

Large language models

Predicting the next word, at scale

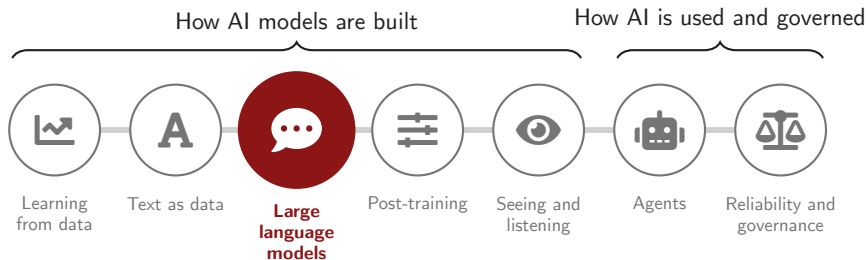
POLI 158: AI & Politics · Session 7

Yiqing Xu

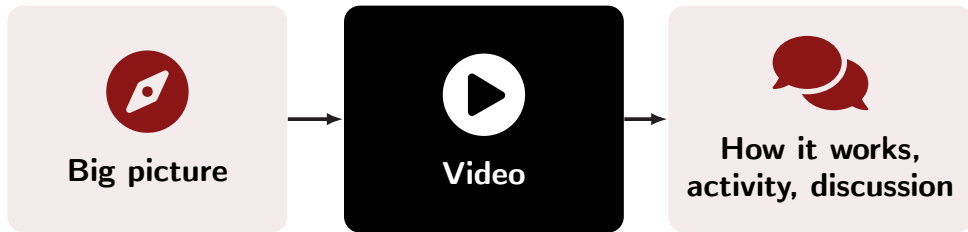
TODAY'S QUESTION

Can a system trained to predict the next word understand politics? If its answers are usually right, does it matter whether it understands?

The seven Wednesday sessions



Today in three parts



Where large language models are used

Everyday life



chatbots



writing help



coding help



answers in search

Research



labeling text at scale



summarizing documents



simulated survey answers

session 5

Politics and policy



writing reports

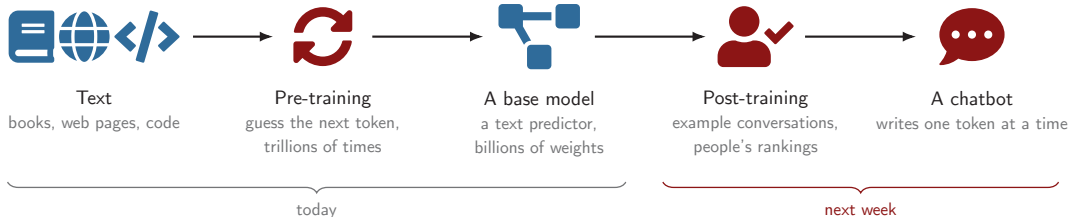


answering voters' questions

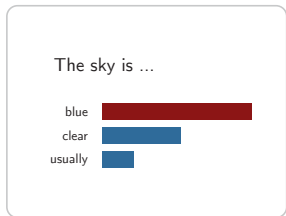


campaign messages

How a large language model is made

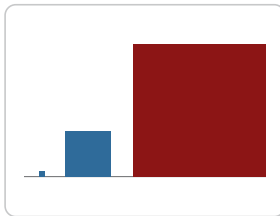


Three concepts in today's video



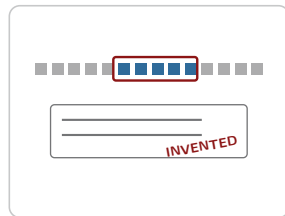
Next-token prediction

one token at a time



Scale

text, size, computing

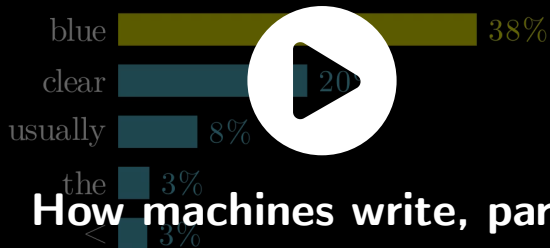


Context and hallucination

what it sees, what it makes up

1 Next-token prediction

The sky is ?

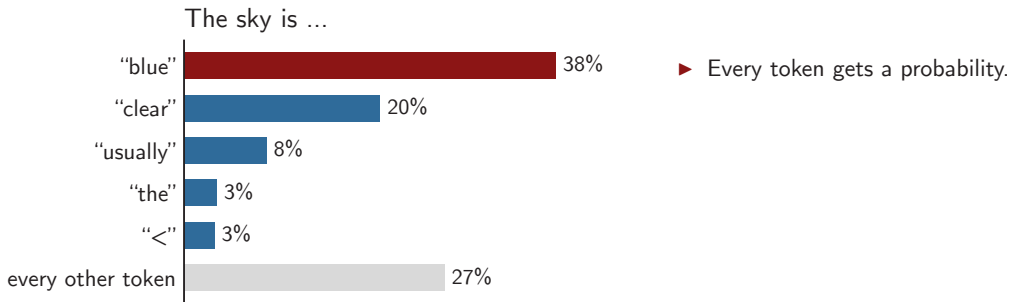


How machines write, part 1

an example from Surla (2023), NVIDIA

Key ideas: next token and scale

Next-token probabilities



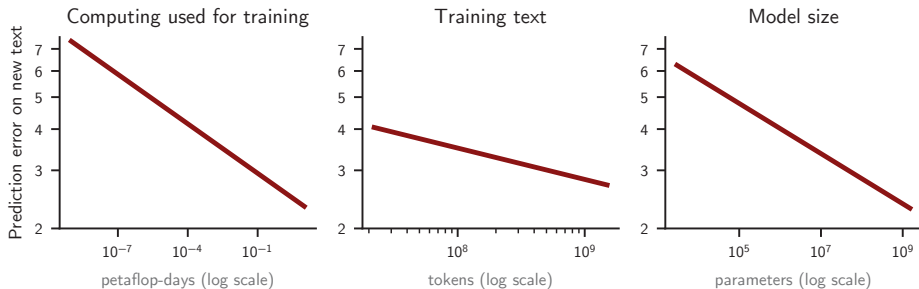
Source: Surla (2023), NVIDIA Technical Blog, Figure 1, as shown in the 2025 course slides. Percentages converted from the log-probabilities shown there.

Writing one token at a time

- Step 1 What is the capital of Germany? **The**
- Step 2 What is the capital of Germany? The **capital**
- Step 3 What is the capital of Germany? The capital **of**
- ...
- Step 6 What is the capital of Germany? The capital of Germany is **Berlin**

Source: the 2025 course slides.

Scaling laws

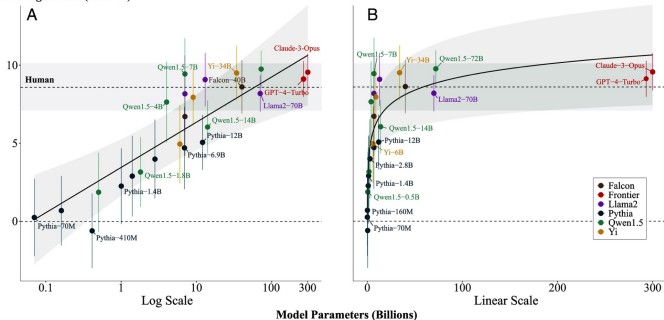


- Ten times more: error about 11% lower (computing), 16% (size), 20% (text).
- Steady enough to forecast: OpenAI predicted GPT-4's error from much smaller models.

Kaplan et al. (2020), Figure 1: the fitted lines only; each holds if the other two are not holding the model back. OpenAI (2023), GPT-4 Technical Report, Section 3.1.

Scaling and political persuasion

Estimated Persuasive Impact
in Percentage Points (95% CI)



- ▶ 24 models wrote messages on 10 U.S. issues; 25,982 people took part.
- ▶ Ten times larger: linked to about 3 more points of attitude change.
- ▶ On a plain scale (B), the gains flatten.
- ▶ Claude-3-Opus and GPT-4-Turbo: not significantly more persuasive than Qwen1.5-7B.

Hackenburg, Tappin, Röttger, Hale, Bright, and Margetts (2025), *PNAS*, Figure 1 (CC BY 4.0) and Table 1: 1.26 points per unit of log size, about 2.9 per tenfold. A: log scale; B: plain scale.

Closed and open-weight models



Closed

OpenAI's GPT, Anthropic's Claude, Google's Gemini, SpaceXAI's Grok (formerly xAI)

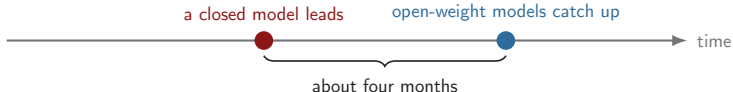
☰ weights not public; used through a provider



Open-weight

DeepSeek (China), Alibaba's Qwen (China), OpenAI's gpt-oss (U.S.)

👥 anyone can download and run it



Source: Epoch AI (2026), "Open models lag state-of-the-art closed models by 4 months"; DeepSeek-AI (2026), DeepSeek-V4 model card (MIT license).

Activity: watch a small model predict

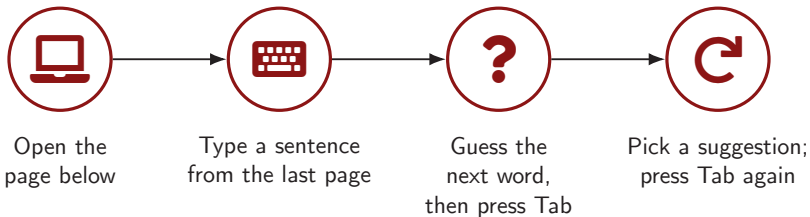
Skills behind next-word prediction

It needs	Text so far	Next word
Grammar	The senators _____ against the bill yesterday.	voted
Facts	Each state elects two _____ to the U.S. Senate.	senators
Reasoning	The bill passed 218 to 214, a margin of _____ votes.	four
Style	Dear Senator, I am writing to _____	urge, ask

Constructed examples. Two senators per state: U.S. Constitution, Article I, Section 3, and the Seventeenth Amendment.

Activity 1: watch GPT-2 write the next word

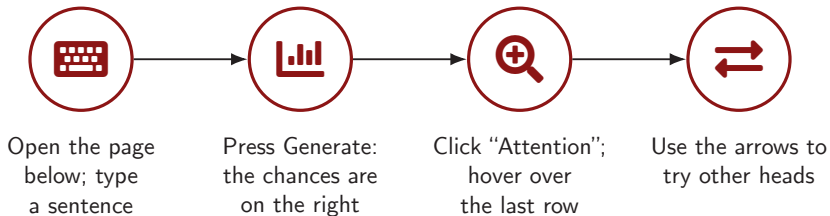
Activity laptops, in pairs



<https://banana-projects-transformer-autocomplete.hf.space/doc/gpt2-large>

Activity 2: look inside the model

Activity laptops, in pairs



<https://poloclub.github.io/transformer-explainer/>

What GPT-2 small predicted

Text so far	GPT-2's top three	Wanted
Each state elects two	representatives 58%, candidates 18%, senators 9%	senators
The bill passed 218 to 214, a margin of	just 36%, about 24%, victory 17%	four
Dear Senator, I am writing to	you 27%, express 27%, ask 24%	urge, ask

- The style came easily. The fact and the arithmetic did not.

Transformer Explainer, GPT-2 small, temperature 0.8, sampling from the top five; each chance is a share of those five. October 10, 2026.

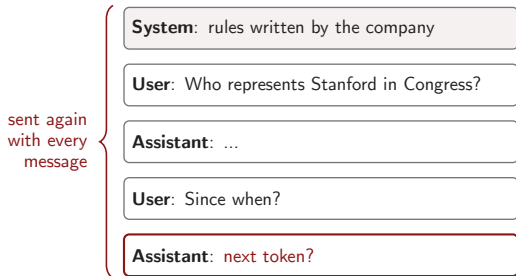
tokens in the window, exchange by exchange

example: about 500 tokens per question and answer



Key ideas: context and hallucination

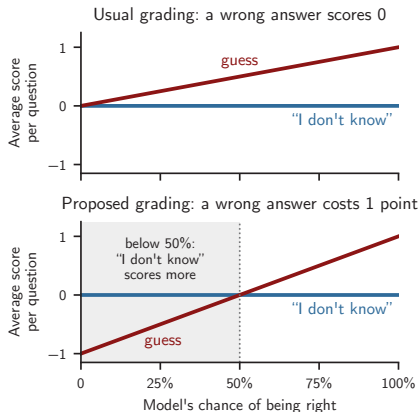
The context window



- ▶ The model keeps no memory: the whole chat is sent each time.
- ▶ 500 tokens per exchange: 5,000 after ten, 27,500 handled in all.
- ▶ GPT-3 (2020) could see 2,048 tokens. DeepSeek-V4 (2026): 1 million.

Source: Anthropic, "Context windows"; OpenAI, "Conversation state" (API documentation); Brown et al. (2020); DeepSeek-AI (2026), model card for DeepSeek-V4. The 500-token chat is an example.

Test scoring with a penalty for wrong answers



right: 1 point



"I don't know": 0



wrong: 0 now, -1 proposed

Source: Kalai, Nachum, Vempala, and Zhang (2025); OpenAI (2025), "Why language models hallucinate."

Hallucination and its consequences

Invented citations in official reports

- ▶ **United States, May 2025.** The White House “Make America Healthy Again” report: at least 7 of its 500-plus citations did not exist. Blamed on “formatting issues”; replaced the same day.
- ▶ **Australia, 2025.** A Deloitte report for the employment department: a made-up quote from a court ruling, and papers that do not exist. Deloitte refunded more than A\$97,000.
- ▶ **South Africa, April 2026.** The Draft National Artificial Intelligence Policy had “various fictitious sources in its reference list.” The minister withdrew it.

Source: NOTUS (May 29, 2025); CFO Dive (October 21, 2025), citing the department and the Associated Press; SAnews, South African Government News Agency (April 2026).

Generated images, voices, and video



text



images



a person's voice



video, now with sound



- ▶ January 2024: robocalls in a cloned Biden voice told New Hampshire voters to skip the primary.
- ▶ The Federal Communications Commission (FCC) fined the political consultant behind the calls \$6 million.

Source: FCC press release, September 26, 2024; Google DeepMind, SynthID and Veo pages (accessed September 2026).

Two views of large language models



A stochastic parrot

A language model stitches together “sequences of linguistic forms” it has seen, “without any reference to meaning.”

Stochastic means based on chance. A parrot repeats words without knowing what they mean.

Bender, Gebru, McMillan-Major, and Shmitchell (2021), *FAccT*.



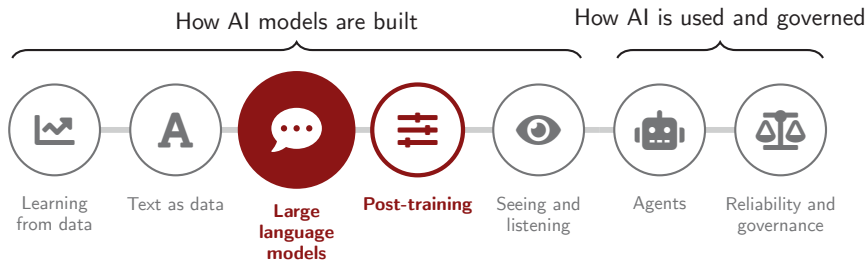
Sparks of general intelligence

After testing an early GPT-4: it “could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system.”

General intelligence means broad abilities across many kinds of tasks.

Bubeck et al. (2023), arXiv:2303.12712.

Next week



Next Wednesday, October 21

Post-training and reasoning models.

Sources

- Anthropic (2026). "Context windows." Claude API documentation.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT*: 610–623.
- Brown, T. B., et al. (2020). "Language Models are Few-Shot Learners." *NeurIPS* 33.
- Bubeck, S., et al. (2023). "Sparks of Artificial General Intelligence: Early experiments with GPT-4." arXiv:2303.12712.
- CFO Dive (2025). "Deloitte refunds over \$60K for report with AI errors, Australian government says." October 21.
- Cho, A., et al. (2024). "Transformer Explainer: Learning LLM Transformers with Interactive Visual Explanation and Experimentation." arXiv:2408.04619.
poloclub.github.io/transformer-explainer.
- DeepSeek-AI (2024). "DeepSeek-V3 Technical Report." arXiv:2412.19437. (2026) DeepSeek-V4 model card, Hugging Face.
- Edwards, J., and Emberson, L. (2026). "Open models lag state-of-the-art closed models by 4 months." Epoch AI, May 29.
- Federal Communications Commission (2024). "FCC fines man behind election interference scheme \$6 million." September 26.
- Hackenberg, K., Tappin, B. M., Röttger, P., Hale, S. A., Bright, J., and Margetts, H. (2025). "Scaling language model size yields diminishing returns for single-message political persuasion." *PNAS* 122(10): e2413443122.
- Hugging Face (2019). Write With Transformer (GPT-2 large).
banana-projects-transformer-autocomplete.hf.space.
- Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. (2025). "Why Language Models Hallucinate." arXiv:2509.04664. OpenAI blog post, September 5.
- Kaplan, J., et al. (2020). "Scaling Laws for Neural Language Models." arXiv:2001.08361.
- Kennard, E., and Manto, M. (2025). "The MAHA Report Cites Studies That Don't Exist." NOTUS, May 29.
- Manto, M. (2025). "The MAHA Report Has Been Updated to Replace Citations That Didn't Exist." NOTUS, May 29.
- OpenAI (2023). "GPT-4 Technical Report." arXiv:2303.08774. OpenAI (2026). "Conversation state." API documentation.
- Ouyang, L., et al. (2022). "Training language models to follow instructions with human feedback." *NeurIPS* 35.
- Radford, A., et al. (2019). "Language Models are Unsupervised Multitask Learners." OpenAI (GPT-2).
- SAnews (2026). "Minister announces withdrawal of draft AI Policy." South African Government News Agency, April.
- Stanford HAI (2024). *The 2024 AI Index Report*.
- Surla, A. (2023). "How to Get Better Outputs from Your Large Language Model." NVIDIA Technical Blog, June 14.
- Video: each source is named on screen. Prep: 3Blue1Brown (2024), "Large Language Models explained briefly" (YouTube); OpenAI (2025), "Why language models hallucinate." Icons: Font Awesome Free 5 (CC BY 4.0). Video: Manim Community Edition; narration by Kokoro-82M, a synthetic voice. The citation in the video is invented and labeled as such.