What is AI (Artificial Intelligence)?

Artificial Intelligence (AI) is the field of computer science that builds systems that can perform tasks that normally require human intelligence.

□ In simple words:

AI is when machines can **think, learn, and make decisions** like humans (to some extent).

□ What AI can do:

- Understand speech (like Siri, Alexa)
- Recognize images (face unlock on phones)
- Translate languages (Google Translate)
- Recommend videos (YouTube, Netflix)
- Chat and answer questions (chatbots like me)

□ What are AI Models?

□ Simple meaning:

An **AI model** is the "brain" or "trained system" inside AI that has learned patterns from data.

□ Think of AI as a *student*, and the AI model as the *trained knowledge the student has learned*.

□ Easy analogy:

Imagine teaching a child:

1. You show many apples □ and bananas □
2. The child learns the difference
3. Now the child can identify new fruits

□ The **trained knowledge in the child's brain = AI model**

□ How an AI model is created

Step 1: Data collection □

- Gather lots of examples (images, text, numbers)

Step 2: Training □

- The model learns patterns from data
- Example: what makes a cat a cat □

Step 3: Testing □

- Check if it makes correct predictions

Step 4: Prediction □

- Use the model to solve real problems

□ Examples of AI Models

1. Image recognition model □

- Input: image
- Output: “cat”, “dog”, “car”

2. Language model □

- Input: text
- Output: answers or generated text
  (Example: ChatGPT)

3. Recommendation model □

- Input: user behavior
- Output: suggested movies/videos

4. Prediction model □

- Input: past data
- Output: future prediction (like house prices)

□ Simple way to remember:

- **AI** = the whole intelligence system
- **AI Model** = the trained brain inside AI
- **Data** = what the model learns from

What is a Language Model?

A **language model** is an AI system that is trained to **understand, predict, and generate human language** (text or speech).

□ Simple definition:

A language model is a system that learns patterns in language and can:

- predict the next word
- complete sentences
- answer questions
- generate new text

Easy analogy:

Think of typing on your phone:

When you type:

“I am going to the …”

Your phone suggests:

“market”, “school”, “office”

□ That prediction is done by a **language model**.

It has learned from millions of sentences what words usually come next.

How a Language Model works (simple idea):-

1. It reads huge amounts of text

- Books
- Articles
- Websites
- Conversations

2. It learns patterns

- Which words often come together
- How sentences are structured
- Meaning of words in context

3. It predicts text ✍

- Given a sentence, it guesses what comes next

Here are some **popular open-source (or open-weight) Large Language Models (LLMs)** you can explore:

□ □ Open-source / open-weight LLMs

□ 1. LLaMA (Meta)

- Developed by **Meta**
- Very popular foundation model family
- Versions: LLaMA 2, LLaMA 3
- Strong performance for chat and reasoning

□ 2. Mistral AI models

- **Mistral 7B**
- **Mixtral (Mixture of Experts model)**
- Known for being **fast + efficient**

□ 3. Falcon (TII - UAE)

- Models like Falcon 7B, 40B
- Good for research and general tasks

□ 4. BLOOM (BigScience project)

- Truly open collaborative model
- Supports many languages
- Trained by global researchers

□ 5. Gemma (Google)

- Lightweight models from Google
- Based on Gemini research
- Good for developers and experiments

□ 6. Phi models (Microsoft)

- Phi-2, Phi-3
- Small but surprisingly powerful
- Designed for efficiency

□□ 7. Qwen (Alibaba)

- Strong multilingual performance
- Good at coding and reasoning
- Popular in open-source community

Here's a **simple beginner guide to use Gemma 2B with Ollama** on your laptop □

What is what?

- **Ollama** → a tool that lets you run AI models locally (offline)
- **Gemma 2B** → a small, fast AI language model by Google
- Together → you get a **local chatbot on your laptop**

Step-by-step: Run Gemma 2B with Ollama

□ 1. Install Ollama

□ Windows / Mac / Linux:

Go to:
□ https://ollama.com

Download and install it like a normal app.

□ 2. Open Terminal / Command Prompt

- Windows: Command Prompt or PowerShell
- Mac: Terminal
- Linux: Terminal

□ 3. Run Gemma 2B model

Type this command:

```
ollama run gemma:2b
```

Then press Enter.

□ What happens next?

- Ollama will **download the model (first time only)**
- Then it will open a **chat interface in your terminal**

You can now talk to it like:

```
You: What is AI?
Model: AI is...
```

🧪 Example usage

```
ollama run gemma:2b
```

Then:

🧑 You:

Explain machine learning simply

🤖 AI:

Machine learning is when computers learn from data...

Useful Ollama commands

▶ Run model

```
ollama run gemma:2b
```

📋 List installed models

```
ollama list
```

⬇ Download model only

```
ollama pull gemma:2b
```

A **Large Language Model (LLM)** is an AI system that learns patterns from huge amounts of text and then uses those patterns to **predict and generate language**. The key idea is simple:

An LLM does not "think" like a human — it **predicts the next word repeatedly** in a very smart way.

Below is a **step-by-step detailed but beginner-friendly explanation** of how it works.

---

1. Big Idea of an LLM

An LLM works like this:

Given some text → it predicts the **most likely next word** → then repeats this again and again

Example:

Input:

"The sky is"

Model predicts:

"blue"

Then:

"The sky is blue and …"

It continues word by word.

---

2. Step 1: Training (Learning from Data)

□ What happens?

The model reads **massive amounts of text**:

- Books □
- Websites □
- Articles □
- Code □

- Conversations

This can be **trillions of words**.

What it learns:

It does NOT memorize sentences. Instead, it learns:

- Which words often appear together
- Grammar patterns
- Meaning from context
- Relationships between words

Simple analogy:

Like reading millions of books and learning:

“In most cases, after ‘I am going to’, people say ‘school’, ‘work’, etc.”

---

3. Step 2: Tokenization (Breaking text)

Before processing, text is broken into small pieces called:

Tokens

Example:

Sentence:

“I love AI”

Tokens might be:

```
["I", "love", "AI"]
```

Sometimes words are split:

```
“unbelievable” → ["un", "believ", "able"]
```

---

4. Step 3: Understanding meaning (Embeddings)

Each token is converted into numbers called:

Vectors (embeddings)

These numbers represent meaning.

Example idea:

- “king” and “queen” → similar vectors
- “cat” and “dog” → close in meaning space

---

5. Step 4: Neural Network Processing (Transformer)

Modern LLMs use a special architecture called:

Transformer

This is the “brain” of the model.

What Transformer does:

It figures out:

- Which words are important
- Context of the sentence
- Relationships between words far apart

Example:

Sentence:

“The cat that I saw on the street was hungry”

The model understands:

- “cat” is related to “was hungry”

- even though words are far apart

---

6. Step 5: Predicting the next word

Now the model calculates probabilities:

Input:

“The sky is”

Output probabilities:

- blue → 70%
- cloudy → 20%
- dark → 5%
- pizza → 0.001%

It picks the most likely word.

---

7. Step 6: Repeating process (generation)

After choosing a word, it repeats:

“The sky is blue”

Then:

“The sky is blue and”

Then:

“The sky is blue and clear”

This continues until it finishes the response.

---

8. Why LLMs are powerful

Because they:

- Learn from **huge data**
- Understand **context**
- Can generate **human-like responses**
- Adapt to many tasks:
    - writing
    - coding
    - translation
    - reasoning (to some extent)