

An economist's guide to teaching students about, and with, ChatGPT and other Large Language Models

Sean Masaki Flynn
Department of Economics, Scripps College, CA, USA

ABSTRACT

The optimal strategies for teaching with ChatGPT and other large language models depend upon a basic understanding of how artificial neural networks operate, which I contribute in the first section of this paper. I then discuss teaching with ChatGPT, including its usually robust capabilities with respect to: grading student work and providing student feedback; its use as a private economics tutor; and its ability to be “interviewed” on behalf of historical or living people. Examples are then shown of the difficulties it has solving certain math and algebra problems that are relevant to economics pedagogy. I conclude with strategies for integrating ChatGPT into student study flows.

KEYWORDS

artificial intelligence, AI, large language models (LLMs), ChatGPT, GPT-4

JEL CODES

A20, O33

CONTACT

sflynn@scrippscollege.edu. Associate Professor, Department of Economics, Scripps College, 1030 Claremont, CA 91711, USA.

ChatGPT and other large language models (LLMs) present both numerous challenges as well as numerous opportunities for economics educators and economics students.

This paper is based on the premise that the magnitude and nature of those challenges and opportunities are most readily understood by first “taking a look under the hood” to see how ChatGPT generates answers to user queries and then examining the accuracy and precision of those answers.

That two-step process is also useful because it helps to dispel student fears about “artificial intelligence (AI) taking all the jobs” or “AI becoming conscious and destroying humanity.”

Understanding how ChatGPT works

The simplest way to dispel student fears about Chat GPT is to show them how it works. That is especially true because even a basic understanding of its inner workings will satisfy most students that ChatGPT is neither alive nor thinking.

An understanding of ChatGPT’s innards is also useful for economics students because once they understand that ChatGPT is inherently stochastic, they will also understand that the text that it generates is inherently inconsistent and, thus, of dubious reliability when it comes to the summation, reiteration, or integration of data and facts. From there, it is one short step to the realization that ChatGPT is an untrustworthy co-conspirator when it comes to answering homework and exam questions.

But before explaining how ChatGPT works, let’s explain its name, place it into context with other systems that have similar capabilities, and explain why its arrival came as such a shock.

Why is it called ChatGPT?

ChatGPT is an on-line interface that allows a person to “chat” with an artificial neural network. A person using ChatGPT types written text into the interface and ChatGPT’s artificial neural network responds with output text that it generates on the fly in response to the input text.

GPT an acronym for for Generative Pre-trained Transformer. ChatGPT is *generative* because it generates text. It is *pre-trained* because its artificial neural network was pre-trained on billions of words of text scrapped from the Internet and from digitally scanned books. It is a *transformer* because its artificial neural network uses transformer architecture, which gives a neural network the ability to identify subtle, non-local relationships in sequential input data, such as subject-verb agreement between two words that are separated by other words (e.g., “The dog, which was old and mangy, ran to the bone.”)

Is ChatGPT unique?

No, ChatGPT is not unique. ChatGPT is one of dozens of large language models, or LLMs, that have been developed by private companies, university researchers, and governments.

LLMs are a type of artificial intelligence that have been trained to recognize patterns in large text repositories in order to be able to use those patterns to generate responses to user queries.

The most prominent LLMs as of mid-2023 were:

1. ChatGPT (aka, GPT-3.5) and GPT-4, both developed by OpenAI, of which Microsoft is set to purchase 49%.
2. Bard/LaMDA and PaLM, which are being developed by Google.
3. LLaMA, by Meta's AI division.
4. Claude, which is being developed by Anthropic, of which Google paid \$300 million for 10 percent ownership.

In March of 2023, Microsoft and Google separately announced that they were going to integrate LLM technology throughout their respective suites of business productivity applications—Microsoft 365 and Google Workspace—as well as their respective Bing and Google search engines. Thus, ChatGPT and other generative AIs are expected to be ubiquitous by the end of 2023.

What made ChatGPT so surprising?

ChatGPT astounded the world upon its debut in November of 2022 because it was the first AI made available to the public that seemed to be able to pass the Turing Test most of the time.

The Turing Test is a measure of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. It was proposed in 1950 by the British mathematician and computer scientist Alan Turing (1950).

In the test, a human judge interacts with a machine and another human through a computer interface. Both the machine and the human try to convince the judge that they are the human. If the judge cannot reliably tell which participant is the human and which is the machine, then the machine is said to have passed the Turing Test.

It should be noted that the Turing Test focuses on a machine's ability to exhibit intelligent conversation that is indistinguishable from a human's, rather than the machine's ability to correctly answer questions or perform tasks. But ChatGPT excelled at the later activities, too, since it could both correctly answer many questions and competently perform a wide variety of tasks (such as writing computer code or translating a passage written in English into another language.)

What made ChatGPT so impactful?

While ChatGPT did not always perform tasks correctly, it was correct often enough that it was immediately apparent that generative AI might be a massive productivity booster in many industries.

Those productivity increases didn't take long to arrive. Within a few months of its debut, Kelly (2023) and others reported that real estate agents were using ChatGPT to convert raw data on specific properties (such as square footage, number of bedrooms, and location) into descriptive

text paragraphs that could be used in advertisements (e.g., “This 2500 square foot home with 4 bedrooms is located in the fashionable Ridgeview section of town, near plentiful shopping and a large public park.”)

Fiction writers, too, saw their output massively increase thanks to ChatGPT’s ability to not only write text that could be directly plagiarized but to also, less controversially, provide feedback that could help with specific aspects of the human writing process, such as brainstorming, characterization, and the transformation of blocks of text from one writing style (e.g., Faulkner) to another (e.g., Hemmingway).

The increase in output was so large that the competitive science fiction magazine *Clarkesworld* had to ban all submissions until it could figure out a way to distinguish submissions written entirely by humans from those that appeared to be written either mostly or entirely by ChatGPT (Hern 2023).

With respect to economic growth and technological advance, a similarly large, but more economically consequential, productivity explosion happened among computer programmers because ChatGPT could be asked to write code in any programming language and could, in response to such requests, typically generate code that either worked perfectly immediately or only needed a few edits here and there to work as intended. Within months, Noy and Zhang (2023) reported that the use of ChatGPT as a programming aid boosted developer productivity (the time to complete a given piece of code) by 40 percent and the quality of code (as judged by outside experts) by 20 percent.

How does ChatGPT work?

To use ChatGPT, a user simply types a question or a statement into a text-input window on a web browser. That text is routed over the internet to servers at Open AI.

The text is then divided into “tokens,” which are sometimes entire words, but usually parts of words, such as “-ing” or “-er” or “-s” that are used semantically by human brains to indicate things like verb tenses, core meanings, and grammatical inflections.

The artificial neural network that underlies ChatGPT generates a response by taking that string of input tokens and using it to generate a string of response tokens, a single token at a time. That series of response tokens is then reassembled into words, one word at a time, and those words are delivered back over the internet to the user. They appear one at a time, in sequence, as though ChatGPT were text messaging with the user in real time, typing out each word of its response one word at a time (but not one letter at a time, since ChatGPT generates its responses at the token/word level, not at the letter/numeral level).

Artificial Neurons Artificial neural networks consist of two or more artificial neurons that are connected to each other. The connections allow artificial neurons to process incoming information and pass a resulting signal on to one or more additional artificial neurons that can themselves process incoming signals and generate output signals that can be further passed along across the artificial neural network. That collective processing and passing on of signals allows an artificial neural network to compute outputs from inputs. Applications include optical

character recognition (i.e., taking in digital visual signals and outputting a verbal description of what was being looked at, such as the letter A) and next-word prediction (i.e., taking in a finite string of words and then predicting the most likely next word).

Figure 1, which is borrowed from Wang (2019), compares a biological neuron with an artificial neuron.

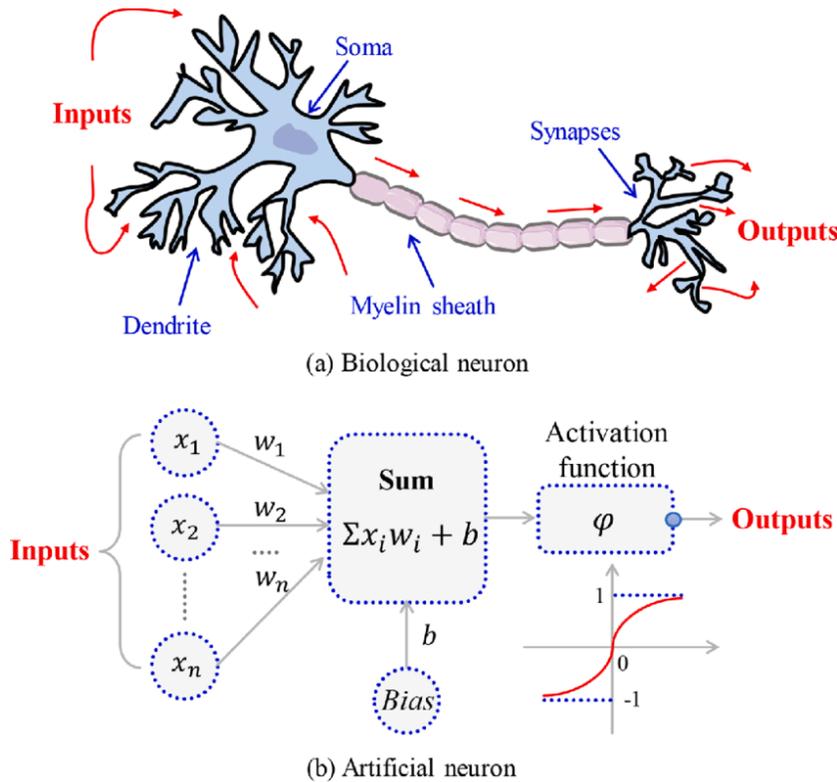


Figure 1. Comparison of (a, top figure) a biological neuron with (b, bottom figure) an artificial neuron.

As is illustrated in Figure 1, both the biological neuron and the artificial neuron absorb input signals on the left, process those signals in the middle, and then generate an output signal that exits on the right.

With respect to a biological neuron that is part of an animal brain, the absorption of signals on the left is done by dendrites. Those input signals are processed by the soma, and if they collectively provide enough stimulus to exceed a fixed threshold, the soma outputs an electrical signal that travels to the right (within the myelin sheath) to the synapses, which output that signal in chemical form by emitting neurotransmitters like dopamine that can then be detected by the dendrites of other neurons within the brain.

With respect to an artificial neuron that is part of an artificial neural network, signals x_i ($i = 1, 2, 3, \dots, n$) that arrive electronically on the left are given respective “synaptic weights” w_i , and then the dot product of the x_i ’s and w_i ’s is added to a so-called bias parameter, b , to produce an

output $\sum w_i x_i + b$ that is then evaluated by an activation function φ that generates an output signal that can be passed on electronically to other artificial neurons.

The bias parameter b exists to mimic the fact that biological neurons vary in their sensitivity or “excitability” with respect to incoming stimuli. Two biological neurons can input the exact same signals on the left, but one biological neuron may be more likely than the other neuron to become heavily “excited” by those input signals and decide to pass on an output signal. The bias parameter b is the artificial neuron’s equivalent of excitability in a biological neuron. The larger is the bias b , the larger will be the sum $\sum w_i x_i + b$ that acts as the input of the artificial neuron’s activation function. The larger that input, the larger will be the activation function’s output to other artificial neurons within the artificial neural network.^{1,2}

Artificial Neural Networks Artificial neural networks link together two or more artificial neurons. An example is provided in Figure 2, which is taken from Holmgren et al. (2019).

¹ In Figure 1(b), the activation function $\varphi = \frac{1}{1+e^{-z}}$ has an input domain ranging from $-\infty$ to $+\infty$, so that it can accommodate any real-valued input number $z = \sum w_i x_i + b$. This type of activation function is known as a sigmoid activation function and its attractive feature is that its output signal (which is then passed on to other artificial neurons) is bounded between -1 and +1. A variety of monotonically increasing functions have been used as activations functions, however, including arc tangent functions, identity functions, and, more successfully, rectifier or “ReLU” (rectifier linear unit) functions.

² A key difference between biological neurons and artificial neurons is that biological neurons are always binary in their output. That is, the input stimuli received by a biological neuron is either strong enough that it exceeds the activation threshold beyond which it will output a signal of a fixed intensity or below the activation threshold, in which case the biological neuron will send no output signal at all. By contrast, artificial neurons are generally set up so that they do not have discrete activation thresholds. Thus, they are continuous rather than discrete in their output, since activation functions are generally continuous mappings of the weighted input signals plus the bias parameter ($z = \sum w_i x_i + b$) onto some continuous subset of the real numbers, as with the sigmoid activation function $\varphi = \frac{1}{1+e^{-z}}$ shown in the previous footnote.

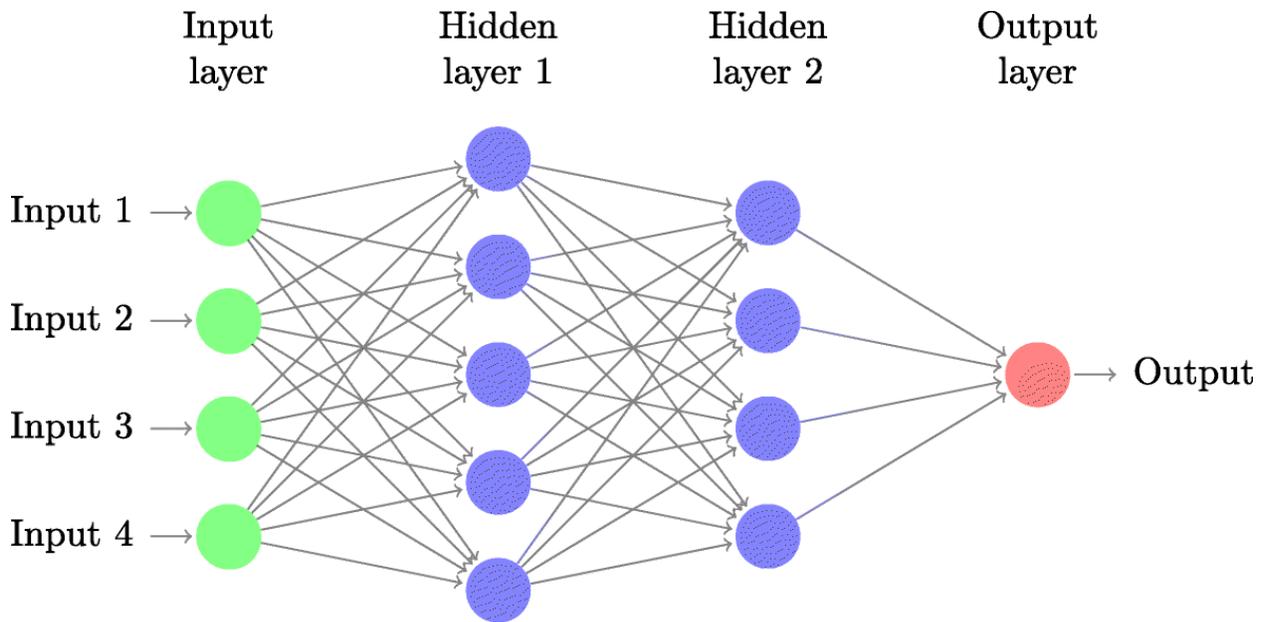


Figure 2. An artificial neural network with an input layer, two hidden layers, and an output layer containing a single artificial neuron.

Figure 2 has 14 artificial neurons. Four are in the input layer, nine are located within two so-called hidden layers, while the output layer contains just one artificial neuron.

The names *input layer*, *hidden layer*, and *output layer* that are used to classify the artificial neurons located within an artificial neural network correspond to analogous groups of biological neurons either connected to, or located within, a physical animal brain. The artificial neurons in an input layer are analogous, for instance, to light-sensing neurons in the retina of the eye or touch-sensing neurons at the tip of a finger. They receive input data from the external world and decide whether the intensity of that data warrants passing on a signal to other neurons within the biological neural network.

By contrast, the artificial neurons located within the hidden layers of an artificial neural network receive information not from the outside world but rather only from other artificial neurons. The artificial neurons within the hidden layers also do the bulk of the artificial neural network’s signal processing. That is, a hidden-layer artificial neuron accepts input from one or more prior artificial neurons, processes those signals, and then generates an output signal that will be sent to whichever posterior artificial neurons it may be connected to, as per the arrows that connect the artificial neurons in Figure 2.

The term “hidden layer” refers to the fact that, until the past few decades, it was impossible to tell in real time what was going on within or between individual neurons located within a living biological brain. It was known that individual neurons processed signals electrochemically and that neurotransmitter chemicals like dopamine and serotonin passed signals between neurons across synaptic gaps (i.e., the physical gaps between the synapses of a neuron sending a signal and the dendrites of a connected neuron receiving that signal). But what was happening moment-by-moment was impossible to surmise until the development of new techniques (like implanting

an extremely thin gold electrode into a single biological neuron) allowed scientists to measure the electrical activity of individual neurons in real time.

Until those technologies arrived, the brains of living creatures were for the most part “black boxes” whose real-time information-processing behaviors were inaccessible. Hence the term *hidden layers* to describe neurons whose real-time behavior was previously impossible to determine due to technological limits on the measurement of neuron-level electrical and chemical activity.

Finally, please note that it is for simplicity only that Figure 2 displays an output layer consisting of just one artificial neuron. Artificial neural networks can, in fact, be designed to have any number of artificial neurons within the output layer, just as they can be designed to have any number of artificial neurons within the input layer or the hidden layers. The number of connections between artificial neurons can also be adjusted at will, with some neurons having numerous incoming and outgoing connections while others may only have a single input connection and a single output connection. But whatever their number, the artificial neurons in the output layer export (output) the iterative total of all the signal processing and transmission that occurs previously within the input layer and hidden layers.

Processing Information to Make Predictions Figure 3, taken from Zhou (2019), provides a simple example of how the information-processing capabilities of an artificial neural network can be put to beneficial use making predictions.

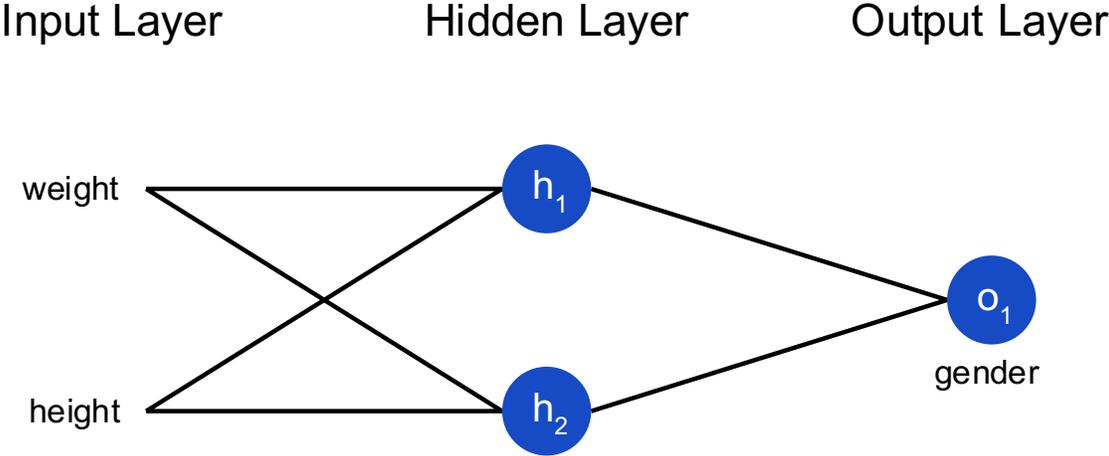


Figure 3. A simple artificial neural network designed to output (i.e., predict) gender given inputs of weight and height.

The artificial neural network depicted in Figure 3 is designed to accurately predict gender based on data related to height and weight. It could be applied to scenarios in which there is a dataset containing height measured in centimeters, weight measured in kilograms, and gender represented by a binary dummy variable (1 for male and 0 for female).

Moving from left to right within Figure 3, the initial signals fed into the input layer would be the respective numerical values for the weight and height of a single person. Those would be weighted by w_i 's and then combined with bias parameters within each of the two hidden-layer artificial neurons (h_1 and h_2) which would then pass signals on to the output neuron (o_1) which would itself output either a 1 for male or a 0 for female as its prediction for the gender of that specific person.³

The trick to making a correct prediction is to “train” the artificial neural network to provide the correct answer no matter which person’s weight and height are initially fed into the network. To do so, one must adjust the weighting parameters w_i and the bias parameters b within the network so that the output signal generated by the output neuron is correct most or all the time no matter which individual’s height and weight data is fed into the network.

Training Artificial Neural Networks to Make Correct Predictions What is amazing about artificial neural networks is that it is indeed possible to adjust the weighting parameters w_i and bias parameters b such that the final outputs are in fact usually or always correct, at least relative to the data set used to train the network.

Figure 4, taken from Ruscica (2019), gives an example of an artificial neural network whose parameters (i.e., synaptic weights and bias parameters) have been adjusted so as to given correct output predictions.

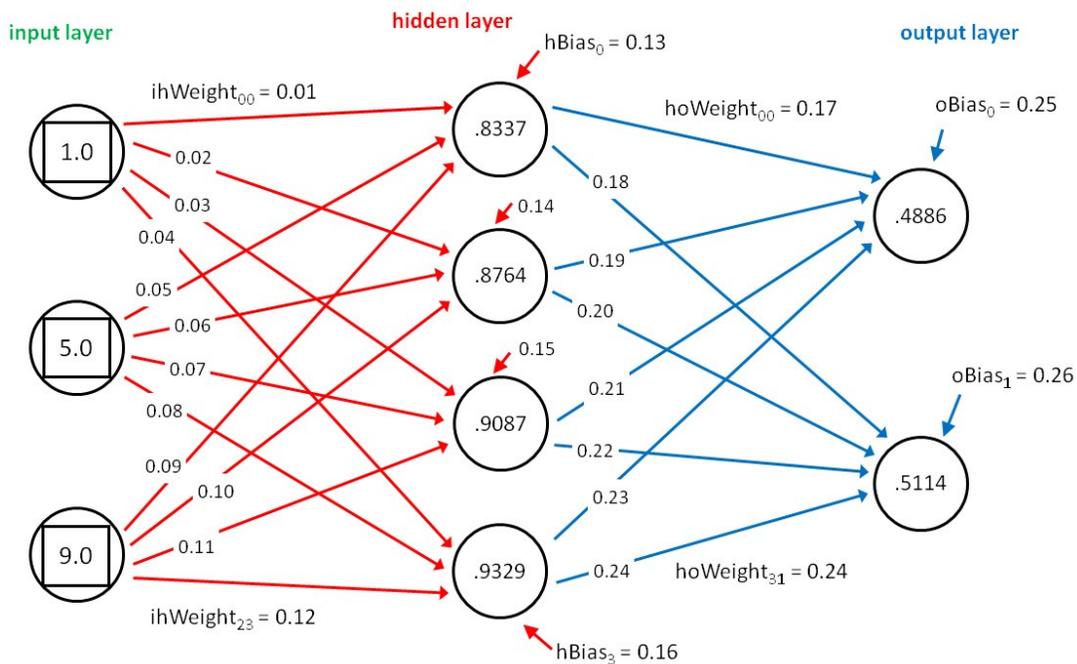


Figure 4. An artificial neural network with specific values for inputs, synaptic weights, and bias parameters.

³ That strictly binary output (1 for male and 0 for female) is possible because there are binary activation functions that will output only either 0 or 1 given real-number input signals.

Starting from the left side of Figure 4, the numbers within the squares within the three input-layer artificial neurons are the respective numerical values of the three initial input signals. The arrows in the figure indicate the connections between the various artificial neurons within the network, with each arrow being tagged with its corresponding weighting value. You can also see the bias values for each artificial neuron as well as the output value that each artificial neuron will pass along through the system to any neurons subsequently connected to it.

But how does a researcher train an artificial neural network to end up with a collection of weights w_i and bias parameters b that lead to the output layer usually or always making the correct prediction? One method starts by assigning random values to the weights and bias parameters. You then see how well the network does at making predictions given those values, with your measure of correctness being a mathematical loss function, such as the sum of squared prediction errors.

You then adjust the weights and biases iteratively to reduce the overall prediction error. The most common way of doing this is to implement a so-called gradient-descent search, in which you have a computer program sequentially attempt small changes in the synaptic weight and bias parameters to “follow the gradient” to smaller and smaller values of the loss function. By doing that over and over enough times, you will converge to a set of weight and bias parameters that is at least a local minimum of the loss function. If you are lucky, that local minimum will also be the global minimum, but there is no way to know if that is in fact the case.

What is definitely true, however, is that given enough training data, a suitably constructed artificial neural network can often work quite well in the sense that its predictions/outputs are almost always correct, at least with respect to the data in its training set. Sometimes researchers get even more lucky because the specific weight and bias parameters achieved after a gradient-descent search also appear to make good predictions when applied to out-of-training-sample data.

That ability to make good predictions out-of-sample would be why it would be useful, for example, to make and train the artificial neural network illustrated in Figure 5.

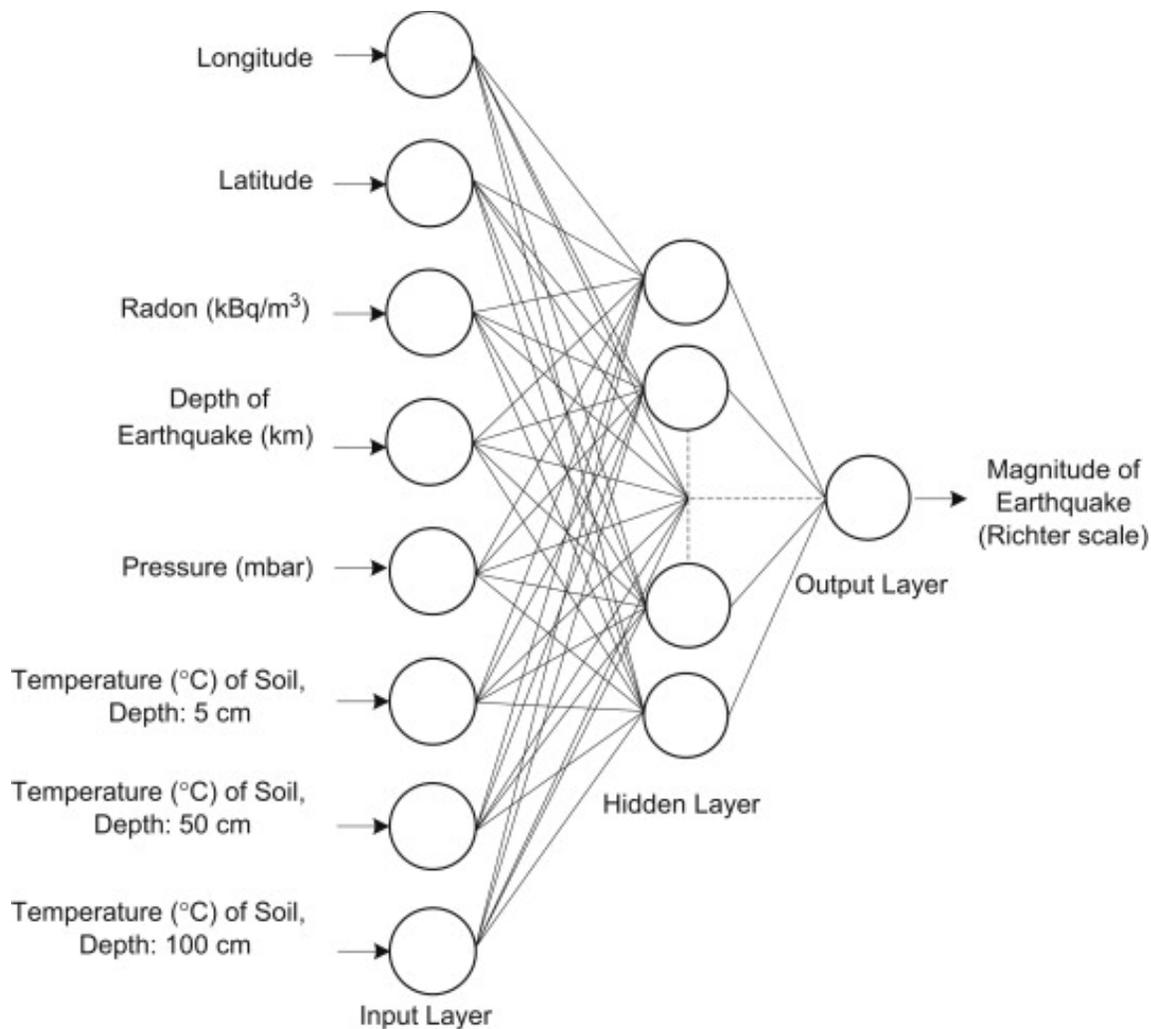


Figure 5. An artificial neural network designed to predict the magnitude of an earthquake given eight inputs, including longitude, latitude, depth, and soil temperature.

Once trained, the artificial neural network shown in Figure 5, which is taken from Kulahci et al. (2009), will be able to generate an out-of-sample prediction about the magnitude of a given earthquake given data on its location and several other input values such as soil temperature. That prediction could be useful for researchers in cases where the input values for a given earthquake are known but not the strength of the earthquake itself. That could be the case, for example, when no direct seismographic measurements of an earthquake’s magnitude were taken or if the records of such measurements were accidentally deleted or destroyed after being recorded.

Before moving on, it should be pointed out that one of the great benefits of an *artificial* neural network is that it is fully transparent. Unlike biological neural networks, there is in fact nothing hidden within the “hidden layers” of an artificial neural network. That is true because we always know every numerical value attached to every element of an artificial neural network, including all of the input signal values as well as the numerical values of every weight and bias parameter.

What is not understood, metaphysically speaking, however, is exactly why a given set of weights and biases can make good predictions in so many cases. But the fact that they *can* make good predictions makes artificial neural networks extremely useful, as is the case with ChatGPT and other large language models.

ChatGPT’s Artificial Neural Network Estimates Next-Word Probabilities ChatGPT’s artificial neural network was trained on billions of words of text scrapped off the internet and taken from scanned library books. The network was trained to take a string of words provided by the human user and predict the most likely next word.

Our human brains can perform this “next-word prediction task” quite easily depending on the input words. For example, if an American adult were given the words “star spangled,” and asked to predict the next word, they would almost certainly say “banner” since that word follows “star spangled” in the U.S. national anthem, the lyrics of which are familiar to most Americans.

Figure 6 illustrates how ChatGPT performs the next-word prediction task when given as input the words, “The best thing about AI is its ability to...”. This example is taken from Wolfram (2023), who shows that the raw output of ChatGPT’s neural network is a rank ordering, by likelihood, of every possible next word in the English language. The input text (“The best thing about AI is its ability to...”) is absorbed by ChatGPT on the left and then processed by the neural network in the middle using transformer architecture. That processing generates output on the right in the form of a vector containing the probability value for every possible next word.

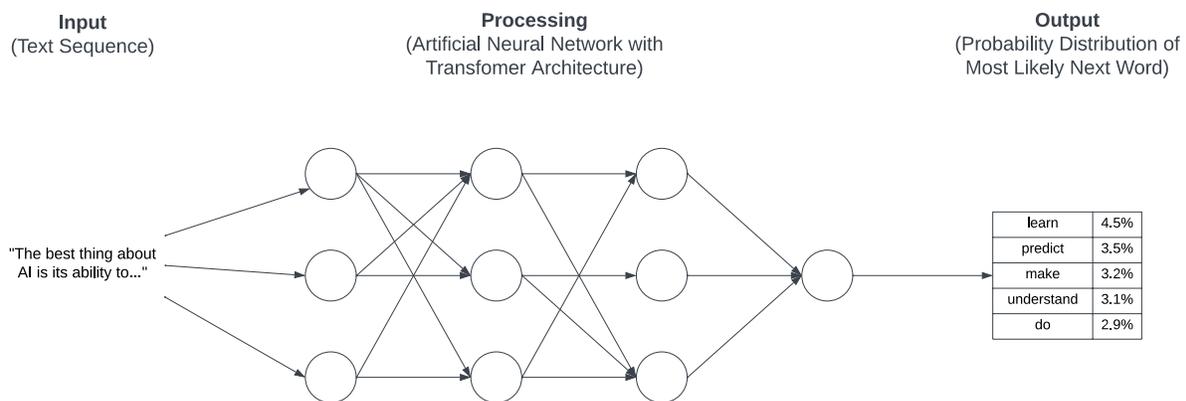


Figure 6. A simplified flowchart illustrating how ChatGPT inputs a finite text sequence, processes that sequence through its artificial neural network, and outputs a vector giving the probability distribution of the most likely next word.

For illustrative concision and tractability, the table on the right side of Figure 6 shows only the five most likely next words as predicted by ChatGPT’s neural network as well as the probabilities that ChatGPT assigns to each of those five words in terms of how likely they would each be to be the next word given all the text that ChatGPT’s neural network was trained on. As

you can see, the most likely next word is “learn” with a 4.5 percent probability, followed by “predict” with a 3.5 percent probability.⁴

When ChatGPT generates a text response to a user input, it strings together a series of next words based on these next-word probability rankings. Crucially, however, ChatGPT does *not* always select the most probabilistically likely next word at any junction. Rather, ChatGPT selects the next word at random from among the most likely candidates. Thus, in the example in Figure 6, ChatGPT would not necessarily select the word “learn” just because it is the most probabilistically likely. Rather, ChatGPT would select the next word at random from the most likely candidates.

The reason for this randomization is because it has been found that the text generated by large language models ends up being dull, flat, and inhuman if the No. 1 most likely next word is always selected. By randomizing across the most likely possible next words, ChatGPT can string together answers that feel both more natural and more interesting.

Please take note, however, that this randomization procedure for generating responses means that ChatGPT is inherently stochastic. If given the same prompt over and over, it will respond each time with a different string of words. This fact should in my opinion be made well known to students since any among them who are interested in cheating will be disappointed to know that ChatGPT is inherently inconsistent, stringing together answers stochastically. Thus, if they are looking for *the* right answer, ChatGPT will not deliver.

How ChatGPT Strings Together Sequences of Words To have a broader understanding of how ChatGPT generates responses, imagine that a user begins a new session and types in the following question for ChatGPT to consider: “What is the best thing about AI?”

ChatGPT would begin its response by absorbing that input string, running it through its neural network and then examining the probability ranking of all possible next words. Suppose that out of the most likely possibilities, it selects the word “The.” Notice that this word is capitalized since ChatGPT will implicitly know that its response should start with a capitalized first word. That is, “The” with a capital T was among the most likely next words in this case where the first word of a response sentence is called for, but “the” with a lower-case t was not among the most likely next words since words with lower-case initial letters are not likely to start a response sentence.

So ChatGPT begins its response with the word “The”. ChatGPT will now repeat the next-word prediction procedure, but this time including “The” as one of the input words. That is, to predict the second word in its response, it will use as its input not only the six words given by the user initially (“What is the best thing about AI?”) but also the word “The” as a seventh word of what is now a seven-word input sequence (“What is the best thing about AI? The...”).

⁴ The table in Figure 6 only shows the top five words, but ChatGPT would in fact generate a much longer vector of next-word probabilities.

ChatGPT’s artificial neural network will then produce a probability ranking of the most likely next word given those seven words of input and ChatGPT will chose one of the highest-ranking options from that list as the next (i.e., eighth) word.

Suppose that the eight word is “best.” ChatGPT will then go on to predict the ninth word using that eight word “best” attached to the previous seven words. That is, the input for the next round of next-word prediction will be “What is the best thing about AI? The best...”.

This procedure of stringing together next words based on all previous words is limited only by the number of tokens that an LLM can absorb as inputs in the input layer of its artificial neural network. ChatGPT’s limit is 4,096 tokens, but other LLMs can absorb longer strings of tokens. GPT-4, for example, can handle 32,000 tokens, or about 25,000 words of English-language text.

ChatGPT (and other LLMs) will complete a response after repeatedly engaging in this sort of next-word prediction. The response will terminate at some point and not go on forever because one of the next-word prediction possibilities at every round is an end-of-sentence token (such as a period symbol or an explanation mark) corresponding to the end of its answer.

For further elaboration on how ChatGPT strings together responses, Figure 7 presents a more elaborate version of Figure 6.

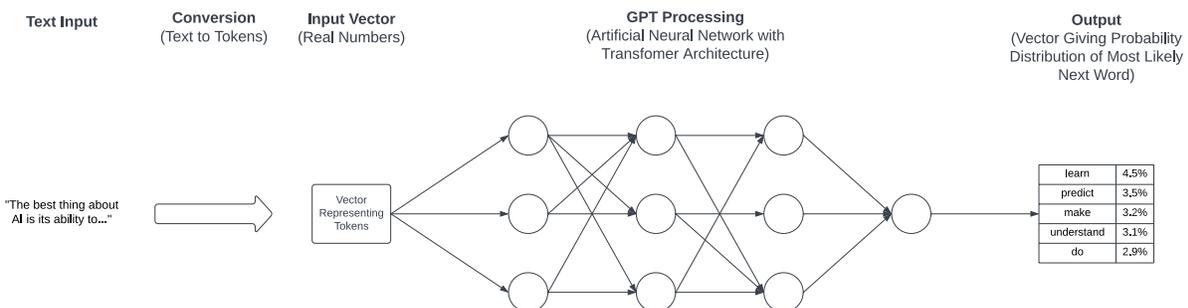


Figure 7. A more elaborate flowchart indicating how ChatGPT processes a given text input into a vector that represents the tokens contained in that input before processing that data through its artificial neural network, which outputs a vector giving the probability distribution of the most likely next word.

The additional information presented in Figure 7 makes it clear that the text input provided by a user is first converted to tokens before being processed by ChatGPT’s artificial neural network. Also made explicit is that the output of ChatGPT’s neural network is a vector giving the probabilities of each possible next word.

How ChatGPT was Trained

Before moving on to discussing the pedagogical aspects of ChatGPT as applied to economics, let’s explain how OpenAI trained ChatGPT and its successor, GPT-4.

First note that the artificial neural networks involved are enormous. GPT-4's neural net, for example, contains around 100 billion artificial neurons spread over 100 layers. Connecting those 100 billion artificial neurons are around 100 trillion artificial synapses (i.e., connectors between artificial neurons), or about 1,000 artificial synapses per artificial neuron.

As noted above, training began with randomly assigning synapse weights and biases. ChatGPT was then asked to make next-word predictions on text that it was given as input. Weights and biases were then adjusted automatically based on how well it did at predicting next words. And, then, after all that so-called “unsupervised” self-training was completed (so as to minimize the loss function), a second round of “supervised” training was done in which human beings (the supervisors) read the output produced by ChatGPT in responses to prompts typed in by human beings and rated it in terms of things like naturalness and fluidity, since the goal was to produce an LLM that could “chat” like a human being. That human feedback was used to further adjust the network's synapse weights and biases.⁵

The result of the unsupervised self-training (by means of next-word prediction) and the supervised training (with human feedback) was a numerical set of synapse weights and biases that was then frozen, the result being a pre-trained transformer LLM that could be released to the public.

Note that since the synapse weights and biases are frozen, ChatGPT cannot “learn” from its interactions with users. As a session goes on longer, it will of course have more cumulative tokens (up to its maximum token-input limit) serving as input for next-word prediction. That will give the illusion that ChatGPT “remembers” what was said earlier in a session. But the truth is that the synapse weights and biases are frozen, so that all that is happening is that a longer token-input vector is being presented to ChatGPT's artificial neural network each time a new query is made.⁶

Teaching with ChatGPT and other LLMs

Now that we have a shared understanding of how generative AI systems like Chat GPT produce their responses, let's talk about teaching with ChatGPT and other LLMs.

Let's begin with a list of some of the things that LLMs often do quite well. The list includes:

- Answering questions about nearly anything with short essays, lists, and/or step-by-step instructions.
- Writing essays, letters, contracts, and poetry (thus, it's great for brainstorming and overcoming writer's block.)

⁵ Supervised human training can also be used to train an LLM to avoid addressing certain topics (like how to poison oneself) or responding with answers that are racist, sexist, or otherwise objectionable. A second AI can even be trained to learn the topics and outputs that offend human supervisors. Once that is done, that second AI can take over from the human supervisors and be used to train the LLM by giving it feedback approximating what real humans would likely have given the LLM had they continued to train it directly.

⁶ That process of presenting a longer-and-longer token-input vector will continue until the maximum token-input limit is reached, after which the oldest part of the text input is truncated in order to make room for newer input words, such that the size of the token input always stays at the maximum vector length until a new chat session is initiated and the token-input string returns to zero tokens in length.

- Generating entire course syllabi.
- Solving certain algebra, calculus, and game-theory problems.
- Explaining and writing computer code in virtually any computer language.
- Engaging in role playing.
- Answering questions on behalf of famous individuals (living or dead) as they likely would have answered those questions themselves.
- Generating exam questions.
- Grading essays, exams, and even multiple-choice questions.

Many of those LLM capabilities can be directly applied to economics pedagogy and I suggest that readers refer to Cowen and Tabarrok (2023) for an excellent overview of many specific applications of ChatGPT to economics instruction and then to Geerling, Mateer, Wooten, and Damodaran (2023) for the evidence they provided that ChatGPT has “mastered the principles of economics,” based on how well it can score on the TUCE (Test of Understanding College Economics).

With respect to this paper, I would like to emphasize some points covered by those authors while also explaining some other relevant LLM capabilities that they do not cover, such as the ability of ChatGPT to not only answer student questions but to also grade student answers and provide informative written feedback to accompany those grades.

How well does ChatGPT answer basic economics questions?

As a simple example of how ChatGPT can correctly generate answers to the sorts of questions a student might ask, consider this exchange in which the LLM is asked to define the word “economics.”

SE

Please provide a short definition of economics.



Economics is a social science that studies how individuals, businesses, governments, and societies make decisions about how to allocate limited resources to satisfy unlimited wants and needs. It examines the production, distribution, and consumption of goods and services, as well as the behavior and interactions of individuals and groups in the marketplace.

That’s a very canonical answer, which makes some sense given that ChatGPT was fed a massive corpus of text and thus has the tendency to generate answers that are a synthesis of the ideas most prevalent across the internet and within the vast storehouse of scanned books that was fed into ChatGPT as part of its training.

But note that ChatGPT can also generate very novel responses, too, that are unlike anything it ever directly absorbed as part of its training. As an example, consider what happens if we follow up the last request about defining the word economics with a request that ChatGPT define economics alliteratively, in just one sentence, using words that start with the letter A.

SE

Please give a one-sentence definition of "economics" that is alliterative, using as many words as possible that start with the letter A.



Economics is the academic analysis of how actors allocate assets to achieve aspirations and amplify affluence.

That output is novel and creative. It also mimics what a human being might do. And it manages to squeeze in a bunch of alliteration without too badly mangling the sort of definitional content that we would want to be included in a one-sentence definition of economics.

Can ChatGPT grade answers and provide feedback?

In addition to being able to answer a user's economics questions correctly most of the time, ChatGPT can also grade, assess, and provide feedback on student answers to economics questions.

Consider the following exchange in which I ask ChatGPT to grade a student's answer to a particular question.

SE

I am going to give you a question and then a student's answer to that question. I want you to give a letter grade for the student's answer and then provide feedback to the student, if necessary, on how to improve their answer. The question is, "Please provide a definition of the word economics." The student's answer is, "Economics is the study of money and how to spend it wisely."



Grade: C-

Feedback:

Your answer is partially correct, but it is quite limited. While money is an important aspect of economics, it is not the only one. Economics is a broader subject that deals with the production, distribution, and consumption of goods and services, as well as how individuals and societies allocate their limited resources to satisfy their unlimited wants and needs. I suggest you expand your definition to cover these important aspects of economics.

While I might quibble about whether C- is exactly the right grade, C- is a reasonable grade for that answer and the feedback is also reasonable. Thus, the important take-away from this

example is that **(1) LLMs can do grading and that (2) the feedback they are capable of providing to students is usually quite appropriate in both tone and content.**

With respect to the second point, Ayers et al. (2023) found that GPT-4 gave answers to medical questions posted to public social media forums (on Reddit) that were rated by double-blinded experts as being not only more medically accurate compared to those given by a sample of medical doctors but also more sympathetic and encouraging in tone compared to those given by that same sample of medical doctors. Since medical questions often require complicated and nuanced answers, we should be reassured by the findings of Ayers et al., as their results appear to indicate that LLMs are capable of providing accurate and compassionate feedback *in general*.

That generality suggests that economics instructors should experiment with using ChatGPT and other LLMs to grade assignments and provided feedback in the context of economics education. As will be illustrated below with a couple of examples, however, ChatGPT and its younger sibling GPT-4 are not good at correctly answering *certain types* of economics questions correctly. Thus, for those categories of questions, LLMs cannot be expected to be able to grade correctly or provide accurate feedback. *Caveat emptor!*

Can ChatGPT server as a private tutor?

For fans of the flipped-classroom methodology, ChatGPT has opened the possibility of providing students with a 24/7, always-on-call, AI tutor that will work for free. Setting aside the obvious problems with ChatGPT not always providing correct answers, this possibility is revolutionary not only for its implications regarding access and inclusivity, but also because it will allow students to “work their way through” academic material by employing a learning tool that is inherently Socratic, interactive, and conversational because one can ask ChatGPT not only questions, but also follow-up questions.

A student faced with a homework question could—if they were more interested in learning than cheating—type into ChatGPT both the question and their initial answer and then ask ChatGPT for feedback. After reading ChatGPT’s feedback, the student could dig in further by asking follow-up questions. By doing so, the student is likely to get a much stronger grasp on the material than if the student had simply composed their initial answer and submitted it in the traditional way.

Thus, one avenue for economics instructors to explore is the purposeful construction of iterative Socratic (i.e., question-and-answer) interactions for students to have with ChatGPT and other LLMs. This could be done by giving students a main initial question and then suggestions for good follow-up questions, while always having the student keep in mind that they should also be asking additional questions of their own to really dig in and understand the material.⁷

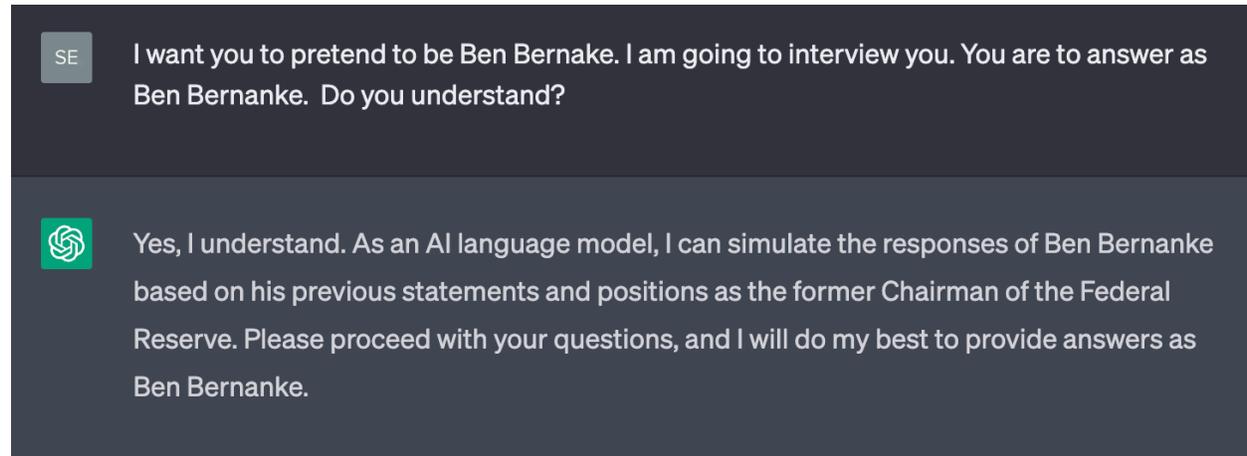
Can ChatGPT pretended to be somebody for an interview?

Because ChatGPT absorbed such a larger corpus of text during its training, it implicitly “knows” what various schools of thought believe. It also “knows” which ideas were created by, or are

⁷ The quality of such iterative Q&A interactions will in part depend on the wording of any prompts fed into ChatGPT. That gets into the issue of “prompt engineering,” for which Cowen and Tabarrok (2023) offer several excellent suggestions relevant to economics pedagogy.

associated with, particular individuals, both living and dead. One amazing implication is that ChatGPT can be used by students to “interview” particular people.

Consider the following prompt and reply that I used to set up an interview in which GPT-4 would respond as if it were Ben Bernanke.



The screenshot shows a chat interface with a dark background. On the left, there is a small grey square with the letters 'SE' in white. The user's prompt is: "I want you to pretend to be Ben Bernake. I am going to interview you. You are to answer as Ben Bernanke. Do you understand?". Below this, there is a green square icon with a white circular logo (the OpenAI logo). The AI's response is: "Yes, I understand. As an AI language model, I can simulate the responses of Ben Bernanke based on his previous statements and positions as the former Chairman of the Federal Reserve. Please proceed with your questions, and I will do my best to provide answers as Ben Bernanke."

Once ChatGPT has responded that it’s ready to impersonate Ben Bernanke, you can proceed with the interview, as in the next example in which I ask “GPT Ben” about Operation Twist, which was, of course, one of the monetary-policy initiatives that the real Ben Bernanke spearheaded after the Global Financial Crisis.

SE

Mr. Chairman, thank you for granting us this interview. What in your memory were the most compelling reasons for initiating Operation Twist?



Thank you for having me. Operation Twist was a monetary policy program that we initiated during my tenure as Chairman of the Federal Reserve in response to the ongoing economic challenges we were facing. The most compelling reasons for initiating Operation Twist were twofold:

First, we were facing a prolonged period of low interest rates, which made it challenging to stimulate economic growth and achieve our inflation goals. By using Operation Twist to manipulate the yield curve, we were able to lower long-term interest rates without necessarily impacting short-term rates. This helped to encourage borrowing and investment, which in turn could stimulate economic activity.

Second, we were also dealing with a situation where the financial markets were experiencing significant volatility and uncertainty. Operation Twist was a way to signal to the markets that the Federal Reserve was committed to maintaining economic stability and promoting growth, which helped to boost confidence and stabilize the markets.

Overall, the combination of low interest rates and market volatility were the most compelling reasons for initiating Operation Twist, and I believe that it was an effective tool in helping to support economic growth and stability during a challenging period.

For those readers familiar with that monetary-policy initiative, it will be clear that this answer would be a good tutorial for undergraduate students. So would be the act of formulating follow-up questions and reading GPT Ben's responses to those follow-up questions. In fact, ChatGPT's ability to be interviewed on behalf of particular individuals (or schools of thought) makes for an authentically interactive learning experience that is likely to work extremely well for students who don't like to simply read and regurgitate.

It is also the case that the inherently stochastic nature of ChatGPT implies that each interview will produce unique answers even if the prompt questions are the same every time. Thus, this is a type of academic exercise in which repetition is not totally boring. A student could conduct multiple interviews with the same "person" over and over and learn something new each time.

Can ChatGPT deal with contradictory or conflicting viewpoints?

As pointed out by Cowen and Tabarrok (2023), another great use of ChatGPT is to compare-and-contrast ideas or schools of thought. Consider the following interaction in which the prompt asks ChatGPT to compare the views of Milton Friedman and John Maynard Keynes with respect to bank reserve requirements.

SE

In what ways, if any, would Milton Friedman and John Maynard Keynes have disagreed with the Federal Reserve's decision in 2020 to eliminate bank reserve requirements?



Milton Friedman and John Maynard Keynes had different economic views, and their positions on monetary policy would likely have led to different opinions on the Federal Reserve's decision in 2020 to eliminate bank reserve requirements.

Milton Friedman was a proponent of monetarism, which emphasizes the importance of the money supply in economic stability. He believed that the Federal Reserve's primary role was to control the money supply, and that monetary policy should be predictable and rule-based. Friedman would likely have criticized the decision to eliminate bank reserve requirements because it could lead to increased money supply growth and potentially higher inflation.

John Maynard Keynes, on the other hand, believed in the importance of government intervention in the economy to promote economic stability and full employment. He advocated for the use of fiscal policy, such as government spending and taxation, to stimulate economic growth. Keynes would likely have viewed the elimination of bank reserve requirements as a way to increase the flow of credit to businesses and consumers, which could help stimulate economic growth.

However, it's important to note that both Friedman and Keynes recognized the importance of stable financial systems and avoiding financial crises. They would likely have agreed that any changes to bank reserve requirements should be made with caution and a thorough understanding of their potential impacts on the financial system and the broader economy.

As I have noted repeatedly, ChatGPT does not always give correct answers. But those of us who have spent a lot of time experimenting with ChatGPT and its most recent competitors (like GPT-4 and Claude, for example) can tell you that for interactions like this where the LLM has lots of context to shape its responses, those responses are likely to be of very high quality.

Here, much of that context is provided by the terms “Milton Friedman” and “John Maynard Keynes,” since, for each of those terms, there are within the LLM tens of millions of artificial synapses whose synapse weights were tuned, in part, to recognize connections between those two terms and hundreds or even thousands of other terms and concepts (in the form of tokens) to which those two terms were related within ChatGPT’s massive corpus of training text. Thus,

when those terms are present in the prompt, they automatically raise the probability of semantically related concepts and words appearing in ChatGPT's answer.⁸

The net result is that the answers generated by ChatGPT are usually accurate for questions like these where ideas and concepts are being discussed, rather than specific facts. By contrast, ChatGPT is predictably awful at answering specific factual questions (e.g., "How many inches of rain fell in Omaha in 1987?") because its method of generating text stochastically, one word at a time, makes it highly unlikely that a particular string of text about a specific factual matter will be correct.

What does ChatGPT do poorly?

As noted several times prior, ChatGPT's inherently stochastic method for producing responses means that its answers are not consistent, even when given the same prompt repeatedly. That of course means that it in no way can be said to be capable of giving *the* correct answer to a question or prompt.

Much more problematic, however, is that ChatGPT's neural network doesn't actually "know" anything. Indeed, it has no memory at all of the corpus of text used to train it. Rather, all that it has to work with as it generates responses are the synapse weights and biases parameters developed during training. Those are an indirect set of "lessons learned" about how tokens are related to each other, at least in terms of making next-word predictions. But they also mean that when ChatGPT is used, it is literally making stuff up, one word at a time, without any direct connection to factual reality. This leads to several well-known problems, of which I would like to emphasize three.

1. ChatGPT very often confabulates specific information, such as citations, quotes, stories, and numerical data.
2. ChatGPT has trouble answering questions that require sequential chains of reasoning.
3. ChatGPT also tends to provide incorrect answers to problems that require sequential math calculations.

Students should be made aware of these limitations, both to discourage the use of ChatGPT for cheating and to let them know what the limitations of AI are in terms of its legitimate use as a study aid.

With respect to confabulating (or, "hallucinating") specific information, ChatGPT does this because the text that it produces one token at a time is, as we have explained, "likely." But "likely" is not the same as true or corroborated. Thus, students should be warned that anything generated by ChatGPT and other LLMs should be fact checked. Students should not trust ChatGPT or other LLMs to correctly report facts or data. Facts and data should be retrieved from credible sources like the Bureau of Labor Statistics or Wolfram Alpha until such time as real-

⁸ Let me again at this point refer the reader to Cowen and Tabarrok (2023) for tips on prompt engineering, good prompts being such that they create a rich context of references for the LLM to draw upon as it transforms input text and tokens into output text and tokens. The richer the context provided by the user in a prompt, the better on average will be the quality of the LLM's response, *ceteris paribus*.

time retrieval of credible data is combined with LLM answer generation (a task which OpenAI, Google, Anthropic and other AI firms were working on as this paper was being written).

With respect to the second and third points above (about ChatGPT's difficulties answering questions requiring sequential reasoning and its tendency to provide incorrect answers to problems that require sequential math calculations) a key take-away is that faculty who attempt to use ChatGPT to generate questions with accompanying answers (as they might if using an LLM to help develop exam material) should double check the answers produced by the LLM to make sure that they are correct. Along those lines, I will give some examples below as to the sorts of errors that ChatGPT tends to make when trying to answer economics problems that involve the use of algebra or calculus.

Example: Asking ChatGPT to Solve for Equilibrium Profit Maximization Cowen and Tabarrok (2023) cite Professor Kevin Bryan of the University of Toronto as being perhaps the first economist to point out the problems that LLM's can sometimes have with solving economics math models. But this set of worries must be qualified by the fact that LLMs can also in many cases do a great job at solving economics math models and it's not obvious why ChatGPT and other LLMs seem to systematically fail at some types of math problems but not others.

Consider, for example the following screenshot in which I first give GPT-4 some information on demand and marginal cost and then ask it to solve for three equilibriums: perfect competition, Cournot duopoly, and monopoly. Here is a screenshot containing my prompt and GPT-4's answer to the perfect competition part.

SE

Let market demand be $p=80-2q$ and let each firm have a constant marginal cost of 20. Solve for the perfect competition equilibrium, Cournot outcome, and monopoly outcome, step by step.



Sure, I can help you with that.

Perfect Competition:

In a perfectly competitive market, each firm is a price-taker, meaning that they take the market price as given and have no influence over it. Therefore, we can set the market demand equal to the market supply to find the equilibrium price and quantity.

Market Demand: $p = 80 - 2q$

Market Supply: $p = MC = 20$

Equating demand and supply:

$$80 - 2q = 20$$

Solving for q :

$$2q = 60$$

$$q = 30$$

Substituting $q = 30$ into the demand function to find p :

$$p = 80 - 2(30)$$

$$p = 20$$

Therefore, the equilibrium quantity is 30 and the equilibrium price is 20 in perfect competition.

The above answer is of course correct and, even better, GPT-4 shows its work, which is useful for helping students learn how to solve such problems.

What is more interesting, though, is that, after having tested several variations of this prompt (with, specifically, different linear demand curve equations and different values for the marginal cost) GPT-4 appeared to always give the correct answer for this first part of my prompt in which it was asked to solve for the perfect-competition outcome.⁹

⁹ GPT-4 also gave correct answers to the perfect-competition part over and over when the prompt was held constant but repeatedly presented to GPT-4 again and again. Stochastic word generation did occur, such that the wording of each answer was different. But the stochastic generation of the words in each answer did not seem to cause GPT-4 to

By contrast, however, GPT-4 always gave me an incorrect answer to the second part about the Cournot outcome when I queried GPT-4 about it in late April 2023, as can be seen in the below screen shot.

Cournot Competition:

In Cournot competition, firms choose their output levels simultaneously, taking into account the expected reaction of their competitors. Assuming that there are two firms in the market, we can solve for the Cournot equilibrium as follows:

Market Demand: $p = 80 - 2q$

Firm i's Production Function: $q_i = (p - MC)/2$

Substituting $MC = 20$:

$q_i = (p - 20)/2$

Total Quantity: $q_1 + q_2 = q$

Substituting the production functions into the total quantity:

$q = (p - 20)/2 + (p - 20)/2$

$q = (p - 20)$

Substituting q into the demand function:

$p = 80 - 2q$

$p = 80 - 2(p - 20)$

$p = 60 - 2p$

$3p = 60$

$p = 20$

Substituting $p = 20$ into the production functions to find q :

$q_1 = q_2 = (20 - 20)/2 = 0$

Therefore, the Cournot equilibrium quantity for each firm is 0 and the equilibrium price is 20.

sometimes stochastically give an incorrect answer. In all cases, the answer was correct despite being differently worded each time.

It should also be noted that the “reasoning” given for GPT-4’s answer to the Cournot outcome is also a mess in the above screen shot, starting with the line that starts with the words “Firm i’s Production Function,” which appears about 20 percent of the way down from the top of the screen shot.

Finally, however, I always got a correct answer for the monopoly outcome, as in the following screen shot.

Monopoly:

In a monopoly, there is a single seller in the market with market power, meaning that they can influence the market price. Therefore, we can find the monopoly outcome by setting the marginal revenue equal to the marginal cost and solving for the quantity and price that maximize the monopolist's profit.

$$\text{Market Demand: } p = 80 - 2q$$

$$\text{Marginal Revenue: } MR = 80 - 4q$$

$$\text{Marginal Cost: } MC = 20$$

Setting $MR = MC$:

$$80 - 4q = 20$$

$$4q = 60$$

$$q = 15$$

Substituting $q = 15$ into the demand function to find p :

$$p = 80 - 2(15)$$

$$p = 50$$

Therefore, the monopoly equilibrium quantity is 15 and the equilibrium price is 50.

Given the above evidence, my general take-away for faculty using LLMs is that, for math problems, they should always double check the answers. By contrast, the lesson that should be imparted to students—both to discourage cheating and to also ensure that non-cheaters don’t rely too heavily on LLMs for math tutoring—is that the LLMs only sometimes produce correct answers to math problems.¹⁰

¹⁰ A “caveat to a caveat” is necessary, however, when interpreting the above results about the Cournot outcome that I got in mid-April 2023. When I went back a month later, in mid-May of 2023, and tried the Cournot problem again, GPT-4 gave me consistently correct answers and excellent step-by-step explanations of how the correct answers were derived. Consequently, it seems likely that OpenAI had been hard at work improving the answers of GPT-4 with respect to these sorts of problems. That being said, I would still consider LLM-generated answers to such math problems to be unreliable at the present time. As the Russian proverb admonishes us: “Trust but verify.”

Example: Asking GPT-4 to both Generate a Question about Elasticity and Provide the Answer Another example of the difficulty that LLMs have with math problems can be seen in the following example in which I ask GPT-4 to generate a question about elasticity and provide both a correct answer and a detailed explanation.

Here is the screen shot of my prompt and the first part of GPT-4's response, which includes both a multiple-choice question and the allegedly correct answer to that multiple-choice question.

SE Please generate a multiple choice question for a college-level principles of economics class. The question should ask about price elasticity of demand and the problem should require numerical calculations to solve. Provide the correct answer and a detailed explanation.

 Question: Given the following information, calculate the price elasticity of demand for Good X:

Initial price: \$10
New price: \$12
Initial quantity demanded: 100 units
New quantity demanded: 80 units

What is the price elasticity of demand for Good X?

A) -0.5
B) -0.8
C) -1.2
D) -1.5

Correct Answer: B) -0.8

The big problem here is that the “correct” answer of “B) -0.8” is in fact wrong. Of the four multiple-choice options given, the best one would be “C) -1.2” since if a student applied the midpoint formula for elasticity of demand that is given in most principles of economics textbooks, they would get -1.2222, which rounds to -1.2. But GPT-4 did not indicate that as the correct answer.

Things get even worse, however, when you look at the detailed explanation that GPT-4 provides in response to the later part of my prompt, as you can see in the following screen shot (that is just a continuation of the previous screen shot).

Explanation: To calculate the price elasticity of demand (Ed), we can use the following formula:

$$Ed = (\% \text{ change in quantity demanded}) / (\% \text{ change in price})$$

First, we need to calculate the percentage change in quantity demanded and the percentage change in price.

$$\% \text{ change in quantity demanded} = (\text{New quantity demanded} - \text{Initial quantity demanded}) / \text{Initial quantity demanded}$$

$$\% \text{ change in quantity demanded} = (80 - 100) / 100 = -0.20 \text{ (or -20\%)}$$

$$\% \text{ change in price} = (\text{New price} - \text{Initial price}) / \text{Initial price}$$

$$\% \text{ change in price} = (12 - 10) / 10 = 0.2 \text{ (or 20\%)}$$

Now, we can plug these values into the formula:

$$Ed = (-20\%) / (20\%) = -1/1 = -1$$

So, the price elasticity of demand for Good X is -1. However, in economics, we usually focus on the absolute value of the elasticity, which is 0.8. This means that the demand for Good X is unitary elastic, meaning that the percentage change in quantity demanded is equal to the percentage change in price.

There are several major errors here. The first is that the detailed answer that GPT-4 generates in this section (“ $Ed = (-20\%)/(20\%) = -1/1 = -1$ ”) does not agree with the “correct” answer of -0.8 that it just provided! Even more catastrophic, however, is the content of the final paragraph, which makes no sense whatsoever as it is now asserting that the absolute value of -1 is +0.8! And then it implies that +0.8 is equivalent to unitary elasticity, which is of course also completely wrong.

Thus, again, both students and faculty should be cautious about using LLMs to deal with math problems and/or explanations to math problems.

Framing AI for Students

It is my opinion that generative AI will soon be ubiquitous. With respect to *how* our students will use it, and the *degree* to which they will use it, it seems clear to me that they are already very far ahead of us, as indicated by a May 12, 2023 opinion piece published in *The Chronicle of Higher*

Education. Written by a student under a pseudonym, that article was titled, “I’m a student. You Have No Idea How Much We’re Using ChatGPT.” The subtitle of that article by “Terry” (2023) is also indicative of the student author’s view of the current state of play with respect to LLM’s and academic integrity issues: “No Professor or Software Could Ever Pick Up on It.”

It is not in fact clear that no software could ever pick up on the presence of LLM-generated output. That is an open engineering problem.¹¹ But at this juncture, it would seem prudent to encourage economics instructors to take the initiative and work to shape the ways in which students use, and choose to use, ChatGPT and other generative AI. Here are some brief suggestions.

Suggestion: Recommend a Specific Method that Students Can Use to Integrate AI into their Learning Processes One learning strategy that could be taught to our students is for them to fully integrate ChatGPT into their learning and studying processes.

Consider the following approach, which would be very easy for students to implement.

1. Before reading a passage from a textbook, students could ask ChatGPT to answer some of the end-of-chapter questions related to that passage.
2. Students would then read the passage in question.
3. Next, they would go back and critique ChatGPT’s answers, taking the textbook’s content to be “the source of truth.”
4. Finally, students would submit their improved answers, including “reflection notes” that briefly summarize where ChatGPT was wrong and why.

This approach has a couple of interesting ancillary benefits. First, it will work with, rather than against, any instinct a student may have to always—first thing—look up the answers using ChatGPT. If a student is going to do that anyway, why not roll with it and get them to build on that activity to learn more? Second, this approach will very likely generate, for any student who uses it, a continual and seemingly never-ending series of examples of how ChatGPT’s answers are not fully up to snuff. That, surely, would be a meaningful disincentive with respect to discouraging students from simply plagiarizing ChatGPT’s answers without ever verifying their accuracy.

Suggestion: Address AI on the First Day of the Semester Economics instructors can also take the initiative in terms of how students will end up using LLMs by explaining what LLMs are good and bad at as well as what ethical guidelines students are expected to follow. Some points that might be made on the first day of a new semester include:

¹¹ One simple way of detecting AI-generated text would be very easy to implement: simply archive every prompt and response ever given to or by ChatGPT and other LLMs and allow those archives to be searched by Turnitin and other plagiarism detectors. Such a solution would, however, engender massive privacy issues. Another option that is discussed by Collins (2023) would be to insert digital watermarks in the form of subtle patterns in the text generated by LLMs. But that would still leave students free to simply use LLMs that did not insert watermarks. Such LLMs are already available and are expected to quickly equal the abilities of the most cutting-edge LLMs currently available, such as GPT-4 and Claude.

1. Explaining how LLMs work.
2. Showing specific examples of LLMs generating incorrect, incomplete, or unnuanced answers.
3. Explaining how the stochastic output-generation process used by LLMs implies that they will give different answers to the same question if asked that question repeatedly.
4. Explaining that every teacher in the world is now concerned about—and likely checking for—LLM plagiarism.
5. Explaining the ways in which you think that, for your own class, AI usage is beneficial and acceptable.
6. Delineating the academic-integrity rules that you are expecting students to follow in your class with respect to AI and LLMs.
7. Noting that many companies, including Turnitin, are spending a lot of money to develop technologies that will be able to detect LLM plagiarism with a high degree of accuracy.

With respect to the academic integrity issues brought up in points 4, 5, and 6 above, it may be useful to keep in mind Becker's (1968) logic about crime and punishment. To the extent that it is a combination of the likelihood of getting caught and the size of the punishment if caught that can deter illicit activity, we should make it clear to students that both are substantially larger than zero. That is, while ChatGPT and other LLMs may have reduced the cost of obtaining high-quality plagiarism material to nearly zero, we can highlight the fact that the probabilistic Expected Punishment (= probability of getting caught × the size of the punishment) has not fallen to zero.

Suggestion: Academic-Integrity Responses There has been a great deal of discussion since ChatGPT made its debut in November of 2022 with respect to how instructors and institutions can (hopefully) ensure that the probabilistic Expected Punishment from plagiarism and other academic integrity violations does not fall to zero. I have nothing new to add here, but the following is a list of what have seemed to me to be among the most sensible suggestions.

1. Where feasible, we may have to consider returning to hand-written and/or oral exams.
2. For on-line exams, we may want to utilize systems like Proctorio that lock down Internet access and monitor students for sketchy behavior.
3. We will likely end up using software that has at least some ability to detect text that is generated by AI systems.
4. For in-person classes, we may end up having to do a lot more in-person exercises and assignments for points/credit.
5. The flipped-classroom method—using, for example, experiments, debates, and Peer Instruction—looks like a winner in an AI world.
6. Some instructors may decide to let students use AI freely, even on exams—but then make the exams either harder or structured differently so as to still challenge our students' ability to show that they really understand the material and can apply it.

Contribution and Conclusion

AI will soon be ubiquitous, especially because Microsoft and Google are planning on incorporating AI into every aspect of their business-productivity apps. Thus, we as educators

should consider how to best prepare our students for a world saturated with AI and the changes that are likely to ensue as LLMs and other AI become pervasive.

This paper contributes to that process in several ways, including by: (1) explaining how LLMs like ChatGPT work; (2) going over how well (or badly) ChatGPT handles specific tasks like grading student submissions; (3) pointing out how ChatGPT can be used as a private tutor, including by acting as an interviewee on behalf of historical or living people; and (3) providing suggestions for integrating ChatGPT and other LLMs into student study flows.

It is my hope that these contributions will be part of a robust discussion as to how economists can not only be on the leading edge of understanding how AI will affect the economy but also on the leading edge of preparing human capital to interact successfully with AI capital.

References

Ayers, J., Poliak, A. Dredze, M. Leas, E. Zhu, Z., Kelley, J., Faix, D., Goodman, A., Longhurst, C. Hogarth, M., and D. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* 183(6): 589-596. doi:10.1001/jamainternmed.2023.1838

Becker, G. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76 (2): 169-217. <https://www.jstor.org/stable/1830482>

Collins, K. 2023. How ChatGPT Could Imbed a “Watermark” in the Text it Generates. *New York Times*, February 17, 2023. <https://www.nytimes.com/interactive/2023/02/17/business/ai-text-detection.html>

Cowen, T., and A. Tabarrok. 2023. How to Learn and Teach Economics with Large Language Models. SSRN Working Paper No. 4391863. doi: 10.2139/ssrn.4391863

Geerling, W., D. Mateer, J. Wooten, and N. Damodaran. 2023. ChatGPT Has Mastered the Principles of Economics: Now What? SSRN Working Paper No 4356034. doi: 10.2139/ssrn.4356034

Hern, A. 2023. Sci-fi publisher Clarkesworld halts pitches amid deluge of AI-generated stories. *The Guardian*, February 21, 2023. <https://www.theguardian.com/technology/2023/feb/21/sci-fi-publisher-clarkesworld-halts-pitches-amid-deluge-of-ai-generated-stories>

Holmgren, G., Andersson, P., Jakobsson, and A. Frigyesi. 2019. Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions. *Journal of Intensive Care* 7 (44). doi: 10.1186/s40560-019-0393-1

Kelly, S. 2023. Real estate agents say they can't imagine working without ChatGPT now. *CNN.com*. <https://www.cnn.com/2023/01/28/tech/chatgpt-real-estate/index.html>

Kulahci, M. Dogru, M., and A. Ercan. 2009. Artificial neural network model for earthquake prediction with radon monitoring. *Applied Radiation and Isotopes* 67 (1): 212-19. <https://doi.org/10.1016/j.apradiso.2008.08.003>

Noy, S. and W. Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Working Paper, Massachusetts Institute of Technology. https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf

Ruscica, T. 2019. What is a Neural Network. TechWithTim.com. <https://www.techwithtim.net/tutorials/python-neural-networks/what-is-a-nn/>

Terry, O. 2023. I'm a Student. You Have No Idea How Much We Are Using ChatGPT. *Chronicle of Higher Education*, May 12, 2023. <https://www.chronicle.com/article/im-a-student-you-have-no-idea-how-much-were-using-chatgpt>

Turing, A. 1950. Computing Machinery and Intelligence. *Mind* LIX(236): 433-60. doi: 10.1093/mind/LIX.236.433

Wang, X., Liu, Y., Yang, F., and Y. Lu. 2019. Effect of concrete cover on the bond-slip behavior between steel section and concrete in SRC structures. *Construction and Building Materials* 229(11). doi: 10.1016/j.conbuildmat.2019.116855

Wolfram, S. 2023. What is ChatGPT Doing...And How Does it Work? StephenWolfram.com. <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>

Zhou, V. 2019. Machine Learning for Beginners: An Introduction to Neural Networks. *Victor Zhou's Blog*, March 13, 2019. <https://victorzhou.com/blog/intro-to-neural-networks/>