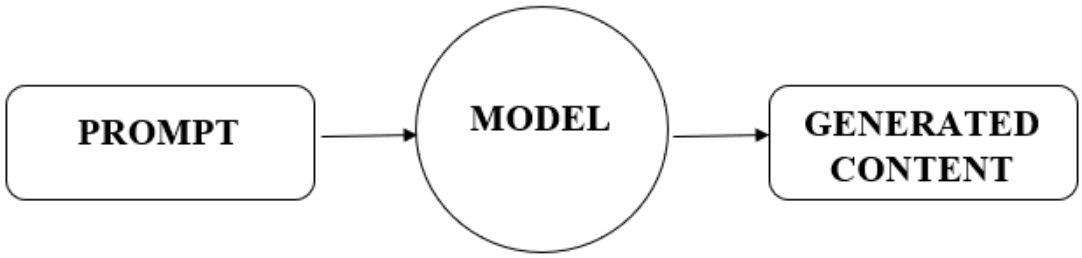
DAY 2

What is Generative AI?

Generative AI is a type of artificial intelligence that can create new content like text, images, music, or code-based on the data it has been trained on. It doesn't just analyze information, it generates original outputs.

Examples: ChatGPT writing an essay, DALL·E creating images, or GitHub Copilot writing code.

Generative AI Architecture



What Are API Models?

An API model is an AI system made available to developers via an Application Programming Interface (API). You don't need to install or train the model; you simply send a request, and the model responds.

Example: Using OpenAI's API to build a chatbot that answers legal questions.

What is a Large Language Model (LLM)?

An LLM is a type of Generative AI trained on vast amounts of text data to generate human-like responses. It can answer questions, translate languages, summarize content, and more-just by predicting the next best word in a sentence.

Think of it as a super-smart autocomplete tool trained on billions of sentences.

Popular LLMs:

- ❖ ChatGPT (OpenAI)
- ❖ Gemini (Google)
- ❖ Claude (Anthropic)
- ❖ LLaMA (Meta)

Base of an LLM

The base of every LLM is built on:

Training Data – The internet, books, articles, etc.

Tokens – Chopped-up pieces of text it understands.

Parameters – Millions to billions of learned weights that help make decisions.

Transformer Architecture – A deep learning structure for language understanding.

Key Terms to Know

TERMS	MEANING
Token	Smallest unit of language, like a word or part of a word.
Parameter	Internal values (like AI "neurons") that learn how to process language.
Prompt	The question or instruction you give the AI.
Fine-Tuning	Teaching an existing model on specific data (e.g., medical or legal).
Inference	The actual output the AI generates.

Analogy: Tokens = letters, parameters = brain cells, inference = final answer.

Tokens are ingredients,parameters are the recipe rules,inference is the dish.

Evolution of LLMs

Year	Model	Parameters	Creator
2018	GPT - 1	117M	OpenAI
2019	BERT	110M	Google
2020	GPT-3	175B	OpenAI
2023	LLaMA,Claude	~70B - ~100B	Meta, Anthropic
2024	GPT-4o,Gemini 1.5	~200B	OpenAI, Google

Inside the LLM: The Transformer Architecture

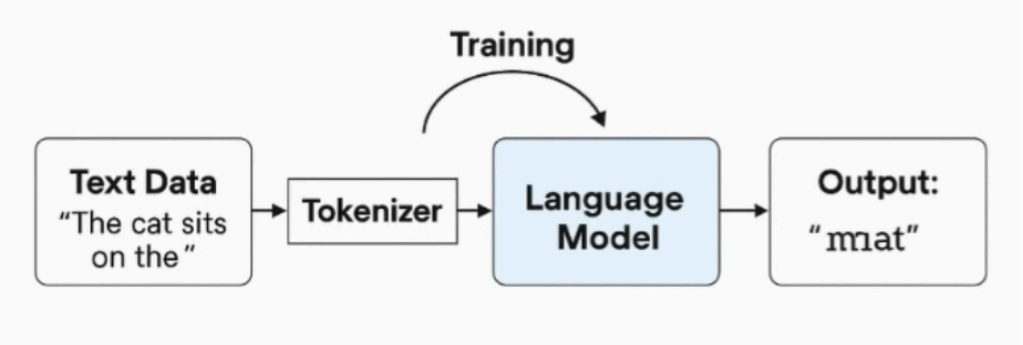
Modern LLMs are built using a **Transformer**, which contains:

Self-Attention – Focuses on which words are most important in a sentence.

Positional Encoding – Understands word order and structure.

Feedforward Networks – Processes and learns deeper meanings.

Think of a Transformer like a super-fast language reader that finds patterns and relationships in words.



Large Language Models

Model	Creator	Description
GPT-3.5 / GPT-4 / GPT-4o	OpenAI	Chat-based models used in ChatGPT; GPT-4o is multimodal.
Claude 1 / 2 / 3	Anthropic	Safe, conversational models with strong reasoning.
Gemini 1.5	Google DeepMind	Successor to Bard; powerful multimodal AI.
LLaMA 2 / LLaMA 3	Meta	Open-source LLMs for research and enterprise.
XGen	Salesforce	Efficient open-source transformer model.
Grok	xAI (Elon Musk)	LLM integrated with X (formerly Twitter).

Code-Specialized Models

Model	Creator	Use
Codex	OpenAI	Powers GitHub Copilot; great for code generation.
StarCoder	BigCode	Open-source code-focused model.
Code LLaMA	Meta	Code generation version of LLaMA.

Multimodal Models (Text + Image + Audio)

Model	Creator	Modalities
GPT-4o	OpenAI	Text, image, audio, video
Gemini 1.5 Pro	Google DeepMind	Text + image + code
Grok Vision	xAI	Image + text
Claude 3 Opus	Anthropic	Handles image + text
DALL·E 3	OpenAI	Image generation from text
Sora (video)	OpenAI	Text-to-video generation
Stable Diffusion	Stability AI	Text-to-image model

How Are LLMs Trained?

Phases of Training:

1. Pretraining – Trained on massive amounts of data to predict words.

2. Fine-Tuning – Adjusted for specific domains or use-cases.

3. RLHF (Reinforcement Learning with Human Feedback) – Trained to give helpful, honest, and safe responses with human feedback.

Example: Like teaching a parrot general words, then training it to have polite conversations.

Applications of LLMs:

LLMs are transforming industries:

- **Customer Support** – 24/7 chatbots
- **Education** – Personalized learning, explanations
- **Healthcare** – Summarizing patient notes
- **Law** – Drafting legal documents
- **Marketing** – Writing ads, captions, blogs
- **Coding** – Auto-completion and debugging (e.g., Copilot)

How to Talk to an LLM – Prompt Engineering

Prompt engineering is the skill of writing effective instructions for AI.Prompt quality really is the deciding factor in how useful and human-like the AI feels.

Prompt Types:

Types	Description	Example

Zero-Shot	Direct request without examples	Translate to Spanish: Good morning.
Few-Shot	Includes examples before asking	English: Hello → Spanish: Hola…
Role-Based	Assigns a role to the model	Act as a travel agent. Recommend a trip to Europe.
Chain-of-Thought	Step-by-step reasoning	Let's solve this step-by-step…
Constraint-Based	Adds rules like word limit or tone	Summarize in 3 bullet points using formal tone.
Reframing	Improves vague prompts	Tell me about biology- Explain DNA in 50 words.

Limitations of LLMs:

Even powerful models have flaws:

- **Hallucinations** – May generate believable but false info.
- **Bias** – Can reflect harmful stereotypes from training data.
- **No true understanding** – They predict, but don't "think."
- **Context window** – Can forget long conversations.
- **Always review outputs** – especially in critical fields like healthcare or law.

Ethical Concerns:

As LLMs become more widespread, they raise important ethical issues:

Misinformation – Can spread false content.

Data Privacy – May memorize sensitive data.

Impersonation – Used for deep fakes or fake profiles.

Responsibility – Who is accountable for AI misuse?

Real case: Microsoft's Tay bot was shut down within 24 hours due to offensive content it learned from users.

The Future of LLMs

The next generation of AI is already emerging:

***Multimodal AI** – Understanding images, audio, and video (e.g., GPT-4o).

***Smaller, faster models** – That work offline or on phones.

***AI Agents** – Autonomous bots that take actions (e.g., Devin).

***Personal AI Assistants** – Customized memory, behavior, and voice.

The future is personal, multimodal, faster, and everywhere.