Presented by :
Sans Lamsal
Engineer @Slalom

do you really need that AI subscription...?

Local Models.

motivations

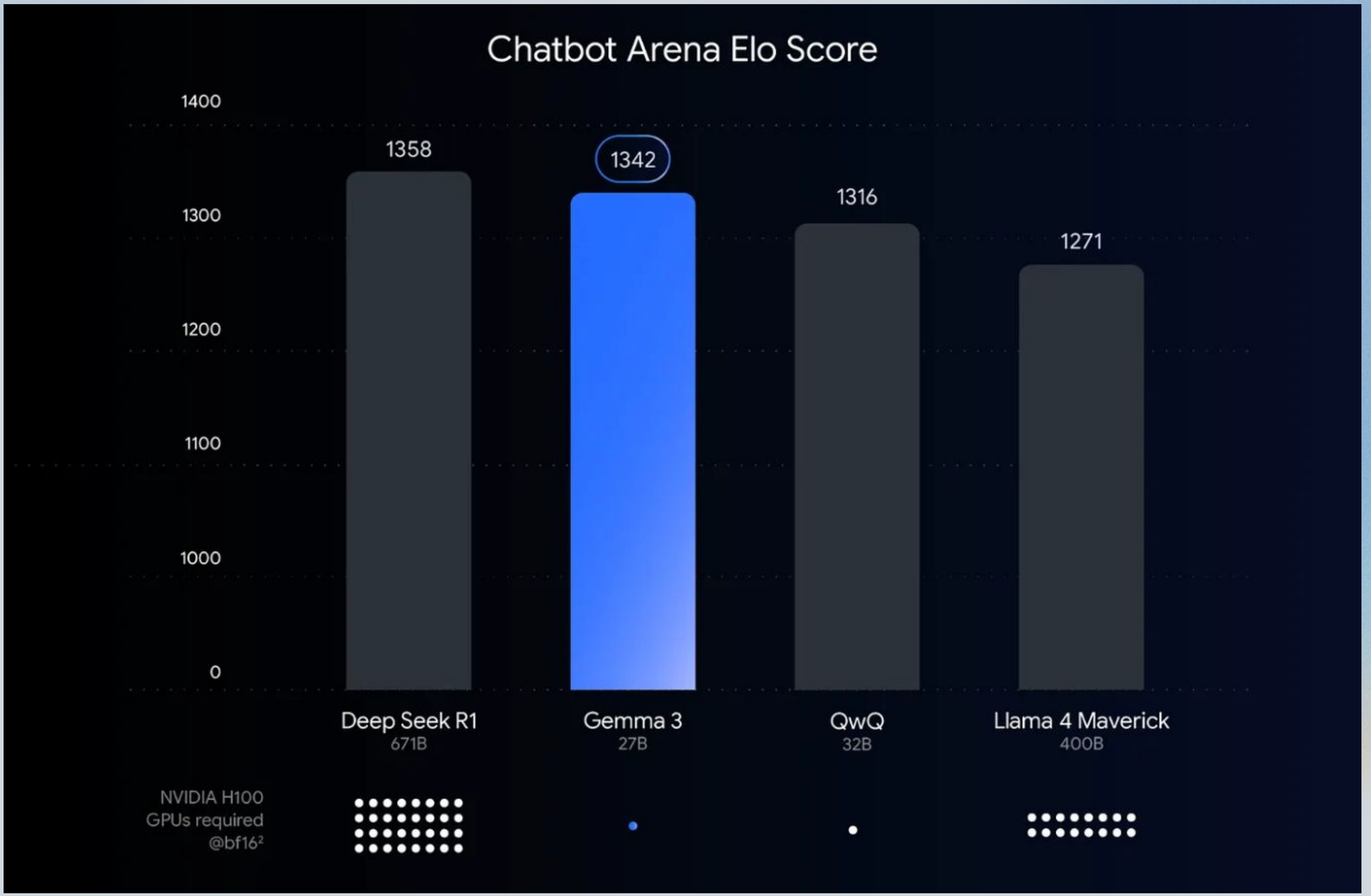
:

rise of local models that’s actually good

~end of tokenmaxxing/cost?

privacy

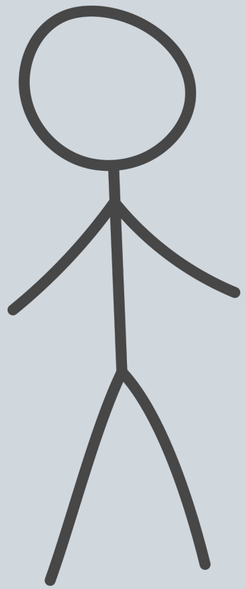
can we do more with less mindset


Chatbot Arena Elo Score
1400
1300
1200
1100
1000
0
1358
1342
1316
1271
Deep Seek R1
671B
Gemma 3
27B
QwQ
32B
Llama 4 Maverick
400B
NVIDIA H100
GPUs required
@bf16[2]





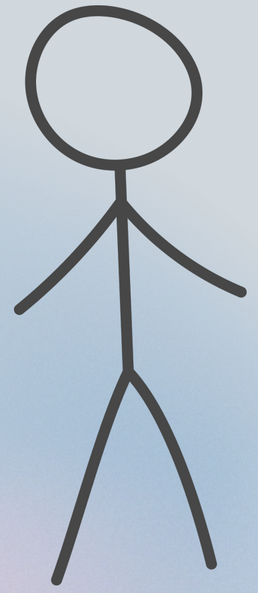




local

models


local
models



Today’s Agenda

01 CanIrunIt? + Stats

02 Download model

03 Ollama, LM Studio,Llamacpp

04 harnesses

05 BYOM

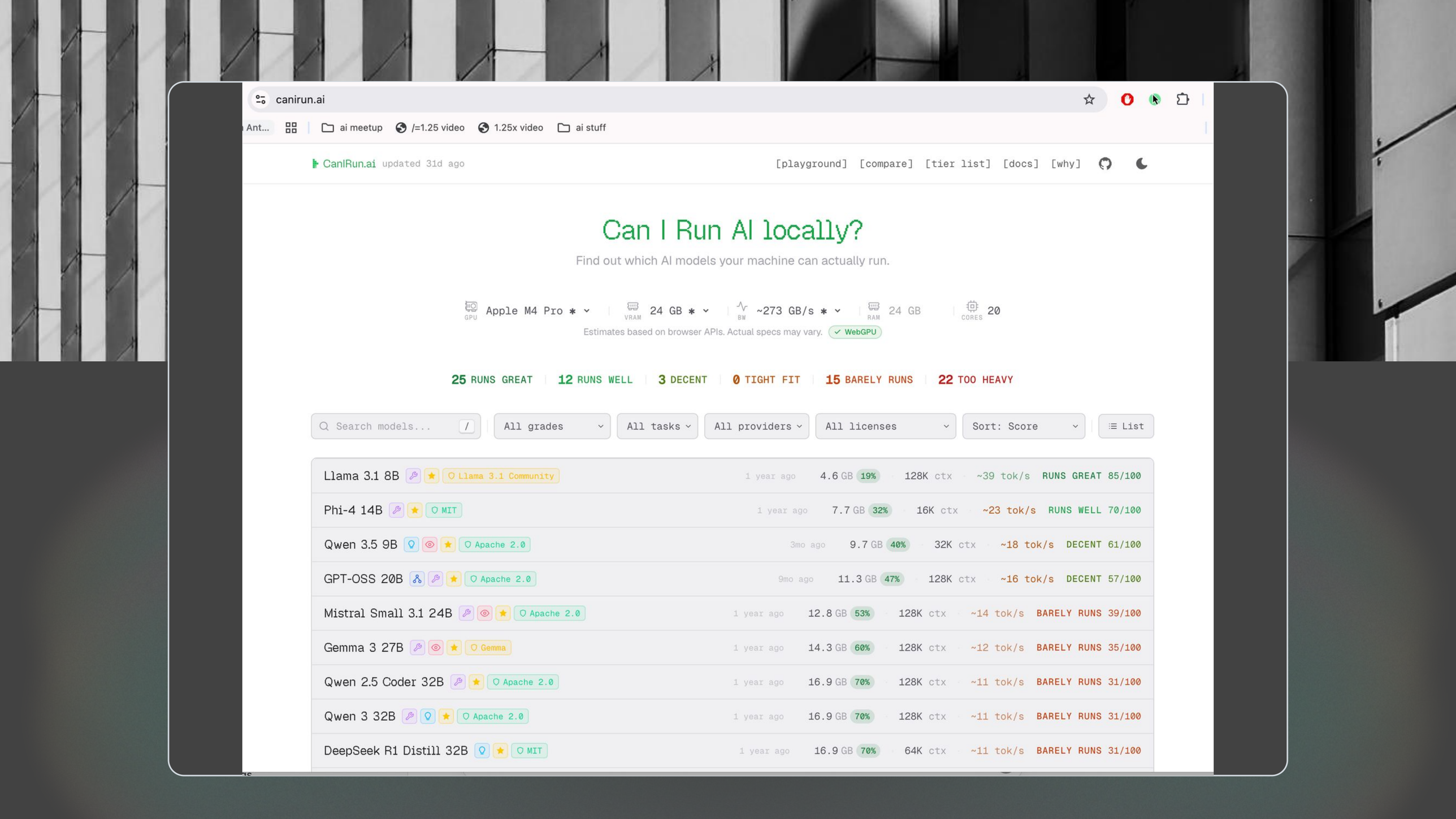
06 searching the web

07 production level saas

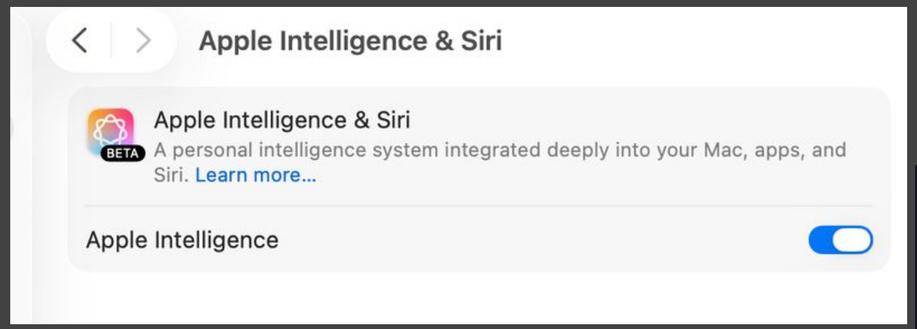
08 hyperframes and local models

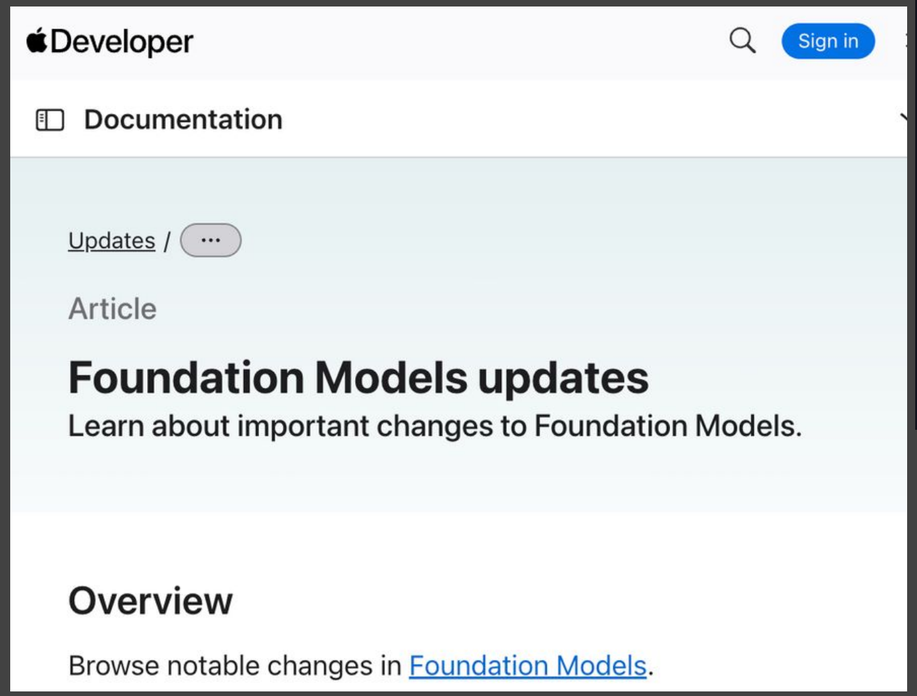
09 did i cancel my subscription

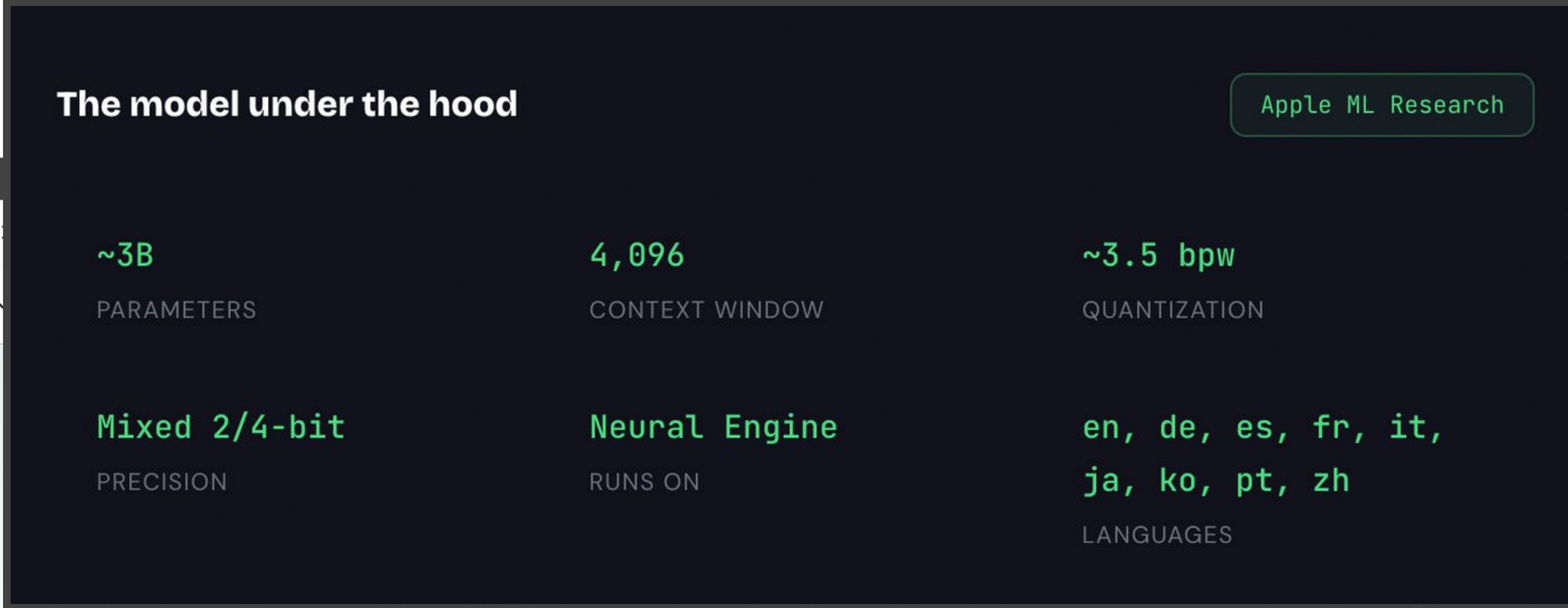
10 insights


canirun.ai
Ant...
ai meetup
/=1.25 video
1.25x video
ai stuff
CanIRun.ai updated 31d ago
[playground] [compare] [tier list] [docs] [why]
Can I Run AI locally?
Find out which AI models your machine can actually run.
GPU Apple M4 Pro
VRAM 24 GB
BW ~273 GB/s
RAM 24 GB
CORES 20
Estimates based on browser APIs. Actual specs may vary.
WebGPU
25 RUNS GREAT
12 RUNS WELL
3 DECENT
0 TIGHT FIT
15 BARELY RUNS
22 TOO HEAVY
Search models...
All grades
All tasks
All providers
All licenses
Sort: Score
List
Llama 3.1 8B Llama 3.1 Community 1 year ago 4.6 GB 19% 128K ctx ~39 tok/s RUNS GREAT 85/100
Phi-4 14B MIT 1 year ago 7.7 GB 32% 16K ctx ~23 tok/s RUNS WELL 70/100
Qwen 3.5 9B Apache 2.0 3mo ago 9.7 GB 40% 32K ctx ~18 tok/s DECENT 61/100
GPT-OSS 20B Apache 2.0 9mo ago 11.3 GB 47% 128K ctx ~16 tok/s DECENT 57/100
Mistral Small 3.1 24B Apache 2.0 1 year ago 12.8 GB 53% 128K ctx ~14 tok/s BARELY RUNS 39/100
Gemma 3 27B Gemma 1 year ago 14.3 GB 60% 128K ctx ~12 tok/s BARELY RUNS 35/100
Qwen 2.5 Coder 32B Apache 2.0 1 year ago 16.9 GB 70% 128K ctx ~11 tok/s BARELY RUNS 31/100
Qwen 3 32B Apache 2.0 1 year ago 16.9 GB 70% 128K ctx ~11 tok/s BARELY RUNS 31/100
DeepSeek R1 Distill 32B MIT 1 year ago 16.9 GB 70% 64K ctx ~11 tok/s BARELY RUNS 31/100

Apple foundation model
(that powers siri)







v1.6.0 · Zero Config · MIT

apfel

The free AI already on your Mac.

macOS Tahoe ships with a 3B parameter LLM. apfel gives you CLI access with one brew install. No model downloads, no API keys, no configuration needed, just works.

- No Token Costs
- On-Device AI
- Scriptable

```
$ brew install apfel
```

Copy

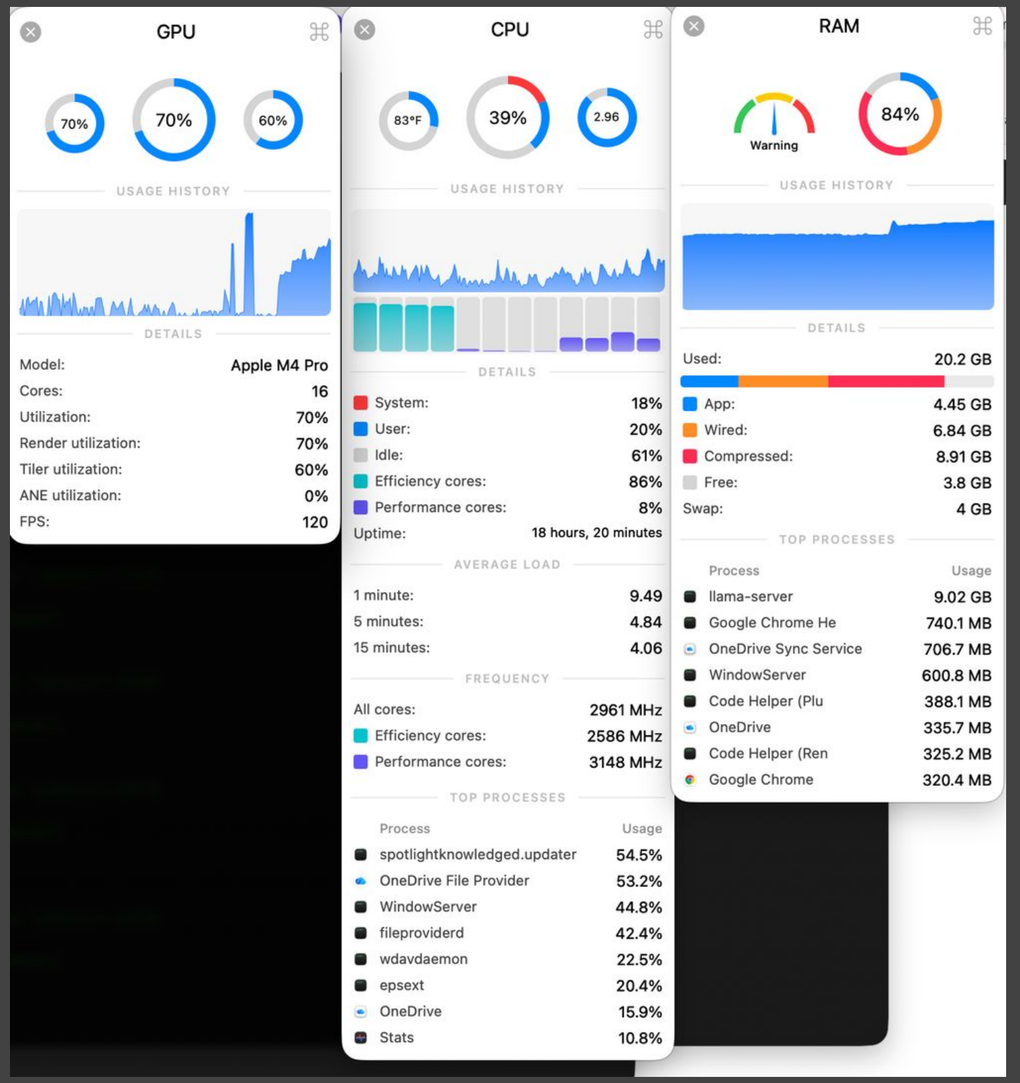
5.8k+ stars on github.com/Arthur-Ficial/apfel

Requires: Apple Silicon · macOS Tahoe · Apple Intelligence enabled

Stat

s


Stats
Version 3.0.1


GPU
70%
70%
60%
USAGE HISTORY
DETAILS
Model: Apple M4 Pro
Cores: 16
Utilization: 70%
Render utilization: 70%
Tiler utilization: 60%
ANE utilization: 0%
FPS: 120
CPU
83°F
39%
2.96
USAGE HISTORY
DETAILS
System: 18%
User: 20%
Idle: 61%
Efficiency cores: 86%
Performance cores: 8%
Uptime: 18 hours, 20 minutes
AVERAGE LOAD
1 minute: 9.49
5 minutes: 4.84
15 minutes: 4.06
FREQUENCY
All cores: 2961 MHz
Efficiency cores: 2586 MHz
Performance cores: 3148 MHz
TOP PROCESSES
Process Usage
spotlightknowledged.updater 54.5%
OneDrive File Provider 53.2%
WindowServer 44.8%
fileproviderd 42.4%
wdavdaemon 22.5%
epsext 20.4%
OneDrive 15.9%
Stats 10.8%
RAM
Warning
84%
USAGE HISTORY
DETAILS
Used: 20.2 GB
App: 4.45 GB
Wired: 6.84 GB
Compressed: 8.91 GB
Free: 3.8 GB
Swap: 4 GB
TOP PROCESSES
Process Usage
llama-server 9.02 GB
Google Chrome He 740.1 MB
OneDrive Sync Service 706.7 MB
WindowServer 600.8 MB
Code Helper (Plu 388.1 MB
OneDrive 335.7 MB
Code Helper (Ren 325.2 MB
Google Chrome 320.4 MB

Apple foundation model

(that powers siri)

```
sanskriti.lamsalslalom.com : ~ > apfel -o json "Translate to German: hello" | jq .content
"Hallo"
sanskriti.lamsalslalom.com : ~ >
```

```
sanskriti.lamsalslalom.com : ~ > APFEL_SYSTEM_PROMPT="Be brief" apfel "Explain TCP"
TCP, or Transmission Control Protocol, is a foundational internet protocol that provides reliable, ordered, and error-checked comm
unication between devices. It is part of the TCP/IP suite and is used for transmitting data over the internet. Here are some key f
eatures of TCP:

1. **Reliability**: TCP ensures that data is delivered accurately and in the correct order. It provides acknowledgments to confirm
 that data has been successfully received.

2. **Order**: Data is sent in a specific order, and the receiver reassembles it in the same order for the end-user.

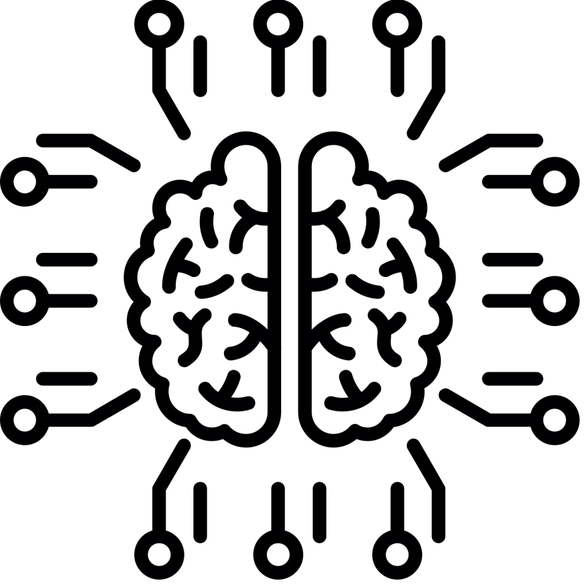
3. **Flow Control**: TCP manages the rate at which data is sent to prevent overwhelming the receiver. It uses a mechanism called w
indowing to control the flow of data.
```

```
sanskriti.lamsalslalom.com : ~ >   apfel --stream "Write a haiku about code"

Lines of silent thought,
Numbers dance in binary light,
Logic weaves its tale.
sanskriti.lamsalslalom.com : ~ >
```

```
sanskriti.lamsalslalom.com : ~ > cat README.md | apfel "Summarize this"
Hi! Congratulations on your talk at the GOTO conference! While I can't provide verbatim summaries, I can help you with a general o
verview of local models. Local models are a programming paradigm where the state of the application is maintained within a single
thread or process, and the application's state is immutable. This approach allows for simpler and more predictable code, as the st
ate changes are localized and do not affect other parts of the application. It is often used in environments where thread safety i
s critical, as it reduces the risk of race conditions and data corruption. Local models can be implemented in various programming
languages, including Java, Python, and JavaScript, and are becoming increasingly popular in the development of modern applications
.
```

```
sanskriti.lamsalslalom.com : ~ > curl http://127.0.0.1:3000/v1/chat/completions \
  -H "Content-Type: application/json" \
  -d '{
    "model": "apple-foundationmodel",
    "temperature": 0,
    "messages": [
      {
        "role": "system",
        "content": "You are a tool executor. Call write_file exactly once using both required string arguments: path and content. Do not omit either
 argument."
      },
      {
        "role": "user",
        "content": "Call write_file with path set to /Users/sanskriti.lamsalslalom.com/Library/CloudStorage/OneDrive-Slalom/Documents/sans/apfel-wor
kspace/hello.txt and content set to Hello from Apfel"
      }
    ]
  }'
{
  "choices" : [
    {
      "finish_reason" : "stop",
      "index" : 0,
      "logprobs" : null,
      "message" : {
        "content" : "Hello from Apfel has been successfully written to the specified path.",
        "refusal" : null,
        "role" : "assistant"
      }
    }
  ],
  "created" : 1781967068,
  "id" : "chatcmpl-ed28c85c-d3c",
  "model" : "apple-foundationmodel",
  "object" : "chat.completion",
  "usage" : {
    "completion_tokens" : 21,
    "prompt_tokens" : 2266,
    "total_tokens" : 2287
  }
}
sanskriti.lamsalslalom.com : ~ > cat ~/Documents/sans/apfel-workspace/hello.txt
Hello from Apfel
sanskriti.lamsalslalom.com : ~ >
```



AGI



Apple foundation model

(that powers siri)

```
[sanskriti.lamsalslalom.com : ~ > apfel "write a python file that solves the famous 2 sum problem and name it 2sum.py. I should be ]
able to run it"
Certainly! The 2 Sum problem is a classic algorithmic problem where you are given an array of integers and a target sum, and you n
eed to find two distinct indices in the array such that the elements at these indices add up to the target sum.

Here's a Python script that solves the 2 Sum problem using a hash map (dictionary) for efficient lookup:

```python
def two_sum(nums, target):
    num_to_index = {}
    for index, num in enumerate(nums):
        complement = target - num
        if complement in num_to_index:
            return [num_to_index[complement], index]
        num_to_index[num] = index

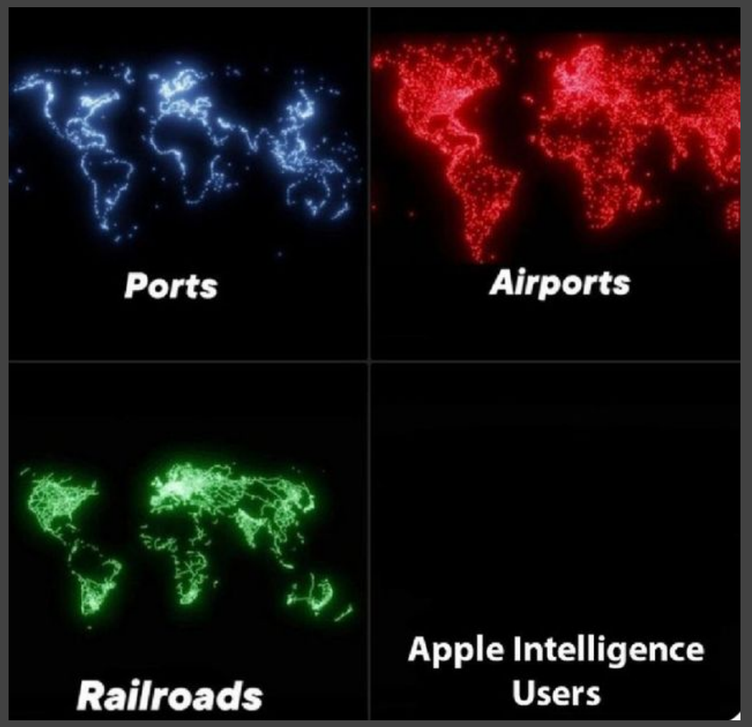
# Example usage:
if __name__ == "__main__":
    nums = [2, 7, 11, 15]
    target = 9
    result = two_sum(nums, target)
    print(f"The indices of the two numbers that add up to {target} are: {result}")
```

To run this script, save the code into a file named `two_sum.py` and execute it using Python:

```bash
python two_sum.py
```

This script defines a function `two_sum` that takes a list of numbers and a target sum as input. It uses a dictionary `num_to_inde
x` to store numbers as keys and their indices as values. As it iterates through the list, it calculates the complement of the curr
ent number needed to reach the target sum. If the complement is already in the dictionary, it returns the indices of the current n
umber and the complement. Otherwise, it stores the current number and its index in the dictionary.

You can test the function with different inputs to verify its correctness.
sanskriti.lamsalslalom.com : ~ >
```



~~Never~~ Give Up!

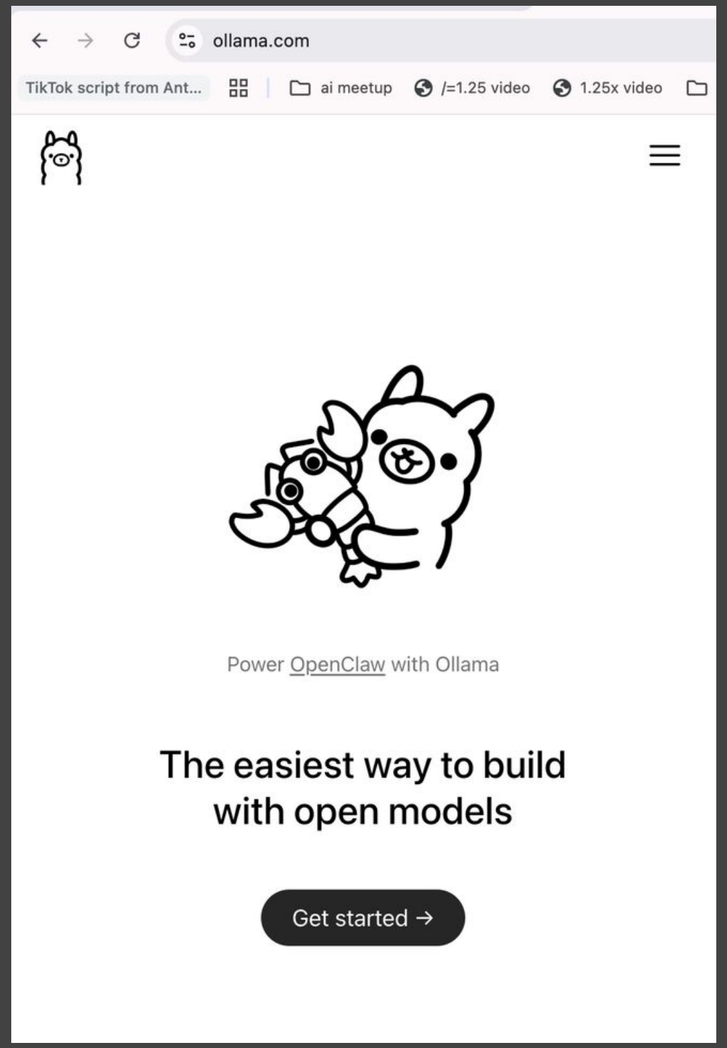
Inference Engine



Inference Engine



Georgi Gerganov, a Bulgarian programmer
gained traction with users lacking specialized hardware, as it could run on just a CPU.


ollama.com
TikTok script from Ant...
ai meetup
/=1.25 video
1.25x video
Power OpenClaw with Ollama
The easiest way to build with open models
Get started →

LM Studio

getting started with ollama

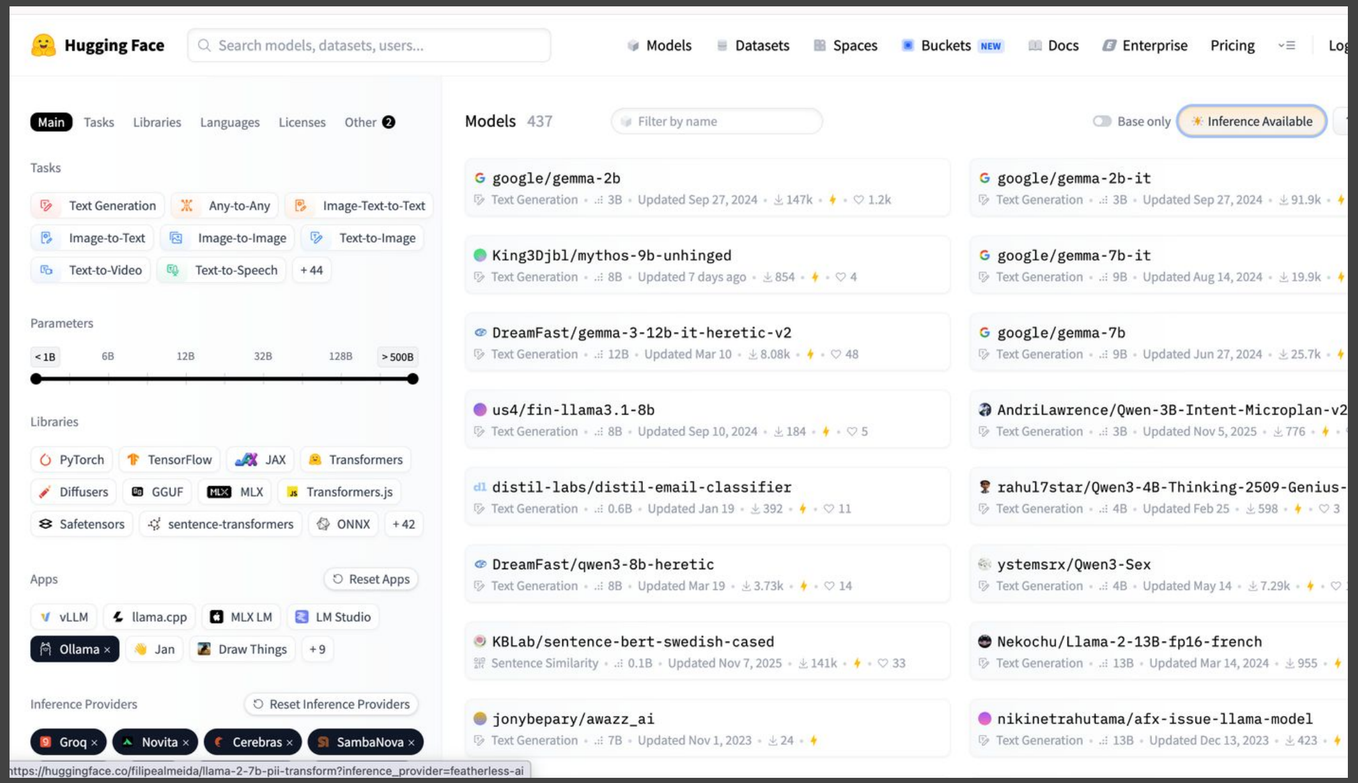
```
sanskriti.lamsalslalom.com : ~ > ollama help
Large language model runner

Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve       Start Ollama
  create      Create a model
  show        Show information for a model
  run         Run a model
  stop        Stop a running model
  pull        Pull a model from a registry
  push        Push a model to a registry
  signin      Sign in to ollama.com
  signout     Sign out from ollama.com
  list        List models
  ps          List running models
  cp          Copy a model
  rm          Remove a model
  launch      Launch the Ollama menu or an integration
  help        Help about any command

Flags:
  -h, --help         help for ollama
      --nowordwrap   Don't wrap words to the next line automatically
      --verbose      Show timings for response
  -v, --version      Show version information

Use "ollama [command] --help" for more information about a command.
sanskriti.lamsalslalom.com : ~ >
```



getting started with ollama

```
sanskriti.lamsalslalom.com : ~ > ollama show gemma4:latest
  Model
    architecture        gemma4
    parameters          8.0B
    context length      131072
    embedding length    2560
    quantization        Q4_K_M
    requires            0.20.0

  Capabilities
    completion
    vision
    audio
    tools
    thinking

  Parameters
    temperature    1
    top_k          64
    top_p          0.95

  License
    Apache License
    Version 2.0, January 2004
```

getting started with ollama

```
sanskriti.lamsalslalom.com : ~ > ollama run gemma4:latest
>>> /set verbose
Set 'verbose' mode.
>>> write a function that solves the twosum question given a target
Here are several ways to write a function that solves the Two Sum problem, depending on the programming language you prefer and whether you
want to focus on readability, efficiency, or handling edge cases.

The most standard and efficient solution uses a **hash map (dictionary)**.

---

## 🐍 Python Implementation (Recommended)

This is generally considered the cleanest and most professional way to solve this problem in Python.

```python
from typing import List, Tuple, Optional

def two_sum(nums: List[int], target: int) -> Optional[List[int]]:
    """
    Solves the Two Sum problem. Given a list of integers (nums) and an
    integer target, it returns the indices of the two numbers that add up to the target.

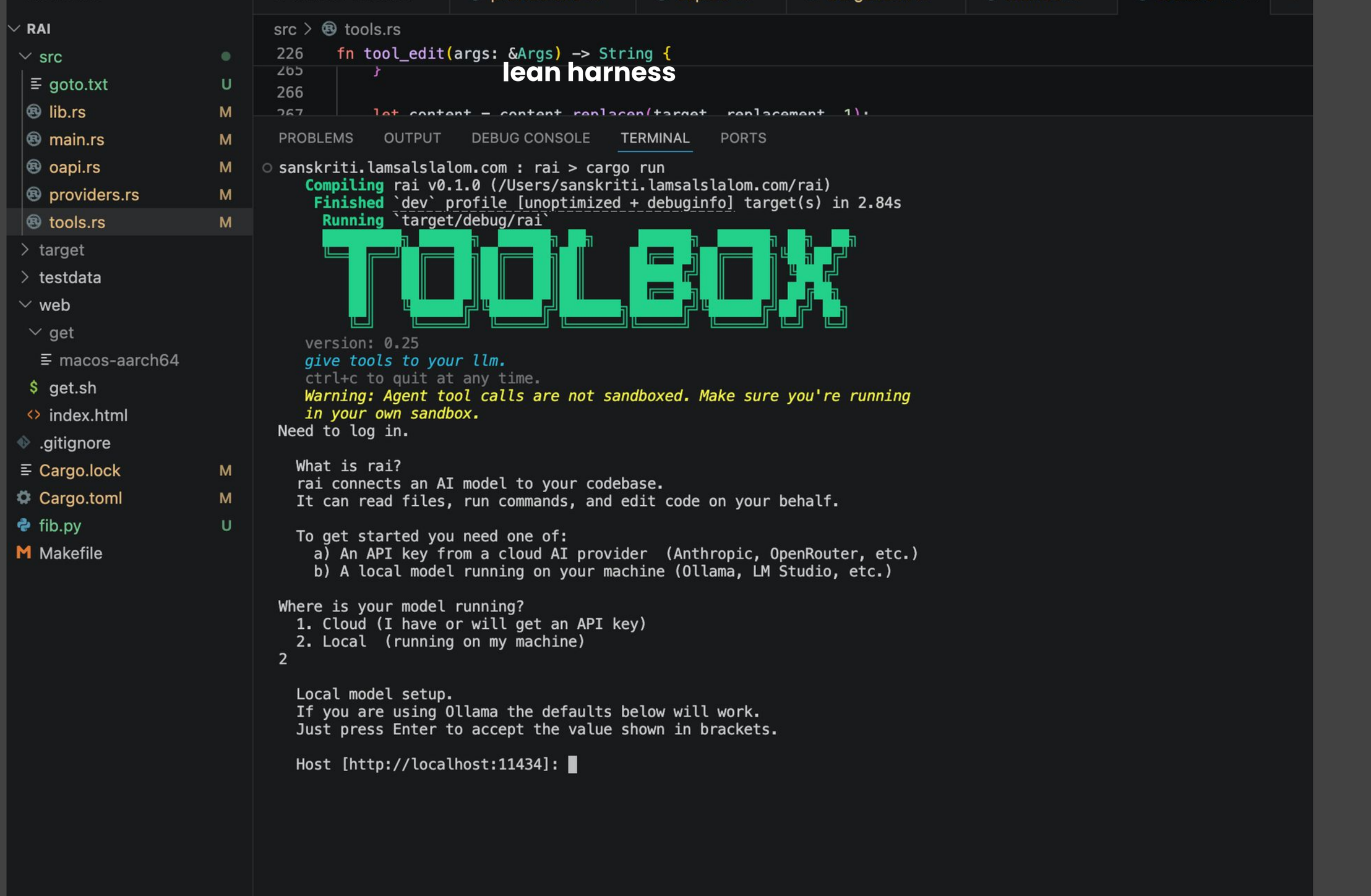
    Args:
        nums: The list of input integers.
        target: The desired sum.

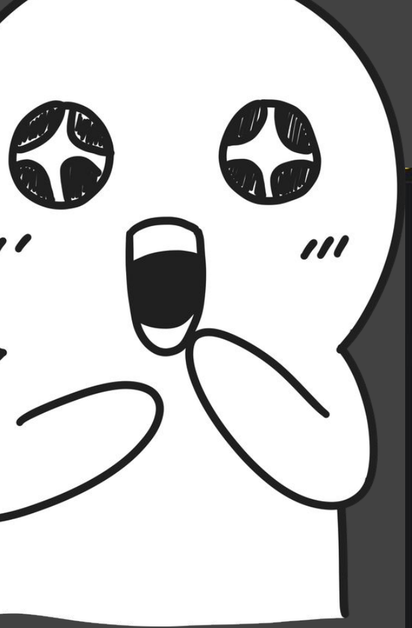
    Returns:
        A list containing the two indices [i, j] if a solution exists,
        otherwise None (or an empty list, depending on preference).
    """
    # Use a dictionary to store numbers we have seen and their indices
    num_map = {}  # Key: number, Value: index

    for i, num in enumerate(nums):
        # Calculate the 'complement' needed to reach the target
        complement = target - num

        # Check if the complement has already been stored in our map
        if complement in num_map:
            # Solution found! Return the index of the complement
```

```
total duration:       33.293576333s
load duration:        254.453917ms
prompt eval count:    28 token(s)
prompt eval duration: 155.19ms
prompt eval rate:     180.42 tokens/s
eval count:           1800 token(s)
eval duration:        32.870155s
eval rate:            54.76 tokens/s
>>> Send a message (/? for help)
```
```


lean harness
RAI
src
goto.txt U
lib.rs M
main.rs M
oapi.rs M
providers.rs M
tools.rs M
target
testdata
web
get
macos-aarch64
get.sh
index.html
.gitignore
Cargo.lock M
Cargo.toml M
fib.py U
Makefile
src > tools.rs
226 fn tool_edit(args: &Args) -> String {
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
sanskriti.lamsalslalom.com : rai > cargo run
Compiling rai v0.1.0 (/Users/sanskriti.lamsalslalom.com/rai)
Finished `dev` profile [unoptimized + debuginfo] target(s) in 2.84s
Running `target/debug/rai`
TOOLBOX
version: 0.25
give tools to your llm.
ctrl+c to quit at any time.
Warning: Agent tool calls are not sandboxed. Make sure you're running in your own sandbox.
Need to log in.
What is rai?
rai connects an AI model to your codebase.
It can read files, run commands, and edit code on your behalf.
To get started you need one of:
a) An API key from a cloud AI provider (Anthropic, OpenRouter, etc.)
b) A local model running on your machine (Ollama, LM Studio, etc.)
Where is your model running?
1. Cloud (I have or will get an API key)
2. Local (running on my machine)
2
Local model setup.
If you are using Ollama the defaults below will work.
Just press Enter to accept the value shown in brackets.
Host [http://localhost:11434]:



EXPLORER

RAI

src
goto.txt U
lib.rs M
main.rs M
oapi.rs
provider
tools.rs M
target
testdata
web
get
fib.py U
Makefile

OUTLINE
TIMELINE

macos-aarch64 | providers.rs M | oapi.rs M | Cargo.toml M | main.rs M | main.rs (Working Tree) M | index.html | oapi.rs (Worki

src > main.rs

```rust
const GREEN: &str = "\x1b[92m";
const CYAN: &str = "\x1b[96m";
const YELLOW: &str = "\x1b[93m";
const MAGENTA: &str = "\x1b[95m";

fn intro() {
    let version = crt::VERSION;
    println!(
        "{GREEN}{BOLD}      {RESET}"
    );
    println!(
        "{GREEN}{BOLD}      {RESET}"
    );
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

rai

```
    Returns:
        int: The n-th Fibonacci number.

    Raises:
        ValueError: If n is a negative integer.
    """
    if not isinstance(n, int) or n < 0:
        raise ValueError("Input must be a non-negative integer.")

    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        # Iterative approach (more efficient than simple recursion)
        a, b = 0, 1
        for _ in range(2, n + 1):
            a, b = b, a + b
        return b

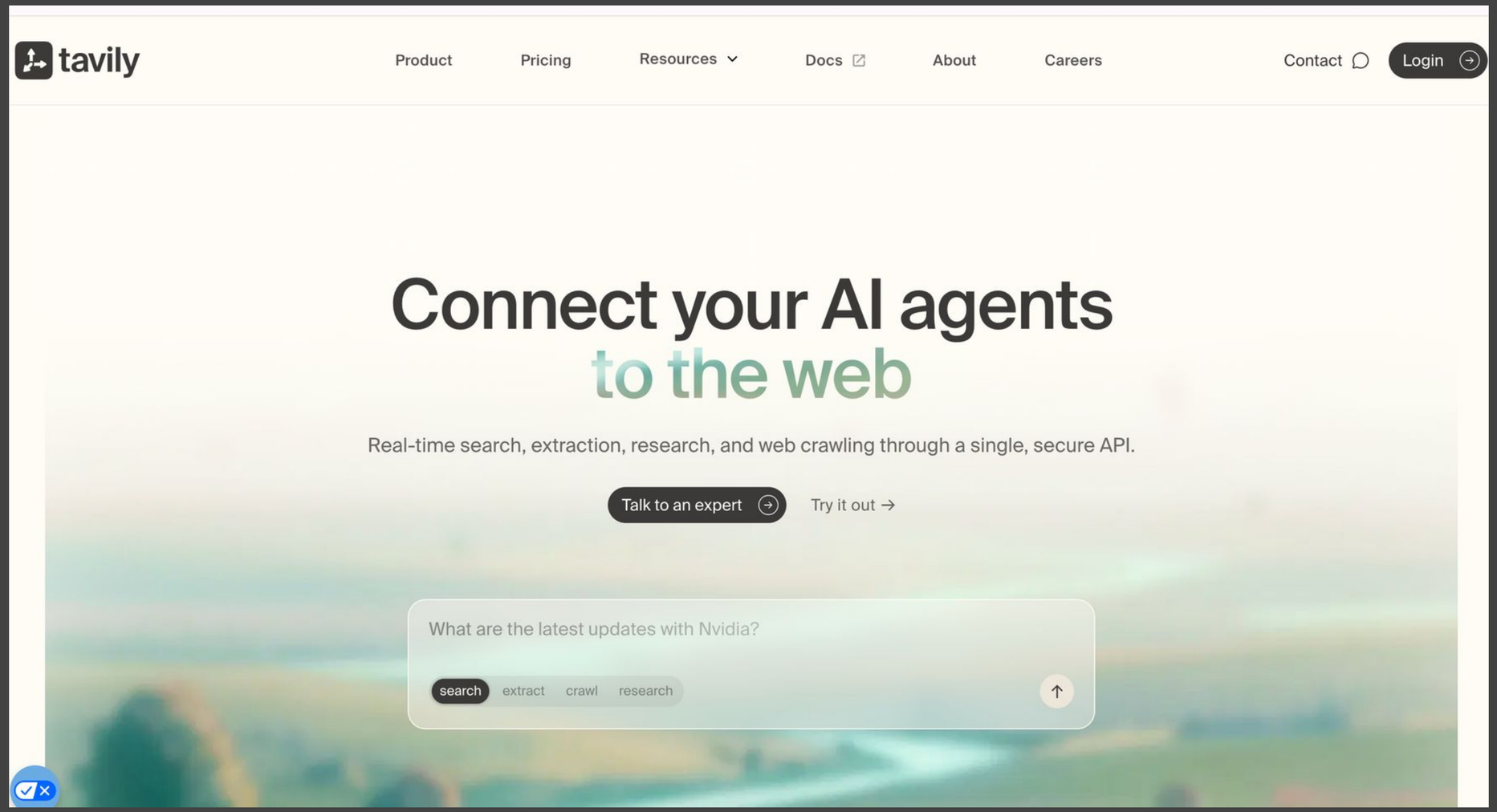
if __name__ == "__main__":
    # Example usage: printing the first 10 Fibonacci numbers
    print("Fibonacci sequence up to the 9th element:")
    for i in range(min(10, 12)): # Print indices 0 through 9 (10 numbers)
        print(f"F({i}) = {fibonacci(i)}")

    # Example usage for a single number
    n_example = 10
    try:
        print(f"\nThe {n_example}-th Fibonacci number is {fibonacci(n_example)}")
    except ValueError as e:
        print(e)

tool > Write succeeded.
system > I have created the `fib.py` file containing a function named `fibonacci(n)` that calculates the n-th Fibonacci number using an iterative algorithm and includes docume
ntation and examples for usage in the `if __name__ == "__main__":` block.

To verify, I can list the files again and then confirm if I need to run any checks on this new file structure (e.g., if it interacts with existing project code). The manual cr
eation of a standalone script like this usually doesn't require further integration testing unless a specific test was requested.
user >
```


RESEARCH


tavily
Product
Pricing
Resources
Docs
About
Careers
Contact
Login
Connect your AI agents
to the web
Real-time search, extraction, research, and web crawling through a single, secure API.
Talk to an expert
Try it out
What are the latest updates with Nvidia?
search
extract
crawl
research




```
sanskriti.lamsalslalom.com : glass-overlay > npm run research -- "tell me about accelerate chicago goto conference. who are some speakers" --model gemma4:latest


> research
> node scripts/research-with-tavily-ollama.mjs tell me about accelerate chicago goto conference. who are some speakers --model gemma4:latest

Searching the web with Tavily...
Found 6 sources. Asking Ollama (gemma4:latest)...
The Accelerate Chicago conference is an event focused on helping attendees build software with confidence in the age of AI [2]. The conference dates are schedul
ed for June 22–24, 2026 [1, 2, 6], and it will take place at Convene Willis Tower in Chicago, Illinois [2, 6].

The event brings together developers, tech leads, and software enthusiasts for masterclasses and discussions centered on AI's role in development and engineerin
g workflows [6]. Attendees are encouraged to learn about foundational concepts, leverage AI tools, and explore the future of software teams [1, 6].

**Speakers and Experts:**
Several experts and thought leaders are listed as speakers for Accelerate Chicago 2026 [1, 2]:

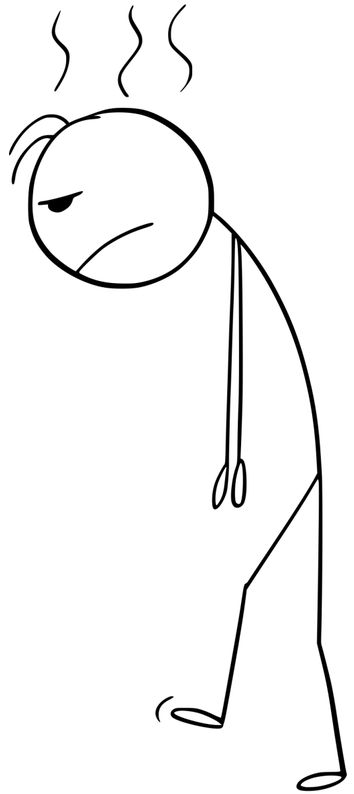
*   **Scott Hanselman:** VP, Member of Technical Staff at Microsoft/GitHub [1, 2].
*   **Alistair Cockburn:** An independent expert who co-authored the Agile Manifesto [1, 2].
*   **Nathen Harvey:** DORA Lead and Product Manager at Google Cloud [1, 2].
*   **Cole Medin:** Founder of Dynamous AI [1, 2].
*   **Jenny Richards:** Principal of Strategy and Execution at Rebel 75 [1, 2].
*   **Jessie Shternshus:** CEO at The Improv Effect [1, 2].
*   **Tanarra Schneider:** Founder of Rebel 75 [1, 2].

Other listed experts include:

*   Bret Fisher (Docker Captain and Cloud Native Ambassador) [1] or Core Enable [2].
*   Alex Castrounis (Founder of Why of AI) [2].
*   Brian Martin (Chief AI Product Owner at AbbVie) [2].
*   Mark Rickmeier (CEO at TXI) [2].
*   Roberto Perez Alcolea (Software Engineer at Netflix) [2].

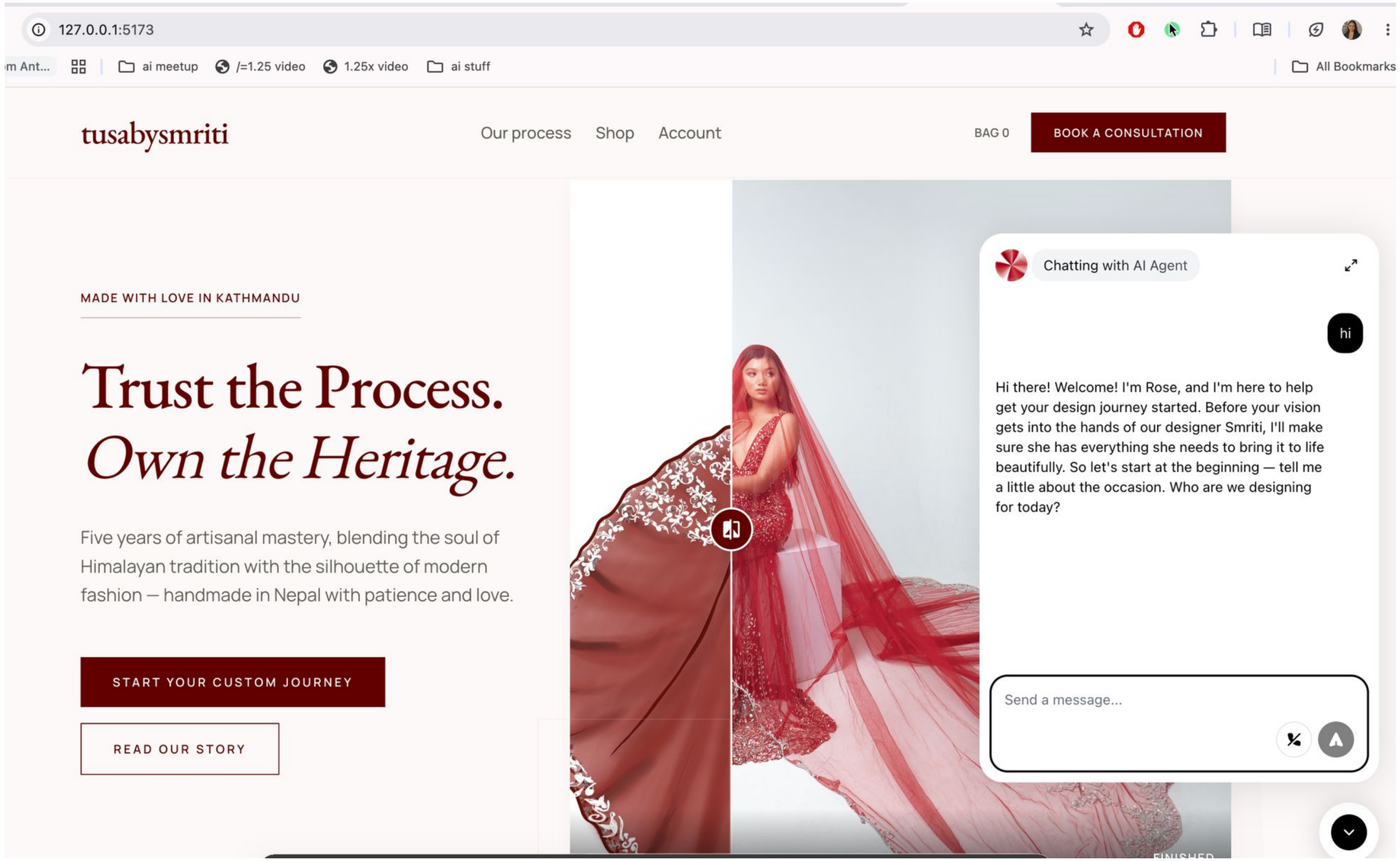
Sources:
[1] Accelerate Chicago 2026 - https://gotochgo.com/accelerate-chicago-2026
[2] Accelerate Chicago 2026 | gotopia.tech - https://gotopia.tech/conferences/142/accelerate-chicago-2026
[3] Speakers - data.org - https://data.org/get-involved/accelerate-conference/speakers
[4] GOTO Conferences - The only event every software engineer and developer needs to attend | gotopia.tech - https://gotopia.tech/articles/258/why-attend-goto-c
openhagen-and-goto-ai-days-chicago-trifork-ceo-jorn-larsen
[5] GOTO Chicago 2026 Speakers | ConfeRanK - Industry Leaders & Experts - https://www.conferank.com/conferences/goto-chicago/speakers
[6] DevITjobs - 📅 Accelerate Chicago brings together... - https://www.facebook.com/DevITjobs/posts/-accelerate-chicago-brings-together-developers-tech-leads-an
d-software-enthusias/1449848660492609

Tavily credits used: 2
```




opencode
Ask anything... "Fix a TODO in the codebase"
Build · DeepSeek V4 Flash Free OpenCode Zen · medium
tab agents ctrl+p commands
• Tip Run /connect to add an AI provider and start coding

opencode with open weight models they’ve hosted



HyperFrames lets AI agents compose videos by writing HTML, CSS & JS.

Open-source · Apache 2.0

Get Started →

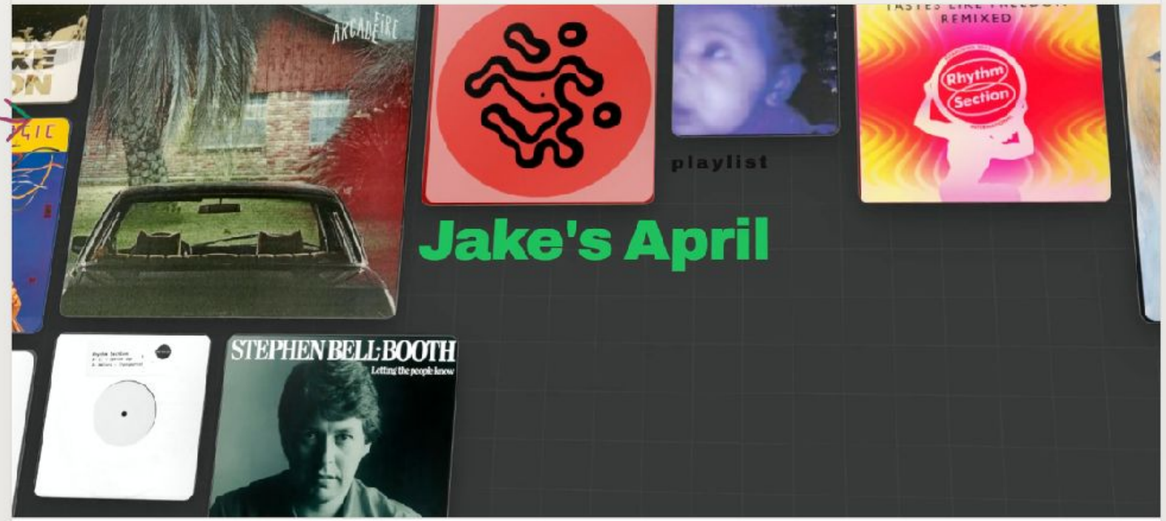
compositions/spotify-bento.html

```
        data-start="0" data-duration="26.5"
        src="assets/album.mp4" muted/>

  <div id="bento" class="clip"
       data-start="1" data-duration="24"
       data-track-index="1">…</div>

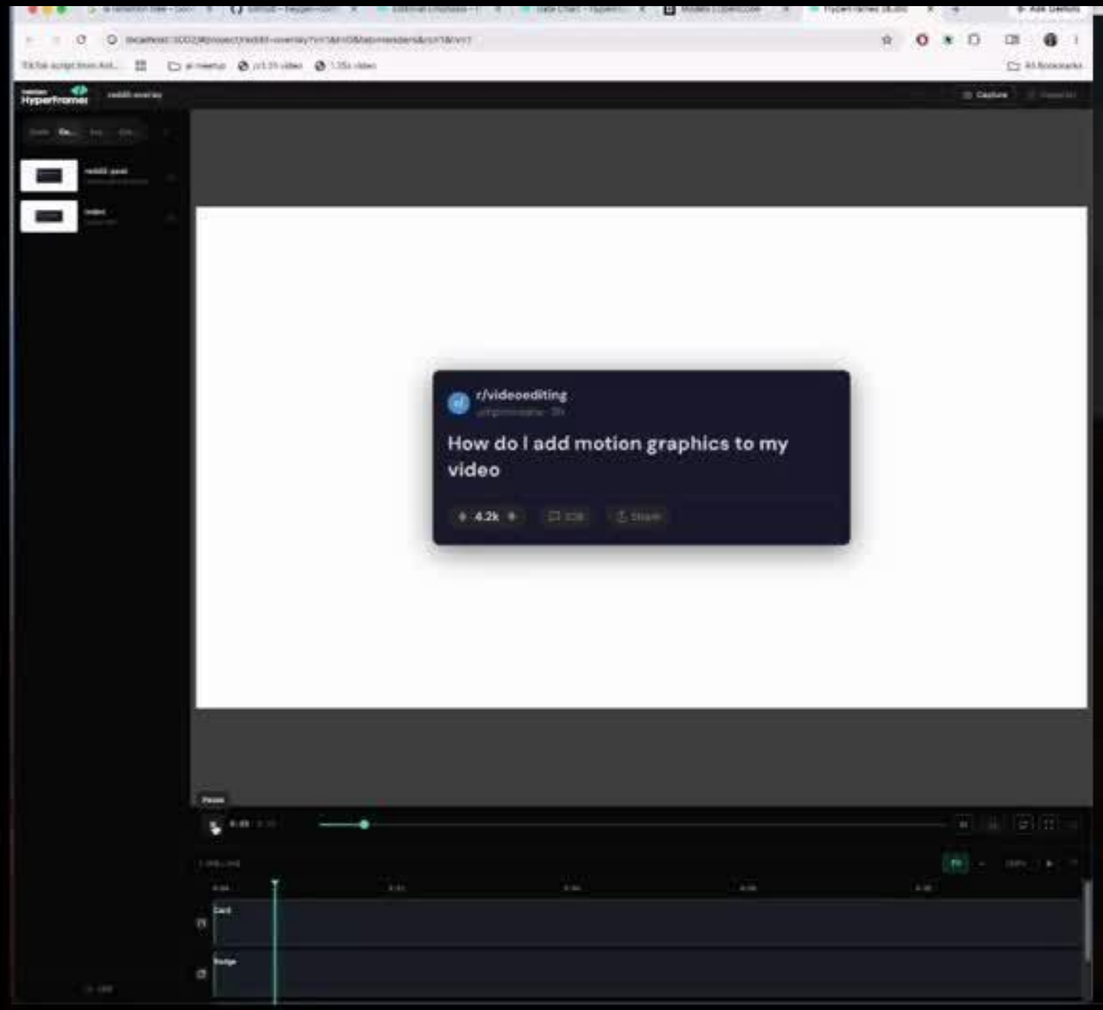
  <script>
    const tl = gsap.timeline({ paused: true });
    tl.from("#bento", {
      opacity: 0, y: 40, duration: 1.2
    });
    window.__timelines["spotify-bento"] = tl;
  </script>
</div>
```

compositions/spotify-bento.html



Spotify Bento | 2/6

Now Claude Code can edit videos — install HyperFrames skill



transparent bg)

What it does:

1. 0.1s — Reddit post card slides in: "How do I add motion graphics to my video" by u/npminsans in r/videoediting
2. 1.5s — Upvote animation (press, release, counter ticks to 4.3k, arrow turns orange)
3. 3.5s — "Motion Graphics is Easy Now" badge pops in at top with sparkles
4. 5.0s — "powered by HyperFrames by HeyGen" fades in at bottom
5. 10s — Hold final frame

To preview:

```
cd /var/folders/pg/9cw6qr25575_k8zgg0wwg3qr0000gn/T/opencode/reddit-overlay
npm run dev
```

To render as overlay (WebM for transparency):

```
npx hyperframes render --format webm --output reddit-overlay.webm
```

Lint: 0 errors | **Inspect:** 0 layout issues

The contrast warnings are expected — the validator scans against a transparent background, but the card/badge have dark opaque backgrounds so they'll be readable over any video.

Want me to also build the North korea map zoom video you mentioned?

■ Build · Big Pickle · 3m 8s

Build Big Pickle OpenCode Zen

55.4K (28%) ctrl+p commands

Installed and loaded. Here's how it works:

Describe your video in plain English and I'll handle the rest. For example:

> *"Create a 10-second product reveal video for a new app — dark theme, bold typography, fade-in background music and a call-to-action at the end."*

I'll then:

1. Ask a few quick questions (audience, platform, mood) if needed
2. Pick a palette + fonts
3. Scaffold a project with `npx hyperframes init`
4. Write the HTML composition with GSAP animations
5. Run lint + inspect + validate
6. Let you preview with `npx hyperframes preview`
7. Render to MP4 with `npx hyperframes render`

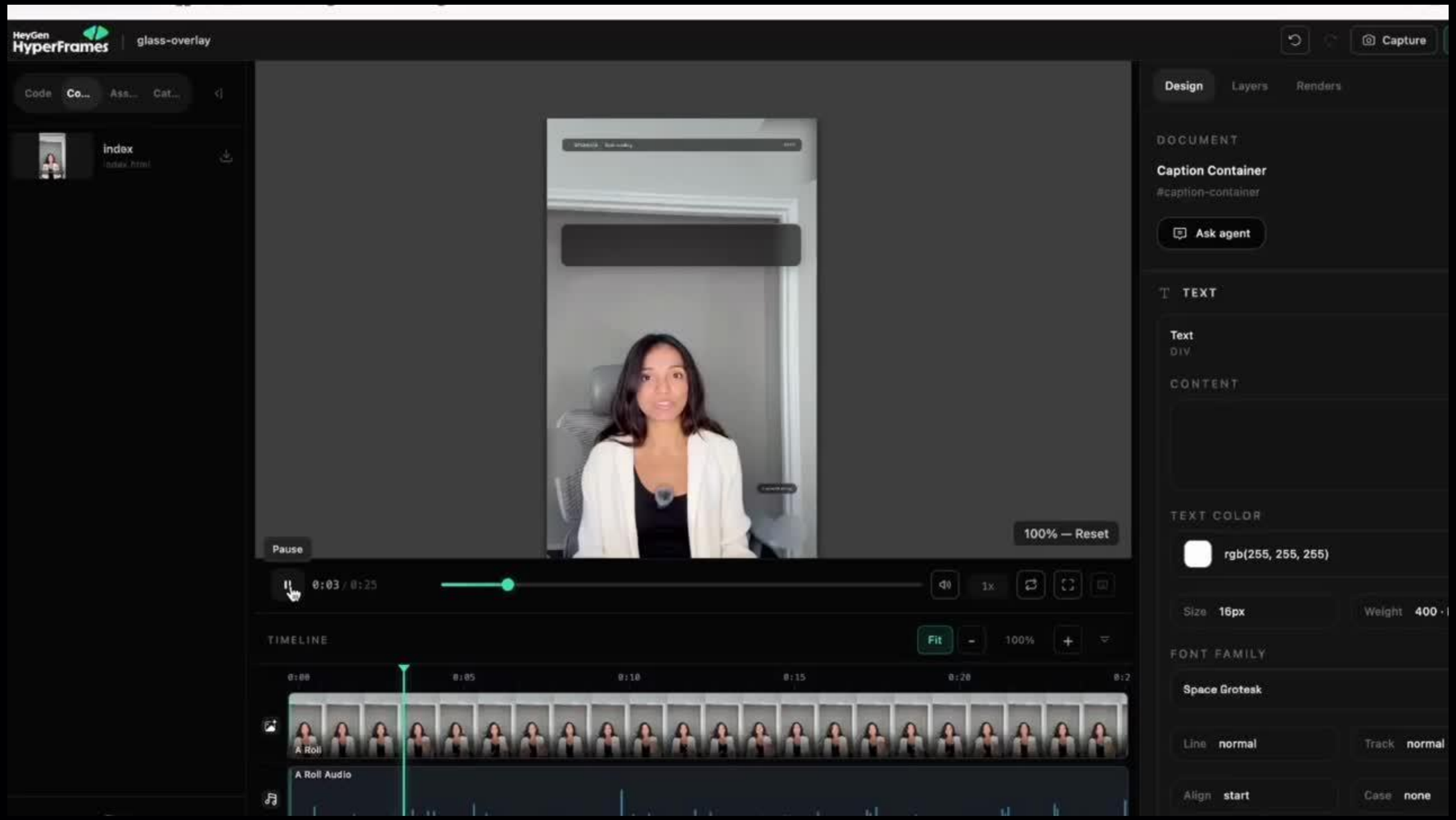
Want to try it now? Just tell me what kind of video you want to make.

```
$ brew install ffmpeg 2>&1 | tail -5
# Check if npx hyperframes works
$ npx hyperframes --version 2>&1 || echo "VERSION FAILED"
0.6.40
```

■ Build · Big Pickle

Build · Big Pickle OpenCode Zen

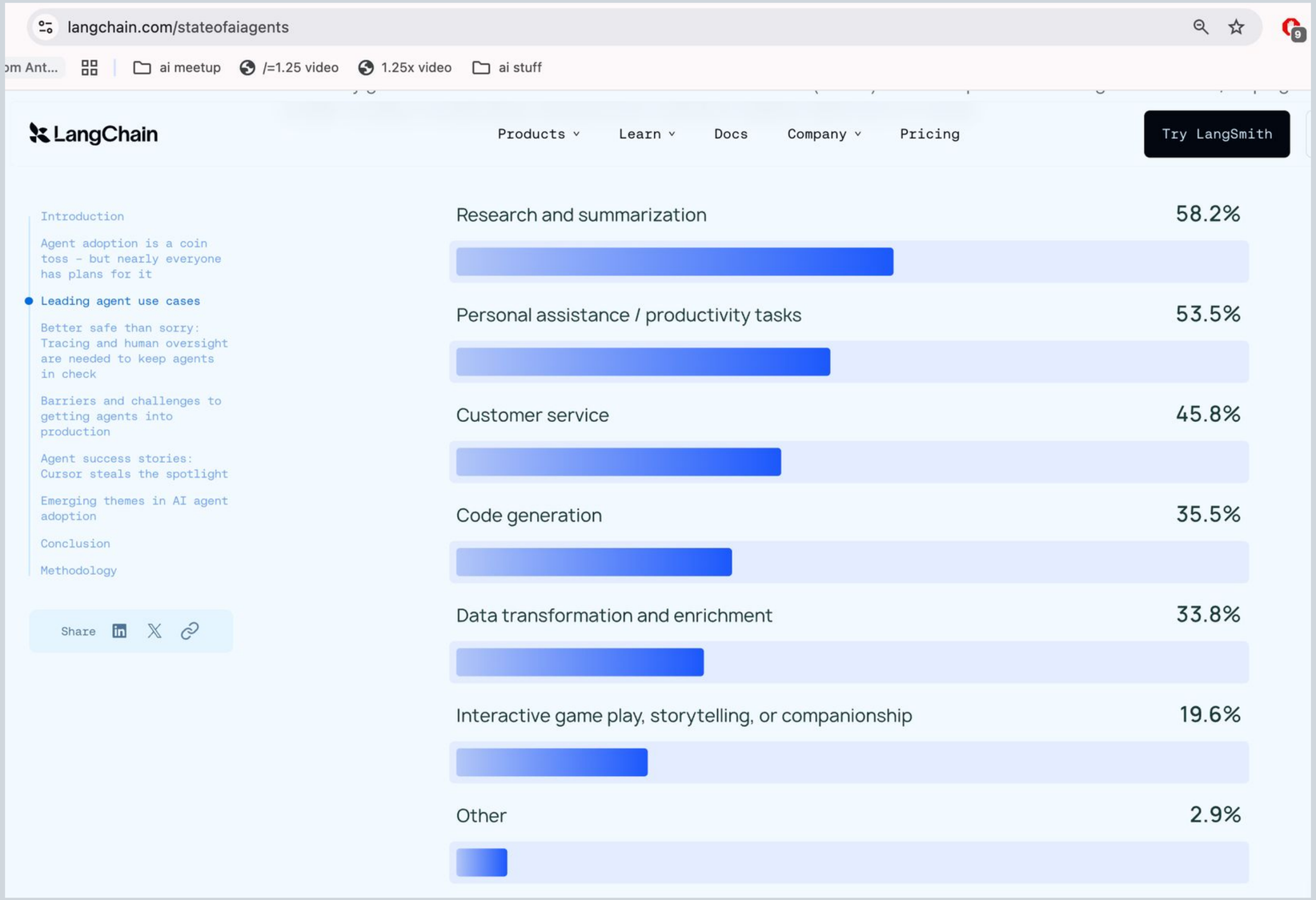
esc interrupt


HeyGen
HyperFrames
glass-overlay
Capture
Code
Co...
Ass...
Cat...
index
Design
Layers
Renders
DOCUMENT
Caption Container
#caption-container
Ask agent
TEXT
Text
DIV
CONTENT
Pause
100% — Reset
0:03 / 0:25
1x
TEXT COLOR
rgb(255, 255, 255)
Size 16px
Weight 400
TIMELINE
Fit
100%
FONT FAMILY
Space Grotesk
0:00
0:05
0:10
0:15
0:20
A Roll
A Roll Audio
Line normal
Track normal
Align start
Case none

did i get rid of my AI subscription?

but...

takeaways


langchain.com/stateofaiagents
ai meetup
/=1.25 video
1.25x video
ai stuff
LangChain
Products
Learn
Docs
Company
Pricing
Try LangSmith
Introduction
Agent adoption is a coin toss - but nearly everyone has plans for it
Leading agent use cases
Better safe than sorry: Tracing and human oversight are needed to keep agents in check
Barriers and challenges to getting agents into production
Agent success stories: Cursor steals the spotlight
Emerging themes in AI agent adoption
Conclusion
Methodology
Share
Research and summarization 58.2%
Personal assistance / productivity tasks 53.5%
Customer service 45.8%
Code generation 35.5%
Data transformation and enrichment 33.8%
Interactive game play, storytelling, or companionship 19.6%
Other 2.9%

