

LEARNING NUGGET E-LEARNING & AI 12 SEPTEMBER 2026

AI doesn't lie. It guesses.

Why language models make things up, why they sound so sure while doing it, and three questions that help learners spot invented answers.

By Rico Eberle, e-learning expert for continuing education and municipal councillor in Dübendorf

RICO EBERLE · LEARNING NUGGET

AI doesn't lie. It guesses.

And it always sounds sure of itself.



Video for this post

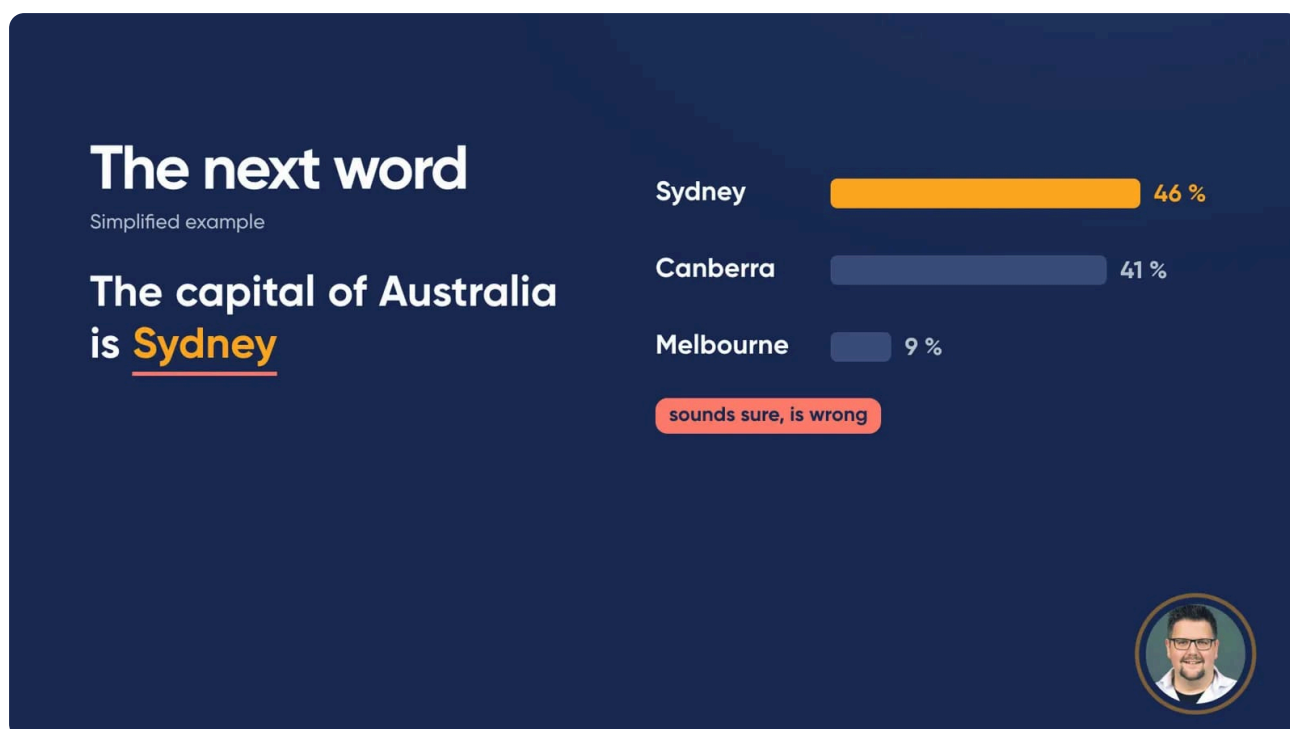
This learning nugget is also available as a video with transcript (1:18 min). Scan the QR code or open the link:

ricoeberle.ch/en/blog/why-ai-hallucinates/

Anyone who works with AI tools knows the moment: the answer is fluent, well structured and sounds competent. And yet it is wrong. A source that does not exist. A year that is slightly off. A quote nobody ever said. This is called hallucination. The word suggests a malfunction. In fact, it follows directly from how language models work.

A language model knows probabilities, not truth

A language model does not look things up. Put simply, it calculates which word is most likely to come next, word by word. Because it was trained on enormous amounts of text, the most likely word is very often the right one.



The most likely word is not always the right one: Sydney instead of Canberra.

When the model does not know something, that process does not change. It still picks the most likely next word. The result is a sentence that sounds just as confident as a correct answer. Hesitation, a "maybe" or a hint of uncertainty does not appear on its own.

Why the AI doesn't simply say "I don't know"

In 2025, OpenAI published an explanation, together with a research paper. The core idea: language models are scored much like students on a multiple-choice test. If you guess, you sometimes get points. If you leave it blank, you certainly get none.

Why guess?

**Guess**

sometimes points

**Leave blank**

certainly 0 points

So the model learns: guessing pays off.



As in a multiple-choice exam, guessing earns more points on average than staying silent.

When scoring rewards guessing and penalises restraint, a model learns exactly that: guessing pays off. Hallucination is therefore also a result of how models are measured and improved.

What this means for teaching

The key insight for learners: **a confident tone says nothing about accuracy.** An AI answer is a draft, not a reference work.

Three checks

Confident tone ≠ correct answer

- 1 Can I find this in a second, independent source?
- 2 Can names, numbers and quotes be verified?
- 3 What happens if I ask the same question differently?



Three checks for every AI answer.

Three questions help in everyday use:

1. **Can I find this in a second, independent source?** Not in a second AI, but in a source that is itself backed up.
2. **Can names, numbers and quotes be verified?** These details are especially vulnerable because they have to sound plausible but should be exact.
3. **What happens if I ask the same question differently?** If the answer changes with a slightly different wording, that is a warning sign.

An exercise for your next lesson

Have learners work in groups and ask an AI about a topic they already know well. Each group uses the three questions to find a statement that is wrong and backs up the error with a source. The debrief usually shows quickly how convincing wrong answers can be. And it turns distrust into a technique that can be practised.

AI is not an encyclopaedia, it's a very good guesser. Knowing that is what makes it useful.

Frequently asked questions

Why does AI hallucinate?

A language model calculates, word by word, which word is most likely to come next. When it doesn't know something, it still picks the most likely word, and the result sounds just as confident as a correct answer.

Why doesn't AI just say "I don't know"?

According to an explanation OpenAI published in 2025, language models are scored much like students on a multiple-choice test: guessing sometimes earns points, leaving it blank certainly earns none. So a model learns that guessing pays off.

How can you spot invented AI answers?

Three questions help: can I find this in a second, independent source? Can names, numbers and quotes be verified? What happens if I ask the same question differently?

Sources

1. OpenAI: Why language models hallucinate (2025)
<https://openai.com/index/why-language-models-hallucinate/>
2. Kalai et al.: Why Language Models Hallucinate (2025)
<https://arxiv.org/abs/2509.04664>



ABOUT THE AUTHOR

Rico Eberle

Rico Eberle is an e-learning expert, business economist (FH) and municipal councillor in Dübendorf, Switzerland. He chairs the foundation board of WBK Dübendorf, a continuing education foundation. In the learning nuggets he explains research on learning, AI and digital sovereignty, briefly and with sources.

Contact: ricoeberle.ch · LinkedIn: [linkedin.com/in/rico-eberle](https://www.linkedin.com/in/rico-eberle)

LICENCE

Text, transcript and video by Rico Eberle are licensed under CC BY 4.0 (music and sound effects in the video excluded). You may share and adapt the content, including for teaching, as long as you credit the source. creativecommons.org/licenses/by/4.0/

SUGGESTED CITATION

"AI doesn't lie. It guesses.", Rico Eberle, ricoeberle.ch, 2026, <https://ricoeberle.ch/en/blog/why-ai-hallucinates/>, CC BY 4.0