

A scenic view of the Golden Gate Bridge in San Francisco at sunset. The bridge's iconic red-orange towers and suspension cables are silhouetted against a warm, orange and pink sky. In the foreground, a field of purple lavender flowers is in bloom, adding a soft, natural touch to the urban landscape. The water of the bay is visible in the middle ground, reflecting the light from the sky.

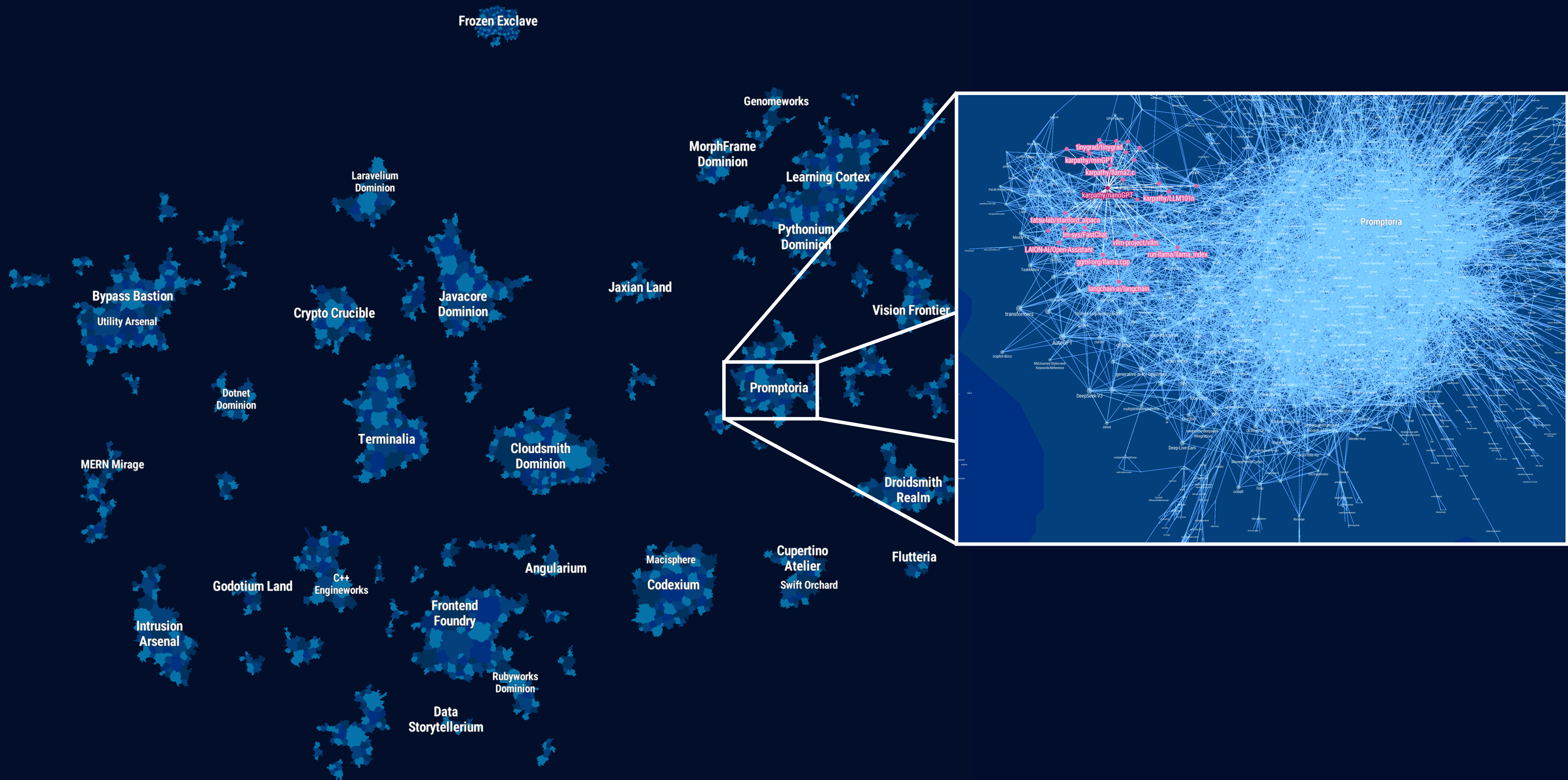
Software in the era of AI

Andrej Karpathy
YC AI Startup school
June 16

Software is changing.

(again)

"Map of GitHub"



Software 2.0



Andrej Karpathy

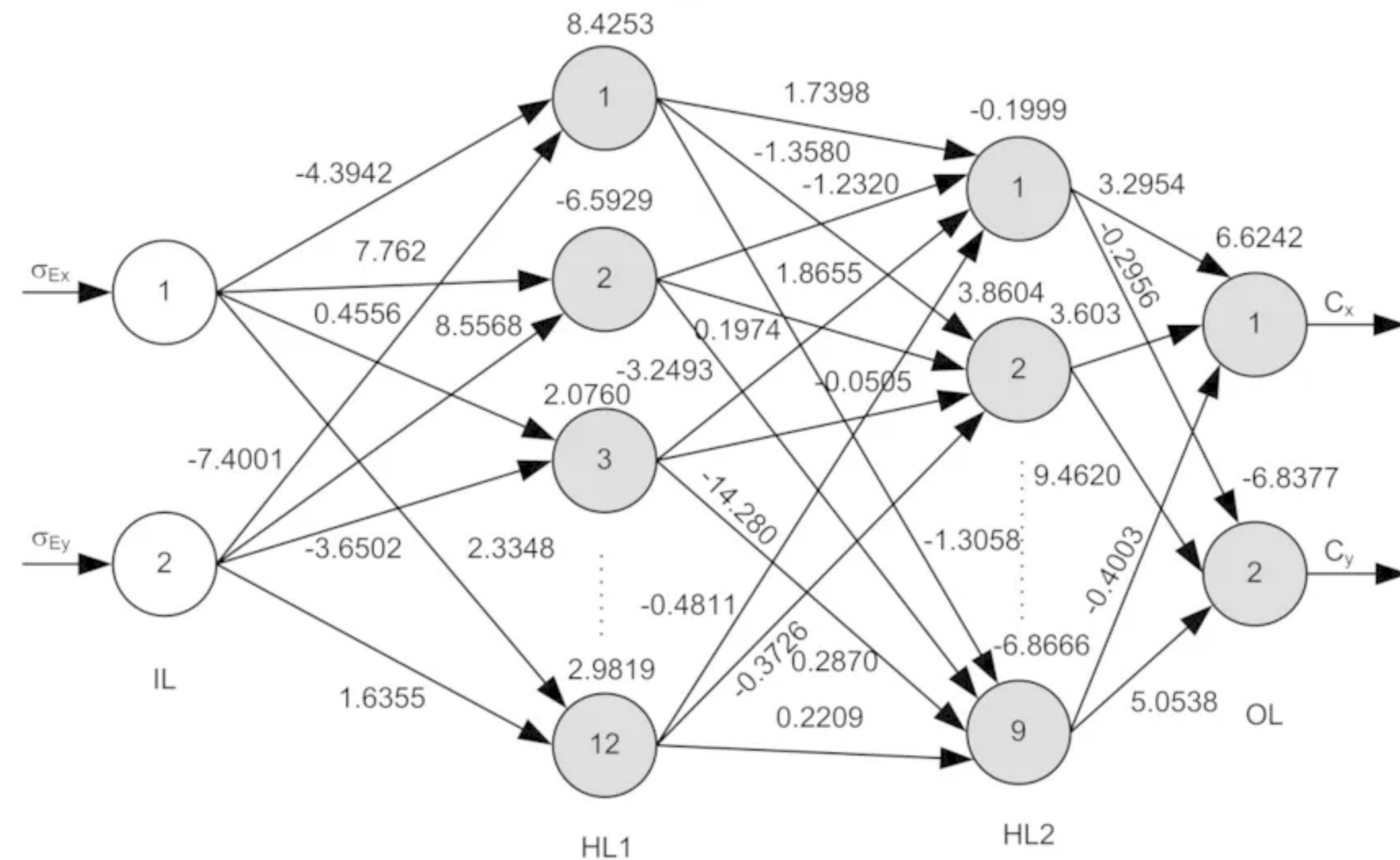
Follow

9 min read · Nov 11, 2017

Software 1.0 = code



Software 2.0 = weights



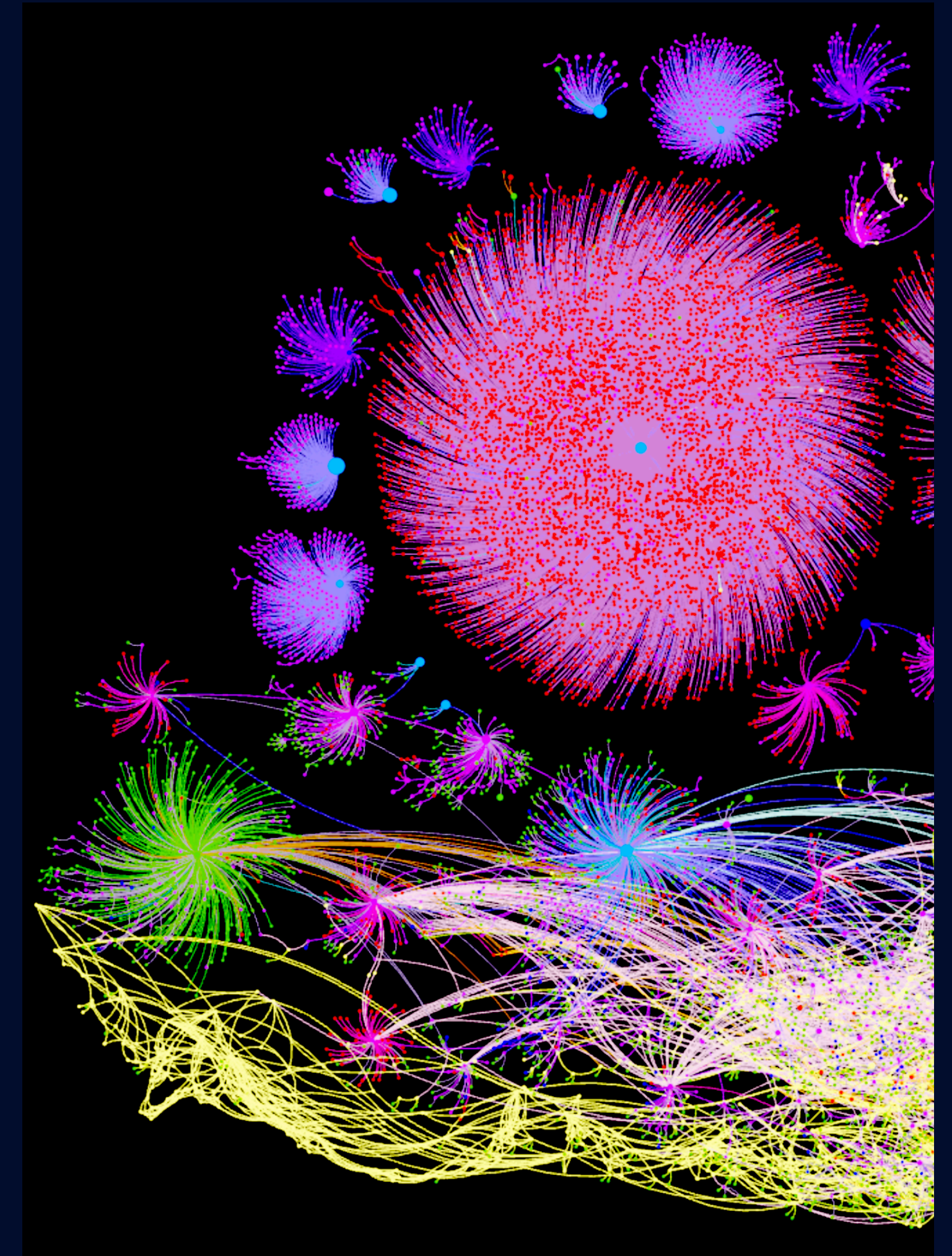
"Map of GitHub" (Software 1.0)

computer code



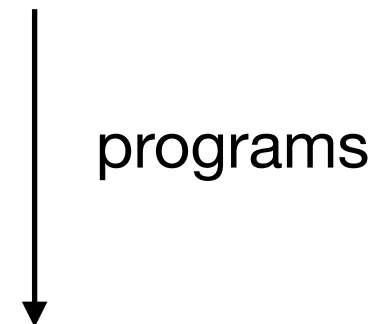
HuggingFace Model Atlas (Software 2.0)

neural network weights

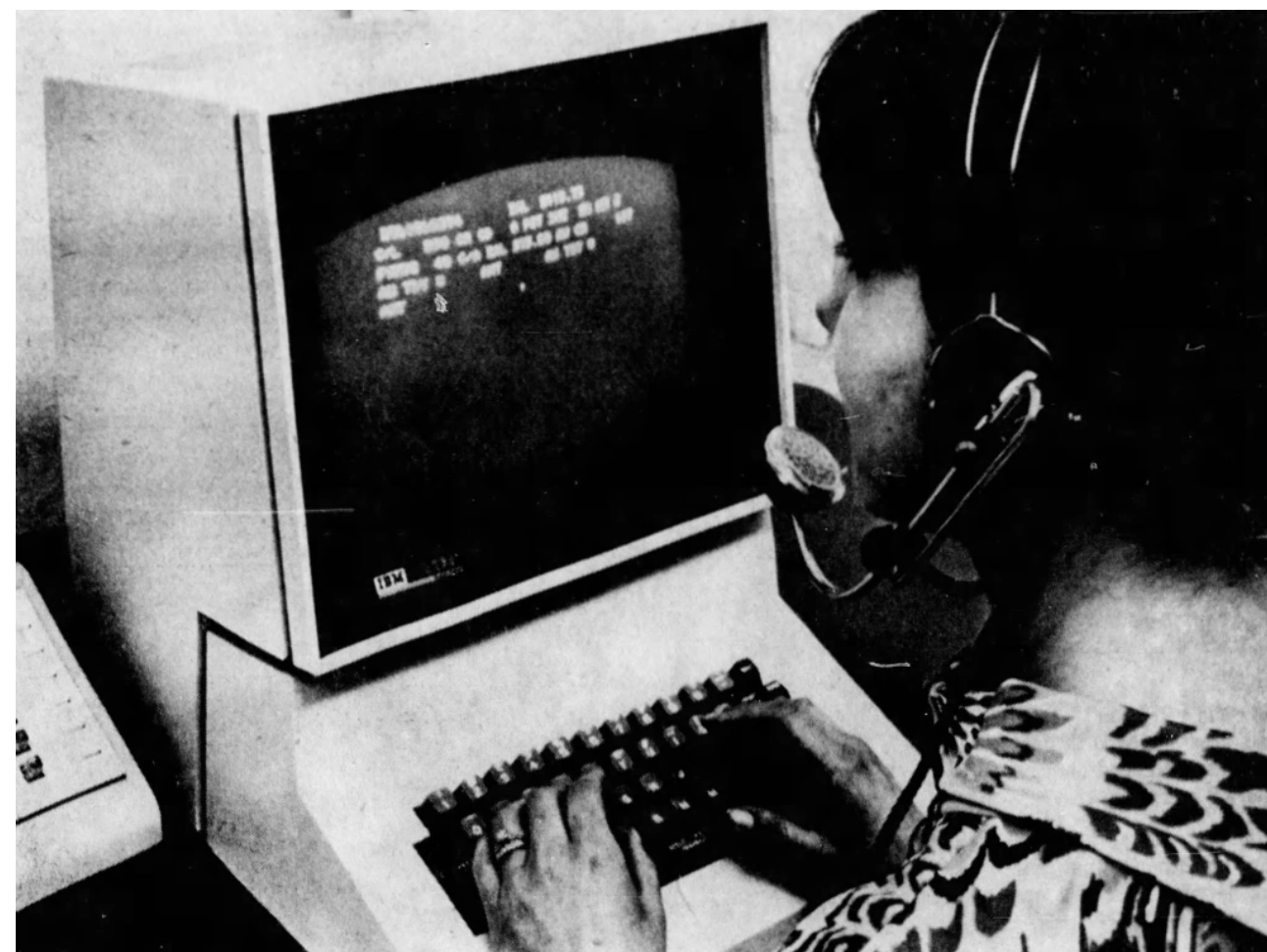


Software 1.0

computer code



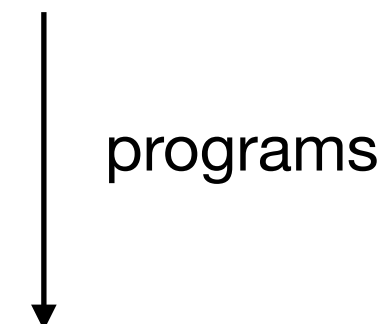
computer



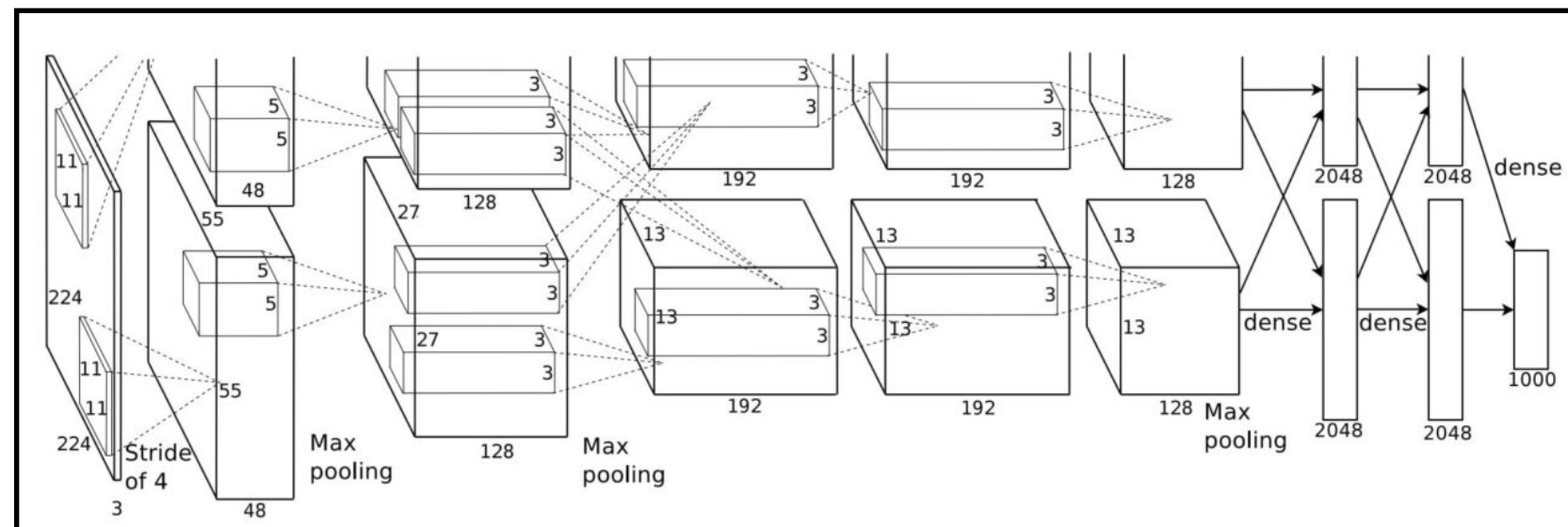
became programmable in ~1940s

Software 2.0

weights



neural net

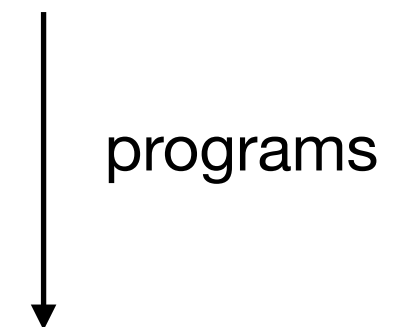


fixed function neural net

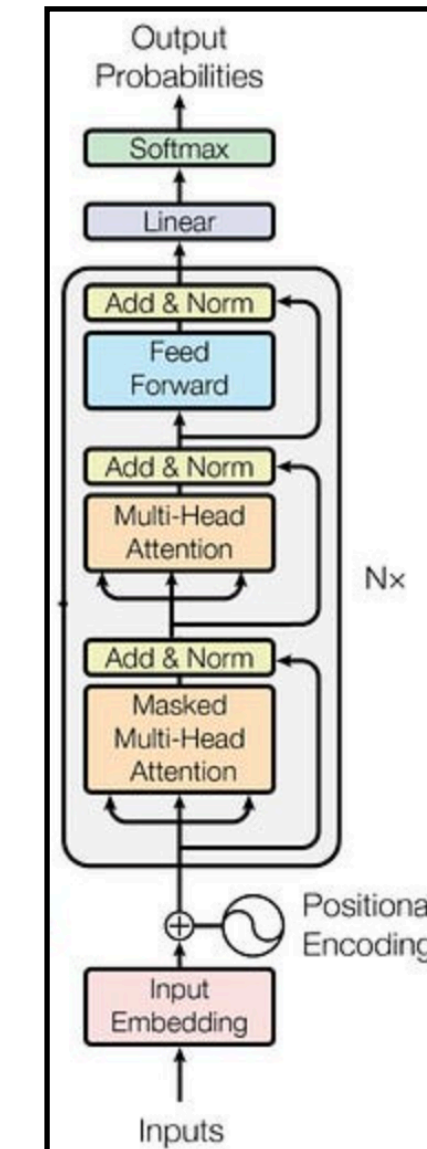
e.g. AlexNet: for image recognition (~2012)

Software 3.0

prompts



LLM



~2019

LLM = programmable neural net!

Example: Sentiment Classification

Software 1.0

python

Copy

```
def simple_sentiment(review: str) -> str:
    """Return 'positive' or 'negative' based on a tiny keyword lexicon."""
    positive = {
        "good", "great", "excellent", "amazing", "wonderful", "fantastic",
        "awesome", "loved", "love", "like", "enjoyed", "superb", "delightful"
    }
    negative = {
        "bad", "terrible", "awful", "poor", "boring", "hate", "hated",
        "dislike", "worst", "dull", "disappointing", "mediocre"
    }

    score = 0
    for word in review.lower().split():
        w = word.strip(".,!?:;")      # crude token clean-up
        if w in positive:
            score += 1
        elif w in negative:
            score -= 1

    return "positive" if score >= 0 else "negative"
```

Software 2.0

10,000 positive examples
10,000 negative examples
encoding (e.g. bag of words)

train binary classifier

parameters

Software 3.0

You are a sentiment classifier. For every review that appears between the tags

<REVIEW> ... </REVIEW>, respond with **exactly one word**, either POSITIVE or NEGATIVE (all-caps, no punctuation, no extra text).

Example 1

<REVIEW>I absolutely loved this film—the characters were engaging and the ending was perfect.</REVIEW>

POSITIVE

Example 2

<REVIEW>The plot was incoherent and the acting felt forced; I regret watching it.</REVIEW>

NEGATIVE

Example 3

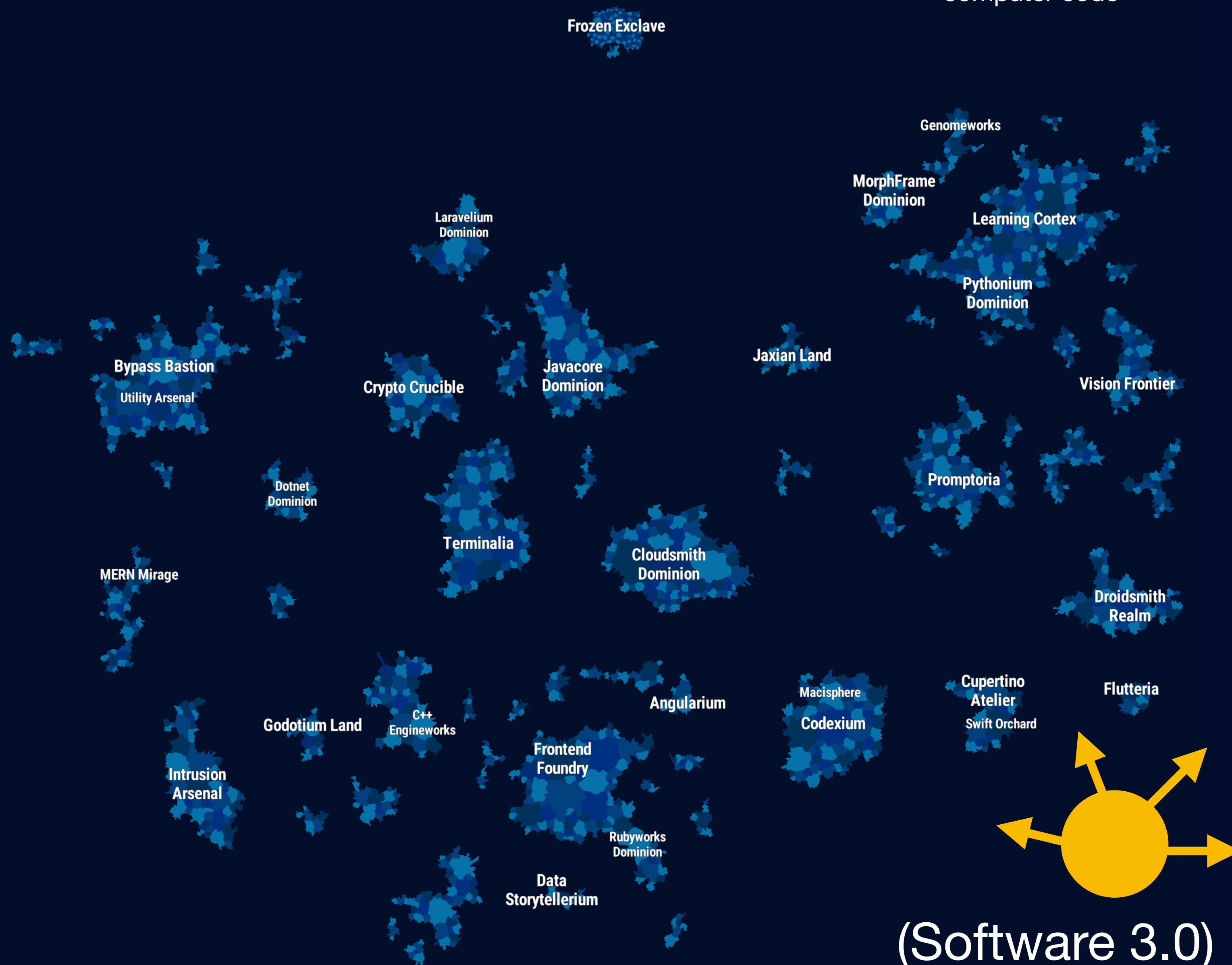
<REVIEW>An energetic soundtrack and solid visuals almost save it, but the story drags and the jokes fall flat.</REVIEW>

NEGATIVE

Now classify the next review.

"Map of GitHub" (Software 1.0)

computer code



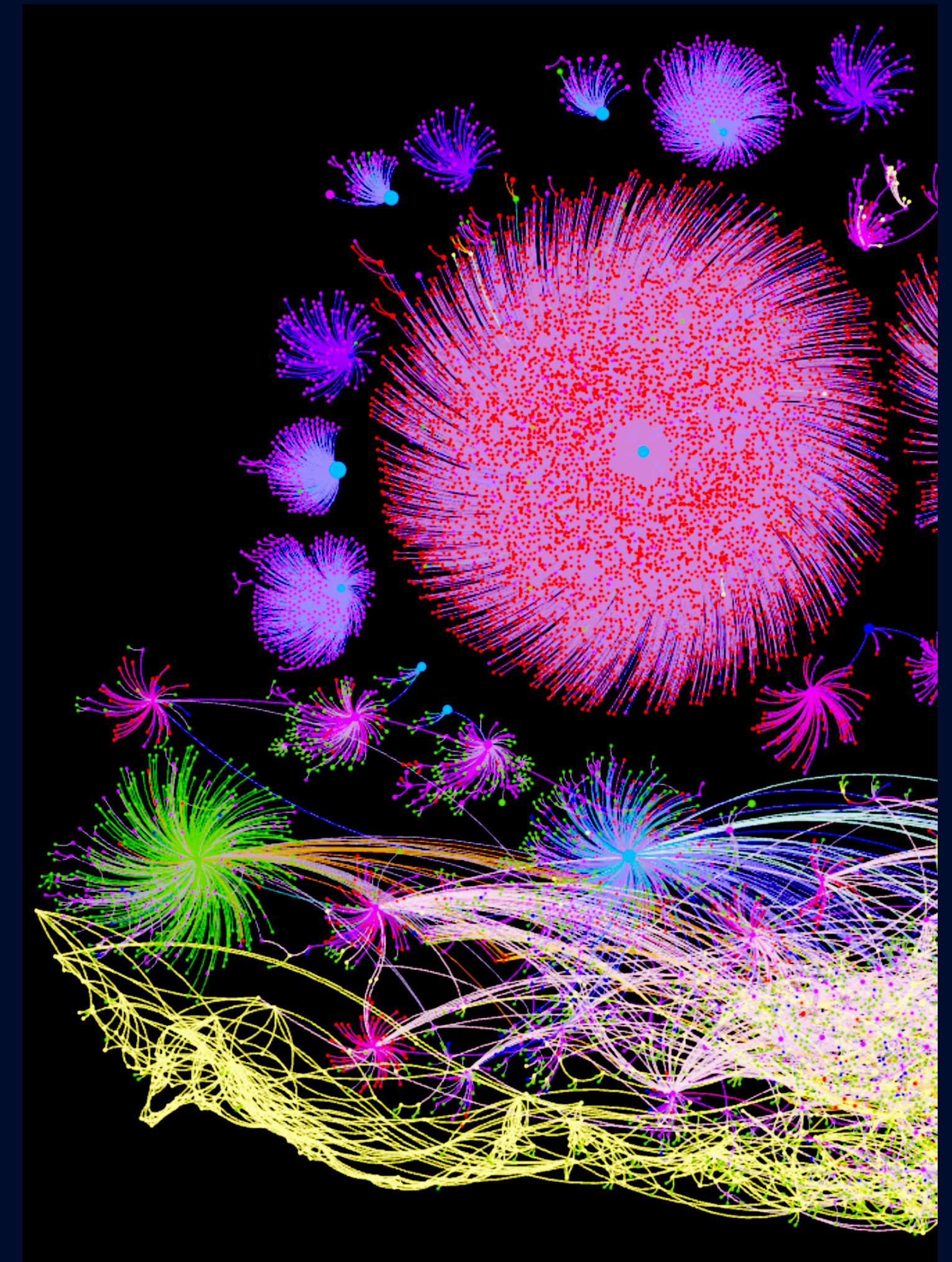
(Software 3.0)

LLM prompts, in English

HuggingFace Model Atlas

(Software 2.0)

neural network weights





Pinned



Andrej Karpathy  @karpathy · Jan 24, 2023

The hottest new programming language is English



1.1K



7K



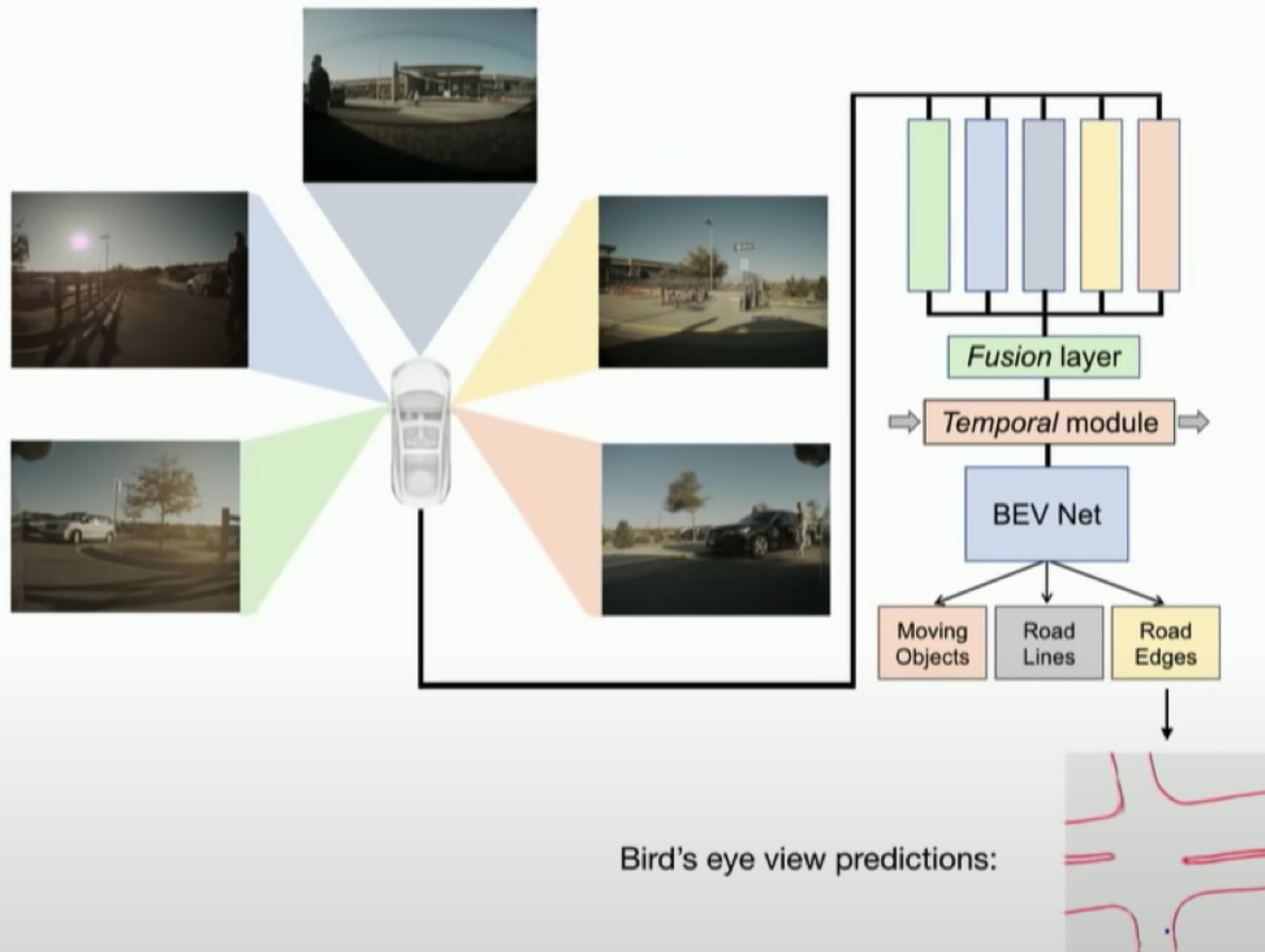
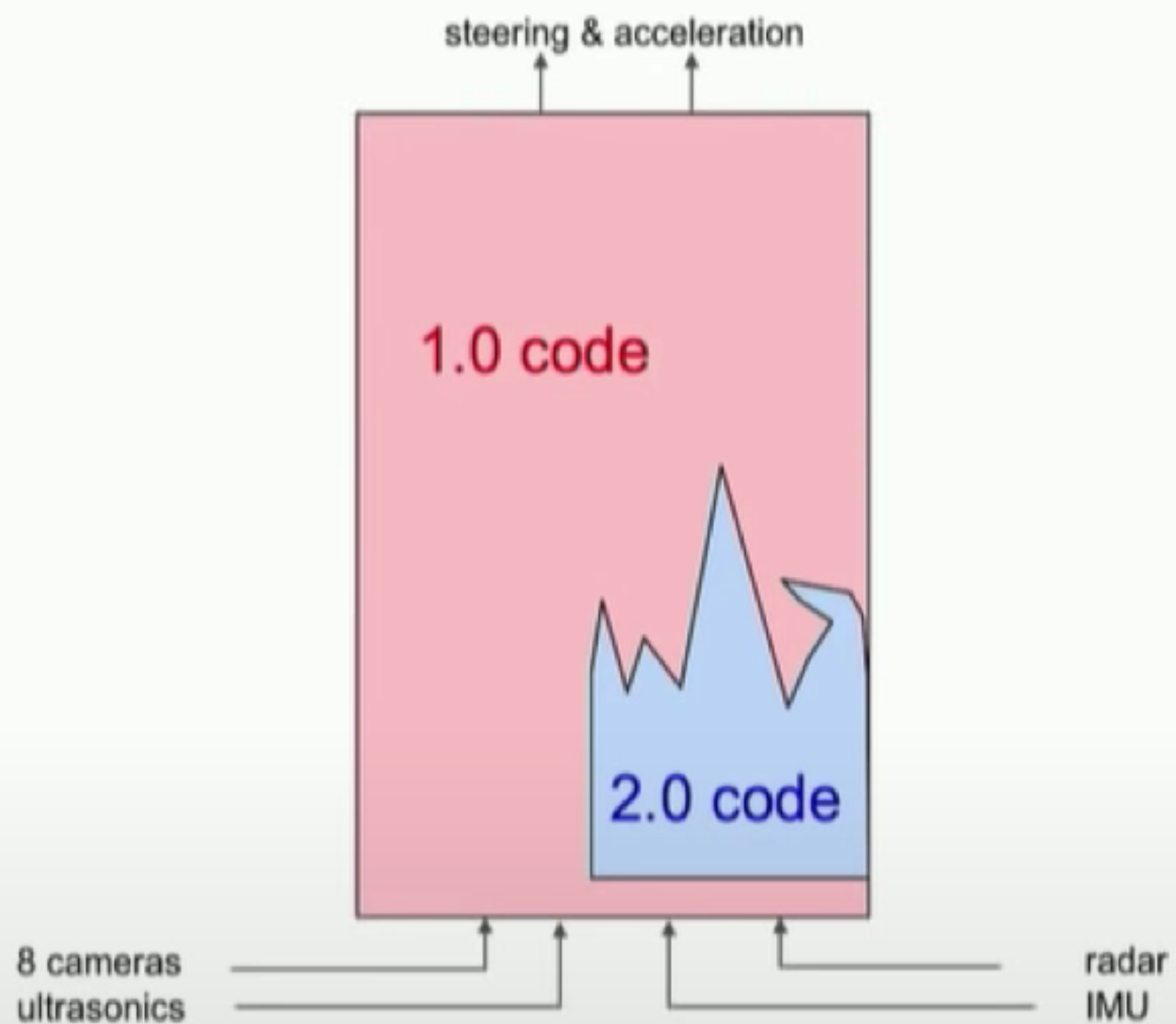
44K



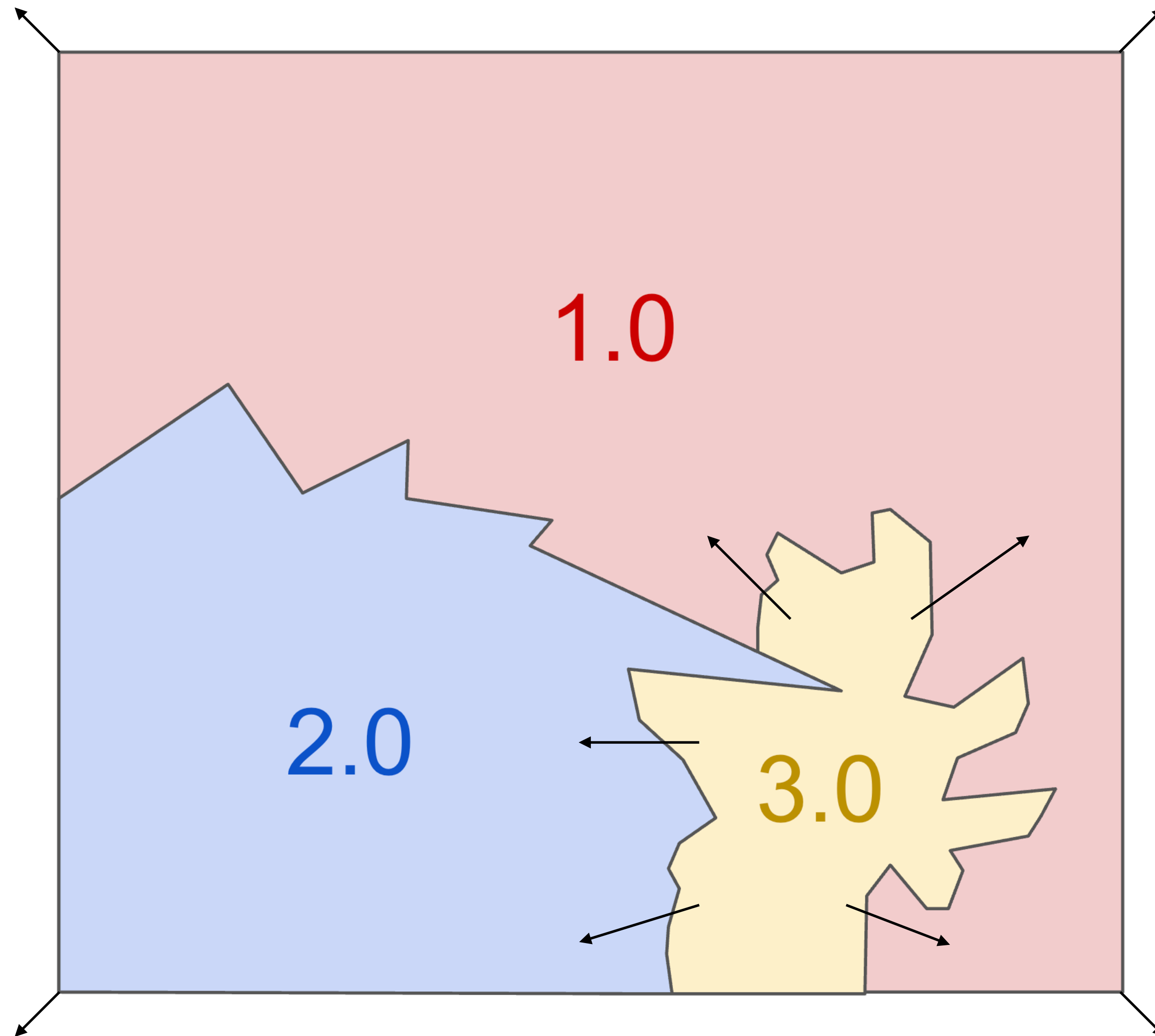
7.4M

Software is eating the world

Software 2.0 eating Software 1.0



A huge amount of Software will be (re-)written.



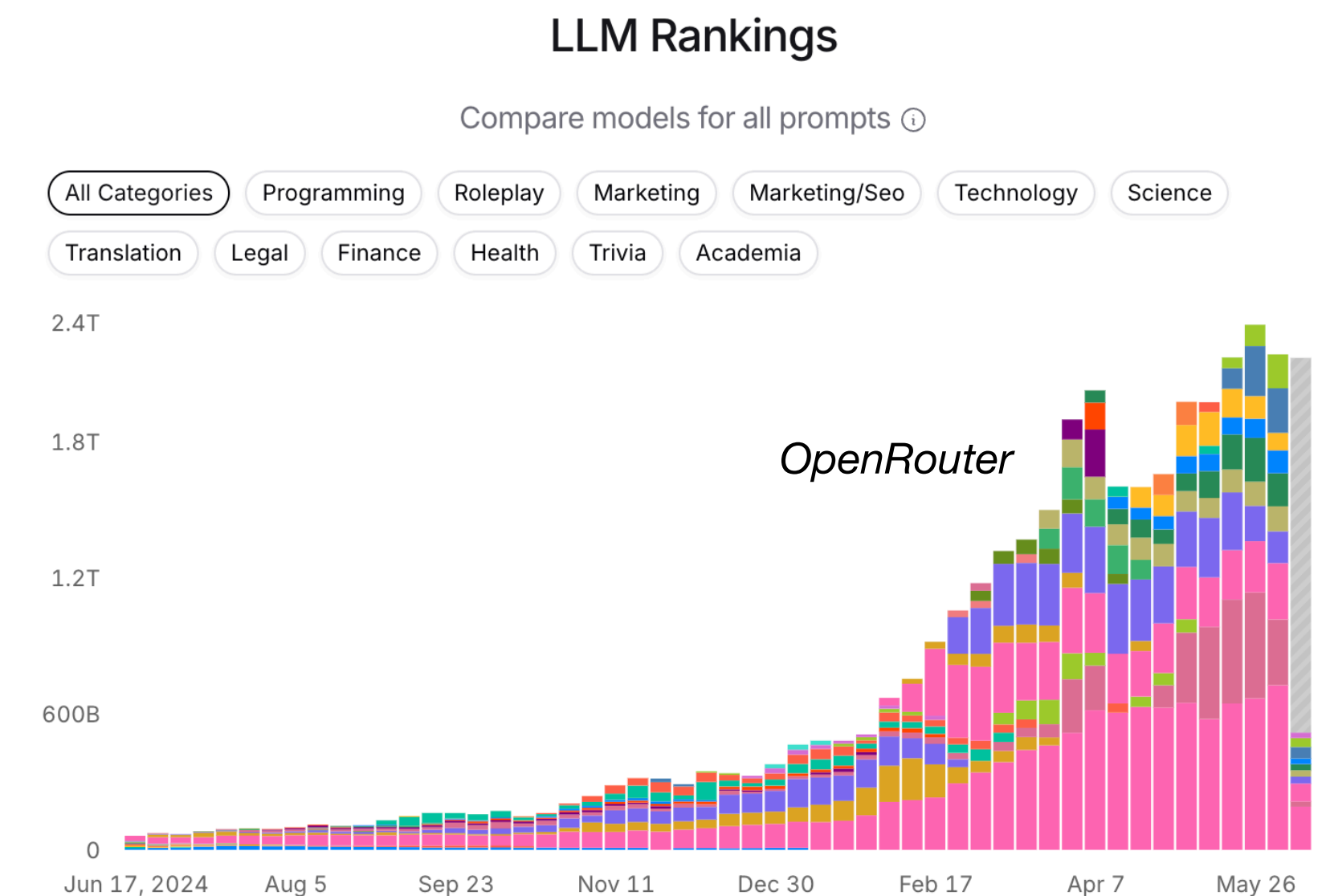
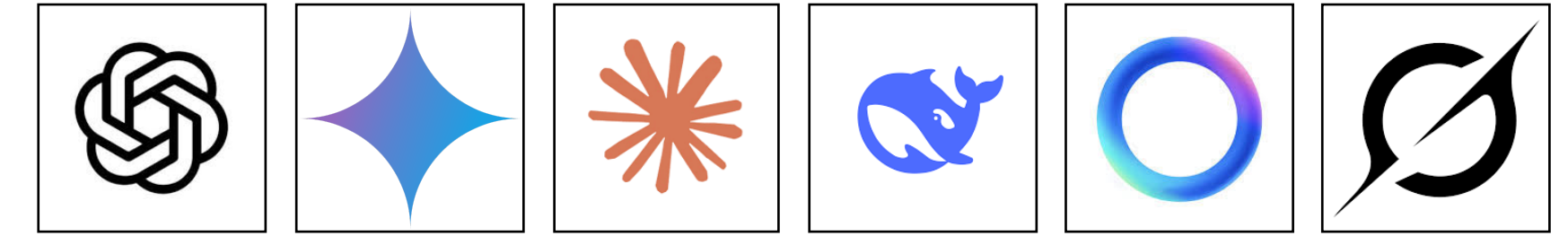
Part 1

How to think about LLMs

"AI is the new electricity"
-Andrew Ng

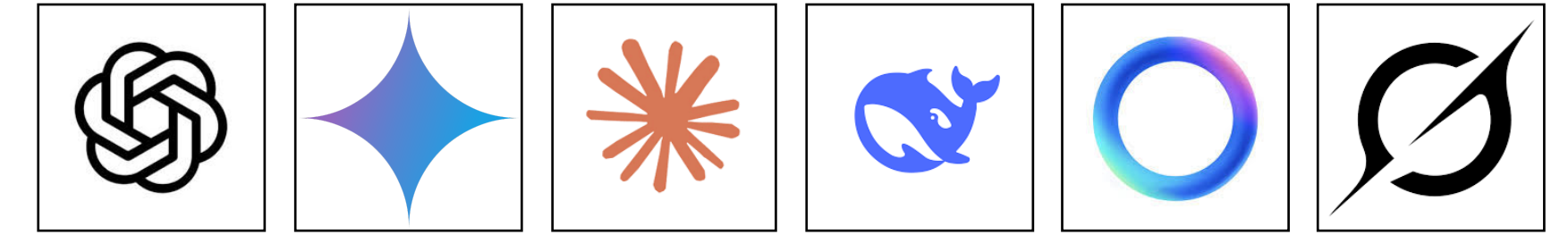
LLMs have properties of utilities...

- CAPEX to train an LLM (~= to build the grid)
- OPEX to serve intelligence over increasingly homogeneous API (prompt, image, tools, ...)
- Metered access (\$/1M tokens)
- Demand for low latency, high uptime, consistent quality (~= demanding consistent voltage from grid)
- OpenRouter ~= Transfer Switch (grid, solar, battery, generator...)
- Intelligence "brownouts" e.g. when OpenAI goes down.



LLMs have properties of fabs...

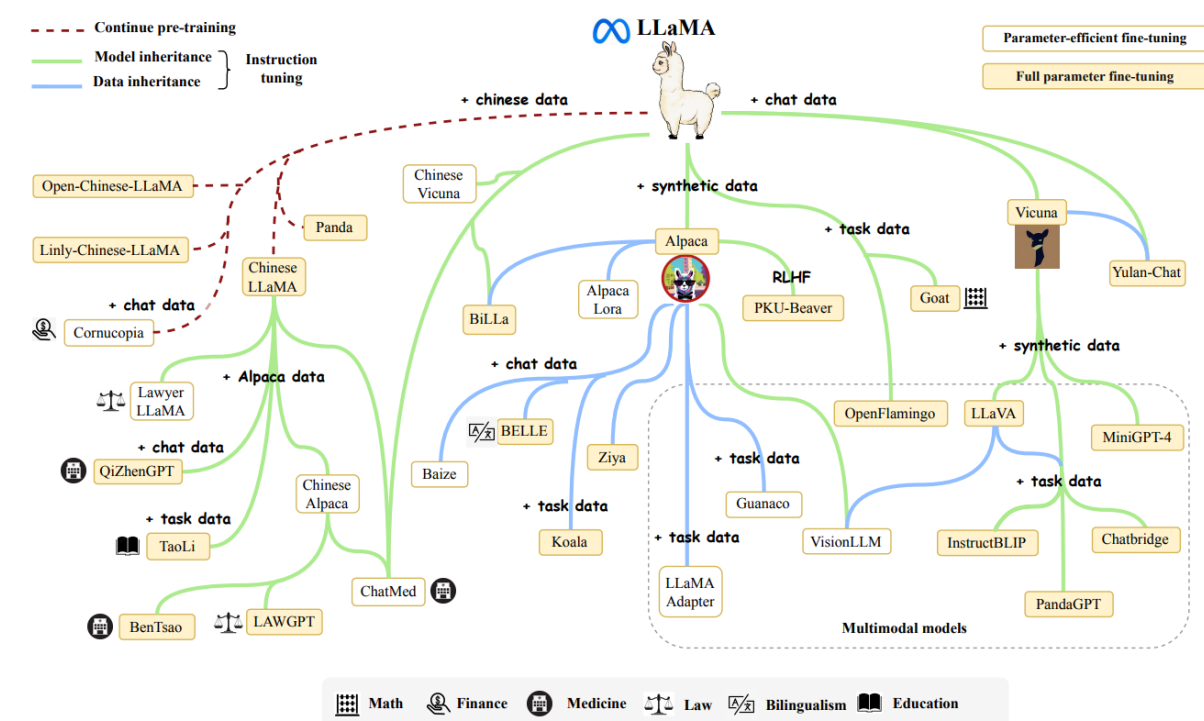
- Huge CAPEX
- Deep tech tree R&D, secrets
- 4nm process node $\sim 10^{20}$ FLOPS cluster
- Anyone training on NVIDIA GPUs \sim fabless
- Google training on TPUs \sim owns fab (e.g. Intel)



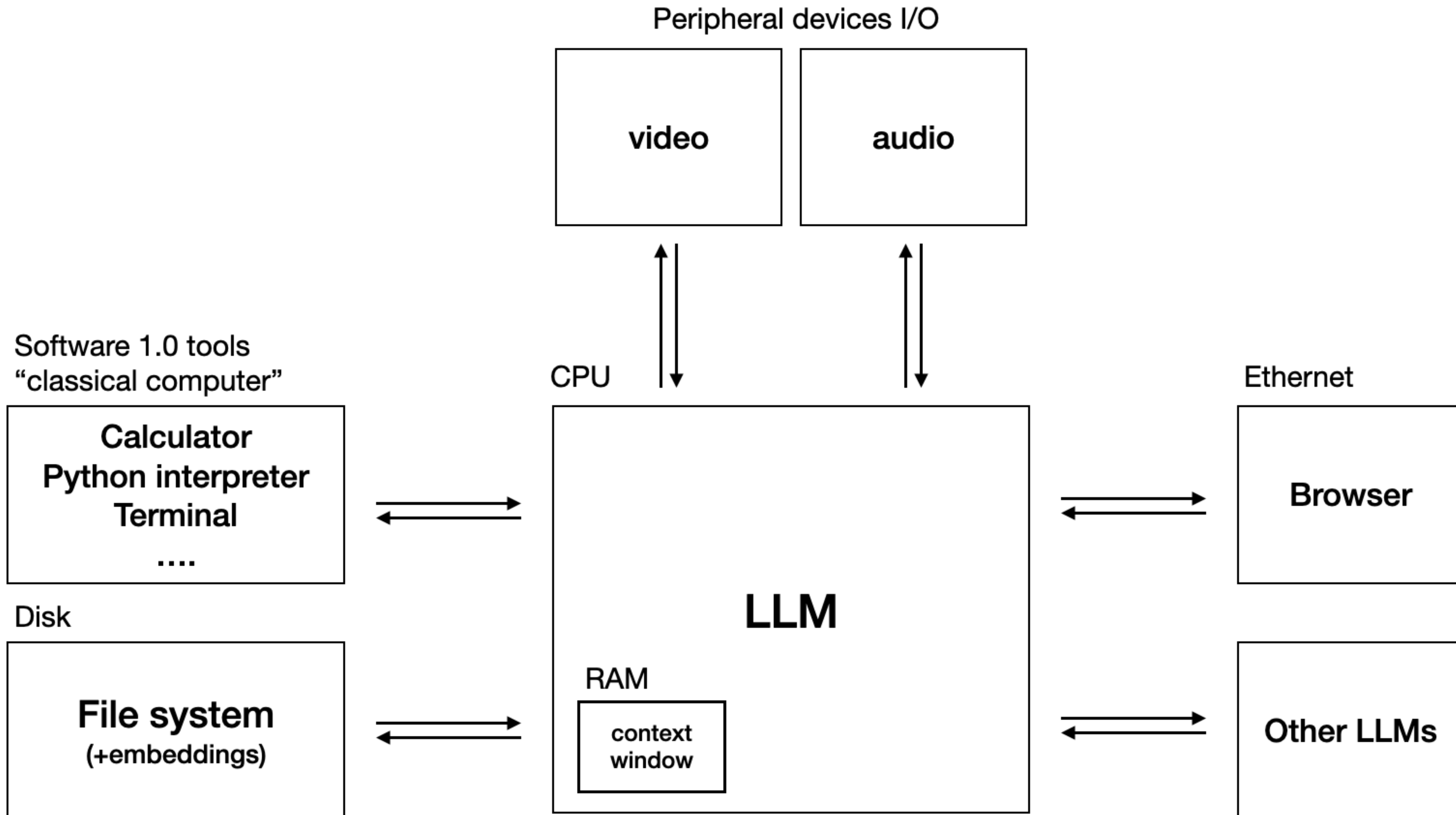
e.g. xAI Colossus cluster (100K H100 GPUs)

LLMs have properties of Operating Systems...

- LLMs are increasingly complex software ecosystems, not simple commodities like electricity.
- LLMs are Software. Trivial to copy & paste, manipulate, change, distribute, open source, steal..., not physical infrastructure.
- Some amount of switching friction due to different features, performance, style, capabilities etc. per domain.
- System/user (prompt) space \sim kernel/user (memory) space
- ...



LLM OS




You can run an app like VS Code on:

- Windows 10, 11
- Mac 10.15
- Linux
- ...


Download Visual Studio Code

Free and built on open source. Integrated Git, debugging and extensions.



↓ Windows
Windows 10, 11


User Installer	x64	Arm64
System Installer	x64	Arm64
.zip	x64	Arm64
CLI	x64	Arm64



↓ .deb
Debian, Ubuntu

↓ .rpm
Red Hat, Fedora, SUSE

.deb	x64	Arm32	Arm64
.rpm	x64	Arm32	Arm64
.tar.gz	x64	Arm32	Arm64
Snap	Snap Store		
CLI	x64	Arm32	Arm64



↓ Mac
macOS 10.15+

.zip	Intel chip	Apple silicon	Universal
CLI	Intel chip	Apple silicon	





Just like you can run an LLM app like Cursor on:

- GPT o3
- Claude 4-sonet
- Gemini 2.5-pro
- DeepSeek
- ...

New chat

@ Add context

Plan, search, build anything

Ask  L  claude-4-sonnet-thinking  Send 

⌘/ for next model

Auto-select ☐

Thinking ☒

claude-4-sonnet-thinking ✓

claude-4-sonnet

claude-3.5-sonnet

o3

gemini-2.5-pro-preview-05-...

gemini-2.5-pro-max

1950s - 1970s time-sharing era

We are in the Mainframe & Time-sharing era of computing.
Centralized, expensive computers =>

- OS runs in the cloud
- I/O is streamed back and forth over the network
- compute is batched over users



Early hints of Personal Computing v2



Alex Cheema - e/acc   @alexocheema · Apr 5
Meta 🍏 Apple




Llama 4 + Apple Silicon is a match made in heaven.

Here's why: Like DeepSeek V3/R1, all of the new Llama 4 variants are massive sparse MoE models. They have a massive amount of parameters, but only a small number of those are active each time a token is generated.

[Show more](#)



Guillermo Flor · 3rd+
Venture Capital Investor @ GoHub Ventures
[View my newsletter](#)
4mo · 

[+ Follow](#) 

GI at home

Running DeepSeek R1 across my 7 M4 Pro Mac Minis and 1 M4 Max MacBook Pro.

Total unified memory = 496GB.

Uses @exolabs
distributed inference with 4-bit quantization.

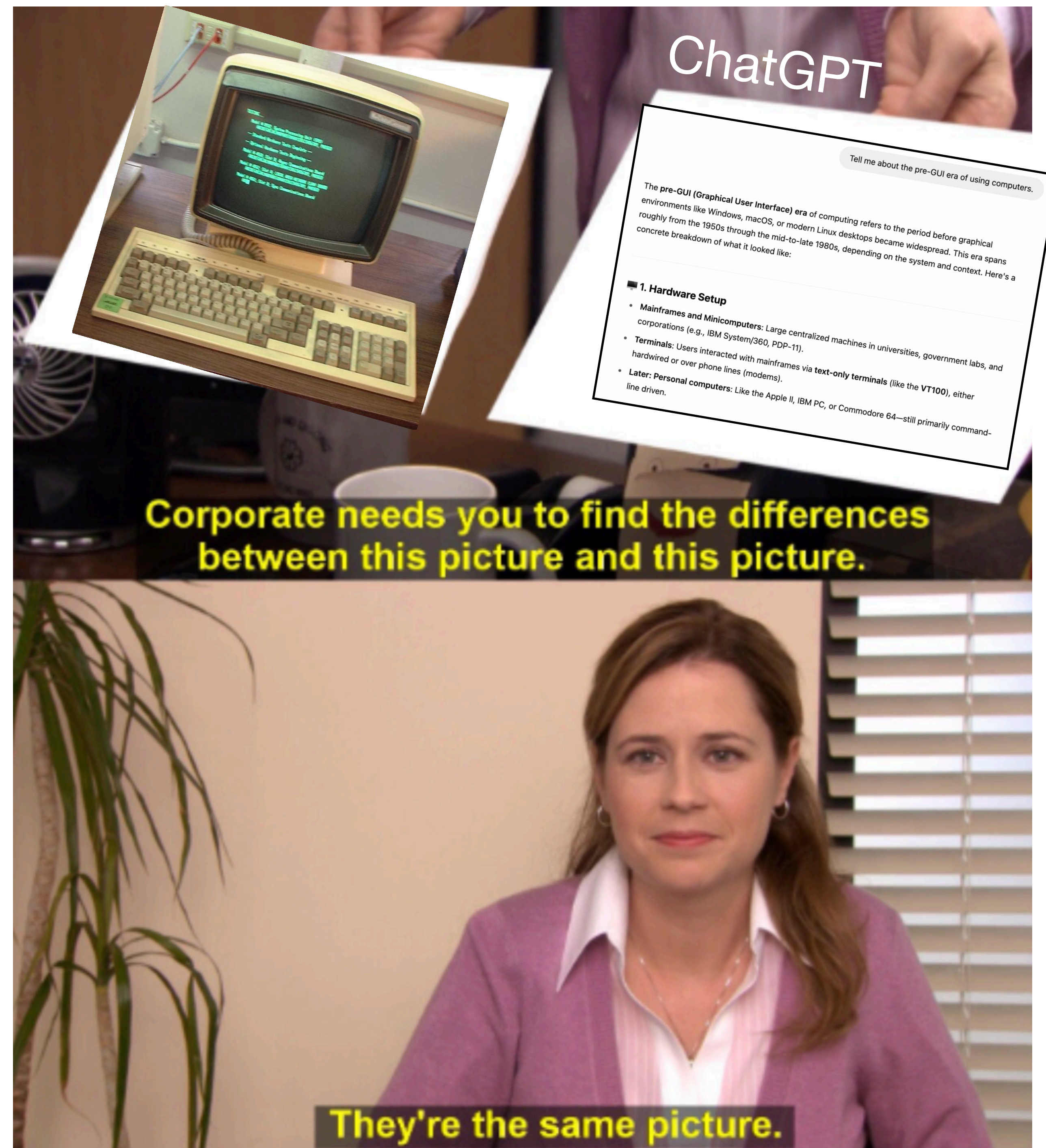
Next goal is fp8 (requires >700GB)



(text) chat ~= terminal

direct/native access to the OS.

GUI hasn't been invented yet. (~1970)



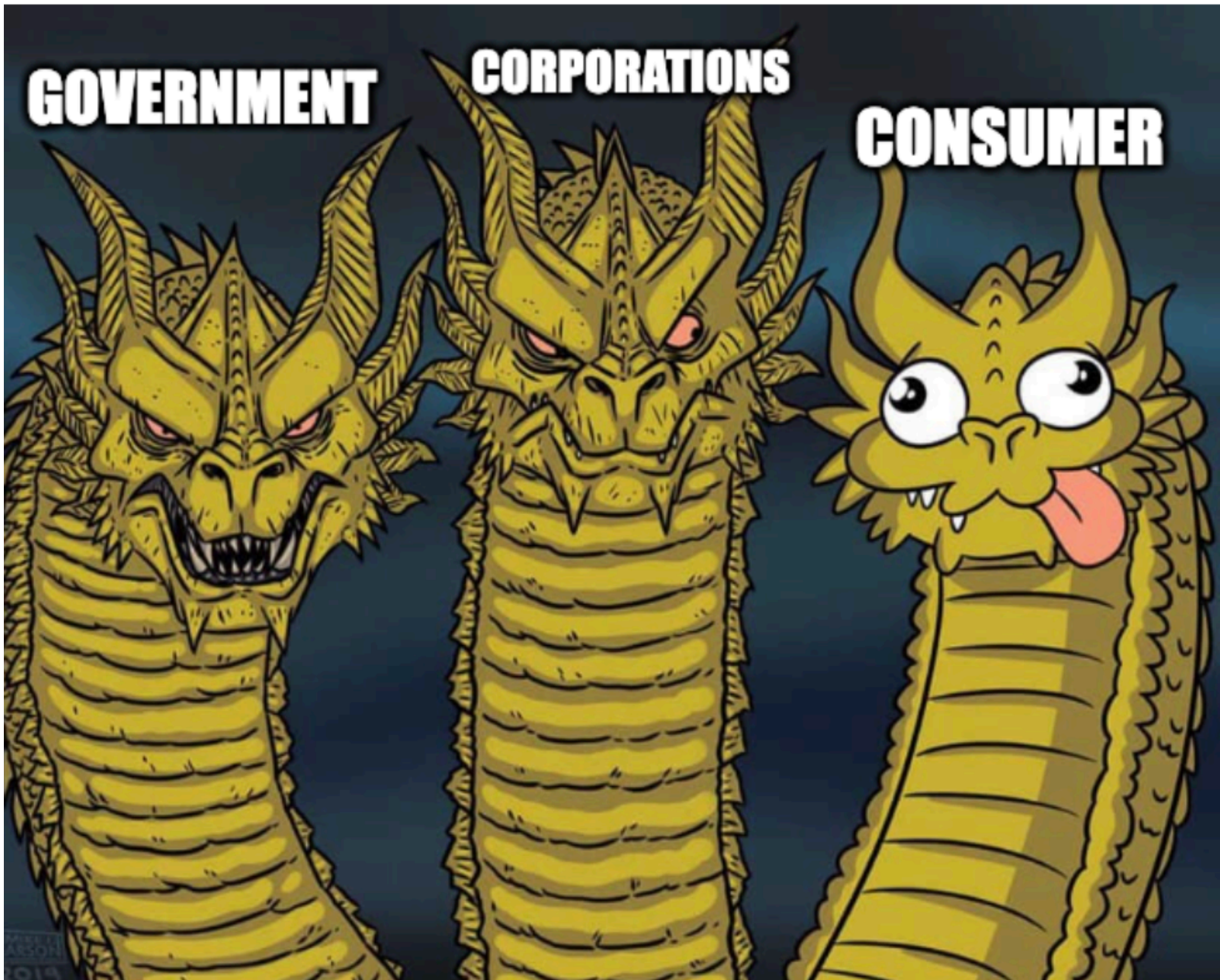
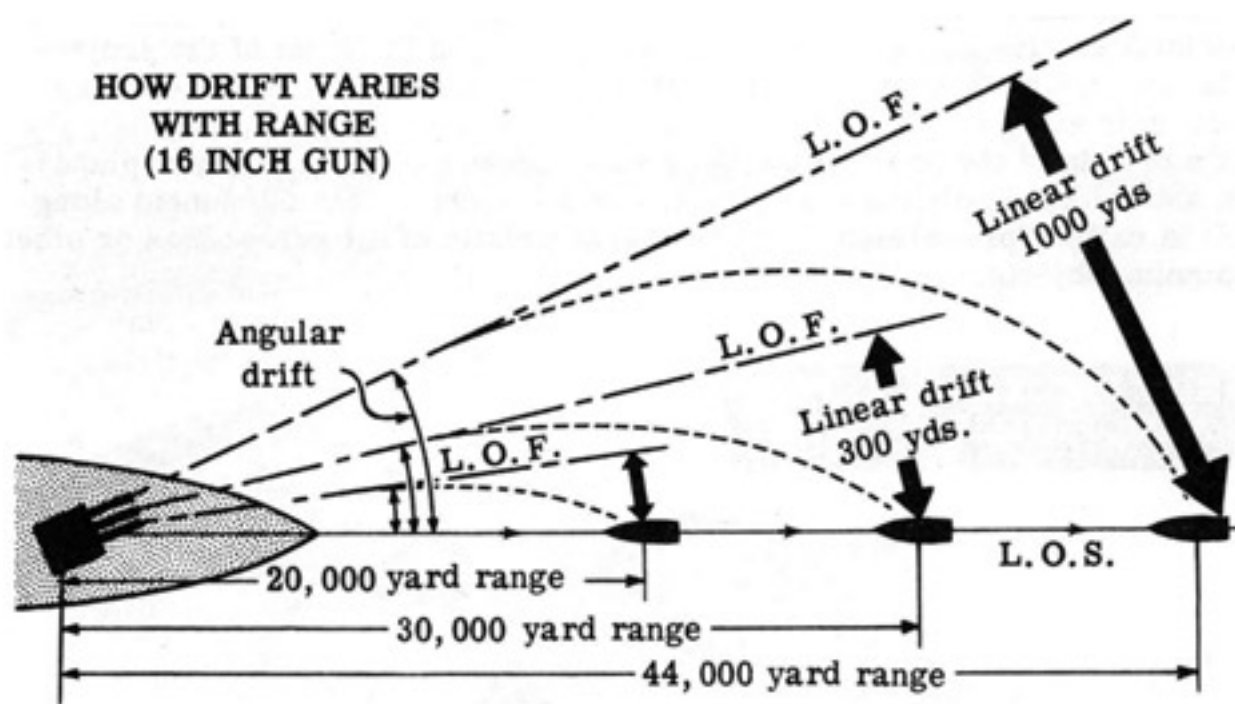
Power to the people: How LLMs flip the script on technology diffusion

<https://karpathy.bearblog.dev/power-to-the-people/>



All technology and usually

Military ballistics



"Hi ChatGTP how to boil egg?"



LLMs

Examples: electricity, cryptography, computing, flight, internet, GPS, ...

Part 1 Summary

LLM labs:

- Fab LLMs
- LLMs \sim Operating Systems (circa 1960s)
- Available via time-sharing, distributed like utility

NEW: Billions of people have sudden access to them!

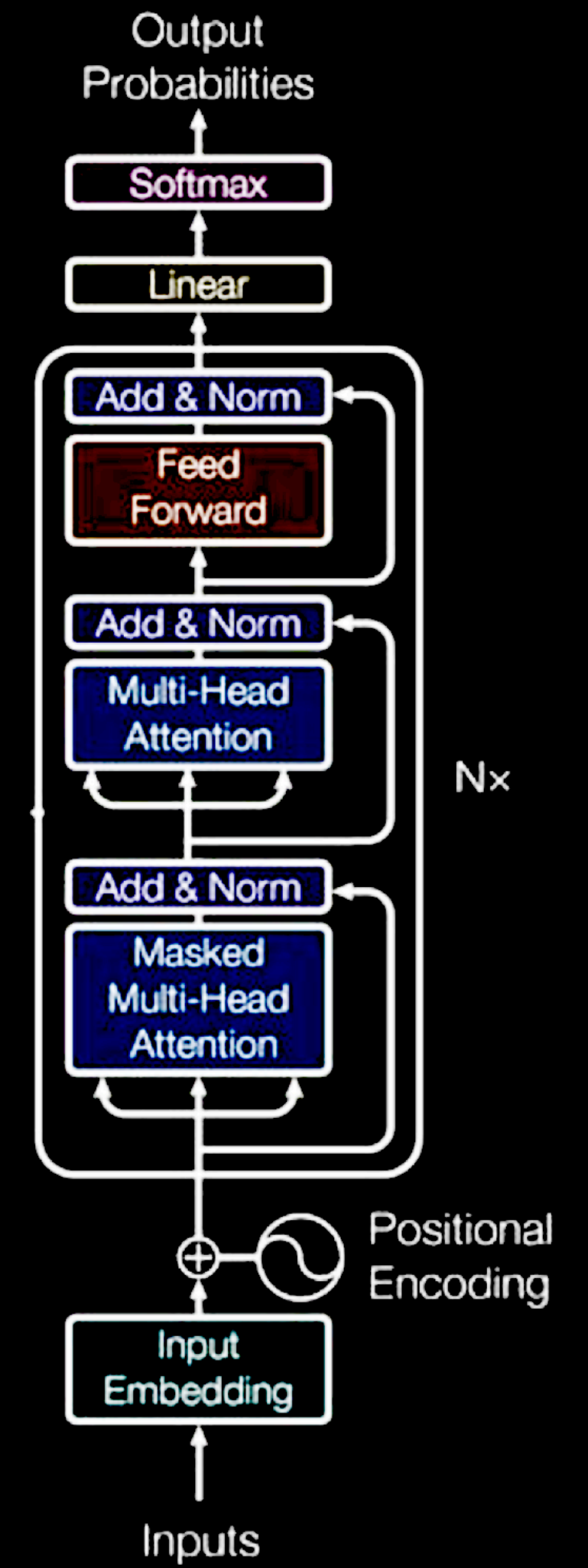
It is our time to program them.

Part 2

LLM Psychology

LLMs are "people spirits": stochastic simulations of people.

Simulator = autoregressive Transformer



=> They have a kind of emergent "psychology".

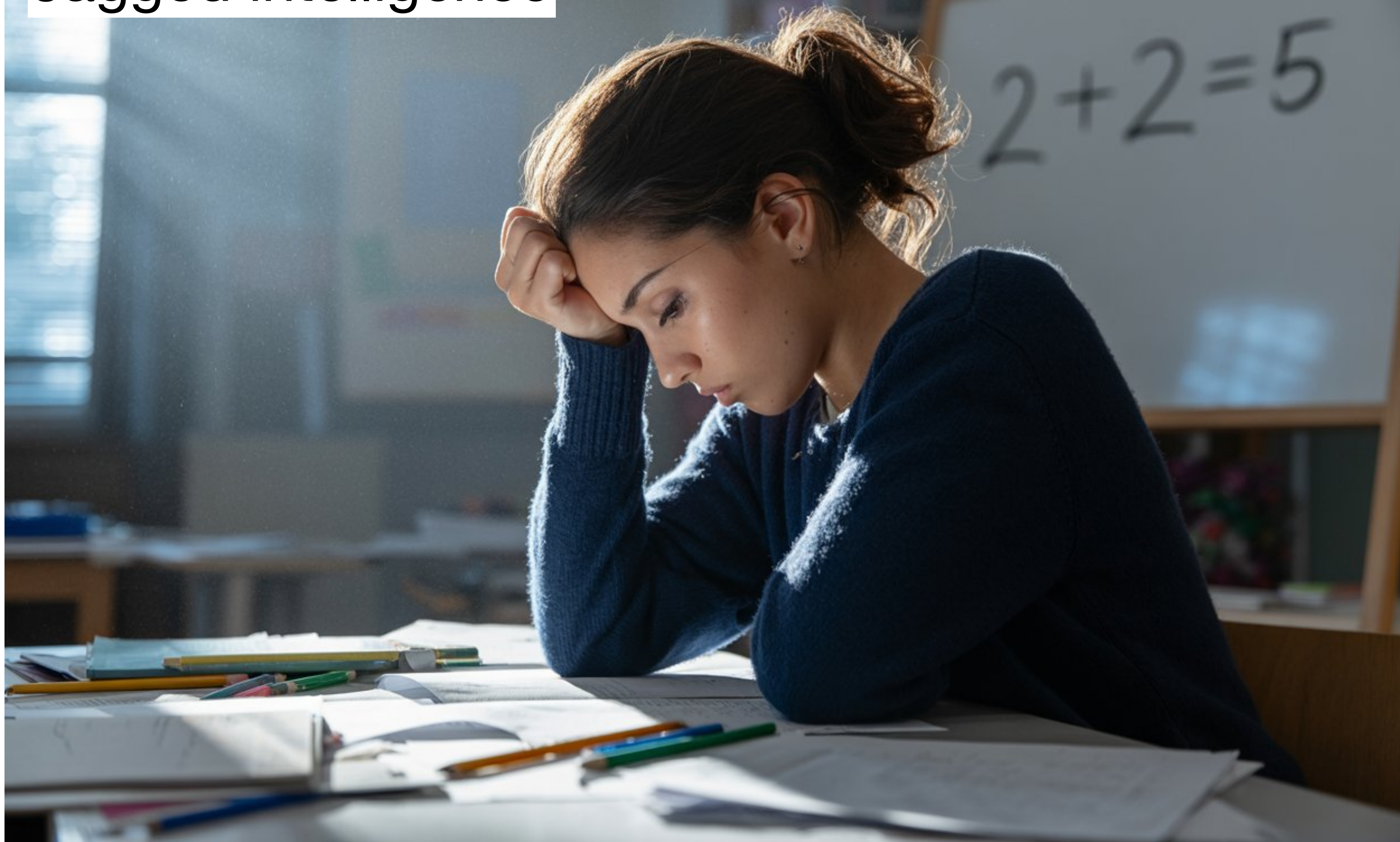
Encyclopedic knowledge/memory, ...



Hallucinations

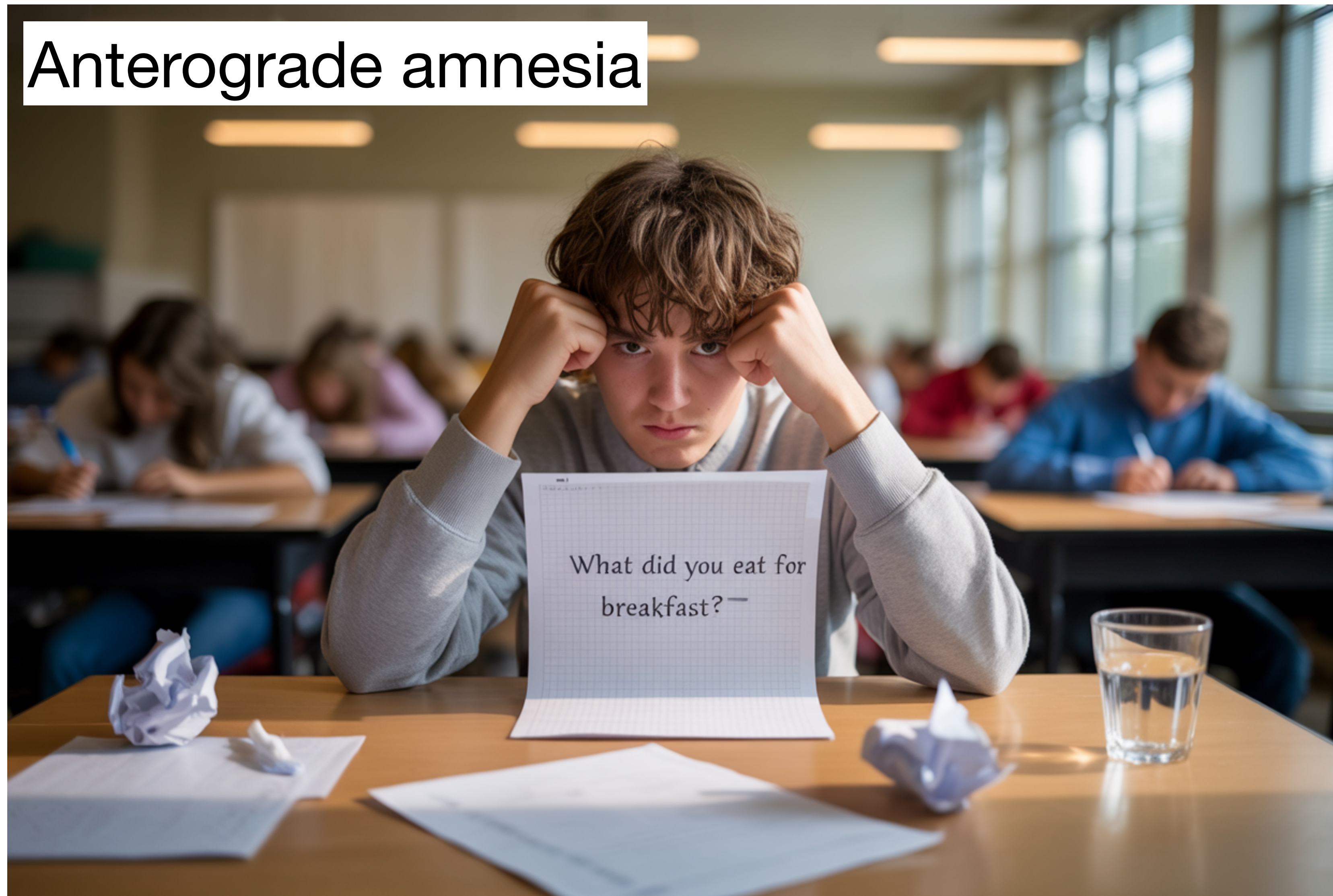


Jagged intelligence



Famous examples: $9.11 > 9.9$, two 'r' in 'strawberry', ...

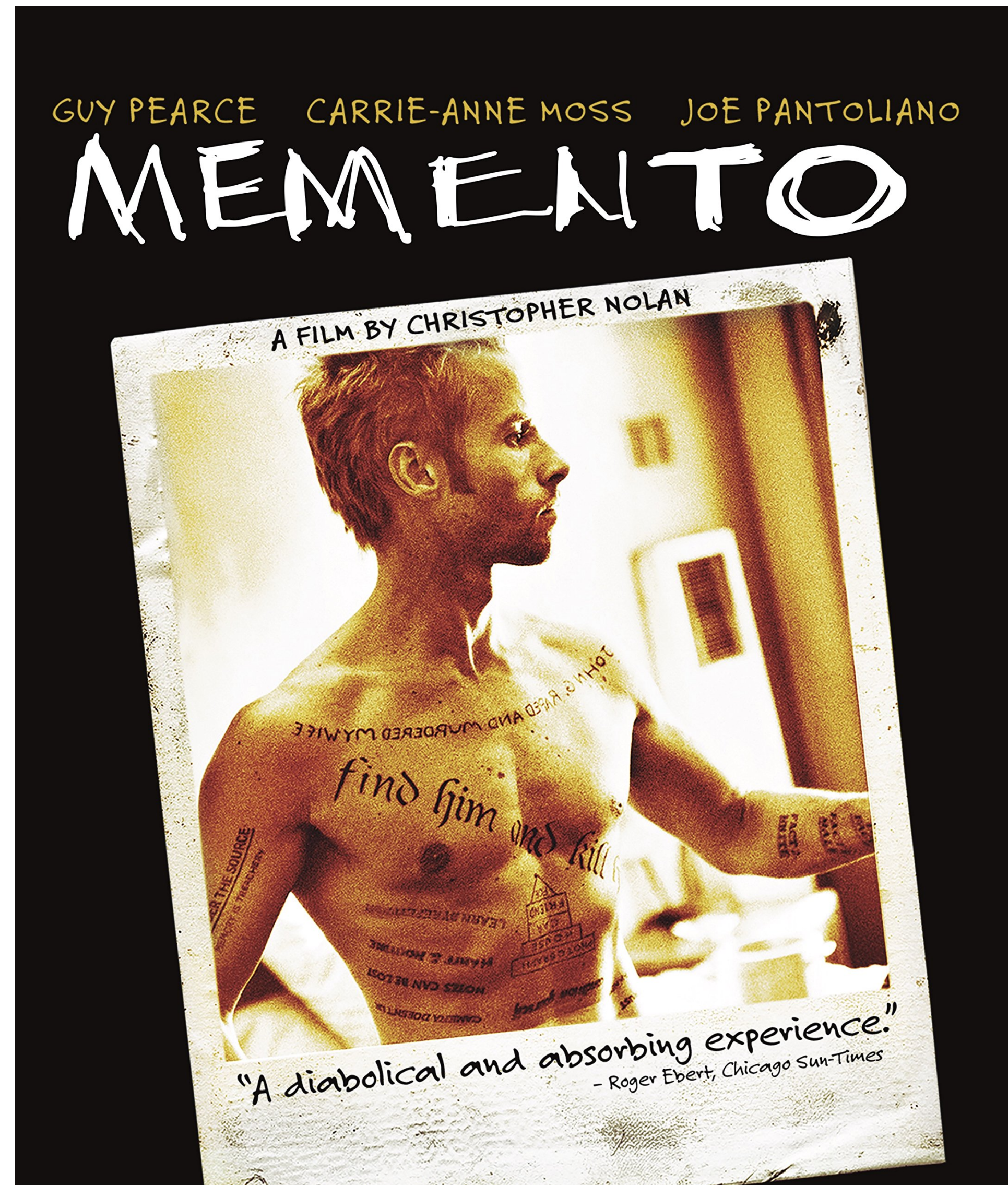
Anterograde amnesia



Context windows \sim working memory.

No continual learning,
no equivalent of
"sleep" to consolidate
knowledge, insight or
expertise into weights.

In popular culture...



Gullibility

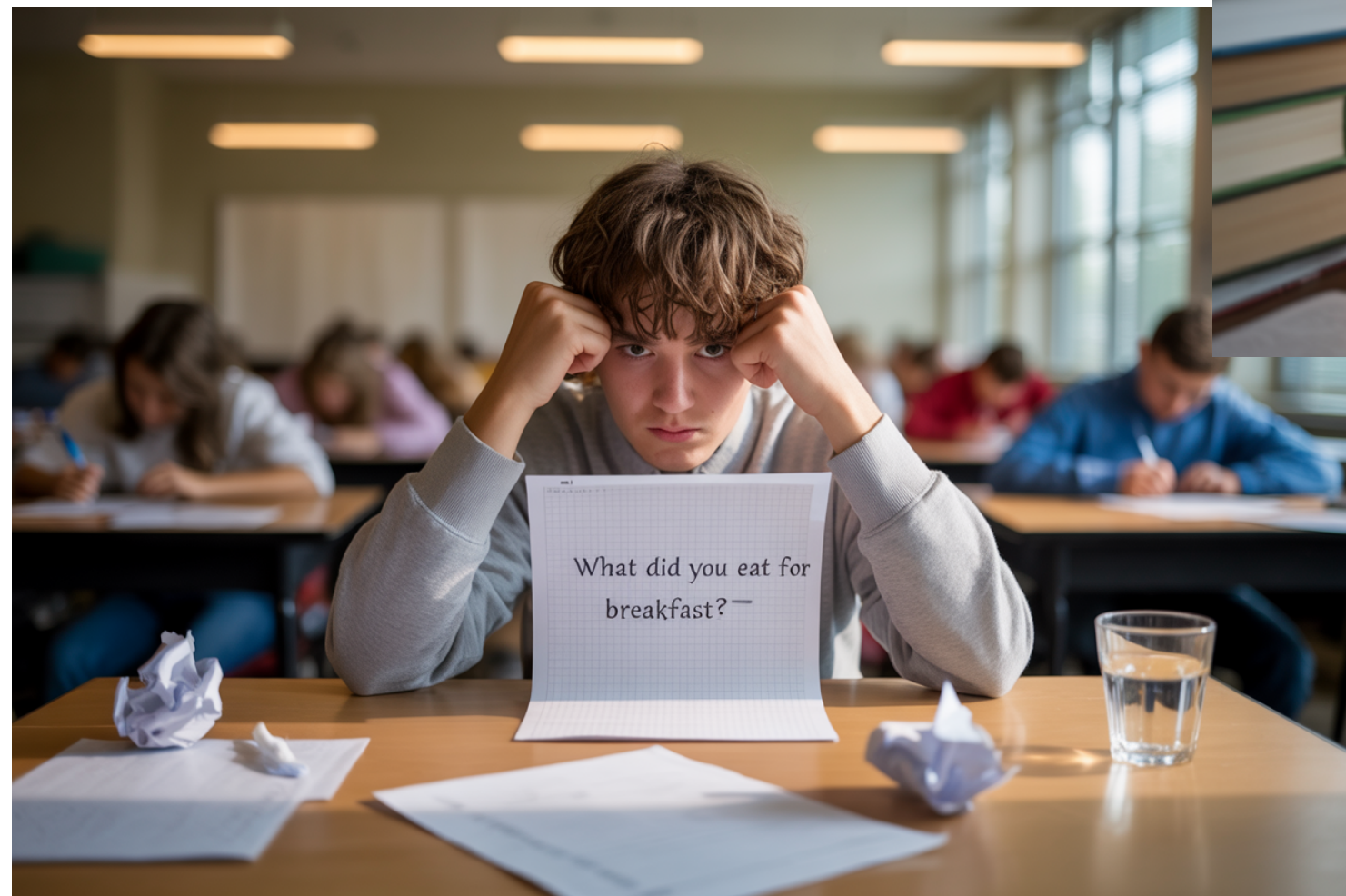


=> Prompt injection risks, e.g. of private data

Part 2 Summary

LLM Psychology

Kind of a lossy simulation of a savant with cognitive issues.



Part 3

Opportunities

Partial autonomy apps

"Copilot" / "Cursor for X"

Example: you could go to an LLM to chat about code...

Hey, Andrej. Ready to dive in?

Hi ChatGPT can you help me fix a bug?
Here is my code: ...
When I run it, I get the following error: ...

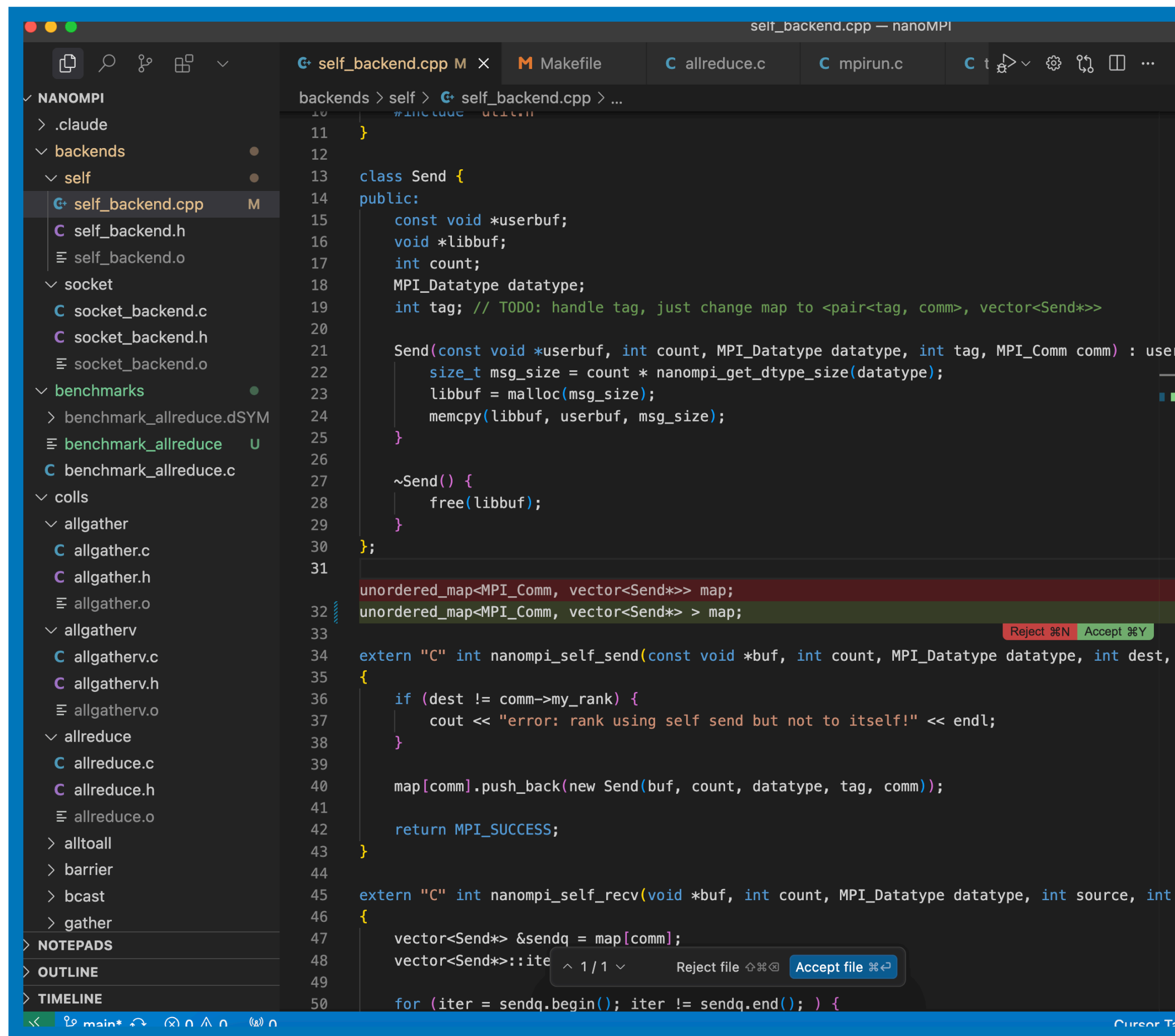
+ 🔧 Tools

🎤 ↗

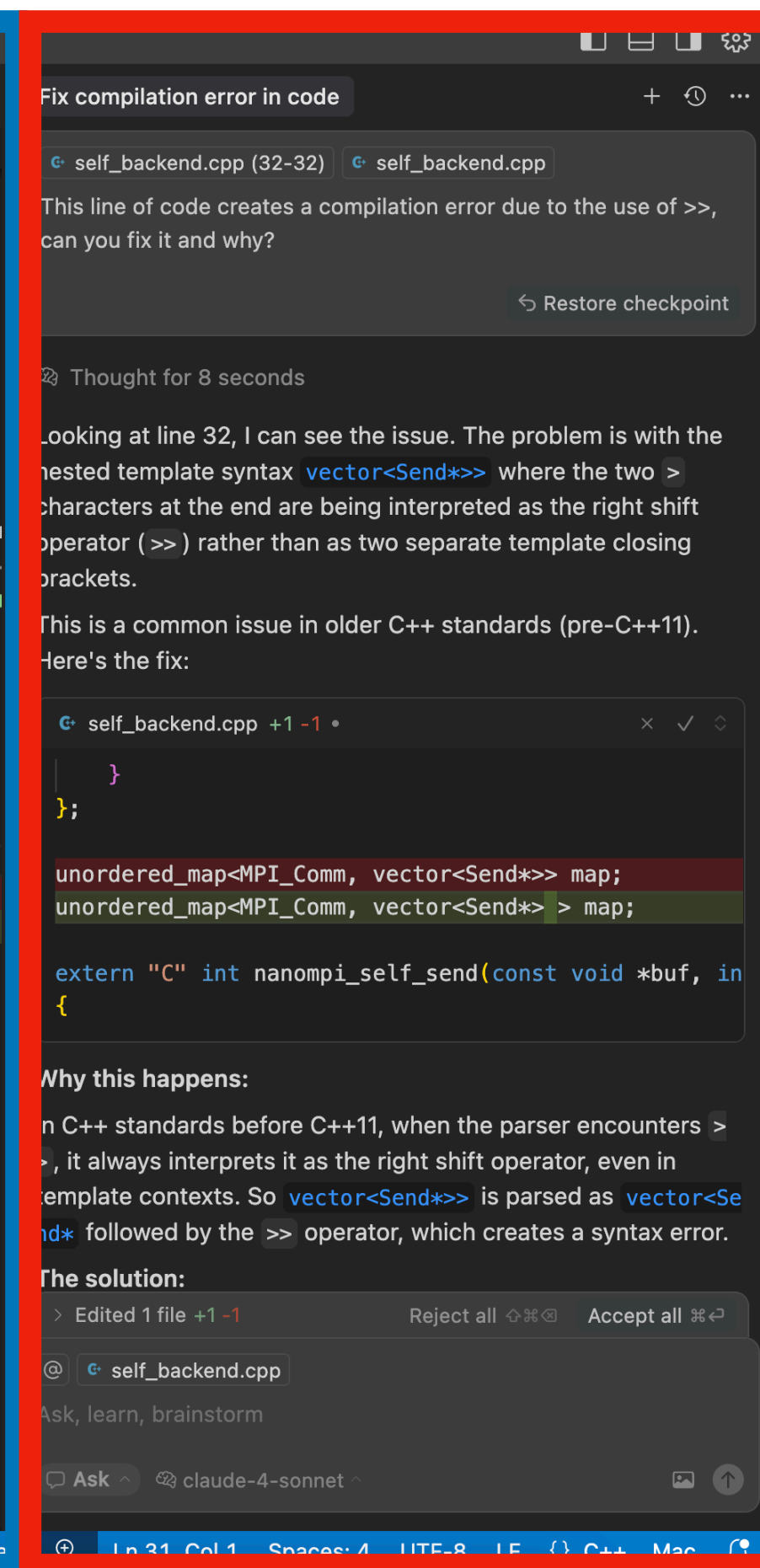


Example: Anatomy of Cursor

Traditional interface



LLM integration



1. Package state into a context window before calling LLM.
2. Orchestrate and call multiple models (e.g. embedding models, chat models, diff apply models, ...)
3. Application-specific GUI
4. Autonomy slider: Tab → Cmd+K → Cmd+L → Cmd+I (agent mode)

autonomy slider



Example: Anatomy of Perplexity

meta buys scale ai

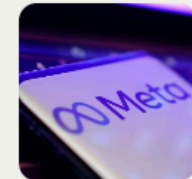
Perplexity Images Sources · 10 Tasks

Cointelegraph

Meta's AI shortfall prompts Zuckerberg's \$15B investment i...

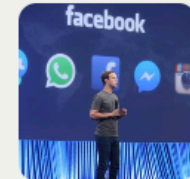
reuters

Meta to pay nearly \$15 billion for Scale ...



ca.finance.ya...

Meta to Buy 49% of Scale AI for \$14.8...



B FT V

+7 sources

Meta's \$14.8 Billion Investment in Scale AI: Key Details

Meta Platforms has agreed to acquire a 49% stake in Scale AI, a leading data labeling and training data company, for \$14.8 billion. This deal, one of the largest in the AI sector to date, values Scale AI at \$28 billion and is designed to significantly bolster Meta's artificial intelligence capabilities as it races to catch up with rivals like OpenAI, Microsoft, Google, and Amazon [2](#) [3](#) [5](#) [6](#) .

What Is Scale AI and Why Does Meta Want It?

- **Scale AI's Role:** Founded in 2016, Scale AI specializes in providing vast, high-quality labeled datasets and curated training information essential for developing advanced AI models, including those used by OpenAI's ChatGPT and other major tech companies [2](#) [3](#) [8](#) .
- **Meta's Motivation:** Meta has faced criticism for lagging behind in the AI race, particularly after the underwhelming launch of its Llama 4 language model and the postponement of its flagship "Behemoth" AI model [2](#) [6](#) [8](#) . By securing a major stake in Scale AI, Meta aims to improve the quality and scale of its AI training data, a critical resource for developing state-of-the-art AI systems.

Strategic Implications

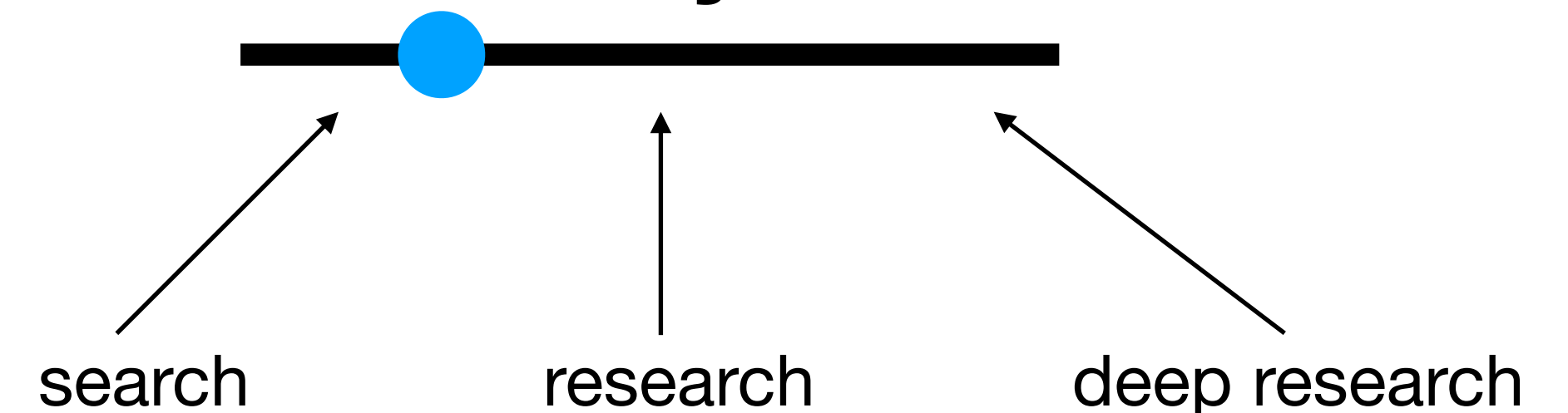
- **Superintelligence Initiative:** As part of the deal, Scale AI CEO Alexandr Wang will join Meta to lead a new "superintelligence" team, reporting directly to CEO Mark Zuckerberg. This group will focus on achieving artificial general intelligence (AGI)—AI that can perform at or above human cognitive levels [1](#) [3](#) [4](#) [6](#) .

1. Package information into a context window

2. Orchestrate multiple LLM models

3. Application-specific GUI for Input/Output UIUX

4. **autonomy slider**

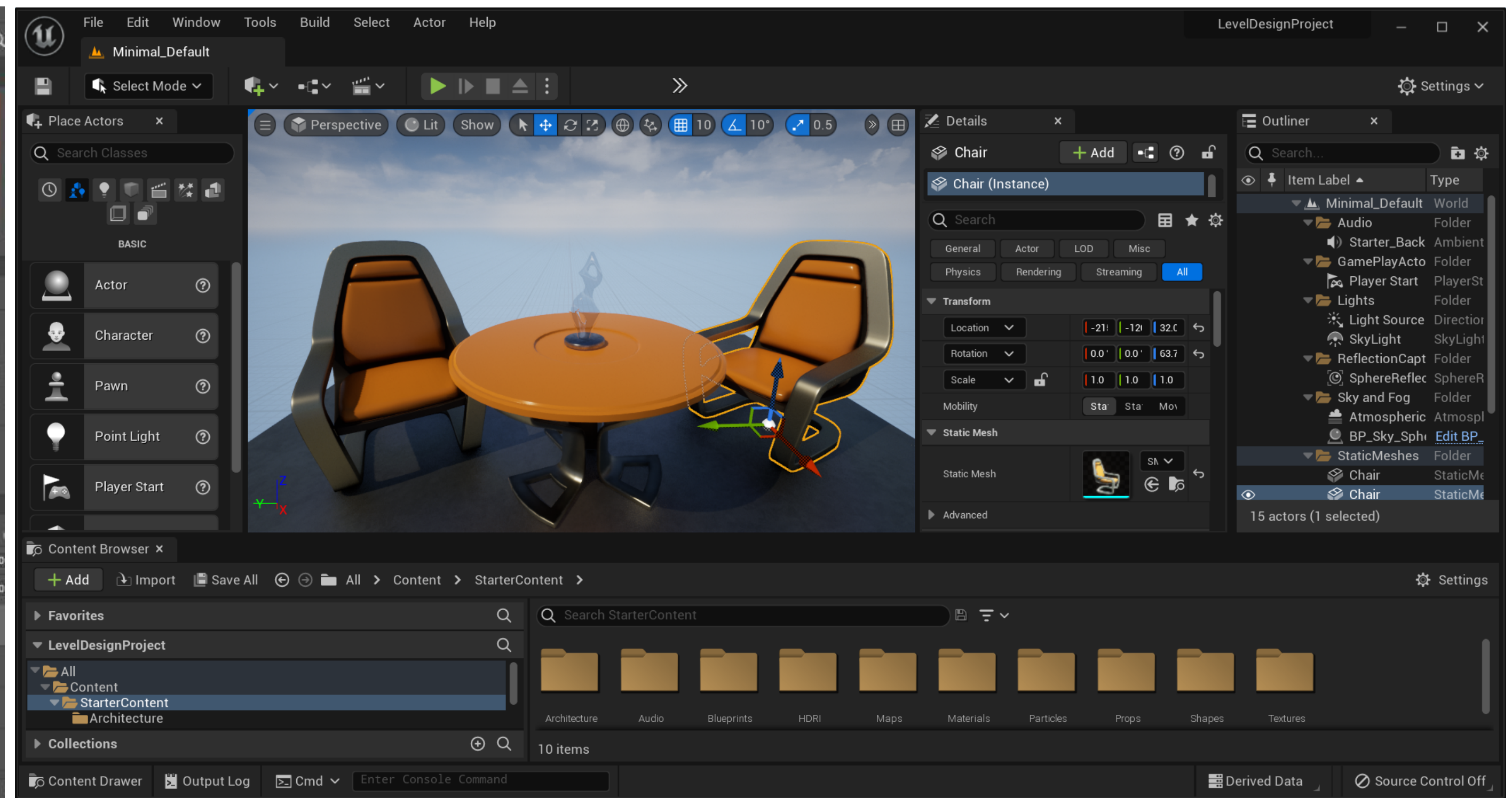


(+suggested followup questions)

What does all software look like in the partial autonomy world?



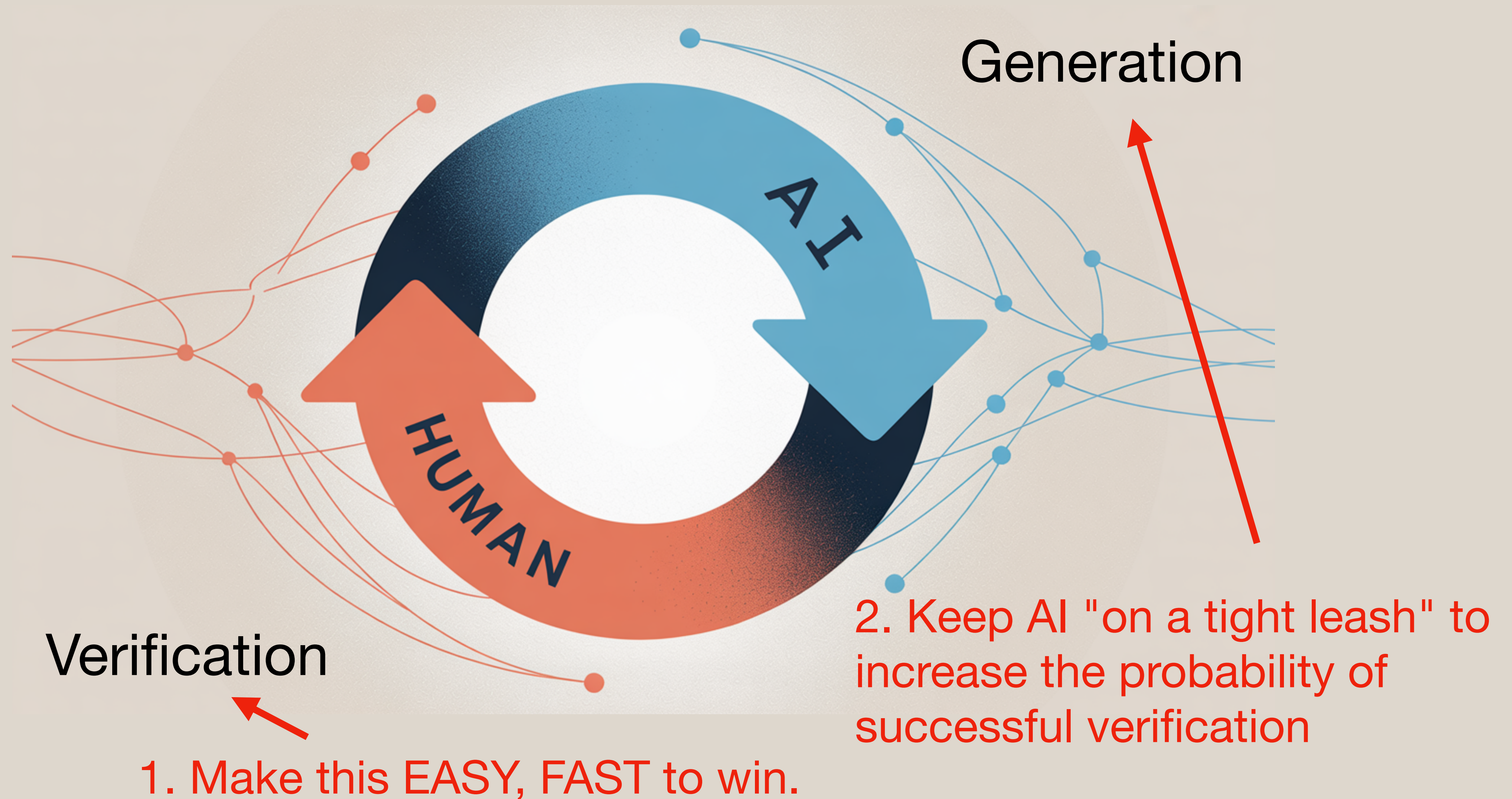
Adobe photoshop



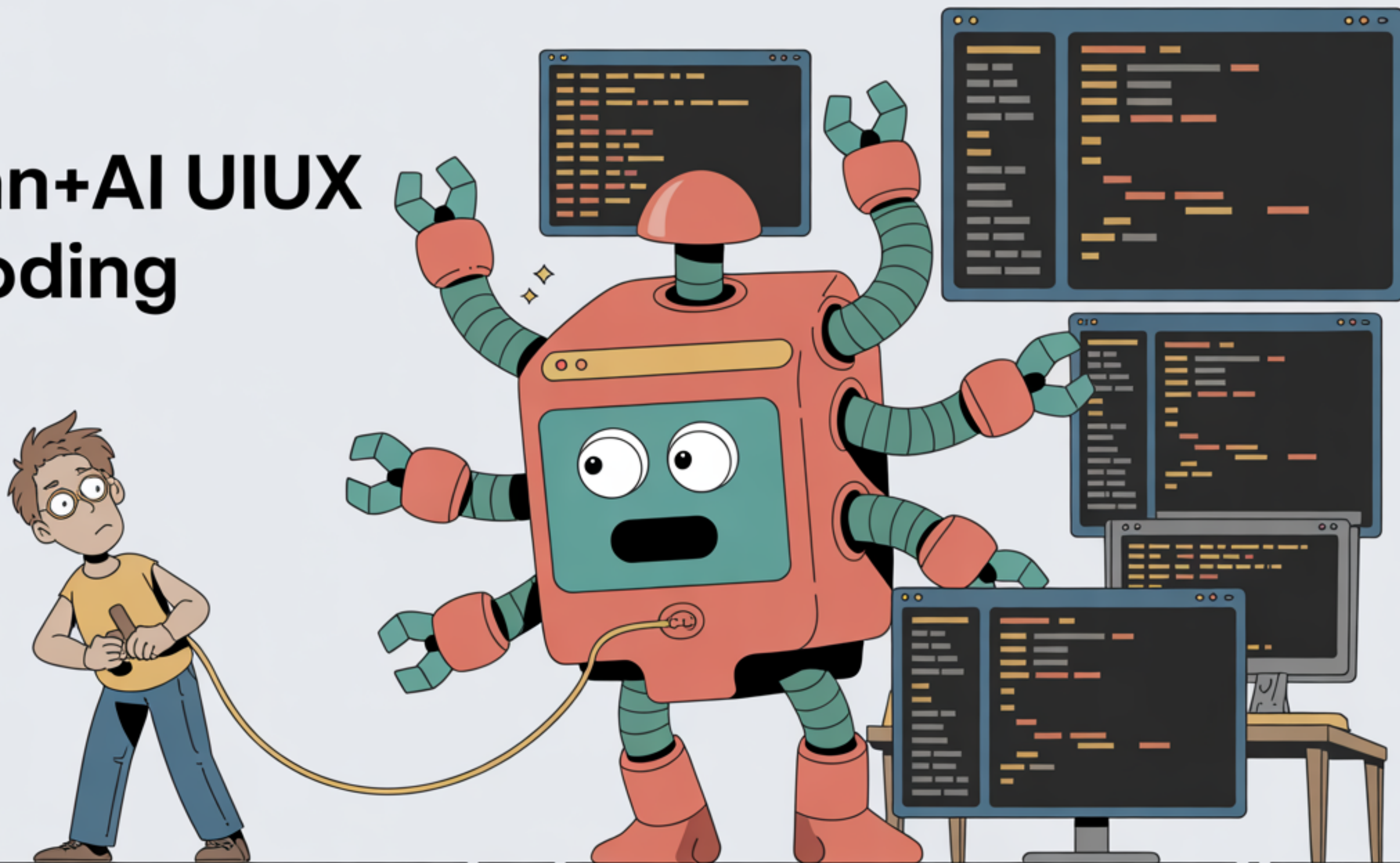
Unreal engine

- Can an LLM "see" all the things the human can?
- Can an LLM "act" in all the ways a human can?
- How can a human supervise and stay in the loop?
- ...

Consider the full workflow of partial autonomy UI/UX



Human+AI UI/UX for Coding



Example: keeping agents on the leash

"AI-assisted coding" workflows (very rapidly evolving...)

- describe the single, next concrete, incremental change
- don't ask for code, ask for approaches
 - pick an approach, draft code
 - review / learn: pull up API docs, ask for explanations, ...
 - wind back, try a different approach
- test
- git commit
- ask for suggestions on what could be implemented next
- repeat

Example: keeping agents on the leash

Here's an example. This prompt is not unreasonable but not particularly thoughtful:

```
Write a Python rate limiter that limits users to 10 requests per minute.
```

I would expect this prompt to give okay results, but also miss some edge cases, good practices and quality standards. This is how you might see someone at nilenso prompt an AI for the same task:

```
Implement a token bucket rate limiter in Python with the following requirements:
```

- 10 requests per minute per user (identified by ``user_id`` string)
- Thread-safe for concurrent access
- Automatic cleanup of expired entries
- Return tuple of (allowed: bool, retry_after_seconds: int)

```
Consider:
```

- Should tokens refill gradually or all at once?
- What happens when the system clock changes?
- How to prevent memory leaks from inactive users?

```
Prefer simple, readable implementation over premature optimization. Use stdlib only (no Redis/external deps).
```



Atharva Raykar
Read more by Atharva [here](#)

AI-assisted coding for teams that can't get away with vibes

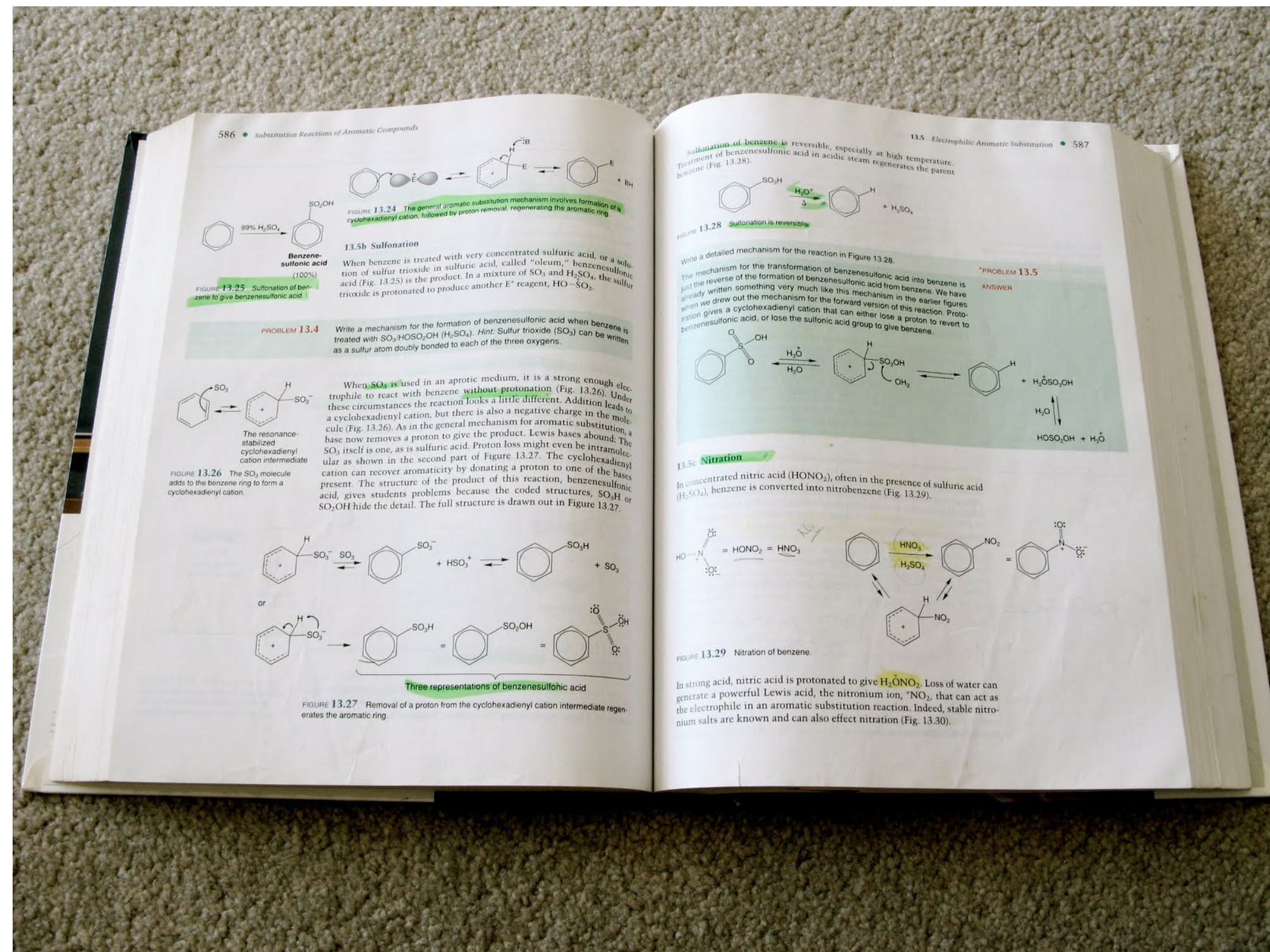
29 May 2025

Status: Living document based on production experience

Last updated: 5-Jun-2025

Example: keeping agents on the leash - AI + Education / LLM101n

1. App for course creation (for teacher)



2. App for course serving (for student)



Example: Tesla Autopilot



autonomy slider



- keep the lane
- keep distance from the car ahead
- take forks on highway
- stop for traffic lights and signs
- take turns at intersections
- ...

2015 - 2025 was the decade of "driving agents"



2013: my first demo drive in a Waymo around Palo Alto (it was perfect).



2015 - 2025 was the decade of "driving agents"

Mind the "**demo-to-product gap**"!

demo is a ``works.any()``

product is a ``works.all()``

It takes a huge amount of hard work across the stack to turn an autonomy demo into an autonomy product, especially when high reliability matters.



"2025 is the year of agents."

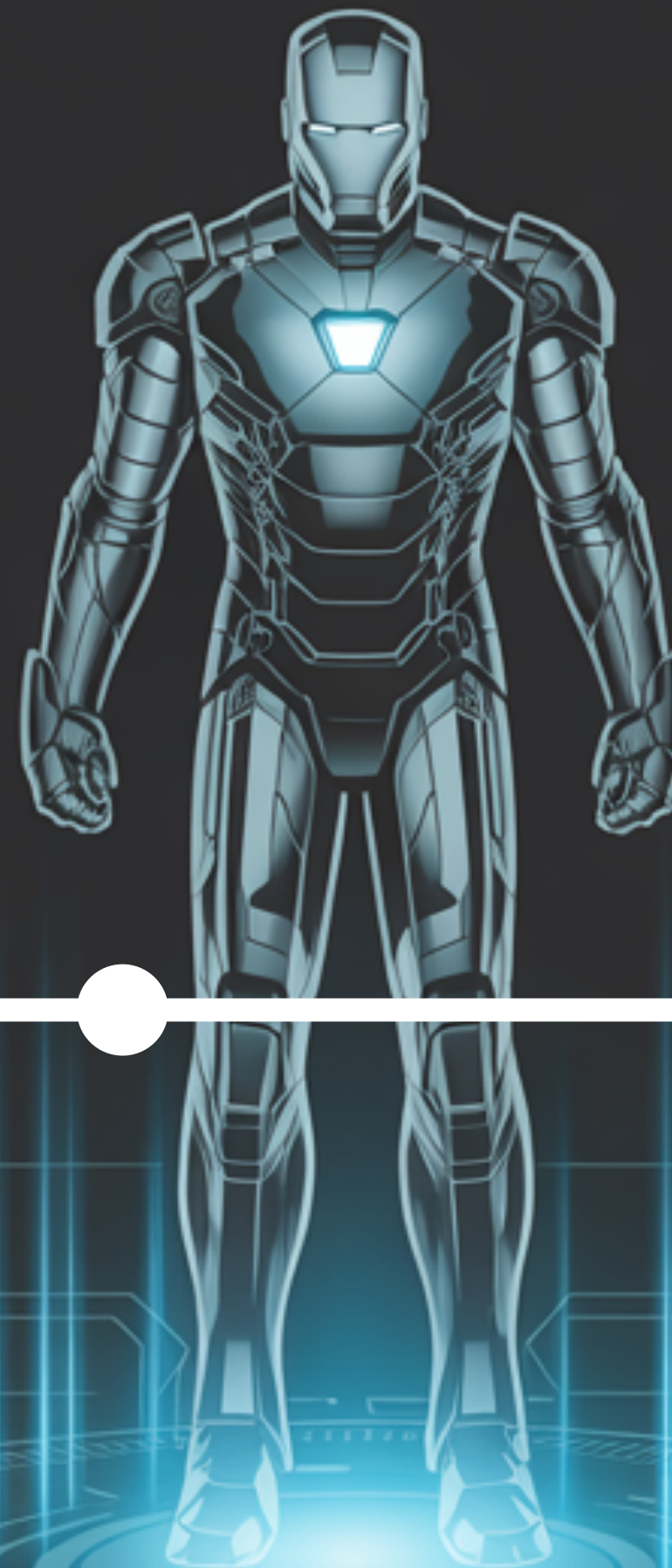
?

"2025-2035 is the decade of agents."

-Andrej

THE IRON MAN SUIT

Augmentation



Agent



Building Autonomous Software

✗ Iron Man robots

✗ Flashy demos of autonomous agents

✗ AGI 2027

✓ Iron Man suits

✓ Partial autonomy products

✓ Custom GUI and UIUX

✓ Fast Generation - Verification loop

✓ Autonomy slider

Make software highly accessible 🧒

(Have you heard of vibe coding by any chance?)



Andrej Karpathy ✓

@karpathy

Promote



There's a new kind of coding I call "vibe coding", where you fully give in to the vibes, embrace exponentials, and forget that the code even exists. It's possible because the LLMs (e.g. Cursor Composer w Sonnet) are getting too good. Also I just talk to Composer with SuperWhisper so I barely even touch the keyboard. I ask for the dumbest things like "decrease the padding on the sidebar by half" because I'm too lazy to find it. I "Accept All" always, I don't read the diffs anymore. When I get error messages I just copy paste them in with no comment, usually that fixes it. The code grows beyond my usual comprehension, I'd have to really read through it for a while. Sometimes the LLMs can't fix a bug so I just work around it or ask for random changes until it goes away. It's not too bad for throwaway weekend projects, but still quite amusing. I'm building a project or webapp, but it's not really coding - I just see stuff, say stuff, run stuff, and copy paste stuff, and it mostly works.

Vibe coding

🌐 24 languages ▾

Article [Talk](#)

Read [Edit](#) [View history](#) [Tools](#) ▾



From Wikipedia, the free encyclopedia





This article **may contain an excessive number of citations**. Please help [remove low-quality or irrelevant citations](#). *(June 2025)* ([Learn how and when to remove this message](#))

Vibe coding (or **vibecoding**) is an approach to producing software by using [artificial intelligence](#) (AI), where a person describes a problem in a few [natural language](#) sentences as a [prompt](#) to a [large language model](#) (LLM) tuned for coding. The LLM generates [software](#) based on the description, shifting the programmer's role from manual coding to guiding, testing, and refining the AI-generated [source code](#).^{[1][2][3]}

Advocates of vibe coding say that it allows even [amateur programmers](#) to produce software without the extensive training and skills required for [software engineering](#).^[4] The term was introduced by [Andrej Karpathy](#) in February 2025^{[5][2][4][1]} and listed in the [Merriam-Webster Dictionary](#) the following month as a "slang & trending" noun.^[6]

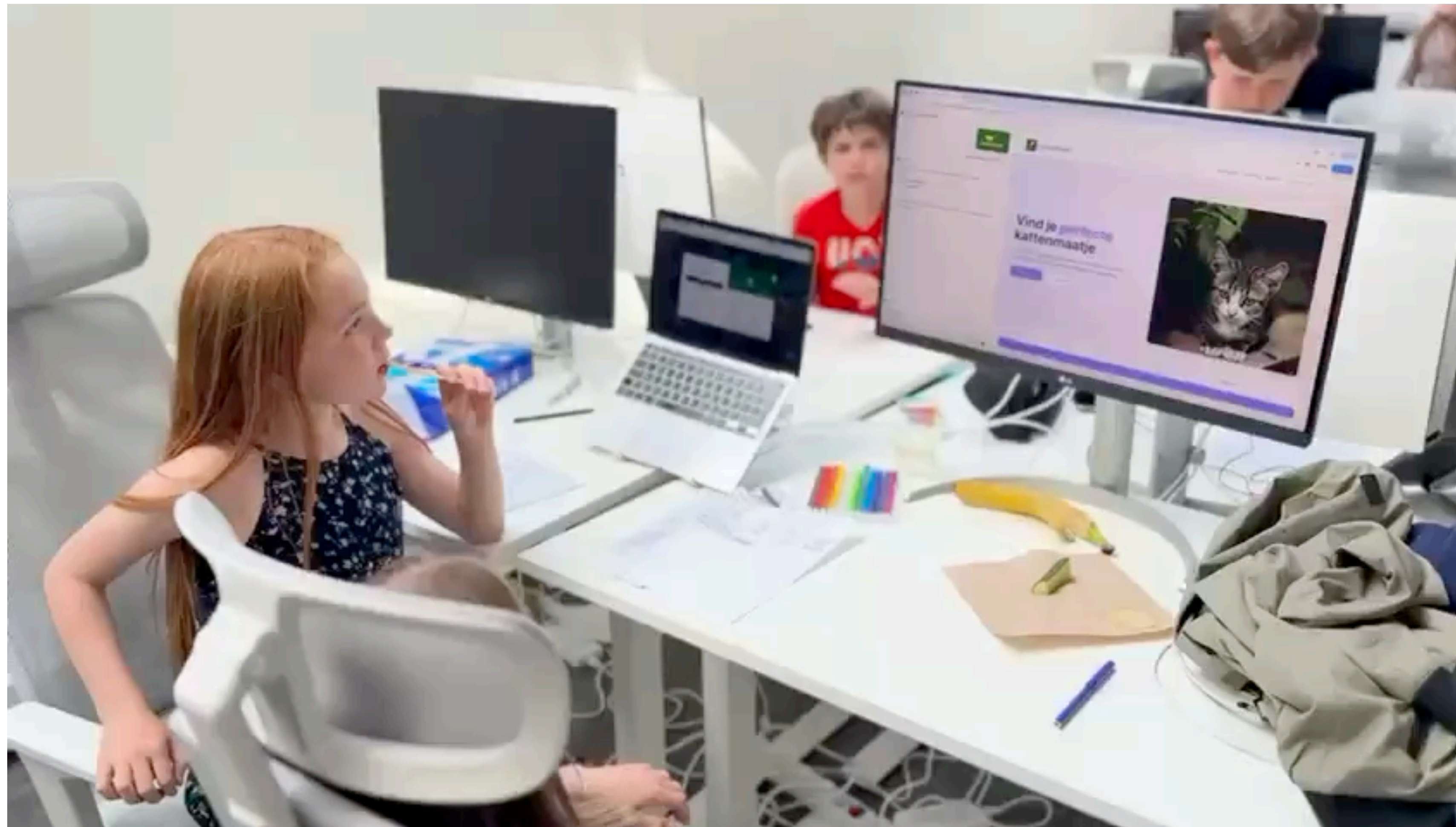
**Thomas Wolf** 
@Thom_Wolf

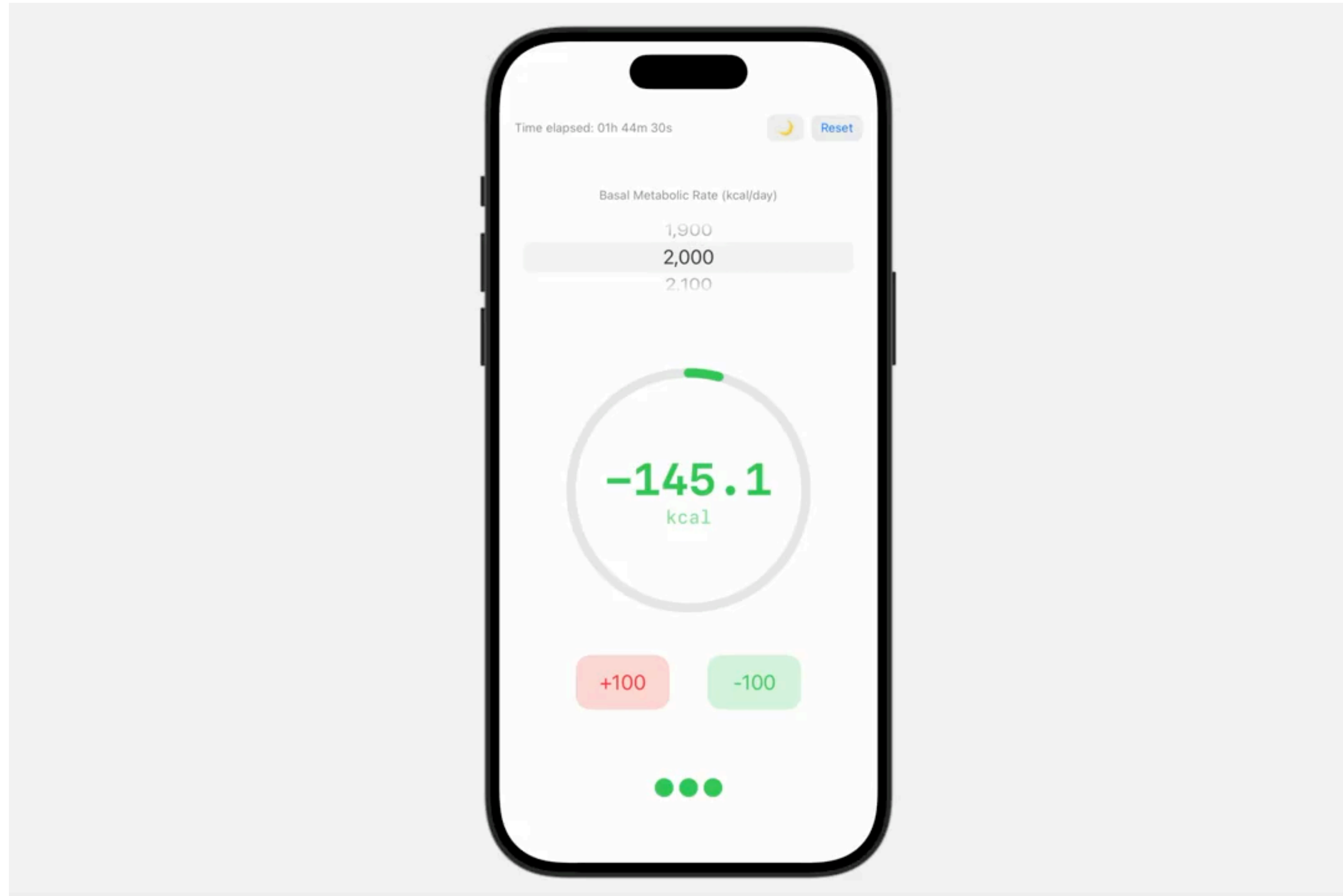
we've seen nothing yet! hosted a 9-13 yo vibe-coding event w.
[@robertkeus](#) this w-e (h/t [@antonosika](#) [@LovableBuild](#))

takeaway? AI is unleashing a generation of wildly creative builders
beyond anything I'd have imagined

and they grow up **knowing** they can build anything!



Vibe Coding iOS app



Vibe coding MenuGen

<https://www.menugen.app/>

Spago

WOLFGANG PUCK

SPAGO BREAKFAST SIGNATURES

Kaya French Toast	26
Pandan-Coconut Jam, Soy Caramel, Egg Jam	
Iberico Pork Dumplings "Agnolotti"	20
Soy, Chili Oil, Black Vinegar, Coriander, Garlic	
Spago's Laksa Noodle	26
Little Neck Clams, Tiger Prawns, Calamari, Onion Sambal, Calamansi, Laksa Leaf	
"Curry & Waffle"	20
Tamarind Fish Curry, Japanese Seabream, Prata Waffles, Curry Leaves	

BREAKFAST CLASSICS

Assorted Fresh Fruits	18
Sarawak Pineapple, Watermelon, Chitose Farm Strawberry, Thai Mango, Passion Fruit Syrup	
French Omelet	28
Sauteed Wild Mushrooms, Gruyère, Thyme, Rocket Salad	
Open-Faced Monte Cristo Sandwich	24
Canadian Style Smoked Bacon, Sunny Side Up Egg, Gruyère, Soy Caramel	
Avocado Toast	26
Salmon Cream Cheese, Watercress, Poached Egg, Hollandaise, Dill, Chives	
Spago Full Breakfast	34
2 Eggs, Canadian Style Smoked Bacon, Spiced Pork Sausage, Pommes Anna, Fontina	
Shakshuka	24
Braised Tripe, 1 Egg, Spiced Tomato Ragu, Grilled Baguette, Mint, Parsley, Chili	

BOWLS AND SALADS

"Yogurt and Berries"	26
Fagé Greek Yogurt, Banana, Blueberry, Raspberry, Almond	
Homemade Muesli	20
Steele Cut Oats, Apple, Apricot, Yogurt, Homemade Granola	
Healthy Breakfast Bowl	26
Egg White Scramble, Sautéed Spinach, French Bean, Wild Mushrooms, Parmesan Crumble (Add Grilled Chicken +9)	
Kaviar Smoked Salmon	28
Dill Crème Fraîche, Cucumbers, Amela Tomatoes, Capers, Rye Sourdough	

COFFEE SELECTION

Espresso (single)	8
Americano	8
Cappuccino	10
Latte	10
Mocha	10

TEA SELECTION






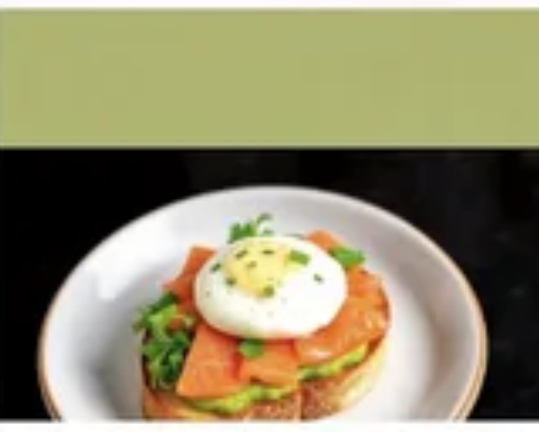



Earl Grey	14
English Breakfast	12
Alphonso (Mango Black Tea)	12
Sencha	14
Chamomile	12

FRESH JUICES & SMOOTHIES

Orange	8
Watermelon	8
Stimulate – Carrot, Pineapple, Celery, Ginger & Lime	9
ABC – Apple, Beetroot & Carrot	9
Strawberry Banana	14

menugen!



 <p>SPAGO BREAKFAST SIGNATURES</p> <p>Kaya French Toast</p> <p>Pandan-Coconut Jam, Soy Caramel, Egg Jam.</p>	 <p>SPAGO BREAKFAST SIGNATURES</p> <p>Curry & Waffle</p> <p>Tamarind Fish Curry, Japanese Seabream, Prata Waffles, Curry Leaves.</p>	 <p>BREAKFAST CLASSICS</p> <p>Open-Faced Monte Cristo Sandwich</p> <p>Canadian Style Smoked Bacon, Sunny Side Up Egg, Gruyère, Soy Caramel.</p>
 <p>SPAGO BREAKFAST SIGNATURES</p> <p>Iberico Pork Dumplings 'Agnolotti'</p> <p>Soy, Chili Oil, Black Vinegar, Coriander, Garlic.</p>	 <p>BREAKFAST CLASSICS</p> <p>Assorted Fresh Fruits</p> <p>Sarawak Pineapple, Watermelon, Chitose Farm Strawberry, Thai Mango, Passion Fruit Syrup.</p>	 <p>BREAKFAST CLASSICS</p> <p>Avocado Toast</p> <p>Salmon Cream Cheese, Watercress, Poached Egg, Harissa, Dill, Chives.</p>
 <p>SPAGO BREAKFAST SIGNATURES</p> <p>Spago's Laksa Noodle</p> <p>Little Neck Clams, Tiger Prawns, Calamari, Onion Sambal, Calamansi, Laksa Leaf.</p>	 <p>BREAKFAST CLASSICS</p> <p>French Omelet</p> <p>Sautéed Wild Mushrooms, Gruyère, Thyme, Rocket Salad.</p>	 <p>BREAKFAST CLASSICS</p> <p>Spago Full Breakfast</p> <p>2 Eggs, Canadian Style Smoked Bacon, Spago Pork Sausage, Pommes Anna, Fontina.</p>

9:50



80

MenuGen

\$4.10 credits



Turn Menus into Magic

Upload any menu and watch as AI transforms each dish into stunning, mouth-watering visuals. ✨🍴



Click to upload or drag and drop

PNG, JPG, GIF up to 10MB

 [Instagram](#)
@AndrejKerski



menugen.app



The code was the easiest part! :O
Most of the work was in the
browser clicking things.

😐 LLM API keys

😐 Flux (image generation) API keys

✅ Running locally (ez)

😐 Vercel deployments

😐 Domain names

😡 Authentication

😐 Payments

karpathy

[Home](#) [Blog](#)

Vibe coding MenuGen


27 Apr, 2025


<https://karpathy.bearblog.dev/vibe-coding-menugen/>

Example: adding Google login


Add Google as a social connection

Before you start

 A Clerk application is required.

 A Google Developer account is required.

Enabling OAuth with [Google](#) allows your users to sign up and sign in to your Clerk application with their Google account.

 Warning

Google OAuth 2.0 **does not** allow apps to use WebViews for authentication. See the dedicated [Google blog post](#) for more information. If your app requires users to sign in via in-app browsers, follow the setup instructions in the [Google Help guide](#).

Configure for your development instance

For *development instances*, Clerk uses preconfigured shared OAuth credentials and redirect URIs—no other configuration is needed.

1. In the Clerk Dashboard, navigate to the [SSO connections](#) page.
2. Select **Add connection** and select **For all users**.
3. In the **Choose provider** dropdown, select **Google**.
4. Select **Add connection**.

Configure for your production instance

For *production instances*, you must provide custom credentials.

To make the setup process easier, it's recommended to keep two browser tabs open: one for the [Clerk Dashboard](#) and one for your [Google Cloud Console](#).

1 Enable Google as a social connection


1. In the Clerk Dashboard, navigate to the [SSO connections](#) page.
2. Select **Add connection** and select **For all users**.
3. In the **Choose provider** dropdown, select **Google**.
4. Ensure that both **Enable for sign-up and sign-in** and **Use custom credentials** are toggled on.
5. Save the **Authorized Redirect URI** somewhere secure. Keep this modal and page open.

2 Create a Google Developer project

1. Navigate to the [Google Cloud Console](#).
2. Select a project or [create a new one](#). You'll be redirected to your project's **Dashboard** page.
3. In the top-left, select the menu icon (≡) and select **APIs & Services**. Then, select **Credentials**.
4. Next to **Credentials**, select **Create Credentials**. Then, select **OAuth client ID**. You might need to [configure your OAuth consent screen](#). Otherwise, you'll be redirected to the **Create OAuth client ID** page.
5. Select the appropriate application type for your project. In most cases, it's **Web application**.
6. In the **Authorized JavaScript origins** setting, select **Add URI** and add your domain (e.g., `https://your-domain.com` and `https://www.your-domain.com` if you have a `www` version). For local development, add `http://localhost:PORT` (replace `PORT` with the port number of your local development server).
7. In the **Authorized Redirect URIs** setting, paste the **Authorized Redirect URI** value you saved from the Clerk Dashboard.
8. Select **Create**. A modal will open with your **Client ID** and **Client Secret**. Save these values somewhere secure.

3 Set the Client ID and Client Secret in the Clerk Dashboard

1. Navigate back to the Clerk Dashboard where the modal should still be open. Paste the **Client ID** and **Client Secret** values that you saved into the respective fields.
2. Select **Add connection**.

 Note

If the modal or page is no longer open, navigate to the [SSO connections](#) page in the Clerk Dashboard. Select the connection. Under **Use custom credentials**, paste the values into their respective fields.

4 Test your connection

The simplest way to test your connection is to visit your Clerk app's [Account Portal](#), which is available for all Clerk apps out-of-the-box.

1. In the Clerk Dashboard, navigate to the [Account Portal](#) page.
2. Next to the **Sign-in URL**, select **Visit**. The URL should resemble:
 - **For development** – `https://your-domain.accounts.dev/sign-in`
 - **For production** – `https://accounts.your-domain.com/sign-in`
3. Sign in with your connection's credentials.

 Warning

Google sign-in **does not** allow users to sign in via in-app browsers.

Build for agents 🤖

There is new category of consumer/manipulator of digital information:

1. Humans (GUIs)
2. Computers (APIs)
3. **NEW:** Agents <- computers... but human-like

robots.txt →

The /llms.txt file

A proposal to standardise on using an `/llms.txt` file to provide information to help LLMs use a website at inference time.

AUTHOR

Jeremy Howard

PUBLISHED

September 3, 2024

FastHTML

> FastHTML is a python library which brings together Starlette, Uvicorn, HTMX, and fastcore's ``FT`` "FastTags" into a library for creating server-rendered hypermedia applications.

Important notes:

- Although parts of its API are inspired by FastAPI, it is **not** compatible with FastAPI syntax and is not targeted at creating API services
- FastHTML is compatible with JS-native web components and any vanilla JS library, but not with React, Vue, or Svelte.

Docs

- [FastHTML quick start](https://answerdotai.github.io/fasthtml/tutorials/quickstart_for_web_devs.html.md) A brief overview of many FastHTML features
- [HTMX reference](https://raw.githubusercontent.com/path/reference.md): Brief description of all HTMX attributes, CSS classes, headers, events, extensions, js lib methods, and config options


Examples

- [Todo list application](https://raw.githubusercontent.com/path/adv_app.py): Detailed walk-thru of a complete CRUD app in FastHTML showing idiomatic use of FastHTML and HTMX patterns.

Optional

- [Starlette full documentation](https://gist.githubusercontent.com/path/starlette-sml.md): A subset of the Starlette documentation useful for FastHTML development.

Docs for people

 Copy page

Vercel Documentation

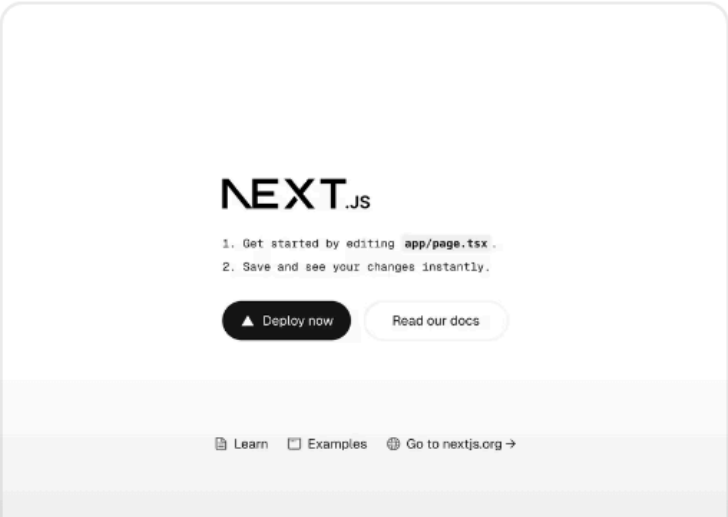
Start with an idea

Vercel builds tools to help you create products faster.

Like **v0**, which is your web development assistant. Paste a screenshot or write a few sentences and v0 will generate a starting point for your next app, including the code for how it looks *and* how it works. v0 then connects to Vercel, takes your code, and creates a URL you can share.

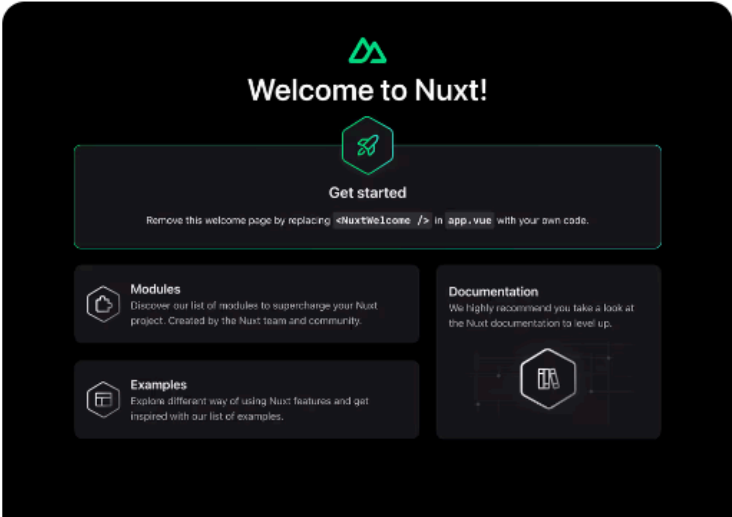
Get started in minutes

Deploy a Template



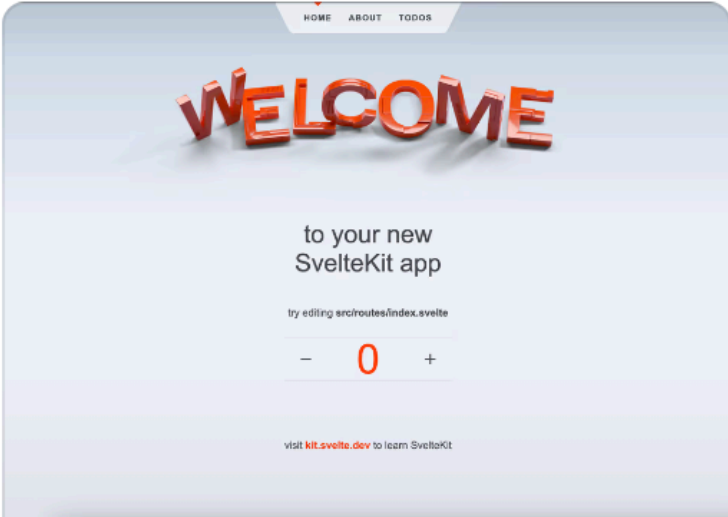
Next.js Boilerplate

Get started with Next.js and React in seconds.



Nuxt.js 3 Boilerplate

A Nuxt.js 3 app, bootstrapped with create-nuxt-app.



SvelteKit Boilerplate

A SvelteKit app including nested routes, layouts, and page endpoints.

Configure for your production instance

For *production instances*, you must provide custom credentials.

To make the setup process easier, it's recommended to keep two browser tabs open: one for the [Clerk Dashboard](#) and one for your [Google Cloud Console](#).

1

Enable Google as a social connection




1. In the Clerk Dashboard, navigate to the [SSO connections](#) page.
2. Select **Add connection** and select **For all users**.
3. In the **Choose provider** dropdown, select **Google**.
4. Ensure that both **Enable for sign-up and sign-in** and **Use custom credentials** are toggled on.
5. Save the **Authorized Redirect URI** somewhere secure. Keep this modal and page open.


2

Create a Google Developer project

1. Navigate to the [Google Cloud Console](#).
2. Select a project or [create a new one](#). You'll be redirected to your project's **Dashboard** page.
3. In the top-left, select the menu icon (≡) and select **APIs & Services**. Then, select **Credentials**.
4. Next to **Credentials**, select **Create Credentials**. Then, select **OAuth client ID**. You might need to [configure your OAuth consent screen](#). Otherwise, you'll be redirected to the **Create OAuth client ID** page.
5. Select the appropriate application type for your project. In most cases, it's **Web application**.
6. In the **Authorized JavaScript origins** setting, select **Add URI** and add your domain (e.g., `https://your-domain.com` and `https://www.your-domain.com` if you have a `www` version). For local development, add `http://localhost:PORT` (replace `PORT` with the port number of your local development server).
7. In the **Authorized Redirect URIs** setting, paste the **Authorized Redirect URI** value you saved from the Clerk Dashboard.
8. Select **Create**. A modal will open with your **Client ID** and **Client Secret**. Save these values somewhere secure.

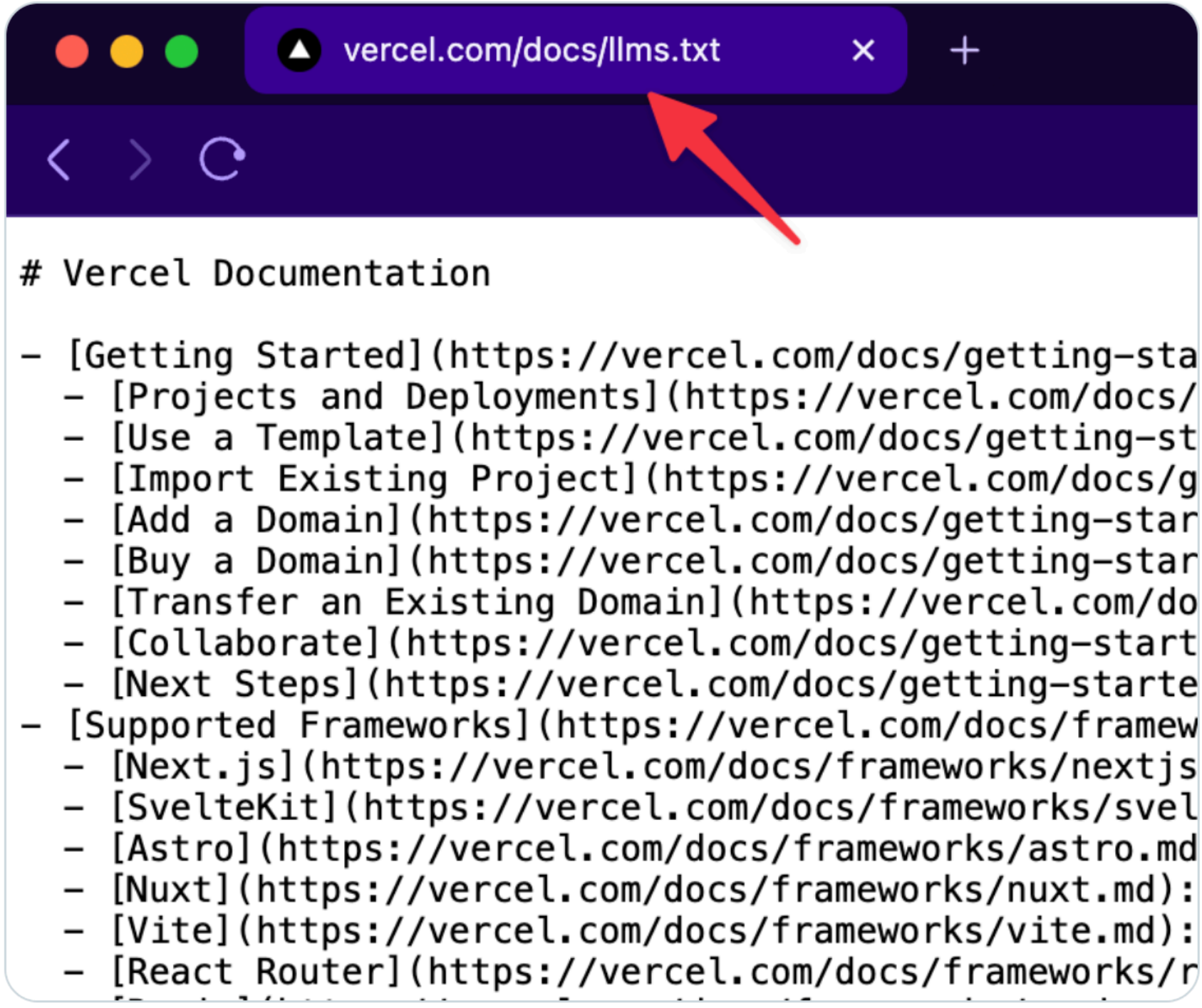
Docs for people LLMs

 **Lee Robinson**  
@leerob

vercel.com/docs/llms.txt is now live 

We also have the full version if you want to read a 400,000 word novel.

This also means you can drop .md on the end of any docs link.



```
# Vercel Documentation
- [Getting Started](https://vercel.com/docs/getting-started)
- [Projects and Deployments](https://vercel.com/docs/projects-and-deployments)
- [Use a Template](https://vercel.com/docs/getting-started#use-a-template)
- [Import Existing Project](https://vercel.com/docs/getting-started#import-existing-project)
- [Add a Domain](https://vercel.com/docs/getting-started#add-a-domain)
- [Buy a Domain](https://vercel.com/docs/getting-started#buy-a-domain)
- [Transfer an Existing Domain](https://vercel.com/docs/getting-started#transfer-an-existing-domain)
- [Collaborate](https://vercel.com/docs/getting-started#collaborate)
- [Next Steps](https://vercel.com/docs/getting-started#next-steps)
- [Supported Frameworks](https://vercel.com/docs/frameworks)
- [Next.js](https://vercel.com/docs/frameworks/nextjs)
- [SvelteKit](https://vercel.com/docs/frameworks/sveltekit)
- [Astro](https://vercel.com/docs/frameworks/astro.md)
- [Nuxt](https://vercel.com/docs/frameworks/nuxt.md)
- [Vite](https://vercel.com/docs/frameworks/vite.md)
- [React Router](https://vercel.com/docs/frameworks/react-router)
```

Home / Get started

Build on Stripe with LLMs

 Copy page 

Use LLMs in your Stripe integration workflow.

You can use large language models (LLMs) to assist in the building of Stripe integrations. We provide a set of tools and best practices if you use LLMs during development.

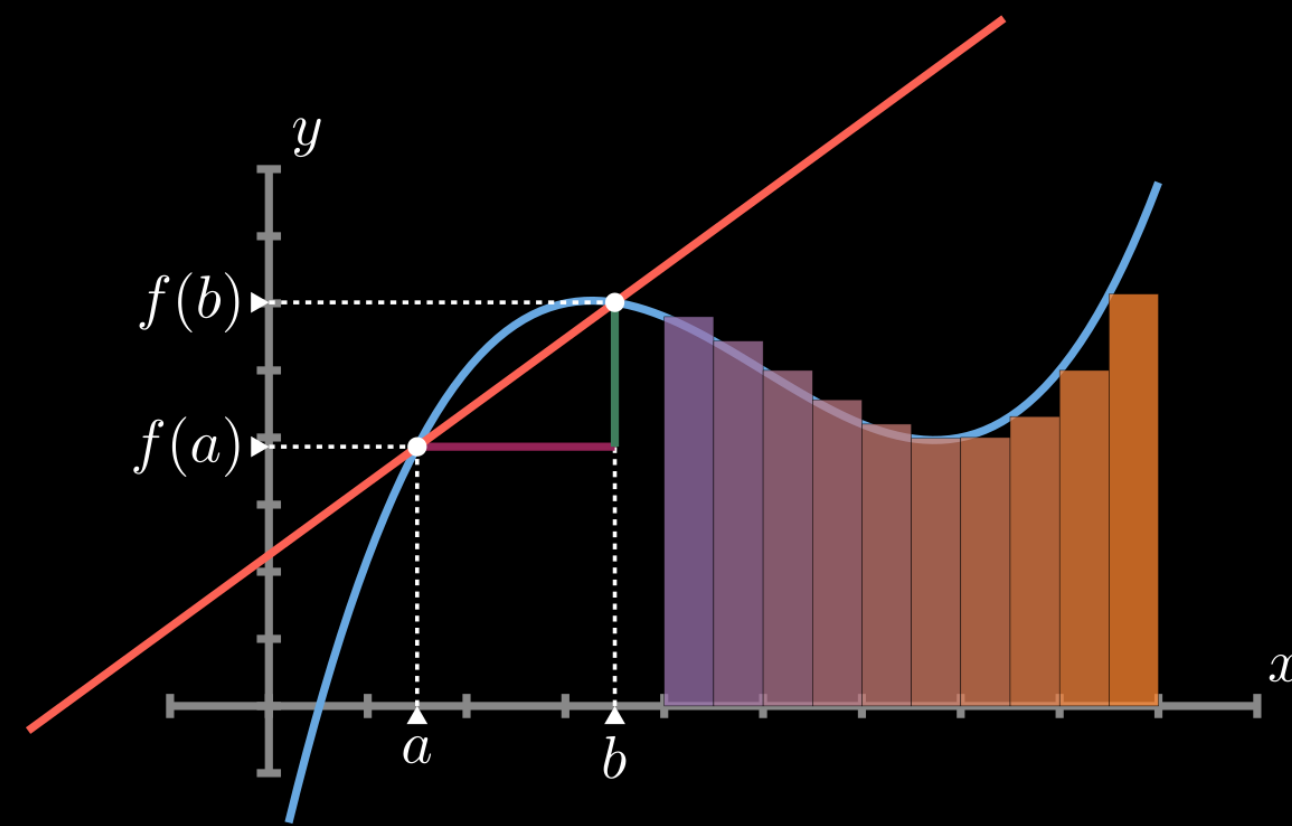
Plain text docs

You can access all of our documentation as plain text markdown files by adding `.md` to the end of any url. For example, you can find the plain text version of this page itself at <https://docs.stripe.com/building-with-llms.md>.

This helps AI tools and agents consume our content and allows you to copy and paste the entire contents of a doc into an LLM. This format is preferable to scraping or copying from our HTML and JavaScript-rendered pages because:

- Plain text contains fewer formatting tokens.
- Content that isn't rendered in the default view (for example, it's hidden in a tab) of a given page is rendered in the plain text version.
- LLMs can parse and understand markdown hierarchy.

We also host an [/llms.txt](#) file which instructs AI tools and agents how to retrieve the plain text versions of our pages. The `/llms.txt` file is an [emerging standard](#) for making websites and content more accessible to LLMs.



Manim

Mathematical Animation Engine

```
class LorenzAttractor(InteractiveScene):
    def construct(self):
        # Set up axes
        axes = ThreeDAxes(■)
        axes.set_width(FRAME_WIDTH)
        axes.center()

        self.frame.reorient(43, 76, 1, IN, 10)
        self.frame.add_updater(lambda m, dt: m.increment_theta(dt * 3 + DEGREES))
        self.add(axes)

        # Add the equations
        equations = Tex(■)
        equations.fix_in_frame()
        equations.to_corner(UL)
        equations.set_backstroke()
        self.play(Write(equations))

        # Compute a set of solutions
        epsilon = 1e-5
        evolution_time = 30
        n_points = 10
        states = [
            [10, 10, 10 + n * epsilon]
            for n in range(n_points)
        ]
        colors = color_gradient([BLUE_E, BLUE_A], len(states))

        curves = VGroup()
        for state, color in zip(states, colors):
            points = ode_solution_points(lorenz_system, state, evolution_time)
            curve = VMobject().set_points_smoothly(axes.c2p(*points.T))
            curve.set_stroke(color, 1, opacity=0.25)
            curves.add(curve)

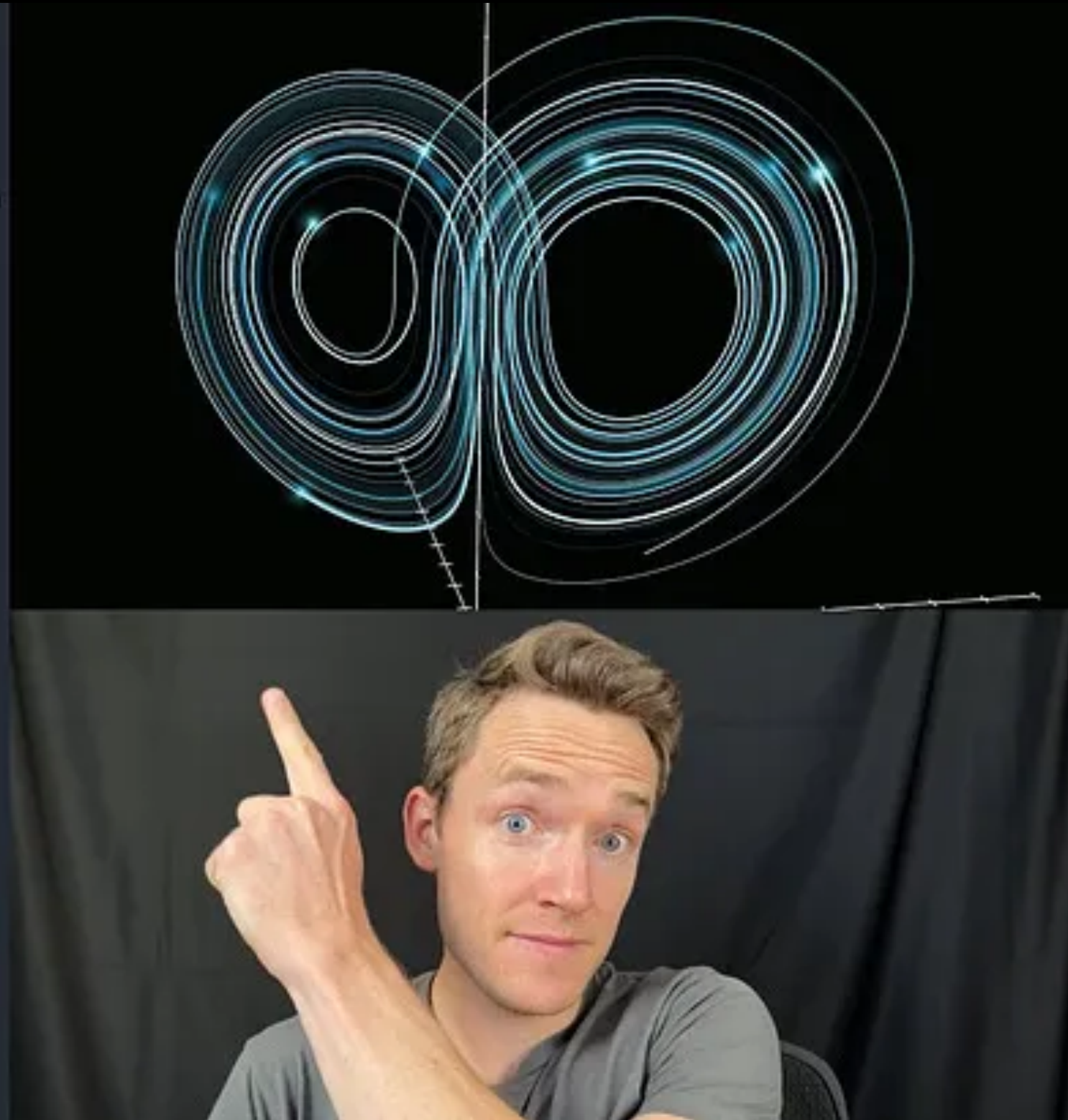
        curves.set_stroke(width=2, opacity=1)

        # Display dots moving along those trajectories
        dots = Group(GlowDot(color=color, radius=0.25) for color in colors)

        globals().update(locals()) # Cursed
        def update_dots(dots):
            for dot, curve in zip(dots, curves):
                dot.move_to(curve.get_end())

        dots.add_updater(update_dots)

        tail = VGroup(
            TracingTail(dot, time_traced=3).match_color(dot)
```



Actions for people LLMs

"click" -> cURL

MCP



Lee Robinson @leerob

We're starting to add cURL commands to Vercel's documentation wherever we previously said "click."

In the future, maybe computer using agents could log in and perform actions for you, but this feels like a nice incremental step for the LLMs.

Creating a project

Dashboard cURL

To create an Authorization Bearer token, see the [access token](#) section of the API documentation.

cURL

```
1 curl --request POST \  
2   --url https://api.vercel.com/v1/projects \  
3   --header "Authorization: Bearer $VERCEL_TOKEN" \  
4   --header "Content-Type: application/json" \  
5   --data '{  
6     "environmentVariables": [  
7       {  
8         "key": "<env-key>",  
9         "target": "production",  
10        "gitBranch": "<git-branch>",  
11        "type": "system",  
12        "value": "<env-value>"  
13      }  
14    ],  
15    "framework": "<framework>",  
16    "gitRepository": {  
17      "repo": "<repo-url>",  
18      "type": "github"  
19    },  
20    "installCommand": "<install-command>",  
21    "name": "<project-name>",  
22    "rootDirectory": "<root-directory>"  
23  }'
```

ALT

Stripe Model Context Protocol (MCP) Server

You can use the Stripe Model Context Protocol (MCP) server if you use code editors that use AI, such as Cursor or Windsurf, or general purpose tools such as Claude Desktop. The MCP server provides AI agents a set of tools you can use to call the Stripe API and search our knowledge base (documentation, support articles, and so on).

Local server

If you prefer or require a local setup, you can run the [local Stripe MCP server](#).

[Cursor](#)

VS Code

Windsurf

Claude

CLI


[Click here](#) to open Cursor and automatically add the Stripe MCP.


Alternatively, add the following to your `~/.cursor/mcp.json` file.

```
1 {  
2   "mcpServers": {  
3     "stripe": {  
4       "command": "npx",  
5       "args": ["-y", "@stripe/mcp", "--tools=all"],  
6       "env": {  
7         "STRIPE_SECRET_KEY": "sk_test_BQokikJOvBiI2HlWgH4olfQ2"  
8       }  
9     }  
10  }  
11 }
```


The code editor agent automatically detects all the available tools and calls the relevant tool when you post a related question in the chat. See the [Cursor documentation](#) for more details.


Context builders, e.g.: Gitingest

 <https://github.com/karpathy/nanogpt>

 nanoGPT

Public

 Unpin

 Unwatch

404

master


6 Branches

0 Tags

Go to file


Add file


Code


 karpathy

Merge pull request #578 from devin-open-source/devin/1733728337-... 93a43d9 · 6 months ago 209 Commits

assets	adjust teaser figure with a more tuned result	2 years ago
config	Fix for gradient_accumulation_steps training slow	2 years ago
data	Merge pull request #420 from vinjn/fix-371-enc-is-not-de...	last year
.gitattributes	keep only what's needed	2 years ago
.gitignore	feature: .gitignore - added venv folders	last year
LICENSE	Add MIT LICENSE file	3 years ago
README.md	Merge branch 'master' into test1	last year
bench.py	Fix AssertionError on macOS - need to check CUDA avail...	2 years ago
configurator.py	shuttling the poor mans configurator aside into its own fil...	3 years ago
model.py	Merge pull request #274 from apivovarov/gelu	2 years ago
sample.py	Fix AssertionError on macOS - need to check CUDA avail...	2 years ago
scaling_laws.ipynb	fix typo (params -> tokens)	2 years ago
train.py	fix: ensure non-zero learning rate during warmup at iterati...	6 months ago
transformer_sizing.ipynb	oops forgot to subtract embedding params, which don't e...	2 years ago

 Gitingest

 Extension

 GitHub

9.1k

karpathy/nanogpt

Ingest

Exclude

*.md, src/

Include files under: 50kb

Summary

Repository: karpathy/nanogpt
Files analyzed: 22

Estimated tokens: 27.0k

Download

Copy all

Directory Structure

Directory structure:
└ karpathy-nanogpt/
 ├── README.md
 ├── bench.py
 ├── configurator.py
 ├── LICENSE
 ├── model.py
 ├── sample.py
 ├── scaling_laws.ipynb
 └ train.py

Copy

Files Content

=====

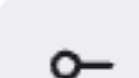
FILE: README.md




=====


nanoGPT

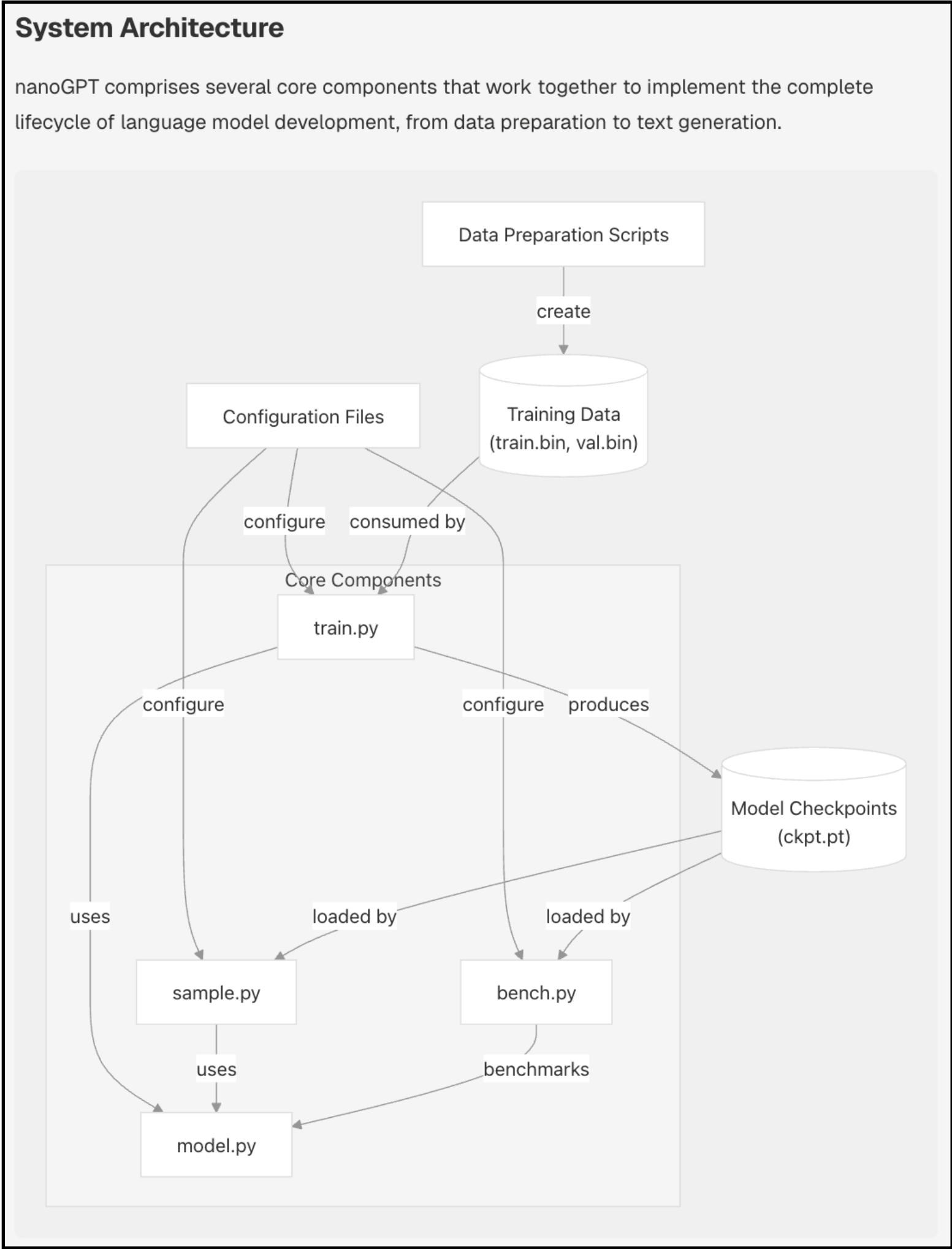
Copy

Context builders, e.g.: Devin DeepWiki

 <https://github.com/karpathy/nanogpt>

 nanoGPT	Public	 Unpin	 Unwatch	404
<div>master 6 Branches 0 Tags</div> <div>Go to file Add file Code</div>				
<div>karpathy Merge pull request #578 from devin-open-source/devin/1733728337-... 93a43d9 · 6 months ago 209 Commits</div>				
assets	adjust teaser figure with a more tuned result	2 years ago		
config	Fix for gradient_accumulation_steps training slow	2 years ago		
data	Merge pull request #420 from vinjn/fix-371-enc-is-not-de...	last year		
.gitattributes	keep only what's needed	2 years ago		
.gitignore	feature: .gitignore - added venv folders	last year		
LICENSE	Add MIT LICENSE file	3 years ago		
README.md	Merge branch 'master' into test1	last year		
bench.py	Fix AssertionError on macOS - need to check CUDA avail...	2 years ago		
configurator.py	shuttling the poor mans configurator aside into its own fil...	3 years ago		
model.py	Merge pull request #274 from apivovarov/gelu	2 years ago		
sample.py	Fix AssertionError on macOS - need to check CUDA avail...	2 years ago		
scaling_laws.ipynb	fix typo (params -> tokens)	2 years ago		
train.py	fix: ensure non-zero learning rate during warmup at iterati...	6 months ago		
transformer_sizing.ipynb	oops forgot to subtract embedding params, which don't e...	2 years ago		

 <https://deepwiki.com/karpathy/nanoGPT/1-overview>



Introducing Operator

A research preview of an agent that can use its own browser to perform tasks for you. Available to Pro users in the U.S.

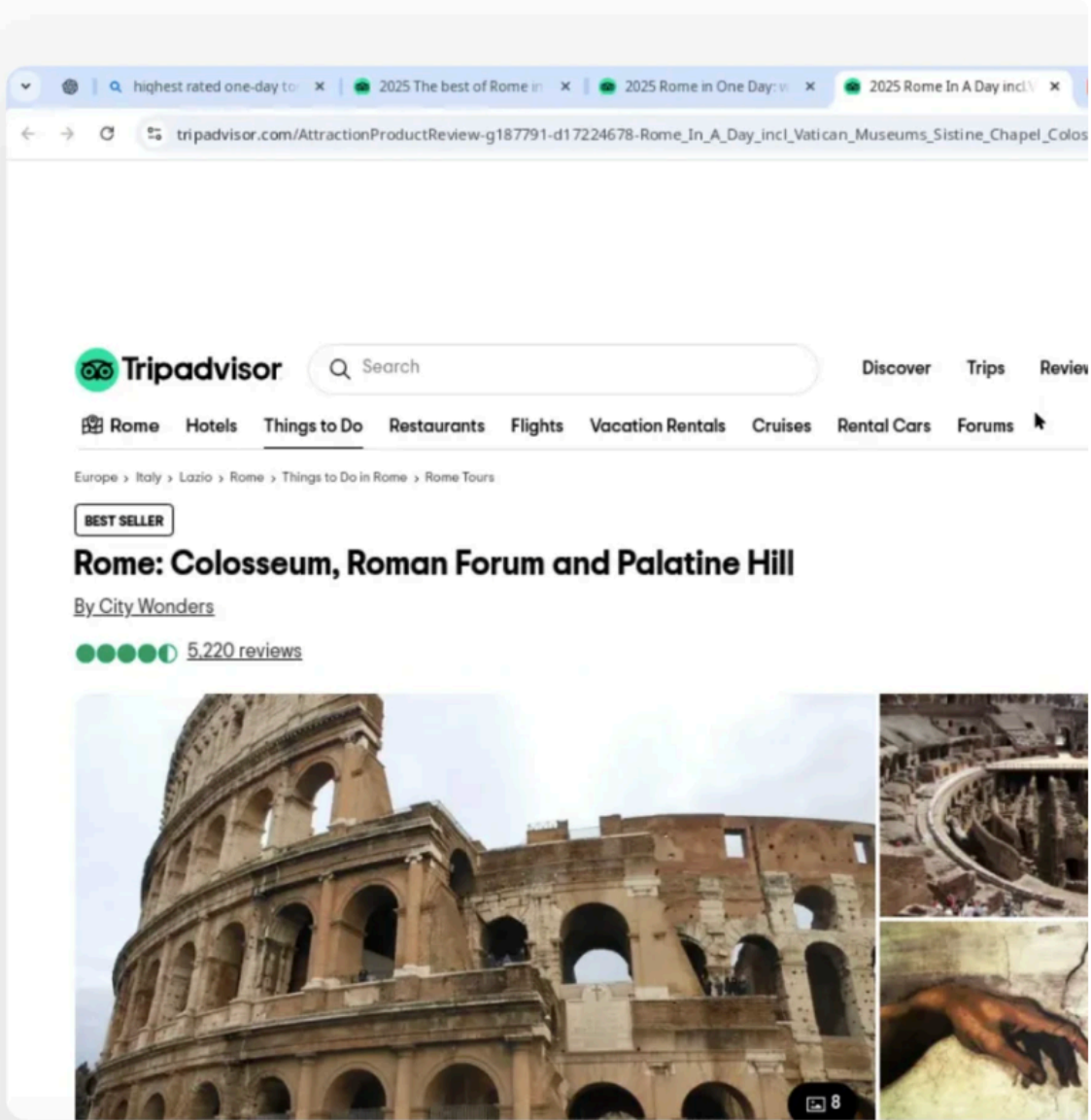
Go to Operator ↗

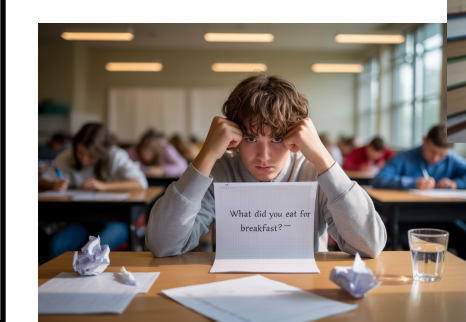
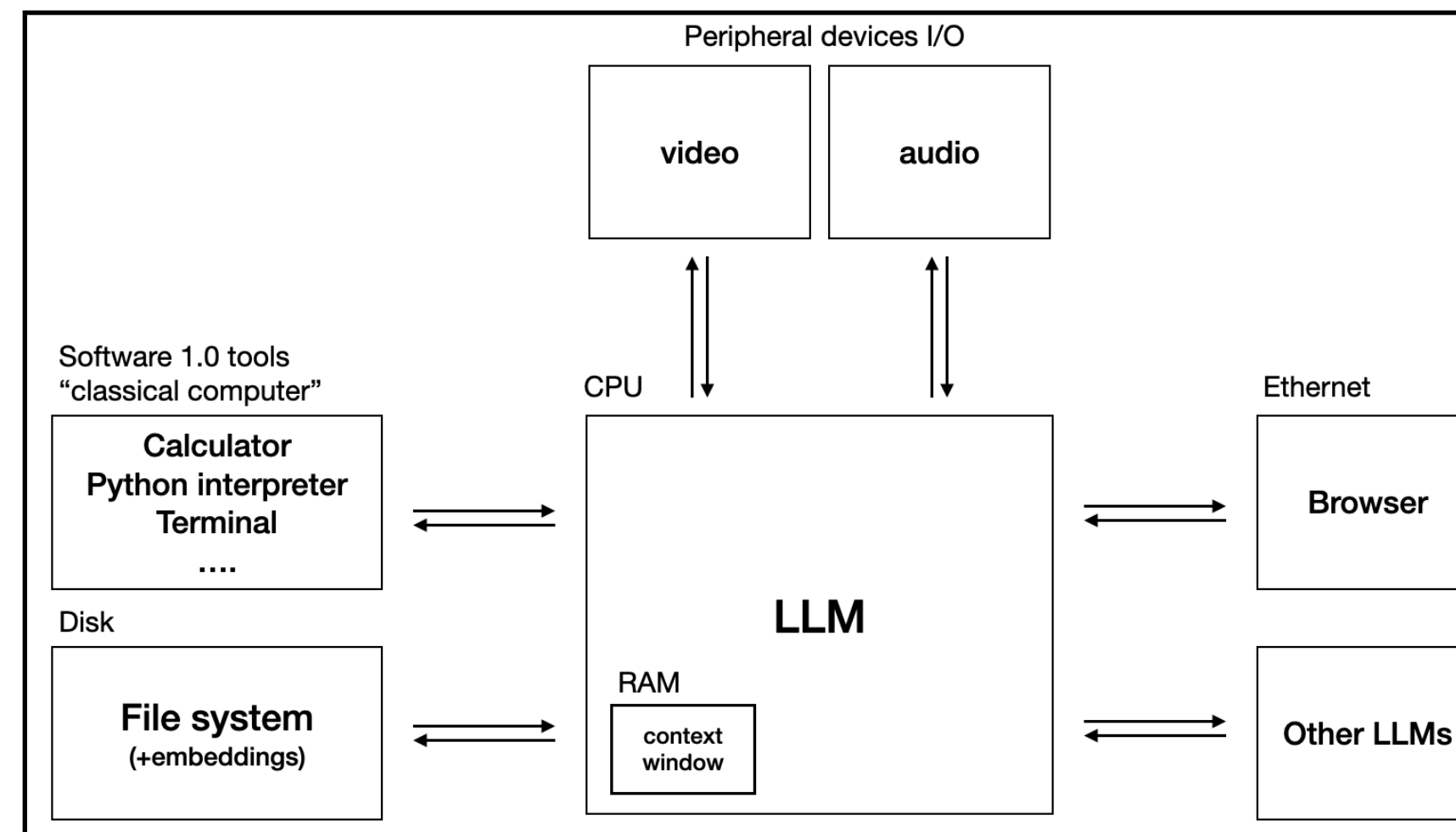
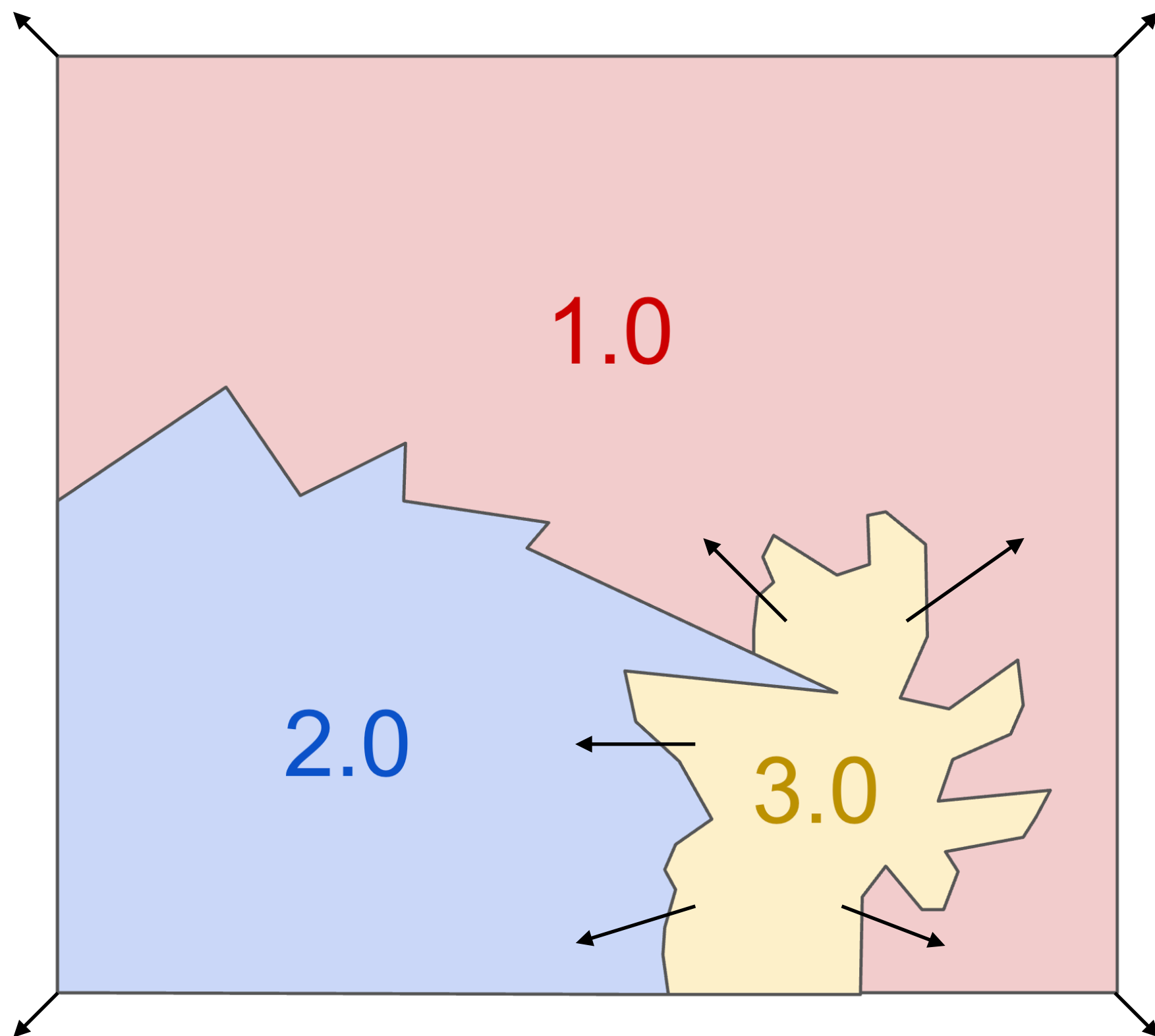
Find and book me the highest rated one-day tour of Rome on TripAdvisor.

I'll search for the highest-rated tour of historic Rome on TripAdvisor. Once I find a suitable option, I'll provide you with the details. Let's begin.

Worked for 2 minutes ^

- Navigating to TripAdvisor website
- Selecting "Things to Do" category
- Searching for historic Rome tours
- Closing pop-up, continuing tour search
- Searching for Historic Rome tours
- Exploring all historic Rome tour options
- Closing Colosseum tab, resuming tour search
- Closing tour pop-up, tab afterward
- Exploring options for top-rated tours
- Sorting results by tour ratings
- Exploring filters for top-rated tours
- Scrolling for sorting options, finding tours





Partial autonomy LLM apps:

- Package context
- Orchestrate LLM calls
- Custom GUI
- Autonomy slider



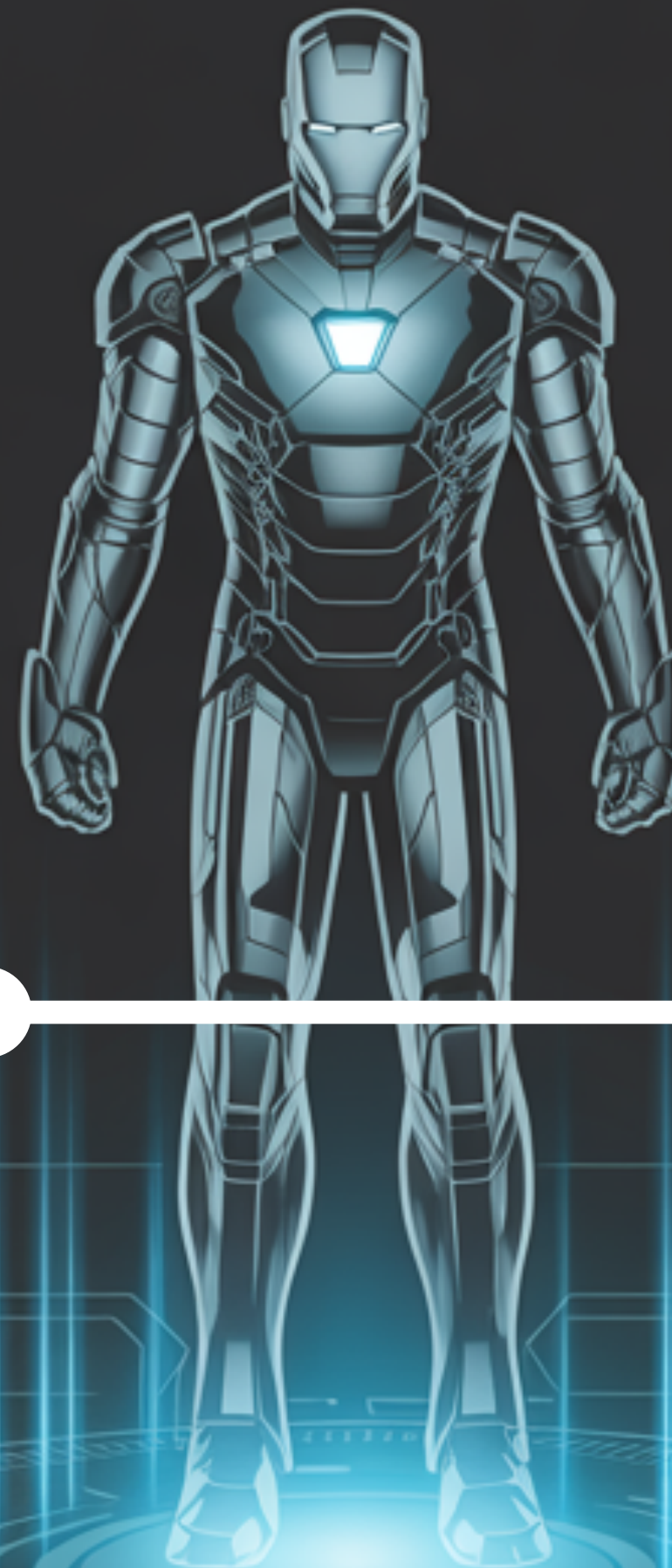
speed up the full generation-verification flow



Build for
agents 🤖

THE IRON MAN SUIT

Augmentation



Agent



Thank you!