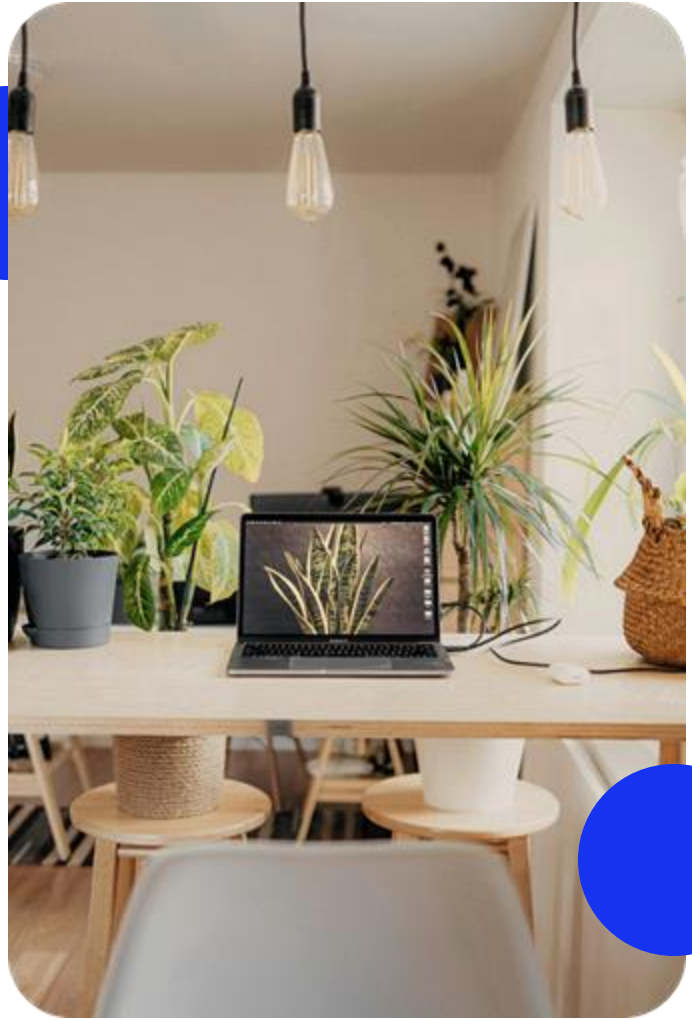
# AI Masterclass

Technical Generative AI Concepts Explained Simply



github.com/Esteban-D-Lopez/TheAIWalkthrough

# Learning Journey Roadmap

| | | |
|---|---|---|
| 01 | Technical Generative AI Foundations | Introduce foundational technical knowledge about AI and large language models (LLMs), laying the groundwork for understanding Generative AI. |
| 02 | Model Optimization Techniques 1 | An overview of key LLM optimization techniques, with a deep dive into Fine-Tuning and Prompt Engineering. |
| 03 | Model Optimization Techniques 2 | An overview of key LLM optimization techniques, with a deep dive into Retrieval Augmented Generation (RAG) and Agentic AI. |
| 04 | Generative AI Monitoring and Evaluation | An overview of key implications and practical considerations of bringing Generative AI products to life safely and efficiently. |

## Goals

- Understand how GenAI technology works
- Feel comfortable exploring with GenAI tools
- Start applying GenAI technology safely and responsibly

# Chapter 1: Technical Generative AI Foundations

# What is Generative AI?

Learning Objective

## Generative AI Defined

A field of **artificial intelligence** that uses **machine learning** techniques, particularly **neural networks**, to create new content by identifying and replicating patterns in data. Popular modern systems are often built using **Large Language Models (LLMs)** that apply **natural language processing (NLP)** to understand prompts and generate human-like text, code, or other media*.

Simply put, Generative AI learns from large sets of data to not just analyze, classify, or predict, but also to generate human-like text, code, and other creative content.

*Image generation tools use **diffusion models**, which learn patterns from millions of images and then generates new ones based on a text prompt.

# Demystifying Generative AI Concepts

## AI

Broad field of intelligent systems

## Machine Learning

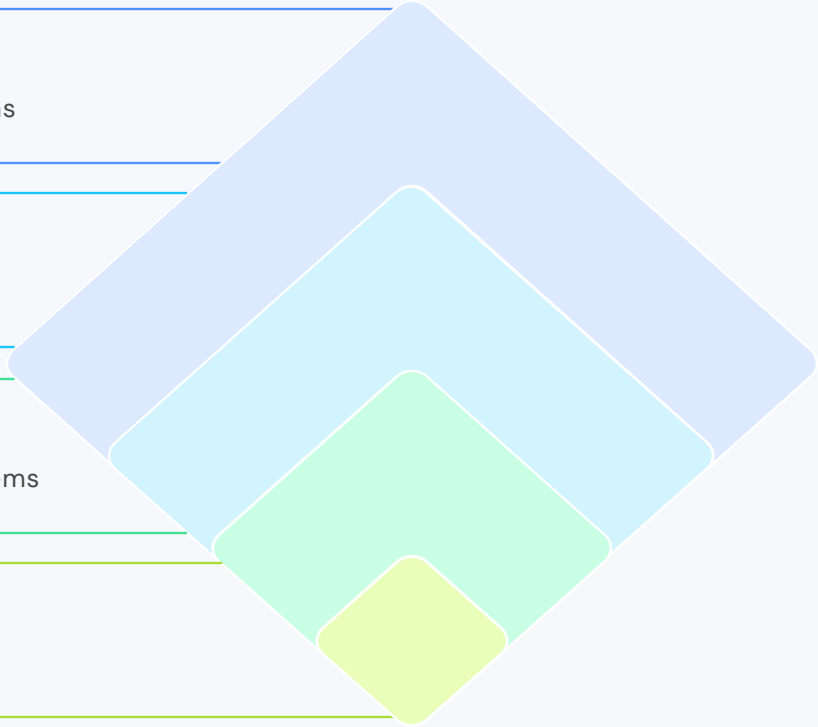
Algorithms learning from data

## Neural Networks

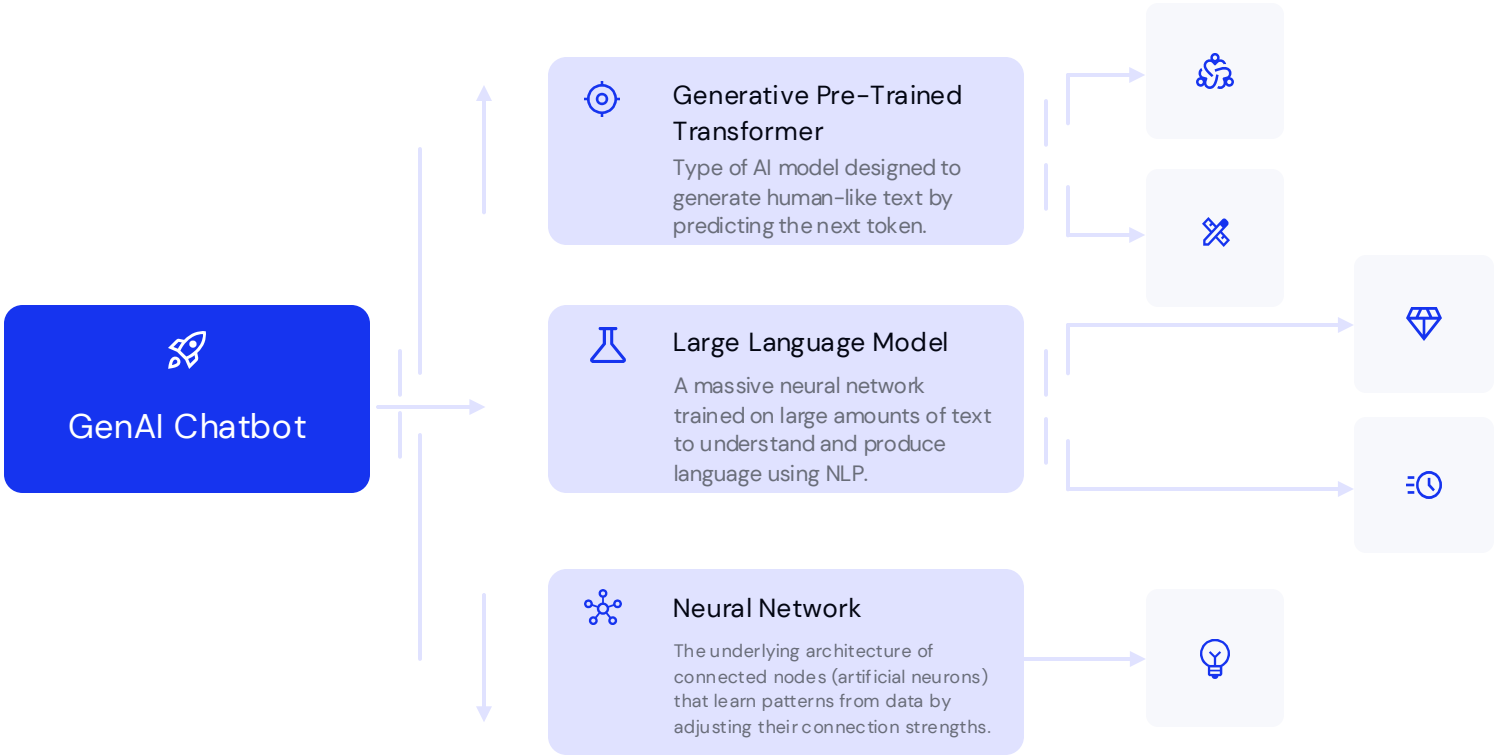
Complex data processing systems

## Generative AI

Branch creating new content

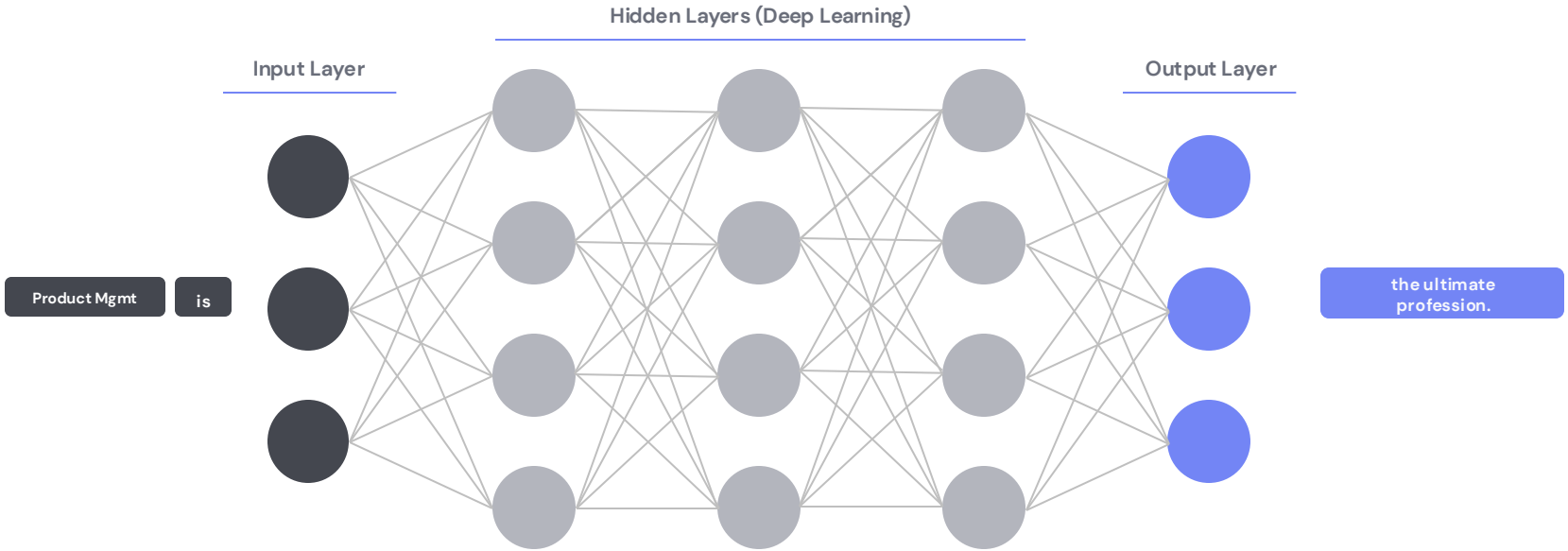


# How do AI Chatbots work?



# Neural Networks

Neural networks are a type of machine learning model that mimic the operations of a human brain to recognize patterns by adjusting the strengths of their connections based on data.




They consist of layers of nodes, or artificial neurons, and layers. Each node connects to others and has its own associated weight and threshold. Have been around since the 1950s.

# Illustrating Neural Network Layers

### Input Layer

*Takes in raw data to initiate the process.* This layer is crucial for feeding the network.

### Hidden Layers

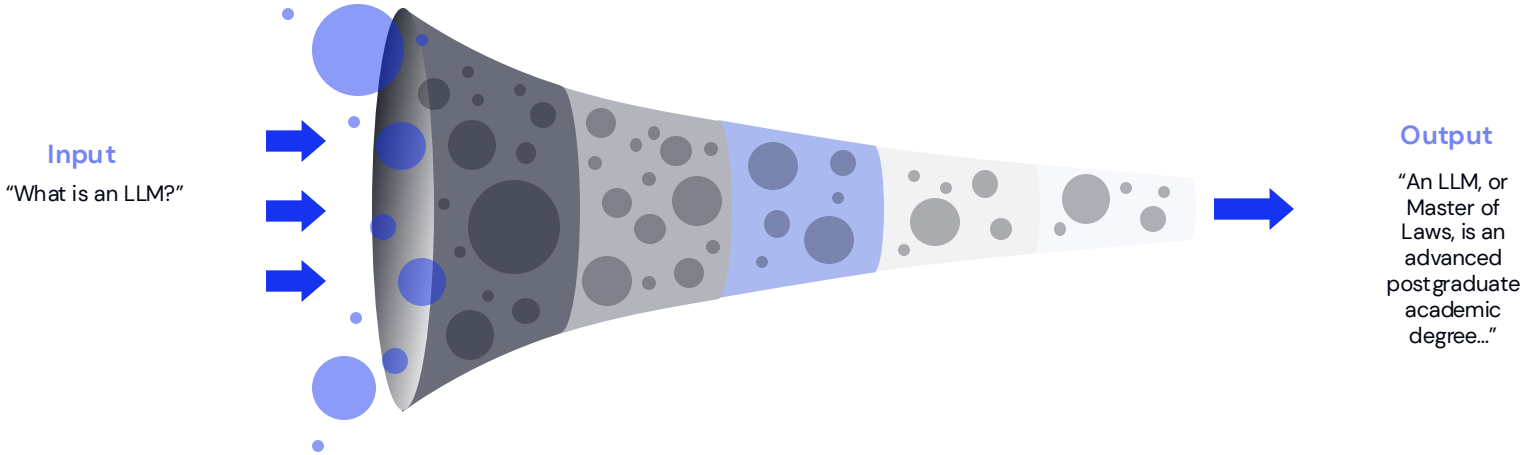
*Processes input data using interconnected neurons.* Learns patterns and derives insights.

### Output Layer

*Delivers the final result or prediction.* Combines outcomes for an actionable output.

# Large Language Models (LLMs)

Are massive neural networks trained on vast amounts of text to understand and generate human-like language by predicting the next token in a sequence.




Although LLMs seem intelligent, they don’t truly understand their outputs. They simply predict likely next tokens based on patterns in their training data.

More on LLMs? Check out: https://www.datacamp.com/blog/what-is-an-llm-a-guide-on-large-language-models

# Visualizing a Large Language Model

01

### Understanding Inputs

**Input Data:** Represents the queries, consisting of text prompts provided by the user.

02

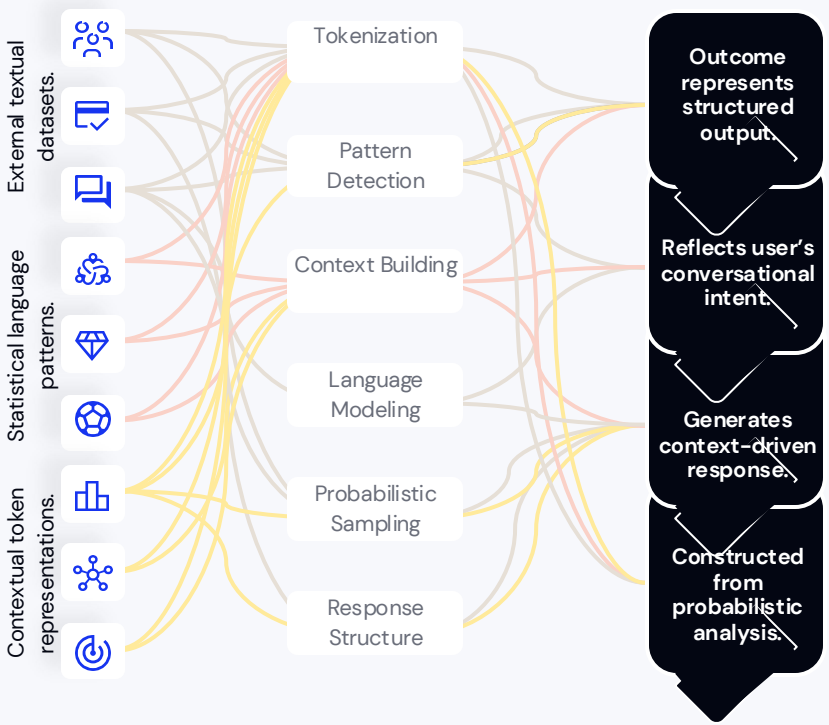
### Processing Complexity

**Processing:** Actions taken by the model to understand and structure input into response.

03

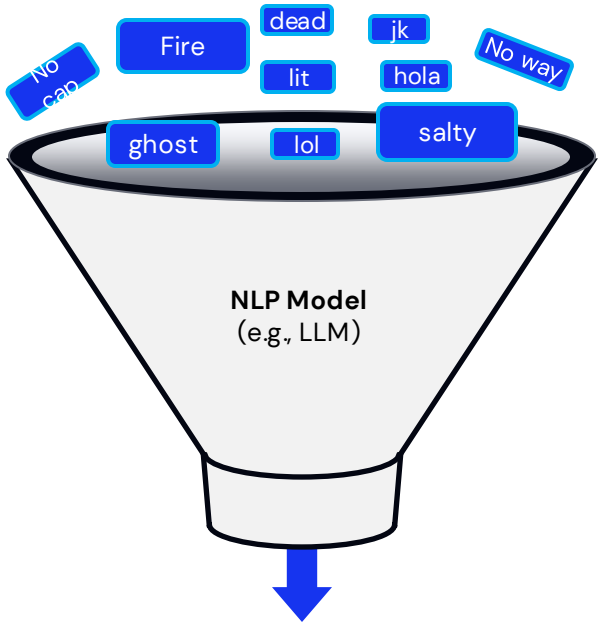
### Outcome Generation

**Output:** Results generated to answer user queries, crafted intelligently via processing logic.



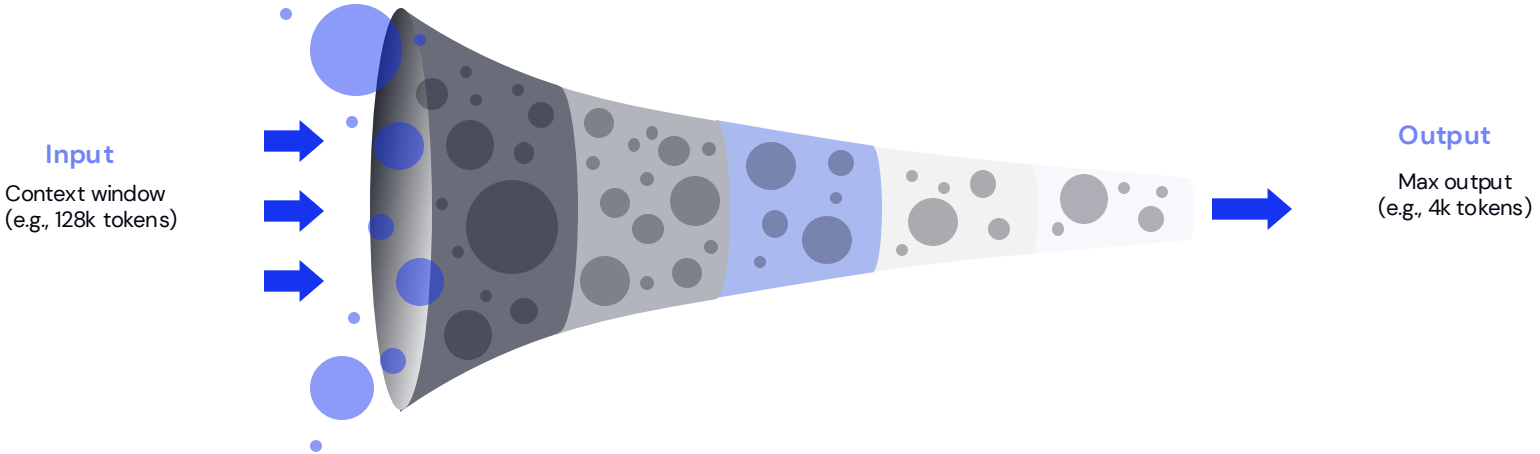
# Natural Language Processing (NLP)

Branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language by combining linguistics, computer science, and machine learning techniques to process text and speech in a way that is meaningful and useful.



# Limited Context

LLMs can only consider a fixed window of tokens at once, and when they lose relevant information outside that window, they may fill the gaps by guessing, often leading to hallucinations.



# Tokenization

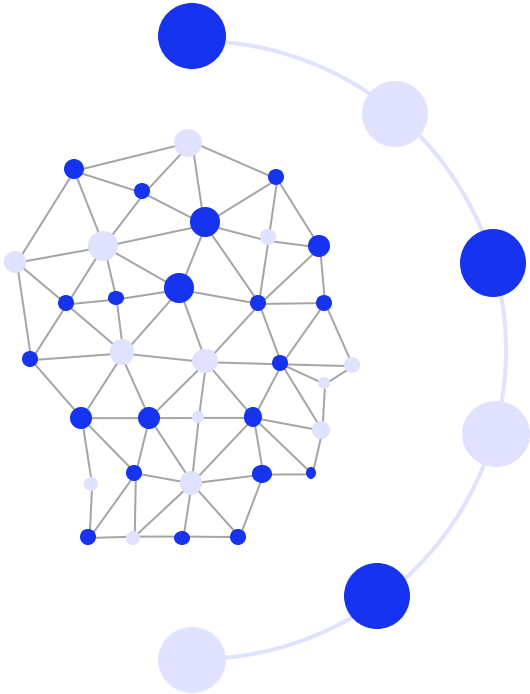
Tokens are the small chunks of text (parts of words, words, or punctuation) that large language models read and generate, and all their input and output is processed as sequences of these tokens.


One token generally corresponds to ~4 characters of text for English text (i.e., ¾ of a word).

https://platform.openai.com/tokenizer

# Transformer Architecture

The Transformer model architecture revolutionized Generative AI by introducing parallel processing and self-attention mechanisms, significantly boosting efficiency and contextual understanding.



## Self-Attention

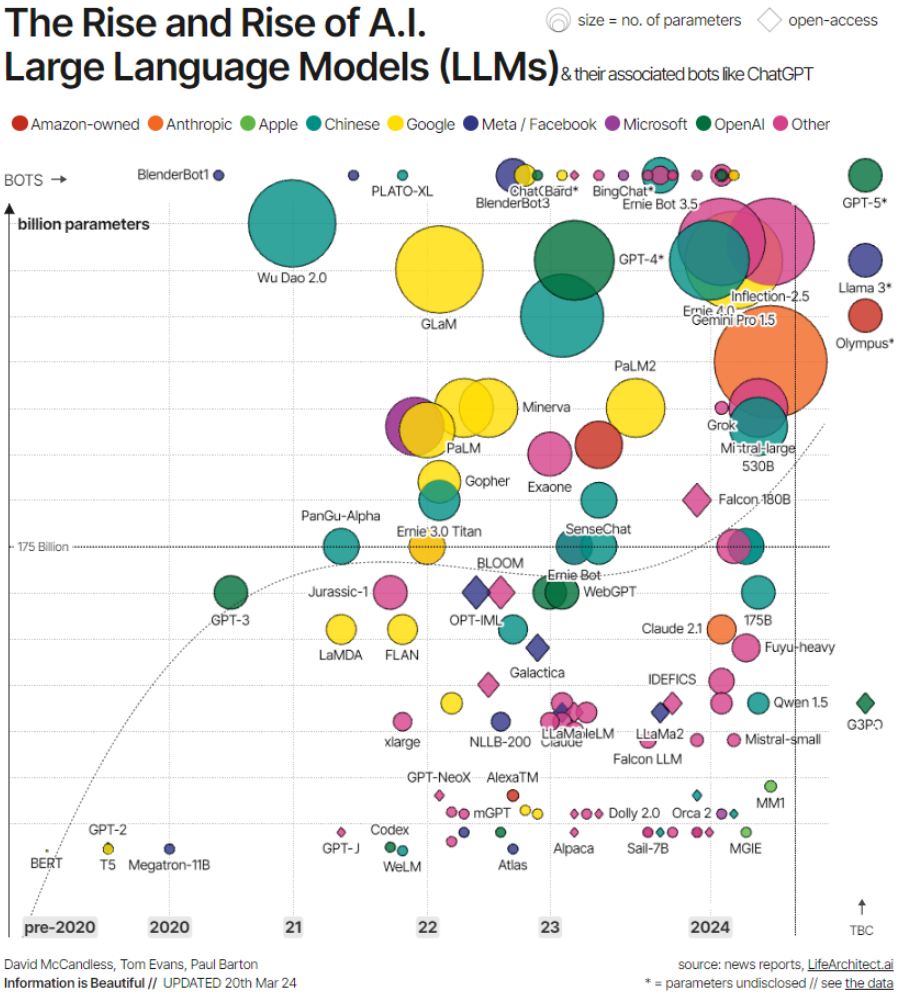
Doesn't treat all words equally, focuses on the meaning of each word and weights its importance relative to others to enhance context understanding.

## Parallel Processing

Reads multiple words simultaneously instead of sequentially to reduce training times and improving the model's ability to learn from vast datasets.

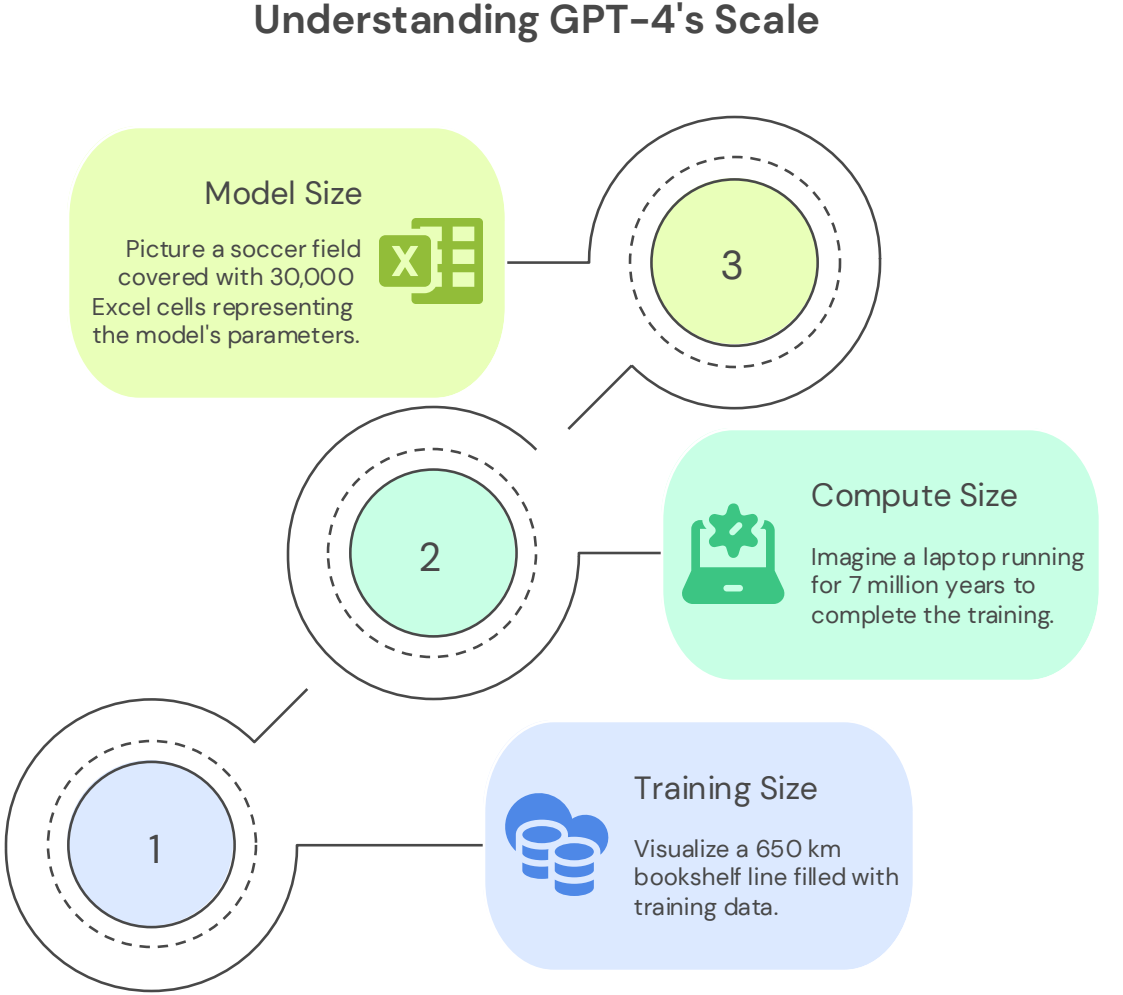
## Scalability

Highly scalable, fueling introduction of LLMs.




# HOW LARGE

are **Large** Language Models?




# HOW LARGE

are **Large** Language Models?

Source: Medium

# AI Energy Consumption

**1%**

**Data Centers**: Consume 1% of global electricity.

**62k**

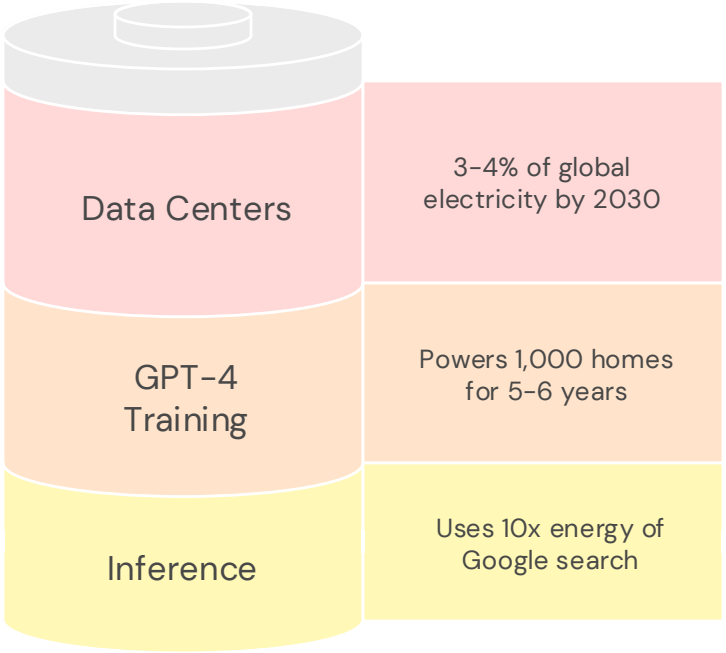
**Training**: GPT-4 consumed 62,000 MWh.

**10x**

**Inference**: Uses 10 times search energy.

**5y**

**Power**: Equal to homes' usage for 5 years.




Sources: Estimated energy metrics from OpenAI, DeepMind, and 2022 IEEE standards.



# Model Optimization

Because large language models are so large, complex, and resource-intensive, it’s crucial to apply optimization techniques to guide them effectively, yielding more accurate, safer results while using resources far more efficiently.

# Strengths and Challenges of Generative AI

| | |
|---|---|
| Efficiency | • **Improved productivity:** Generative AI automates complex tasks, saving time and resources. |
| Versatility | • **Adaptation:** Can be tailored across domains for tasks like content creation or predictive analytics. |
| Opportunities | • **Revolutionization:** Improves resource allocation, expands creativity, and aids decision-making. |
| Ethical Concerns | • **Privacy risks:** Data handling raises questions about security and user privacy. |
| Bias in Outputs | • **Fairness issues:** Outputs can reflect and propagate biases in training data. |
| Environmental Impact | • **Energy consumption:** Training and operation may have significant environmental costs. |