

MODULE 1 · WEEK 1 · LESSON 5

Inside an LLM: Tokens, Context and Next-Word Prediction

Demystify the chatbot in 8 minutes

WHAT YOU WILL BE ABLE TO DO

- Explain tokens and why models count them
- Describe context windows and why long chats degrade
- Explain temperature and what 'creativity' settings really do

LESSON CONTENT

The model predicts the next token

An LLM reads your text as tokens - chunks roughly three-quarters of a word - and computes probabilities for what token comes next. It samples one, appends it, and predicts again. Everything a chatbot does is this loop dressed in a friendly interface.

Context window: the model's short-term memory

The model can only 'see' a limited number of tokens at once, the context window. Your instructions, the conversation history and its own reply all compete for that space. When a long chat gets confusing, the window is crowded: restate the goal instead of saying 'as I said before'.

Temperature is a risk dial, not intelligence

Temperature controls how strongly the model favours likely tokens over unlikely ones. Low temperature: focused, repeatable. High: varied, surprising, more prone to nonsense. For facts and forms use low; for brainstorming, raise it a little. It changes randomness, not knowledge.

KEY TERMS

Token	The text chunk an LLM reads and predicts; ~0.75 of a word.
Context window	How many tokens the model can consider at once.
Next-token prediction	The core loop generating all LLM output.
Temperature	Sampling randomness control; higher = more varied output.

HANDS-ON TASK

Token experiment · Tools: ChatGPT / Gemini / Claude free tier

1. Paste a 200-word paragraph into any free chatbot.
2. Ask: 'Roughly how many tokens is this, and why?'
3. Then ask the same question twice more; compare how wording changes answers.

KEY TAKEAWAYS

- ✓ Before you go, three takeaways from this lesson.
- ✓ One: Explain tokens and why models count them.

- ✓ Two: Describe context windows and why long chats degrade.
- ✓ Three: Explain temperature and what 'creativity' settings really do.
- ✓ And your hands-on task: token experiment.
- ✓ Paste a 200-word paragraph into any free chatbot.
- ✓ When it is done, mark the lesson complete on your dashboard, keep the artifact for your portfolio, and step into the next lesson.
- ✓ I will see you there.

Watch the lesson video on your dashboard, complete the task, then mark the lesson complete. Your artifact belongs in your Portfolio of Evidence.