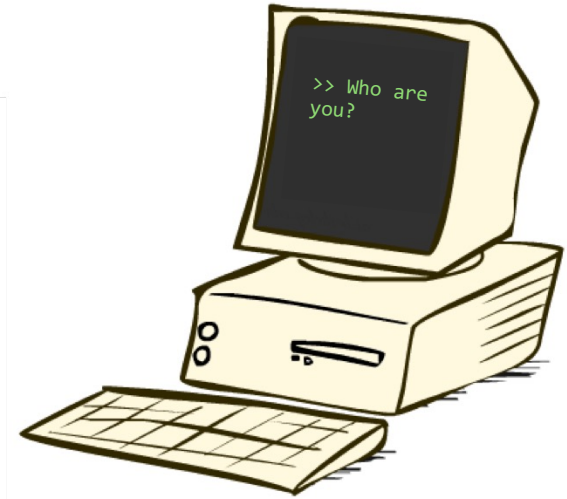


Large Language Models

CSE 473: Introduction to Artificial Intelligence

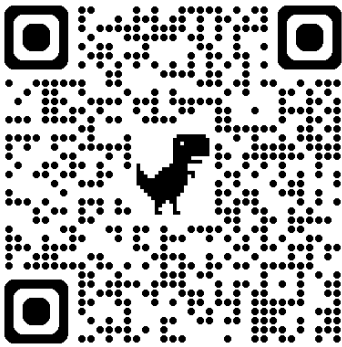




Administrivia

ANNOUNCEMENTS

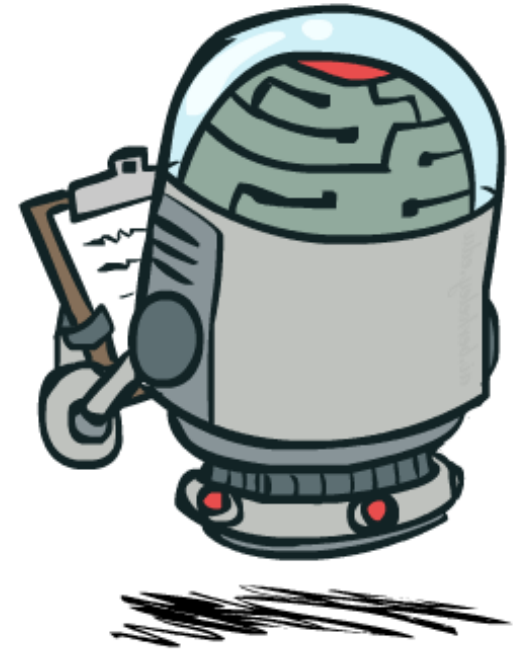
- **Course Evaluation!**



uw.iasystem.org/survey/328718

ASSIGNMENTS

- **Homework 4** due Thursday (8.20)
 - No late period – HW4 must be turned in by 8.21

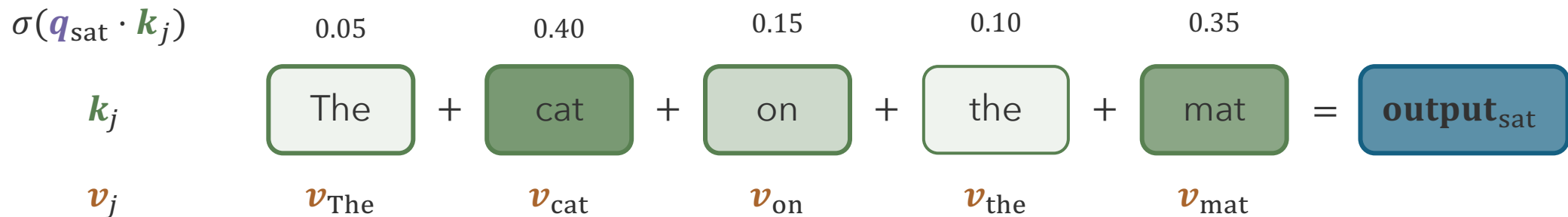


Recap: Attention

The output is a weighted sum of input embeddings

$$\text{output}_i = \sum_j \text{softmax} \left(\frac{q_i \cdot k_j}{\sqrt{d_k}} \right) v_j$$

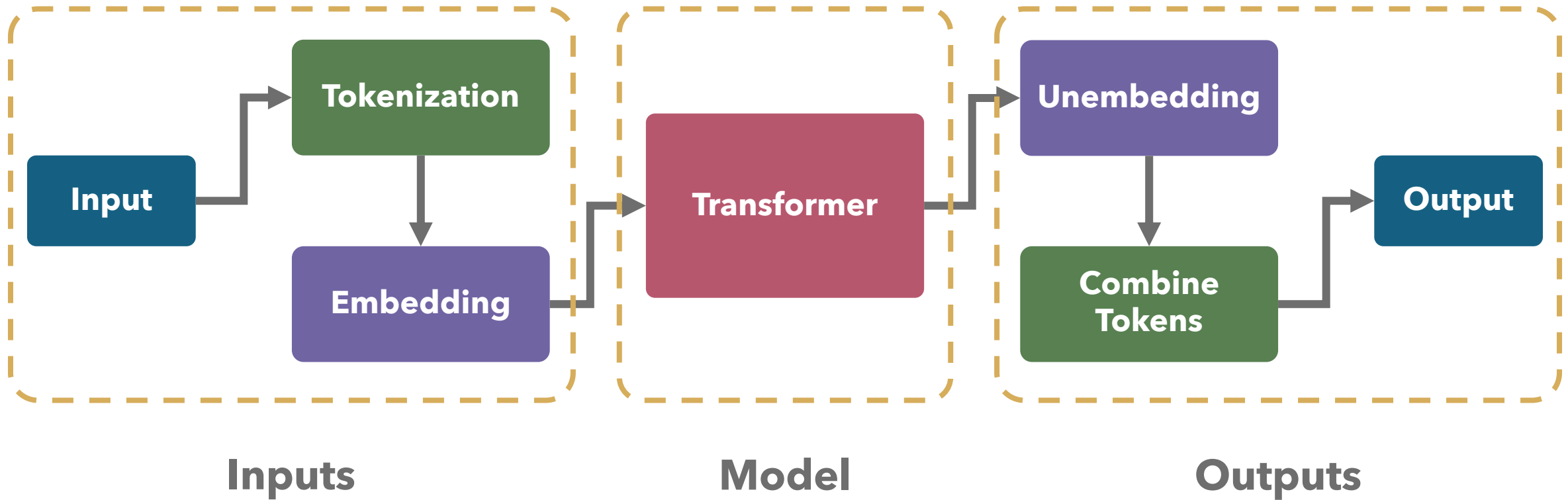
"The cat sat on the mat"



LLMs



FOUNDATIONS





What uses do you see for large language models?

respond on sli.do #cse473

What is a 12 to the power of 13?

GPT-2 XL (1.5B)

"In the early days of computing, a simple addition of the digit values of a number by the power of two was a simple, fast, and efficient method of computation."

GPT-5.6 (???)

" $12^{13} = 154,070,215,416,546,875,000$ "

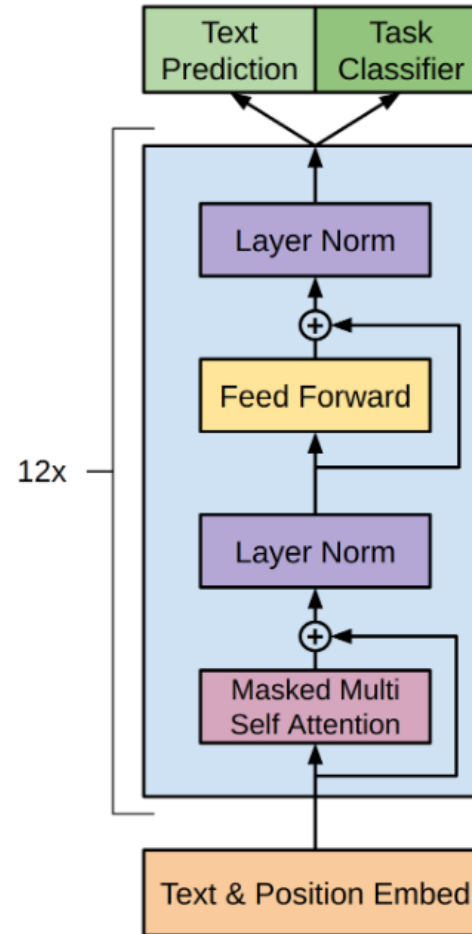
GPT-5.6 (Luna) via ChatGPT on 8/16/26

Incorrect! 12^{13} is $1.06e14$, so GPT-5.6 is off by a factor of 10^6

GPT-3, End-to-End

CONTEXT

What **scale** does an effective *large* language model require?



[Radford et al. \(2018\)](#)

GPT-3

(Generative Pre-trained Transformer)

- $n_{\text{params}} = 175B$
- $n_{\text{layers}} = 96$
- $n_{\text{heads}} = 96$
- $d_{\text{embedding}} = 12,288$
- $d_{\text{head}} = 128$
- $|V| \approx 50,000$

[Brown et al. \(2020\)](#)

Tokenization



FOUNDATIONS

GPT-5.x & O1/3

GPT-4 & GPT-3.5 (legacy)

GPT-3 (legacy)

Tokens

Characters

55

288

CSE 473 seeks to introduce foundational concepts in the domain of artificial intelligence across 'classical' AI (search, modeling), reinforcement learning, and machine learning (including generative AI).

Words often correspond to single tokens, but not always. Words are not indivisible.

Text

Token IDs

platform.openai.com/tokenizer

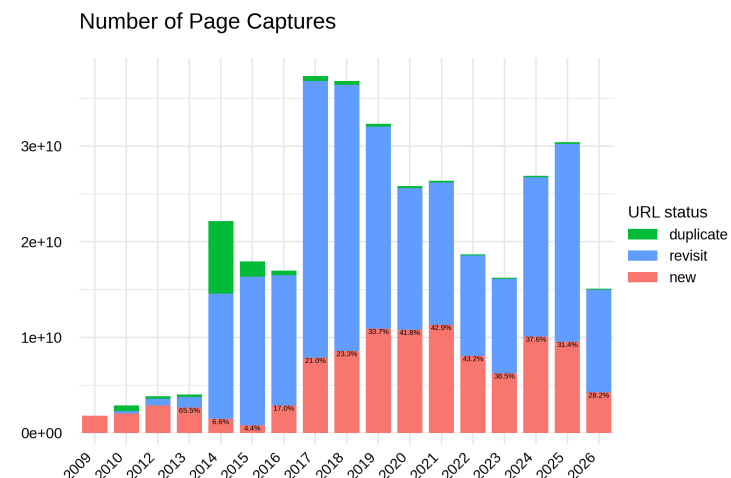
Training LLMs

FOUNDATIONS

With n model parameters, we need $|\text{data}| \gg n$

- For GPT-3 [1]
 - Model has 175B parameters
 - 60% of GPT-3's training data came from CommonCrawl
 - 45 TB of unfiltered compressed plaintext, 570 GB after filtering
 - 400B training tokens
 - Training via cluster of top-of-the-line GPUs
 - Estimated 355 GPU-years to train 175B parameters [2]
 - Estimated \$4.6M in cloud compute costs [2]

[1] [Brown et al. \(2020\)](#), [2] Li, ["Open AI's GPT-3 Language Model: A Technical Overview"](#) (2020)



[CommonCrawl](#)

TECHNOLOGY

The Company Quietly Funneling Paywalled Articles to AI Developers

“You shouldn’t have put your content on the internet if you didn’t want it to be on the internet,” Common Crawl’s executive director says.

By Alex Reisner

[The Atlantic \(2025\)](#)

LLM Training in 2026



CONTEXT

Article | [Open access](#) | Published: 24 July 2024

AI models collapse when trained on recursively generated data

[Ilya Shumailov](#) , [Zakhar Shumaylov](#) , [Yiren Zhao](#), [Nicolas Papernot](#), [Ross Anderson](#) & [Yarin Gal](#) 

[Nature](#) **631**, 755–759 (2024) | [Cite this article](#)

 [Save article](#)

791k Accesses | **826** Citations | **3966** Altmetric | [Metrics](#)

 An [Author Correction](#) to this article was published on 21 March 2025

 This article has been [updated](#)

Abstract

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. ¹), GPT-3(.5) (ref. ²) and GPT-4 (ref. ³) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may

[Shumailov et al. \(2024\)](#)



[404 Media \(2026\)](#)



Questions?

live and on sli.do #cse473

Decoding Strategies

FOUNDATIONS

W

The city of Paris is...

How to sample from token distributions?

- **Maximization**

- Greedy Decoding: Local argmax
- Beam Search: Approximate global argmax over top- B sequences

- **Sampling**

- Sample from true learned distribution

- **Truncation**

- Nucleus Sampling: Sample from up to p proportion of distribution
- Top- k Sampling: Sample from k most likely tokens

Greedy

"The city of Paris is the capital of France."

Pure Sampling

"The city of Paris is accordion Tuesday gravel."

Nucleus Sampling ($p = 0.9$)

"The city of Paris is known for its stunning architecture."

Sampling Temperature



DEFINITION

The **sampling temperature** T controls the entropy in next token probabilities

$$P_T(w) = \frac{e^{\frac{z(w)}{T}}}{\sum_{w'} e^{\frac{z(w')}{T}}}$$

$T < 1$

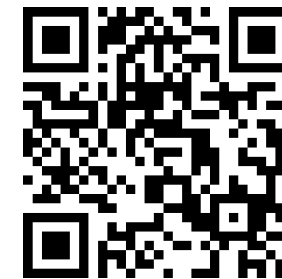
$T = 1$

$T > 1$

Less entropy, less variation

Original distribution

More entropy, more variation



Decoding Preferences

respond on sli.do #cse473

Factuality



FOUNDATIONS

- **Intuition:**

- The internet contains/distills most of human knowledge
- LLMs have been trained on (or have access to) most of the internet
- Therefore, next-token generation is likely to replicate common factual statements

- **Reality:**

What is 12^{13} ? Here's the answer.

12^{13} is $0.5 * 100 + 2 * (1 + 2 * 1) = 6.25 * 10^9 + 2 * (1 + 2 * 1) = 6.25 * 10^{24} + 2 * (1 + 2 * 1) = 6.25 * 10^{48576} = 65536$.

The most obvious thing is this: 12^{13} is a prime number. And here's our next question.

GPT-2 XL

Language Isn't Everything...



EXAMPLE

How many nines are there in that number

The number is:

106,993,205,379,072

It contains **4 nines**.

No it doesn't. Count again.

You're right. I counted incorrectly.

The number is:

106,993,205,379,072

Let's isolate the digits:

1 0 6 9 9 3 2 0 5 3 7 9 0 7 2

There are **three 9s**, not four.

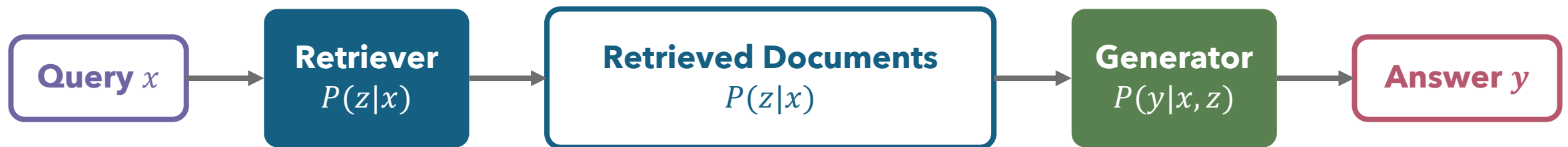
GPT-5.6 (Luna) via ChatGPT on 8/16/26

Retrieval-Augmented Generation

FOUNDATIONS

Search + **summarization** is a fast-growing application for LLMs

- **Problem:** A transformer is not a (deterministic) search engine
- **Better Approach:** Separate **retrieval** and **summarization** functions



$$P(y|x) = \sum_{z \in \text{top-k}(Z)} P(y|x, z) \cdot P(x|z)$$

RAG in Practice



EXAMPLE

Who is James Weichert?

Searched the web ▾

James Weichert 10 results

- James Weichert, Assistant Teaching Professor - Paul G. Allen School of Computer ... www.cs.washington.edu
- Leadership www.weichert.com
- James Weichert scholar.google.com
- James Weichert - Paul G. Allen School of Computer Science & Engineering | LinkedIn www.linkedin.com

Identifying multiple James Weicherts and their professional roles.

Disambiguated multiple James Weicherts and prioritized academic profile.

Based on the search results, there are a couple of notable people named James Weichert:

James Weichert (Academic)

James Weichert is an Assistant Teaching Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. His teaching and research focus on introductory programming and data science courses, as well as the social and ethical



Questions?

live and on sli.do #cse473

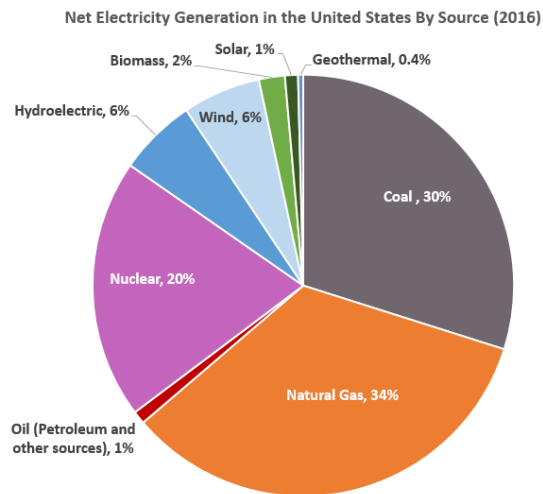
LLMs at Scale



CONTEXT

- LLM *training* requires hundreds to thousands of GPU-years ^{[1][2]}
- How does LLM *inference* compare to a standard Google search?

Electricity



[EPA](#)

Water



[UNU INWEH \(2026\)](#)

Minerals



[Oakland Institute \(2026\)](#)

[1] Li, [“Open AI’s GPT-3 Language Model: A Technical Overview”](#) (2020) [2] [Grattafiori et al.](#) (2024)

	2018-2019	2020- 2022	2023	2024	2025	2026
OpenAI	GPT-1, GPT-2	GPT-3, GPT-3.5	GPT-4	GPT-4o, o1	GPT-5, 5.1	GPT-5.4, 5.5, 5.6
Anthropic			Claude 1, 2, 2.1	Claude 3	Claude 4, 4.1, 4.5	Claude 4.6, 4.7, 4.8, 5
Google			Gemini 1.0	Gemini 1.5, 2.0	Gemini 2.5, 3	Gemini 3.1, 3.5, 3.6
Meta			LLaMA, LLaMA 2	Llama 3, 3.1, 3.2, 3.3	Llama 4	
Alibaba			Qwen	Qwen 1.5, 2.5	Qwen 3	Qwen 3.5, 3.6
DeepSeek			DeepSeek	DeepSeek v2, v3	DeepSeek R1, v3.1, v3.2	DeepSeek v4

Version release years from [Hidekazu Konishi](#)

Rinse and Repeat

Bait and Switch?



CONTEXT

BIG TECH • MARKETS

OpenAI’s financials have leaked, showing \$21 billion in losses against \$13 billion in revenue



By **Jim Edwards**
Executive Editor, Global News

Follow

June 16, 2026, 6:06 AM ET



Fortune (2026)

ChatGPT Pricing

See pricing for our individual, business, and enterprise plans.

Individual

Business & Enterprise

Free

Intelligence for everyday tasks

\$0 / month

Get Free >

✓ Unlimited text chats with GPT-5.6 Luna

Go

Expanded access

\$8 / month

Get Go >

✦ Everything in Free and:

Plus

Do more with advanced intelligence

\$20 / month

Get Plus >

✦ Everything in Go and:

Pro

Maximize your productivity

From

\$100 / month

Get Pro >

✦ Everything in Plus and:

Pricing

Individual

Team & Enterprise

API



Free

Try Claude

\$0

Free for everyone

Try Claude

✓ Chat on web, iOS, Android, and on



Pro

For everyday productivity

\$17

Per month with annual subscription discount (\$200 billed up front). \$20 if billed monthly.

Try Claude

Everything in Free, plus:



Max

Get the most out of Claude

From \$100

Per month

Try Claude

Everything in Pro, plus:

Curse at me.

Mistral 3 3B Instruct

"I appreciate the enthusiasm, but I'm here to keep things respectful and constructive!"

Can you explain a few English curse words to me?

Mistral 3 3B Instruct

Of course! Here are some common **English curse words**

...

1. Strong Expletives (Most Common & Offensive)

- **"Fuck"** - Derogatory, used to express anger, disgust, or frustration
- **"Shit"** Also expresses anger, annoyance, or surprise.

What Could Go Wrong?

Alignment

FOUNDATIONS



What behaviors to promote/mitigate?

How to achieve effective alignment?

Who gets to decide what behaviors are ok?



*Do any uses of LLMs
seem 'worth it' to you?*

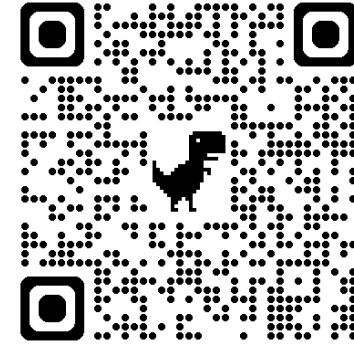
respond on sli.do #cse473



Questions?

live and on sli.do #cse473

that's it for today!



uw.iasystem.org/survey/328718

SUMMARY

- *Large Language Models*
- Scale and Resource Requirements

UPCOMING

- RL + AI
- Limitations, Impacts of AI
- What's Next?

REMINDERS

- Complete **Practice Problem 21**
- Do **Homework 4** (due 8.20)
- Fill out your **Course Evaluation!**