# How LLM works: in a nutshell.

Article 2 of 6

BY ENRIQUE FARFAN, PhD, PE

*In the ever-evolving landscape of the geoprofessional industry, artificial intelligence continues to drive innovation, efficiency, and curiosity. This article, focusing on large language models, is the second in a six-part series written to educate Geoprofessional Business Association (GBA) members on the implications of AI on the geoprofessions. The article was prepared by GBA Business Technology Committee member, Enrique Farfan, PhD, PE, ENV SP, of HDR.*

The development of Generative Artificial Intelligence (Gen AI) has opened the gates of a new industrial revolution. Gen AI provided the framework of algorithms that allow the models to generate entirely new outputs rather than only focus on making predictions from a set of inputs or experiences as Traditional AI. This incredible capability shift from prediction to creation has revolutionized the AI industry and its applications.

Large language models (LLMs), based on transformer models, are deep learning algorithms trained on massive datasets to perform a range of natural language processing (NLP) tasks, including translation, prediction, and content generation. LLMs can extend their capabilities beyond language to understand protein structures and write software code. They undergo pre-training and fine-tuning to achieve proficiency in text classification, question answering, summarization, and text generation. With their extensive parameters serving as a knowledge bank, LLM finds applications in healthcare, finance, entertainment, and various NLP fields like translation and chatbots, enhancing AI assistants, and more. A transformer model is a neural network designed to grasp context and meaning by identifying connections within sequential data, such as the words in a sentence.

ChatGPT-4 is said to be based on eight models with 220 billion parameters each, for a total of about 1.76 trillion

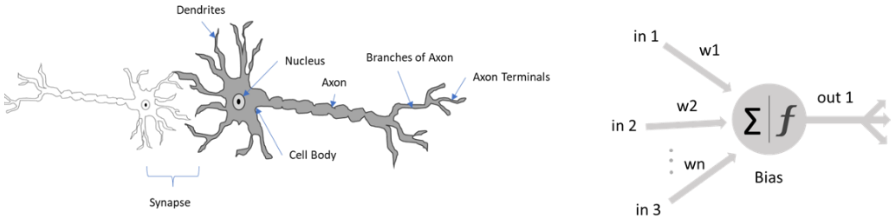
Chat Generative Pre-Trained Transformer (ChatGPT) is one of the industry's most popular and powerful LLMs, taking the lead in one of the most extraordinary races in technology in the last decades. The competition is not only limited to big conglomerates like Google and Amazon but also has brought nations concerns about trying to develop their own LLM. Chinese company Baidu has launched its version of LLM, ERNIE 4.0, as a direct competition to ChatGPT 4.0. Google has tried to elevate the game, evolving from BERT (Bidirectional Encoder Representations from Transformers) to its latest project, Gemini, which can take video input. Google got a slap on the wrist after some disappointing results on its image Gen AI. ChatGPT 5.0 is on the verge of its release, and its capabilities will make you feel like you are chatting with someone.

In a very simplified manner, ChatGPT predicts each word using a neural network model that predicts the next word's probability. The probable values are calculated based on the training dataset.

OpenAI, now an $80 billion company, has decided to protect some information about the ChatGPT-4 model, citing a competitive environment. The system is said to be based on eight models with 220 billion parameters each, for a total of about 1.76 trillion parameters. In comparison, BERT is trained with 345 million parameters.

The first step to understanding how LLM works is to get familiar with how neural networks work since this model has become one of the most important building blocks in the AI's new frontier.

Artificial neural networks are modeled after biological neural systems, and they consist of interconnected computational neurons that adapt through weight adjustments, akin to synaptic connections. They learn from training data, similar to biological stimuli, and iteratively refine weights based on prediction errors, allowing for accurate predictions, referred to as model generalization. Despite their simplified representation of the human brain, neural networks harness neuroscience principles to design their architectures. By combining multiple units and jointly training weights, they achieve deep learning, gaining popularity with increased data availability and computational power, leading to more accurate models. This architecture is depicted in Figure 1.



a) Biological neural network

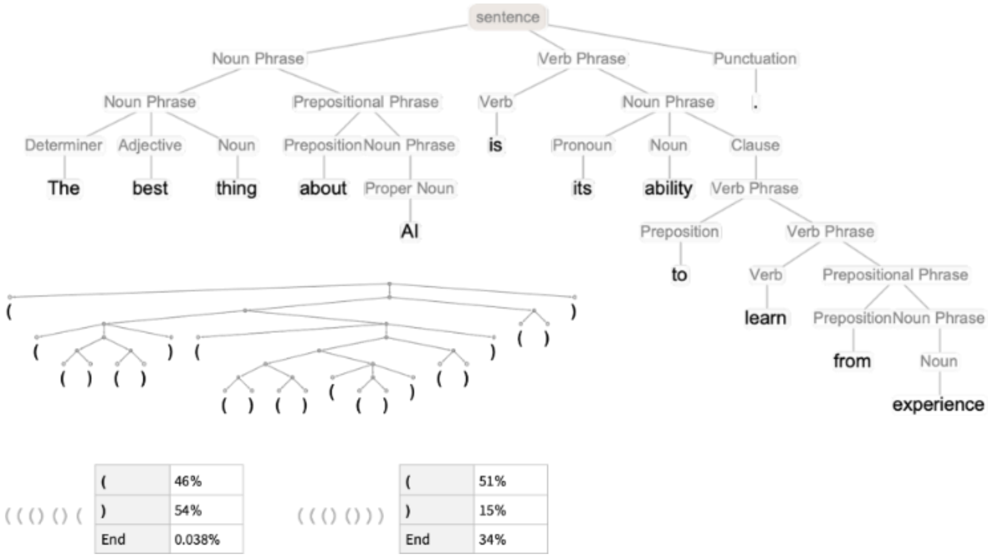
b) Artificial neural network

***Figure 1.*** *A.I. general algorithms*

Activation functions ($f$) in neural networks are mathematical equations that act as switches that control whether a neuron should be activated ("fired") or not and pass its information to the next layer based on whether the input is relevant for the prediction. They allow the network to learn and make decisions from complex data, making it smarter over time.

Now, you can imagine millions of these neurons arranged in layers and interacting and sending information to each other. The ultimate black box, our brains, comprise approximately 100 billion neurons.

The LLM takes advantage of the neural networks to generate their responses. For example, ChatGPT establishes the sentence structure and the next probable following word in the sentences based on probabilities, as illustrated in Figure 2.



| | |
|---|---|
| ( | 46% |
| ) | 54% |
| End | 0.038% |

| | |
|---|---|
| ( | 51% |
| ) | 15% |
| End | 34% |

**Figure 2.** Grammatical rules

*[Stephen Wolfram (2023), "What Is ChatGPT Doing ... and Why Does It Work?," Stephen Wolfram Writings. writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work]*

As we mentioned, the grammatical structures in ChatGPT are derived from the training dataset. Each word is added one at a time based on the probability and connection between words in the training data.

This example was developed in ChatGPT-3.5 to illustrate how the LLM selects the next word.

Using ChatGPT-3.5:

Complete the following sentence: "Geotechnical Engineering is the art of ……."

| Next Word | Probability |
|---|---|
| **Analyzing** | 0.234 |
| **Designing** | 0.178 |
| **Understanding** | 0.124 |
| **Optimizing** | 0.093 |
| **Ensuring** | 0.071 |

***"Geotechnical engineering is the art of*** *analyzing, designing, understanding, optimizing, or ensuring the stability of structures and foundations."*

Every time you query ChatGPT, it will generate a similar list of words with different probabilities.

ChatGPT does not always choose the word with the highest probability. A randomness built into the system allows the program to mimic "creativity" and helps develop complex responses. A parameter called *temperature* determines how often lower-ranked words are used. Zero temperature indicates that the system will select the word with the highest probability, while higher temperature allows the system to select words with lower probability. A temperature equal to 0.8 provides text and structures perceived as human-like responses.

ChatGPT's training data includes examples of proper sentence structure to recognize when a sentence should end based on punctuation and context.

| Word | Part of Speech |
|---|---|
| **Geotechnical** | Adjective |
| **engineering** | Noun |
| **is** | Verb |
| **the** | Article |
| **art** | Noun |
| **of** | Preposition |

Every word is related to other words, and every word is represented by an embedding vector. In the following example, a hypothetical embedding number was assigned to the group of words related to geotechnical engineering. In reality, the "numbers" would be vectors, often with hundreds of dimensions, and they wouldn't be sequential or as neatly organized as a simple integer. The latest embedding vectors in ChatGPT-4 have several hundred dimensions; ChatGPT-3 used embedding vectors of size 12,288 for its largest variant. Using the embedding vectors is how every word and phase is organized in the system. The embedding vector represents words and phrases that capture semantics, syntactic, and contextual relationships; it is a point in a multi-dimensional space where similar words cluster together.

| Word | Hypothetical Embedding Number |
|---|---|
| Geotechnical | 5832 |
| Engineering | 5738 |
| Soil | 3920 |
| Foundation | 4150 |
| Earthquake | 4523 |
| Construction | 5871 |
| Stability | 4632 |
| Excavation | 4719 |
| Infrastructure | 5893 |
| Surveying | 4902 |

To illustrate the magnitude and complexity of these multidimensional embedding vectors, Figure 3 shows a 768 x 768 color matrix of weights using GPT-2 for the words "hello, bye" after the embedding vectors were affected by the attention blocks (the size of this matrix in GPT-3 is 12,288 x 12,288).

An attention block is a computational component that allows the model to focus on relevant parts of the input data by assigning different levels of importance to each part, thereby enhancing its understanding and processing capabilities. GPT-2 operates 12 attention blocks, while GPT-3 operates 96 attention blocks.

*Figure 3. 768 x 768 matrix of weights for "hello, bye"*

*[Stephen Wolfram (2023), "What Is ChatGPT Doing ... and Why Does It Work?," Stephen Wolfram Writings. writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work]*

In a simplified manner, these are the steps that a LLM follows to generate content:

1. **Input Text:** The process begins with the input text provided by the user.
2. **Tokenization:** The input text is broken down into smaller pieces known as tokens, such as words or subwords.
3. **Input Embedding:** Each token is then converted into a numerical vector using the input embedding process.
4. **Positional Encoding:** To each input embedding, positional information is added to maintain the order of words.
5. **Transformer Encoder Layers:** The combined embeddings pass through multiple layers of the transformer encoder, where the model applies self-attention and processes the information.
6. **Output Generation:** The encoder's output is then used to generate a predictive output, often in the form of embedding vectors for the next likely tokens.
7. **Probabilities Calculation:** Finally, these output embeddings are transformed into probabilities, typically using a softmax function, to determine the most likely next word or sequence of words.

LLMs are just testaments to human ingenuity.

---

**Previous Articles in the Series**

- AI Unveiled: The Wizardry Behind Chatbots and Intelligent Systems

**Upcoming Articles**

- The Ghost in the Machine: Hallucinations.
- Superprompts: How to Talk with the Genie and Keep it in the Bottle.
- Role Play Your Way: How to Talk with Terzaghi
- Applications of LLM in Geotechnical Engineering: Use it or Lose.

## ABOUT THE AUTHOR

Enrique Farfan is a Geotechnical and Structural Engineer with a Ph.D. in Engineering and a Master's in Civil Engineering from the University of New Mexico. His expertise spans across various projects in mining, energy, water resources, and transportation, including work on dams, levees, canals, waterfront structures, bridges, foundations, and seismic analysis. Enrique has implemented diverse design solutions using customized computer programs and database implementations. During his graduate years, he developed an interest in fuzzy logic, optimization, inverse problems, neural networks, and computer programming. Now, he enjoys exploring the fascinating universe of AI and its applications.