Lecture 01: Introduction to GenAI

1: 3100, 3200, 3300, 3400, ?

2: Sun Rises from the ____?

3: Twinkle twinkle little ___?

3: Modi is prime minister of ___?

4: President of USA is ___?

5: roses are red, violets are ___?

6: अक्ल बड़ी या बयस or अक़्ल बड़ी या भैंस

7: धोबी का कुतका, न घर का न घाट का or धोबी का कुत्ता, न घर का न घाट का

8: Take unusual pattern: 17, 360

It do prediction

it can’t calculate (like human)

it can’t run the code

it can’t search on the internet

It can’t do reasoning

What is Generative AI?

Core Idea: AI that predicts the next word/token based on patterns learned from massive data.

Key Point: It does NOT think, understand, or reason. It only predicts.

Experience It Yourself

Try completing these:

1. "100, 200, 300, ___" → You said **400** (instant!)
2. "Twinkle twinkle little ___" → You said **star** (no thinking!)
3. "Roses are red, violets are ___" → You said **blue** (but violets are purple!)

What happened?

- You predicted based on patterns you've seen many times
- You didn't think, you just knew
- Sometimes frequency beats facts (violets are purple, not blue)

This is EXACTLY how LLMs work!

The Two Modes

Humans Have Two Modes:

Mode 1: Pattern Recognition (Fast)

- "100, 200, 300, ___" → 400 (instant)
- "The Sun rises in the ___" → east (automatic)

Mode 2: Reasoning (Slow)

- "37, 38, 42, 51, 67, ___" → (you think: differences are 1, 4, 9, 16... next is 25... answer: 92)

LLMs Only Have Mode 1:

- Fast pattern matching
- NO actual reasoning
- Can simulate reasoning if they saw similar patterns in training

---

1. What is Generation?

Simple explanation:
When you write an essay, you write one word at a time:

- "Climate"
- "Climate change"
- "Climate change is"
- "Climate change is a"
- "Climate change is a serious"

How do you decide the next word? Based on patterns you've learned from reading.

LLMs do the same:

- Predict one token at a time
- Build sequences through prediction
- Based on patterns from training data

- **LLM (Large Language Model):**
  - *Input:* Text Tokens.
  - *Internal Math:* Calculates relationships between words.
  - *Output:* Text Tokens.
  - *Principle:* It predicts the next **word**.
- **LMM (Large Multimodal Model):**
  - *Input:* Text, Images, Audio, Video.
  - *Internal Math:* Calculates relationships between *concepts* (regardless of format).
  - *Output:* Text, Images, Audio, Video.
  - *Principle:* It predicts the next **piece of data** (whether that data is a word, a pixel, or a sound wave).

---

2. Tokens, Not Words

Problem: Computers only understand numbers, not words.

Solution: Break text into tokens (smaller pieces).

Examples:

```
"I like cats" → ["I", " like", " cats"]

"I don't like pineapple" → ["I", " don", "'t", " like", " pine", "apple"]

"I love dosa" → ["I", " love", " d", "osa"]
```

Why It Matters:

Q: How many letters in "strawberry"?
You: 10 letters (s-t-r-a-w-b-e-r-r-y)

ChatGPT often gets it wrong! Why?

- It sees tokens like ["straw", "berry"], not individual letters
- Can't count letters because it never saw them separately

This explains:

- ❌ Can't reverse words letter-by-letter
- ❌ Can't count specific letters
- ❌ Struggles with spelling

---

3. Training vs Inference

Training Phase (Learning):

- Feed billions of text examples
- For each: hide next word, make model predict
- When wrong, adjust internal numbers slightly

- Repeat billions of times
- **Takes:** Months, millions of dollars
- **Done:** Once

Inference Phase (Using):

- Use learned patterns (frozen, no new learning)
- Just predict based on what was learned
- **Takes:** Milliseconds
- **Done:** Every time you chat

Key Point: The ChatGPT you use is NOT learning from you. It's using patterns learned months ago.

---

4. Context Window (Memory Limit)

What is it?
Everything the AI can "see" right now:

- Your current message
- Previous messages
- Uploaded files

The Limit:

- 4K tokens ≈ 3,000 words (short chat)
- 32K tokens ≈ 24,000 words (small book chapter)
- 200K tokens ≈ 150,000 words (entire novel)

What happens when you exceed?

```
Window size: 10 tokens
Your input: 15 tokens
[1][2][3][4][5][6][7][8][9][10][11][12][13][14][15]
```

```
AI sees only: [6][7][8][9][10][11][12][13][14][15]
              ↑ First 5 dropped!
```

This explains:

- Why it forgets early parts of long conversations
- Why you sometimes need to repeat information
- Why new chat = complete fresh start

---

5. Temperature (Randomness Control)

Problem: "The capital of France is ___"

Probabilities:

- "Paris" → 98%
- "Lyon" → 1%
- "London" → 0.01%

Should it ALWAYS pick "Paris"? Then every answer is identical.

Temperature Settings:

Temperature = 0 (No randomness):

- Always picks highest probability
- Same input → same output every time
- **Use for:** Math, facts, code

Temperature = 0.7 (Medium):

- Usually high probability, sometimes lower
- **Use for:** General writing, explanations

Temperature = 1.5 (High):

- More random, creative
- Can be nonsensical

- **Use for:** Creative writing, brainstorming

Common Myths About LLMs

❌ Myth 1: LLMs search the internet

Truth: They don't connect to internet. Only use patterns from training (frozen months ago).

❌ Myth 2: LLMs think/understand

Truth: No thinking. Pure mathematical prediction, token by token.

❌ Myth 3: LLMs calculate math

Truth: They PREDICT digits, not calculate.

- 2+2 = correct (seen millions of times)
- 8437×6829 = often wrong (predicting digits)

❌ Myth 4: LLMs remember forever

Truth: Only remember within context window. New chat = forgot everything.

❌ Myth 5: LLMs learn from your corrections

Truth: They use corrections in current chat only. Not updating their knowledge.

❌ Myth 6: LLM-generated code always works

Truth: Often has bugs. Always test. Never blindly trust.

Who Uses LLMs? (Real Problems Solved)

1. GitHub Copilot

Problem: Developers spend 60% time on repetitive code
Solution: AI suggests code as you type
Impact: 55% faster coding

2. Duolingo

Problem: Can't personalize for millions of students
Solution: AI tutor adjusts to YOUR level
Impact: Personalized learning at scale

3. Intercom (Customer Support)

Problem: 70% of tickets are repetitive questions
Solution: AI chatbot answers instantly
Impact: 50% queries resolved by AI, 24/7 availability

4. Notion AI

Problem: Hours wasted on documentation
Solution: AI generates summaries, action items
Impact: Save 2-3 hours/week

5. Khan Academy

Problem: Teachers can't give 1-on-1 attention
Solution: AI tutor (Khanmigo) guides through questions
Impact: Every student gets personal tutor

6. Grammarly

Problem: Poor writing hurts professionalism
Solution: AI fixes grammar, improves clarity
Impact: Better communication, saves time

7. Harvey AI (Legal)

Problem: Lawyers spend 60% time reading documents
Solution: AI summarizes cases, finds precedents
Impact: 10 hours work → 1 hour

How It All Works Together

When you type: "Write a function to add two numbers"

1. **Tokenization:** Text → tokens → numbers

2. **Context:** Your message fits in context window
3. **Prediction:** Model calculates probabilities for next token
4. **Temperature:** Picks based on randomness setting
5. **Generation:** One token at a time:

> functionfunction addfunction add(function add(afunction add(a,function add(a, bfunction add(a, b) {...

No thinking. Just prediction after prediction.

Key Takeaways

1. ✅ LLMs predict patterns, don't think
2. ✅ Work on tokens, not words/letters
3. ✅ Training = learning (past), Inference = using (now)
4. ✅ Context window = temporary memory
5. ✅ Temperature = randomness control
6. ✅ Not perfect - always verify outputs
7. ✅ Used everywhere - transforming every industry

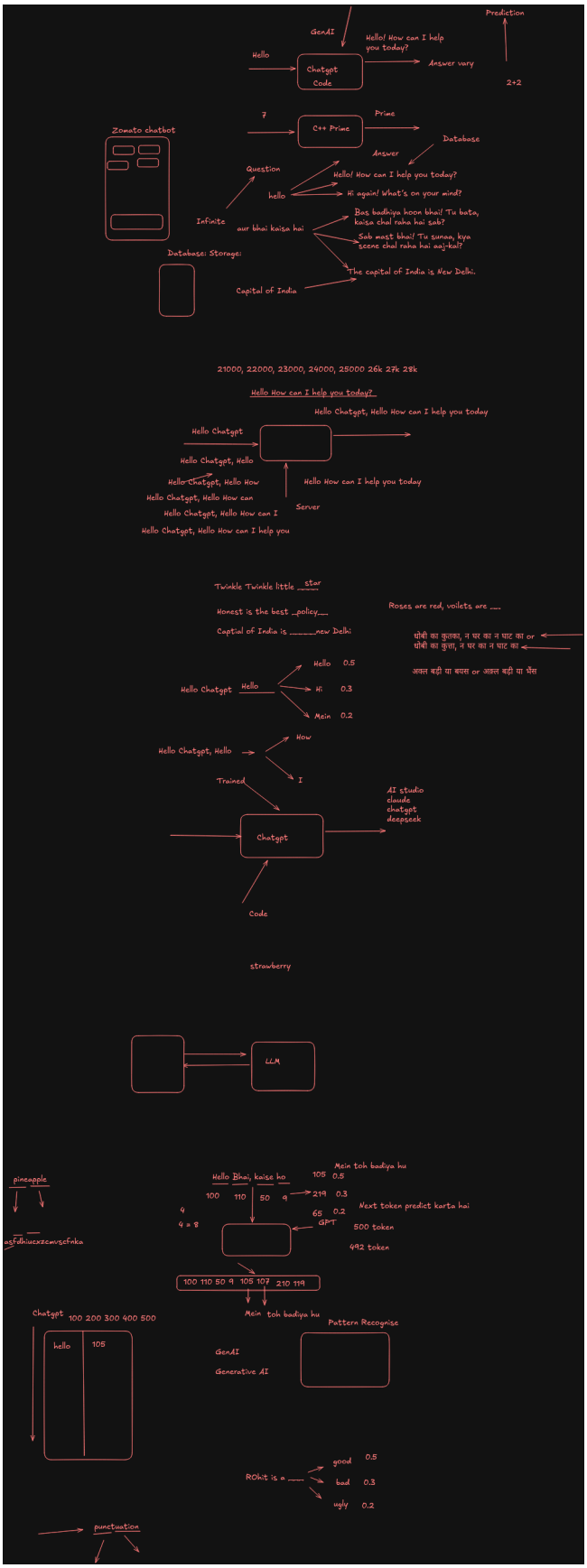
Remember:

"LLMs are powerful pattern predictors, not magic intelligence boxes."

Understanding how they work helps you:

- Use them effectively
- Know their limitations
- Build products with them
- Stay relevant in job market

End of Notes


GenAI
Hello! How can I help you today?
Hello
Chatgpt
Code
Answer vary
Prediction
2+2
Zomato chatbot
7
C++ Prime
Prime
Database
Answer
Question
Hello! How can I help you today?
hello
Hi again! What's on your mind?
Infinite
aur bhai kaisa hai
Bas badhiya hoon bhai! Tu bata, kaisa chal raha hai sab?
Sab mast bhai! Tu sunaa, kya scene chal raha hai aaj-kal?
Database: Storage:
The capital of India is New Delhi.
Capital of India
21000, 22000, 23000, 24000, 25000 26k 27k 28k
Hello How can I help you today?
Hello Chatgpt, Hello How can I help you today
Hello Chatgpt
Hello Chatgpt, Hello
Hello Chatgpt, Hello How
Hello Chatgpt, Hello How can
Hello Chatgpt, Hello How can I
Hello Chatgpt, Hello How can I help you
Hello How can I help you today
Server
Twinkle Twinkle little _star
Honest is the best _policy_
Captial of India is _____new Delhi
Roses are red, voilets are ___
धोबी का कुत्ता, न घर का न घाट का or
धोबी का कुत्ता, न घर का न घाट का
अक्ल बड़ी या भैंस or अक्ल बड़ी या भैंस
Hello Chatgpt
Hello
Hello 0.5
Hi 0.3
Mein 0.2
Hello Chatgpt, Hello
How
I
Trained
AI studio
claude
chatgpt
deepseek
Chatgpt
Code
strawberry
LLM
pineapple
asfdhiuexzcmvsefnka
Hello Bhai, kaise ho
100 110 50 9
219 0.3
65 0.2
GPT
Next token predict karta hai
500 token
492 token
4
4 = 8
Mein toh badiya hu
105 0.5
100 110 50 9 105 107 210 119
Mein toh badiya hu
Chatgpt 100 200 300 400 500
hello 105
GenAI
Generative AI
Pattern Recognise
ROhit is a ____
good 0.5
bad 0.3
ugly 0.2
punctuation