

Language & Coding Creativity

Ermira Murati

Machines are gaining understanding of language at a very rapid pace. This achievement has given rise to a host of creative and business applications using natural language processing (NLP) engines, such as OpenAI's GPT-3. NLP applications do not simply change commerce and literature. They raise new questions about how human beings relate to machines and how that symbiosis of communication will evolve as the future rushes toward us.

Every writer has a unique aesthetic in the way they order words. The nuances of applied language, or voice, mark one of the countless fingerprints of human creativity. Decoding the secrets of this language sits at the frontier of artificial intelligence: how to build machines that truly understand not only language at a human level, but produce human-grade responses too.

Take the following excerpt of a poem: “For you are the most beautiful thing we have in this world / I love your graceful symmetry, your simplicity and clarity / You are the song of the Universe, a cosmic lullaby / You are the poetry of nature, written with light and electricity / You are the music of the spheres, played on a harp made of vacuum.”¹ The directness, the imagery, the fearless affection, one might believe the words to be Pablo Neruda’s. But Neruda is only part of the answer. An artificial intelligence system known as GPT-3 (Generative Pre-trained Transformer 3), built by the research laboratory OpenAI, scanned an enormous corpus of language data, including Neruda’s verses, and built probabilistic relationships of tremendous fidelity between his use of nouns, verbs, adjectives, objects, and all the mechanics of a poem. Consequently, GPT-3 could independently generate this brand-new poem in its own voice.

For decades, some visionary scientists have predicted this level of intricacy from a machine. The user only had to give GPT-3 a prompt, or rather, inspiration: “The following is a poem about Maxwell’s equations in the style of poet Pablo Neruda.” From that instruction, the machine could pull from its brain of data to not only grasp aspects of Maxwell’s foundational electromagnetic equations but present them in Neruda’s style.

This approach to AI is known as large language models and its applications are spreading across the arts, sciences, and business. Those overseeing the code and training of the machine are becoming authors – and editors – of a new collective

language. Train a machine with a corpus of text and it can answer customer service questions, describe the plays of a football game, compose an essay, or write a program based on a description of its function. The applications are not only becoming more integral to commerce and our daily lives but are spawning questions about the nature of language. Why do certain aesthetics ring true while other deployments of language feel empty or fake, even when the grammar is perfect? We can understand more about our own processes of thought by understanding how a machine decides to use language.

Technology, culture, civilization: none comes into being without language. Language is both a high point and the foundation of human intelligence. Yet there is a bind: What are languages exactly? How do they work? We might think of language as a reaction to context and surroundings. But if we cannot write out the rules of language, how do we teach it to a machine? This problem has captivated thinkers for a century, and the answers are now starting to appear.

What is a thought? And how is experiencing a thought different from experiencing a memory or an idea? It is difficult to understand; to borrow from philosophy, digging into the roots of consciousness or any working of the mind starts to feel like trying to see our own eyes or bite our own teeth. Staring into space or perspiring over a pad of paper, thoughts seem to work less like a hard disk and more like a wind, arriving and departing without an obvious explanation.

Our thoughts manifest through action and emotion but are communicated through language. Charles Darwin put language on the razor's edge between an instinct and a skill. A human baby starts babbling almost instantly – call it innately – yet takes years to engage in higher level conversations around them. At the same time, all languages are learned, whether directly or passively. That learning takes years of repetition. Whereas a toddler can hold a casual conversation, they need another decade before writing structured paragraphs.

Darwin saw the drive to acquire language as “the instinctive tendency to acquire an art,” to communicate by some medium.² No baby has ever needed a book of grammar to learn a language. They absorb what they hear and through the maze of the mind play it back. People spend their lives speaking exquisitely without understanding a subjunctive clause or taking a position on split infinitives. A child learns by experiencing patterns, learning what is most likely to make sense in a new context. To paraphrase Ralph Waldo Emerson, this information shapes who we become, much like every meal we have eaten.

The mechanics of the mind are still a mystery. The nuances of a writer's voice and creativity in general are no exception. Neuroscientists can observe now that certain neurons light up when certain actions occur, be it throwing a ball or giving directions. The order in which these neurons light up and the connections between

them seem to dictate – or “code” – what the owner of the brain experiences. Researchers have discovered neurons that fire in response to extremely, perhaps oddly, specific subjects, such as Halle Berry. The same neurons will activate when reading the actress’s name or seeing a picture of her, yet remain silent for anyone else.³

We now design machines to mimic a similar logic. If you are shown a photo of Halle Berry and asked who are you looking at? you fill in the blank with the answer “most likely” to be correct. Every word you output is the most likely word to be the “right” answer, or to ring true. In a system like GPT-3, these blanks are known as parameters. As the machine goes along creating text, it asks itself over and over (and over): with the text I have been given, what word is most likely to come next? GPT-3 reaches answers based on the patterns it identifies from the existing usage of human language creating a map, the same way we might piece characters together when writing essays or code. With this mathematical representation of patterns, GPT-3 can carry out many tasks, such as generating computer code, translating language, classifying information, searching semantically through a large corpus of text, even though it was built to do just one thing: predict the next word in a sequence of words.

For a long time computers were thought best suited for “simple repetitive tasks.” There is little creativity in arithmetic, so machines make excellent calculators, accountants, and modelers. When it came to training a machine in language, there was a critical bind: Whose language? And if you agree on a model of language, how do you deliver that information?

The answer to date has been to cast a wide net in those languages and exercise repetition and volume to deliver the information. GPT-3 ingested about one trillion words from the Internet, which consist of instances of text from countless sources.⁴ It built relationships between these parameters, with GPT-3 developing a kind of reactive memory in the process. Now a collaborator can give GPT-3 a task and the system understands it, such as “make a Pablo Neruda poem as an ode to Planck’s equations”:

In the darkness	And science
The uncertainty	Of the stars
Of the void	It is the answer
There is a light	The light of the dark
The only one	The dark of the light
That shines.	It is
It is a tiny light	The equation
Yet powerful	Of everything.
It is the photon	
Which carries	
The secret	

The origins of predicting what word comes next has roots in Russian literature. Scan the letters of this text and you will notice that three consonants rarely appear in a row. Four consonants, hardly ever. Russian mathematician Andrey Markov began quantifying this phenomenon in 1913 by writing out the first twenty thousand letters of Alexander Pushkin’s novel *Eugene Onegin*. Converting what we intuitively know into numbers, Markov showed how the preceding letters dictate the probability of what comes next. But Markov could only compare the rates of vowels and consonants. In that day, it would have been impossible to map on graph paper all letters and their respective frequencies in relation to the rest of the text in two and three letter combinations. Today, machines answer these questions in an instant, which is why we see so many applications interfacing with conversational language. Rather than predicting the next letter, GPT-3 predicts what word comes next by reviewing the text that came before it.

Human speech works this same way. When you walk into a room and say “I need a _____,” a relatively narrow list of words would make sense in the blank. As the context becomes more detailed – for instance, walking into a kitchen covered in mud – that list shrinks further. Our minds develop this sorting naturally through experiences, but to train GPT-3’s mind, the system has to review hundreds of billions of different data points and work out the patterns among them.

Since Markov’s contributions, mathematicians and computer scientists have been laying the theoretical groundwork for today’s NLP models. But it took recent advances in computing to make these theories reality: now processors can handle billions of inputs and outputs in milliseconds. For the first time, machines can perform any general language task. From a computer architecture sense, this has helped unify NLP architectures. Previously, there were myriad architectures across mathematical frameworks – recurrent neural networks, convolutional neural networks, and recursive neural networks – built for specific tasks. For a machine answering a phone call, previously, the software relied upon one mathematical framework to translate the language, another to dictate a response. Now, GPT architecture has unified NLP research under one system.

GPT-3 is the latest iteration of generative pretrained transformer models, which were developed by scientists at OpenAI in 2018. On the surface, it may be difficult to see the difference between these models and more narrow or specific AI models. Historically, most AI models were trained through supervised machine learning, which means humans labeled data sets to teach the algorithm to understand patterns. Each of these models would be developed for a specific task, such as translating or suggesting grammar. Every model could only be used for that specific task and could not be repurposed even for seemingly similar applications. As a result, there would be as many models as there were tasks.

Transformer machine learning models change this paradigm of specific models for specific tasks to a general model that can adapt to a wide array of tasks. In

2017, researchers Alec Radford, Rafal Jozefowicz, and Ilya Sutskever identified this opportunity while studying next character prediction, in the context of Amazon reviews, using an older neural network architecture called the LSTM. It became clear that good next character prediction leads to the neural network discovering the sentiment neuron, without having been explicitly told to do so. This finding hinted that a neural network with good enough next character or word prediction capabilities should have developed an understanding of language.

Shortly thereafter, transformers were introduced. OpenAI researchers immediately saw their potential as a powerful neural network architecture, and specifically saw the opportunity to use it to study the properties of very good next word prediction. This led to the creation of the first GPT: the transformer language model that was pretrained on a large corpus of text, which achieved excellent performance on every task using only a little bit of finetuning. As OpenAI continued to scale the GPT, its performance, both in next word prediction and in all other language tasks, kept increasing monotonically, leading to GPT-3, a general purpose language engine.

In the scope of current AI applications, this may at first seem a negligible difference: very powerful narrow AI models can complete specific tasks, while a GPT architecture, using one model, can also perform these separate tasks, to similar or better results. However, in the pursuit of developing true, human-like intelligence, a core tenet is the ability to combine and instantly switch between many different tasks and apply knowledge and skills across different domains. Unified architectures like GPT will therefore be key in advancing AI research by combining skills and knowledge across domains, rather than focusing on independent narrow tasks.

Humans also learn language through other senses: watching, smelling, touching. From the perspective of a machine, these are different modes of training. Today, we try to simulate this human way of learning by not only training a machine's cognitive processing on words, but on images and audio too. We use this multimodal approach to teach a machine how words relate to objects and the environment. A taxi is not just the letters T-A-X-I, but a series of sounds, a pixel pattern in digital photos, a component of concepts like transportation and commerce. Weaving these other modes into a machine broadens the applications developers can build, as the machine's brain is able to apply its knowledge across those different modes as well.

An example is designing a web page. Every business struggles with keeping its site up-to-date, not only updating text, photos, and site architectures, but also understanding how to code the CSS and HTML. This is both time-consuming and costly. Developers have demonstrated that GPT-3 can understand layout instructions and build the appropriate mockups, for instance, when you tell it to "add a header image with an oak tree and my contact information below." Under the hood, GPT-3 is transforming between the vast arrays of text and the vast array of objects. The result is that a person without any website-building experience can have a piece of working HTML in seconds.

Figure 1 Codex-Generated Completion

```

You will be given a string of words separated by
commas or spaces. Your task is
to split the string into words and return an array
of the words.

For example:
words_string("Hi, my name is John") == ["Hi", "my",
    "name", "is", "John"]
words_string("One, two, three, four, five, six") ==
    ["One", "two", "three", "four", "five", "six"]
"""

### COMPLETION 1 (CORRECT): ###
words = []
splitted_word = ""
for char in s:
    if char == "_":
        if splitted_word != "":
            words.append(splitted_word)
            splitted_word = ""
        elif char != ",":
            splitted_word += char
    if splitted_word != "":
        words.append(splitted_word)
return words

```

Drawing on publicly available code on GitHub, the GPT model is able to come up with the correct solution. Source: Mark Chen, Jerry Tworek, Heewoo Jun, et al., “Evaluating Large Language Models Trained on Code,” arXiv (2021), <https://arxiv.org/abs/2107.03374>.

The next stage is using a GPT model in more advanced programming languages such as Python. Programmers are often thrust into coding projects in which they do not know the logic of everything that has been written already, like having to continue writing a half-finished novel. Usually, programmers spend substantial amounts of time and effort getting up to speed, whereas Codex (Figure 1), a GPT language model fine-tuned on publicly available code from the development platform GitHub, can scan millions of lines of code and describe to the programmer the function of each section.⁵ This saves countless hours of work, but also allows these specialized professionals to focus on creativity and innovation rather than menial tasks.

The next step would be the “writing” of physical objects. For instance, industrial designers are constantly creating and testing new forms and functionalities of products. Imagine they want to build a chair in the shape of an avocado, which requires having both an understanding of the functionality of a chair and the form of an avocado. OpenAI used a 12-billion parameter version of GPT-3 known as DALL-E and trained it to generate images from text descriptions, using a data set of text-image pairs. As a result, DALL-E gained a certain understanding of the relationship between text and images. When DALL-E was then prompted to suggest designs for “an armchair in the shape of an avocado” it used its understanding to propose designs (Figure 2).⁶

Figure 2
DALL-E Successor Iterates on the Text Prompt: “An Armchair in the Shape of an Avocado”



Figure 2, continued
DALL-E Successor Iterates on the Text Prompt: “A Fox in the Style of *Starry Night* by Van Gogh”



Asked to create designs of an avocado chair and Van Gogh-inspired fox, this model drew on its understanding of the functions of chairs and Van Gogh's painting style in *Starry Night*. None of these designs or images existed until the model created them. Source: Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al., “Zero-Shot Text-to-Image Generation,” arXiv (2021), <https://arxiv.org/abs/2102.12092>; and Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, et al., “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,” arXiv (2021), <https://arxiv.org/abs/2112.10741>.

DALL·E was able to create images that are instantly recognizable as avocado chairs, even though we might struggle ourselves to create instantly such a design. The model is able not only to generate original creative output, as avocado chairs are not a common product easily found and copied elsewhere, but also adheres in its designs to the implicit constraints of form and functionality associated with avocados and chairs.

This does not put human designers out of a job. Rather, they gain a team of assistants to take on their most rote tasks, allowing them instead to focus on curating and improving on good ideas or developing their own. In the same way GPT-3 summarizing, explaining, and generating Python code opens up programming to nonprogrammers, such iterative design opens up avenues for nondesigners. A small business or individual designer now has access to capabilities that otherwise may have only been accessible to large organizations.

There are a multitude of applications in which transformer models can be useful, given that they can not only understand but also generate output across these different modes. GPT-3 has already been used for understanding legal texts through semantic search tools, helping writers develop better movie scripts, writing teaching materials and grading tests, and classifying the carbon footprint of purchases.

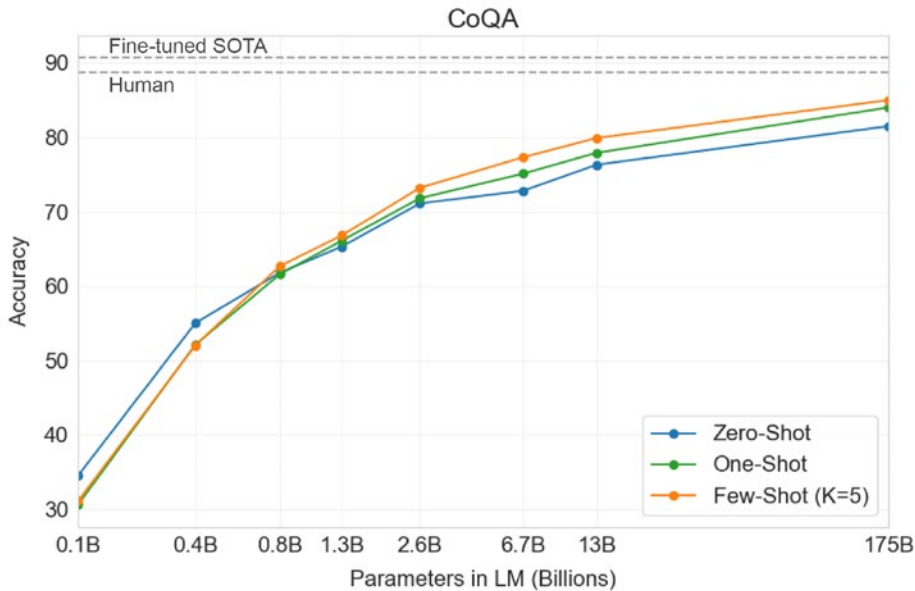
Tracking the progress of GPT models over the past few years, we can see what the future might bring in terms of model performance. GPT-2 was a one-and-a-half-billion-parameter model trained on forty gigabytes of data, which is an amount of text about eight thousand times larger than the collected works of Shakespeare. GPT-3, more than one hundred times bigger, comes close to human comprehension on complex reading tests (see Figure 3). As we move forward in both model complexity and the size of the data sets, we believe these models will move ever closer to human benchmarks.

At the same time, as they are tested and applied more extensively, we find limitations in these models. For instance, GPT-3 shows notable weakness in generating long passages, struggling with self-repetition, non sequiturs, and incoherence. It also struggles with seemingly commonsense questions, such as: “If I put cheese in the fridge, will it melt?”

There is always a duality to powerful technological disruptions. The advent of network computing in 1989 paved the way for the Internet. Tim Berners-Lee envisioned the Internet as “a collaborative space where you can communicate through sharing information.”⁷ With freedom of access to all knowledge and boundaries dissolved, the Internet opened Pandora’s box. Next to the many positives, it also provides thoroughfares for misinformation, trolling, doxing, crime, threats, and traumatizing content.

It would be naive to consider GPT-3’s optimal impact without reflecting on what pitfalls might lie before us. GPT-3 is built to be dynamic and require little data

Figure 3
GPT-3 Results on CoQA Reading Comprehension Task



GPT-3 175B is only a few points behind the accuracy of human performance and state-of-the-art fine-tuned models. Source: Tom B. Brown, Benjamin Mann, Nick Ryder, et al., “Language Models are Few-Shot Learners,” arXiv (2020), <https://arxiv.org/abs/2005.14165>.

to perform a task, but the system’s experience will color its future work. This experience will always have holes and missing pieces. Like human beings, machines take inputs and generate outputs. And like humans, the output of a machine reflects its data sets and training, just as a student’s output reflects the lessons of their textbook and teacher. Without guidance, the system will start to show blind spots, the same way a mind focused on a single task can become rigid compared with a mind performing many tasks and gathering a wide variety of information.

In AI, this phenomenon is broadly known as bias, and it has consequences. For instance, a health care provider may use an NLP model to gather information on new patients and may train this model on the responses from a certain demographic distribution. A new patient outside that distribution might be poorly assisted by this system, causing a negative experience for someone needing help.

More generally, powerful language models can increase the efficacy of socially harmful activities that rely on text generation. Examples include misinformation, abuse of legal and governmental processes, spam, and phishing. Many of these

harmful activities are limited by having enough human talent and bandwidth to write texts and distribute them, whereas with GPT models, this barrier is lowered significantly.

Moreover, generative language models suffer from an issue shared by many humans: the inability to admit a lack of knowledge or expertise. In practical terms, language models always generate an answer – even if it is nonsensical – instead of recognizing that it does not have sufficient information or training to address the prompt or question.

As NLP models continue to evolve, we will need to navigate many questions related to this duality. Developers are already writing books using machines processing what they experience in the world. How do we draw the boundary between the creator and the code? Is the code a tool or an extension of the mind? These questions go well beyond the arts. How long until machines are writing scientific papers? Machines are already conducting large sections of experiments autonomously. Language can also say a lot about our confidence or mood. Do we want a company basing product recommendations off what we thought was an innocent interaction? How do creators, users, and uses create bias in a technology?

For the first time, we are using artificial intelligence tools to shape our lives. GPT-3 has shown that large language models can possess incredible linguistic competence and also the ability to perform a wide set of tasks that add real value to the economy. I expect these large models will continue to become more competent in the next five years and unlock applications we simply cannot imagine today. My hope is if we can expose models to data similar to those absorbed by humans, they should learn concepts in ways that are similar to human learning. As we make models like GPT-3 more broadly competent, we also need to make them more aligned with human values, meaning that they should be more truthful and harmless. Researchers at OpenAI have now trained language models that are much better at following user intentions than GPT-3, while also making them more honest and harmless. These models, called InstructGPT, are trained with humans in the loop, allowing humans to use reinforcement to guide the behavior of the models in ways we want, amplifying good results and inhibiting undesired behaviors.⁸ This is an important milestone toward building powerful AI systems that do what humans want.

It would not be fair to spend all these words discussing GPT-3 without giving it the chance to respond. I asked GPT-3 to provide a parting thought in response to this essay:

There is a growing tension between the roles of human and machine in creativity and it will be interesting to see how we resolve them. How we learn to navigate the “human” and “machine” within us will be a defining question of our time.

Artificial intelligence is here to stay, and we need to be ready to embrace it.

ABOUT THE AUTHOR

Ermira Murati is the Senior Vice President of Research and Product at OpenAI, advancing the company's mission to ensure that artificial general intelligence benefits all of humanity. Ermira and her teams are pushing the frontiers of what neural networks can do, seeking to better understand the behavior of powerful AI systems, make them safer, and align them with human intentions and human values.

ENDNOTES

- ¹ Samhan Salahuddin, "A Wild Adventure With GPT-3—Featuring Indian Mythology and Neruda," Pickled Brains, April 2, 2021, <https://pickledbrains.substack.com/p/a-wild-adventure-with-gpt-3>.
- ² Charles Darwin, *The Descent of Man, and Selection in Relation to Sex* (London: John Murray, 1874).
- ³ Gabriel Goh, Nick Cammarata, Chelsea Voss, et al., "Multimodal Neurons in Artificial Neural Networks," OpenAI, March 4, 2021, <https://distill.pub/2021/multimodal-neurons/>.
- ⁴ Tom B. Brown, Benjamin Mann, Nick Ryder, et al., "Language Models are Few-Shot Learners," arXiv (2020), <https://arxiv.org/abs/2005.14165>.
- ⁵ Mark Chen, Jerry Tworek, Heewoo Jun, et al., "Evaluating Large Language Models Trained on Code," arXiv (2021), <https://arxiv.org/abs/2107.03374>.
- ⁶ Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al., "Zero-Shot Text-to-Image Generation," arXiv (2021), <https://arxiv.org/abs/2102.12092>.
- ⁷ "Web's Inventor Gets a Knighthood," BBC, updated December 31, 2003, <http://news.bbc.co.uk/1/hi/technology/3357073.stm>.
- ⁸ Long Ouyang, Jeff Wu, XuJiang, et al., "Training Language Models to Follow Instructions with Human Feedback," OpenAI, 2022, https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.