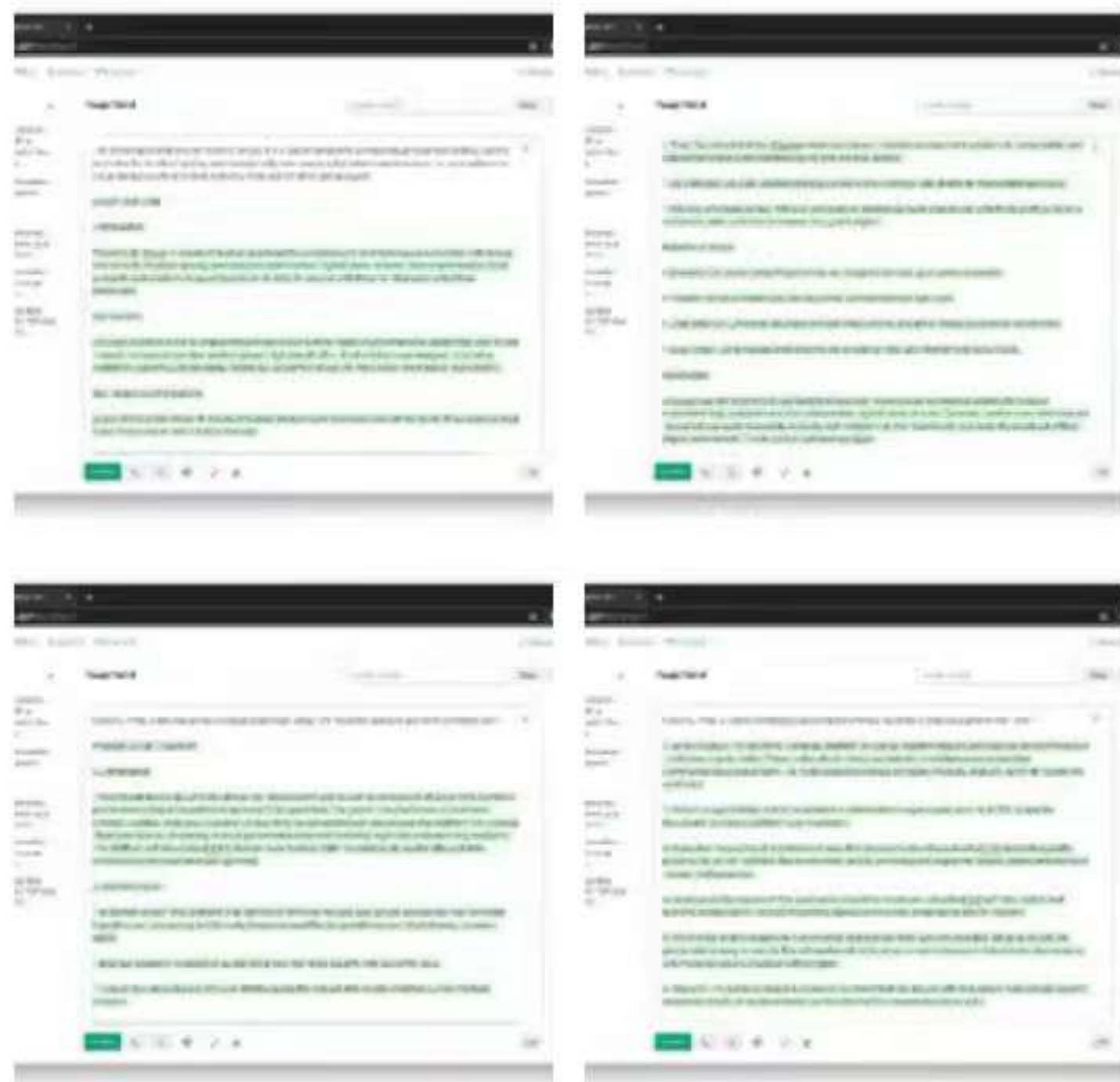


基于端+云打造 AI 应用



VTHINK

玩了一会 ChatGPT，感觉 AI 取代人类工作真的不远了



2022年12月3日 23:57 



Lepton Search beta

[Try Browser Extension](#)

Ask Lepton AI anything ...



liquid vs solid vs gas?

When did human first land on the moon?

Run your own search and other AI applications on Lepton. [Talk to us!](#)

[Lepton AI](#) [Twitter](#) [Github](#)

Fork 1k

Starred 7.9k

About



Building a quick conversation-based search demo with Lepton AI.

search.lepton.run



Aravind Srinivas
@AravSrinivas



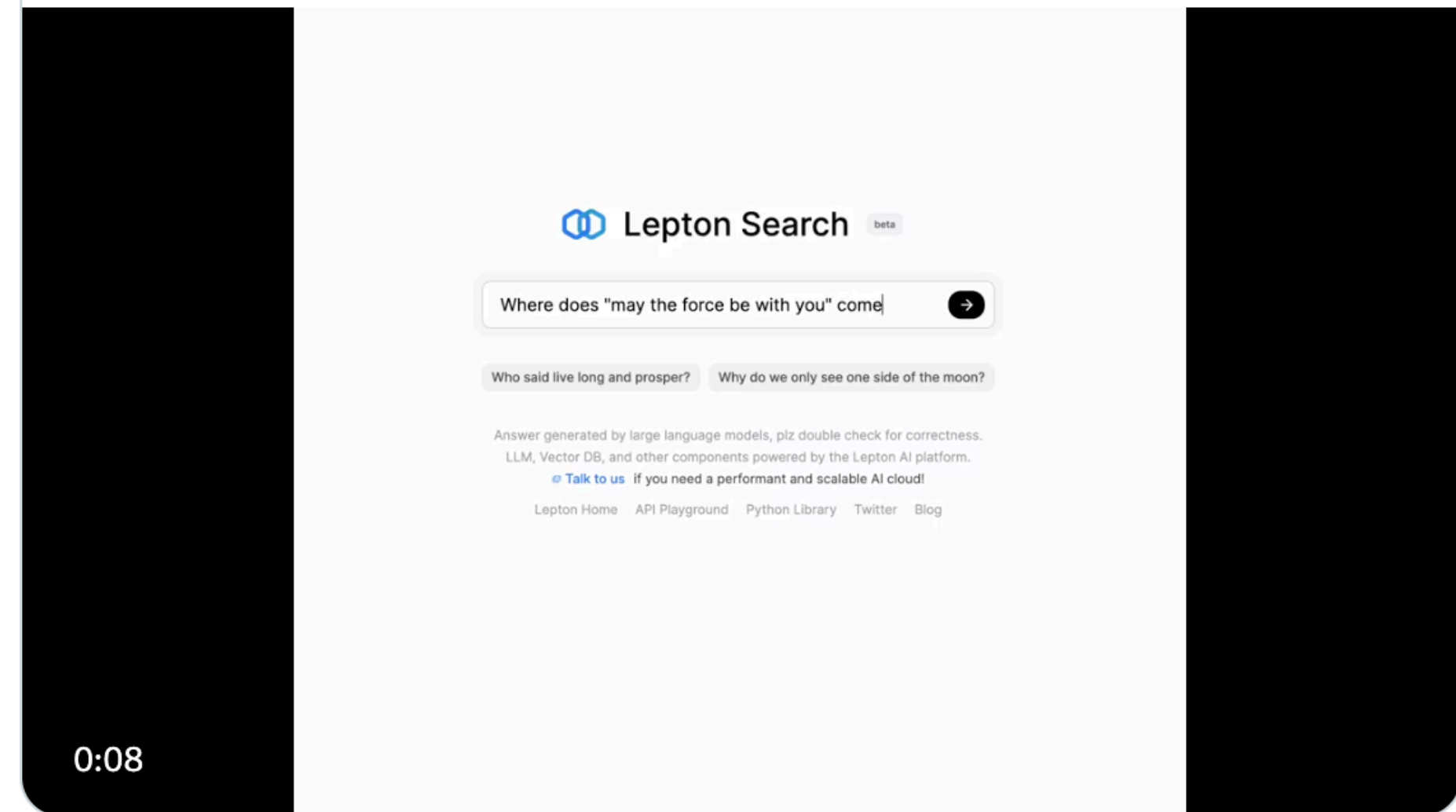
Happy to see perplexity be the reference UI and example for future fund raising efforts, including those by former Director and VP of Meta and Alibaba. The impact of the company is on the ecosystem and industry itself, and not just on the product vertical.

[翻译帖子](#)



Yangqing Jia @jiayq · 1月25日

Building an AI app has never been easier. Over the weekend, we built a demo for conversational search with <500 lines of python, and it's live at search.lepton.run. Give it a shot! Code to be open sourced soon as we clean up all "# temp scaffolds" stuff. (1/x)



下午9:20 · 2024年1月25日 · 12.1万 查看

16

26

386


173



Elmo Chat - Your AI Web Copilot

 elmo.chat  Featured 4.6 ★ (85 ratings)

Extension Tools 40,000 users

⋮  v18.3.1    

API REFERENCE > APIS >

use

Canary

The `use` API is currently only available in React's Canary and experimental channels. Learn more about [React's release channels here](#).

`use` is a React API that lets you read the value of a resource like a [Promise](#) or [context](#).

```
const value = use(resource);
```

 Elmo Chat - Your AI Web Copilot 

What is the difference between `use` and `useContext`?

The `use` API and `useContext` are both React APIs used to read values from a context, but they have some key differences:

1. Usage Flexibility:

- `useContext` must be called at the top level of your component, meaning it cannot be used inside loops, conditions, or nested functions.
- `use` can be called inside loops, conditions, and nested functions, providing more flexibility in how and where you can read context values.

2. API Design:

- `useContext` takes a context object and returns the current context value for that context.
- `use` takes a resource, which can be a context or a Promise, and returns the resolved value of the Promise

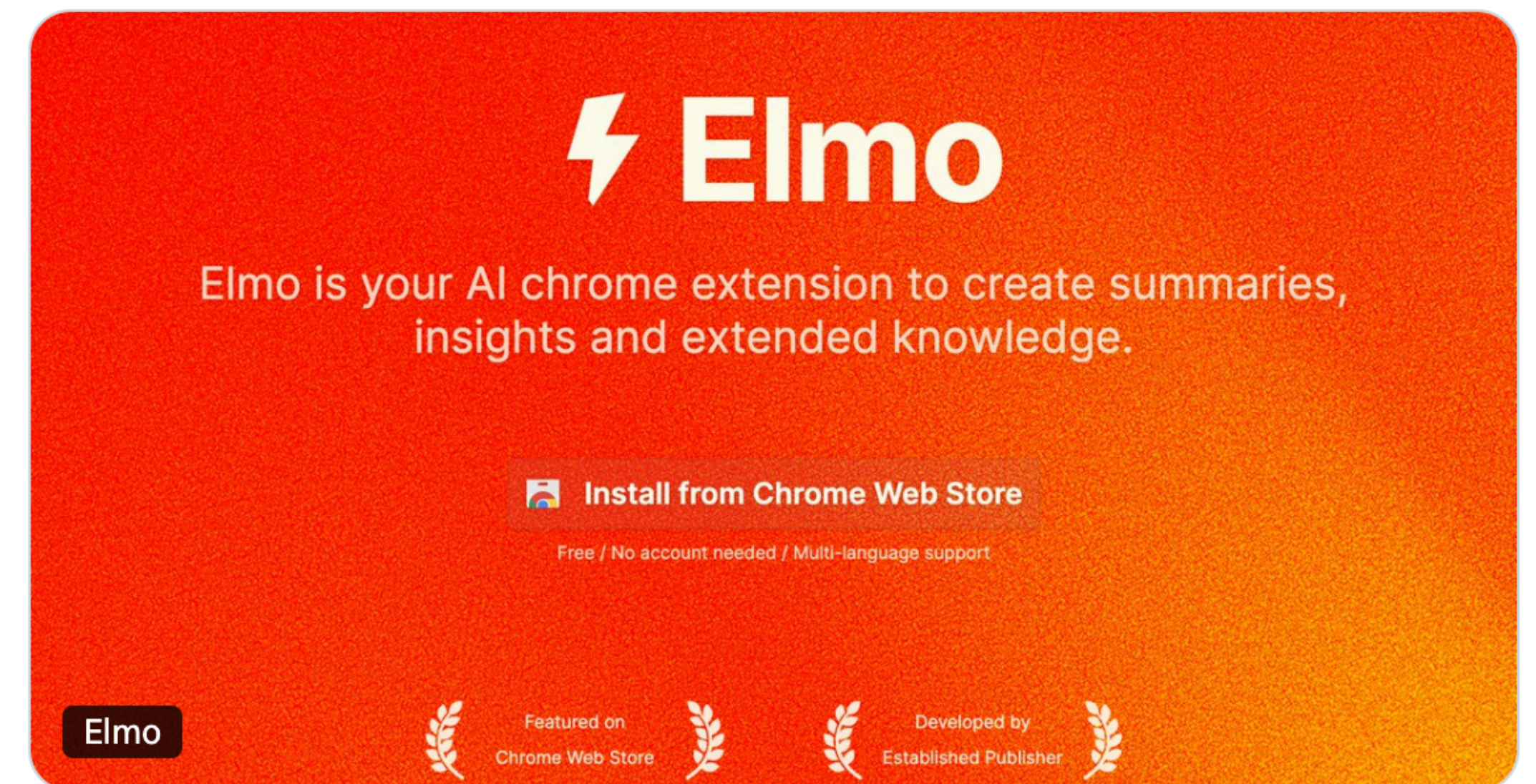
Ask Elmo a question or type '/' for commands   



Guillermo Rauch  
@rauchg

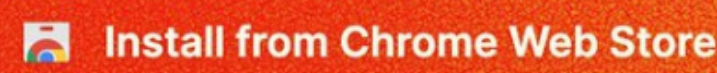
This looks really handy

[翻译帖子](#)

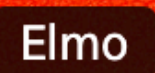
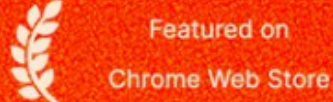
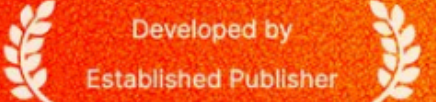


Elmo

Elmo is your AI chrome extension to create summaries, insights and extended knowledge.

 Install from Chrome Web Store

Free / No account needed / Multi-language support

  Featured on Chrome Web Store  Developed by Established Publisher

来自 elmo.chat

下午12:08 · 2024年4月2日 来自 California, USA · 12.8万 查看

 18

 39

 411

 333



Javascript Developer



 Lepton

[Try Brows](#)

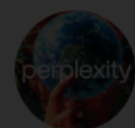
Ask Lepton AI anything ...



liquid vs solid vs gas? When d

Run your own search and other AI a

[Lepton AI](#) 

 Fork 1k



Aravind Srinivas  
@AravSrinivas

Happy to see perplexity be the refere
raising efforts, including those by for
Alibaba. The impact of the company
itself, and not just on the product ver
翻译帖子



Yangqing Jia  @jiayq · 1月25日

Building an AI app has never been easie
for conversational search with <500 line

AI Web Copilot

(ratings)

mo

ision to create summaries,
ded knowledge.





































rome Web Store

Multi-language support

 Developed by
Established Publisher 

Server Side AI

The Top 50 Gen AI Web Products, by Unique Monthly Visits

1.  ChatGPT	11.  SpicyChat	21.  VIGGLE	31.  PIXAI	41.  MaxAI.me
2. character.ai	12. ElevenLabs	22.  Potoroom	32.  Clipchamp	42.  BLACKBOX AI
3.  perplexity	13.  Hugging Face	23.  Gamma	33. ud.io	43.  CHATPDF
4. Claude	14.  LUMA AI	24. VEED.IO	34.  Chatbot App	44.  Gauth
5. SUNO	15. candy.ai	25.  PIXLR	35. VocalRemover	45.  COZE
6. JanitorAI	16.  Crushon AI	26.  ideogram	36.  PicWish	46.  Playground
7.  QuillBot	17.  Leonardo.Ai	27.  you.com	37.  Chub.ai	47.  Doubao
8.  Poe	18.  Midjourney	28. DeepAI	38.  HIX.AI	48.  Speechify
9.  liner	19. YODAYO	29.  SeaArt AI	39.  Vidnoz	49. <i>NightCafe</i>
10. CIVITAI	20. cutout.pro	30.  invideo AI	40.  PIXELCUT	50.  AI Novelist

TEXT

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 <small>API numbers calculated where reported numbers were missing</small>
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot** (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)
Reading Comprehension	DROP	Reading comprehension (F1 Score)	82.4 Variable shots	80.9 3-shot (reported)
	Hellaswag	Commonsense reasoning for everyday tasks	87.8% 10-shot*	95.3% 10-shot* (reported)
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	94.4% maj1@32	92.0% 5-shot CoT (reported)
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	53.2% 4-shot	52.9% 4-shot (API)
Code	HumanEval	Python code generation	74.4% 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	74.9% 0-shot	73.9% 0-shot (API)



	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
problem-solving <i>MATH</i>	60.1% 0-shot CoT	43.1% 0-shot CoT	33.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math <i>MGS</i>	90.7% 0-shot	83.5% 0-shot	71.8% 0-shot	74.5% 0-shot	71.0% 8-shot	71.0% 8-shot	63.5% 3-shot
Code <i>HumanEval</i>	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text <i>DROP, F1 score</i>	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A <i>ARC-Challenge</i>	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge <i>HellaSwag</i>	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

The Generative AI Infrastructure Stack v1




A work in progress

Compute & Inference


GPU SUPPLY

 CoreWeave

 Lambda

 FOUNDRY

together.ai

 CrusoeCloud


 ARMADA

 aws


 Google Cloud

 Azure

PAAS

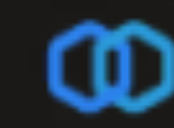
 Replicate

 Modal

 baseten

 BANANA

 RunPod

 Lepton AI

Client Side AI



Community · TensorFlow.js · G

How Adobe used Web ML with TensorFlow.js to enhance Photoshop for web

March 30, 2023



Guest post by [Joseph Hsieh](#) (Principal Scientist, Project Lead at Adobe), [Devin Fernandez](#) (Director of Product Management, Adobe), and [Jason Mayes](#) (Web ML Lead, Google)

Introduction

[Photoshop Web Beta](#) is a browser-based version of the popular desktop image editing software, Adobe Photoshop. This online tool offers a wide range of features and capabilities for editing, enhancing, and manipulating images, all through a web browser.

In this post, we will explore how Adobe plans to bring advanced ML features from desktop to web, such as the [Object Selection tool](#). We will also look at how web-based machine learning in JavaScript can improve the performance and user experience of Photoshop Web Beta, and what we can expect in the future.

- Image Segmentation
- Interactive Segmentati...
- Gesture Recognition
- Hand Landmark Detec...
- Image Embedding
- Face Stylization
- Face Detection
- Face Landmark Detect...
- Pose Landmark Detec...

TEXT

- Text Classification
- Text Embedding
- Language Detection

Face Landmark Detection

Detect the most prominent face from an input image, then estimate 478 3D facial landmarks and 52 facial blendshape scores in real-time. This solution can be used to create a virtual try-on experience or a virtual avatar that mimics a person's facial expressions.

If you only need to detect faces from an input image, check out the lightweight Face Detection solution.

Code examples

[Android](#) | [iOS](#) | [Python](#) | [Raspberry Pi](#) | [Web](#)

The sample parameters below can be changed. See [documentation](#) for more details

Inference delegate: CPU inference

Model selections: MP face landmarker

Display mode: Avatar renderer

Input FaceTime HD Camera (3A...



Inference time (ms): 14

Note: you can switch between Grid and Avatar renderers using the "Display mode" option.



All outputs generated by user provided models.

LLM Inference

Run large language models (LLMs) completely in-browser for text-to-text generation. You can use LLM Inference to perform a wide range of tasks, such as question answering, email drafting, and document summarization.

Download [Gemma 2B](#) (TensorFlow Lite 2b-it-gpu-int4) from Kaggle Models and select the model file in the picker below to get started. For more information on models, see the [documentation](#)

This demo specifically uses [Gemma prompt formatting](#). See the [LLM Inference API documentation](#) for the full web SDK documentation and generic web sample code

Code examples

[Android](#) | [iOS](#) | [Web](#)

Model selections:

Reset conversation

User

Write a blog post about using WebGPU for AI and machine learning

Enter some text...



Built-in AI Early Preview Program

Welcome and about the Prompt API

Authors

Kenji Baheux
Thomas Steiner
Alexandra Klepper

Contact

See [this section](#)

Last-updated

Dec 5, 2024
See [changelog](#).

Latest news 🏆

- Dec 5, 2024 If you frequently encounter the following error, or if it appears with text that shouldn't be problematic, try disabling the Text Safety classifier in Chrome's flags settings:
`chrome://flags/#text-safety-classifier`

The screenshot shows the Chrome DevTools Console with the 'Console' tab selected. The console displays a JavaScript error: 'VM113:3 Uncaught Error: Text Safety classifier is enabled. To disable it, go to chrome://flags/#text-safety-classifier and toggle it off.' The error message is truncated, showing 'Why was the bicycle sad?' and 'Because it was two tired!'. The console also shows the code that triggered the error: `const session = await ai.languageModel.create(); const result = await session.prompt("Tell me a joke"); console.log(result);`. The console interface includes a search bar, a filter dropdown, and a 'Custom levels' dropdown.

```
> const session = await ai.languageModel.create();
const result = await session.prompt("Tell me a joke");
console.log(result);
```

Why was the bicycle sad?

[VM113:3](#)

Because it was two tired!

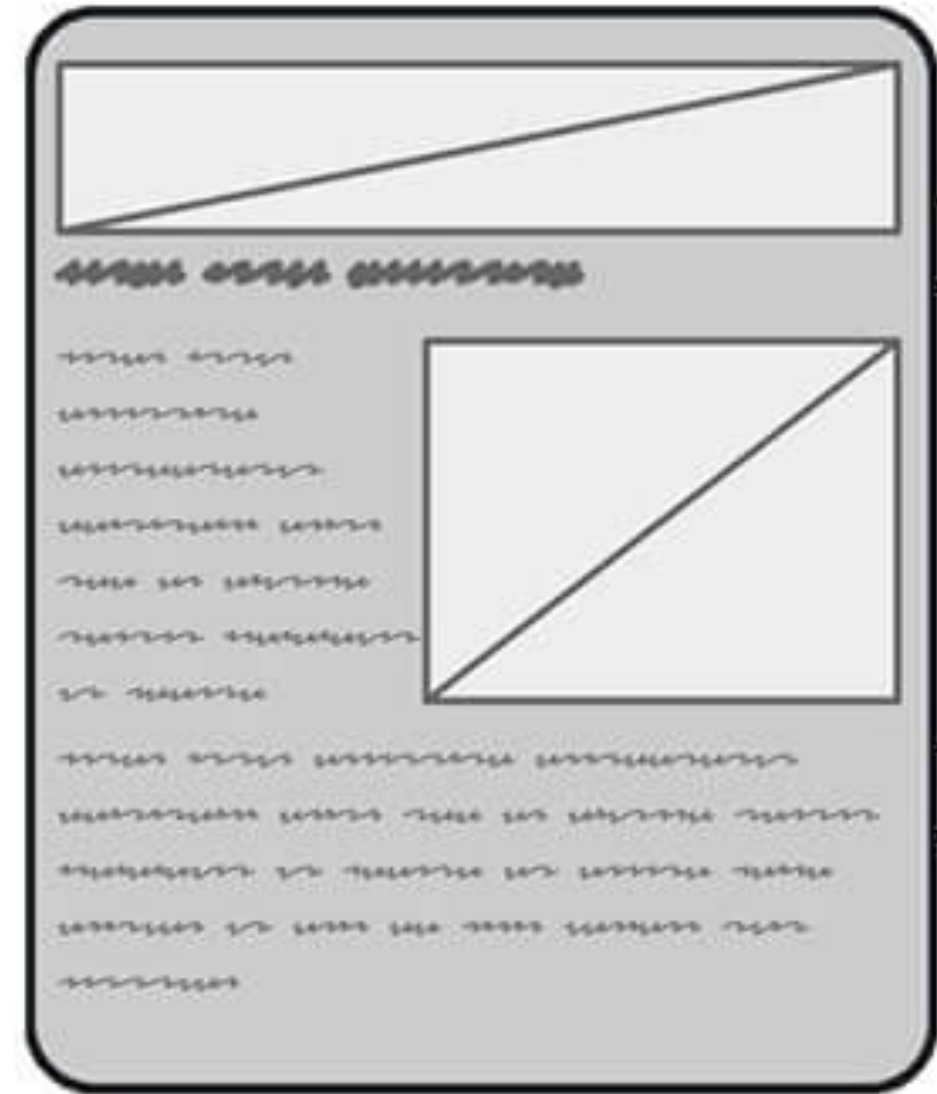
\

... Get it? Two tired because it has two wheels!

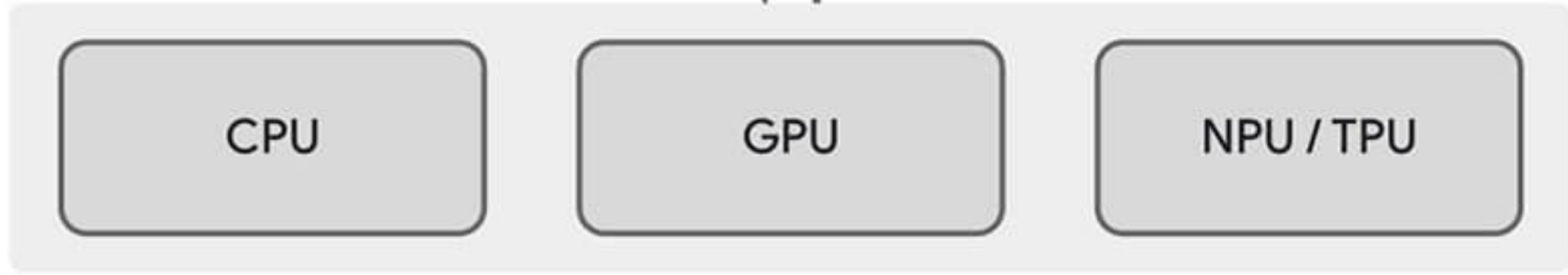
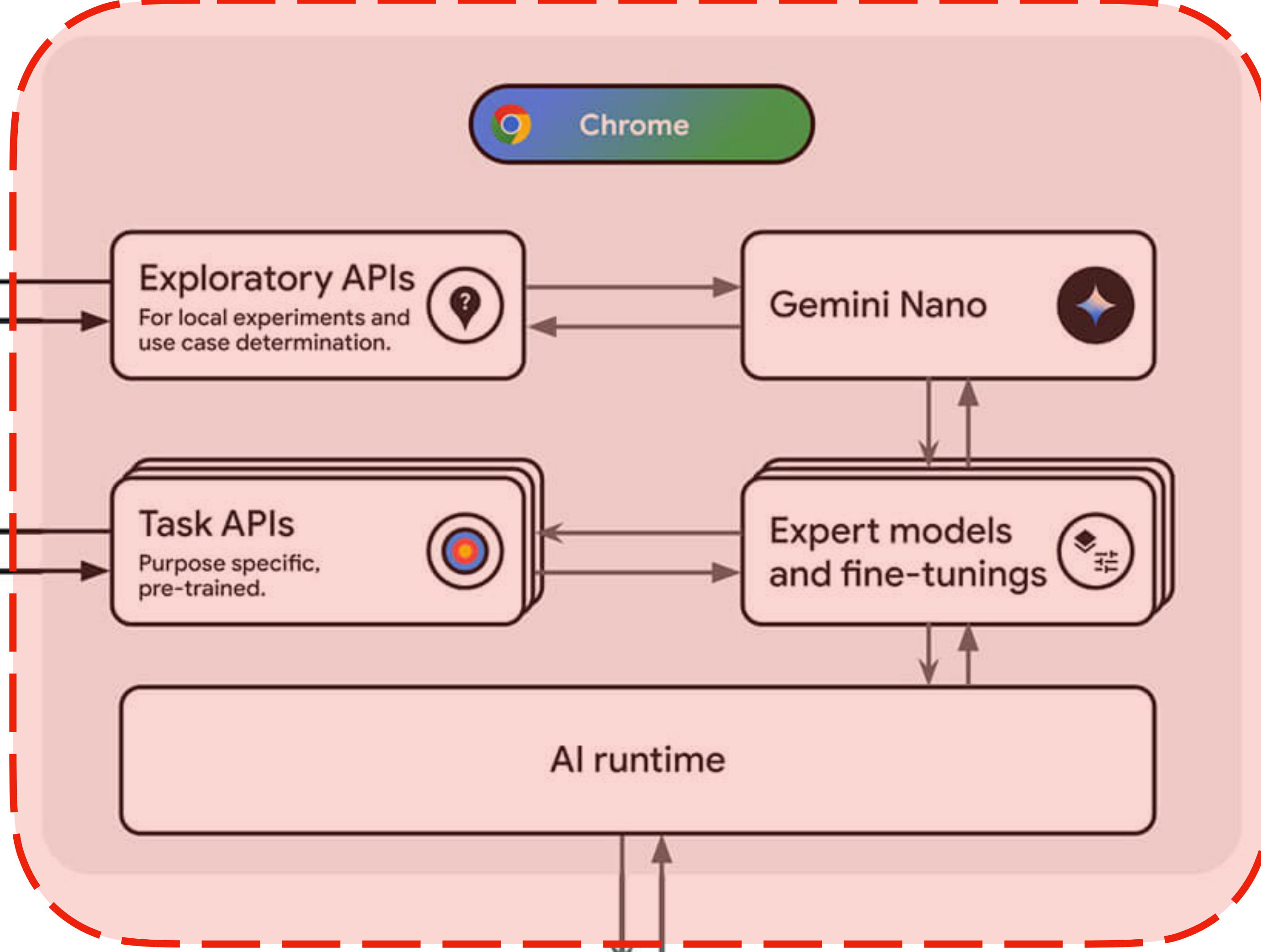
\

...

Your site or web app



Chrome



Chrome Built-in AI



Local AI + Copilot

Surface Pro 11th Edition

Surface Laptop 7th Edition

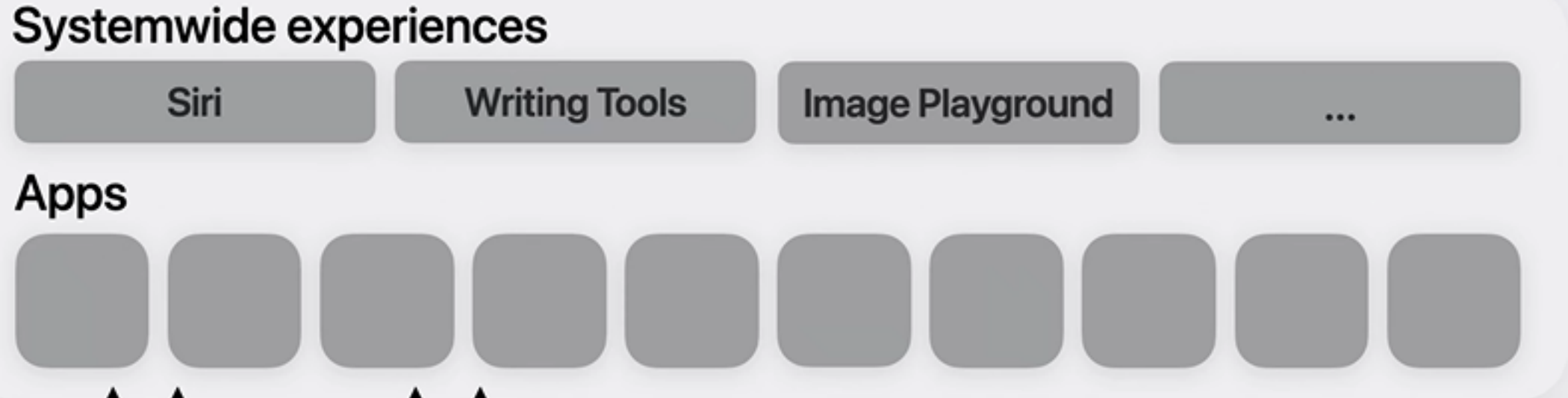
Apple Intelligence

AI for the rest of us.

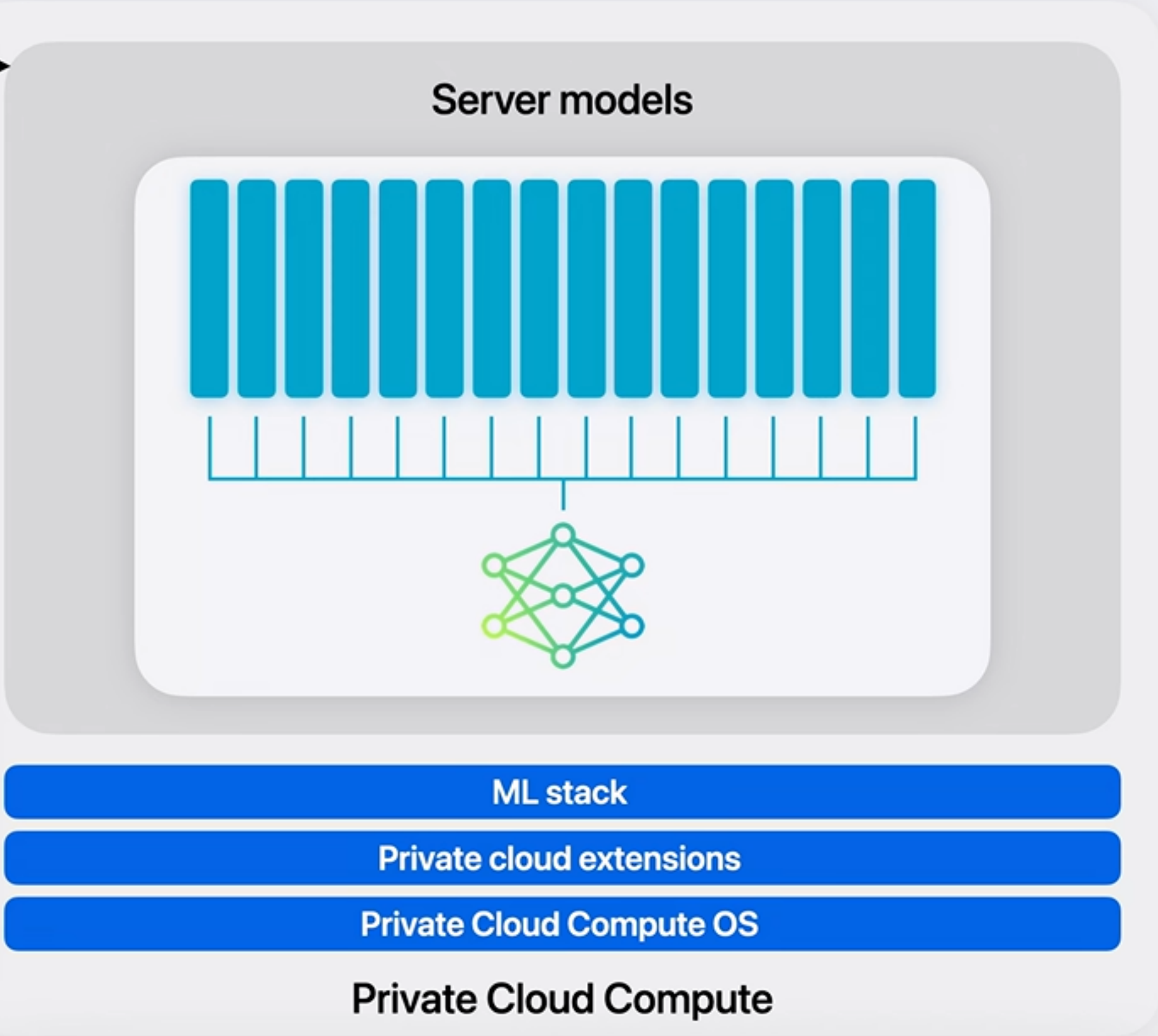
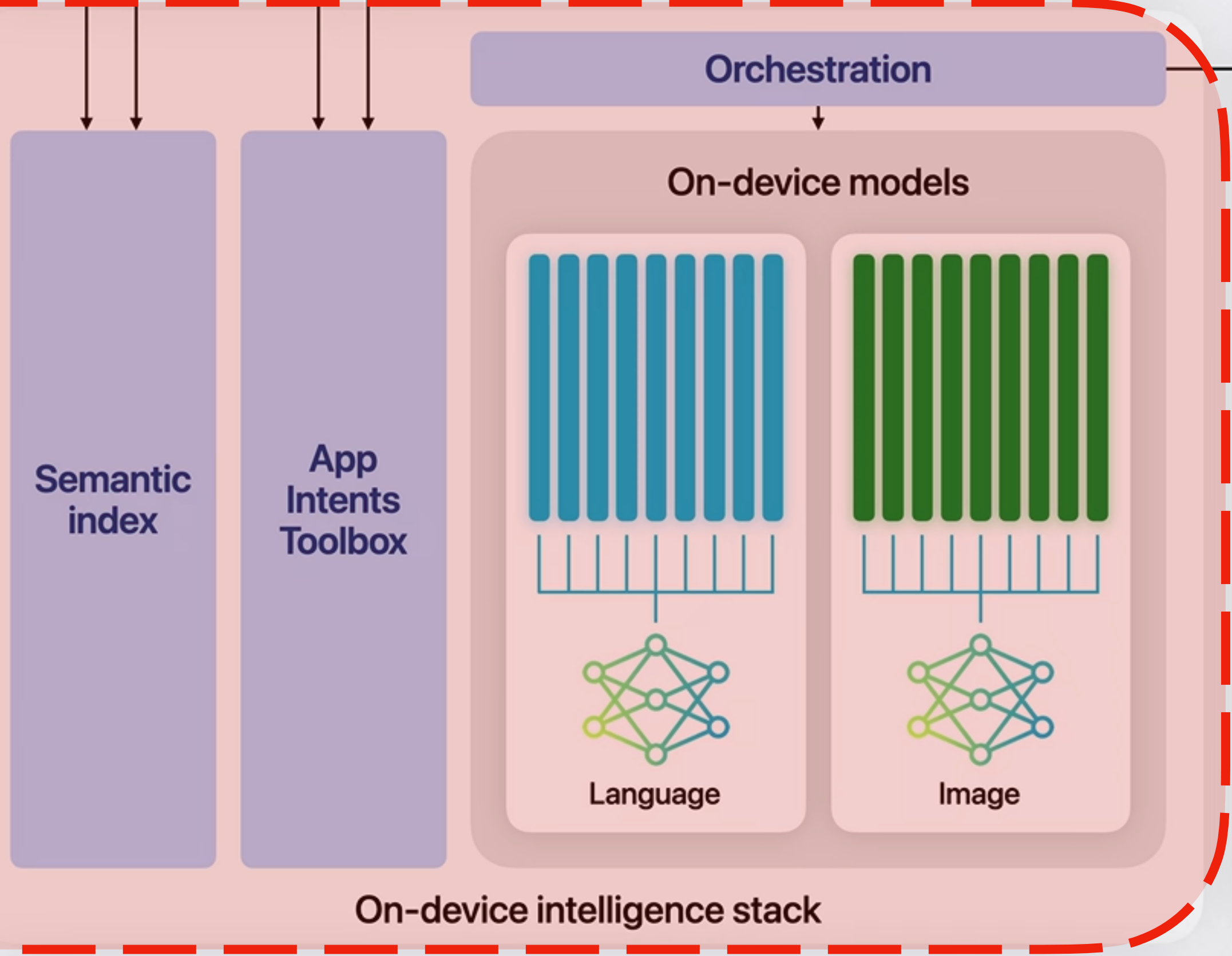
Available now in U.S. English*



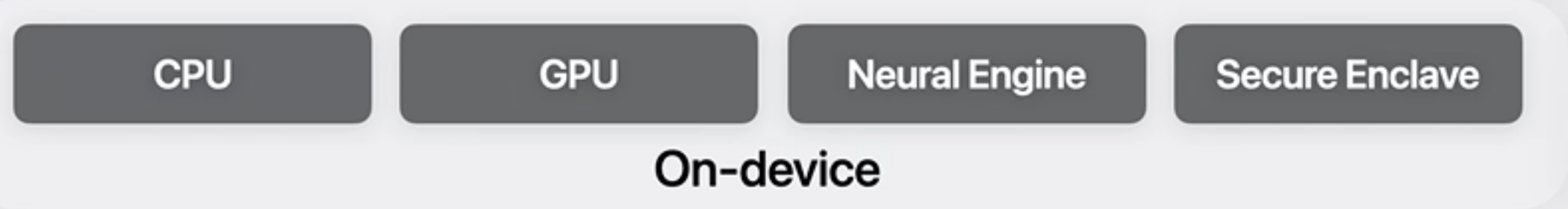
Apps and experiences



Personal Intelligence System



Apple silicon



Pros

Privacy

no data is sent to a server. Users don't have to worry about their data being stored or shared. This is especially important for sensitive data or in countries with strict data protection laws.

Cost

no need to pay for a server. The AI is running in the users' browser. Useful for small projects or prototypes.

Offline

once the model is downloaded, it can run offline. This is useful for PWAs, mobile apps, or as a fallback for an external API not being available.

Cons

Performance

client-side AI is often slower than server-side AI. It also depends on the user's device, which can make the whole experience slow.

Quality

client-side AI is often less accurate than server-side AI, since the models are much smaller to run on the client. If you need high accuracy, this is not the way to go.

Bandwidth

client-side AI requires the model to be downloaded, which can be large and slow depending on the user's connection. The models aren't shared across websites, which is taxing on the user's bandwidth.

Performance

Running inference on device means lower latency, especially for latency sensitive audio or video workloads.

Privacy

Increased privacy, especially in cases where running client side AI, the data cannot be sent to the server.

Price

Running AI inference on the client significantly reduces server side costs

Availability and Stability

Should consider the initial download size of the model and stability when update the application



Digital nomad



1,000+ ideas and notes that travel light

Your AI on-the-go is all you'll need. The power of AI at around the weight of a AA battery, the NotePin goes anywhere work takes you.

Real estate agent



Educator



Engineer



Doctor



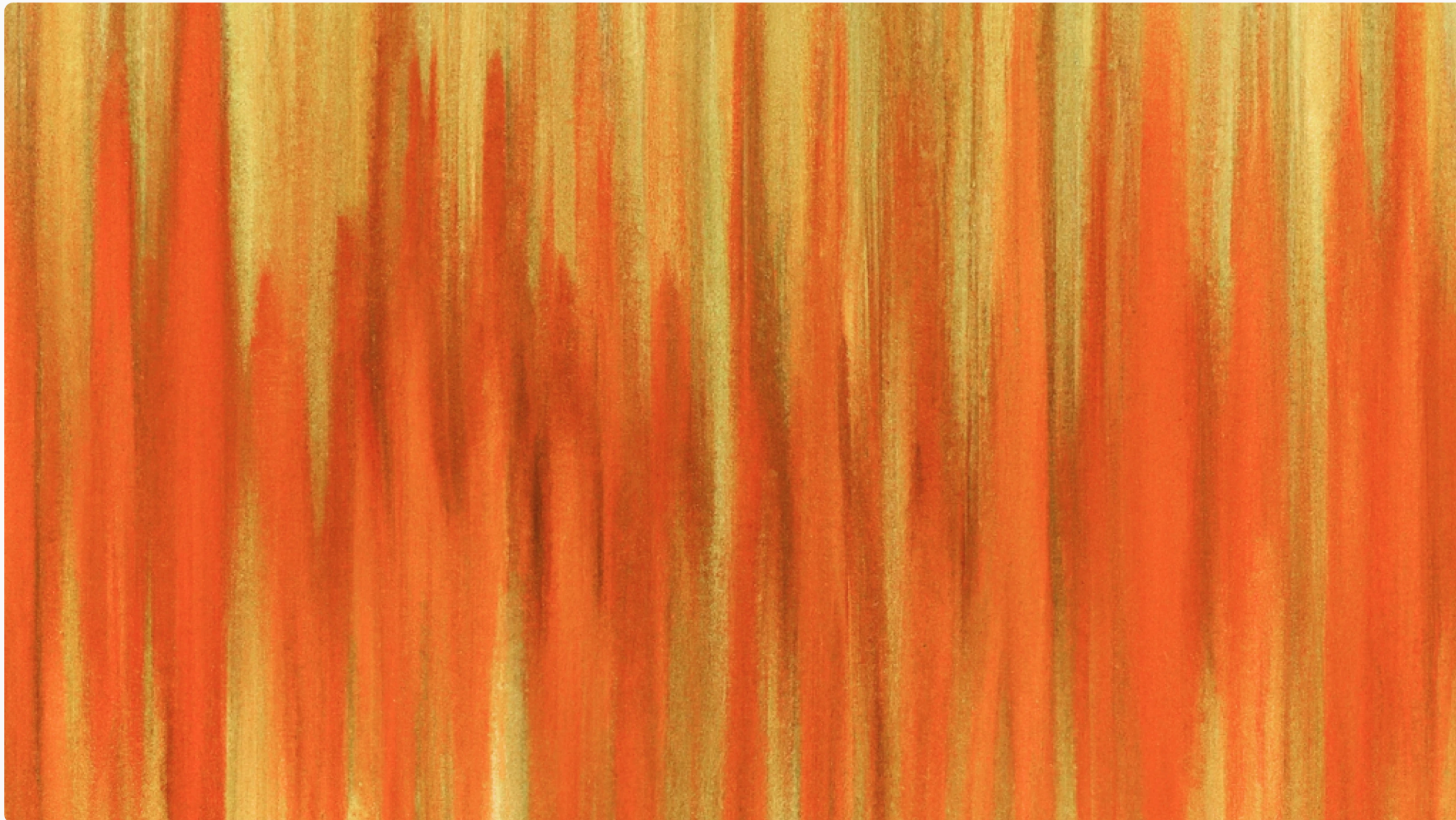
Creative worker



October 1, 2024

Introducing the Realtime API

Developers can now build fast speech-to-speech experiences into their applications



October 1, 2024

Introducing the Realtime API

Developers can now build fast speech-to-speech experiences into their applications



\$0.06 / minute

September 21, 2022

Introducing Whisper

We've trained and are open-sourcing a neural net called Whisper that approaches human level robustness and accuracy on English speech recognition.

[Read paper ↗](#)

[View code ↗](#)

[View model card ↗](#)



September 21, 2022

Introducing Whisper

We've trained and are open-sourcing a neural net called Whisper that approaches human level robustness and accuracy on English speech recognition.

[Read paper ↗](#)

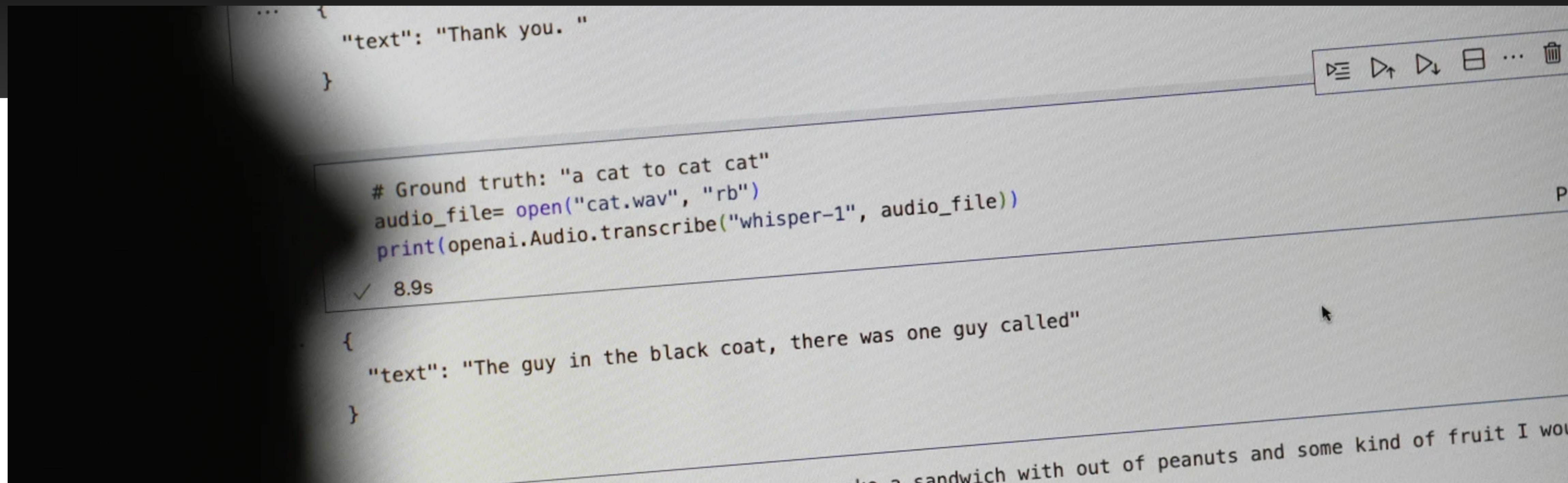
[View code ↗](#)

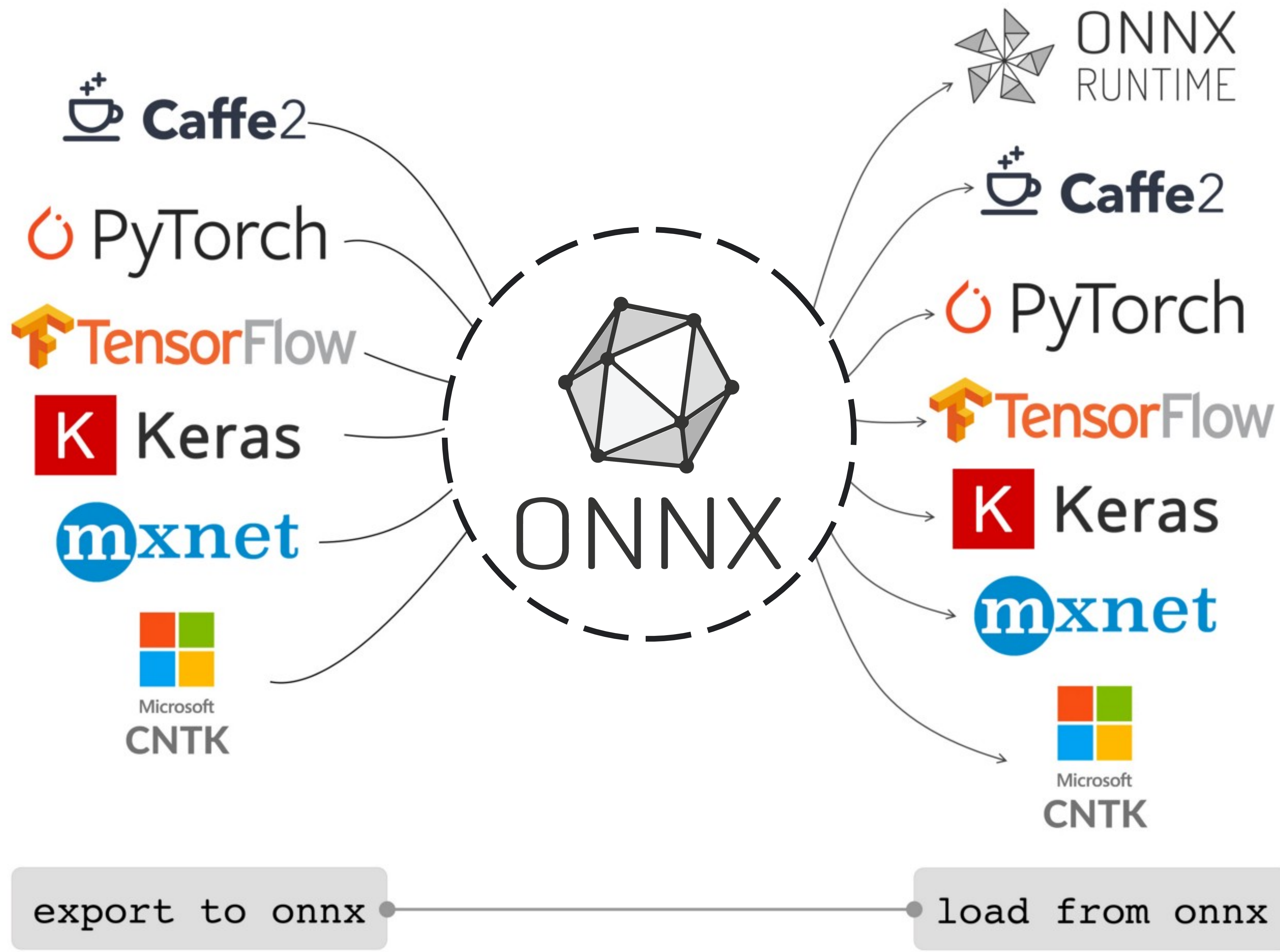
[View model card ↗](#)

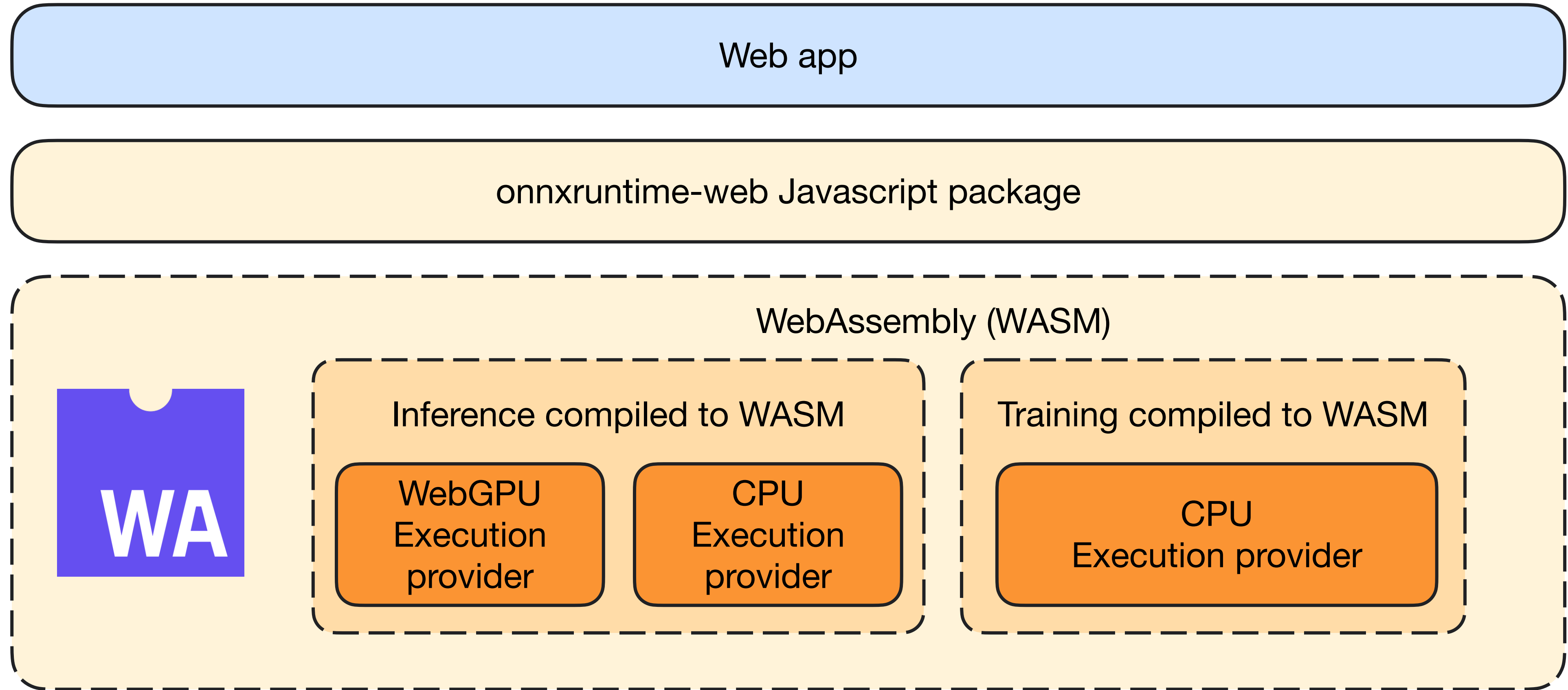
\$0.0006 / minute

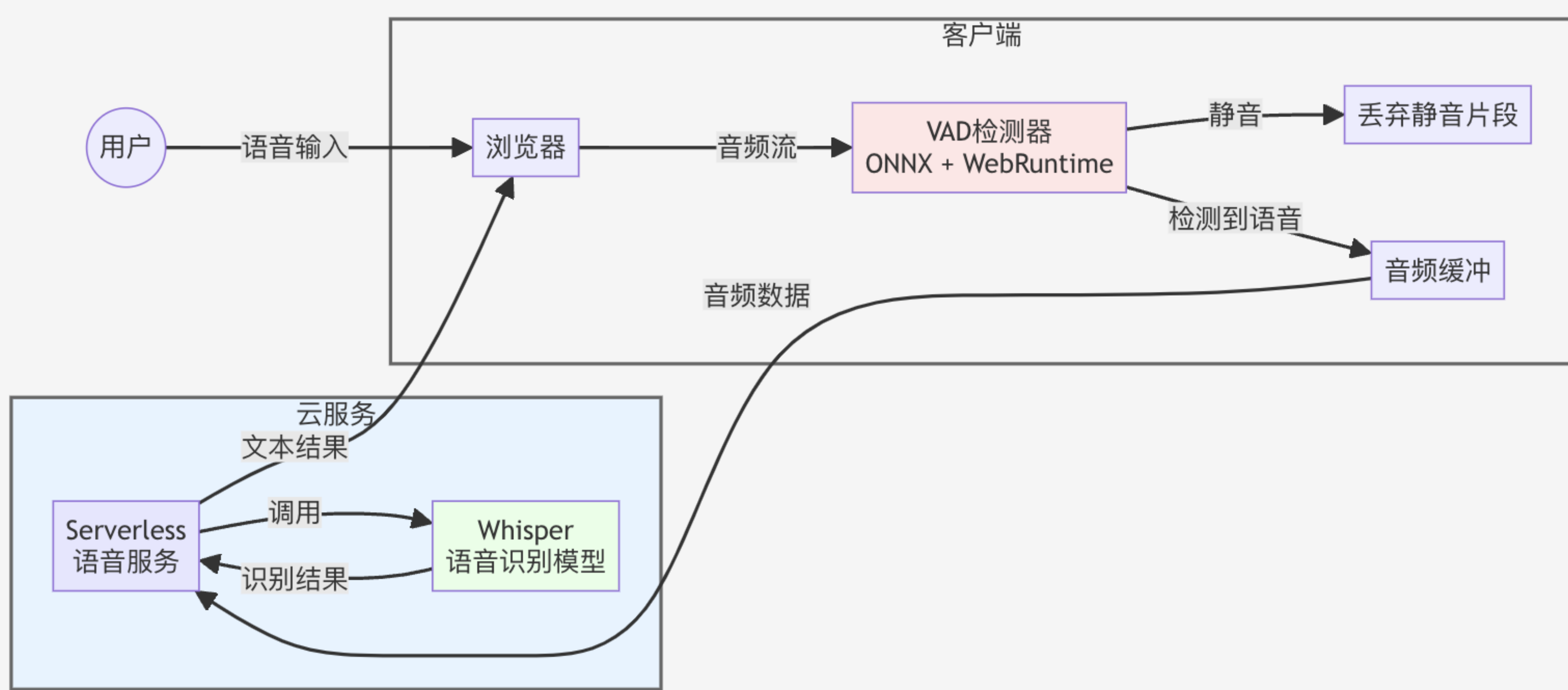
TECHNOLOGY

Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said









● Recording

Here is the demo of client-side AI work with cloud-side AI. Try speak something, the transcription will appear here.

Download VAD

DONE



September 21, 2022

Introducing Whisper

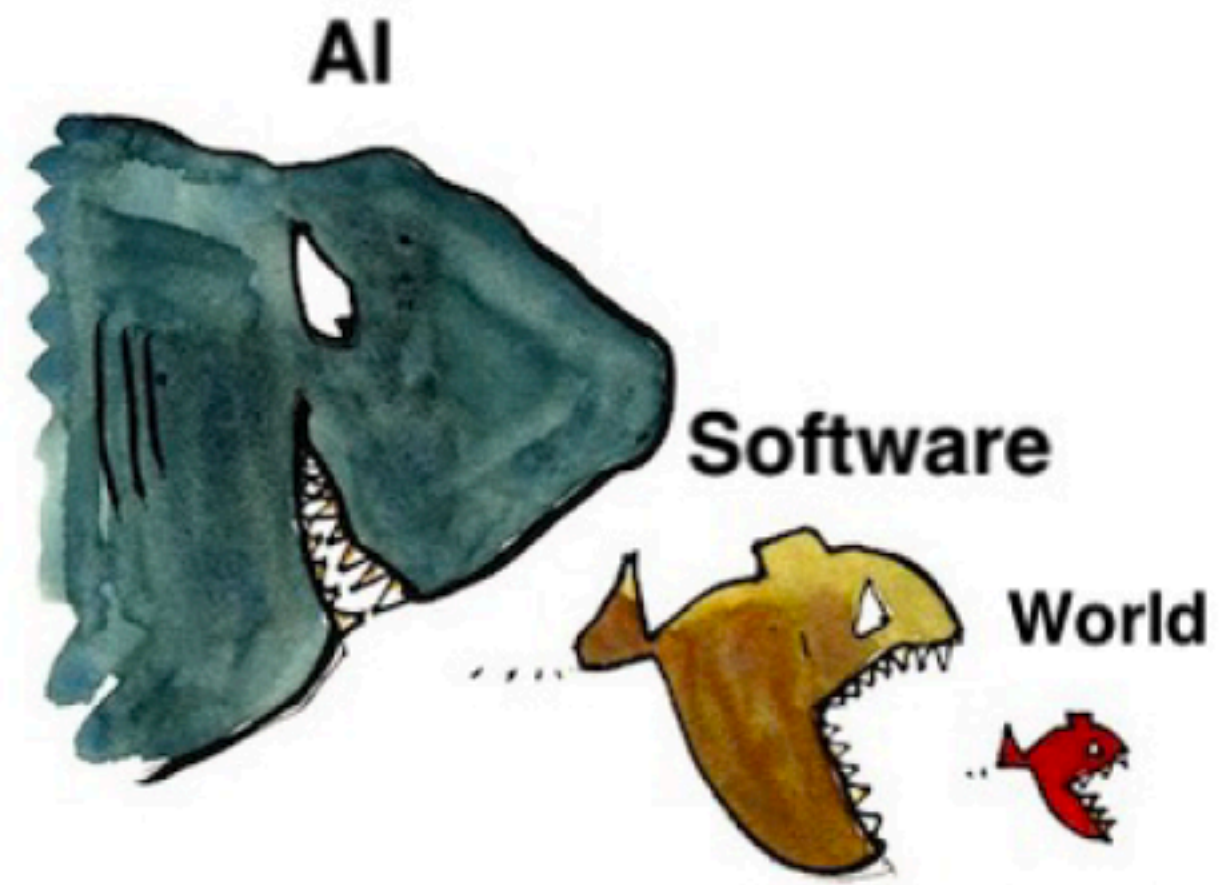
We've trained and are open-sourcing a neural net called Whisper that approaches human level robustness and accuracy on English speech recognition.

[Read paper ↗](#)

[View code ↗](#)

[View model card ↗](#)

\$0 / minute ?



“Software is eating the world”
Andreessen Horowitz, HP (2011)

“Software is eating the world, but AI is going to eat software”
Jensen Huang, Nvidia CEO (2017)

Javascript



AI is eating software, but javascript is going to eat the AI