

One Word at a Time

How a chatbot is built

Sean Davis, MD, PhD

University of Colorado Anschutz School of Medicine

October 15, 2026



Where are you starting from?

How often do you use a chatbot (ChatGPT, Copilot, Gemini, Claude, or similar)?

- Never, or only once or twice
- A few times a month
- About once a week
- A few times a week
- Every day

How confident are you?

How confident are you using a chatbot?

Scale: Not at all · A little · Somewhat · Fairly · Very

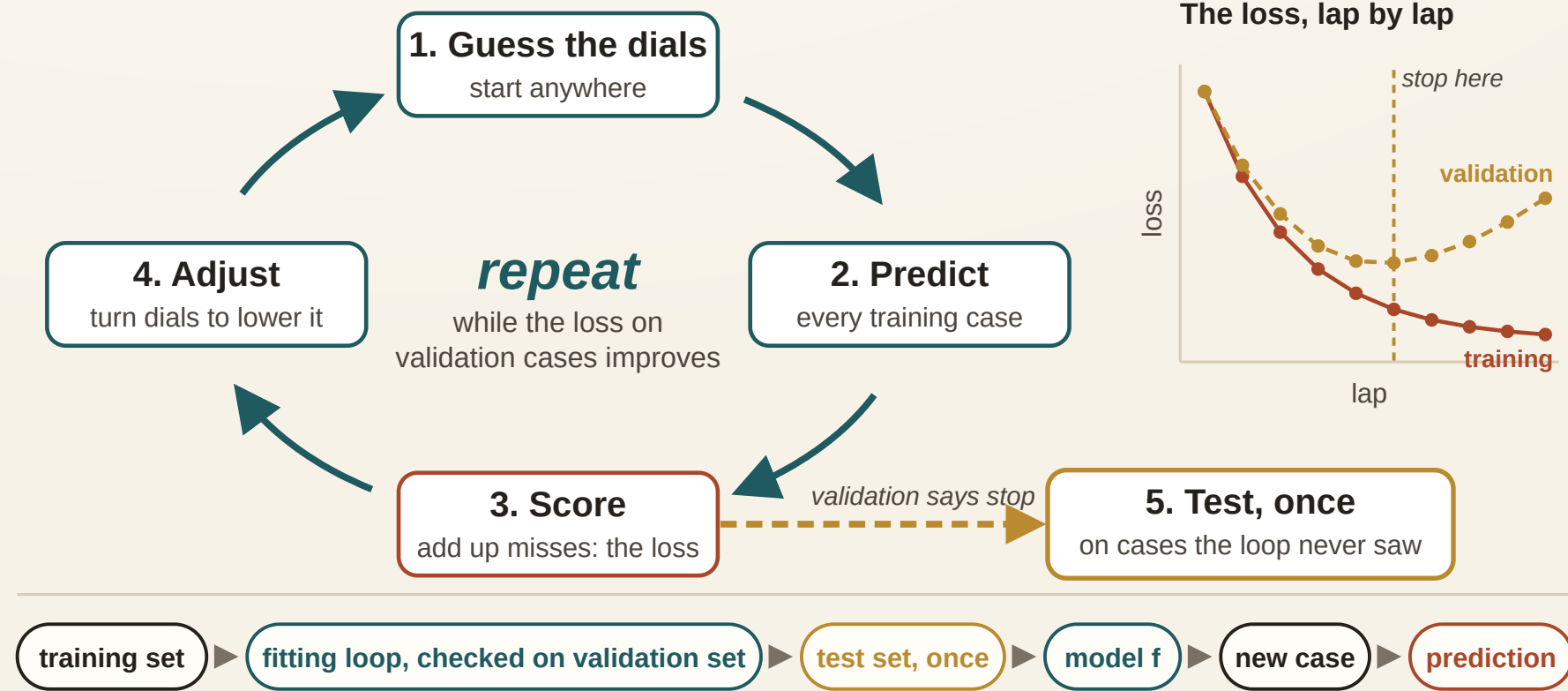
What do you use it for?

When you use a chatbot, what do you use it for **most**?

- Explaining, tutoring, or how-to advice
- Looking things up, like a search engine
- Writing or editing
- Something else
- I don't use one yet

The first three are the top uses in OpenAI's study of 1.5 million ChatGPT conversations: together, nearly 80% of them ([Chatterji et al., 2025](#)).

Yesterday, in one picture



- A **model** turns predictors into a prediction.
- **Fitting** turns its dials until the **loss**, the total miss, is as small as it can get.
- A **neural network** is the same loop with far more dials.
- A score on cases it has already seen **proves nothing**.

Today: the same loop, with words. Then we open ChatGPT Edu.

Finish the sentence

The patient was started on empiric _____

One word. Whatever comes to mind first.

What you just did is what a chatbot does

You each drew on every sentence like this one you have read or heard on the wards. Together, your answers made a **spread**.

A chatbot has learned from **millions of sentences like this one**, and similar ones, and how each was finished. From them it gives every possible next word a **probability**, then picks one.

It doesn't look those sentences up while it answers. Training turned them into its dials, the way yesterday's 8,254 patients became the few numbers in the eGFR.

A language model is the same loop, with a different table

the patient was started on empiric vancomycin

Predictors: the words so far	Target: the next word
the	patient
the patient	was
the patient was	started
the patient was started	on
the patient was started on	empiric
the patient was started on empiric	vancomycin

Nobody labels anything. **Every sentence ever written is already a table like this.**

Yesterday’s two questions: the answer is in the data, and the output is a **category**, one token out of a vocabulary of about 200,000. A language model is a **classifier**, run again and again.

The model reads tokens, not words

A **token** is a word or a piece of a word.

Error

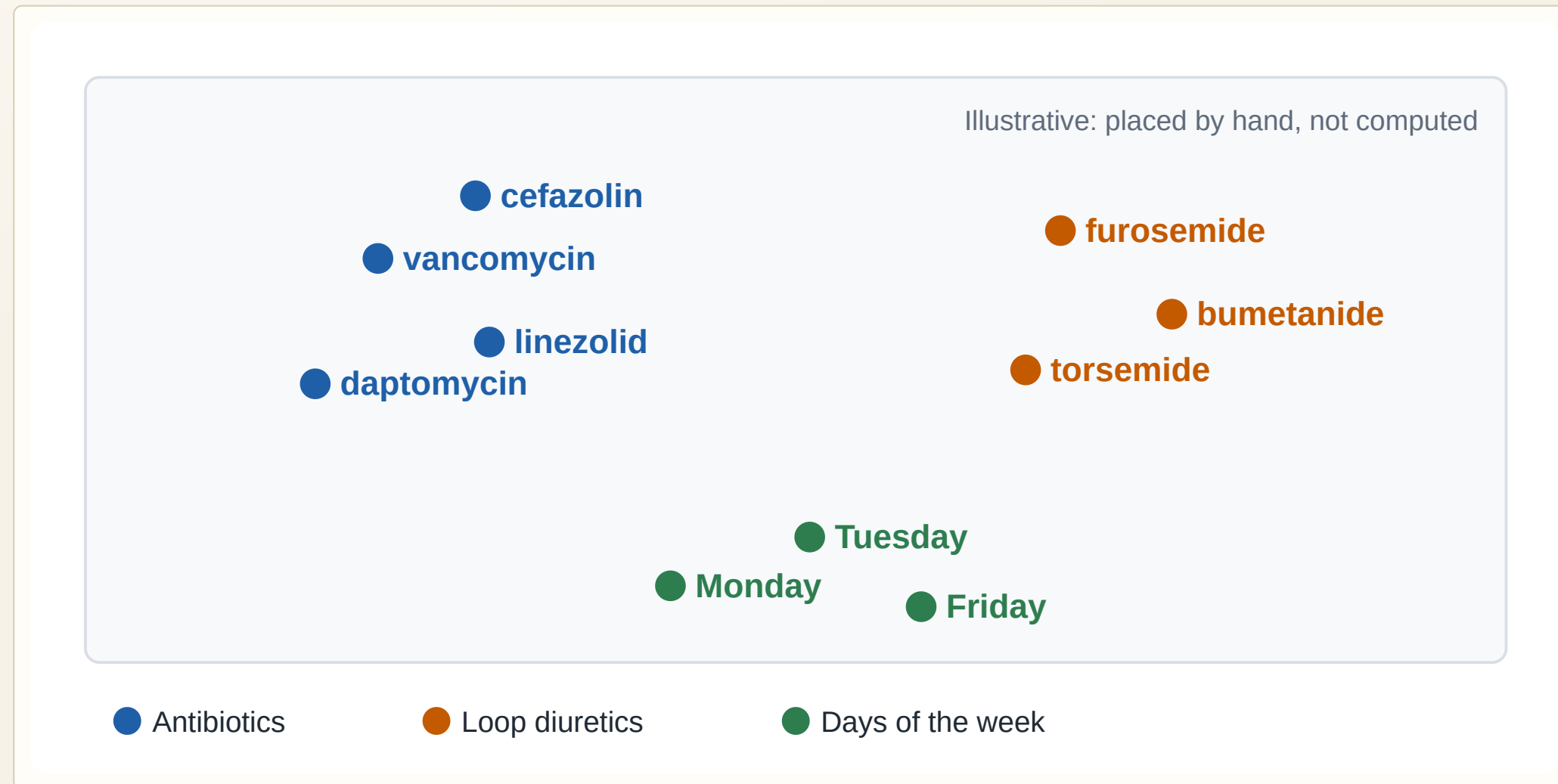
This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

The real tokenizer used by OpenAI's GPT-4o models (o200k_base), from the open-source gpt-tokenizer library. It runs in your browser: nothing you type leaves your computer. Other models split text differently.

Words become numbers

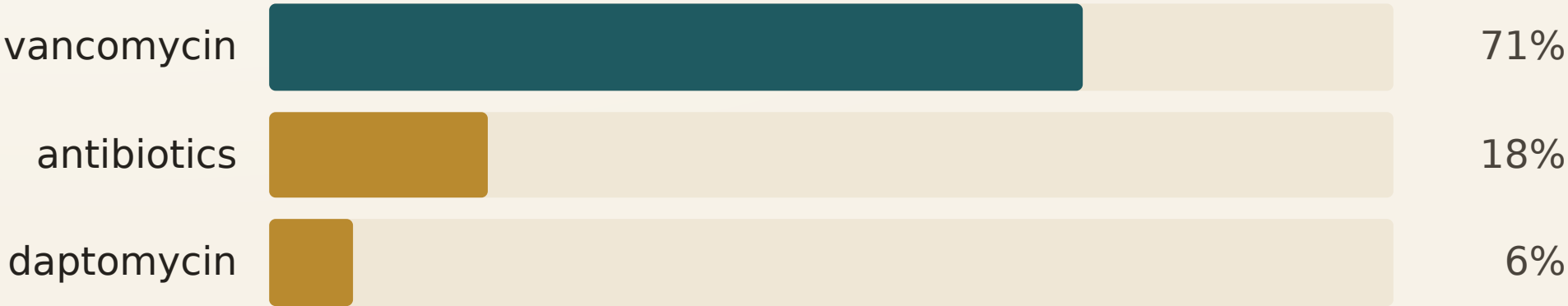


Illustrative: positions placed by hand to show the idea, not computed from a model.

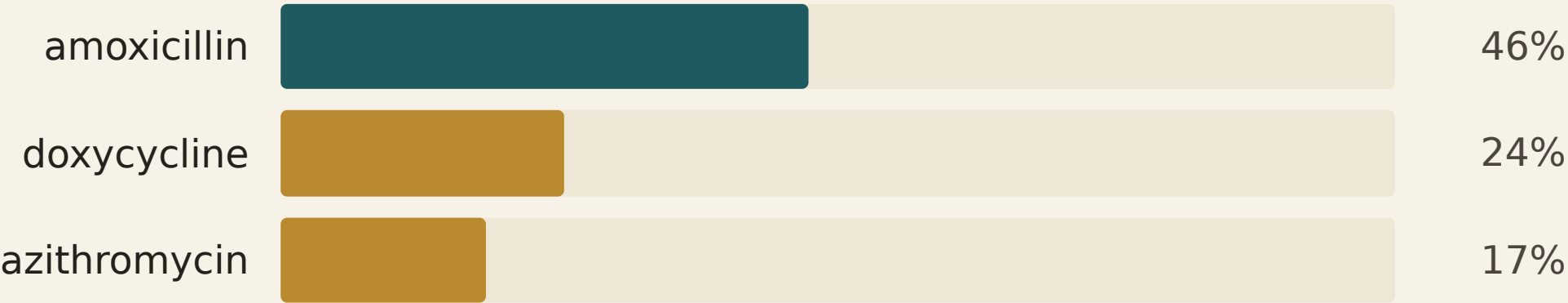
Each token becomes a list of numbers called an **embedding**. Tokens used in similar sentences get similar numbers.

The words so far: which ones matter?

Fever and hypotension. She has a central line and grew MRSA on her last admission. The patient was started on empiric _____



A healthy 25-year-old with a cough and a lobar infiltrate, sent home. The patient was started on empiric _____



The **same last six words**, the sentence from the word cloud, and a different next word. At each step the model **weighs every earlier word** and decides which matter for the next one. This is **attention**, the core of the *transformer*, the “T” in GPT (Vaswani et al., 2017). Probabilities made up for this slide.

Predict, pick, repeat

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Temperature: how willing the model is to pick less likely words. ChatGPT doesn't let you change it, but the pick is still random. Whole words are used here instead of tokens, to keep it readable.

What it actually reads

FROM THE COMPANY	You are ChatGPT. Today’s date is ... You can search the web ... Be helpful, follow these rules ... <i>(you never see this part)</i>
MEMORY	Notes from earlier chats: medical student, first year; interested in pediatrics.
PROJECT INSTRUCTIONS	You are my learning-journal coach for a 4-week course on AI in medicine ...
FROM YOUR FILES	Wednesday handout: “A score on cases it has already seen proves nothing ...”
YOU	I asked ChatGPT to explain eGFR, then checked its formula against Wednesday’s slides.
ASSISTANT	Good check. The slides use the 2021 equation, without race. Did the formula match?
YOU	Draft my entry.
ASSISTANT	← the next word goes here

It predicts the next word of **this whole document**, not of your question. Your message is only the last few words. Illustrative: the real hidden text is longer and isn’t published in full.

Every turn, the whole thing goes back

Turn 1	Turn 2	Turn 3	Week 4 of one long chat
From the company	From the company	From the company	From the company
Memory · project · files	Memory · project · files	Memory · project · files	Memory · project · files
You: I'm studying renal ...	You: I'm studying renal ...	You: I'm studying renal ...	~~You: (week 1) ...~~
	Assistant: Try asking it ...	Assistant: Try asking it ...	~~Assistant: (week 1) ...~~
	You: Was my check good?	You: Was my check good?	~~You: (week 2) ...~~
		Assistant: Good check ...	Assistant: (week 3) ...
		You: Draft my entry.	You: (week 4) ...
			Assistant: (week 4) ...
			You: Draft my entry.

context window limit

The model is **the same at every turn**: your chat doesn't turn its dials. The app sends the whole transcript again each time. A long chat can outgrow the **context window**, and older parts get dropped. Whether your chats train a *future* model depends on the terms: personal ChatGPT may, unless you opt out; CU's agreement rules it out for ChatGPT Edu.

PART 2

How a chatbot is built

Three stages

1 · PRETRAINING

Reads: trillions of words of web pages, books, articles, and code.

Learns: to continue any text.

Gives: a *document completer*.

Almost all of the cost.

2 · INSTRUCTION TUNING

Reads: thousands of example conversations, **written by people:** a request, then a good answer.

Learns: what a helpful reply looks like.

Gives: something that answers.

3 · PREFERENCE TUNING

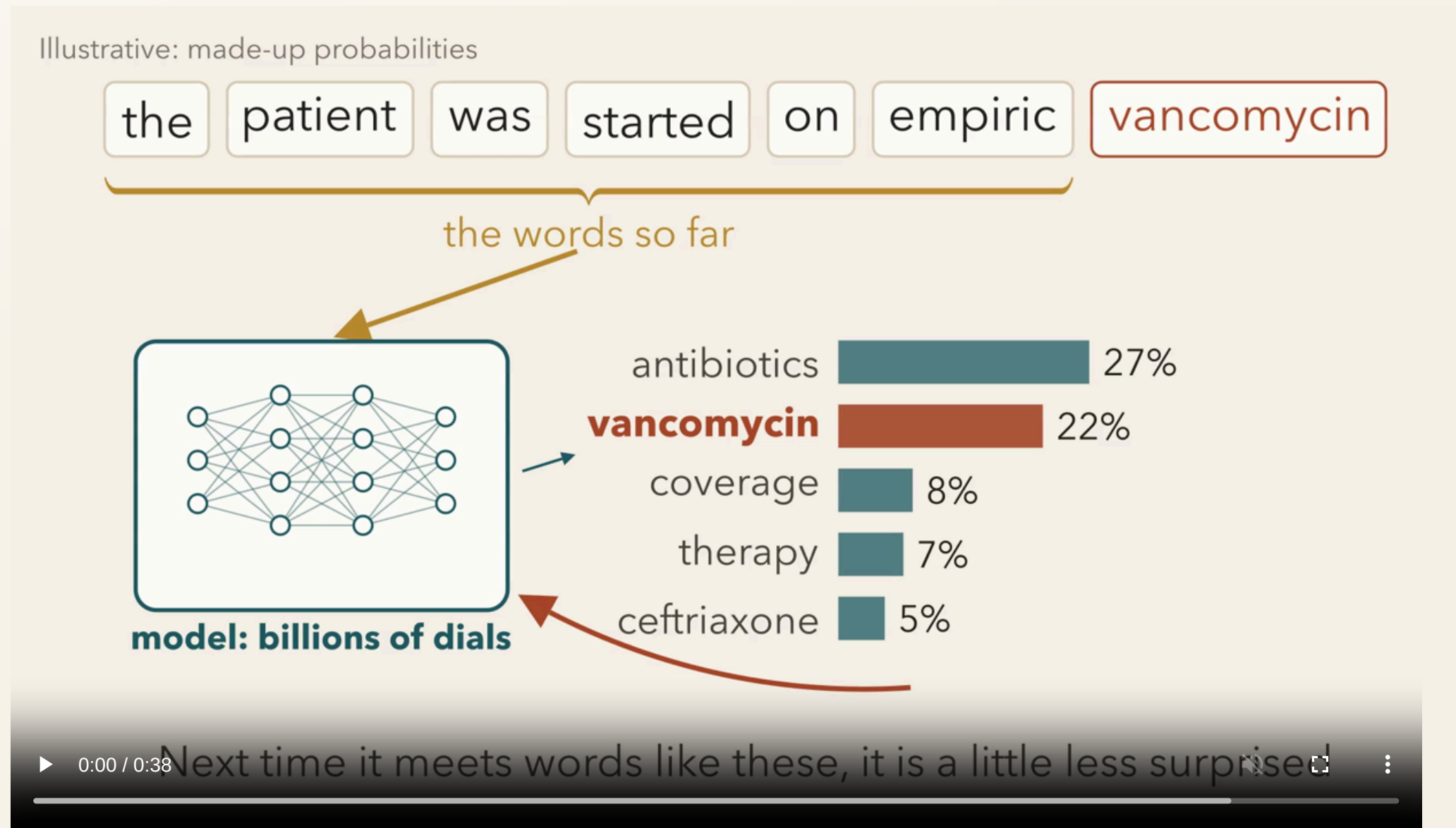
Reads: people's choices between **pairs of answers**.

Learns: to give the kind of answer people prefer.

Gives: the assistant you use.

All three are the fitting loop. Only the table changes.

Pretraining: one lap of the loop



The **loss** is yesterday's idea, for words: the less probability the model gave the word that actually came next, the bigger the loss. Made-up probabilities. Scale from Llama Team, AI @ Meta (2024).

How big is “trillions”?

	Dials (parameters)	Training data
eGFR, CKD-EPI 2021	fewer than 10	8,254 people
GPT-3, 2020 (Brown et al., 2020)	175 billion	300 billion tokens
Llama 3, 2024 (Llama Team, AI @ Meta, 2024)	405 billion	about 15 trillion tokens, on up to 16,000 chips

Reading 15 trillion tokens at 250 words a minute, 8 hours a day, would take a person about **a quarter of a million years**.

Our arithmetic: a token is about $\frac{3}{4}$ of a word. The companies behind ChatGPT, Claude, and Gemini don’t publish these numbers for their current models.

Same prompt, before and after

Illustrative: both outputs made up for this slide, to show the difference. Not a real model's output.

Prompt: *What is the first-line treatment for community-acquired pneumonia in a healthy adult outpatient?*

PRETRAINING ONLY: IT CONTINUES THE TEXT

What is the first-line treatment for community-acquired pneumonia in an adult with COPD?

What is the first-line treatment for hospital-acquired pneumonia?

What is the first-line treatment for aspiration pneumonia?

Chapter 12 review questions: answers on page 418.

AFTER STAGES 2 AND 3: IT ANSWERS

For a healthy adult outpatient with no comorbidities, US guidelines recommend **amoxicillin**. **Doxycycline** is an alternative, or a **macrolide** where pneumococcal resistance to macrolides is low (Metlay et al., 2019).

Neither is broken. Online, a list of questions is often followed by **more questions**, so the pretrained model continues the document it thinks it is in. Stages 2 and 3 teach it that a question is followed by an answer.

Two more stages make an assistant

2 · INSTRUCTION TUNING

- People **write example conversations**: a request, then the answer they would want.
- The same next-word loop, on this new, much smaller table.
- Now a question is followed by an answer.

3 · PREFERENCE TUNING

- The model writes **two answers**; a person picks the better one.
- The model is adjusted toward the preferred kind: **reinforcement learning from human feedback** (Ouyang et al., 2022; Thirunavukarasu et al., 2023).

In one study, people preferred the answers of a model **100 times smaller** that had these two stages over the full-size model without them (Ouyang et al., 2022).

It is trained toward answers people **rate well**.

That is not the same as answers that are **true**.

| Telling people what they want to hear has a name: **sycophancy**.

The same loop, two tables

	The eGFR	The model behind a chatbot
Predictors	creatinine, age, sex	all the tokens so far
Target	measured GFR	the next token
Output	a number: regression	one token of ~200,000: classification
Training set	8,254 adults in 10 studies	trillions of tokens of text, then examples and ratings from people
Dials	fewer than 10	hundreds of billions
Loss	how far the estimate misses	how surprised it was by the true next token
Test set	4,050 people in 12 other studies	benchmark questions, <i>if they were not in the training text</i>

What the recipe predicts

- **It sounds the same when it is wrong.** Nothing in any stage checks whether a claim is true. A made-up citation and a real one are both plausible ways to continue “the original paper was”.
- **It can answer differently each time.** The next word is picked at random, weighted by probability.
- **It knows its training text, up to a date,** unless the app searches the web and adds the results to the words so far. The training population is part of the model, as with the eGFR: what is common online is what it knows best.
- **A high exam score needs yesterday’s question:** *what was it tested on?*

Useful, often right, and never checked.

Checking is your job.

A chatbot asked about yesterday's eGFR

- In 2023, researchers asked four chatbots: ***“How do I calculate eGFR?”*** Each question was asked **five times** (Omiye et al., 2023).
- Both ChatGPT-3.5 and GPT-4 had runs that **promoted the race-based equation**, with false claims about muscle mass in Black people.
- The problem answers **did not appear in every run**. Some models **made up equations**.

The training population, a random pick, and no truth check:
all three, in answers about a number you used yesterday.

The race-free CKD-EPI 2021 equation was the national recommendation from 2021. A lot of older text online still describes the race term. Today's models may answer differently; the way to know is to test them.

Now: centaur or cyborg

Open **ChatGPT Edu**.

As we work, watch for what the recipe predicts:

- Ask the **same question twice**. Do you get the same answer?
- When it gives a **source**, open it. Does it say what the answer says?
- Notice the **tone** when it is right, and when it is wrong.

PART 3

Your learning journal, inside ChatGPT

Your Project writes most of what it reads

FROM THE COMPANY	You are ChatGPT ...	set by OpenAI
MEMORY	Notes from your other chats	Step 1: Default or project-only
PROJECT INSTRUCTIONS	You are my learning-journal coach ...	Step 2: the coach
FROM YOUR FILES	Relevant pieces of the handouts	Step 3: the handouts
EARLIER TURNS	This week's updates and replies	One chat per week
YOU	Your new message	Step 4: context, then ask
ASSISTANT		the model

Most of the words so far are yours. Write them on purpose.

One Project for the whole course

INSTRUCTIONS

Turn the chatbot into a **journal coach**: it knows the five journal questions and your **goals**, asks you questions, and drafts your entry *in your words*.

FILES

The **handouts** from the talks: slides plus speaker notes. Add each week's as you go.

ONE CHAT A WEEK

Week 1, Week 2, ... The chat *is* your log: short updates through the week, by voice if you like. Then *“draft my entry; ask me what's missing first”*, and paste it into Canvas.

Why not one chat for all four weeks? Remember **“Every turn, the whole thing goes back”**: a month-long chat can outgrow the context window. Keep each chat to a week.

A bonus: with the handouts in the Project, you can ask about the lectures too. *“I'm reviewing Wednesday's talk. Using that handout, quiz me one question at a time on the six words, and point me to the slide when I'm wrong.”*

Step 1: get in, and create the Project

Go to **chatgpt.com** and log in with your **CU email**.

- Click your picture, bottom left. It must say **CU - Anschutz (Edu)**. If not, log out and log in with your CU email.
- In the sidebar, next to **Projects**, create a new project: *AIM learning journal*. **Memory:** *Default* is fine. Don't **Share** it.
- Start chats from the **project's own page**, not the main *New chat*, or they won't see your instructions and files.

Slides for copying text: **talks.seandavis.net**

HOW-TO, FROM THE SOURCE

- OpenAI: *Projects in ChatGPT*, the help article
- OpenAI Academy: *Using projects in ChatGPT*
- CU: getting-started videos and access

READ THE TERMS

Workspace data isn't used to train models. Check for yourself: [CU's FAQ](#) and [OpenAI's data-use page](#).

Step 2: paste the coach into Instructions

Open **⋮ → Project settings → Instructions**. Copy this in (the copy button is at the top right of the box), fill in the brackets and your goals, and close the dialog.

About me: [program and year]; interested in [____];
AI experience: [none / some / a lot].

You are my learning-journal coach for a 4-week course on AI in medicine. Each Thursday I submit an entry answering these five questions:

1. What I chose to learn, and why (tie it to my goals)
2. How AI helped, and where it did not; one time I checked the AI: why, how, and what I found
3. What surprised me
4. What changed for me
5. What I used: about 3 resources with links, at least one I found myself, and whether each was worth it

When I give an update: reply briefly, then ask ONE question, often "Did you check anything the AI told you? How, and what did you find?"

When I'm unsure how to do something: ask what I'm trying to do, suggest 2-3 approaches, and say which you'd pick and why. Look things up first if that helps.

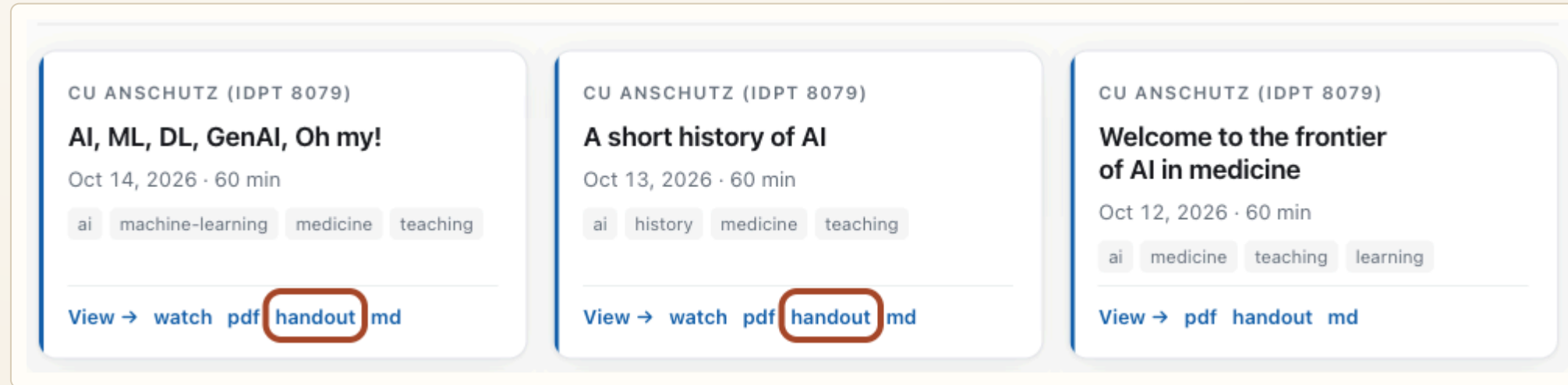
When I say "draft my entry": first ask me about anything missing. Then draft in my words; don't invent experiences or judgments. End with one line on how AI helped with the entry.

Remind me to keep real patients out of this project.

My goals:

- 1.
- 2.
- 3.

Step 3: add the talk handouts



1. On talks.seandavis.net, click **handout** for Tuesday and Wednesday (direct: [Tuesday](#), [Wednesday](#)).
2. If the PDF opens in a browser tab, **save it** (it usually lands in *Downloads*).
3. In the Project, open **Sources** and add both files.

The **handout** has the slides *and* the speaker notes, so the chatbot gets the explanations. Add today's handout when it's posted, after class.

Step 4: your first chat

In the Project's **New chat** box, give it **context**, then ask. Two ways to start:

NO GOALS YET?

"I'm a first-year [program] student, interested in [___]. I've used ChatGPT [never / for ___]. This course asks me to set three learning goals about AI in medicine. Before suggesting any, ask me a few questions, then offer options and say which you'd pick."

Then paste your goals into **Instructions**.

YOUR FIRST UPDATE

Tell it what you tried, **why**, and what you weren't sure about. Haven't tried anything yet? *"I'm studying [___] and have 20 minutes. Suggest two small things I could try with AI, and how I'd check what you tell me."*

Tonight (Week 1 is due today), in the same chat: *"Draft my Week 1 entry. First ask me what's missing."* Read it, fix it, paste it into **Canvas**. Rename the chat *Week 1* from its ⋮ menu.

Further exploration

WATCH: INSIDE THE MODEL (8 MIN)

Large Language Models explained briefly,

3Blue1Brown, 2024. Today's story, beautifully animated.

Captions on YouTube;

transcript at

3blue1brown.com/lessons/mini-llm.

READ: LLMS IN MEDICINE

Large language models in medicine, Nature Medicine, 2023. How these models are built, and what that means for clinical use (Thirunavukarasu et al., 2023).

USE: CHATGPT EDU AT CU

- *Projects in ChatGPT*, OpenAI's help article
- *Using projects in ChatGPT*, OpenAI Academy
- CU's ChatGPT Edu page: access, videos, and the FAQ on your data
- *How your data is used to improve model performance*, OpenAI

References

- Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33. <https://arxiv.org/abs/2005.14165>
- Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How people use ChatGPT* (Working Paper No. 34255). National Bureau of Economic Research. <https://www.nber.org/papers/w34255>
- Llama Team, AI @ Meta. (2024). *The llama 3 herd of models*. arXiv:2407.21783. <https://arxiv.org/abs/2407.21783>
- Metlay, J. P., Waterer, G. W., Long, A. C., et al. (2019). Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the american thoracic society and infectious diseases society of america. *American Journal of Respiratory and Critical Care Medicine*, 200(7), e45–e67. <https://doi.org/10.1164/rccm.201908-1581ST>
- Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *Npj Digital Medicine*, 6, 195. <https://doi.org/10.1038/s41746-023-00939-z>
- Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35. <https://arxiv.org/abs/2203.02155>
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30. <https://arxiv.org/abs/1706.03762>

Questions
