

Understanding Large Language Models (LLMs): How Do They Collect Data?

[Assem Hijazi – Experience, Knowledge & Ideas](#)

I'll continue writing about LLMs (Large Language Models) to help explain how they work in simple terms. With each article, we'll dive a bit deeper to build a solid understanding, particularly of **generic LLMs**. We'll also explore their limitations.

How Does an LLM Collect Data?

LLMs like GPT are trained on massive amounts of data, which helps them understand and generate human-like text. But how exactly do they collect this data, and where does it come from?

1. Data Collection Methods

LLMs don't actively browse the web to gather information in real-time. Instead, during their training phase, they are exposed to large datasets that are compiled from publicly available sources. These sources may include:

- **Web pages:** Publicly accessible websites that are scraped for training purposes.
- **Books and research articles:** Digitized books and academic papers provide structured knowledge.
- **Social media:** Public posts and conversations that reflect everyday language.
- **Documents and public reports:** Official documents, Wikipedia entries, and other publicly available knowledge repositories.

This process, known as **scraping**, involves collecting and organizing huge amounts of data from these sources to create the model's training dataset.

2. Does the LLM Store Source Information?

When an LLM is trained, it processes a lot of data from various sources. However, the important thing to know is that the LLM doesn't actually **store the original sources**. Instead, it learns **patterns and relationships between words** during training and uses those to generate its responses.

3. Data Usage: Datasets, Not Direct Sources

Once the data has been collected, the LLM doesn't store it as an exact copy. Instead, it's processed into **datasets** that allow the model to learn **language patterns**. The LLM doesn't go back and pull specific content from these datasets when answering questions. Rather, it has already absorbed the patterns and generates responses based on what it learned during training.

4. The Problem with Trustworthiness

Not all data is created equal. LLMs are trained on a wide range of topics to build a broad understanding of language, but the **trustworthiness** of the sources they are trained on can be a critical issue. Since LLMs pull from a vast range of data, including **unverified sources**, the accuracy of the information can vary.

- Because the LLM doesn't store or track the original sources, there's no way to **verify** where it got its information from. This means that sometimes, the model might provide an answer that sounds confident but could be based on **unreliable** or uncertified data.

- For instance, the LLM might have learned something from a **reliable academic journal**, but it could also have learned from a **random blog**. Without tracking sources, you can't always be sure of the **accuracy** or **trustworthiness** of the answer.

5. Mathematical Data Structures in LLMs

The LLM is built on a complex **mathematical structure** made up of **tokens**, **vectors**, and **parameters**. Each word (or token) is represented by a **vector**—a series of **1,024 numbers**—which changes dynamically based on the input and the relationships between words. The LLM processes these relationships through **multiple layers**, refining them to generate responses that fit the context of the prompt.

Looking Ahead: Retrieval-Augmented Generation (RAG)

New techniques like **Retrieval-Augmented Generation (RAG)** aim to overcome some of the limitations of LLMs by allowing them to retrieve and incorporate **real-time information** from external sources. This helps keep the data up to date and more relevant, addressing some of the challenges with **outdated** or **unverified data**.

In the next article, we'll dive into **RAG** and **fine-tuning** in simple terms, explaining how they enhance LLMs and make them more effective in specialized contexts.