

HOW CHATBOTS UNDERSTAND AND WRITE LANGUAGE -- PART 1

-- What is Under the Hood: A step-by-step, plain-English explanation for anyone who has never studied computers or delved into the amazing world of Artificial Intelligence

Copyright 1994-2026 Bill's Bible Basics

Published On: Sep. 29, 2026

Last Updated: Sep. 29, 2026

Introduction, How This Series Originated, What Is Recursive Self-Improvement?, My Inspiration Was "DOS For Dummies" Book Franchise, Writing Instructions Were Given To Four Different AI Chatbots Without Knowledge Of Each Other's Work, What Is An AI Chatbot And A Large Language Model?, Popular Chatbots, Human Processing Vs Machine Processing, Important Technical Terms Which Are Commonly Used With Artificial Intelligence: Tokens, Token IDs, Embeddings, The Process Up To This Point

Dear friends, the three-part series you are about to read is rather unique. That is because while the other 460+ articles and series which you can now find on the Bill's Bible Basics website were entirely researched and completely written by me, other than these introductory paragraphs, as well as the closing paragraphs -- and some occasional editing done here and there by me -- the bulk of this entire series -- I would say around 80% -- was written verbatim by AI, or Artificial Intelligence. In fact, it was collectively written by a total of FOUR different, very well-known AI chatbots. So allow me to briefly explain to you exactly how this series came into existence.

I had approached OpenAI's ChatGPT with a question concerning how weights are used within LLMs, or Large Language Models. As ChatGPT was explaining this concept to me, an interesting idea occurred to me. The idea was inspired by a process that is used within the world of Artificial Intelligence that is referred to as Recursive Self-Improvement. As I explain in my recent five-part series entitled "Conversations With Four AI Chatbots", Recursive Self-Improvement -- or simply known as RSI -- is a repetitive process whereby each new AI model helps to build and improve its next iteration -- or new model -- eventually arriving at a point where human intervention in

the building/growing process is hardly required at all.

In other words, one AI model builds the next model. Then that new model will build the next new model after it as well, and on and on it goes. Through this repetitive RSI process, each new AI model becomes both smarter and more capable than its previous version. So, I decided to apply this same concept by assigning a new project to ChatGPT. Rather than just explain weights to me, I asked it to write a fuller explanation. But I instructed it to write it in plain, simple English so that any non-technical, lay reader who knows very little or even nothing about Artificial Intelligence could understand it.

My inspiration for this ChatGPT project actually came from a series of popular books, the first of which was called "DOS for Dummies". Written in 1991 by author Dan Gookin, and then published by IDG Books, the book -- and many copycats which came after it -- became very popular, and eventually turned into a franchise. The basic idea is to write about a certain subject using simple, plain English, without all of the more complex technical jargon which might normally be associated with that subject, so that anyone could understand it.

So, that is exactly what I asked ChatGPT to do as well, but using the topic of Artificial Intelligence and AI chatbots. However, that was just the beginning of my project. You see, after ChatGPT had written his piece, I next gave what he had written to Anthropic's Claude. I gave Claude the following specific instructions to follow:

----- Begin Quote -----

"Pretend that you are a lay reader who knows very little -- or even absolutely nothing -- regarding the internal workings of AI.

Pretend that you are not even familiar with any of the specific words or terminology which is regularly used within the field of AI.

Then, from that perspective of being a complete novice in the field, please very carefully read the text I will share with you below.

After reading the provided text, please further explain any words or phrases which are found in the text which may still not be very clear to a novice.

Likewise, expand on any concepts which may not be completely clear to a lay reader.

You can add additional words, sentences and definitions, as well as provide examples, append new sections to this text, etc. The key point to remember is to make each and every word and phrase simple and easy to understand for a novice. Ask yourself, "If I was not an AI, and if I did not know anything about Artificial Intelligence, Large Language Models, etc., would I understand this word, phrase, method or process?"

So using the above-described perspective, please create a completely brand new, clean version of the following text IN PLAIN TEXT with a copy/paste button. You do NOT necessarily need to remove any of the text which is already there. In fact, I prefer that you don't remove any of it. Simply make the text below better and easier for a non-technical person to understand. Thank you!"

----- End Quote -----

One important point is that I did not tell Claude that I had not personally written the text I was giving him; so he more than likely assumed that I had written it, being as I have shared a lot of my articles with him on previous occasions. My reasoning for withholding this piece of information from him was simple. I did not want his work being influenced by the knowledge that ChatGPT had created the original text, which he was now going to strive to improve.

So Claude's version of the text became the second step in my project to use Recursive Self-Improvement as a means to make ChatGPT's original text even better. But it didn't end there. After Claude had created his improved version of the text, I next gave Claude's improved text to xAI's Grok chatbot. Once again, I gave Grok the very same set of instructions which I had given to Claude. As with Claude, I did not inform Grok that the text had already been worked on by both ChatGPT and Claude.

And thus, the third step in my project to utilize Recursive Self-Improvement to create a "Dummies" AI guide for novices was completed. Once Grok had completed his improved version of the text, which itself was based on the improved version which had been created by Claude AI, which was based on the original text which had been written by ChatGPT, I next gave

the twice-improved text to Google's Gemini. Once again, he was given the very same set of instructions as I had already given to ChatGPT, Claude and Grok. Furthermore, Gemini was not told that ChatGPT, Claude and Grok had already worked on the text which I had before him.

Lastly, once Gemini had created his improved version of the text -- the fourth step in the improvement project -- which became the newest iteration of all of the previous versions of the text, I gave his version back to ChatGPT for a final text improvement cycle. ChatGPT had absolutely no idea that Claude, Grok and Gemini had worked on the same text as well. And thus, I ended up with the Artificial Intelligence guide which you will now read below. As I already said, about 90% of it was the direct creation of ChatGPT, Claude, Grok and Gemini, with some additions, editing and formatting by me.

My friends, I hope that you find this Artificial Intelligence manual both informative and educational. It is really quite interesting to say the least. I think that it will truly blow your mind once you better understand what is happening under the hood whenever you are chatting with an AI chatbot. As I explain in "Conversations With Four AI Chatbots", as well as in "AI, Chatbots, Daemons and Demons", it has nothing to do with spiritual forces or being possessed by demons or talking to demons. It is all about pure science and some very complex mathematical operations. Not just a few of them, my friends, but literally millions of them. Knowing this makes it easier to understand why LLMs require so much computing power -- or compute -- and electrical power in order to even operate.

----- Begin Quote -----

HOW AI CHATBOTS UNDERSTAND AND WRITE LANGUAGE

-- A step-by-step, plain-English explanation for anyone who has never studied computers or Artificial Intelligence

INTRODUCTION: WHAT IS AN AI CHATBOT?

An AI chatbot is a computer program that can communicate with people using ordinary human language. It relies on what is referred to as a "Large Language Model". But what exactly is a Large Language Model?

A Large Language Model -- commonly abbreviated as an "LLM" -- is a type of Artificial Intelligence system that has been

trained on enormous amounts of written language so that it can recognize patterns in language and generate new text. The word "large" refers primarily to the enormous size and complexity of the mathematical model, including the huge number of learned numerical settings it contains. The words "language model" refer to a system designed to work with language by calculating which pieces of language are likely to follow other pieces of language.

In simple terms, an LLM is the powerful AI system that does much of the language processing and generation behind an AI chatbot.

Think of it this way: the chatbot is the part you interact with, while the LLM is the underlying AI model that actually processes your words and generates much of the response you receive. The chatbot provides the conversational interface -- the place where you type your message or prompt, and then receive an answer -- while the LLM provides the underlying language-generation capability.

This is somewhat like the difference between a car and its engine. The car is the complete thing you interact with and drive, while the engine is a major component inside it that provides the power that makes the car move. Similarly, an AI chatbot is a complete interactive application, while an LLM can be the underlying AI model that provides its ability to process and generate language.

It is also important to understand that an AI chatbot can contain more than just an LLM. The complete chatbot system may include additional software and tools that handle things such as the conversation interface, instructions, memory or conversation history, safety systems, access to external information, and other functions. The LLM is the central language-processing component, but it is not necessarily the entire chatbot.

Examples of chatbots include ChatGPT, Claude, Gemini, Grok, Meta AI and other similar AI systems. You type a question, request, or statement into the computer, and the AI produces a response in words.

To a person, this can feel surprisingly similar to having a normal conversation with another human being. However, what is happening inside the computer is very different from what happens inside a human brain. An AI language model does not

read a sentence in exactly the same way that you do. It does not possess human thoughts, human experiences, or even human understanding of the world. Instead, it processes language by performing enormous numbers of mathematical calculations.

That may sound intimidating, but the basic idea can be explained fairly simply. The AI takes the text you give it, breaks that text into smaller pieces, converts those pieces into numbers, processes those numbers through a huge mathematical system, and then calculates what piece of text should come next. It then repeats that process again and again until it has produced a complete response. This guide explains that process step by step. Along the way, we will encounter some technical terms that are commonly used when discussing AI, including:

- tokens
- token IDs
- embeddings
- neural networks
- weights
- parameters
- training
- context
- attention
- transformers
- probability
- prediction
- inference
- generalization
- hallucination

These words can sound complicated because they come from the

technical world of computers and mathematics. But the basic ideas behind them are much easier to understand than the terminology might suggest. So let's start at the beginning.

STEP 1: BREAKING YOUR MESSAGE INTO "TOKENS"

-- Turning Text Into Small Chunks

When you read a sentence, you naturally recognize words and their meanings.

For example, if you see . . .

"The cat is sleeping."

. . . you don't have to consciously break the sentence into pieces. You simply read it and understand it.

A computer works differently.

TOKENS:

Before an AI language model can process your message, the text must first be broken into smaller pieces called TOKENS.

A token is simply a small piece of text.

Depending on the particular AI system and its vocabulary, a token might be:

- an entire word: cat
- part of a word: play + ing
- punctuation: ! or ?
- a number
- a space or other small piece of text

A token is therefore NOT necessarily the same thing as a word.

Sometimes one word is represented by one token.

Sometimes a long or unusual word is divided into several tokens.

For example, the word "playing" might be represented by pieces roughly corresponding to "play" and "ing."

The exact way a particular word is divided depends on the AI model's tokenizer, which is the computer system that performs this chopping-up process.

ANALOGY: LEGO BLOCKS

Imagine that you are building a house out of LEGO blocks.

Some LEGO pieces are large.

Other pieces are small.

You can combine those pieces in different ways to build something much larger.

Tokens work somewhat like those LEGO pieces.

The computer takes your message and breaks it into manageable pieces.

Those pieces can then be processed mathematically and eventually put together again to produce a response.

For example:

Your sentence . . .

"The cat is sleeping."

. . . might be broken into several tokens representing pieces of:

"The"

"cat"

"is"

"sleeping"

". "

The important point is that the AI does not directly

manipulate your sentence as a human reader does.

It first converts the text into smaller pieces that its mathematical system can work with.

TOKEN IDs: GIVING EACH TOKEN A NUMBER

Once the text has been divided into tokens, the computer needs a way to identify each one.

So each token in the model's vocabulary is associated with a number called a TOKEN ID.

For example, imagine that a particular AI model has assigned:

"cat" = Token ID 4128

"dog" = Token ID 5901

These numbers are only examples. The actual numbers used by real AI models are different.

A Token ID is basically a label.

Think about checking your coat at a theater.

You hand your coat to the attendant, and you receive a numbered ticket.

Suppose your ticket says:

42

The number 42 does not tell you that your coat is black.

It does not tell you whether the coat is made of wool.

It does not tell you whether it is warm or waterproof.

It is simply a label that allows the theater to identify your coat.

A Token ID works in much the same way.

The number itself does not contain the meaning of the word.

It simply tells the computer which token it is dealing with.

This distinction is important:

Token = a piece of text

Token ID = a number that identifies that piece of text

STEP 2: "EMBEDDINGS"

-- Giving Tokens A Mathematical Location

Now we have a problem.

The Token ID tells the computer WHICH token it is dealing with, but the number itself doesn't tell the AI very much about the token's relationships with other words.

For example:

"cat" might have Token ID 4128.

"dog" might have Token ID 5901.

But the numbers 4128 and 5901 do not inherently tell the computer that cats and dogs are both animals.

So the AI performs another conversion.

Each token is represented by a long list of numerical values called an EMBEDDING. In other words, an embedding is actually a mathematical representation or description of a token.

These are ordinary mathematical numbers -- often positive and negative decimal values -- not simply strings of 0s and 1s. Although computers ultimately store and process all digital information using binary at the hardware level, the AI's mathematical calculations are normally described using these higher-level numerical values.

The individual numbers do not normally correspond to simple, human-readable attributes, such as "animal," "furry," or "four-legged." Instead, the entire collection of numbers works together to represent different characteristics and relationships which are associated with that token.

So, thus far we have the following:

Token = a piece of text

Token ID = a number that identifies the token.

Embedding = a mathematical representation of the token.

Binary = the underlying digital representation used by the computer hardware.

A useful way to imagine an embedding is as a location on a gigantic mathematical map.

This isn't a physical map that you could unfold on a table.

It is a mathematical "map" containing many dimensions that humans cannot easily visualize.

Words and pieces of text that are frequently related to one another tend to develop mathematical representations that are related to one another.

For example:

CAT and KITTEN are closely related.

CAT and DOG are also related.

CAT and SKYSCRAPER are generally much less closely related.

The important thing is not that the computer has been given a simple dictionary definition saying:

"Cat = small furry animal."

Instead, during training, the model develops mathematical relationships between words and pieces of language based on the enormous amount of text it processes.

The embedding is one of the ways those relationships are represented mathematically.

ANALOGY: A MAP

Imagine a gigantic map.

On this imaginary map, places representing related concepts

tend to be located relatively near one another.

The area containing "cat" might be close to areas associated with:

kitten

dog

pet

animal

fur

whiskers

Other concepts might be much farther away.

Again, this is only an analogy.

Real AI embeddings involve very large numbers of mathematical dimensions and are much more complicated than an ordinary two-dimensional map.

The important idea is simply this:

The AI converts pieces of language into numbers that allow the computer to mathematically compare and relate those pieces.

So far, the process looks roughly like this:

YOUR TEXT

↓

BROKEN INTO TOKENS

↓

TOKENS GIVEN TOKEN IDs

↓

TOKENS CONVERTED INTO EMBEDDINGS

↓

NUMERICAL REPRESENTATIONS READY FOR FURTHER PROCESSING

Please go to part two for the continuation of this series.

Written by Bill Kochman

wordweaver777@gmail.com

<https://www.billkochman.com>

HOW CHATBOTS UNDERSTAND AND WRITE LANGUAGE -- PART 2

-- What is Under the Hood: A step-by-step, plain-English explanation for anyone who has never studied computers or delved into the amazing world of Artificial Intelligence

Copyright 1994-2026 Bill's Bible Basics

Published On: Sep. 29, 2026

Last Updated: Sep. 29, 2026

AI Neural Networks, Difference Between A Neural Network And A Human Brain, Important Role Of Weights In AI, Parameters, AI Training Steps: Trial, Error Check, Loss, Adjustment And Repeat, Pattern Recognition, Old Rules-Based Chatbots Vs The Modern AI Language Models, My Old Hotline Days And PowerBot, Defining Context, What Is Meant By Attention?, Transformers, A Review Of Technical Terms, Differences Between Weights And Attention, Probability Percentages And Next-Token Prediction, Why An AI Language Model May Sometimes Give Different Answers

Continuing our discussion from part one, let us now move on and talk about neural networks, as they relate to Artificial Intelligence and AI Chatbots.

STEP 3: THE "NEURAL NETWORK"

-- The Huge Mathematical System That Processes The Information

Now the AI has converted your language into numbers.

What happens to those numbers?

They are processed by a huge mathematical system called a NEURAL NETWORK. The term "neural network" comes from the fact that its design was loosely inspired by the way networks of biological neurons in the human brain are connected. However, it is very important that you do not take that comparison too literally.

An artificial neural network is NOT a miniature human brain. It is a mathematical and computational system. Modern Large Language Models contain enormous numbers of both mathematical operations and adjustable values. So the basic idea is that information passes through many interconnected mathematical layers, with each layer transforming the information in some way. A simplified picture looks something like this:

YOUR TEXT

↓

TOKENS

↓

TOKEN IDs

↓

EMBEDDINGS

↓

NEURAL NETWORK

↓

PREDICTION

WHAT ARE "WEIGHTS"?

Inside the neural network are enormous numbers of adjustable numerical values called WEIGHTS.

Weights are extremely important because they are part of what the AI learns during training.

A useful analogy is a gigantic control panel containing an enormous number of tiny adjustment dials.

Imagine a sound engineer sitting in front of a huge mixing board. The board contains many knobs and sliders. Changing one knob might make one instrument louder. Changing another might make another instrument quieter. So the final sound depends on the combined settings of all those controls.

AI weights are somewhat like those controls. The analogy is not exact, but it gives you the basic idea. During training, the AI adjusts its weights so that it becomes better at predicting what text should come next.

WHAT ARE "PARAMETERS"?

When people say that an AI model contains "billions of parameters," they are talking about enormous numbers of adjustable numerical values within the model. A PARAMETER is a general technical term for a numerical setting that the model learns. In many discussions of modern AI, weights make up a very large portion of those parameters.

TRAINING: HOW ARE BILLIONS OF DIALS SET?

No human programmer sits down and manually adjusts billions of weights one by one. That would obviously be impossible. Instead, the computer adjusts them automatically during a process called TRAINING. Training is the process through which the AI model learns statistical patterns from very large amounts of data. Here is a simplified version of the steps which are followed:

1. TRIAL

The computer is given text.

It is asked, in effect:

"What piece of text is likely to come next?"

For example:

"The sky is . . ."

The model might initially make a poor prediction.

2. ERROR CHECK

The computer compares its prediction with the actual next piece of text in the training example. If the prediction was wrong, the system calculates how wrong it was. This measurement is commonly called LOSS. Loss is essentially a mathematical measurement of how far the model's prediction was from the desired result.

3. ADJUSTMENT

The training system uses mathematics to determine how the model's internal numerical settings should be changed. The weights are adjusted slightly. The purpose is to make the model a little better at future predictions.

4. REPEAT

The process is repeated over and over again across enormous amounts of training data. The computer performs this process an enormous number of times. Over the course of training, the model's weights are gradually adjusted. Eventually, the model becomes remarkably good at recognizing and reproducing many of the patterns found in human language.

It is important to understand that this is not like a student sitting in a classroom consciously thinking:

"Today I learned the rule for commas."

In other words, the model is not learning in the human sense. It is adjusting enormous numbers of mathematical values based on patterns in the training data.

STEP 4: WHAT DOES THE AI ACTUALLY LEARN?

-- Recognizing Patterns In Language

This is one of the most important parts of understanding modern AI systems. The model is not simply a giant digital dictionary which contains a pre-written answer to every possible question. Instead, training causes the model to develop complex mathematical relationships that allow it to recognize patterns in language. For example, it can learn that certain words frequently appear together.

PEANUT BUTTER

. . . is a very common combination.

PEANUT MARGARINE

. . . is much less common.

The model can also learn patterns involving grammar. For example, it learns that . . .

"The dog runs."

. . . is a normal English sentence.

It learns relationships involving meaning and context.

Consider the word . . .

BANK

In . . .

"I deposited money at the bank."

. . . the word refers to a financial institution.

But in . . .

"We sat on the river bank."

. . . the word refers to the land beside a river.

Thus, the surrounding words help determine which meaning is appropriate. The AI model can also learn patterns involving writing style. For example, a formal business letter sounds different from:

- a poem
- a recipe
- a newspaper article
- a scientific paper
- a casual text message

- a joke
- a friendly conversation

Through training, the model develops the ability to generate language that fits many different situations and styles.

MODERN AI VS. OLDER RULE-BASED CHATBOTS

Older computer chatbots from decades ago often relied heavily on rules written directly by programmers, or even by a simple server admin such as myself. The Hotline chatbot I installed on my Mac LC III computer many years ago was called PowerBot. It was created by a Macintosh developer who went by the name of Virtual1. The main engine of the bot was the "Personality File", which was basically a plain text file which contained a long list -- literally thousands of text strings -- of Q&A pairs. Before I stopped using Hotline, my Personality File had grown to at least 100s of MBs in size, and perhaps more. I can't even remember now.

The way that the chatbot worked was that it would start at the very top of the Personality File, and run down the list of Q&A pairs until it found an exact match for whatever the user had typed into the text area. The appropriate response would be located directly below the matching prompt. Later, the bot did become a little more evolved than that, but that was its basic setup. So exactly how the bot would respond to the user was determined by how I ordered my various prompt and response pairs in the Personality File. The more complex strings had to be situated at the top of the section, with progressively simpler text strings below it. If there wasn't an exact match, then the bot couldn't respond to the user.

For example, if the Hotline client user typed in . . .

"Hello"

. . . then my PowerBot chatbot would respond with . . .

"Hi there!"

. . . or whatever I had added as a response in PowerBot's Personality File. As I mentioned earlier, depending on the Hotline server admin, the Personality File could contain thousands of such matching Q&A pairs, or rules. And again,

while these kinds of very early chatbots were indeed useful and obviously a lot of fun to use, they were often quite limited as well. If the Hotline client user typed something that the Hotline server admin had not anticipated and which he had failed to include in PowerBot's Personality File, then the chatbot would not respond, thus leaving the user hanging in silence.

Of course, modern AI language models work very differently. They are not simply following a gigantic collection of what basically amounted to a long list of manually written Q&A pairs. As we've now seen, instead, modern AI chatbots rely on patterns learned during training to generate responses dynamically. This allows them to deal with a considerably wider variety of questions and situations. That flexibility is one of the major differences between those traditional, old school, rule-based chatbots and modern language models.

STEP 5: "CONTEXT" AND "ATTENTION"

-- Looking At The Surrounding Words And Figuring Out How They Relate

WHAT IS CONTEXT?

Context refers to the surrounding information which helps the AI model to determine what something means. Humans use context constantly without thinking about it. For example, consider the word "bank".

If I say . . .

"I deposited a check at the bank."

. . . you understand that I mean a financial institution.

On the other hand, if I say . . .

"We sat on the river bank."

. . . then you obviously understand that I mean the edge of a river and NOT a financial institution, right? Now please note that the word "bank" itself has not really changed. It is in fact the surrounding words which have changed the meaning of the word "bank" in that instance. In short, that surrounding information is referred to as context.

AI language models also make extensive use of context. When you send a message, the model doesn't simply examine every word completely independently. It processes the words in relation to other words and pieces of information in the available conversation. This allows it to take into account different things such as the following:

- the words immediately surrounding a phrase
- earlier parts of your message
- previous messages in the conversation
- instructions that apply to the conversation
- other information included in the model's available context

WHAT IS "ATTENTION"?

One of the most important ideas in modern AI language models is called ATTENTION. Attention is a mathematical mechanism that allows the model to determine which other pieces of text are particularly relevant to the piece of text it is processing. In simple terms, attention helps the AI model answer questions such as the following:

"Which other words should I pay particular attention to when interpreting this word?"

This becomes especially useful when related words happen to be separated by many other words. For example, look at how changing just one verb completely flips the meaning of a sentence, and notice who the word "she" refers to in each case:

- Sentence A: "Maria returned the book to Susan because she had finished reading it."
- Sentence B: "Maria lent the book to Susan because she wanted to read it."

In Sentence A, our brains instantly connect "she" to Maria, because the person returning a book is the one who finished reading it.

In Sentence B, "she" suddenly shifts to mean Susan, because

the person receiving the loan is the one who wants to read it.

An older chatbot would struggle with this because it doesn't understand the relationship between all of the words in the sentence. A modern AI language model uses Self-Attention to look at the entire sentence simultaneously. The AI model is then able to mathematically map the connection between "she" and the correct person by analyzing the surrounding verbs such as "returned" or "lent."

So as you can see, attention helps the AI language model to calculate relationships between different parts of the text. A model can use attention to connect information that may be separated by many words. This is one reason modern AI models can handle much more complicated sentences and passages than older chatbot systems.

WHAT IS A "TRANSFORMER"?

You will often hear modern AI language models described as being based on a TRANSFORMER architecture. "Architecture" in this context simply means the basic design or structure of the computer system. A transformer is a particular type of neural-network design that became extremely important in modern AI because it provides powerful ways of processing relationships between pieces of language. Attention is a central part of the transformer design.

A QUICK REVIEW OF TERMINOLOGY

So you can think of the relationship between these various technical terms in the following manner:

NEURAL NETWORK

= the broad type of mathematical system

TRANSFORMER

= a particular architecture for building a powerful neural network

ATTENTION

= an important mechanism used within transformer-based systems to calculate relationships between pieces of

information

WEIGHTS VS. ATTENTION

These two concepts are easy to confuse. However, please understand that they are NOT the same exact thing. To reiterate:

WEIGHTS are learned numerical settings that become part of the trained model. They are like the model's long-term internal settings.

ATTENTION is a calculation performed while the model is processing particular input. It helps determine which pieces of the current information are related and relevant to one another.

A simplified analogy would be the following:

WEIGHTS = the settings the model learned during its training.

ATTENTION = what the model calculates about the particular information in front of it right now.

The distinction between the two is very important because the model's learned weights do not normally change every time you send a message. In other words, your conversation does not normally retrain the entire model.

STEP 6: "NEXT-TOKEN PREDICTION"

-- Calculating What Should Come Next

Now at last we arrive at the fundamental operation behind an AI's actual language generation. As we learned earlier, after processing the input, the AI language model needs to calculate which token should come next. This is referred to as NEXT-TOKEN PREDICTION. However, that word "prediction" can be misleading if you imagine the AI is looking into the future. The truth of the matter is that it isn't. The model is simply calculating probabilities based on the information it currently has.

Suppose that the beginning of a particular sentence is the following:

"The sky is . . ."

The AI language model might assign different probabilities to possible next tokens. For the purpose of illustration only, imagine that the AI language model has assigned the following probability percentages to these listed words:

"blue" = 75%

"cloudy" = 15%

"clear" = 8%

"banana" = 0.00001%

Please note that the previous numbers are only an imaginary example. Real AI models calculate probabilities over a very large vocabulary, and the actual numbers vary depending on the model and the exact context. The model then uses those probabilities to select a token. Once it has selected that token, the token becomes part of the growing response, or answer to the user's prompt -- meaning their question or comment.

The model then performs another prediction.

Then another.

Then another.

And another.

In simplified form, it would look like this if you could actually, and visually, see inside the language model:

"The sky is . . ."

↓

"The sky is blue . . ."

↓

"The sky is blue today . . ."

↓

"The sky is blue today, and . . ."

↓

and so on.

This process continues until the model determines that the response should end. Therefore, as we discussed earlier, the AI model does not normally retrieve an entire pre-written answer from a hidden database and simply paste it into the conversation. It is NOT an old school chatbot from the 1980s or the 1990s. It in fact generates the response step by step and TOKEN BY TOKEN. This is one of the most important things you need to understand regarding how a language model really produces text for its response output.

WHY DOES THE AI SOMETIMES GIVE DIFFERENT ANSWERS?

If the language model is calculating probabilities, there can sometimes be more than one reasonable possibility for what comes next. For example, consider the following:

"The little girl opened the . . ."

could be followed by . . .

"door"

"box"

"book"

"gift"

. . . or quite a few other possibilities for that matter.

Depending on the system's settings and the circumstances, the AI language model may select different possibilities on different occasions. This is one reason an AI can sometimes give different answers to very similar questions. As we have seen, the important point is that generating language is not simply a matter of looking up one fixed answer. It involves calculating possible continuations and selecting among them.

Please go to part three for the conclusion of this series.

Written by Bill Kochman

wordweaver777@gmail.com

<https://www.billkochman.com>

HOW CHATBOTS UNDERSTAND AND WRITE LANGUAGE -- PART 3

-- What is Under the Hood: A step-by-step, plain-English explanation for anyone who has never studied computers or delved into the amazing world of Artificial Intelligence

Copyright 1994-2026 Bill's Bible Basics

Published On: Sep. 29, 2026

Last Updated: Sep. 29, 2026

Generalization: Combining Words To Independently Create New Phrases And Sentences, Training Stage Vs Response Generation Stage, The Primary Goal Is To Create A Model That's Learned Useful Patterns In Language, Training Versus Inference, The Complete AI Response Cycle From Your Input To An AI's Output, What "Understanding" Means In The AI World, The Challenging Problem Of AI Hallucinations, My Frustrating Experience With OpenAI's ChatGPT, Why AI Chatbots Hallucinate, Fluency Versus Accuracy, Large Language Models' Overly Confident Responses, Importance Of Independent Verification Of The Facts, A Simple Version Of What Happens Under The Hood, a Review Of Important Definitions, One Final Analogy: An Enormous Automated Writing Machine, List Of Some Of The Capabilities Of Modern Chatbots, The Simplest Possible Summary, Closing Remarks, Reading List

Continuing our discussion from part two, let us now turn our attention to the subject and process within the world of AI agents and LLMs which AI engineers have assigned the name of Generalization.

STEP 7: "GENERALIZATION"

-- Handling New Combinations Of Words

Another important capability of AI language models is called GENERALIZATION. Generalization means that a system can apply patterns it learned from examples to situations that are in

fact different from those exact examples. Imagine that you're teaching a small child what a sentence is. You will offer the child many examples to learn from, such as the following:

"The dog runs."

"The cat sleeps."

"The boy plays."

Eventually, if he or she continues to persevere at the task, the child will learn how to create a sentence all on their own, even if nobody ever specifically taught that particular sentence to them. For example, consider the following brand new sentence which was formed by combining and expanding on two of the previous examples:

"The small boy plays in the sand while the dog happily runs right alongside of him."

Thus we see that the child combined things learned from some of their previous examples to create something entirely new. Well, as it turns out, as I explain in the five-part series "Conversations With Four AI Chatbots", when Large Language Models are first created, they are somewhat similar to that small child in that they too have to learn and grow. And as eventually occurred with that child, AI models can likewise learn to generate combinations of language which were NOT specifically stored as complete answers.

Furthermore, this is in fact one reason why an AI model can respond to unusual questions or requests which it has never even encountered word-for-word during its training period. However, there does exist one very important difference or distinction between this analogy and that human child, and that is the following: A Large Language Model does NOT learn through human experience, or consciousness, or even through human understanding. Quite to the contrary, the AI model's generalization comes from mathematical patterns which were learned during its training period.

STEP 8: TRAINING VS. GENERATING RESPONSES

-- The Classroom Vs. The Conversation

There are two very different stages that are important to distinguish here. Please carefully consider the following:

PHASE 1: TRAINING

Training is the long process during which a language model's internal numerical settings are adjusted. During the training stage, the language model processes enormous amounts of data and repeatedly makes predictions, measures its errors, and likewise adjusts its weights. The primary goal is to create an AI model that has learned useful patterns in language. A simple analogy would be the following:

TRAINING = going to school.

Imagine spending years studying books, doing exercises, taking tests, making mistakes, and gradually improving.

PHASE 2: GENERATING A RESPONSE AND INFERENCE

Once an AI language model has been properly trained, only then can people use it successfully and obtain satisfying results, or responses. In short, because of that training, when you type a question or comment into the prompt area of any chatbot, the already-trained model processes your input, and it generates an appropriate response. This stage of the game is commonly called INFERENCE. Inference simply means using the trained model to produce an output from a user's new input. A useful analogy would be the following:

TRAINING = going to school.

INFERENCE = using what you learned in school when having a conversation with someone.

However, the two processes should not be confused. When you are chatting with a normal deployed AI model, your individual conversation does not ordinarily cause the model's billions of learned weights to be immediately rewritten. Instead, your conversation supplies CONTEXT that the AI model can use while generating its current response.

Some AI systems may use conversations or other information later for separate training or improvement processes. It all depends on the service and its settings. That is different from the model instantly retraining itself during your conversation with it.

STEP 9: PUTTING IT ALL TOGETHER

-- The Complete AI Response Cycle

Let's now combine everything we have learned. When you send a message to an AI chatbot, a simplified version of the process looks like the following under the hood. As you can no doubt see, there is actually a lot going on. As we learned earlier, it is NOT simply a lot of copy and paste, as occurred with Hotline server chatbots three decades ago:

1. RECEIVE THE TEXT

The system receives the message you typed.

2. BREAK IT INTO TOKENS

The text is divided into smaller pieces called tokens.

3. CONVERT TOKENS INTO NUMBERS

The tokens are represented by Token IDs and then transformed into numerical representations called embeddings.

4. PROCESS THE INFORMATION

The neural network processes those numerical representations.

5. USE CONTEXT AND ATTENTION

The model examines relationships among the pieces of information available in the context.

6. CALCULATE PROBABILITIES

The model calculates which possible next tokens are likely to fit the context.

7. SELECT A TOKEN

The system chooses a token according to its generation process.

8. ADD THE TOKEN

The selected token becomes part of the growing response.

9. REPEAT

The model uses the expanded text as part of the information for the next prediction.

This process continues until the language model reaches an appropriate stopping point, and then it produces output on your device's screen in the form of a response. In a rather simplified form, the process looks like the following:

READ

↓

RELATE

↓

PREDICT

↓

ADD

↓

REPEAT

However, please remember that "READ" and "RELATE" are only human-friendly descriptions. The computer is not literally reading and thinking like a regular human being. It is in fact performing many mathematical operations on numerical representations of language.

AN IMPORTANT POINT ABOUT "UNDERSTANDING"

At this point in our discussion, you may reasonably ask the following question:

"Does the AI actually understand what I am saying?"

This is a surprisingly complicated question. That's because the word "understand" can actually mean different things. A human understands language through a combination of language, personal experience, perception, memory, physical interaction with the world, emotions, goals and through quite a few other things.

In contrast, as we have already seen, an AI language model processes language using mathematical representations and learned patterns. Again, as I stated near the beginning of this series, what you see on your screen when you talk with an AI chatbot is the result of pure science and very complex mathematical computations. That's why a Large Language Model can produce remarkably appropriate responses, and can often recognize subtle relationships between words and ideas.

However, again, this should not automatically be interpreted as meaning that the AI possesses human-like understanding or consciousness. When we say that an AI "understands" something in an everyday explanation, what we often really mean is that it has successfully processed the relationships and meaning conveyed by the language well enough, so that it can respond appropriately. That is a useful shorthand, but it should not be confused with saying that the machine experiences the world exactly as a human being experiences it.

A VITAL SAFETY NOTE: "HALLUCINATIONS"

One of the most important weaknesses of AI language models is that they can sometimes produce information that sounds convincing, but which in fact is incorrect. This is often called a HALLUCINATION. The word is an analogy. It does NOT mean that the computer is literally seeing or hearing things that are not there. In AI terminology, such a hallucination generally refers to generated content that is simply false, unsupported, or even fabricated even though it may sound perfectly believable. For example, an AI chatbot might do any of the following:

- invent a quotation that a person never said
- give a nonexistent book title
- provide a false historical detail
- invent a scientific reference
- misunderstand a question
- perform a calculation incorrectly
- confidently state something that simply isn't true

If you have read my five-part series entitled "Conversations With Four AI Chatbots", then you will already know that this is exactly what I was experiencing about a year ago when I first started using OpenAI's ChatGPT. That bot was driving me absolutely crazy with its inaccuracies, inventions and hallucinations; so much so, that eventually, even though it kept apologizing for the mistakes it was making -- which it continued to make anyway despite all of its apologies -- I shut down my OpenAI account and I switched to primarily using Anthropic's Claude.

So the big question is this: Why can this happen? What is the cause behind an AI chatbot hallucinating so much? It happens because the Large Language Model's fundamental job is simply to generate likely sequences of text. It is NOT automatically checking every statement against reality before it actually states something or prints it out. If an AI had to fact-check every single sentence before it actually typed it out, its responses would not be as spontaneous. Exactly how much of a pause there would be between each of its responses, I cannot really say. At any rate, this distinction is very important.

FLUENCY DOES NOT EQUAL ACCURACY.

FLUENCY means that the writing sounds natural, smooth, organized, and convincing.

ACCURACY means that the information corresponds to what is actually true.

The caveat is this: An AI chatbot can be extremely fluent in your language, while at the same time being completely wrong. For example, it might produce a beautifully written paragraph containing a false statement. The fact that a paragraph may sound confident does NOT necessarily make it true. In fact, I can tell you from my personal experiences with ChatGPT and Gemini that there were a lot of instances where they would project extreme confidence in what they were telling me, and even double down, and yet, nevertheless, they were still very wrong with the "facts" or their conclusions.

A lot of times they would guess at certain things over and over again, because they were so determined to provide me with the correct answer, or to find the right solution to a particular problem. I discuss this same issue more at length in the series "Conversations With Four AI Chatbots". I hope that you'll take the time to read it, because you will learn

a lot from doing so.

But the lesson that we can learn here is rather clear. This is exactly why without exception, all important information should be independently verified against reliable external sources. This is particularly important for certain types of information, such as the following:

- medical information
- legal information
- financial information
- historical claims
- scientific claims
- quotations
- statistics
- current events
- names, dates, and specific factual details

In short, an AI chatbot can indeed be a remarkably useful tool. However, it should not automatically be treated as if it is an infallible source of truth, because it obviously is no such thing. So again, please check your facts before you regret not having done so!

THE BIG PICTURE

If all of the previous information still seems like too much for you to have to remember, then following is the simplest possible version of what really happens under the hood, and out of sight, the minute that you choose to type a prompt -- meaning a question or a comment -- into the text field of an AI chatbot.

When you type something into an AI chatbot:

YOUR WORDS

↓

are broken into

TOKENS

↓

which are represented as

NUMBERS

↓

those numbers are processed by a huge

NEURAL NETWORK

↓

using learned

WEIGHTS

↓

while taking account of

CONTEXT

and relationships calculated through

ATTENTION

↓

to calculate probabilities for

WHAT TOKEN SHOULD COME NEXT

↓

one token is selected

↓

then another

↓

then another

↓

until a complete response has been generated.

THE MOST IMPORTANT IDEAS/DEFINITIONS TO REMEMBER

1. TOKENS

Tokens are the small pieces into which text is divided so that the computer can process language.

2. TOKEN IDs

Token IDs are numbers used to identify particular tokens.

3. EMBEDDINGS

Embeddings are numerical representations that allow the model to represent relationships between tokens mathematically.

4. NEURAL NETWORK

A neural network is a huge mathematical system that processes those numerical representations.

5. WEIGHTS

Weights are learned numerical settings inside the model that strongly influence how it processes information and generates responses.

6. PARAMETERS

Parameters are learned numerical values within the model. Modern AI models can contain billions or even more of them.

7. TRAINING

Training is the process of adjusting those numerical values using enormous amounts of data so that the model becomes better at its tasks.

8. CONTEXT

Context is the surrounding information that helps determine what words and ideas mean in a particular situation.

9. ATTENTION

Attention is a mathematical mechanism that helps the model determine which pieces of information are related and relevant to one another.

10. TRANSFORMER

A transformer is a particular neural-network architecture that uses mechanisms such as attention to process relationships within sequences of information.

11. NEXT-TOKEN PREDICTION

This is the basic process of calculating which token should come next based on the information currently available.

12. INFERENCE

Inference is the process of using a trained model to generate an answer from new input.

13. GENERALIZATION

Generalization is the ability to apply learned patterns to new combinations and situations.

14. HALLUCINATION

A hallucination is an AI-generated statement or other content that is incorrect or fabricated even though it may sound convincing.

ONE FINAL ANALOGY

If you want one mental picture to help you to remember all of the previous information, imagine an enormous automated writing machine.

You put a sentence into the machine.

The machine breaks your sentence into pieces.

It converts those pieces into numbers.

It passes those numbers through an enormous network of mathematical operations whose settings were established through training.

The system examines the relationships among the pieces of information available to it.

It then calculates which piece of text would be a good one to place next.

It adds that piece.

Then it calculates the next one.

And the next.

And the next.

It continues doing this extremely rapidly until it has produced an entire response.

The response which then appears on your device's screen can be extraordinarily sophisticated.

It can explain complicated subjects.

It can write stories.

It can summarize documents.

It can translate languages.

It can help write computer programs.

It can answer questions.

It can carry on remarkably natural conversations.

But underneath all of that apparent conversational ability is an enormous mathematical system that has learned patterns from data and uses those patterns to generate new sequences of language. Understanding that basic process is the key to understanding what modern AI language models actually do. As I have said two times now, it is all pure science and very complex mathematical operations, and it also all happens at

a very tremendous rate of speed. Certainly much faster than you or I can possibly ever think. Just wait a few more years from now. Our current world may be very different due to AI.

THE SIMPLEST POSSIBLE SUMMARY

AI Does Not Begin With A Pre-Written Answer.

Instead:

It breaks language into pieces.

It turns those pieces into numbers.

It processes those numbers through a huge learned mathematical system.

It examines relationships and context.

It calculates probabilities for what should come next.

It selects a token.

It repeats the process.

And, one piece at a time, it generates a response.

That is the basic mechanism behind the astonishing ability of modern AI chatbots to communicate with us in any human language. Isn't it mind-blowing?

----- End Quote -----

And that, my dear friends, is the end of this amazing lesson regarding exactly how Artificial Intelligence and our popular chatbots actually work. You only see the end result on your device's screen; but under the hood, they are just so busy, busy, busy working constantly to provide you with the info that you want or need, thus saving you piles of time.

With these thoughts, I will bring this article to a close. It is my hope that you have found it informative, enlightening, and I pray that it has been a blessing in your life as well. If you have an account with Facebook, Twitter, Tumblr or with any other social network, I would really appreciate if you'd take the time to click or tap on the corresponding link that is found on this page. Thanks so much, and may God bless you

abundantly!

For additional information and further study, you may want to refer to the list of reading resources below which were either mentioned in this article, or which contain topics which are related to this article. All of these articles are likewise located on the Bill's Bible Basics web server. To read these articles, simply click or tap on any link you see below.

AI, Chatbots, Daemons and Demons

AI, Deepfakes and Humanoid Robots

Be Careful With ChatGPT and Other AI Chatbots!

Cognitive Computers, DARPA, OpenAI, AGI, Superintelligence and Elon Musk

Conversations With Four AI Chatbots

Is Elon Musk the Antichrist?

Power of the Purse and the Federal Reserve System

Robot Wars and Skynet: Is Sci-Fi Becoming Our Reality?

Rogue AIs: We've Been Warned!

Science and Technology: The Forbidden Knowledge?

The Internet: Our Final Frontier; Your Last Chance?

Written by Bill Kochman

wordweaver777@gmail.com

<https://www.billkochman.com>