How does a large language model work?

KS3 KS4 Ages 11-18 ⏰ 5 min read

You type a question and within seconds a machine writes back a thoughtful, coherent, often impressively accurate answer. It can write poetry, debug code, summarise documents, and argue philosophical positions. So what's actually happening inside these systems? The honest answer is: a lot of very clever maths, and something that turns out to be surprisingly unlike human thinking.

It starts with prediction

At its core, a large language model (LLM) is a next-word predictor. Given a sequence of words, it predicts what word is most likely to come next — then the next, then the next — building up a response one token at a time. That sounds trivial, but when you train a system to do this prediction on a vast enough scale, something remarkable emerges: the ability to answer questions, reason through problems, and write convincingly in almost any style.

📚 Imagine someone who has read almost every book, article, forum post, and website ever written. They haven't understood any of it in a deep sense — but they've become extraordinarily good at pattern matching: "when a conversation goes like this, it usually continues like that." An LLM is something like that. It has absorbed the patterns of human language so thoroughly that it can reproduce them in a way that looks remarkably like understanding.

How does training work?

LLMs are trained on enormous datasets — billions of pages of text from the internet, books, and other sources. During training, the model repeatedly tries to predict the next word in a piece of text, compares its prediction to the actual word, and adjusts billions of internal numerical weights to do better next time. This process, run across millions of examples on thousands of computer chips over weeks or months, produces a model that has effectively compressed the patterns of human language into its weights.

After this pre-training, models are then fine-tuned using human feedback — trainers rate responses for helpfulness and accuracy, and the model is nudged further in the

direction of useful, safe responses.

Does it actually understand anything?

This is the genuinely contested question. LLMs can fail in ways that suggest no real understanding — making confident factual errors, being thrown by slight rewording of a question they answered correctly, struggling with simple logical puzzles. On the other hand, they perform well on tests designed to measure reasoning and even show emergent capabilities their creators didn't deliberately train in. Most researchers sit somewhere between "definitely not conscious" and "we genuinely don't fully understand what's happening in there." What's clear is that it's not the same as human understanding — but it's also not nothing.