, the admirable conversation of a great many things, and the personal compensation of a gentleman. The thing is that, in the countries, this stranger is the finest of the pleasing people, who have accepted something to say to him, something that would make him constantly intimate with his friend of the little prospect, who to all the world is the least bit desirable.

Sample 2: These things made their whole relation so impersonal that they hadn’t the rules or reasons people found in ordinary friendships. They didn’t care for the things it was supposed necessary to care for in the intercourse of the world. They ended one day—they never knew which of them expressed it first—by throwing out the idea that they didn’t care for each other. Over this idea they grew quite intimate; they rallied to it in a way that marked a fresh start in their confidence. If to feel deeply together about certain things wholly distinct from themselves didn’t constitute a safety, where was safety to be looked for?

Comparison 2:

Sample 1: It is a great pleasure to write the word; but I am not sure there is not a certain impudence in pretending to add anything to it. Venice has been painted and described many thousands of times, and of all the cities of the world is the easiest to visit without going there. Open the first book and you will find a rhapsody about it; step into the first picture-dealer’s and you will find three or four high-coloured “views” of it. There is notoriously nothing more to be said on the subject. Every one has been there, and every one has brought back a collection of photographs.

Sample 2: It is a great pleasure to write the word; but I am not sure there is not a certain impudence in pretending to add anything to it. There is nothing there to do with. The words, after a little, were the words that have broken off; and they are not the only words when they are, for those of the prospects in which they have not been the most performed. They are of the most serious and interesting people, but they are all crowded and accumulated; they are always of a cluster of men.

Comparison 3:

Sample 1: She had a capital and a crooked perversity, but when had she been a person of such a more serious disposition? Hadn't this taken him away and began to make her approach clear such that the dear old man suspected her, who had been so cleverly placed, as it were, in a position so discomfortable that she was in the most intensely sensitive way in an artist's place to see him simply best? In this case she might be with her sensibility, her profit in his connection, but not her faith now that she had now an inferior substitute.

Sample 2: She had stood face at him as if he were too far away from her and took in an intense sound from his face as if he had been absolutely aware of, that at any rate, her companion whose attitude to shirk had the apprehension of a sense of something sufficiently acute. She had seen him at the same time that he could see her, by the same turn of a certain possible challenge of his sister, too strongly to see how much he had expected in having her.

Comparison 4

Sample 1: The book is a singularly rich exhibition of an inward state, the state of private poetic intercourse with things, the kind of current that in a given personal experience flows to and fro between the imagination and the world. It represents the esthetic consciousness, proud of its conquests and discoveries, and yet trying, after all, as with the vexed sense of a want, to look through other windows and eyes. It goes all lengths, as is of course indispensable on behalf of a personage constituting a case.

Sample 2: What it has all amounted to, as I say, is that we have never known otherwise an agent so beautifully organized, organized from within, for a mission, and that such an organization at free play has made us really want never to lift a finger to break the charm. We catch at every turn of our present long-drawn crisis indeed that portentous name: it's displayed to us on a measureless scale that our enemy is organized, organized possibly to the effect of binding us with a spell if anything could keep us passive.

**Turing Test Answer Key**

Comparison 1:

Sample 1: JamesBot

Sample 2: Henry James, Alter of the Dead

Comparison 2:

Sample 1: Henry James, Italian Hours

Sample 2: JamesBot

Comparison 3:

Sample 1: JamesBot

Sample 2: JamesBot

Comparison 4:

Sample 1: Henry James, NOTES ON NOVELISTS WITH SOME OTHER NOTES

Sample 2: Henry James, Within the Rim and Other Essays

1. This statement applies even to people deeply familiar with computers (both on the software and architecture levels) or with contemporary machine learning algorithms – the notion of algorithm *explainability*, or understanding how a machine learning algorithm makes decisions or predictions, is an open area of research and one without universally applicable methods for every black-box algorithm. [↑](#footnote-ref-1)
2. It is worth noting that GPT is not bidirectional, and so cannot capture word dependence contextually if the word is used in a context only revealed later in a passage (like Teddy Bear vs Teddy Roosevelt) – more information here: <https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2>. There are bidirectional models like Google’s BERT model. [↑](#footnote-ref-2)
3. Clearly not entirely: one time the model generated the Gutenberg boilerplate text. [↑](#footnote-ref-3)
4. I began writing this paper before the release of ChatGPT, OpenAI’s newest model. If you wish, replace every instance of GPT-3 in your head with ChatGPT while reading. [↑](#footnote-ref-4)
5. For both, see appendix. [↑](#footnote-ref-5)
6. Thanks to BU students Caroline Wall and Will Chang, and others for participating in the test and giving their thoughtful insight. [↑](#footnote-ref-6)
7. Floyd 19. [↑](#footnote-ref-7)
8. Ibid 21. [↑](#footnote-ref-8)
9. See <https://analyticsindiamag.com/freaky-chatgpt-fails-that-caught-our-eyes/> for examples, including ChatGPT failing at inferential reasoning and advocating for sexual assault. [↑](#footnote-ref-9)
10. Exceptions might include funeral or memorial applause, as the individual is dead and so incapable of feeling accomplished. But that is a rare exception, is still directed at once living people, and is usually considered impolite. [↑](#footnote-ref-10)
11. I should clarify that these most recent LLMs are truly extraordinary: it is easy to find examples of them solving deep inferential problems, bug-fixing code in a way difficult for expert humans, and more. That is, when the model is not accidently “hallucinating”, or producing meaningful sounding garbage. [↑](#footnote-ref-11)
12. See <https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/>. [↑](#footnote-ref-12)
13. For an example, see here: <https://blogs.lse.ac.uk/impactofsocialsciences/2022/05/17/new-ai-tools-that-can-write-student-essays-require-educators-to-rethink-teaching-and-assessment/>. Another writer mentioned that this essay was “graduate level material”. I would wonder what graduate institution they attended that rewarded such simplistic writing, and suspect they were inflating somewhat. [↑](#footnote-ref-13)