Lecture 2: Large Language Models (LLM) Basics

1. What is an LLM?

At its core, a Large Language Model (LLM) is a **neural network** designed to understand, generate, and respond to human-like text.

- **Neural Network Definition:** These models are inspired symbolically by the circuitry of the human brain, featuring input layers, hidden layers of neurons, and output layers.
- **Functionality:** Unlike simple computer programs, LLMs are designed for generic text tasks. Dr. Dander demonstrates this by asking ChatGPT to "plan a relaxing day," showing that it converses almost exactly like a human rather than a robotic script,.
- **Core Definition:** They are deep neural networks trained on massive amounts of data to perform specific tasks such as understanding and generating text.

2. What does "Large" mean?

The term "Large" specifically refers to the **number of parameters** (weights/variables) within the model.

- **The Scale:** We have moved from models with millions of parameters to those with billions and even trillions.
- **Evolution of GPT Parameters:**
    - **GPT-1:** ~100 million parameters.
    - **GPT-2:** ~1.5 billion parameters.
    - **GPT-3:** 175 billion parameters.
- **Historical Context:** A graph cited from the journal *Nature* shows that prior to 2020, models rarely exceeded 100,000 parameters. Around 2020, parameter counts exploded, reaching the billions and trillions seen today.

3. LLMs vs. Earlier NLP Models

The lecture highlights two main differences between modern LLMs and older Natural Language Processing (NLP) models:

- **Generalists vs. Specialists:**
    - **Earlier NLP:** These were "specialists." You needed a specific model for translation and a completely different model for sentiment analysis.
    - **Modern LLMs:** These are "generalists." A single architecture (like GPT) trained for text completion can also perform translation, sentiment analysis, and summarization.
- **Complex Instruction:** Earlier models could not handle complex, creative tasks like "draft an email to a colleague about movie tickets with emojis," whereas modern LLMs find this trivial.

4. The "Secret Sauce": Transformers

Dr. Dander identifies the **Transformer architecture** as the specific reason LLMs have become so powerful.

- **Origin:** The architecture was introduced in the 2017 Google paper titled *"Attention Is All You Need"*.
- **Impact:** This paper revolutionized AI, receiving over 100,000 citations in just five years.
- **Key Concepts:** The paper introduces complex mechanisms such as "positional encoding," "dot product attention," and "key-query-values," which serve as the foundation for modern AI.

5. Demystifying Terminology (The Hierarchy)

The lecture uses a nested umbrella analogy to clarify confusing industry terms,:

1. **Artificial Intelligence (AI):** The broadest bucket. It includes any machine exhibiting intelligence, including simple **rule-based** chatbots (like an airline bot that only responds to button clicks and does not learn),.
2. **Machine Learning (ML):** A subset of AI. These are machines that *learn* and adapt from data. This includes neural networks but also non-neural algorithms like **Decision Trees** (used for medical predictions),.
3. **Deep Learning (DL):** A subset of ML. This specifically refers to **neural networks**. It covers various modalities, such as Convolutional Neural Networks (CNNs) used for image recognition (e.g., identifying a pizza vs. a coffee cup),.
4. **LLMs:** A subset of Deep Learning. These are neural networks designed strictly for **text**.
5. **Generative AI:** Described as a mix of LLMs and Deep Learning. It covers text generation *plus* other modalities like creating images, audio, and video.

6. Major Applications

The lecture outlines five "pillars" of LLM utility:

1. **Content Creation:** Generating new creative text, such as writing a poem about the solar system in the style of a detective story,.
2. **Chatbots:** Powering virtual assistants for banks, airlines, and customer service to replace rule-based systems.
3. **Machine Translation:** Instantly translating text between languages (e.g., English to French) with high accuracy.
4. **Text Generation:** Creating lesson plans, news articles, or reports from scratch,.
5. **Sentiment Analysis:** Analyzing text to detect emotions, such as identifying hate speech on social media or analyzing customer feedback.