TECH · ARTIFICIAL INTELLIGENCE AGES 11–16

How Large Language Models Work

From your typed question to a finished reply, built one predicted token at a time.

You type a prompt

You ask something, such as “Explain why the sky is blue”, and the text is sent to the model.

Text is split into tokens

The prompt is broken into tokens: words, parts of words or punctuation. In English, one token is about ¾ of a word.

Tokens become numbers

Each token has an ID, which is turned into a long list of numbers called an embedding that the network can calculate with.

Self-attention adds context

Transformer layers work out how much each token matters to every other one, for example what “it” refers to.

Next token is predicted

The model gives every token in its vocabulary a probability and picks one, usually from the most likely options.

Is the answer finished?

If not, the new token is added to the text and the model predicts again. It stops when the reply is complete or a length limit is hit.

The answer appears

The chosen tokens are turned back into text and shown on your screen.

Prediction, not fact-checking

An LLM predicts likely text; it does not look answers up. It can “hallucinate”, stating wrong things confidently, so check important facts elsewhere.

Free to share · CC BY-NC 4.0 · Sources at infographhub.com