



Lecture 03: I Built a Chatbot From Scratch

 Practical Implementation of Chatbot using LLM (DSA-Specific)

Project Goal:

ऐसा chatbot बनाना है जो **सिर्फ DSA (Data Structures & Algorithms)** से related सवालों के जवाब दे।

-  Off-topic queries → reject कर देगा
 -  Only DSA Q&A allowed
-

Step 1: LLM कैसे काम करता है – **Token System**


Token क्या होता है?

LLM (ChatGPT, Gemini etc.) किसी भी input को **छोटे-छोटे parts** में तोड़ता है, जिन्हें **tokens** कहते हैं।

Token = Words, Punctuation, या Subwords का एक हिस्सा

Token Usage कैसे होता है?


Total Tokens = Input Tokens + Output Tokens

 Example:

Prompt = 300 tokens

Current user input = 10 tokens

Output by LLM = 20 tokens

 Total = 330 tokens used

Pricing Example:

Tokens Used	Approx. Cost
-------------	-----------------

1 Million Tokens ₹1000

2 Million Tokens ₹2000

🧠 Note:

"Hello kaise ho?" → 3–4 tokens only
→ जितना short prompt, उतना कम खर्चा



Step 2: LLM Context Limit & Chat History का Issue

LLMs एक बार में **सिर्फ limited tokens** तक की जानकारी process कर सकते हैं।
(जैसे GPT-3.5 → 4K tokens, GPT-4 → up to 32K)



Problem:

अगर पुरानी 150–200 messages भेजे, तो model confuse हो जाता है या काम करना बंद कर देता है।



Step 3: Context Selection Techniques (Memory Optimization)

LLM को हर बार **सिर्फ जरूरी history** भेजनी चाहिए।



Method



Description



Method 1 Last 50 messages + current input



Method 2 First 20 + Last 20 messages



Method 3 🔥 **Best:** Summary of all past + Last 50 messages




Best Practice:

पुरानी conversation का **summary generate करो (LLM से ही)**
फिर उसे current message के साथ भेज दो

"Summarize last 250 messages in under 200 tokens."


Step 4: LLM से Summary कैसे बनवाएं?

LLM खुद ही पुराने message को पढ़कर short summary बना सकता है।

 Example:

Summary: User asked mostly about Stack, Queue, and Binary Tree traversal.

Wants to practice recursion and DP questions next.

 हर बार पूरी history भेजने की जगह यह summary भेजना बेहतर होता है।

Step 5: System Instruction (LLM को Train करना)

LLM को एक fixed role देना बहुत ज़रूरी है — ताकि वो **सिर्फ DSA** related answers दे।

 **Example System Prompt:**

You are a DSA instructor chatbot.

Your job is to answer only questions related to Data Structures and Algorithms.

Reject or ignore any question outside this domain.

 यही method Zomato, Amazon जैसी companies भी use करती हैं।

Step 6: Strict Behavior Prompting (DSA-Only Mode)

जब user कुछ off-topic पूछे, तो chatbot politely reject कर दे।

 **Prompt Example:**

Sorry, I am designed to help only with Data Structures & Algorithms.

Please ask something related to DSA.

Final Chatbot Architecture

User Input



Context Manager

↳ Old Summary (by LLM)

↳ Last 50 Messages



System Instruction (DSA-Only Prompt)







LLM Model




Response to User


- ✓ Lightweight
- ✓ Token-Efficient
- ✓ Role-Focused

Key Learnings Recap

Topic	Summary
 Project Goal	Only DSA-based chatbot
 Tokens	हर input/output token count होता है
 Context Limit	LLM को limited memory होती है
 Summarization	पुराने chats का compressed context भेजो

 System Instruction Fixed role से chatbot को control करो

 Off-topic Blocking Prompt से chatbot को strict बनाओ

 Architecture Summary + Input + Role → Response