



LLMs Unplugged

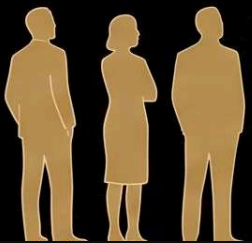
Does 'Unplugged' still matter in the age of AI?

Twenty minutes, and most of it is you doing the work rather than me talking. Introduce yourself briefly — and that you've spent the last couple of years teaching people how language models work with paper and dice.

Don't hand out the sheets yet. They get read too early otherwise, and the reveal is worth protecting.

Everyday life is increasingly regulated by complex technologies that most people neither understand nor believe they can do much to influence.

— Albert Bandura, 2001



Bandura wrote this in 2001, about technology in general, and it has aged into a description of this year. I want to point it at one technology in particular.

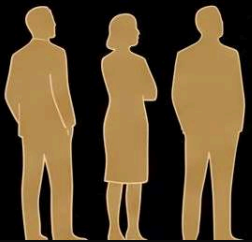
Read it out whole first, with nothing lit up. Then two clicks, because there are two failures in that sentence and they are not equal.

First click, neither understand: ordinary. I don't understand my car's transmission and I get by.

Second click, nor believe they can do much to influence: the expensive one. Having concluded, in advance, that this is not a thing you get a say in. That's the failure AI is currently manufacturing at scale, and it's the one this talk is aimed at.

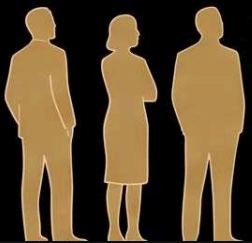
Everyday life is increasingly regulated by complex technologies that most people neither understand nor believe they can do much to influence.

— Albert Bandura, 2001



Everyday life is increasingly regulated by complex technologies that most people neither understand nor believe they can do much to influence.

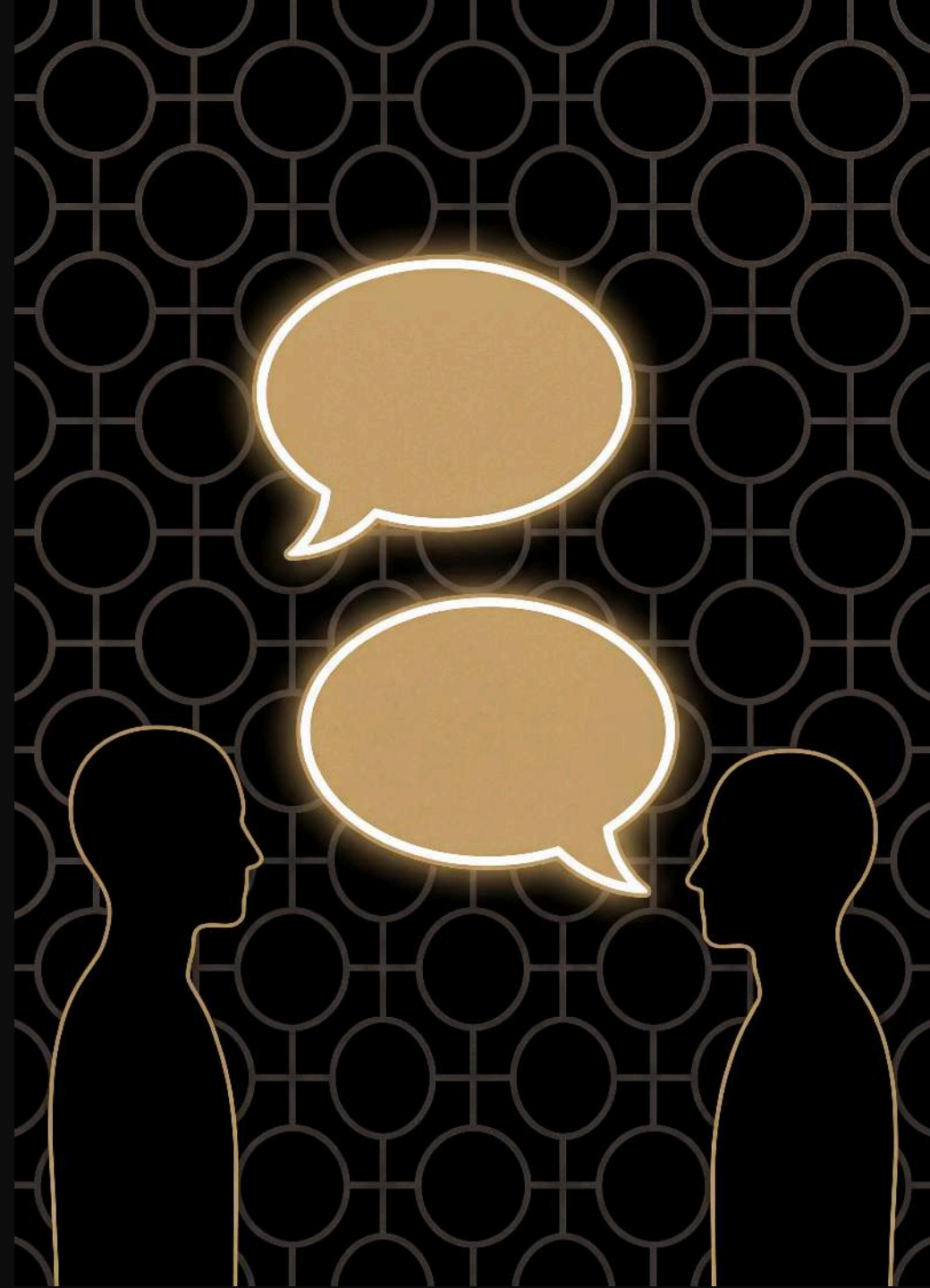
— Albert Bandura, 2001



what we get told about AI

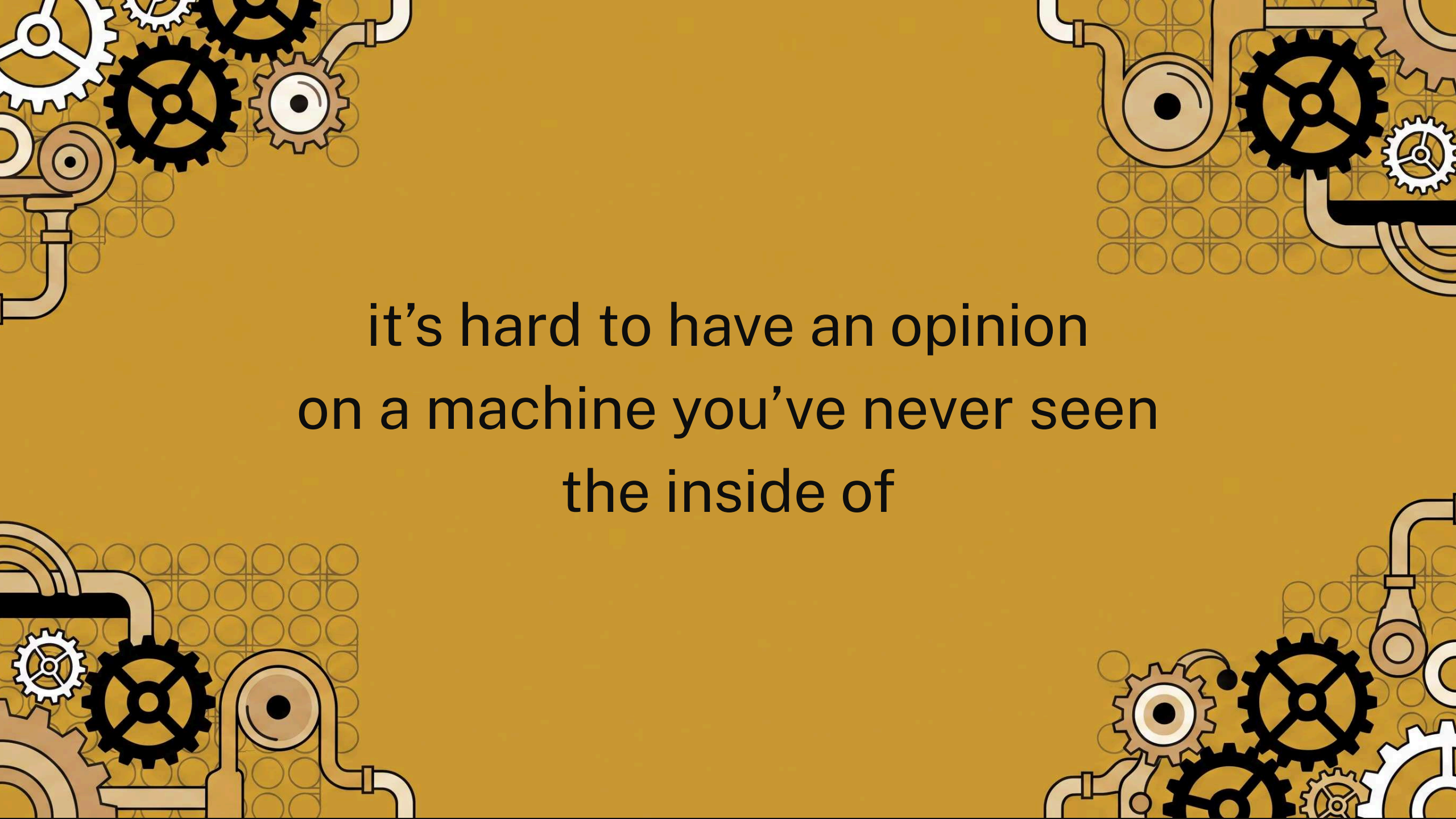
it **understands** you — it's basically a mind

it's **just autocomplete** — it's basically a parlour
trick



Both of these are on offer everywhere, usually from people who are confident. And notice they're opposites that do the same job: each one tells you there's nothing in the middle worth your attention. Either it's beyond you, or it's beneath you. Either way, don't look inside.

Beat. The honest answer is in between and it's more interesting than both, and — this is the bit people don't believe — it fits in a room like this one.



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

This is the whole argument of the talk, stated once, before the demonstration that earns it.

Then: so let's look inside. Not a diagram of the inside — the inside. Have the sheets handed out; once everyone has one, ask them to pick it up.



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.

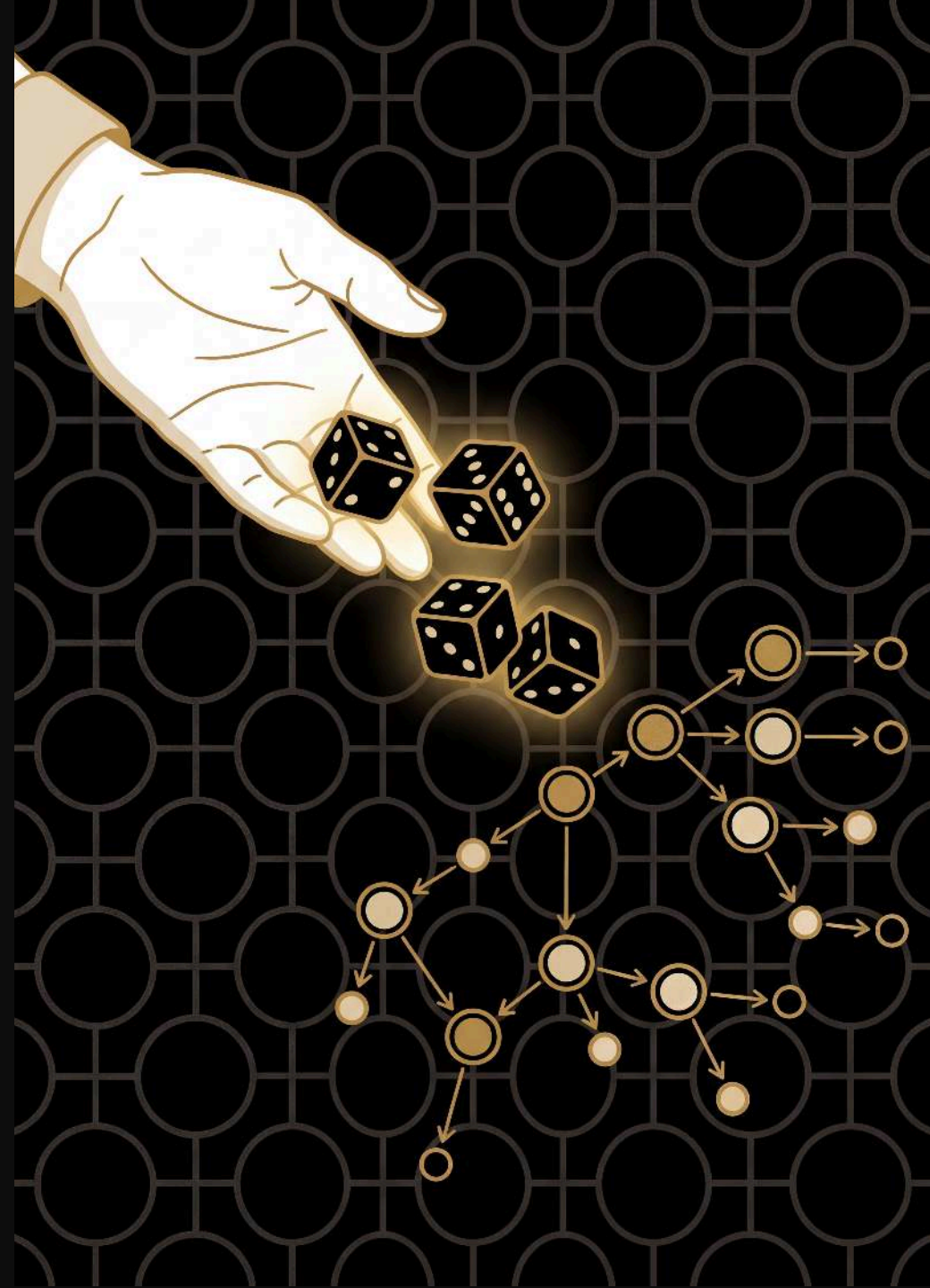
Say it once and move on: a token is one word or one punctuation mark. The role-mapping slide is where it sticks.

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 not in the book



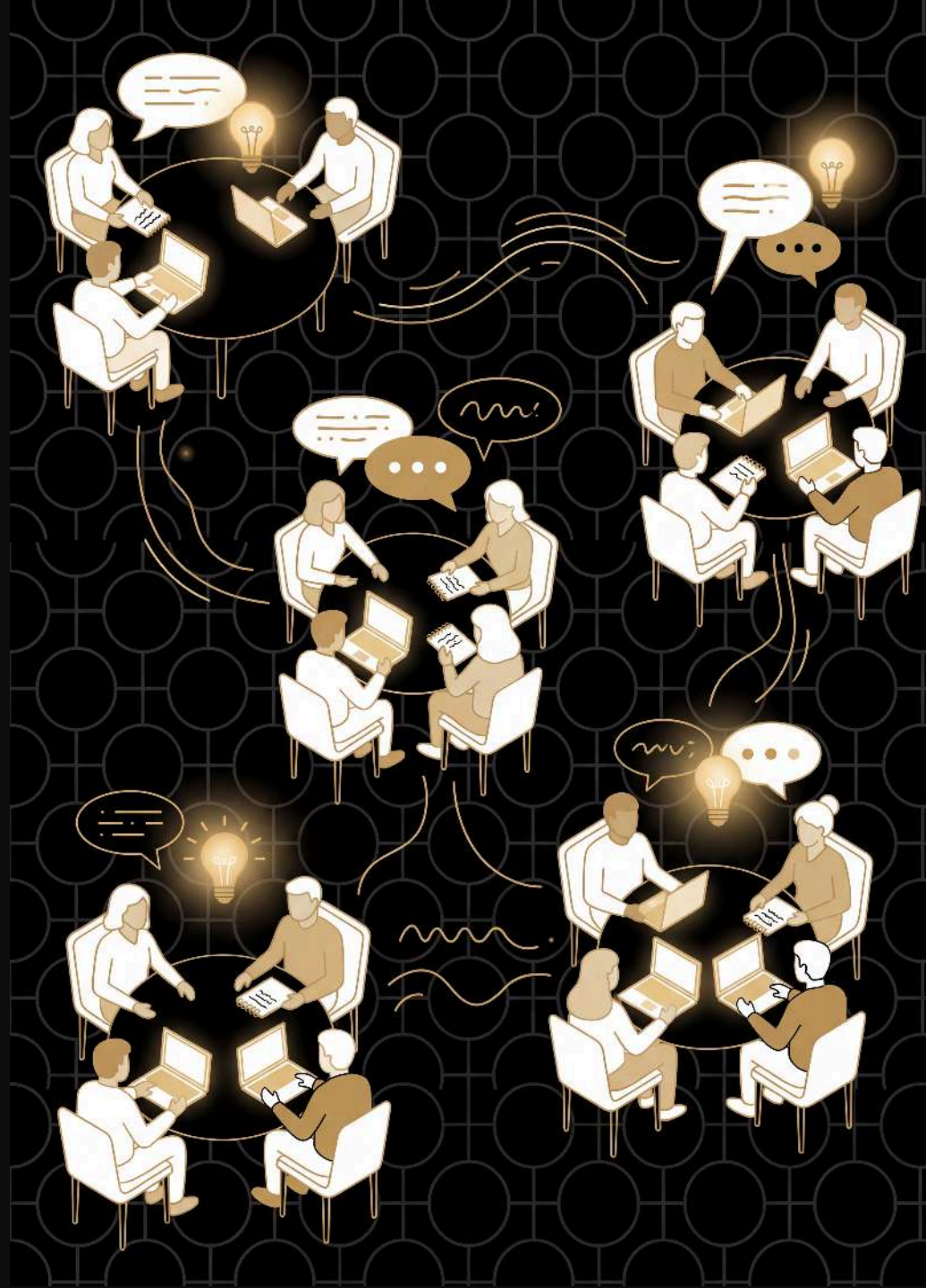
Check it if you like — unless we got very unlucky, that exact sentence isn't in The Cat in the Hat. Nobody wrote it. There was no lookup and no database of sentences. It came out of a pile of paper and a series of coin tosses.

That's the whole trick, and you have now watched it happen from the inside.

between us, we have the whole book

1,962 words · 240 distinct · 1,961 token pairs
dealt round the room

no single sheet has the whole picture



Every one of the 1,961 word-pairs in the book was dealt to exactly one person, and the deal deliberately spreads each context across as many people as it will go. So the model wasn't in my laptop and it wasn't on anyone's sheet — for the last five minutes it existed only across the whole room. If somebody had stayed home, part of it would be missing and some contexts would draw no hands at all.

That's not a metaphor for how a model is distributed. It's literally where the model was.

the same model in grid form

look at me ! look at me ! look at me now !

	look	at	me	!	now
look					
at					
me					
!					
now					

One line out of the book you're all holding — “Look at me! Look at me! Look at me now!” — small enough that the whole model fits on a slide. Every pair from your sheets is a tally mark in a cell: row is the token we're on, column is the token that could come next.

Point at the me row: two marks under !, one under now. That's three hands going up, two of them saying the same thing — the same distribution you were making with your arms five minutes ago.

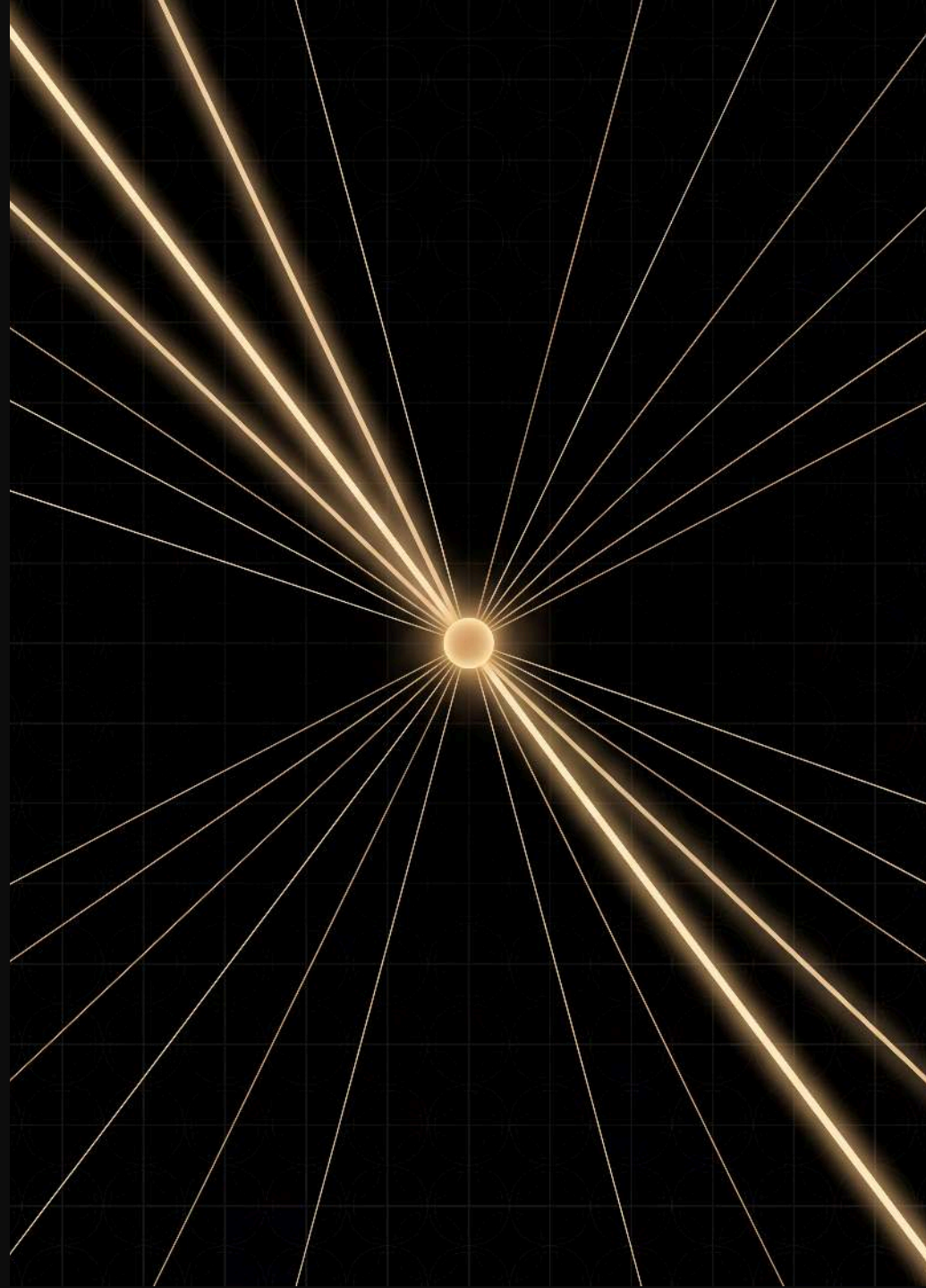
This is the form we use in the workshop, where each table builds their own grid from scratch. Same model, small enough to hold in one pair of hands.

three things Claude* does that we didn't

it looks at **more than one word** of context

it can say what it **never saw** — **cat** teaches it
about **dog**

it's been taught to **answer**, not just continue



The asterisk is yours to explain, in one line: Claude here means any frontier model — ChatGPT, Gemini, Llama. Nothing on this slide is specific to Anthropic's; I just need a name to point at.

Say each in a sentence, resist explaining any of them; each one has a full module on the website and none of them fits in this talk.

One word of context is why our sentence wandered — when I called the, nobody could tell whether we'd just written “sat in the” or “hit the”. Claude looks back over the whole conversation.

Our model rates any pair it never saw as flatly impossible, however sensible. A real model spreads what it learns across billions of numbers instead of separate cells, so what it learns about cat reaches dog.

And ours only ever continues text. Ask it a question and it continues the question — that's not it refusing, answering just isn't a thing it does. Turning that into something that answers you is another round of training, on conversations.

None of these changed the loop. Look at the context, weigh the options, pick one, write it down.

the line on the last slide



the whole room, as a grid

240 distinct tokens · 57,600 cells

5.8 m² of paper — roughly this stage

Read the caption. The little gold square is the slide you were just looking at; the big one is all of us. 240 distinct tokens, so 57,600 cells, and only about a thousand of them have a tally in them at all. The rest are empty — pairs the book never made, which the model treats as impossible.

At one centimetre per cell that's a square about two and a half metres on a side. Roughly this stage.

This is the model we made today. Now put its two kinds of scale beside a frontier LLM. *[next slide]*

today

frontier LLM

context

how much it looks at before each guess



1 token

the whole conversation

training text

how much the model reads

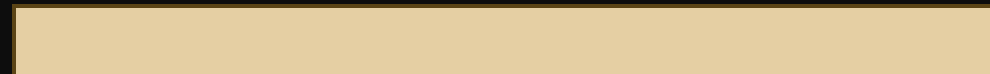


1,962 tokens

15 trillion tokens

model size

how many learned numbers it contains



57,600 grid cells

~1 trillion parameters

three independent scales · different kinds of model

Three independent scales grow. The first is how much the model looks at before each guess. Ours looked at exactly one token — which is why our sentence wandered — and Claude looks back over the whole conversation.

The second is how much text went past the counter. Ours read 1,962 tokens; fifteen trillion is the published figure for a recent open model’s training run. Every one of those tokens went past a counter, the way the book went past yours.

The third is how many learned numbers the model contains. Ours had a 240 by 240 grid: 57,600 places a count could live. A trillion parameters is the right order of magnitude for a frontier model, and the honest answer for the closed ones is “about that, and nobody outside will tell you”.

The bars have separate scales because these quantities can be changed independently. And the endpoints are different kinds of model: grid cells on the left, neural-network parameters on the right. This is a scale comparison, not a claim that one is a compressed version of the other.

much more context

much more training text

many more learned numbers

a different kind of model



Those are the differences to carry forward. Context, training text and learned numbers are all differences of scale, and none of them determines the others. What makes them worth scaling is the last line: spreading what it learns between related words — and looking usefully across a whole conversation — comes from different machinery, not merely from making our grid enormous.

No transformer lesson today. The important continuity is on the next slide.

still just **words in** → **words out**



Every reply from Claude or ChatGPT is produced one word at a time, by the loop you just ran with your hand in the air.

back to Bandura

most people neither understand nor **believe**
they can do much to influence

a room full of people just did, slowly, the thing
you're told needs a data centre



Read the clause off the slide again — it's the one we lit up at the start, and it's the one the last twenty minutes were aimed at.

This is why unplugged still matters, and I'd argue it matters more now than when CS Unplugged started.

The technologies Bandura was worried about in 2001 were at least visibly machines. This one talks back, and the people selling it have an interest in you believing it's either a mind or a magic trick. Both stories end with you having no view.

Twenty minutes of paper and hands in the air is the cheapest known cure for that.

it doesn't make you an expert
it makes you **harder to bullsh*t**



Be careful not to oversell. You can't build Claude now, and nothing here tells you whether it should be marking your students' essays.

What you can do is hear a claim about AI and have somewhere to stand — you know what kind of thing it is, which claims need more than confidence behind them, and which questions are worth asking. For a technology being sold this hard, that's not a small thing to hand a room.

And it's portable: everything here is free, printable, and needs no devices.



llmsunplugged.org/talks/unplugged-in-the-age-of-ai/

Everything — this talk with its running order and what to print, the workshops where you build your own model from scratch and take it home, 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