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 **GenAI Optimization Techniques 1**
An overview of key LLM optimization techniques, with a deep dive into Fine-Tuning and Prompt Engineering.

03 **GenAI 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

Presented by

Esteban D. Lopez

Location	New York, NY
Hometown	Quito, Ecuador
Education	University of New Orleans – BS. Accounting Louisiana State University – MS. Accounting Columbia University – FinTech Bootcamp New York University – MS. AI and Machine Learning (Current)
Professional Career	Hedge Funds Assurance Transformation & Economic Consulting AI & Data Technology Consulting AI Product Management

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

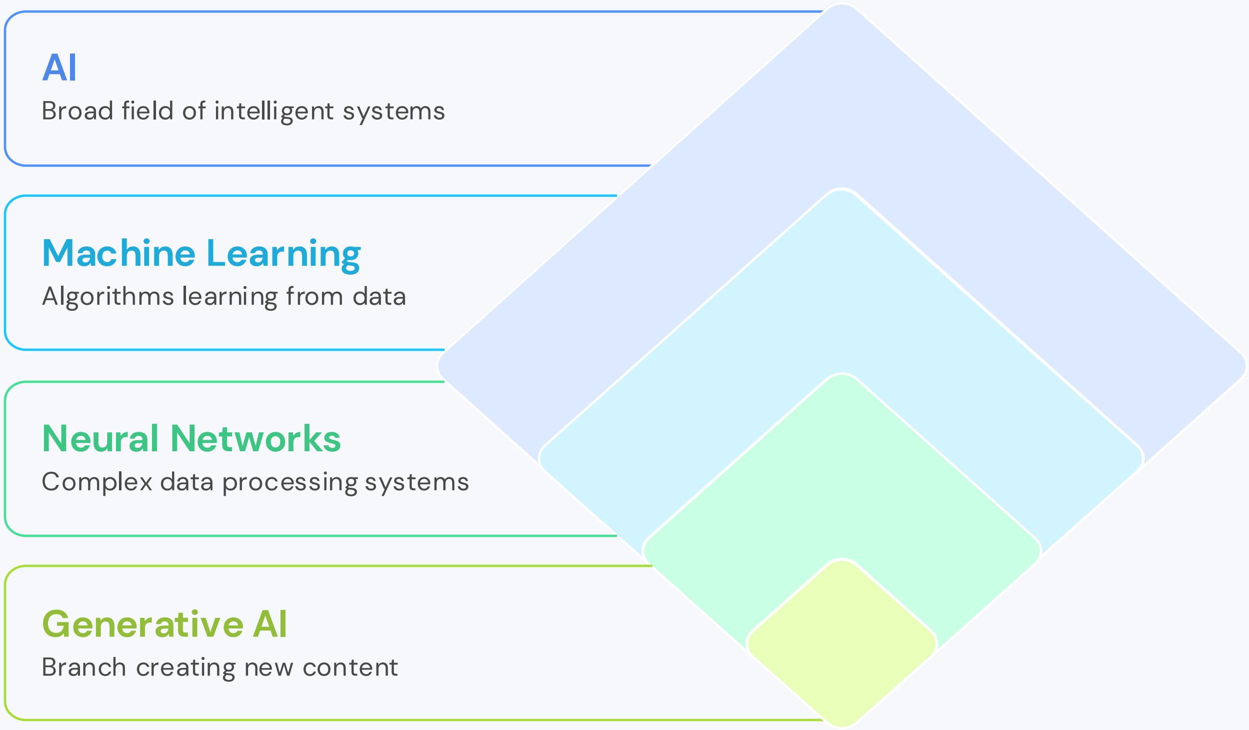
A field of **artificial intelligence** that uses **machine learning** techniques, particularly **neural networks**, to create new content by identifying and replicating patterns in data. Popular modern systems are often built using **Large Language Models (LLMs)** that apply **natural language processing (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.

Technical Foundations

Demystifying Generative AI Concepts



AI
Broad field of intelligent systems

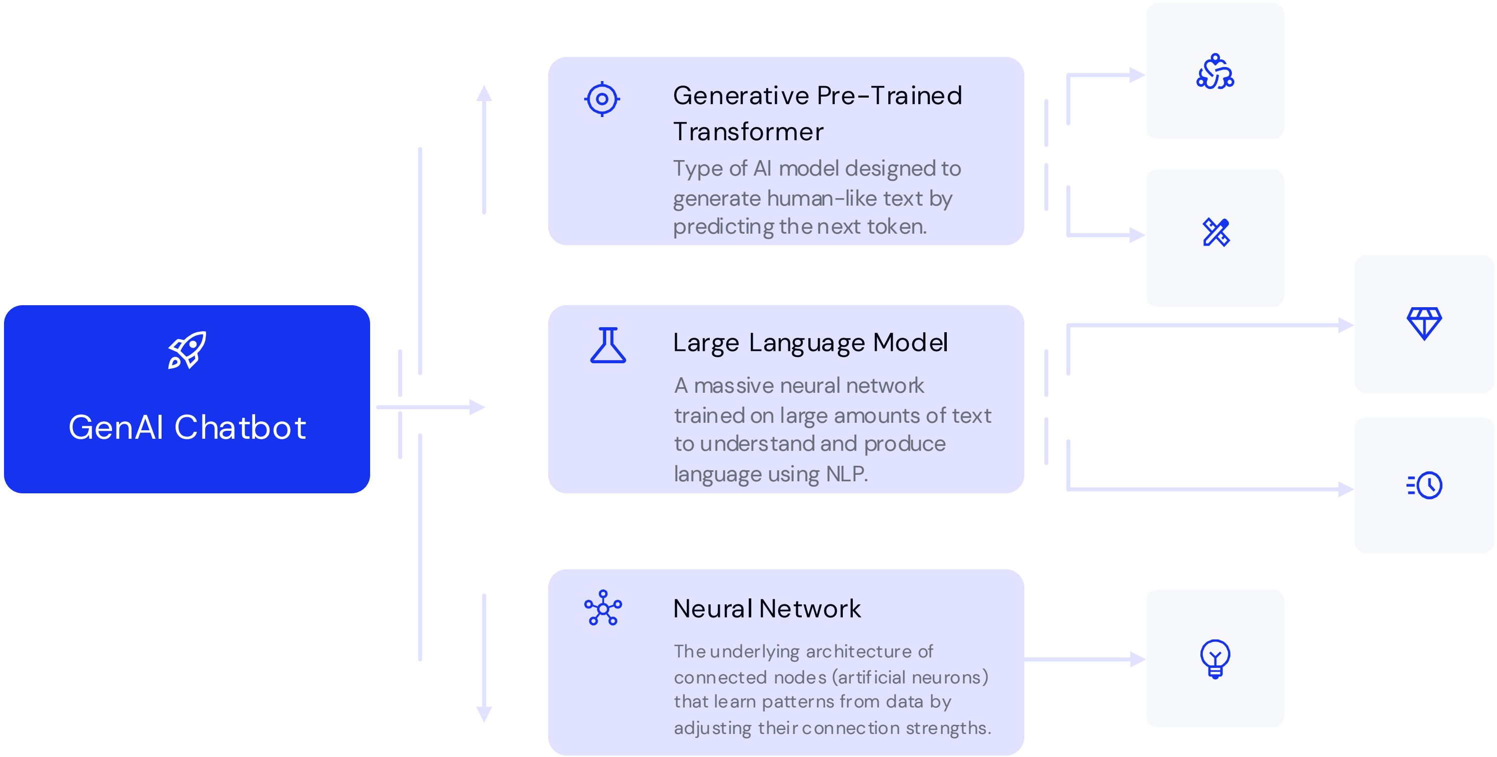
Machine Learning
Algorithms learning from data

Neural Networks
Complex data processing systems

Generative AI
Branch creating new content

Technical Foundations

How do AI Chatbots work?



GenAI Chatbot

Generative Pre-Trained Transformer

Type of AI model designed to generate human-like text by predicting the next token.

Large Language Model

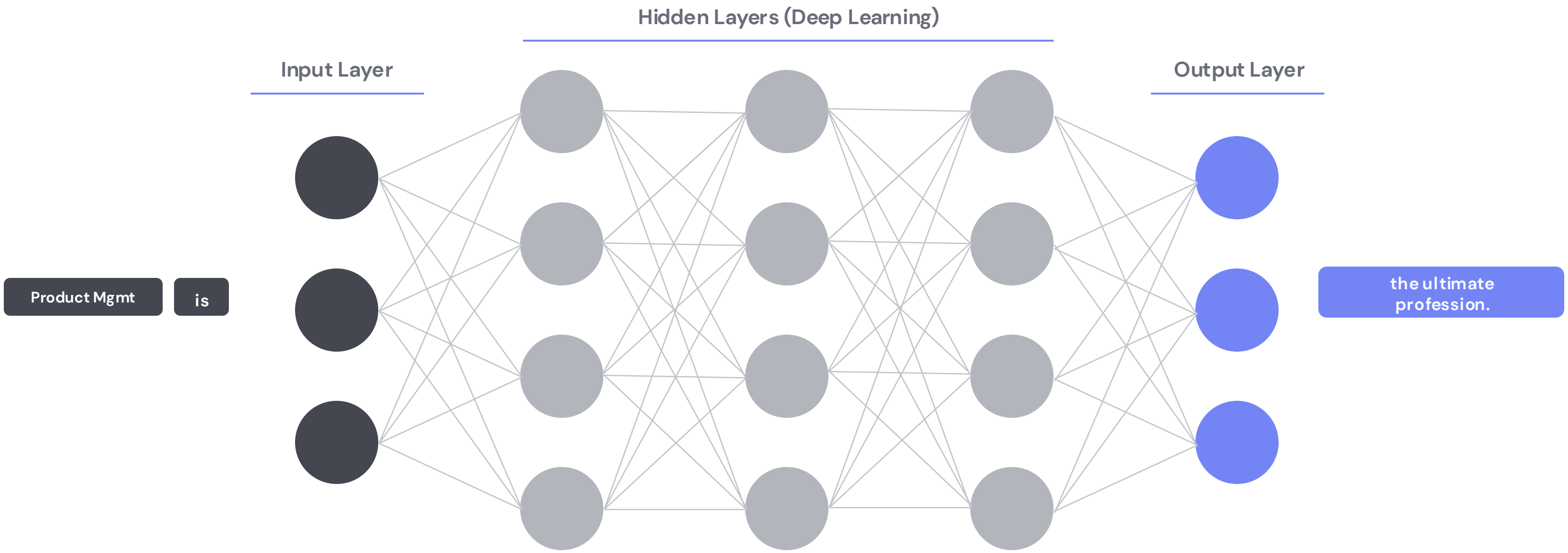
A massive neural network trained on large amounts of text to understand and produce language using NLP.

Neural Network

The underlying architecture of connected nodes (artificial neurons) that learn patterns from data by adjusting their connection strengths.

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.

Technical Foundations

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

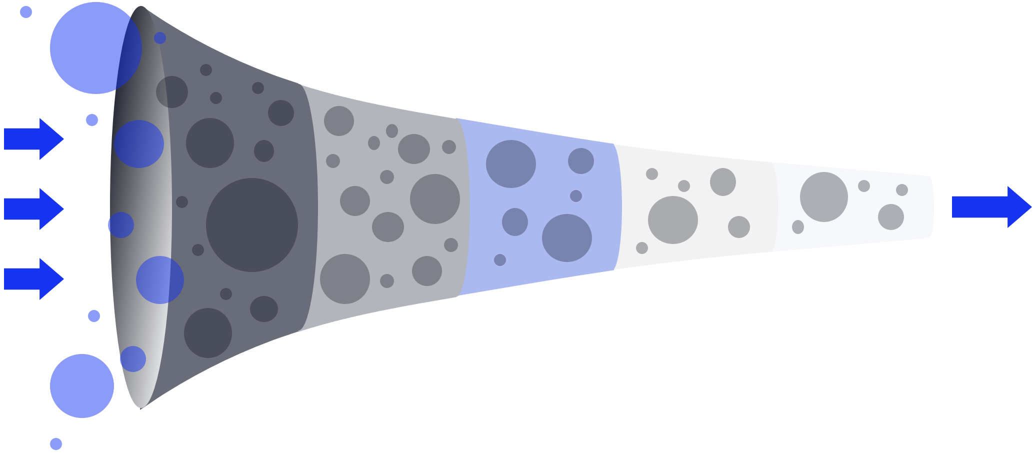
Delivers the final result or prediction. Combines outcomes for an actionable output.

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.

Input

“What is an LLM?”



Output

“An LLM, or Master of Laws, is an advanced postgraduate academic degree…”

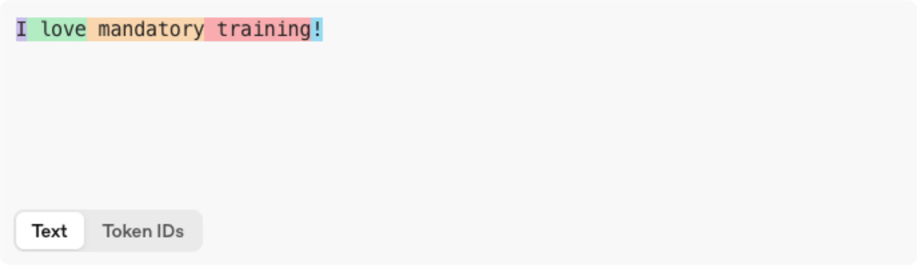
Although LLMs seem intelligent, they don’t truly understand their outputs. They simply predict likely next tokens based on patterns in their training data.

https://www.datacamp.com/blog/what-is-an-llm-a-guide-on-large-language-models

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.

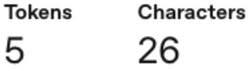
Human-readable Words:



I love mandatory training!

Text | Token IDs

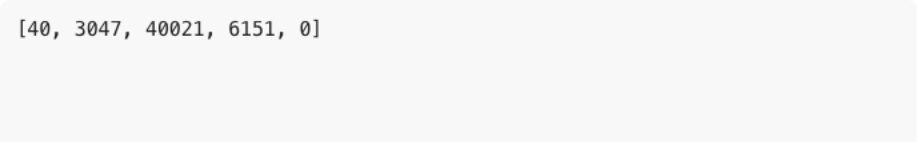
Token Count:



Tokens
5

Characters
26

Token IDs:



```
[40, 3047, 40021, 6151, 0]
```

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

https://platform.openai.com/tokenizer

Large Language Models

Visualizing a Large Language Model

01 Understanding Inputs

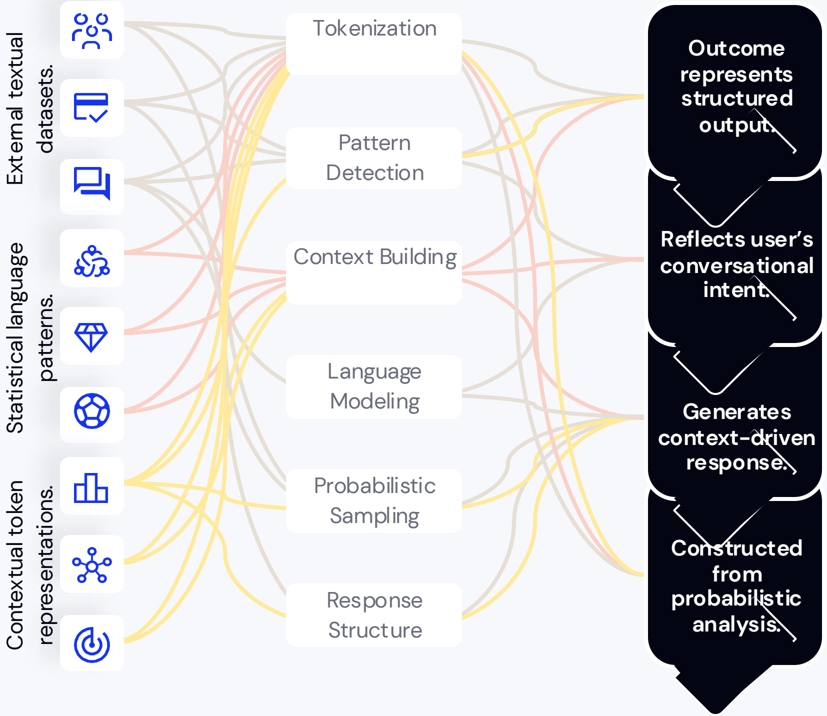
Input Data: Represents the queries, consisting of text prompts provided by the user.

02 Processing Complexity

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

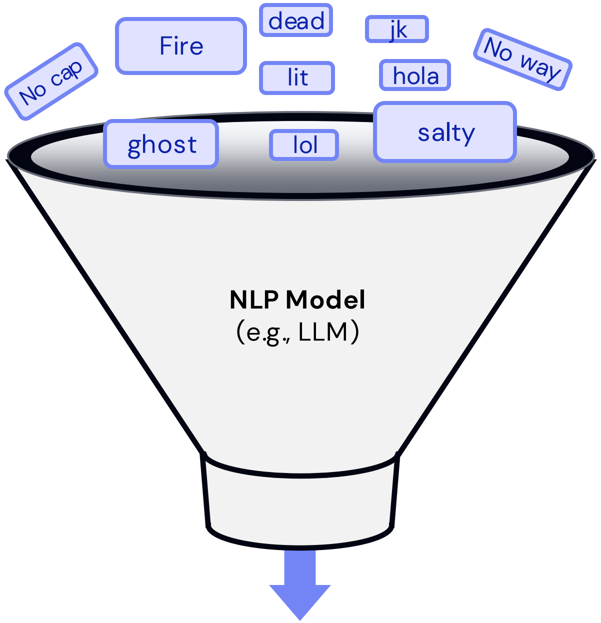
03 Outcome Generation

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


External textual datasets.
Statistical language patterns.
Contextual token representations.
Tokenization
Pattern Detection
Context Building
Language Modeling
Probabilistic Sampling
Response Structure
Outcome represents structured output.
Reflects user's conversational intent.
Generates context-driven response.
Constructed from probabilistic analysis.

Natural Language Processing (NLP)

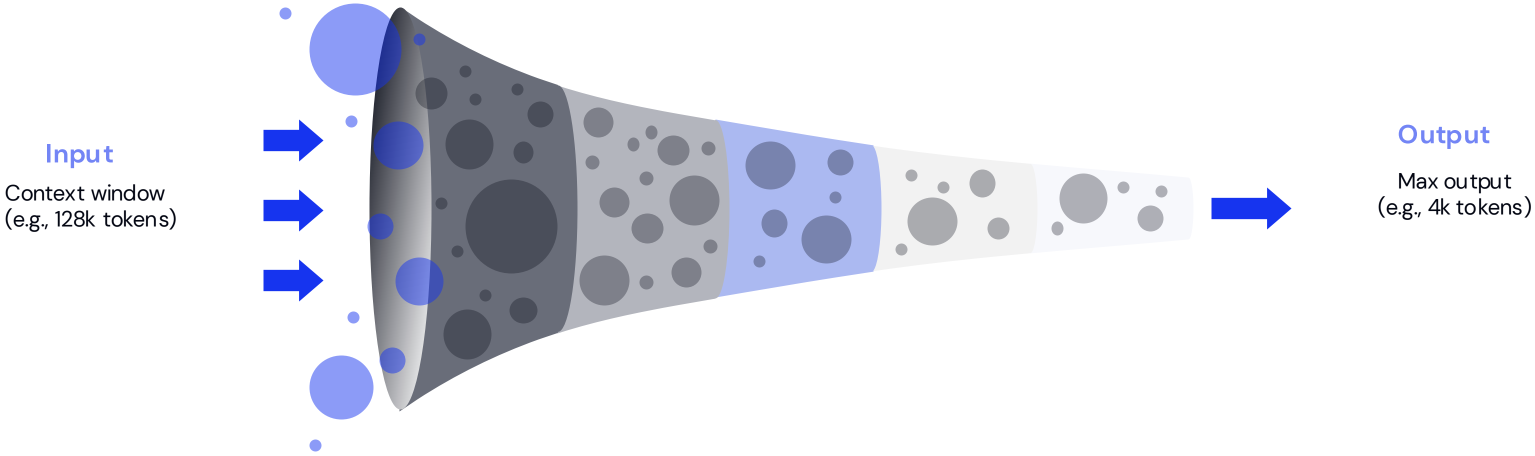
Branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language by combining linguistics, computer science, and machine learning techniques to process text and speech in a way that is meaningful and useful.



Technical Foundations

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.



Input

Context window
(e.g., 128k tokens)

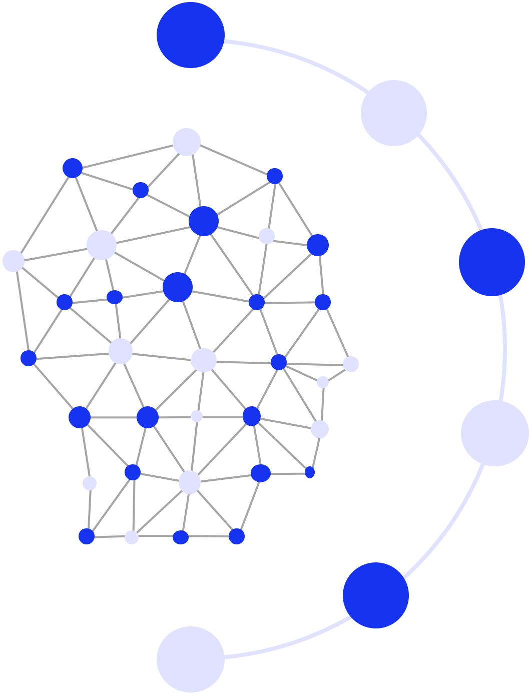
Output

Max output
(e.g., 4k tokens)

Technical Foundations

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

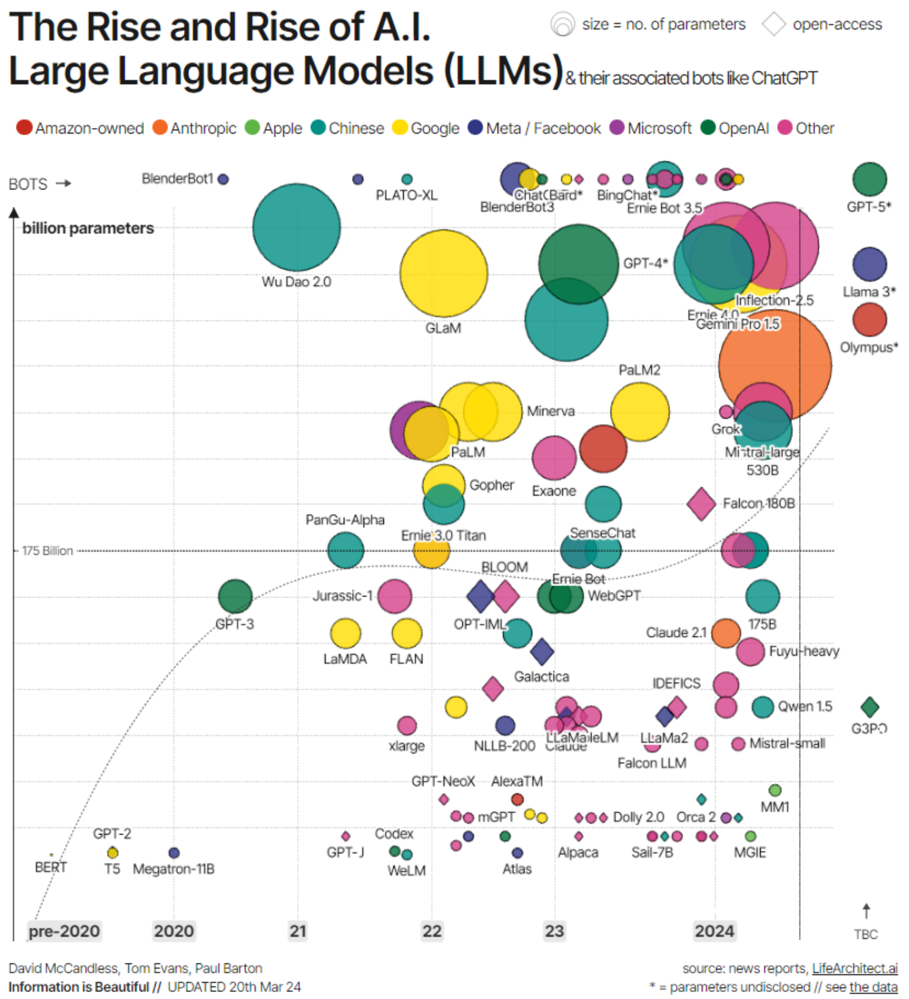
Doesn't treat all words equally, focuses on the meaning of each word and weights its importance relative to others to enhance context understanding.

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.



The Rise and Rise of A.I.
Large Language Models (LLMs) & their associated bots like ChatGPT

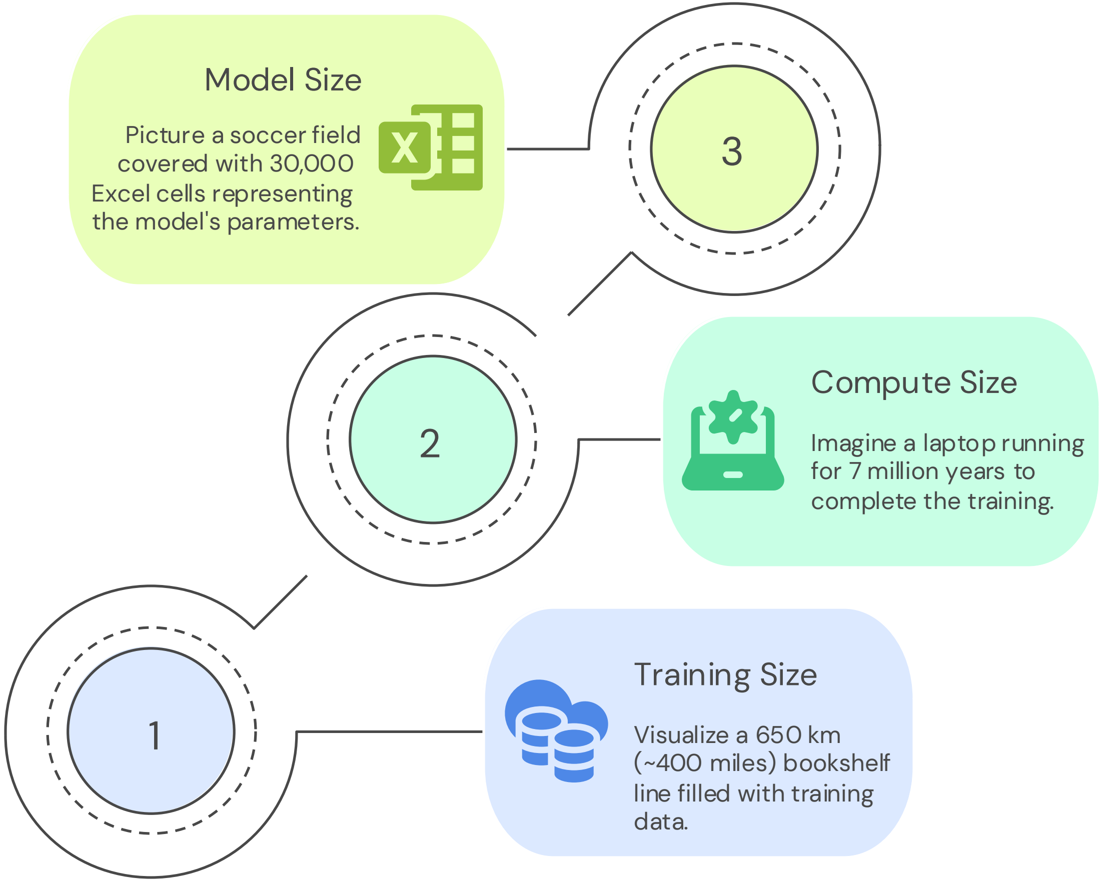
David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 20th Mar 24

source: news reports, LifeArchitect.ai
* = parameters undisclosed // see the data

HOW LARGE

are **Large** Language Models?

Understanding GPT-4's Scale



Model Size

Picture a soccer field covered with 30,000 Excel cells representing the model's parameters.

Compute Size

Imagine a laptop running for 7 million years to complete the training.

Training Size

Visualize a 650 km (~400 miles) bookshelf line filled with training data.

HOW LARGE

are **Large** Language Models?

Source: Medium

AI Energy Consumption

1%

Data Centers: Consume 1% of global electricity.

62k

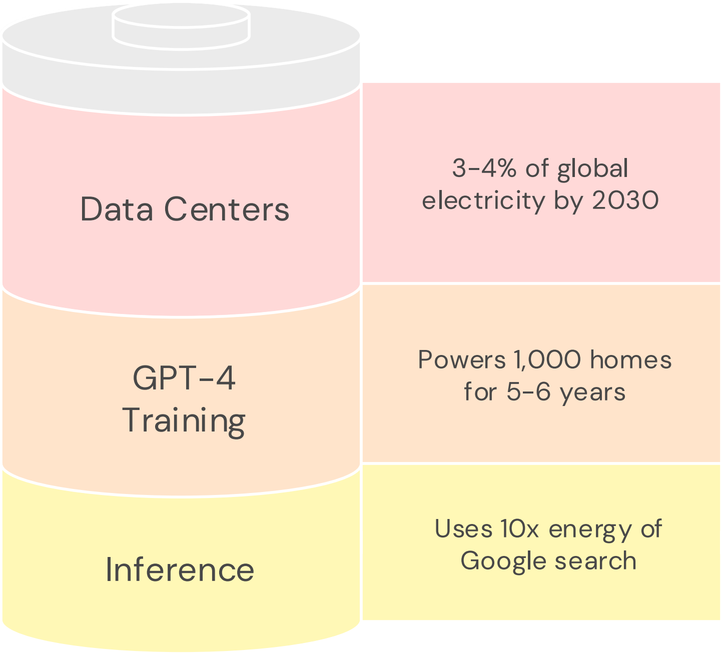
Training: GPT-4 consumed 62,000 MWh.

10x

Inference: Uses 10 times search energy.

5y

Power: Equal to homes' usage for 5 years.



Data Centers

3-4% of global electricity by 2030

GPT-4 Training

Powers 1,000 homes for 5-6 years

Inference

Uses 10x energy of Google search

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.

Technical Foundations

Workflow Evolution

People-Driven Processes

Manual execution with data insights

Technology Adoption

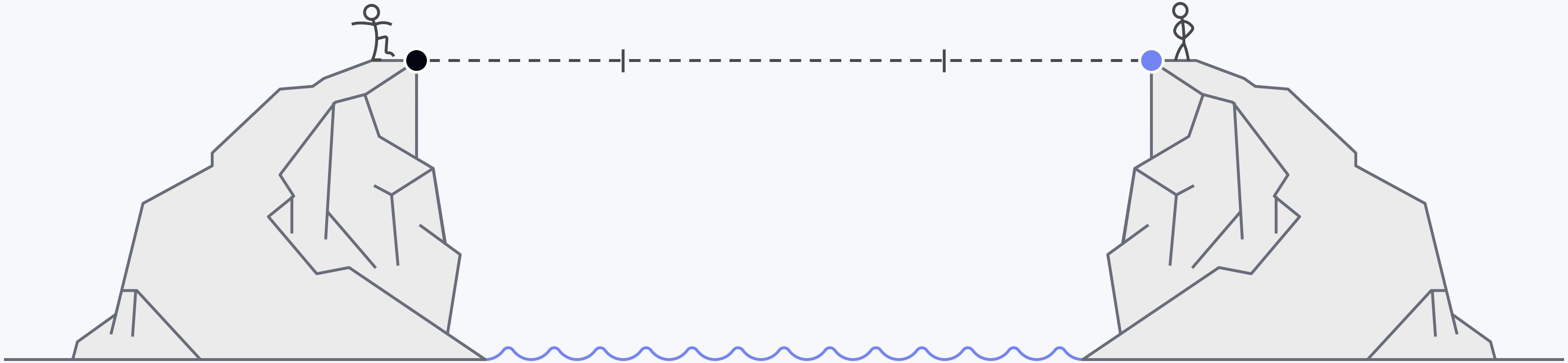
Technology takes over process execution

Data-Driven Execution

Processes driven by data analysis

Technology-Driven Processes

Automated execution with human oversight



Thank you!

Questions?