TRANSFORMERS AND PROMPTING

TRANSFORMERS

Built on Self-Attention : every token can “see” all others.
Processes sequences in parallel (unlike RNNs that go step by step).

LayerNorm → Multi-Head Attention → Feed-Forward → op.

WHAT IS AN LLM API CALL?

LLM = Large Language Model (e.g., Llama, GPT)
You can interact using APIs
The model responds to prompts (your instructions

MAIN PARAMETERS

Stream: Get responses token-by-token (for real-time chat).
Logprobs: Get probability of tokens → good for analysis

Parameter	Description	Tip
temperature	Controls randomness (0–2)	Lower = focused, higher = creative
top_p	Chooses tokens from top probability mass	Lower = more precise
max_tokens	Response length limit	Prevents long outputs
function_call	Lets LLM call defined APIs	Great for structured outputs

sampling vs picking

The sky is ...

day 2 stuff

yea

WHY USE LANGCHAIN?”

Content:

- Easier abstraction over raw API calls
- Adds memory, chains, tools, and prompt templates
- Helps you build chatbots, agents, RAG systems faster

Byeee