How Large Language Models Actually Work

FREE EDUCATIONAL BRIEF · LEARN AI. UNCOVER THE DATA. PROTECT THE TRUTH.

Large language models power almost every AI product you touch — chat assistants, writing tools, coding copilots, customer-service bots. They can seem magical, but the core idea fits in one sentence: an LLM is a system trained to predict the next word. Understanding how that single trick produces essays, answers, and arguments is the first step in learning to build with AI and see through it.

TOKENS: THE ATOMS OF MACHINE LANGUAGE

Models don't read words the way you do. Text is chopped into tokens — chunks of characters that might be a whole word, part of a word, or punctuation. “Investigator” might become three tokens; “AI” is one. Every token is converted into a long list of numbers called an embedding, which captures how that token relates to every other token the model has seen. All of the model's “knowledge” lives in these numerical relationships.

TRAINING: PATTERN-PRESSING AT PLANETARY SCALE

» Pre-training

The model reads enormous amounts of text and, billions of times over, plays the same game: hide the next token, guess it, measure the error, adjust. No facts are stored — only statistical patterns of language.

» Fine-tuning

The raw model is then shaped with curated examples of helpful, safe, on-format answers, teaching it how to behave, not just what words follow what.

» Feedback alignment

Human reviewers rank candidate answers, and the model learns to prefer responses people rate highly — this is why modern assistants feel cooperative.

WHY MODELS HALLUCINATE

Because an LLM predicts plausible text rather than retrieving verified facts, it can produce confident, fluent statements that are simply wrong — a citation that doesn't exist, a date that's off by a decade. This isn't lying; it's the prediction machine doing exactly what it was built to do, without a built-in truth check. That is why serious systems pair models with retrieval from trusted sources, and why human verification remains the final control.

INVESTIGATOR'S TAKEAWAYS

- An LLM predicts the next token; everything else is an emergent effect of that objective.
- Its knowledge is statistical pattern, not a database of verified facts.
- Fluency is not accuracy — confidence in tone tells you nothing about truth.
- Hallucination is structural, so verification workflows are not optional.
- Systems that ground answers in real documents (retrieval) dramatically reduce — but never eliminate — errors.