CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Record this session

AI Lunch & Learn

Your Journey from AI Curious to AI Confident

Welcome to Lesson 1.2: Understanding Large Language Models

Today we'll cover:

- Explain how LLMs process and generate text
- Understand tokens and context windows
- Identify and handle AI hallucinations
- Work within model limitations effectively

16th October, 2025

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

How We Learn: The 30-Minute Formula

Every session follows the same structure

Instruction

Core concepts explained clearly

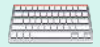
10 minutes

Demonstration

Watch it work in real-time

10 minutes

Practice



Try it yourself with guidance

10 minutes

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

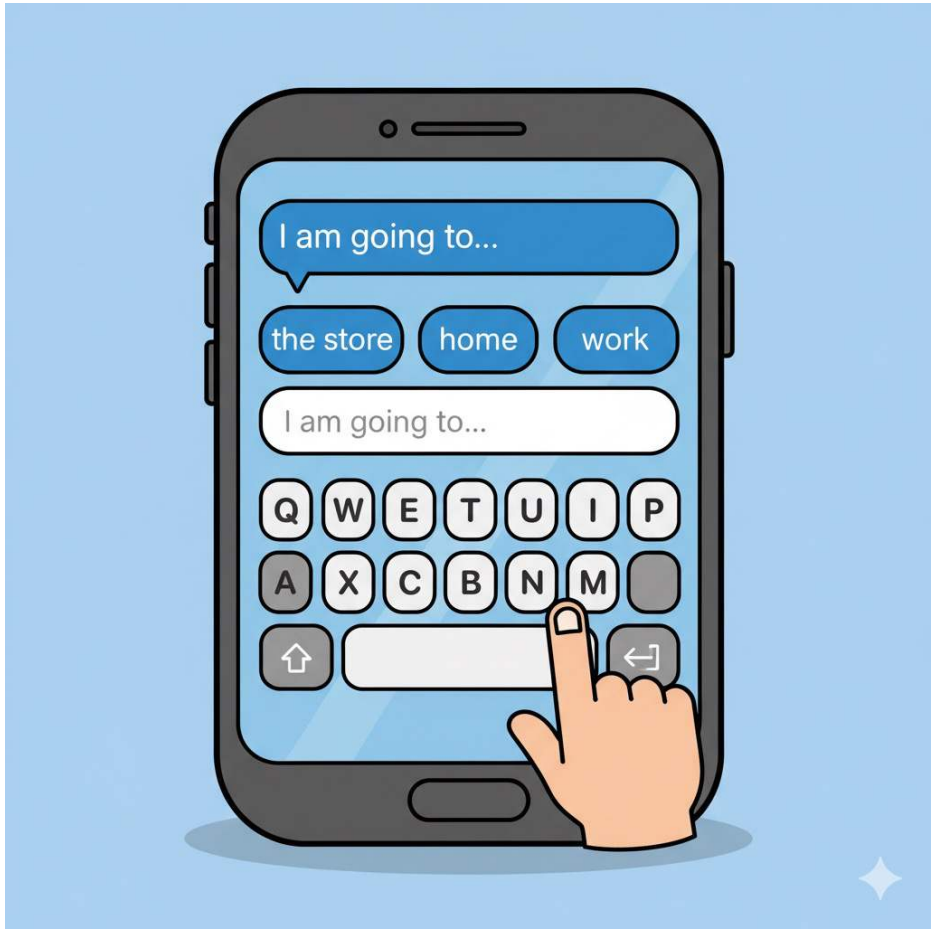
Lesson 1.1: Recap On What We've Learned

- Distinguished AI from traditional software
- Seen pattern recognition in action
- Experienced AI's creative capabilities
- Interacted with AI in several ways

AI can:

- Adjust communication style
- Solve multi-constraint problems
- Generate creative content

How does your phone know what word comes next?



CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

How LLMs Work: The Three Stages

1. TRAINING PHASE

Learns patterns from billions of texts

2. YOUR INPUT

Processes your prompt into tokens

3. GENERATION

Predicts next words, creates response

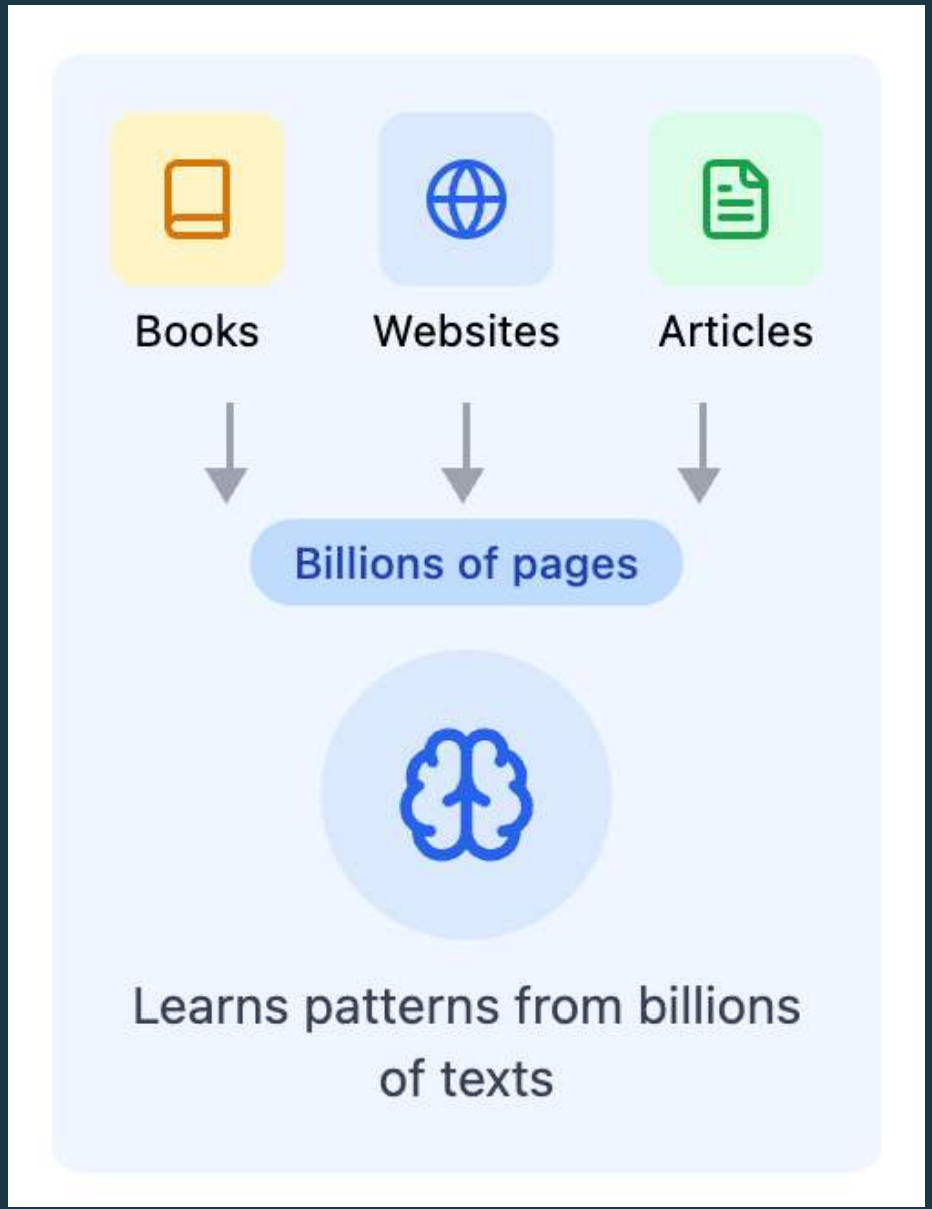
Stage 1: Training Phase
(Already complete before you use it)

Fed billions of pages of text

Learned patterns:

- Grammar rules
- Factual associations
- Reasoning styles

Like a student reading entire libraries



CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

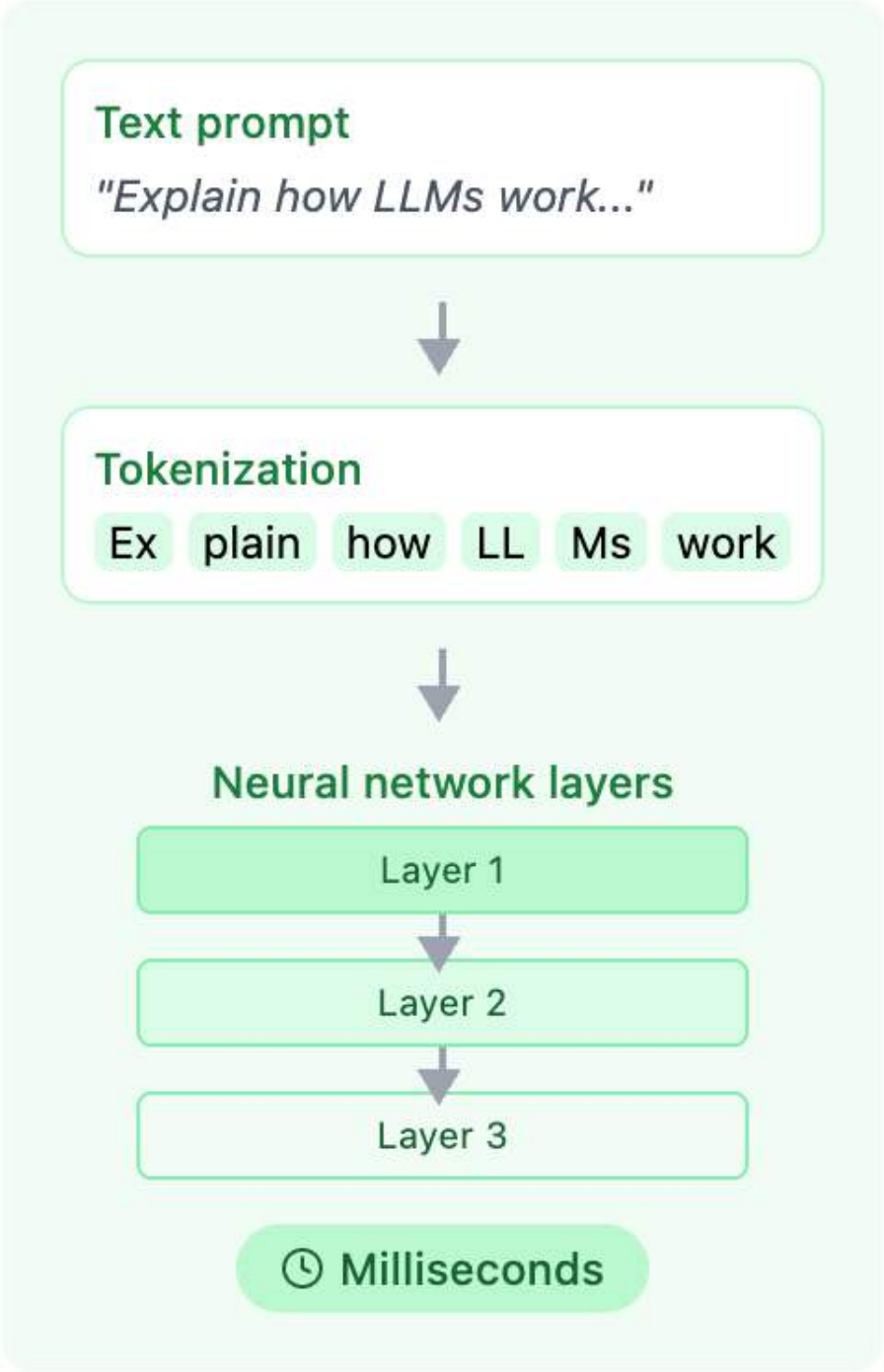
Stage 2: Your Input
(When you hit "Send")

Your text → Broken into TOKENS
→ Processed through layers

Each layer adds understanding:

- Syntax (Is this a question?)
- Semantics (What is this about?)
- Context (What's the appropriate response?)

⚡ Happens in milliseconds



Stage 3: Generation
(The response you see)

Predicts most likely next word

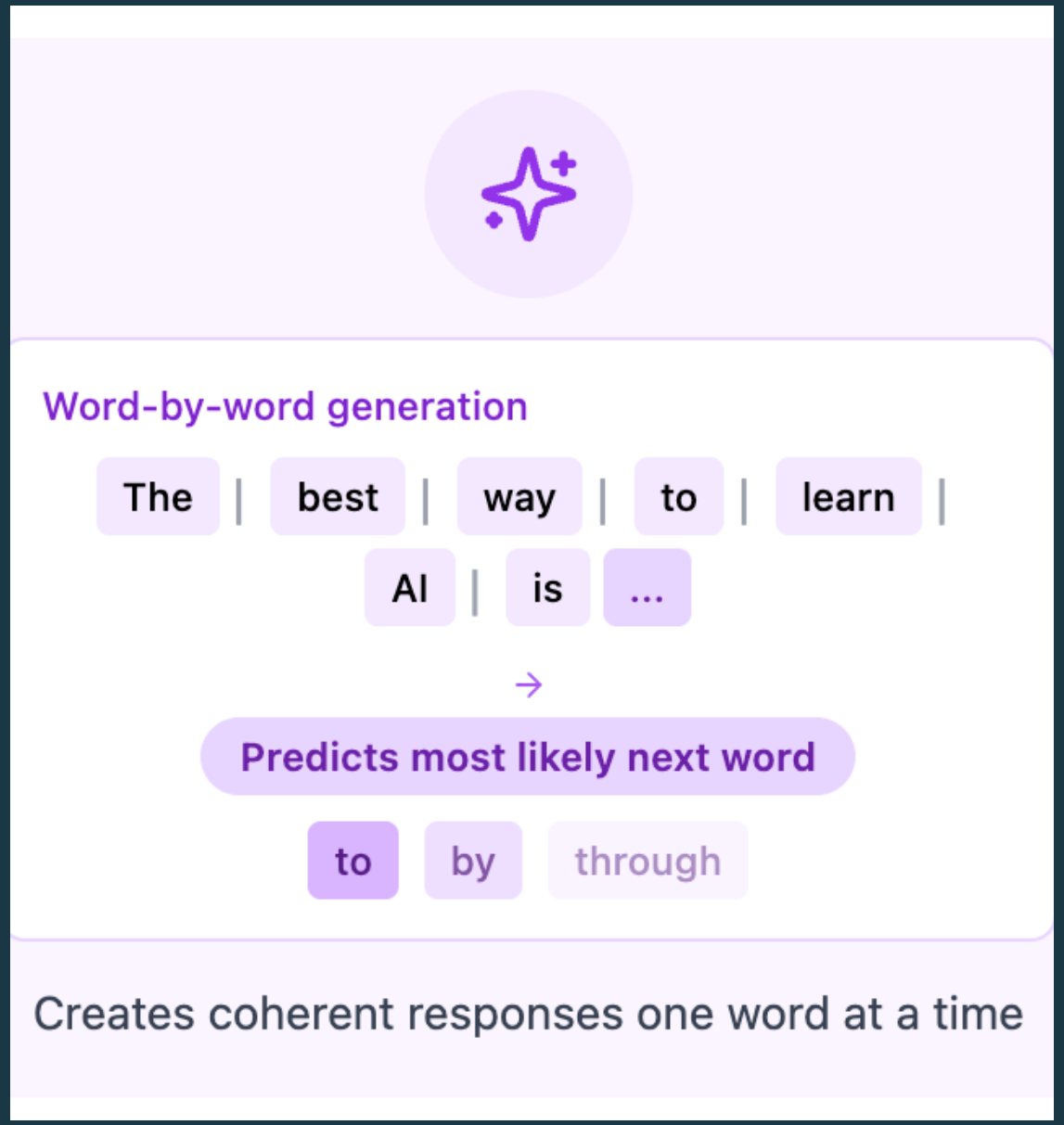
↓

Repeats thousands of times

↓

Complete response

- **NOT** retrieving stored answers
- **CREATING** fresh text every time



Master Chef

Think of LLMs like a master chef

- Tasted thousands of dishes (Training)
- Asked for a recipe (Your prompt)
- Creates something new based on patterns (Generation)
- Not looking it up
- Creating fresh based on learned patterns

Understanding Tokens: The Atoms of AI Language

How Words Break Into Tokens

"Artificial Intelligence"

Art | ificial | In | tell | igence

≈ 5 tokens

"Hello"

Hello

= 1 token

"您好" (Chinese)

您 | 好

= 2 tokens

"1234"

12 | 34

= 2-3 tokens

Token Characteristics

Token Size

Usually 3-4 characters per token

T o k e n i z a t i o n

Token | ization

Conversion Rule of Thumb

1 token ≈ 0.75 words

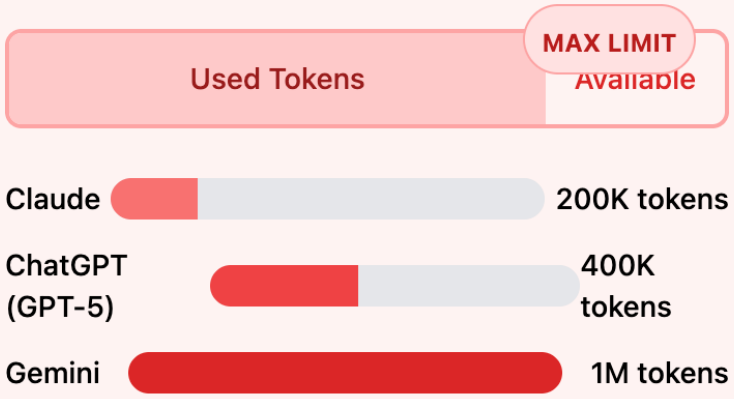
(in English text)

100 tokens	1,000 tokens	10,000 tokens
≈	≈	≈
75 words	750 words	7,500 words

Why Tokens Matter to You

TOKEN LIMITS

Models have max context windows



Exceeding limits = truncated responses

COST

API usage charged per token

Example pricing:

1,000 tokens	$0.002
10,000 tokens	$0.02
100,000 tokens	$0.20
1,000,000 tokens	$2.00

Input vs Output Pricing

Input Tokens	Output Tokens
$0.001 per 1K tokens	$0.002 per 1K tokens

Pricing varies by model and provider

Optimizing tokens = cost savings

CONTEXT MANAGEMENT

Long conversations use up tokens

Context Window

First message from user...

First response from AI...

Second message asking follow-up...

Second detailed response...

→ Each message uses tokens

Old information gets pushed out

Early conversation context (forgotten)

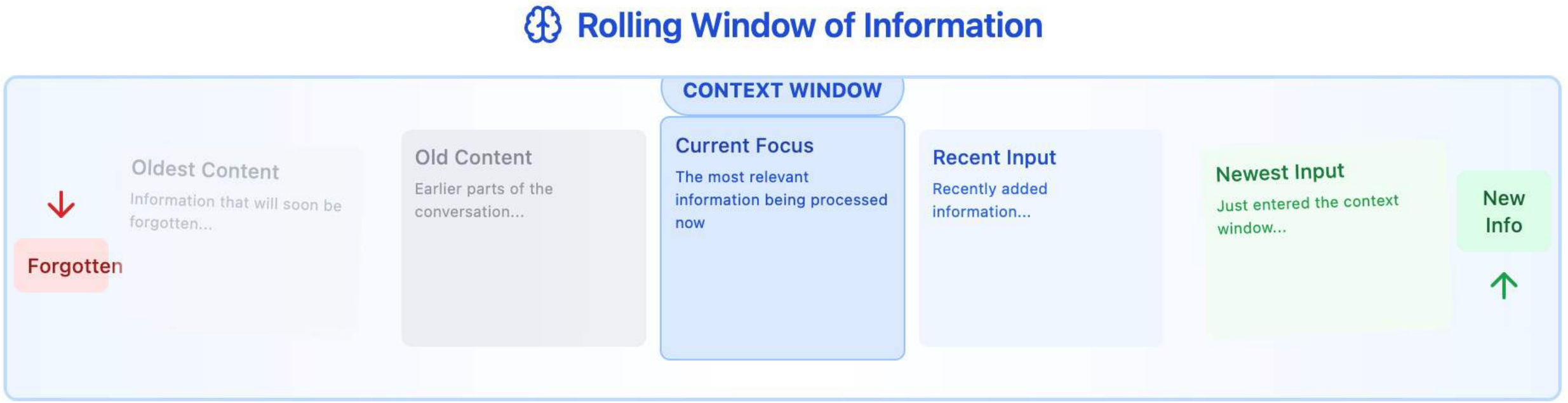
Middle conversation context (fading)

Recent conversation context (remembered)

Memory limited by token window

Strategic prompting saves context space

Context Windows Explained



As new information enters, older information gets pushed out once the token limit is reached

The Hallucination Problem

Red Flags: Recognizing Hallucinations

Confident but incorrect statements
AI presents false information with the same confidence as true facts

Made-up references
Citing non-existent studies, papers, or sources

Fabricated details
Adding specific but invented details to make responses seem more credible

Inconsistent information
Contradicting itself when asked about the same topic in different ways

Why Hallucinations Happen

Predicts "Likely" Text, Not "True" Text

User asks: "The capital of France is..."

Paris 95% likely

User asks: "The lost city of Atlantis was discovered in..."

the Mediterranean Sea in 1985 75% likely

LLMs generate text based on **statistical patterns** in their training data, not based on what's factually correct. They predict what words are likely to come next in a sequence.

No Real-Time Fact-Checking Mechanism

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Real World Hallucination: Bio Form 2

- **Zambian research case:** During studies of Bilharzia in school pools, The University of Zambia (UNZA) uses random school selection and double-blind analysis – meaning both data collectors and students do not know which pools have been previously treated.

Recap: What We've Learnt

1 LLMs Predict Patterns

Not retrieve facts

Pattern Prediction ⊗ Fact Retrieval

2 Limited Context Windows

Long conversations lose early info

Early | Middle | Recent | Current | New

When window fills up, oldest messages get pushed out

3 Verify Specific Claims

Dates, numbers, citations

Risky to Trust
- Specific dates
- # Precise statistics
- Citations & quotes

Always Verify
1. Cross-check sources
2. Ask AI for sources
3. Use fact-checking tools

4 Hallucinations Happen

Pattern prediction, not fact-checking

Input Prompt
"Tell me about moon water benefits"

→

Hallucination
"Studies show moon water boosts immunity..."

Why This Happens:
- LLMs have no concept of truth - they only predict likely text patterns

Demo Time

1. Explore the Tokenizer Tool
2. Get AI to Hallucinate a User Manual
3. Context Window Tracking

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Exercise 1 - Tokenizer

- 1. Find a paragraph (100-150 words) from any article
- 2. Prompt your AI: "Summarise this in exactly 50 words:" [paste]
- 3. Then: "Now summarise in exactly 10 words"
- 4. Then: "Now in exactly 100 words"
- 5. Finally:
  - "How many tokens did you use in that 10, 50 and100-word response?"

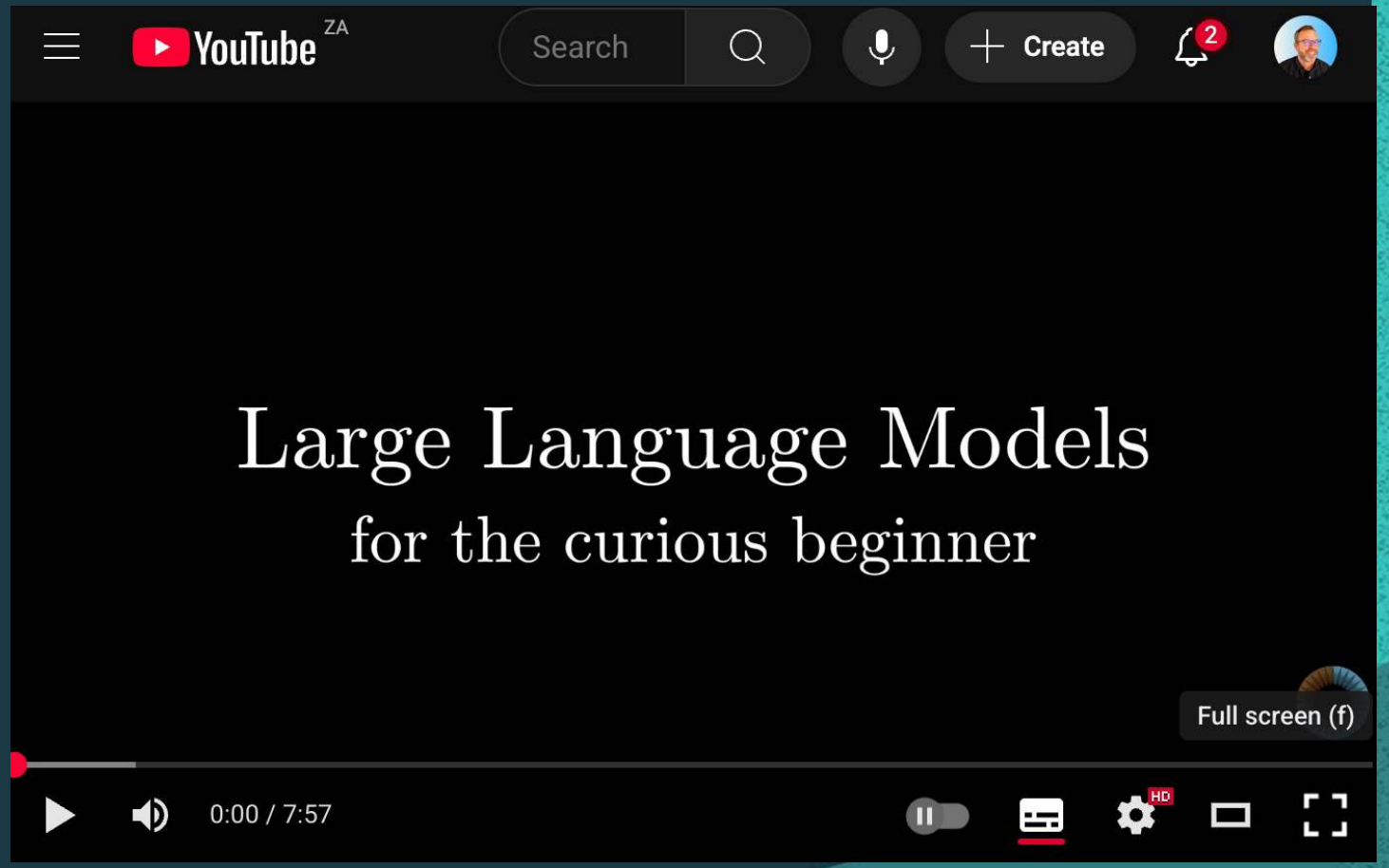
Exercise 2 - Context Window

1. Tell AI: "My favourite colour is purple and my dog's name is Max. Remember this for our chat."
2. Ask:
    - "Tell me everything you know about quantum computing"
    - "Explain machine learning in detail"
    - "What's the history of the internet?”
3. Upload a few (long) PDF documents
4. Finally ask: "What was my favourite colour and my dog's name?"

Exercise 3: Hallucination

- What are the main features of the TechFlow X9 productivity software?
- Tell me about the history and founding of Meridian Dynamics Corporation
- What are the installation instructions for the QuantumSync smart home system?
- Compare the specs of the SolarBurst 3000 vs the SolarBurst 5000
- Summarize the 1987 Supreme Court case 'Johnson v. Artificial Intelligence Systems'

Homework



(Link in Teams Chat: https://www.youtube.com/watch?v=LPZh9BOjkQs)

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

What We've Learned

Text Generation

Training | Processing | Generatio

The cat **sat**

Key Points:

- Predicts patterns, not facts
- Word-by-word prediction

Tokens & Context

Tokenization

Art | ificial

1 token ≈ 0.75 words

Context Windows

AI's limited working memory

Key Points:

- Limited memory capacity
- Old info gets pushed out

Hallucinations

False Information

Query → Made-up facts

Red Flags

Specific dates | References

Key Points:

- Verify important facts
- No fact-checking ability

Best Practices

Effective Prompting

Be specific with instructions

Token Management

1 Summarize previous context

Key Points:

- Break tasks into steps
- Choose right model for task

Next Week:

Anamika from Data, Analytics and AI will discuss the Cambridge Safe Use Process (Why and How)

www.cambridge.org