

WHAT AI ACTUALLY IS (AND ISN'T)

THE BIG IDEA

AI **predicts what sounds right**, and it can be confidently wrong. Its strengths are **jagged**. And **you're the one in charge**. The video's two opening stories (lawyers fined for filing six court cases ChatGPT invented, and an AI search summary suggesting glue on pizza) both come from that first fact.

WHAT A LARGE LANGUAGE MODEL ACTUALLY DOES Video 2:08

A large language model (LLM), the technology behind ChatGPT, Copilot, Gemini, Claude and similar tools, works like a very sophisticated version of the autocomplete on your phone. It looks at the text so far, predicts a likely next piece, adds it, and repeats. Sometimes one continuation is far more likely than any other. Often many are plausible, and the model still has to pick one.

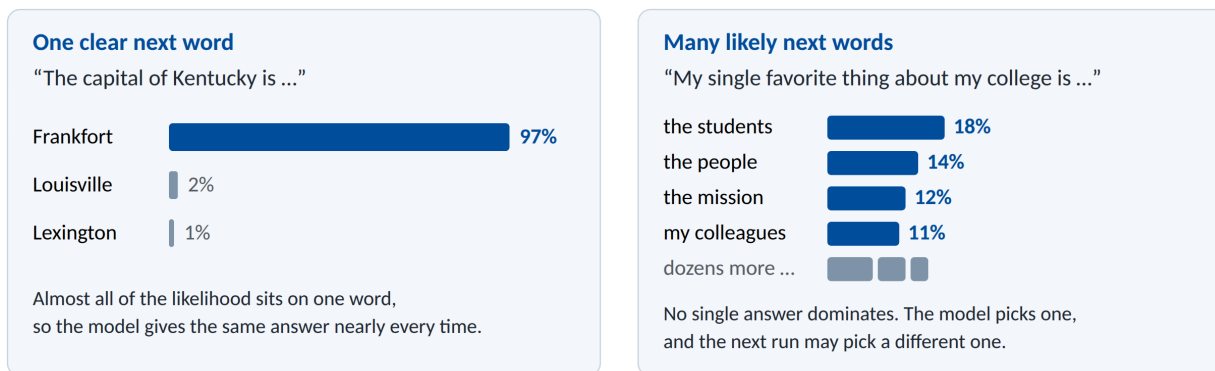
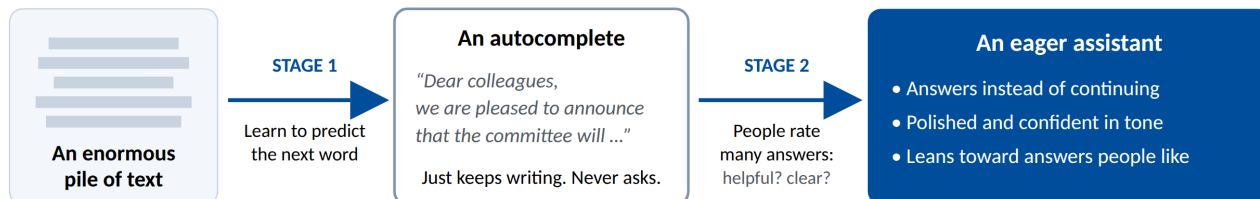


Figure 1. Percentages are illustrative, not from a real model.

CLARIFICATION: IT DOESN'T LOOK THINGS UP

The model doesn't keep a copy of what it read. It rebuilds every answer from patterns it picked up in training. Think of it as working from a **blurry picture** of everything it read: the broad shapes are clear, the fine details are fuzzy, and when a detail is missing it fills in something that looks right. That's why exact names, dates, numbers, and page ranges are where it most often slips.

FROM AUTOCOMPLETE TO ASSISTANT: TWO STAGES OF TRAINING Video 3:53



Notice what's missing: neither stage, on its own, checks whether an answer is true.

Stage two is why the model answers like an eager assistant, sounds polished and sure of itself, and leans toward the answer you'd like to hear. When a lawyer asked ChatGPT whether its cases were real, "yes" was exactly the kind of answer that gets rewarded. Whether any of this adds up to real understanding is an open debate. The practical point: left to itself, the model is predicting *what sounds right and what's likely to be rewarded*, not checking whether it's correct.

SEARCH CHANGES WHAT IT READS, NOT HOW IT WRITES Video 5:24

Many tools now search the web or read files you give them. That helps, but it changes what the model reads, not how it writes: the answer is still predicted. The glue-on-pizza advice came from this step, when Google’s search pulled in an 11-year-old Reddit joke and the AI summary wrote it up as cooking advice (Reid, 2024). The blurry picture also has a date on it, the **training cutoff**. Search can add newer material, but the picture itself doesn’t update.

THREE BEHAVIORS, ONE CAUSE Video 6:12

Behavior	What it looks like	Why it happens
Confident when wrong	A wrong answer in the same assured tone as a right one, with nothing to tip you off.	Polished, confident answers tended to score well when people rated them in stage two.
Invents what fits (a “hallucination”)	Something false but believable: a fake citation, a statistic nobody measured, a policy your college doesn’t have.	It’s very good at producing the shape of an answer. Whether the content is real is a separate question.
Different every time	The same question in two chats gets slightly, or very, different answers.	It chooses among plausible options instead of always taking the single most likely one.

THE CITATION TEST (VIDEO 7:18)

In the video, we asked Claude for three citations on work colleges. One came back as *Stinebrickner & Stinebrickner (2006), Research in Higher Education, 47(6)*. Google Scholar showed the article and journal are real, but the authors are Wolniak and Pascarella (2007), in volume 48. Claude had searched the web first and still got the authors wrong.

Why it matters: a partly right citation is harder to catch than an invented one, because some of it checks out.

CLARIFICATION: WHY GUESS INSTEAD OF “I DON’T KNOW”?

Researchers at OpenAI and Georgia Tech (Kalai et al., 2025) offer one explanation: models are trained and scored like students on a test with no penalty for wrong answers. A guess might earn credit; “I don’t know” never does. So they learn to guess, confidently.

Temperature: the setting that controls how much variety you get. Most chat apps don’t let you change it, and even at the lowest setting answers aren’t perfectly repeatable (He & Thinking Machines Lab, 2025).

WHERE AI HELPS AND WHERE IT TRIPS UP: THE JAGGED FRONTIER Video 10:40

In a field experiment with 758 consultants at Boston Consulting Group (Dell’Acqua et al., 2023), consultants using AI on tasks **inside** the AI’s capabilities completed 12.2% more tasks, finished 25.1% faster, and produced work rated more than 40% higher in quality. On one task chosen to fall just **outside** them, consultants using AI were 19 percentage points less likely to reach the right answer than those working without it.

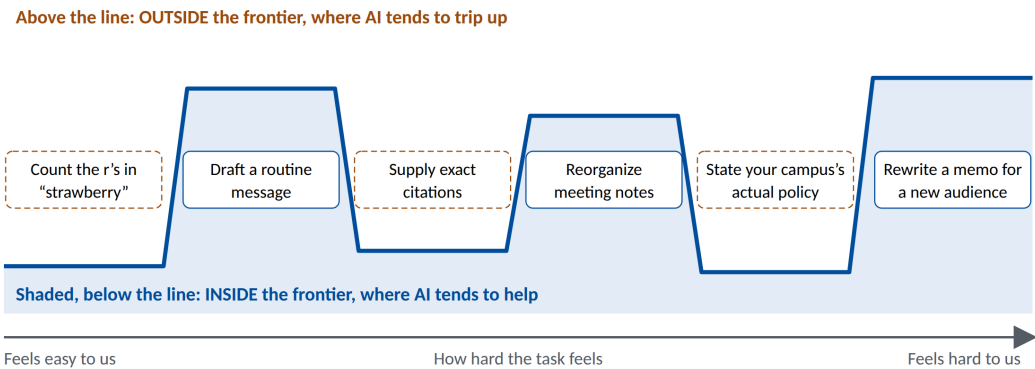


Figure 3. How hard a task feels is a poor guide to which side it lands on. Placement is illustrative and varies by tool.

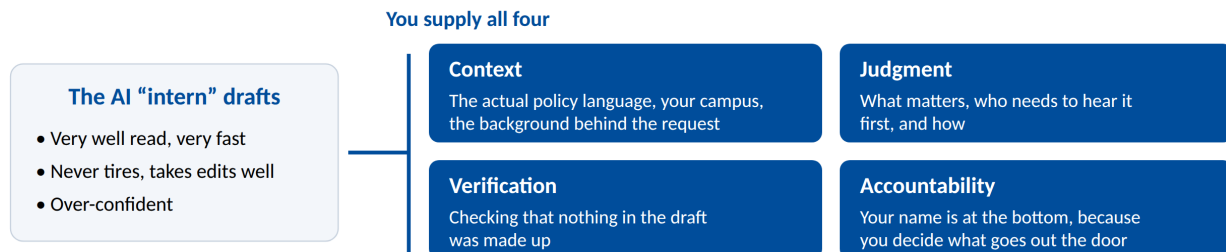
What helps most is knowing the work well enough to direct the AI and check what it gives you. The video’s rough map also lists **getting the gist** of a document you might not otherwise read (where AI tends to help) and **anything too recent** to be in its training (where it tends to trip up).

- **Summaries:** reading is how you come to know something, and whether it matters. A summary is only the model’s guess about what mattered.

- **Why counting trips it up:** the model doesn't see letters, or even whole words. It sees chunks of text called **tokens**. Counting the r's in "strawberry" through tokens is like counting through frosted glass.
- **The better it looks, the less we check:** in another study, recruiters given a better AI tool paid less attention to their own reviews, and their decisions got slightly worse (Dell'Acqua, 2022).

A WORKING MENTAL MODEL: THE OVER-CONFIDENT INTERN Video 13:10

Think of AI as a very well-read, very fast, over-confident intern. It drafts in seconds and never gets tired. But it doesn't know your campus, has no real sense of what matters, and isn't accountable for anything it hands you.



Example: a draft email to faculty about an advising policy change takes seconds, but without the actual policy language it will invent a plausible version. **Where the analogy breaks:** real interns learn from your corrections, have a stake in the work, and can be held responsible; AI has none of those. Whether it remembers past conversations depends on the tool and its settings (many personal accounts do by default, many campus versions don't), so check yours. Either way, accountability stays with you.

CHECK YOURSELF: FIVE TERMS (VIDEO 14:48)

Define each term in a few words, then compare.

LLM: a large language model; it predicts text, one chunk at a time.

Token: one of those chunks, a word or a piece of a word.

Hallucination: something false but believable.

Training cutoff: the date the blurry picture was taken.

Temperature: how much variety the model allows.

TRY IT TOMORROW

Check one thing. Next time a tool hands you a fact, a number, or a name, check one before you use it. It's the check those two lawyers never made.

Ask twice. Same question, two chats. When is the variety useful, and when is it a problem?

Count the r's in "strawberry." Has your model figured it out yet?

Tools change almost monthly; assume today's AI is the worst you'll ever use (Mollick, 2024). Framing the problem, judging the answer, and owning the result won't go out of date.

SOURCES

- Dell'Acqua, F. (2022). *Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters* [Working paper]. Laboratory for Innovation Science at Harvard.
- Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality* (Working Paper No. 24-013). Harvard Business School. <https://doi.org/10.2139/ssrn.4573321>
- He, H., & Thinking Machines Lab. (2025, September). Defeating nondeterminism in LLM inference. *Connectionism*. <https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/>
- Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why language models hallucinate* (arXiv:2509.04664). arXiv. <https://arxiv.org/abs/2509.04664>
- Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443 (S.D.N.Y. 2023).
- Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio.
- Reid, E. (2024, May 30). *AI Overviews: About last week*. Google. <https://blog.google/products-and-platforms/products/search/ai-overviews-update-may-2024/>
- Wolniak, G. C., & Pascarella, E. T. (2007). Initial evidence on the long-term impacts of work colleges. *Research in Higher Education*, 48(1), 39–71.

This handout was created in collaboration with Claude AI (Anthropic).