NO PREP

What Are Large Language Models (LLMs)?

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

What Are Large Language Models (LLMs)?

The Machine That Learned to Read the Internet

If you have ever typed a question into ChatGPT, Gemini, or Claude and gotten a full, coherent answer back,

Name: **Period:** **Date:**

What Are Large Language Models (LLMs)?

The Machine That Learned to Read the Internet

If you have ever typed a question into ChatGPT, Gemini, or Claude and gotten a full, coherent answer back, you were talking to a Large Language Model, or LLM. These are AI systems trained to understand and generate human language, and they are behind almost every AI chatbot, writing assistant, and customer service bot you encounter today. But what exactly are they, and why are they so different from the software that came before? Traditional computer programs follow rules. A programmer writes instructions, and the computer executes them exactly. If you type in a math problem, the program runs a formula. If you ask it something outside its programmed rules, it fails. LLMs work completely differently. Instead of following written rules, they learn patterns from reading enormous amounts of text, and then they use those patterns to predict what words should come next in any given situation.

Trained on Text, Not Programmed With Rules

The word 'large' in Large Language Model refers to two things: the amount of text used to train the model and the number of internal settings, called parameters, the model uses to make predictions. Modern LLMs like GPT-4 have been trained on hundreds of billions of words pulled from books, websites, news articles, Wikipedia, academic papers, and much more. They have hundreds of billions to trillions of parameters. During training, the model reads text and tries to predict the next word in a sentence over and over again, billions of times. Every time it gets it wrong, the system adjusts its internal settings slightly. By the time training is done, the model has essentially memorized the statistical patterns of human language. It knows that after the word 'the,' a noun often follows. It knows that 'president' often appears near words like 'White House,' 'election,' and 'policy.' It has

absorbed the style, structure, and meaning of an enormous amount of human communication. This process is called unsupervised learning. Unlike earlier AI systems that needed humans to label every example, LLMs train themselves by reading raw text and learning from patterns. No one had to label each sentence. The model figured it out on its own.

How LLMs Generate Text

When you type a question into an LLM, it does not look up an answer in a database. It predicts, word by word, what the most statistically likely response would be given everything it learned during training. This is called autoregressive generation. The model picks the most probable next word, adds it to the sentence, then picks the next most probable word after that, and keeps going until the response is complete. This is why LLMs can write in so many different styles, answer questions about almost any topic, and even tell jokes or write poetry. They are not retrieving stored answers. They are constructing responses from scratch using patterns they absorbed during training. It is also why they can make things up convincingly. If the pattern points toward a certain answer, the model will produce it even if the answer is wrong. This is called hallucination, and it is one of the biggest limitations of LLMs.

What Makes an LLM Different From a Search Engine

Many people initially compare LLMs to search engines, but they work very differently. A search engine like Google indexes billions of web pages and retrieves links to pages that match your query. It is finding existing documents. An LLM generates new text on the fly based on patterns it learned. It is composing, not retrieving. This distinction matters. When Google returns a result, you can click on the source and verify it. When an LLM gives you an answer, there is no source link by default. The text was generated, not copied. Some LLMs now include web search tools that let them retrieve real-time information, but the core model itself is always generating, not retrieving. This is why verifying LLM outputs against reliable sources is so important.

Who Builds LLMs and Why It Costs So Much

Building an LLM is not something a student or small company can do. Training a frontier model like GPT-4 or Claude costs tens to hundreds of millions of dollars. The process requires thousands of specialized computer chips called GPUs running

for weeks or months, enormous amounts of electricity, and massive teams of engineers, researchers, and data scientists. Only a handful of organizations in the world, including OpenAI, Google, Anthropic, Meta, and Microsoft, have the resources to train the largest models. Once trained, these models are made available to users through apps and APIs. OpenAI charges monthly subscriptions for ChatGPT Plus. Anthropic charges for access to Claude. Google integrates Gemini into its workspace products. The business of selling access to LLMs has become one of the fastest-growing sectors in the technology industry. For students, this means the tools you use for free today may cost more in the future as companies build out their products.

The Real Limits of What LLMs Know

LLMs have a knowledge cutoff, which means they only know about events that occurred before their training ended. If a major news event happened after training finished, the model has no knowledge of it. This is why ChatGPT can answer questions about historical events in detail but may be uncertain or wrong about things that happened recently. They also do not actually understand language the way humans do. They process statistical patterns, not meaning. If you ask an LLM a trick question or one that requires common-sense reasoning outside their training data, they can fail in ways that a human would not. Researchers call this the difference between statistical intelligence and genuine comprehension. LLMs are extraordinarily good at pattern matching, but pattern matching and true understanding are not the same thing. The concept of emergence is one of the most discussed phenomena in AI research. As language models scale to hundreds of billions of parameters and are trained on increasingly large datasets, they begin to exhibit capabilities that were not explicitly trained for and that do not appear in smaller models at all. Researchers call these emergent abilities. For example, GPT-3, when it reached a certain scale, suddenly demonstrated a meaningful ability to perform arithmetic, answer analogy questions, and translate between languages, none of which were explicitly in its training objective. This emergent behavior suggests that scale itself, rather than targeted training, may be a driver of new capabilities. The business implications of LLMs extend well beyond individual chatbot interactions. Enterprises are integrating LLMs into customer service pipelines, legal document review, software development workflows, marketing content generation,

and medical record summarization. The consulting firm McKinsey estimated in 2023 that generative AI could add between $2.6 trillion and $4.4 trillion annually to the global economy across these use cases. This kind of economic projection, whether or not it proves accurate, explains why investment in LLM development and deployment has accelerated so rapidly despite the significant costs involved.

Key Information: Large Language Models

Detail	Information
What LLM stands for	Large Language Model
Core function	Understand and generate human language using pattern recognition
Training method	Unsupervised learning on massive text datasets
'Large' refers to	Size of training data and number of model parameters
How responses are made	Word-by-word prediction (autoregressive generation)
Major LLM developers	OpenAI, Google, Anthropic, Meta, Microsoft
Key limitation: hallucination	Generating confident but incorrect information
Key limitation: knowledge cutoff	No awareness of events after training ended
Difference from search engines	LLMs generate new text; search engines retrieve existing pages
Training cost	Tens to hundreds of millions of dollars for frontier models

Case Study: Marcus Tries to Use ChatGPT for His History Essay

Marcus is a junior who just discovered ChatGPT. For his history class, his teacher assigned a research essay on a recent political event. Marcus typed his question into ChatGPT and got back a beautifully written, confident-sounding five paragraphs. He was thrilled. He submitted the essay without checking the facts. The next day, his teacher handed the essay back with a failing grade and a note: three of the key facts were wrong, and one of the events described had not actually happened that way. Marcus was confused. The AI had sounded so sure. He went back to ChatGPT and asked it to cite its sources. The model produced a list of sources, but when Marcus clicked on them, two of the links did not exist. He realized he had submitted AI-generated content without understanding how LLMs actually work. He assumed the AI was retrieving real facts from a database. Instead, it had been generating statistically probable text, whether or not that text was accurate.

1. Based on what you read about how LLMs work, explain in your own words why Marcus's essay contained wrong information even though ChatGPT sounded confident.

2. What should Marcus do differently the next time he uses an AI tool for a class assignment?

Multiple Choice

1. What does the word 'large' in Large Language Model primarily refer to?

    a) The size of the training data and number of parameters

    b) The physical size of the computers used to run the model

    c) The length of the responses the model produces

    d) The number of languages the model can speak

2. How does an LLM generate a response to your question?

    a) It copies the text word for word from the most relevant website

    b) It predicts the most probable next word, one word at a time

    c) It searches a database of stored answers

    d) A human employee types a response

3. What is AI hallucination in the context of LLMs?

    a) The AI generating confident but inaccurate information

    b) The AI dreaming or thinking during low-activity periods

    c) The AI refusing to answer certain questions

    d) The AI displaying visual images alongside text

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Large Language Model (LLM)	A) The word-by-word prediction process LLMs use to construct responses
2. ___ Parameters	B) When an AI model produces confident but factually incorrect information
3. ___ Knowledge cutoff	C) Internal numerical settings an AI model adjusts during training to improve predictions
4. ___ Autoregressive generation	D) The date after which an AI model has no knowledge of world events
5. ___ Hallucination	E) An AI system trained to understand and generate human language using pattern recognition

Fill-in-the-Blank

1. LLMs learn by reading enormous amounts of _____ and predicting what word comes next.

2. When an LLM produces a wrong answer confidently, this is called _____.

3. Unlike search engines that retrieve existing pages, LLMs _____ new text from scratch.

Short Answer

Answer each question in 2-4 complete sentences.

1. Explain the difference between how a traditional computer program works and how an LLM works.

2. Why does it cost tens or hundreds of millions of dollars to train a large language model?

3. Why is it important to verify information generated by an LLM rather than trusting it automatically?