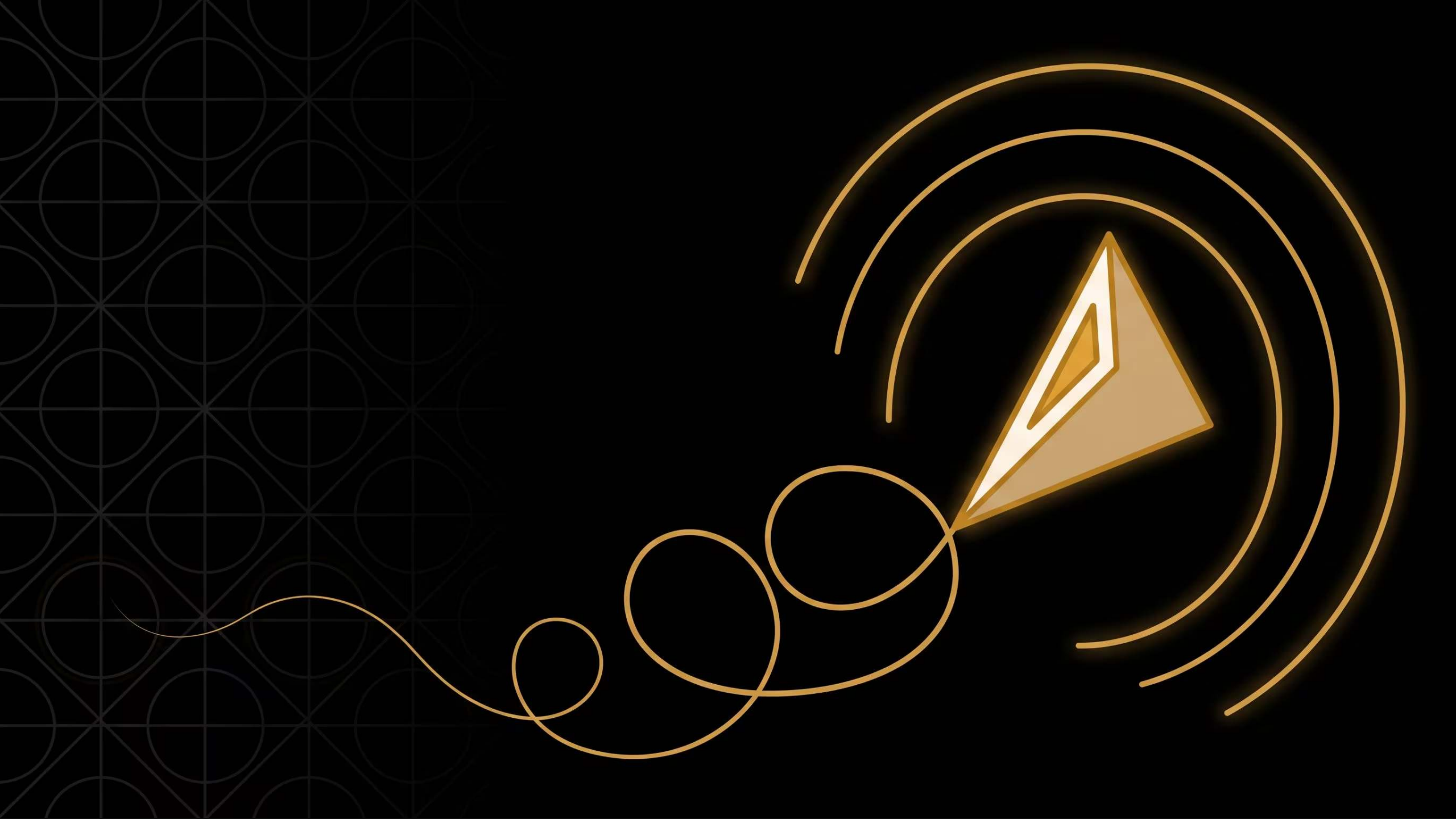




Demystifying large language models

Hands up: used one this week. Keep them up if you could explain to a ten-year-old what it did.

Most hands come down.



SAY: we got the tool before we got the story about the tool.

Not too complicated — it just happened too fast.



SAY: one story says it's basically a mind. The other says it's just autocomplete.

Opposites doing the same job: beyond you or beneath you, either way don't look inside.



it's hard to have an opinion
on a machine you've never seen inside

Cue the house lights up and the ushers to start handing out the sheets, because together we're going to look inside a language model. Not a diagram of the inside — the inside.

Take one, pass the rest along, try not to leave any that aren't in someone's hand.



a page of **token pairs** —
a word, and the word before it

Hold one up: a sheet of token pairs. The boxed token is the word that came before; the one beside it is what followed it, somewhere in the text. Mostly words, some punctuation. Every word keeps the same colour wherever it appears, so you scan by colour and confirm by reading.

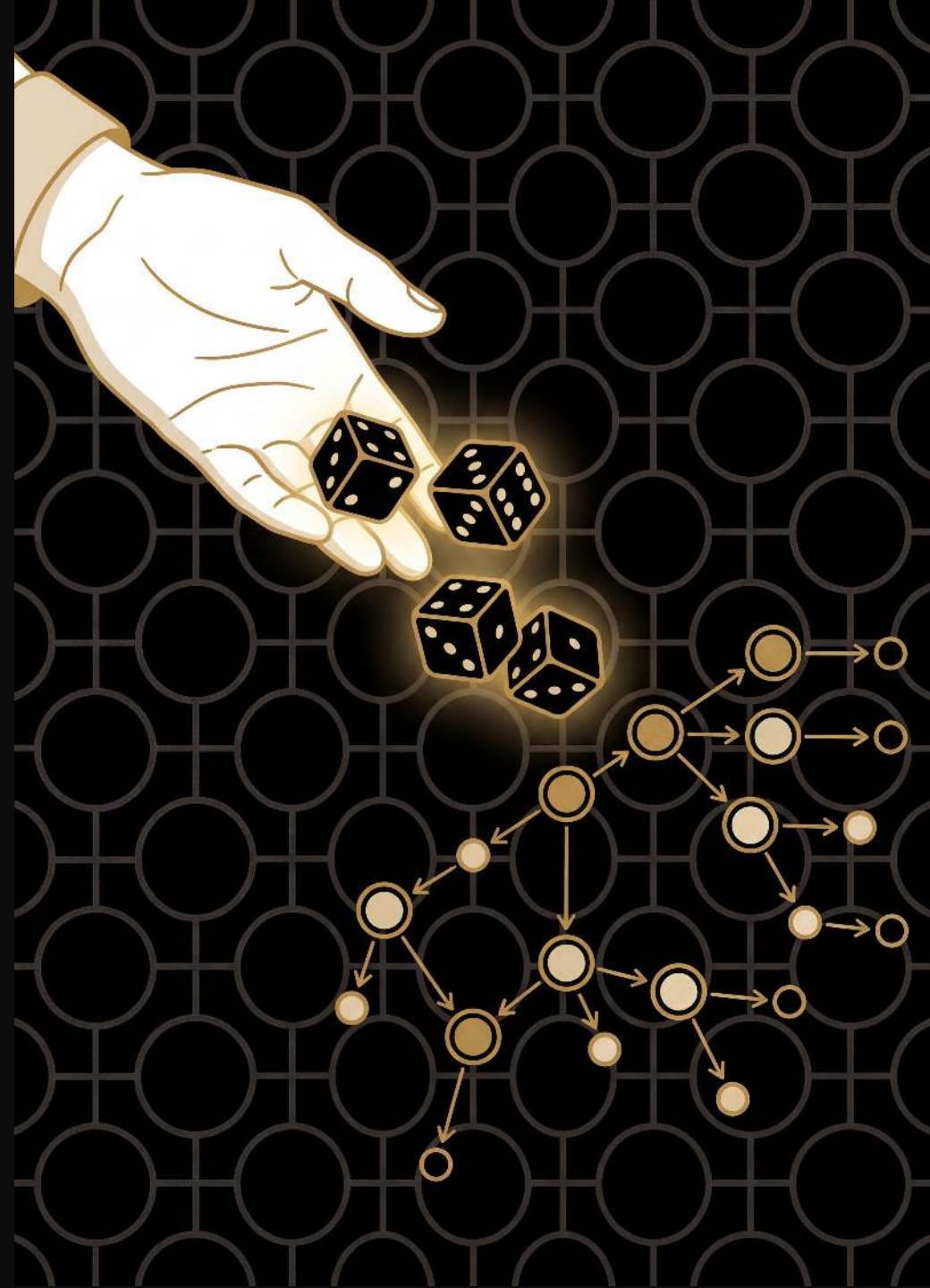
Say it once and move on: a token is one word or one punctuation mark. The role-mapping slide is where it sticks.

The shuffle is worth naming: in text order these pages would just be the book with boxes drawn round it, and you'd read it. Shuffled, no page says anything. Nobody here is holding a story.

Ten seconds to scan the sheet, then eyes up.

the algorithm

1. I **call out** the last token we wrote (and its colour)
2. **hands up** if it's the *boxed* token on your page
3. I **pick a hand at random** — you read out the token beside it
4. I **write it down**, and we go again



Walk it once, slowly. Reading it off the screen is fine.

Step 2: one hand each. More than one match? Go with the first one you spot.

Step 3 is the one to labour, briefly: I pick at random — not whoever's fastest, because that samples the quickest reader in the room rather than the text. Random is what makes the room behave like the model. Don't explain why yet; they'll see it.

Then the seed: someone near the front, sheet in hand, reads out any one pair, boxed word first. It matters that it came off a sheet rather than out of your head — the room supplies the first two words as well as every word after them. If their pair is an odd one (a full stop, a name that appears once), take it anyway: a thin first round is a real property of the model, not a failure. It can't dead-end on you, so you never need a second volunteer.



4:00

start reset +1 -1

Shift-W, write the volunteer's two words, call out the second of them, go. Four minutes, 12–15 tokens, write BIG for the back of the room. No commentary between hands going up and a word landing on the board.

Narrate two things, once each, without stopping the flow. The hands are the distribution: call a common word and a forest goes up; call a rare one and you get two. Nobody here has been taught weighted sampling and the room is doing it anyway — the number of hands is how often that continuation happens in the text. (Pedant's footnote, only if a pedant asks: someone holding several matches for one context raises one hand, which flattens the busiest words slightly.) And the choosing is a coin toss: when five hands are up, the four I don't pick were all real continuations. Different pick, different sentence. Same reason the "regenerate" button gives a different answer.

Hands will go up every round. Every distinct token appears as a boxed word somewhere, so there is no word you can call that has nothing following it — don't plan a "running dry" beat. The one way the room goes quiet is an empty seat: that sheet went home with someone, and its pairs are simply gone. Say exactly that, and reseed. The loud seeds for the corpus are on the brief (page 1 of the printed set); avoid restarting from a full stop, or the first thing the room does is argue about punctuation.

Stop while they still want more, around 15 tokens. Read the sentence back.



we just ran a language model

Say it flatly and let it sit. Not a simulation of one, not an analogy for one: the mechanism, at human speed, with the arithmetic small enough to see.

what role did you play?

each word on your sheet → a **token**

the pairs on your page → the **parameters** of the model

the hands in the air → the **probability distribution** over next tokens

my random pick → **sampling**

the sentence I wrote → **generation**



The jargon that gets used to make this stuff sound unreachable, and they've now got a physical referent for every bit of it. Go along the list at a fair clip — the point is the mapping, not a definition of each term.

Token: one word or one punctuation mark, which is why the full stops were out there getting picked like everything else. (Only if it comes up: a real model chops rarer words into pieces, which is why it can spell a word it has never seen.) Parameters and weights are the same thing under two names; you'll hear both. When a researcher says a model has a trillion parameters, they mean this: a very large pile of the thing on your lap. The number of pairs the room is holding is on the brief — say it.



the sentence we wrote is in no book



Nobody wrote it: no lookup in a big database of sentences, nobody pulled out their phone and searched the web. Everything was decided by the token pairs on our collective sheets.

So where did they come from?

three books

Each token pair on the sheets came from one of these three books. Click them in one at a time and read each out.

The sheets — the model — had no indicator of which token came from which book. It worked because there's token crossover between them: the, and, . are in all three, and the chain drifts from one book into another through them.

three books

“He was an old man who fished
alone in a skiff in the Gulf Stream.”

— *Ernest Hemingway, The Old Man and the Sea (1952)*

three books

“He was an old man who fished
alone in a skiff in the Gulf Stream.”

— *Ernest Hemingway, The Old Man and the Sea (1952)*

“TRUE! nervous, very, very dreadfully
nervous I had been and am.”

— *Edgar Allan Poe, The Tell-Tale Heart (1843)*

three books

“He was an old man who fished
alone in a skiff in the Gulf Stream.”

— *Ernest Hemingway, The Old Man and the Sea (1952)*

“TRUE! nervous, very, very dreadfully
nervous I had been and am.”

— *Edgar Allan Poe, The Tell-Tale Heart (1843)*

“The sun did not shine. It was too wet to play.”

— *Dr. Seuss, The Cat in the Hat (1957)*

a hall full of people just did
the thing you're told needs a data centre



Who now thinks they could give a better answer to a ten-year-old about how language models do what they do?

Obviously it still needs some simplification — but, crucially, it's not beyond them, and it's not beneath them to try to understand.



it doesn't make you an expert

it makes you harder to bullsh*t

It doesn't make you an expert: even with a billion dollars in your pocket you can't go out and build your own Claude-level model from scratch.

But it might make you, well — look at the slide.

That's what LLMs Unplugged is for: helping us all tell better, more accurate stories about how this technology works.



llmsunplugged.org/talks/demystifying-large-language-models/

Everything — this talk with its running order and what to print, the workshops, the printable materials, the tool that made your sheets — is at that address, free and CC-licensed. Hold here long enough for the room to scan it, then finish on the School mark.



Australian
National
University

School of
Cybernetics

www.llmsunplugged.org