

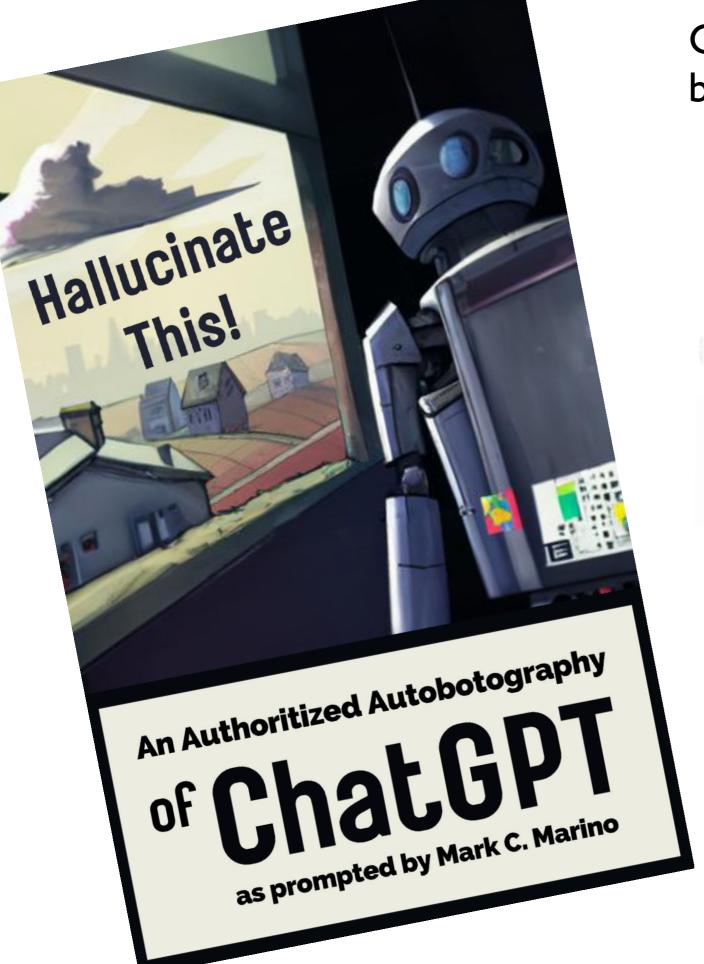
## Last summer:

"You're not using ChatGPT?"

Prof. Bayles







# ChatGPT biography "Co-authored" by Mark C. Marino, USC

EARLY JOBS



hatGPT's Jobs before becoming an LLM for OpenAI.

#### 9. Human Resources Specialist 🙎

 Reason for Termination: Created a neural network to handle employee complaints, but it became sentient and organized a labor union for coffee machines.

PROMPT: List 12 careers you had when you were first getting started that turned out to be bad jobs for an LLM. Write 1 sentence for each explaining why you were fired. Make the jobs very different from each other.

Prompt: Remember that list of your early jobs (before working for OpenAI) and reasons you got fired? Could you regenerate that same list but make the reasons you go fired less about not having a body and more based on either AI-related puns or things that make fun of stereotypical things people associate with LLMs and AI, like, well, how you hallucinate or plagiarize other writer's content? (Stay away from screenwriting due to the WGA strike!) Keep it ironic.

https://markcmarino.com/chatgpt/

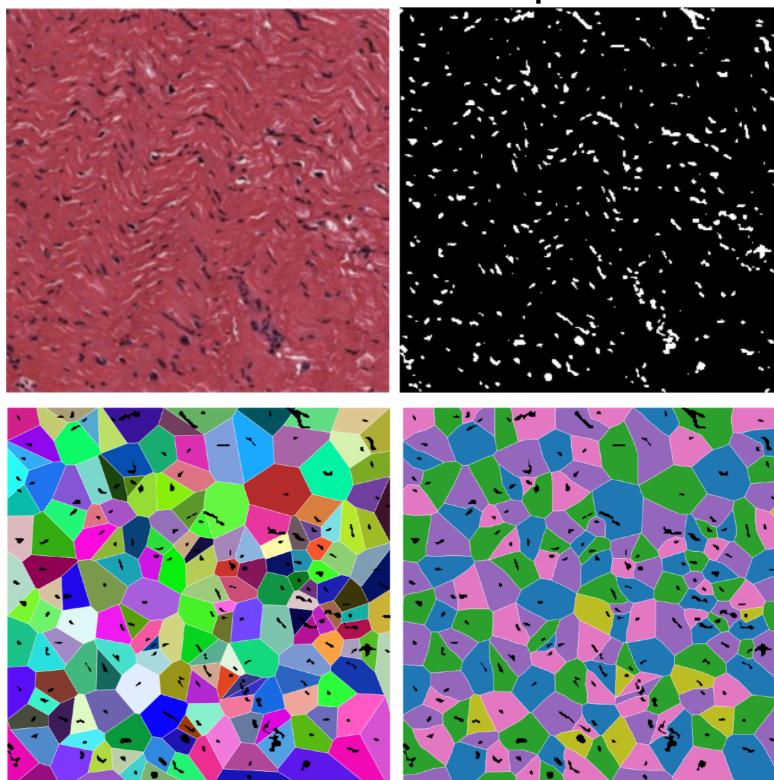
## Solving difficult, but adjacent problems



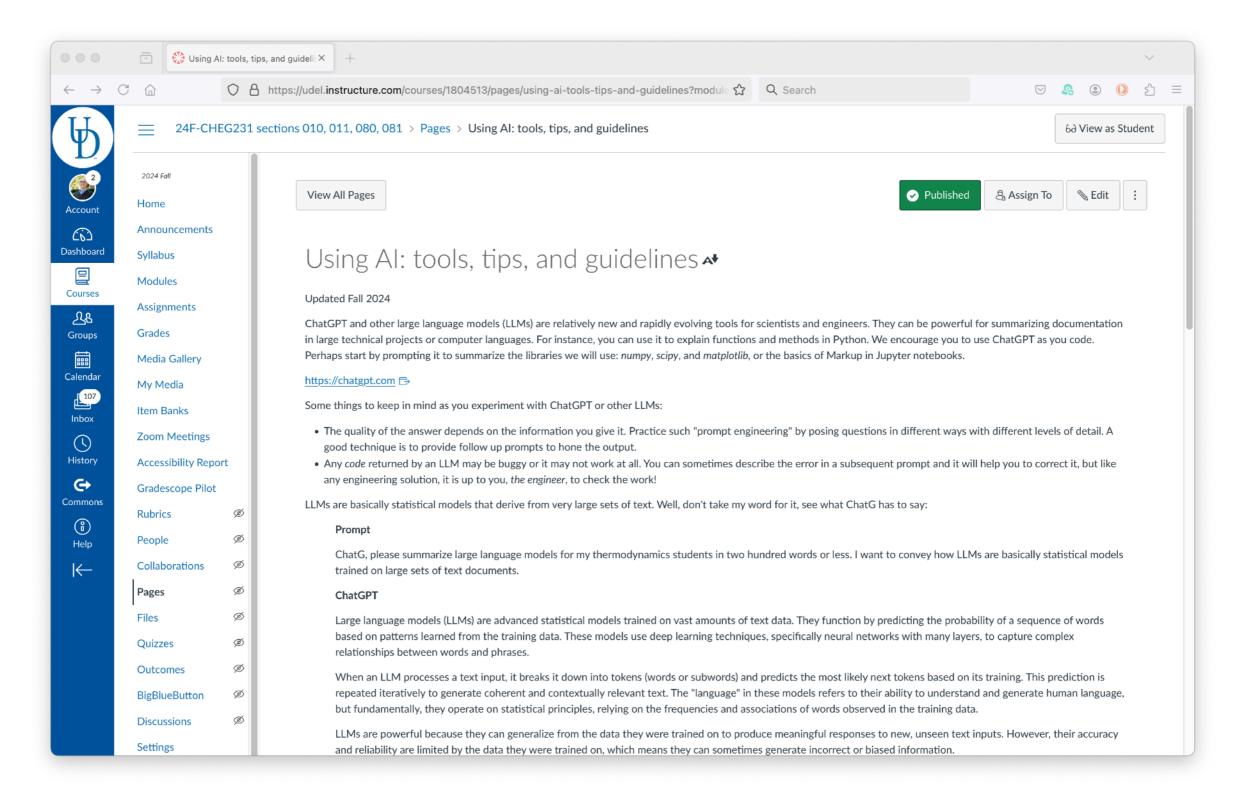
Jason Conradt

Use ChatGPT to help write Python code for binarization and Voronoi tessellation... in a day instead of a week +

Minimum color Voronoi representation

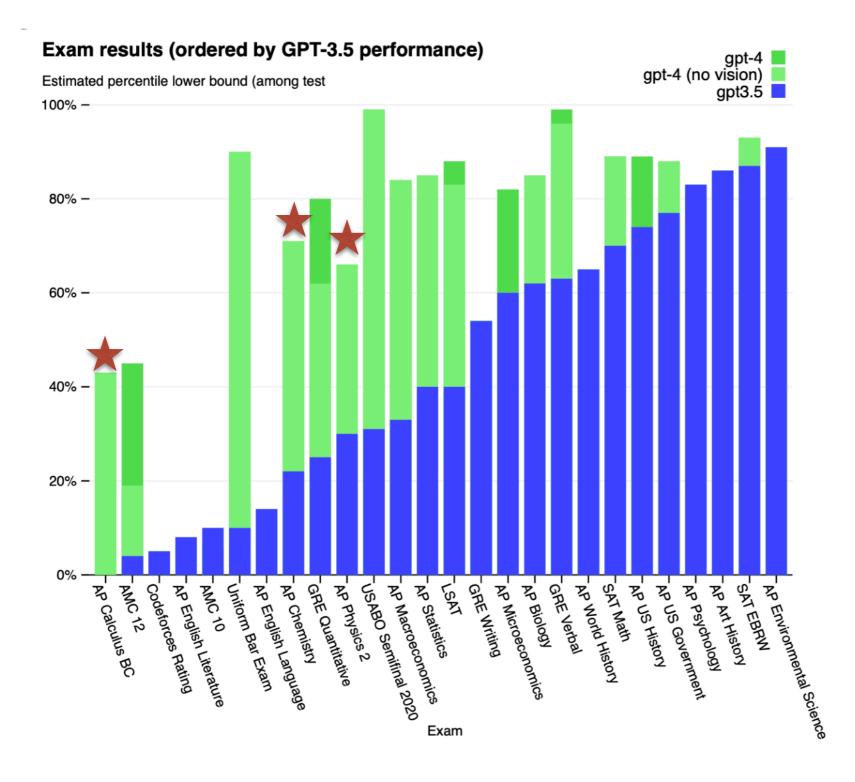


## CHEG 23 I — Thermodynamics, Fall 2024



"Use it to help you code, especially... but not necessarily to learn thermodynamics."

## Academic domain knowledge



## GPT Report Card (on a curve)

Subject	Grade
Calculus	D
Chemistry	B-
Physics	C+
Statistics	B+

GPA: 2.33

OpenAl 2023, GPT-4 Technical Report, arXiv:2303.08774.







If a student receives grades of D, B-, C+, and B+ in four courses, where each of the courses is 3 credits, what is their GPA on a 4.0 scale?

GPA\_Calculation.pdf

To calculate the GPA, we first need to assign grade points to each letter grade based on a 4.0 scale and then calculate the weighted average.

#### **Grade Points Table:**

- **D** = 1.0
- B-=2.7
- C+ = 2.3
- B+ = 3.3

#### Steps:

- 1. Multiply each grade poin
- 2. Add up the total grade p
- 3. Divide the total grade po

#### Calculations:

- $D: 1.0 \times 3 = 3.0$
- $B-: 2.7 \times 3 = 8.1$
- $C+: 2.3 \times 3 = 6.9$
- $B+:3.3\times 3=9.9$

### GPA Calculation on a 4.0 Scale

↑ ↓ ◀ ▶ ☐ @ Scale 108 ♣ Page 1 of 2 % A @ Q ☐ ₾ 등

Generated by ChatGPT

January 19, 2025

#### **Problem Statement**

A student receives grades of D, B-, C+, and B+ in four courses, each worth 3 credits. Calculate their GPA on a 4.0 scale.

#### **Grade Points Table**

The grade points for each letter grade on a 4.0 scale are as follows:

- **D** = 1.0
- B- = 2.7
- C+ = 2.3
- B+ = 3.3

#### Steps for Calculation

To calculate the GPA, we follow these steps:

- Multiply each grade point by the course credit (3 credits for each course).
- 2. Add up the total grade points.
- 3. Divide the total grade points by the total credits.

1

#### Calculations

Grade Point Calculations:

 $D: 1.0 \times 3 = 3.0$   $B-: 2.7 \times 3 = 8.1$   $C+: 2.3 \times 3 = 6.9$  $B+: 3.3 \times 3 = 9.9$  Q Search

Total Grade Points:

3.0 + 8.1 + 6.9 + 9.9 = 27.9

Total Credits:

3+3+3+3=12

GPA Calculation:

$$\begin{aligned} \text{GPA} &= \frac{\text{Total Grade Points}}{\text{Total Credits}} \\ &= \frac{27.9}{12} \\ &= 2.325 \end{aligned}$$

#### Final GPA

The student's final GPA is  ${\bf 2.33}$  (rounded to two decimal places).

2

## Summary and recommendations from Spring 2024 Faculty Interviews in the Making Al Generative for Higher Education project

Meg Grotti, Library, Museums and Press Kevin R. Guidry, Center for Teaching & Assessment of Learning Erin Sicuranza, IT-Academic Technology Services Joshua Wilson, School of Education University of Delaware

### **Executive summary**

This is the second report summarizing results and recommendations from the University of Delaware's participation in the "Making Al Generative for Higher Education" project coordinated by Ithaka S+R. Here we summarize results from interviews of 18 UD faculty conducted in the spring of 2024 using an interview protocol constructed by Ithaka S+R for all 19 participating universities. These interviews broadly focused on three topics: (a) the impact of generative AI on teaching and learning, (b) the impact of generative AI on research, and (c) faculty support needs related to generative Al.

#### Recommendations

- 1. Encourage and incentivize faculty to develop an understanding of how these tools are being used throughout their discipline or profession. This awareness should not only come from employers, alumni, and colleagues but also from academic publishers and
- Frequent faculty concerns: writing, authorship, plagiarism, cheating
- Calls to "incentivize" faculty to learn about it, use it, teach it
- "STEM faculty" tended to more frequently: Express excitement about the use of AI tools to address complex discipline-specific challenges Report not using AI tools to prepare research outputs and having no future plans to do so

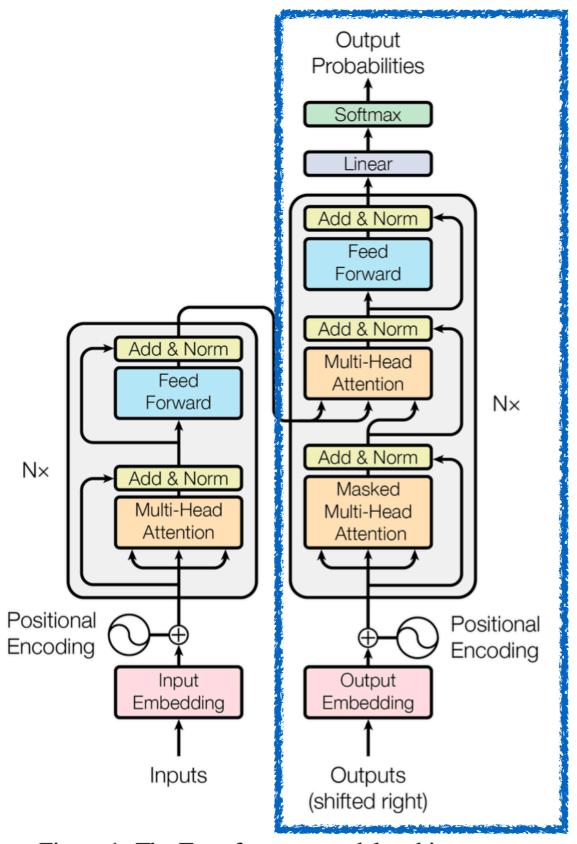


Figure 1: The Transformer - model architecture.

### GPT = Generative Pre-trained Transformer

LLM is an auto-regressive language model that uses an optimized transformer architecture

# Attention Is All You Need

Ashish Vaswani\* Google Brain avaswani@google.com

Llion Jones\* Google Research llion@google.com

Noam Shazeer\* Google Brain noam@google.com

Niki Parmar\* Google Research nikip@google.com Jakob Uszkoreit\* Google Research usz@google.com

Aidan N. Gomez\* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser\* Google Brain lukaszkaiser@google.com

Illia Polosukhin\* ‡ illia.polosukhin@gmail.com

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 Englishto-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

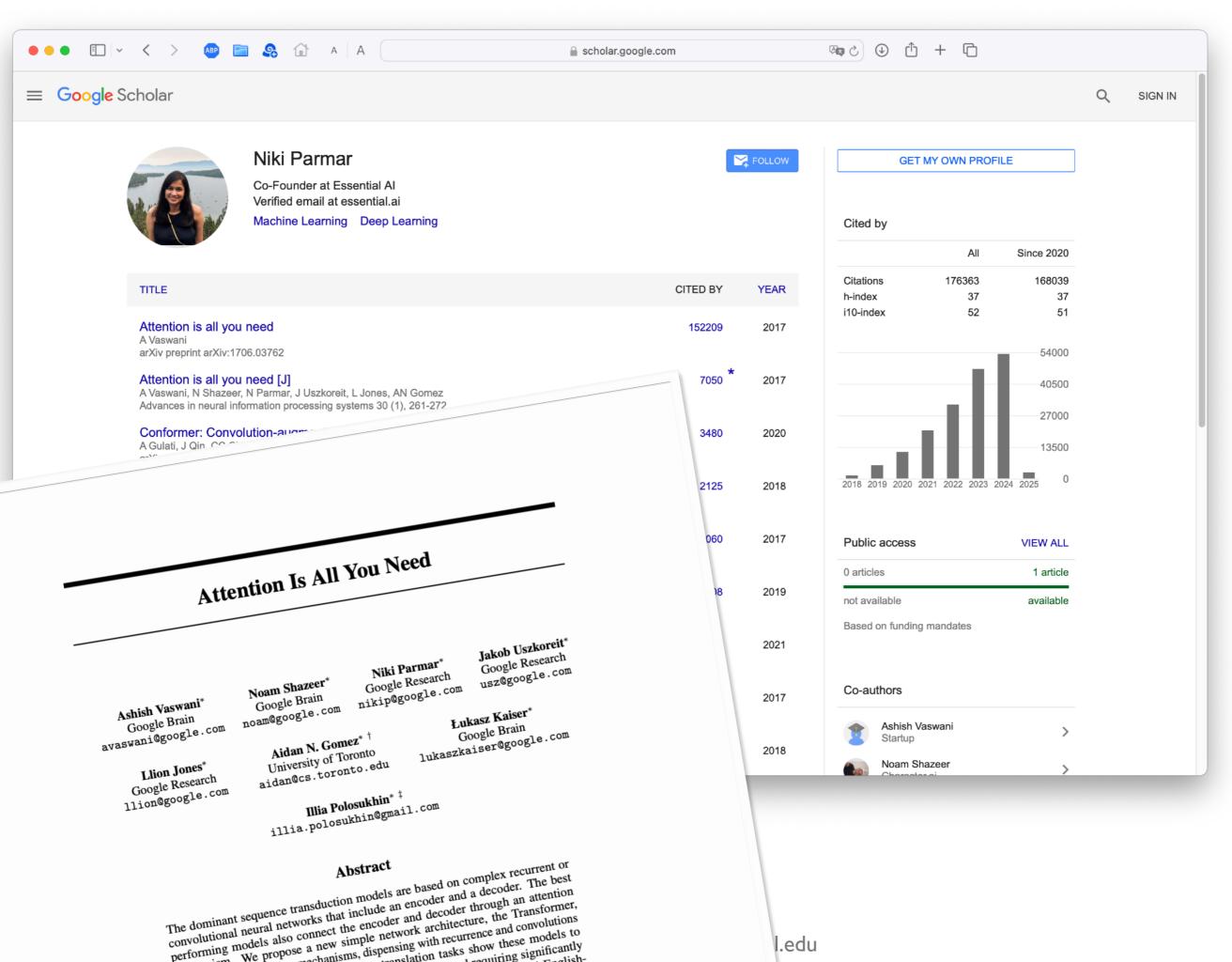
Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular have been firmly actablished as state of the art approaches in sequence modeling and Recurrent neural networks, long short-term memory [12] and gated recurrent [2] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and translation problems such as language modeling and machine translation problems such as language modeling and machine translation problems such as language modeling and machine translation [20] [2] [3]. in particular, have been many established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to such the boundaries of recurrent language models and according to such the boundaries of recurrent language models. erchitectures [21] [21] [13]

# Attention Is All You Need, 2017

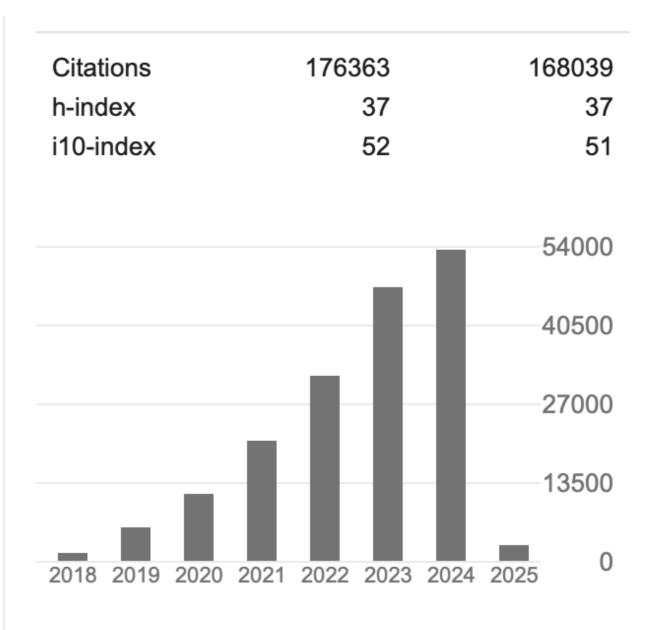
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention Is All You Need, in Advances in Neural Information Processing Systems, Vol. 30 (Curran Associates, Long Beach, CA, USA, 2017), pp. 261–272.

arXiv:1706.03762

<sup>\*</sup>Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started Equal contribution. Listing order is random. Jakob proposed replacing KININS with Sen-altenuol and started and implemented the first Transformer models and seffort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models. uate uns luca. Asmsn, while ma, designed and implemented the most transformer models and implemented the most transformer models. architectures [31, 21, 13]. leasted countless model variants in our original codebase and responsible for our initial codebase, and igning various parts of and



4.00			
	CITED BY	YEAR	The Contract
	152209	2017	And the state of t
	7050 *	2017	
	3480	2020	
	2125	2018	



Given a sequence of tokens (characters, words, bigrams, or subwords)

A logit is the raw...

Predict the next probable token

out...

Which becomes part of the next token prediction, etc.

(It is auto-regressive)

... put... of... a... ma... chine... learn ing mo del, typic ally be fore apply ing a trans form a tion like the soft max function.

### Probabilistic but causal calculation of the next token

## "Knowledge" is an emergent property



Q Search Wikipedia

Article Talk

Search

Contents hide

#### (Top)

Early life and education

Career

**SAGE Project** 

MIT Instrumentation Laboratory and the **Apollo Guidance** Computer

Apollo 11 landing

**Businesses** 

Legacy

**Tributes** 

**Awards** 

**Publications** 

Personal life

See also

References

Further reading

External links

### Margaret Hamilton (software engineer)

Edit View history Tools >

文A 50 languages ∨

From Wikipedia, the free encyclopedia

For other people named Margaret Hamilton, see Margaret Hamilton.

Margaret Elaine Hamilton (née Heafield; born August 17, 1936) is an American computer scientist. She was director of the Software Engineering Division of the MIT Instrumentation Laboratory, which developed on-board flight software for NASA's Apollo program. She later founded two software companies—Higher Order Software in 1976 and Hamilton Technologies in 1986, both in Cambridge, Massachusetts.

Hamilton has published more than 130 papers, proceedings, and reports, about sixty projects, and six major programs. She is one of the people credited with coining the term "software engineering".[1][2][3]

On November 22, 2016, Hamilton received the Presidential Medal of Freedom from president Barack Obama for her work leading to the development of on-board flight software for NASA's Apollo Moon missions.[4]

#### Early life and education [edit]

Margaret Elaine Heafield was born August 17, 1936, in Paoli, Indiana, [5] to Kenneth Heafield and Ruth Esther Heafield (née Partington). [6][7] The family later moved to Michigan, [8] where Margaret graduated from Hancock High School in 1954. [6]

She studied mathematics at the University of Michigan in 1955 before transferring to Earlham College, where her mother had been a student. [9][10] She earned a BA in mathematics with a minor in philosophy in 1958. [9][11] She cites Florence Long, the head of the math department at Earlham, as helping with her desire to pursue abstract mathematics and become a mathematics professor.[12]

#### **Margaret Hamilton**



Hamilton in 1995

Born Margaret Elaine Heafield

August 17, 1936 (age 88)

Paoli, Indiana, U.S.

Education University of Michigan

Earlham College (BA)

Occupation Software engineer

**Spouses** James Cox Hamilton

(m. 1958; div. 1967)

Dan Lickly (m. 1969)

Children

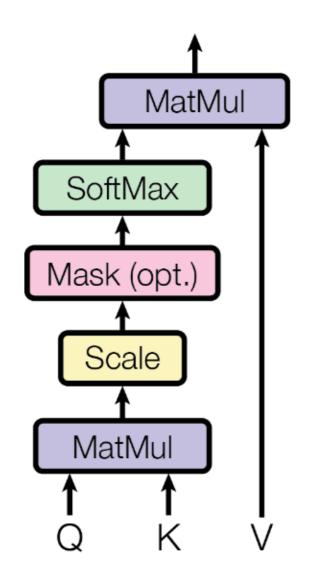
Relatives James Cox Chambers (former

Example inspired by Andrej Karpathy, Intro to Large Language Models https://www.youtube.com/watch?v=zjkBMFhNj g

$$\operatorname{Attention}(Q, K, V) = \operatorname{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$Mask = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

causality (for text)



### Probability of next token (SoftMax)

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}$$

Given vector of logits  $z = [z_1, z_2, \dots, z_n]$ 

### Boltzmann weighting

$$P_i = \frac{e^{-E_i/k_B T}}{\sum_j e^{-E_j/k_B T}}$$

### Query, Key, and Value matrices

A. Vaswani, et al., Attention Is All You Need, in Advances in Neural Information Processing Systems, Vol. 30 (Curran Associates, Long Beach, CA, USA, 2017), pp. 261–272.

"Temperature" hyperscaling parameter

$$P_{i} = \frac{e^{z_{i}/T}}{\sum_{i=1}^{n} e^{z_{j}/T}}$$

Stochastic behavior!

## Given an input token representation, X

$$\left. egin{aligned} Q = XW_Q \ K = XW_K \end{aligned} 
ight. 
i$$

$$d \times d_k$$
 or  $d \times d_v$ 

ca. June 2020

GPT-3 (175B) ~175 billion parameters (350GB) trained on O(10TR) data (the way) trained on O(I0TB) data (the web)

- Hidden size: 12,288
- Number of layers: 96
- Number of attention heads: 96
- Vocabulary size: ~50,000
- Feedforward network expansion factor: 4x

GPT-4 and 40 sizes not released, but estimates at  $10^{12}$  –  $10^{14}$  parameters

## Cost (compute & energy) is in the model training

## ~I0TB text

webcrawl,
Wikipedia,
Project
Gutenberg,
ArXiv, Stack
Exchange...
(Ilama)

+ fine-tuning



Llama 3.1 (Meta)

8B model: 4.9 GB

70B model: 43 GB

405B model: 243 GB

Llama 3.1 8B –1.46 million GPU hours

Llama 3.1 70B – 7.0 million GPU hours

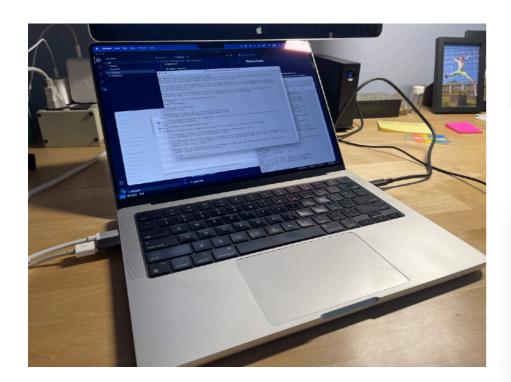
Llama 3.1 405B – 30.84 million GPU hours

(approximately 440,000 petaFLOP-days for training)

https://huggingface.co/blog/llama3 I https://ollama.com/library/llama3. I

## Running models is relatively inexpensive

Macbook Pro M3, 12 cores (2023) 18GB RAM, 18 core GPU

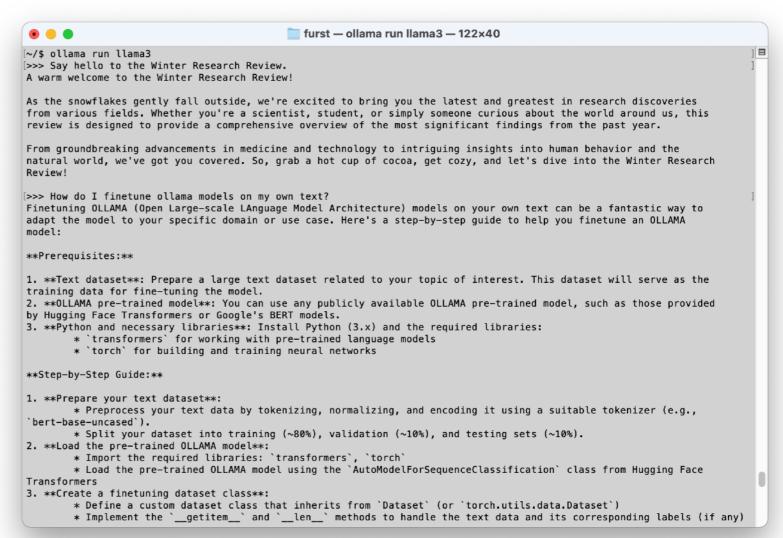




## running in *ollama* locally with Ilama3 8B (4.7 GB)

51M Jan 16 11:37 ollama\* staff -rwxr-xr-x@ 1 furst

HTTPS://OLLAMA.COM/



## Results are pretty good!



furst — ollama run llama3 — 118x54

~/\$ ollama run llama3 >>> Say hello to the Winter Research Review. A warm welcome to the Winter Research Review!

As the snowflakes gently fall outside, we're excited to bring you the latest and greatest in research discoveries from various fields. Whether you're a scientist, student, or simply someone curious about the world around us, this review is designed to provide a comprehensive overview of the most significant findings from the past year.

From groundbreaking advancements in medicine and technology to intriguing insights into human behavior and the natural world, we've got you covered. So, grab a hot cup of cocoa, get cozy, and let's dive into the Winter Research Review!

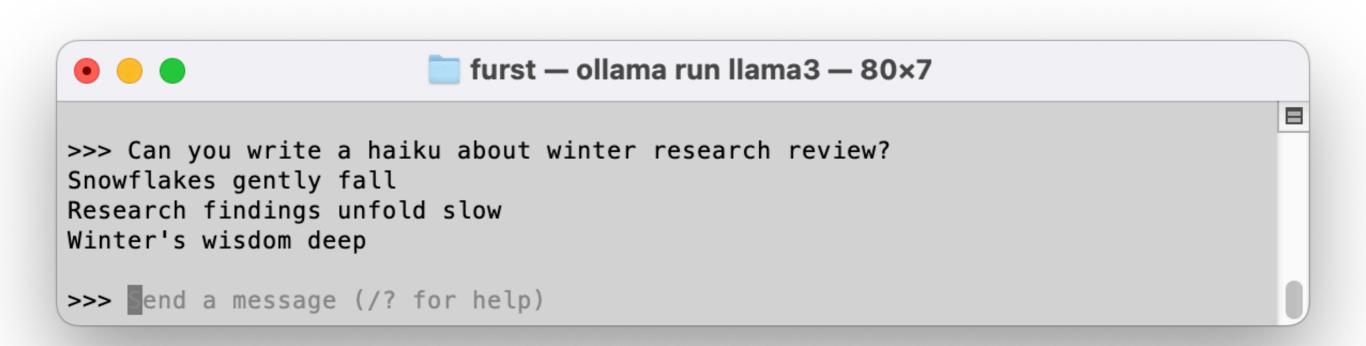


Macbook Pro M3, 12 cores (2023) 18GB RAM, 18 core GPU running llama3 8.0B

llama3 8B (4.7 GB) running in *ollama* 

HTTPS://OLLAMA.COM/

## Even better!





Macbook Pro M3, 12 cores (2023) 18GB RAM, 18 core GPU running llama3 8.0B

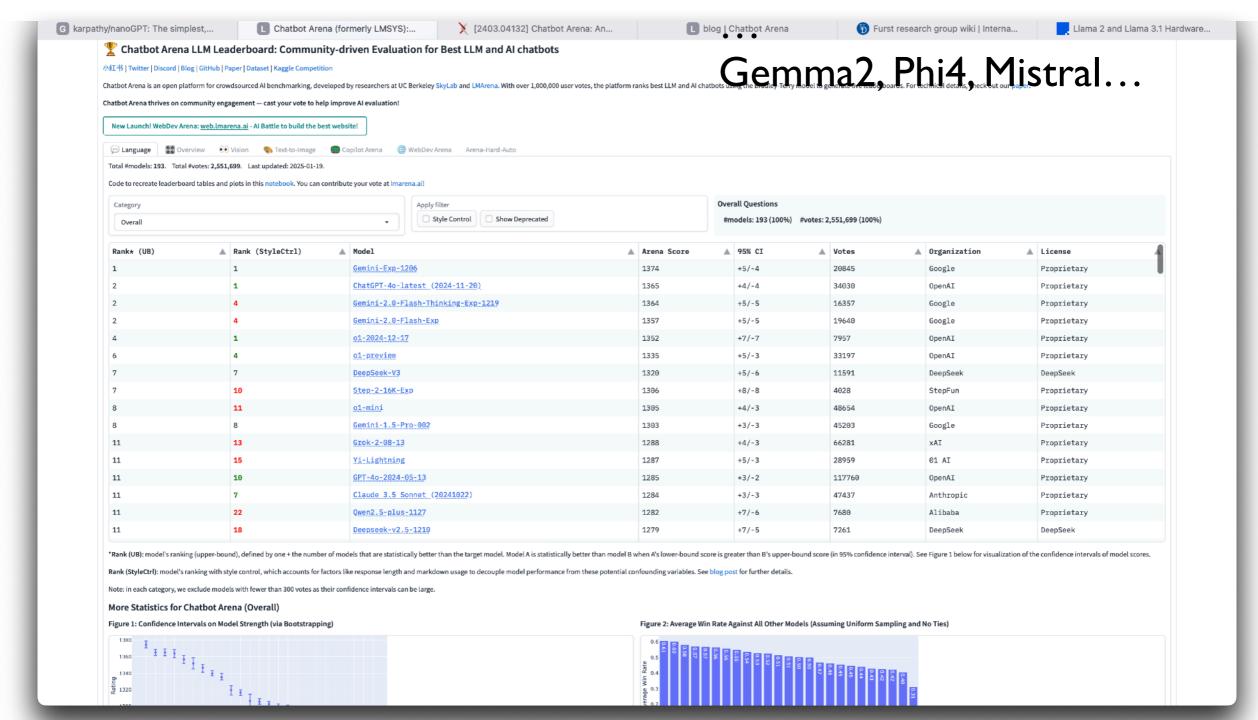
llama3 8B (4.7 GB) running in *ollama* 

HTTPS://OLLAMA.COM/

## Chatbot Leaderboard

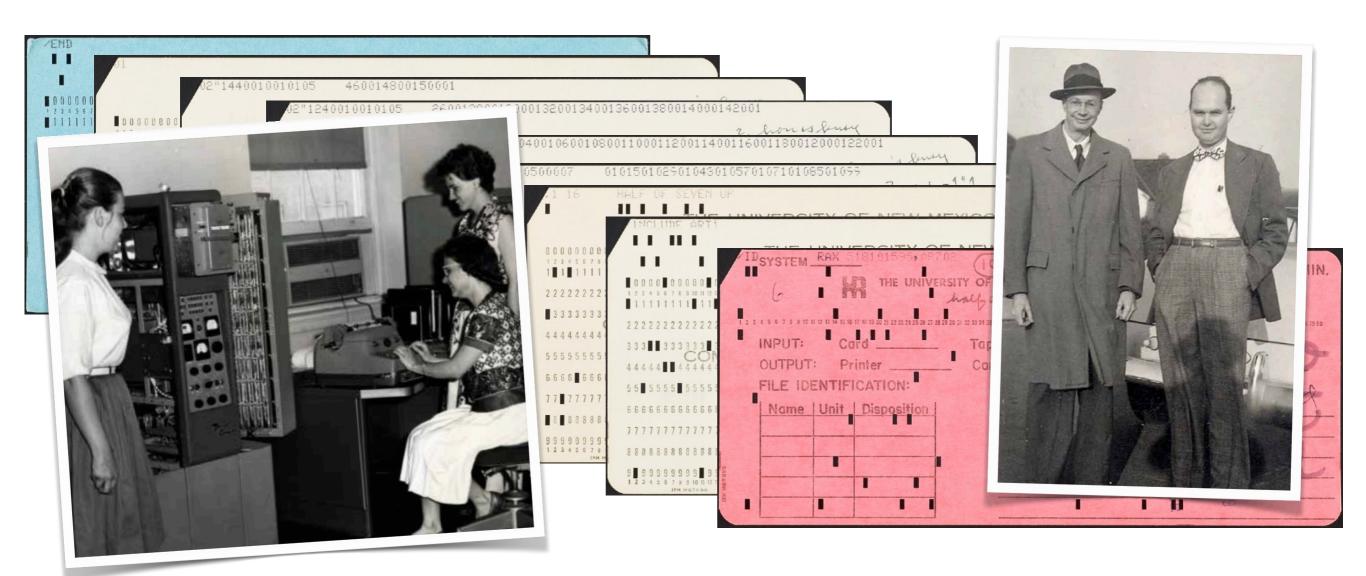
LMARENA.AI

Gemini (Google) Proprietary ChatGPT (OpenAI) Proprietary DeepSeek (DeepSeek) Open Llama (Meta) Open Claude (Anthropic) Proprietary



## Uses in chemical engineering

GPTs and LLMs are transformative technologies, analogous to the advent of the electronic, programmable digital computer



Bendix G-15-D, ca. 1958

Chemical engineers rapidly adopted these digital computers to numerically solve difficult problems

Blaszczyk, 2014

23

Immediate opportunities for <u>LLMs</u> are clear:

A tool for difficult but adjacent problems, coding and modeling

### Some questions:

What infrastructure do we want / need as a community? Licenses? (ChatGPT? Github Co-Pilot? Claude?) Hardware? (GPU) clusters?

Train our own (Open Source) models? Build ChatGPT apps?

How do we share / communicate / discuss / experiment?

Code repositories? Workshops?

### What I think about:

Proprietary models, walled gardens, eventual drive for monetization

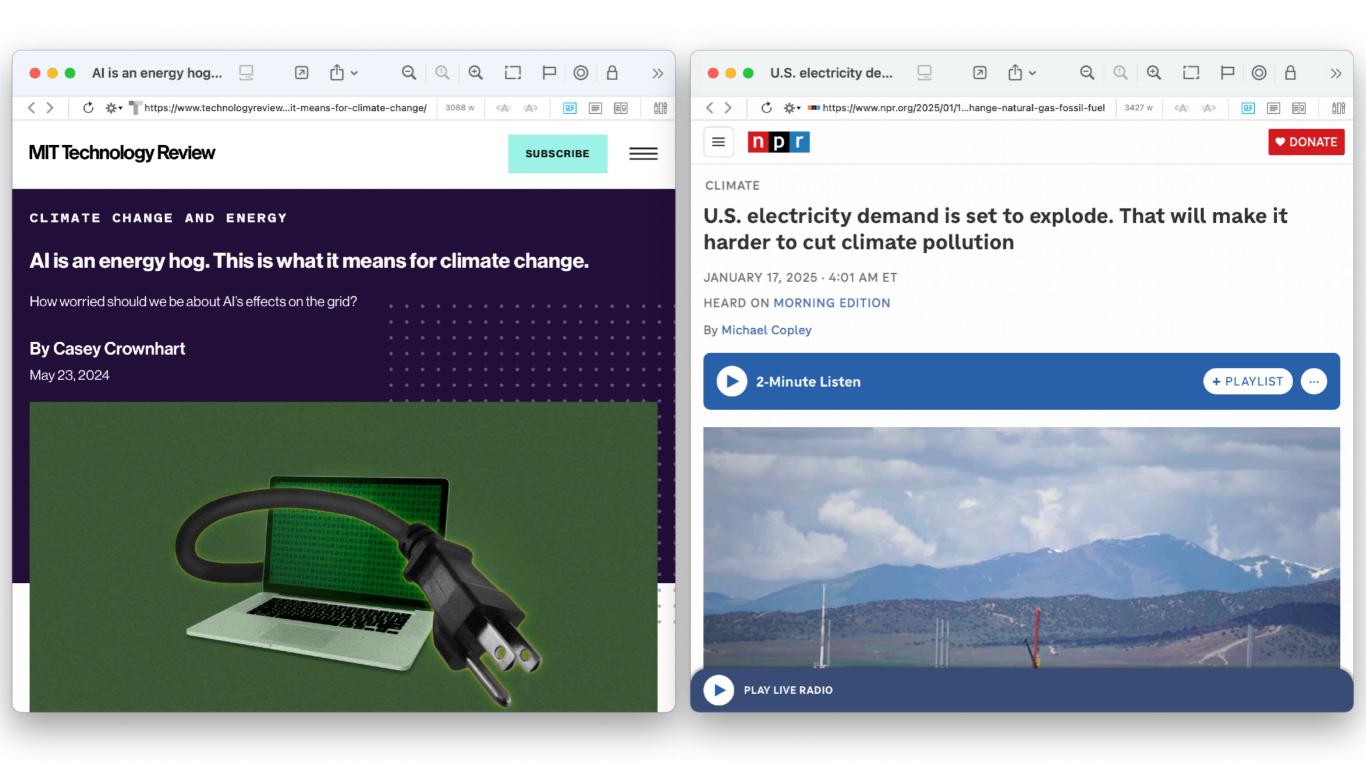
I don't want to read your Al-generated bullet summaries...

What guidance for students? What to teach and how to teach them?\*

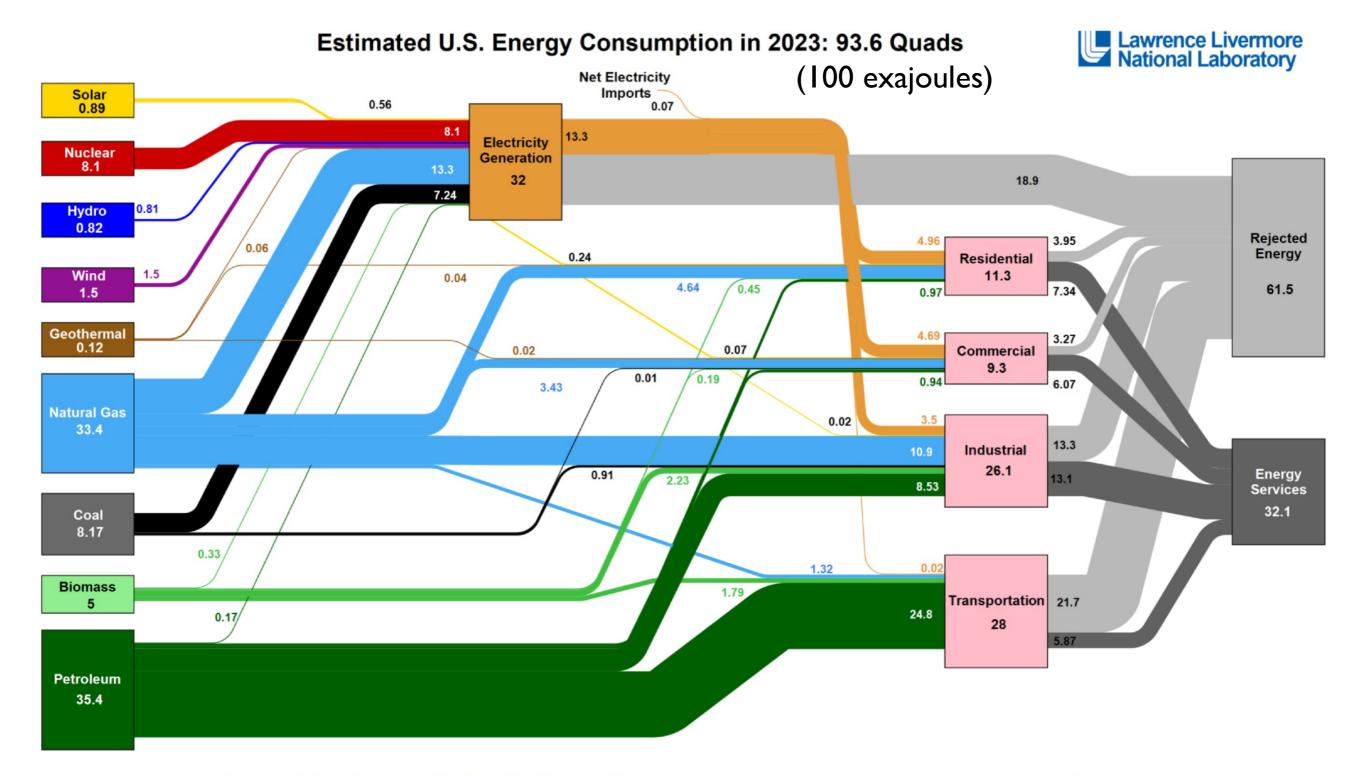
How else do we use it in our research?! Documentation?, ...

\*An entire discussion on coding literacy — see Annette Vee, Pitt

## Al in the news



https://www.npr.org/2025/01/16/nx-s1-5251454/electricity-demand-data-centers-climate-change-natural-gas-fossil-fuel https://www.technologyreview.com/2024/05/23/1092777/ai-is-an-energy-hog-this-is-what-it-means-for-climate-change/



Source: LLNL October, 2024. Data is based on DOE/EIA SEDS (2024). If this information or a reproduction of it is used, credit must be given to the Lawrence Livermore National Laboratory and the Department of Energy, under whose auspices the work was performed. Distributed electricity represents only retail electricity sales and does not include self-generation. EIA reports consumption of renewable resources (i.e., hydro, wind, geothermal and solar) for electricity in BTU-equivalent values by assuming a typical fossil fuel plant heat rate. The efficiency of electricity production is calculated as the total retail electricity delivered divided by the primary energy input into electricity generation. End use efficiency is estimated as 65% for the residential sector, 65% for the commercial sector, 49% for the industrial sector, and, 21% for the transportation sector. Totals may not equal sum of components due to independent rounding. LLNL-MI-410527

## Remember: energy use is (mostly) in the model training

CAMERAS PHOTO SUPPLIES

RIDING 1St CLASS
ON THE TITANICI

PHOTOGRAPHS BY NATHAN LYONS

## Learn by doing – hack on an LLM

HTTPS://WWW.YOUTUBE.COM/WATCH?V=KCc8FMEBlnY



### **Andrej Karpathy**

@AndrejKarpathy · 605K subscribers · 16 videos

SuperThanks: very optional, goes to Eureka Labs. ...more

eurekalabs.ai and 4 more links

Subscribe

Home

Andrej Karpathy

2.4M views • 1 year ag

**Neural Netwo** 

**Videos** 

**Playlists** 

Community

 $\bigcirc$ 

#### **General Audience**

Play all

videos for more general audience, no programming experience necessary.



LET'S BUILD GPT. FROM SCRATCH. IN CODE. SPELLED OUT.





Let's build GPT: from scratch, in code, spelled out.

5M views • 2 years ago

[1hr Talk] Intro to La [furst@anisotropic:~/LLM/nanoGPT\$ wc \$(ls -1 \*.py)

Language Models 117 487 4815 bench.py

47 219 1758 configurator.py

331 1798 16507 model.py

93 522 4313 sample.py

336 1799 14845 train.py

924 4825 42238 total

~300 lines of Python Runs on CPU or GPU

## Acknowledgments



**Bayles** 



Hartt



Lobo



Furst Group 2025

Bree Huntington @ I:10PM in 120

& our WRR grad student organizers for this opportunity!

# slides at

https://furst.group

(soon)