Large Language Models: How They Really Work

Introduction

Large Language Models (LLMs) are becoming an important part of our daily lives. They help us write emails, translate languages, answer questions, and even create music (Coelho, 2025). But how do these models really work? In this essay, I will explain LLMs in simple English, using many examples, so that high school students can understand. I will cover what LLMs are, what data they learn from, what tokens and embeddings are, how LLMs learn patterns, what parameters are, how training and fine-tuning work, what Retrieval-Augmented Generation (RAG) is, why LLMs sometimes make mistakes (hallucinate), their strengths and weaknesses, and how they are used in real offices.

What Are Large Language Models?

A Large Language Model (LLM) is a computer program that can read, understand, and write human language. It is called "large" because it has learned from a huge amount of text and has many internal settings (called parameters). The most popular LLMs are ChatGPT, Gemini, and others. These models can answer questions, write stories, summarize articles, and much more.

For example, when you type a question into ChatGPT, it understands your words and tries to give you a helpful answer. This is possible because the model has learned from reading millions of books, articles, and web pages.

LLMs are a type of artificial intelligence (AI), but not all AI is an LLM. For example, computer vision is another branch of AI that helps computers "see" images (Chowdhury, 2025). LLMs focus on language.

What Data Do LLMs Learn From?

LLMs learn from a huge collection of written text. This text can come from:

- Books (fiction and non-fiction)
- Wikipedia articles
- News websites
- Social media posts
- Scientific papers
- Online forums

Imagine a giant library, filled with every type of book, newspaper, and website you can think of. The LLM "reads" this library to learn how people use language.

For example, if an LLM reads many recipes, it learns the words and steps people use when cooking. If it reads news articles, it learns how people describe events. This is why LLMs can answer so many different types of questions.

However, not all data is perfect. Sometimes, the information can be wrong or biased. That is why LLMs can sometimes give incorrect or unfair answers (Siddharth et al., 2025).

What Are Tokens?

Computers do not understand whole words the way humans do. Instead, LLMs break text into smaller pieces called "tokens." Tokens can be words, parts of words, or even single letters or symbols.

For example: - The sentence "I am happy." might be broken into these tokens: ["I" , "am" , "happy" , "."] - The word "unhappiness" could be split into ["un" , "happiness"]

Tokens help the LLM process language step by step. If you think of a sentence as a train, tokens are the train cars. The model looks at each car to understand the whole train.

What Are Embeddings? (Simple Example)

Once text is split into tokens, each token is turned into a set of numbers called an "embedding." These numbers help the LLM understand the meaning and relationships between words.

Simple Example: Imagine you have three words: "cat" , "dog" , and "car." The LLM gives each word a list of numbers:

- "cat" → [0.1, 0.8, 0.2]
- "dog" → [0.2, 0.7, 0.3]
- "car" → [0.9, 0.1, 0.5]

The numbers for "cat" and "dog" are close to each other, which shows they are both animals. "Car" is different, so its numbers are farther away. These numbers help the LLM "feel" that "cat" and "dog" are similar, but "car" is different.

Embeddings are like coordinates on a map, where similar words are close together.

How LLMs Learn Patterns

LLMs learn patterns in language by reading lots of text and noticing which words often appear together. For example, if the model reads many sen-

tences like:

- “The cat sat on the mat.”
- “The dog sat on the mat.”
- “The bird sat on the mat.”

It learns that “sat on the mat” is a common ending, and “cat,” “dog,” and “bird” are things that can sit.

The model does not simply memorize every sentence. Instead, it learns the rules and patterns behind the language. This is similar to how students learn grammar by reading and practicing, not just by memorizing every sentence in a textbook.

LLMs use a type of artificial neural network called a “transformer” to spot these patterns. It looks at the words before and after a token to guess what comes next.

What Are Parameters?

Parameters are the settings inside the LLM that decide how it processes and produces language. You can think of parameters like the dials and switches inside a radio. When you turn the dials, the radio sounds different. In an LLM, there are millions or even billions of these dials.

During training, the model adjusts its parameters to better predict the next word in a sentence. The more parameters, the more complex patterns the model can learn.

For example, a small LLM might have 100 million parameters. A very large model, like GPT-4, can have over 100 billion parameters.

The Training Process (Explained Simply)

Training an LLM is like teaching a child to finish your sentences. The computer is given lots of sentences with words missing, and its job is to guess the missing word.

For example: - “The sun rises in the **.” (Answer: “east”) - “She drank a glass of** .” (Answer: “water”)

Every time the model guesses correctly, its parameters are adjusted to reinforce that pattern. When it guesses wrong, it learns from the mistake and tries to do better next time.

This process is repeated millions or billions of times with different sentences. Over time, the LLM gets better at predicting the correct word, and thus at understanding and producing language.

Training requires powerful computers and a lot of electricity, which can have environmental impacts (Siddharth et al., 2025).

How Fine-Tuning Works

After the main training, the LLM can be "fine-tuned" for special tasks. Fine-tuning means training the model further on a smaller, more specific set of texts.

For example, if you want an LLM to be good at medical questions, you can fine-tune it on medical books and articles. If you want it to answer legal questions, you can fine-tune it on legal documents.

Fine-tuning helps the model become more accurate in certain areas, like a student who takes extra lessons in math or music to get better at those subjects (Chowdhury, 2025).

What Is Retrieval-Augmented Generation (RAG)?

Sometimes, even a well-trained LLM does not know the latest facts. Retrieval-Augmented Generation (RAG) is a method that helps LLMs find up-to-date or specific information.

Here's how it works:

1. When you ask a question, the LLM searches a database (like Wikipedia or company files) for relevant documents.
2. It reads the most relevant documents.
3. It uses the information from those documents to write a better answer.

Example:
If you ask, "Who won the 2024 Olympics?" the LLM might not know if it was trained before 2024. With RAG, it searches the web for the latest results, then gives you the current answer.

RAG makes LLMs more reliable and accurate, especially for new or changing information.

Why LLMs Hallucinate

Hallucination means the LLM gives an answer that sounds correct but is actually wrong or made up.

Example: - You ask: "Who is the president of Mars?" - The LLM might say: "The president of Mars is John Smith."

Of course, Mars doesn't have a president! The LLM made up the answer because it was trying to fill in the blank with something that "fits" the question, even if it's nonsense.

LLMs hallucinate because:

- They don't really "know" facts; they guess based on patterns.
- If they have not seen the information in their training data, they try to generate something that sounds plausible.
- They may mix up details from different sources (Weichert & Eldardiry, 2025).

This is why it's important to double-check LLM answers, especially for important topics.

Strengths and Weaknesses of LLMs

Strengths

- **Speed:** LLMs can read and write faster than any human.
- **Versatility:** They can answer questions, write stories, translate languages, summarize documents, and more.
- **Availability:** They are always ready to help, 24/7.
- **Personalization:** LLMs can be fine-tuned for special tasks, like medicine, law, or education.

Example:
An LLM can help a student write an essay, a doctor look up medical facts, or a businessperson summarize a long report.

Weaknesses

- **Hallucination:** LLMs sometimes make up facts or give wrong answers.
- **Bias:** If the training data has bias, the model may repeat it in its answers (Siddharth et al., 2025).
- **No Real Understanding:** LLMs do not truly "understand" or "think" like humans. They do not have feelings or beliefs.
- **Environmental Impact:** Training large models uses a lot of electricity and water (Siddharth et al., 2025).
- **Privacy Risks:** LLMs can accidentally reveal private information if it was in their training data (Feffer et al., 2023).

Real Office Examples

LLMs are already used in many offices around the world. Here are some real examples:

1. Customer Support

Many companies use LLMs to answer customer questions online. For example, a bank might use an LLM-powered chatbot to answer questions like

"How do I reset my password?" or "What are your business hours?"

Benefit: Reduces wait times for customers and frees up human workers for more complex tasks.

2. Writing Emails and Reports

Some businesses use LLMs to help write emails, reports, or job descriptions. For example, a human resources manager can ask the LLM to draft a job posting, and then edit the draft as needed.

Benefit: Saves time and helps people who are not confident writers.

3. Research Assistance

In law firms, LLMs can summarize long legal documents or find relevant cases quickly. In medicine, LLMs can help doctors look up the latest research articles.

Benefit: Helps professionals process information faster and make better decisions.

4. Translation

International companies use LLMs to translate emails and documents into different languages, making it easier to work with people from around the world.

5. Creative Work

LLMs are used in creative fields too. For example, musicians use AI to generate new song ideas (Coelho, 2025), and marketers use LLMs to brainstorm advertising slogans.

6. Education

Teachers use LLMs to create quizzes, explain topics, or provide feedback on student writing (Chowdhury, 2025). Students can use LLMs as tutors to get help with homework.

7. Office Automation

Companies use LLMs to read and sort emails, schedule meetings, or generate meeting notes automatically.

Example:
A real estate office might use an LLM to scan incoming emails and automatically sort them into folders: “New Clients,” “Appointments,” “Questions,” etc.

Conclusion

Large Language Models are powerful tools that are changing the way we work, learn, and create. They are trained on massive amounts of text, break words into tokens, use embeddings to understand meaning, and adjust millions of parameters to learn patterns in language. LLMs can be fine-tuned for special tasks and enhanced with RAG to find up-to-date information.

However, LLMs have weaknesses, such as hallucination, bias, and environmental impact. They do not truly understand language but are very good at finding and repeating patterns. In offices, LLMs are used in customer support, writing, research, translation, creative work, education, and more.

As LLMs become more common, it is important for everyone—including high school students—to understand how they work, their strengths, and their risks. This knowledge will help us use these tools wisely and safely in the future.

References

Chowdhury, T. (2025). Computational Thinking with Computer Vision: Developing AI Competency in an Introductory Computer Science Course. Retrieved from https://arxiv.org/pdf/2503.19006v1

Coelho, G. (2025). AI in Music and Sound: Pedagogical Reflections, Post-Structuralist Approaches and Creative Outcomes in Seminar Practice. Proceedings of the 6th Conference on AI Music Creativity (AIMC 2025). Retrieved from https://arxiv.org/pdf/2511.17425v1

Feffer, M., Martelaro, N., & Heidari, H. (2023). The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements. Retrieved from https://arxiv.org/pdf/2310.06269v1

Siddharth, S., Prince, B., Harsh, A., & Ramachandran, S. (2025). The World of AI: A Novel Approach to AI Literacy for First-year Engineering Students. Retrieved from https://arxiv.org/pdf/2506.08041v1

Weichert, J., & Eldardiry, H. (2025). Educating a Responsible AI Workforce: Piloting a Curricular Module on AI Policy in a Graduate Machine Learning Course. Retrieved from https://arxiv.org/pdf/2502.07931v1