

AI eats the world

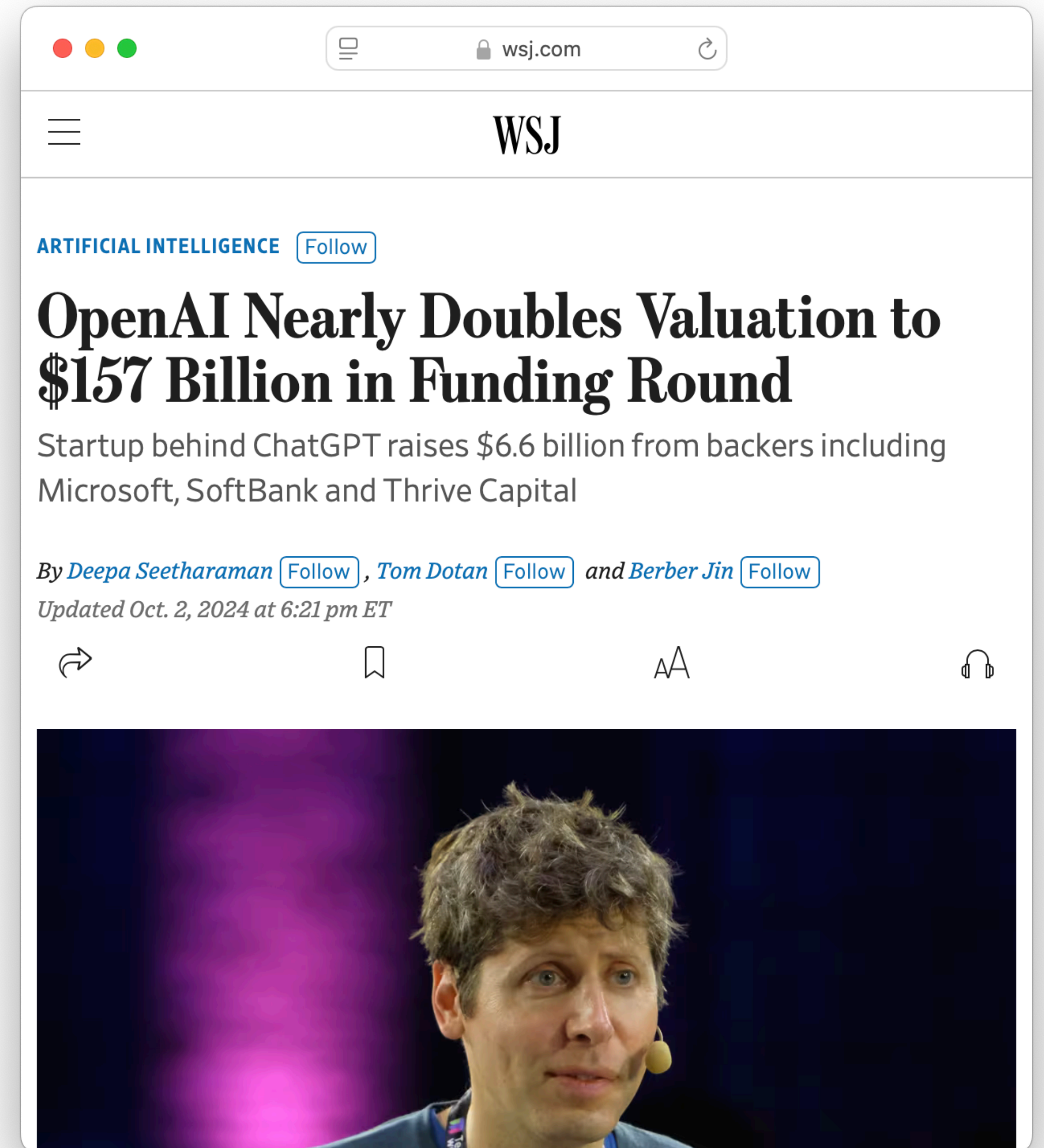
Benedict Evans

November 2024

www.ben-evans.com

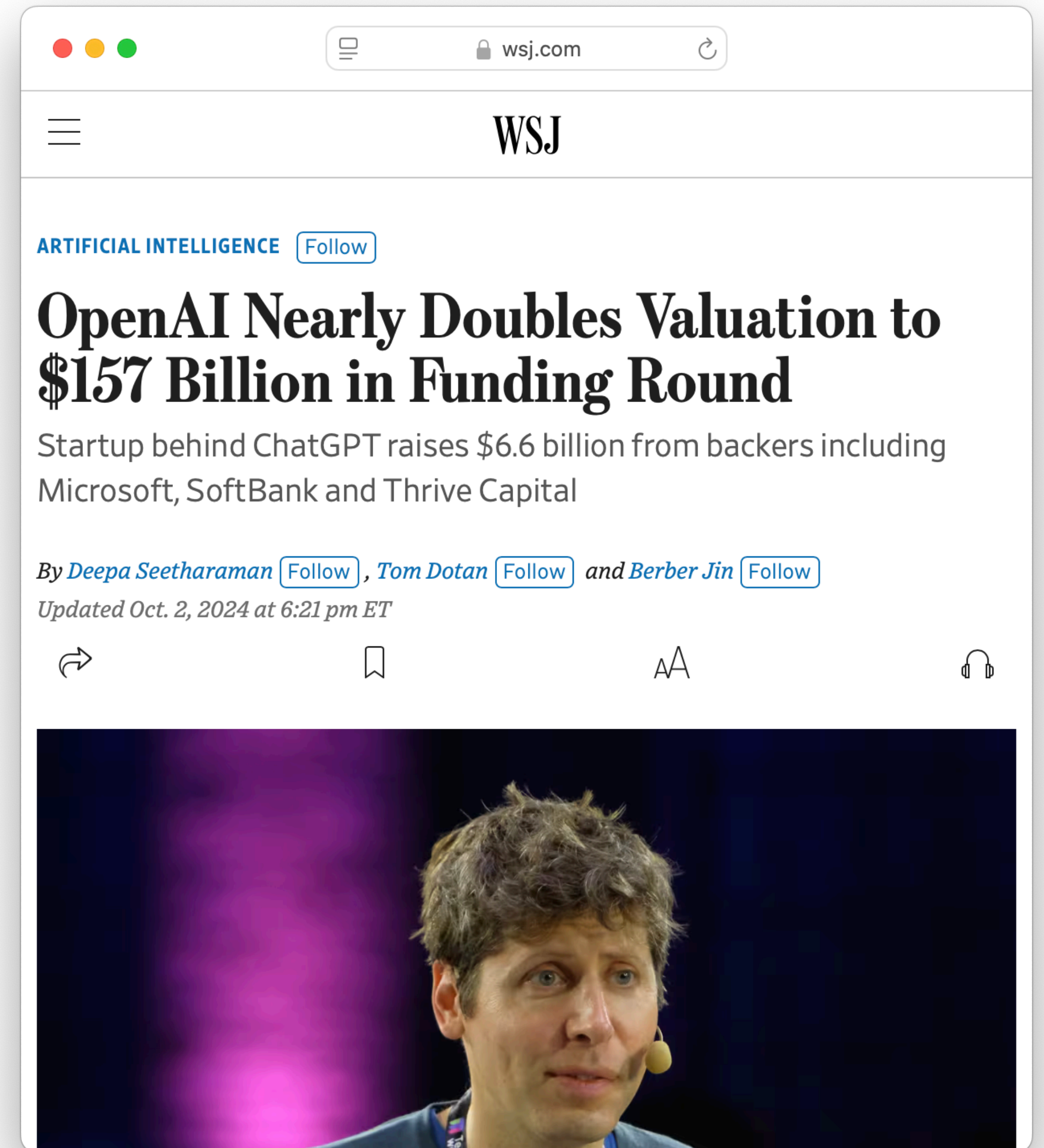
“In my lifetime, I’ve seen two demonstrations of tech that struck me as revolutionary: the GUI and ChatGPT”

Bill Gates, March 2023



(Microsoft took 20 years to reach a \$150bn valuation*)

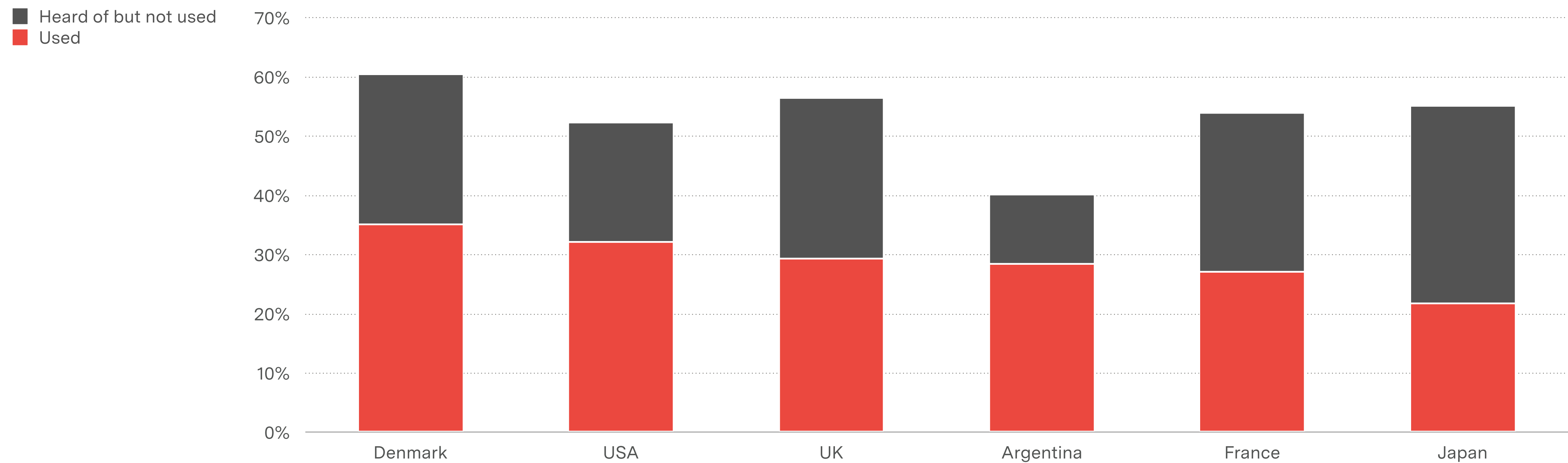
* \$150bn market cap in summer 1996. 2024 dollars.



Huge interest

ChatGPT reached mainstream consciousness with unprecedented speed

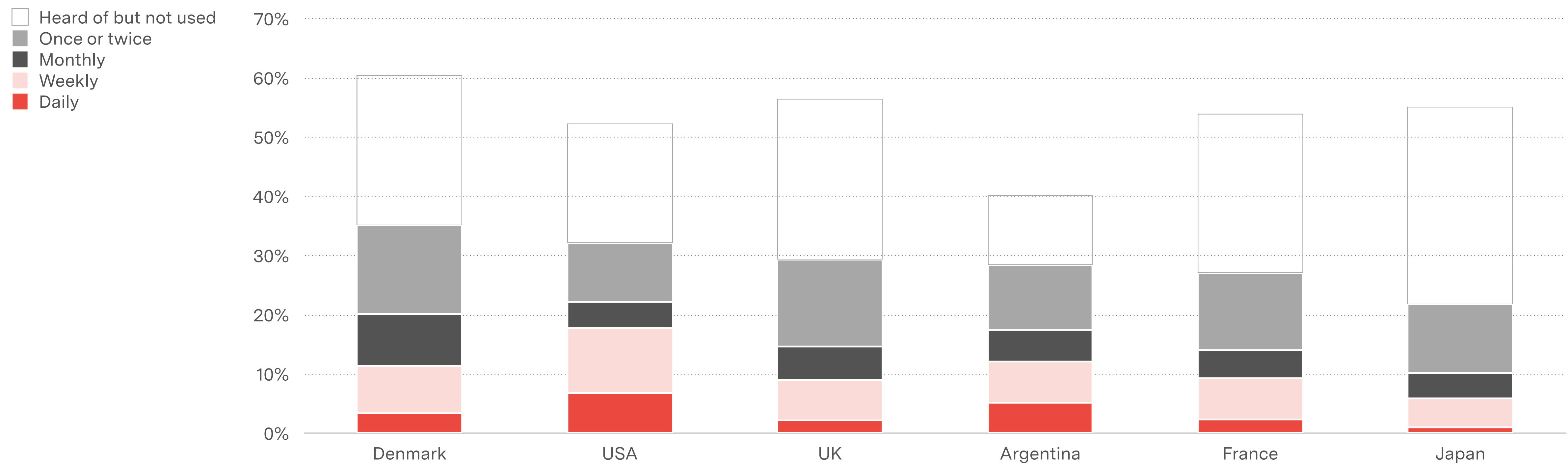
Awareness and use of ChatGPT by country, May 2024



Huge interest, limited use so far

Glass half full / half empty - lots of people have tried it, but few found it useful, so far

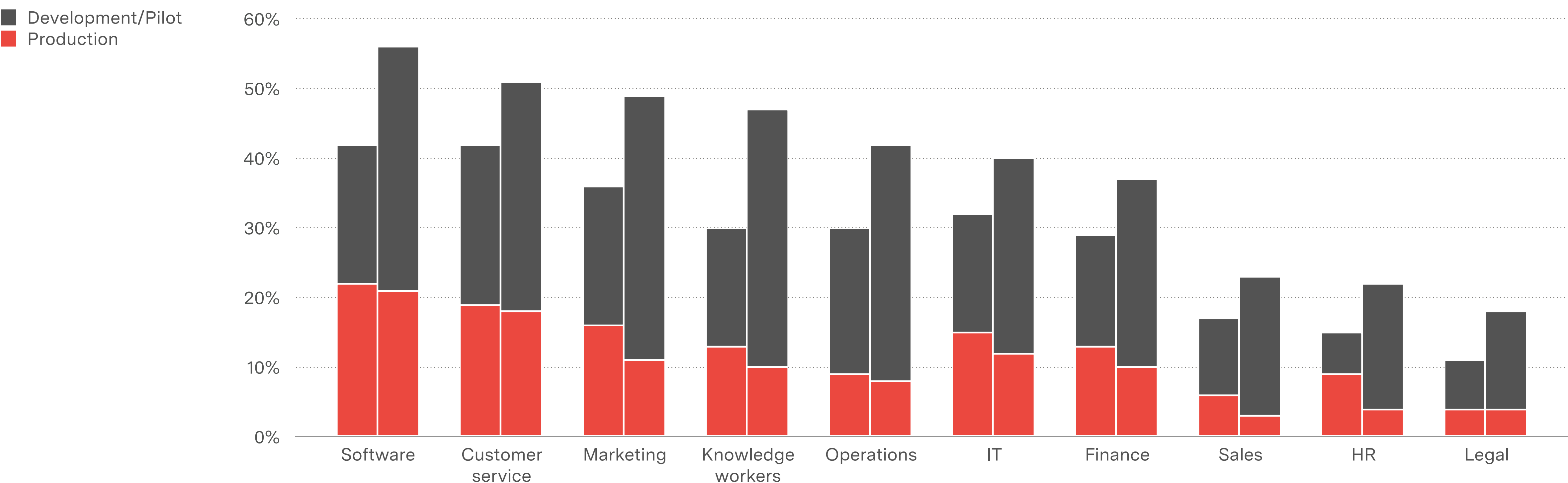
Use of ChatGPT by country, May 2024



Huge interest, limited use so far

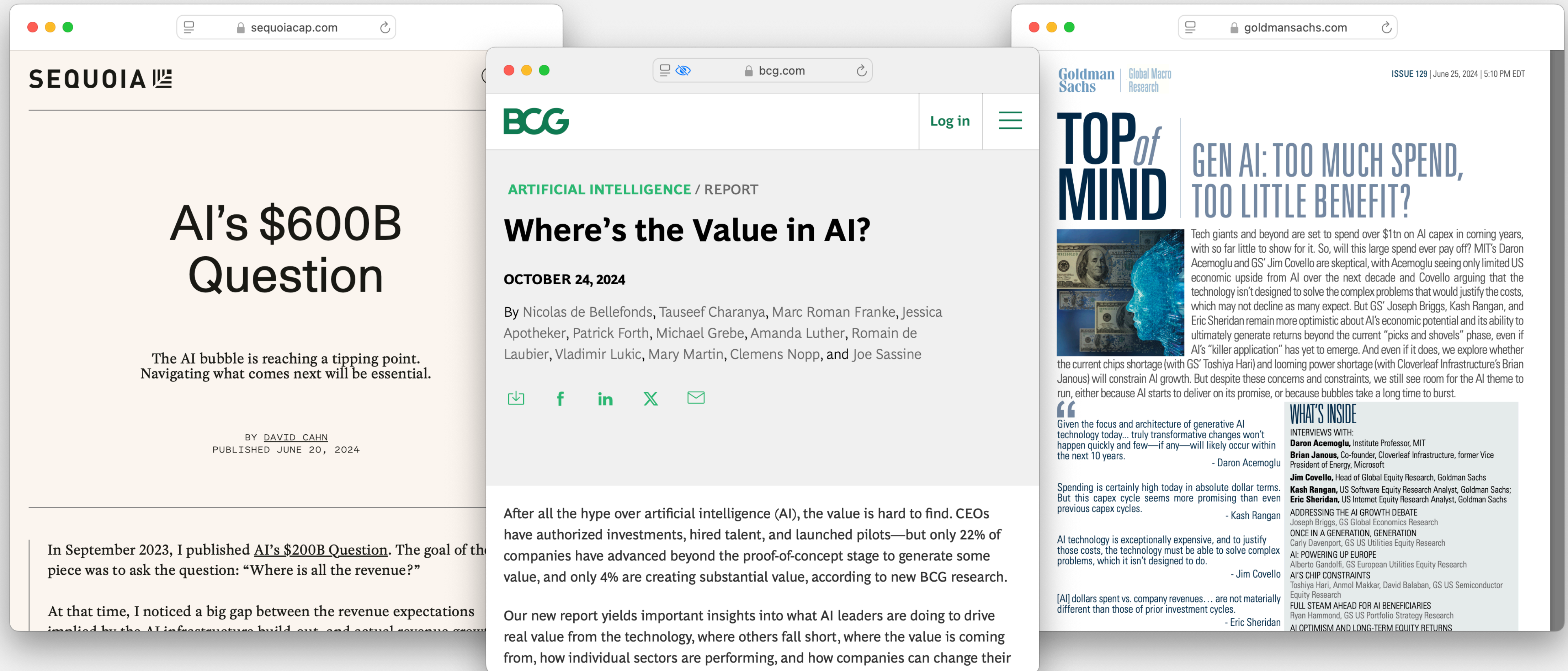
Enterprise software takes time, and come with early disappointments

Enterprise use case adoption rates for generative AI, October 2023 & February 2024



“Wait a minute?”

An investment surge ahead of a proven market prompts the obvious questions



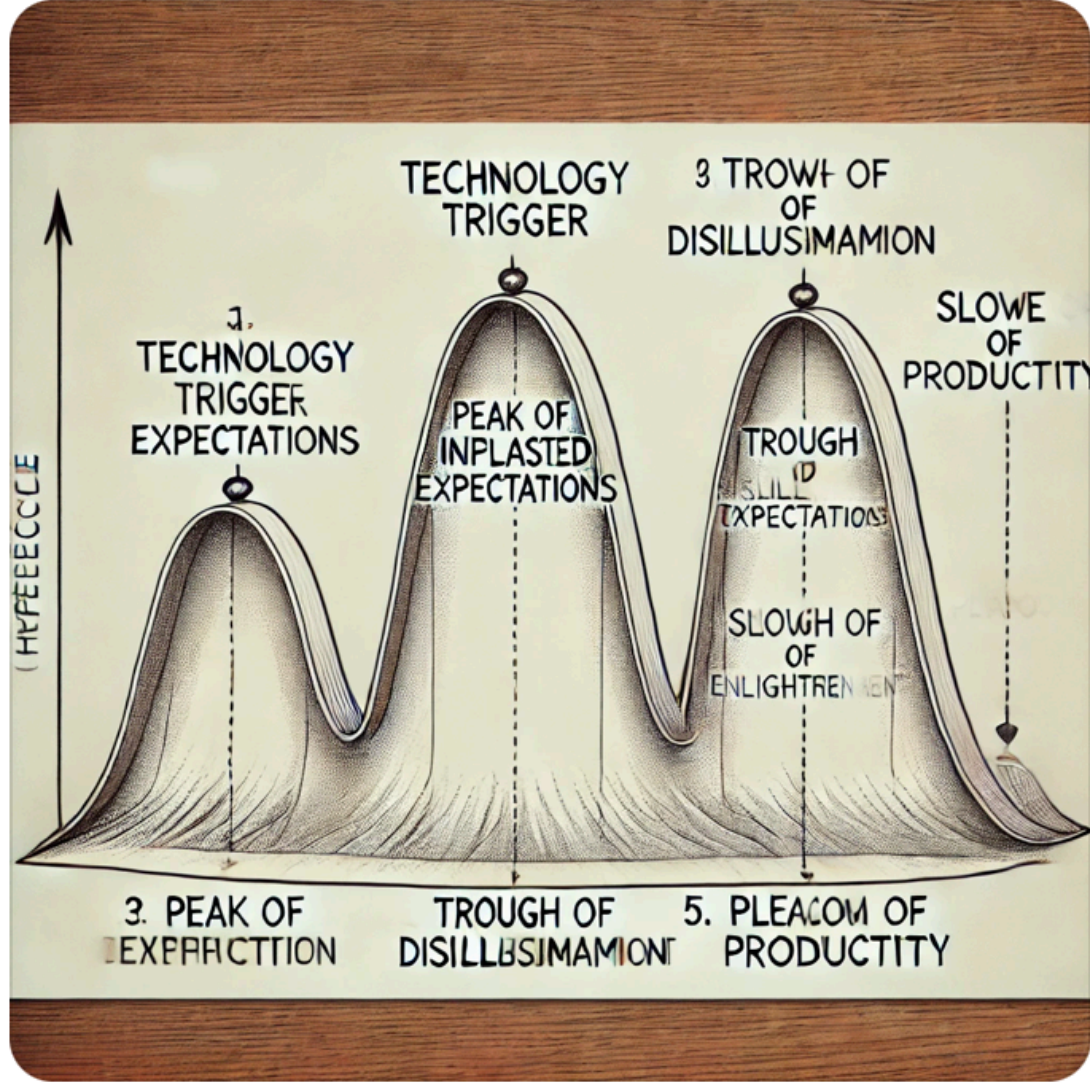
Welcome to the hype cycle?

It always takes time to reach the “Pleacom of Productivity”

chatgpt.com

ChatGPT 4o

Draw a diagram of the Gartner hype cycle



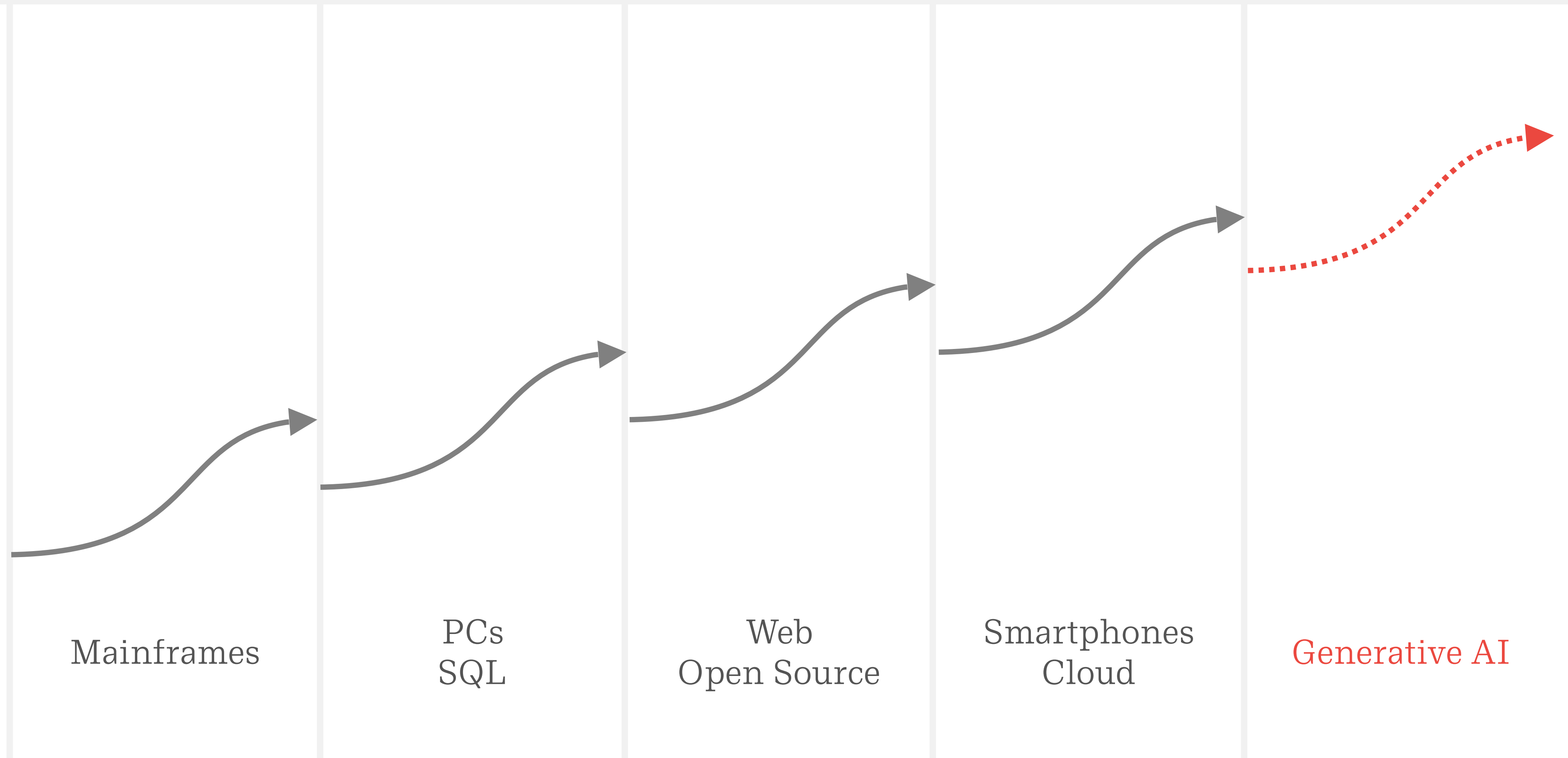
Here is a diagram of the Gartner Hype Cycle with the key phases labeled. Let me know if you'd like any adjustments!

Message ChatGPT

ChatGPT can make mistakes. Check important info.

Beyond the noise - the next platform shift

After the web and smartphones, all tech gets built around generative AI



But everything is wide open

We don't know the answers - we're still working out the questions



How far will
this scale?

How is this
useful?

How do we
deploy this?

How far will this scale?

Foundational question: will LLMs keep scaling?

We got these results by using more and more data and compute - will that keep working?



“Just give yourself the option that what’s been happening for six years now is going to continue”

Kevin Scott, Microsoft CTO

“I don’t know that I would look at the training trends and extrapolate three orders of magnitude ahead blindly from today”

Sergey Brin

Is it slowing down right now?

A sudden blip, or something more?

Reuters

My News

OpenAI and others seek new path to smarter AI as current methods hit limitations

By Krystal Hu and Anna Tong


November 11, 2024 6:30 PM EST · Updated 2 days ago

🔖

Aa

🔗

Sponsored by



Sponsors are not involved in the creation of this or any other Reuters news articles

Summary

Companies

- AI companies face delays and challenges with training new large language models
- Some researchers are focusing on more time for inference in new models
- Shift could impact AI arms race for resources like chips and energy

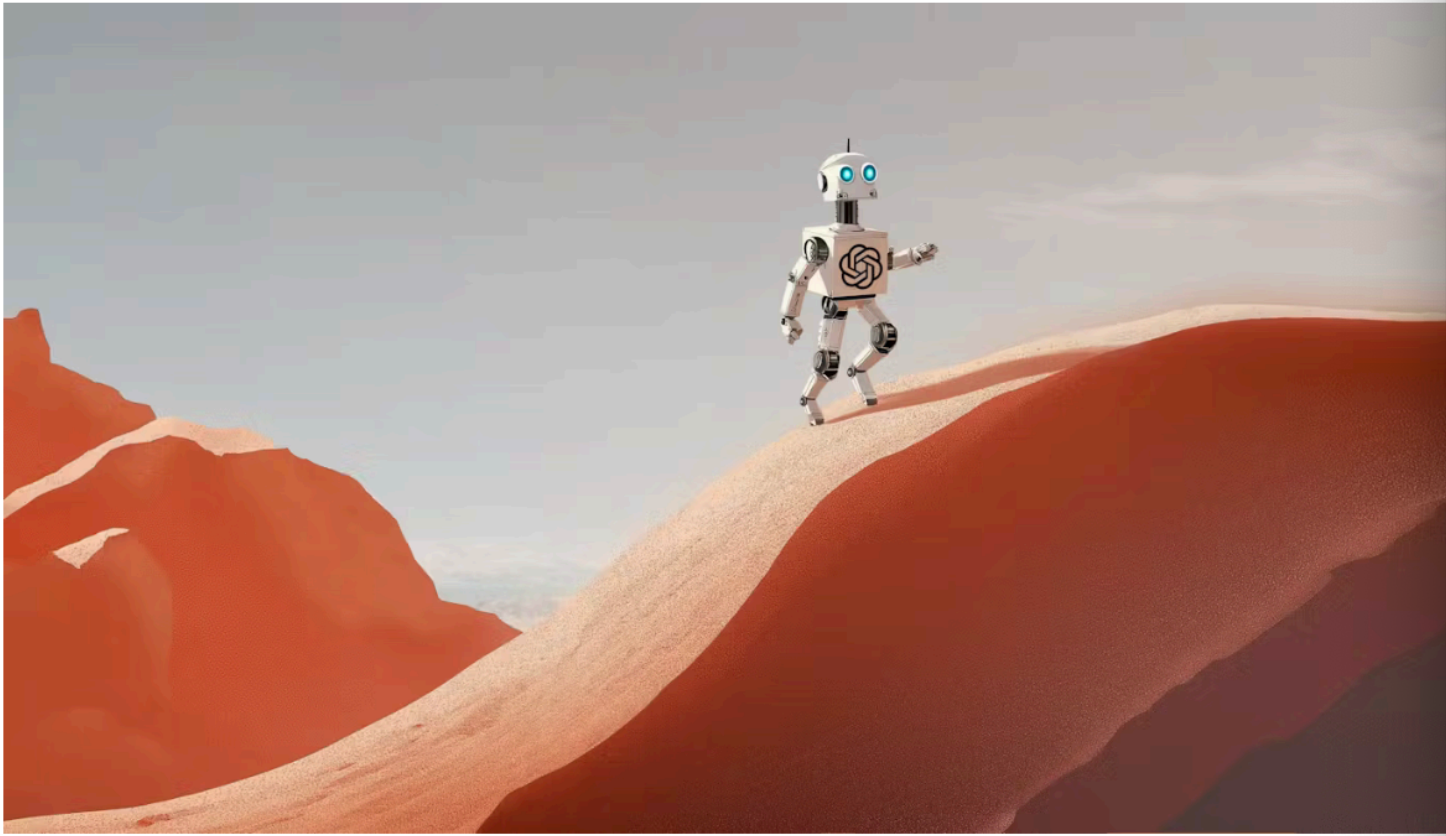
Nov 11 (Reuters) - Artificial intelligence companies like OpenAI are seeking to overcome unexpected delays and challenges in the pursuit of ever-bigger large language models by developing training techniques that use more human-like ways for algorithms to "think".

A dozen AI scientists, researchers and investors told Reuters they believe that these

The Information

Exclusive

OpenAI Shifts Strategy as Rate of ‘GPT’ AI Improvements Slows



Art by Clark Mill

By Stephanie Palazzolo, Erin Woo and Amir Efrati

Share

Bloomberg

The AI Race:

Startups to Watch

AI Glossary

Model Collapse Risk

AI's Real Carbon Fo

Technology

OpenAI, Google and Anthropic Are Struggling to Build More Advanced AI

Three of the leading artificial intelligence companies are seeing diminishing returns from their costly efforts to develop newer models.

By [Rachel Metz](#), [Shirin Ghaffary](#), [Dina Bass](#), and [Julia Love](#)

13 November 2024 at 05:00 GMT-5

🔖 Save

🌐 Translate

OpenAI was on the cusp of a milestone. The startup finished an initial round of training in September for a massive new artificial intelligence model that it hoped would significantly surpass prior versions of the technology behind ChatGPT and move closer to its goal of powerful AI that outperforms humans.

But the model, known internally as Orion, did not hit the company’s desired performance, according to two people familiar with the matter, who spoke on condition of anonymity to discuss company matters. As of late summer, for example, Orion fell short when trying to answer coding

Scaling is hard

Scaling these models has practical challenges and will take time, even before the science questions



Lead-times
for GPUs
and power

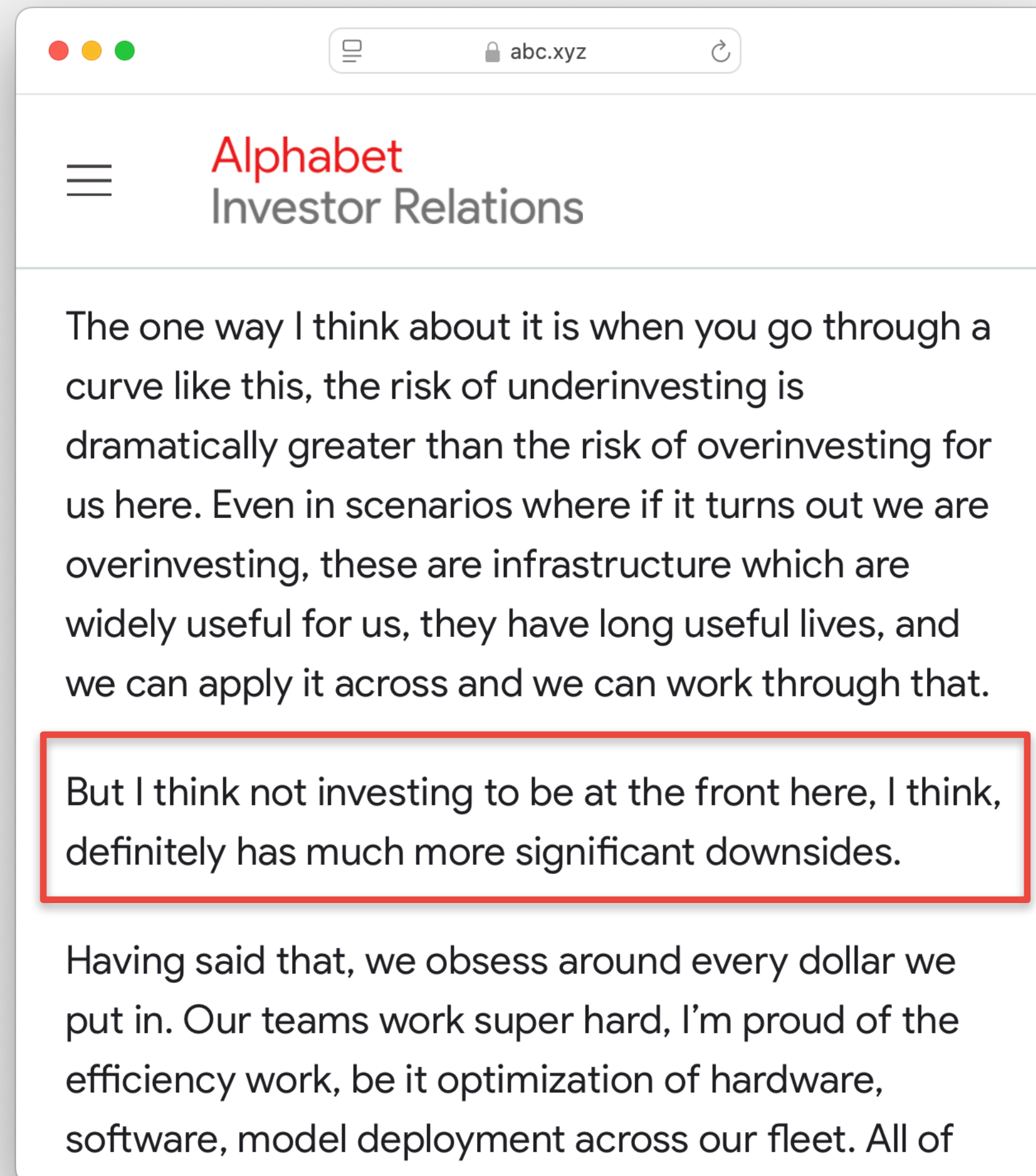
How much
more training
data is there?

Execution &
engineering

And will the
results be
better?

We're going to find out - if only for FOMO

An asymmetric bet - over-spending capex has less downside than losing the next platform?



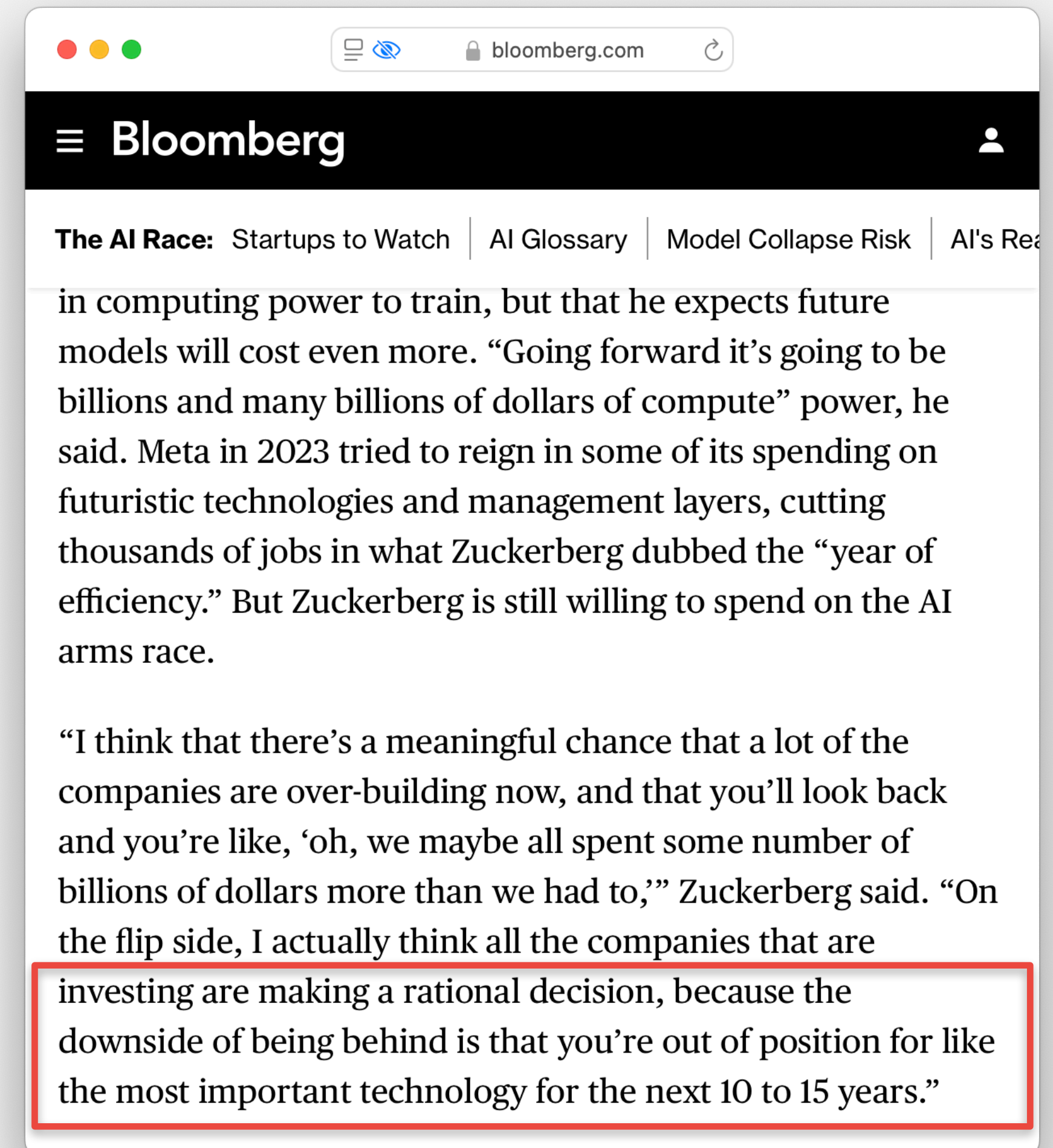
A screenshot of a web browser showing the Alphabet Investor Relations page. The browser's address bar displays 'abc.xyz'. The page header includes the Alphabet logo and 'Investor Relations'. The main text discusses the risk of underinvesting versus overinvesting in infrastructure. A specific quote is highlighted with a red rectangular border.

Alphabet
Investor Relations

The one way I think about it is when you go through a curve like this, the risk of underinvesting is dramatically greater than the risk of overinvesting for us here. Even in scenarios where if it turns out we are overinvesting, these are infrastructure which are widely useful for us, they have long useful lives, and we can apply it across and we can work through that.

But I think not investing to be at the front here, I think, definitely has much more significant downsides.

Having said that, we obsess around every dollar we put in. Our teams work super hard, I'm proud of the efficiency work, be it optimization of hardware, software, model deployment across our fleet. All of



A screenshot of a Bloomberg article titled 'The AI Race'. The browser's address bar shows 'bloomberg.com'. The article discusses the AI arms race, mentioning Meta's spending and Zuckerberg's perspective. A quote about the downside of being behind is highlighted with a red rectangular border.

Bloomberg

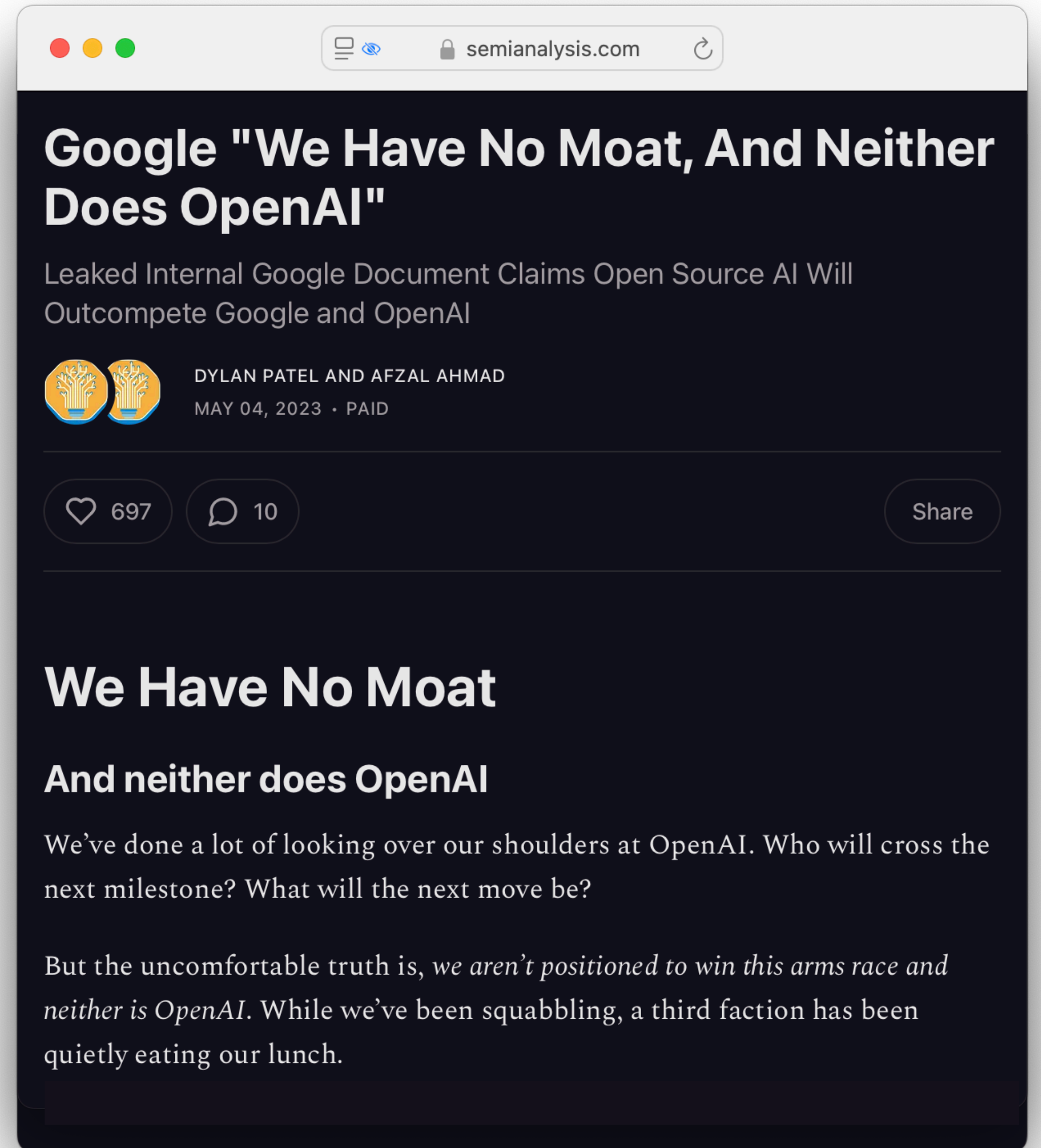
The AI Race: Startups to Watch | AI Glossary | Model Collapse Risk | AI's Realities

in computing power to train, but that he expects future models will cost even more. "Going forward it's going to be billions and many billions of dollars of compute" power, he said. Meta in 2023 tried to reign in some of its spending on futuristic technologies and management layers, cutting thousands of jobs in what Zuckerberg dubbed the "year of efficiency." But Zuckerberg is still willing to spend on the AI arms race.

"I think that there's a meaningful chance that a lot of the companies are over-building now, and that you'll look back and you're like, 'oh, we maybe all spent some number of billions of dollars more than we had to,'" Zuckerberg said. "On the flip side, I actually think all the companies that are investing are making a rational decision, because the downside of being behind is that you're out of position for like the most important technology for the next 10 to 15 years."

“We have no moat”

Internal Google memo, May 2023...



“The models that are in training now ... are closer in cost to \$1bn... and then I think in 2025 and 2026, we’ll get more towards \$5bn or \$10bn”

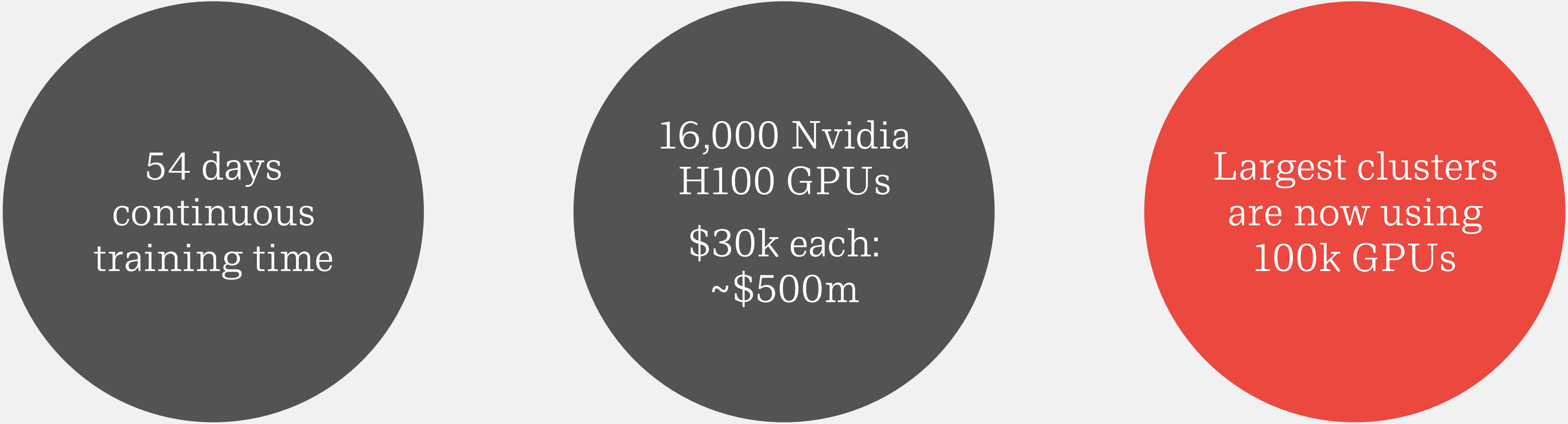
Dario Amodei, Anthropic CEO, April 2024

“The amount of compute needed to train Llama 4 will likely be almost 10x more than what we used to train Llama 3 - and future models will continue to grow beyond that”

Mark Zuckerberg, July 2024

Training Meta's Llama 3.1 SOTA model

Unprecedented computational (and capital) intensity



54 days
continuous
training time

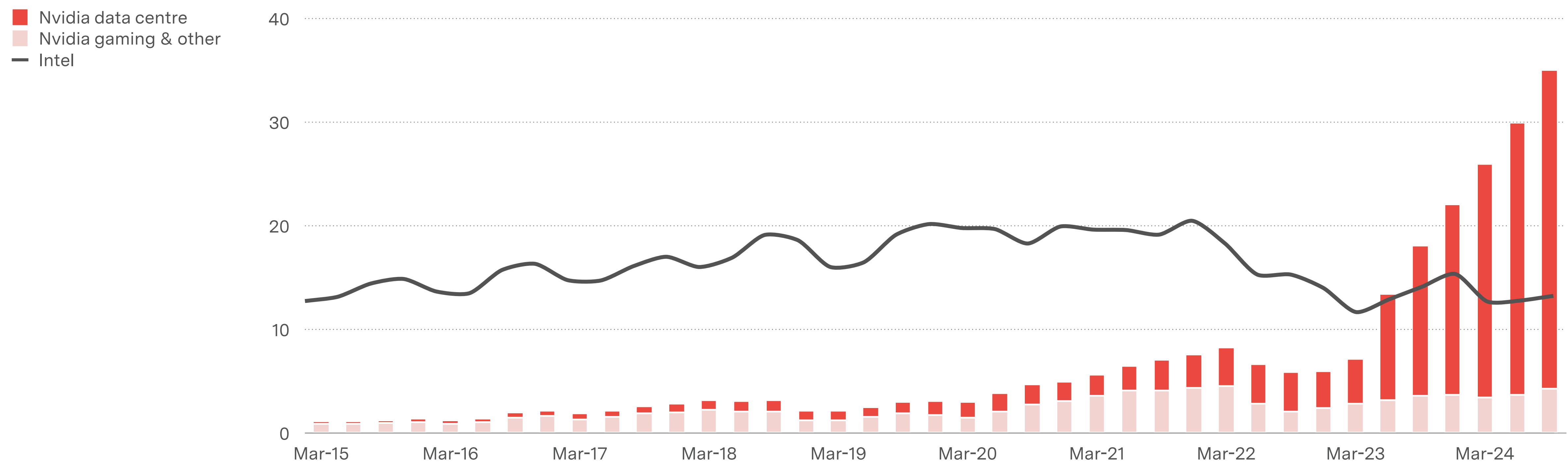
16,000 Nvidia
H100 GPUs
\$30k each:
~\$500m

Largest clusters
are now using
100k GPUs

If the moat is capital...

Nvidia can't keep up with demand - for now (but semiconductors are a cyclical industry)

Quarterly revenue by segment (\$bn)

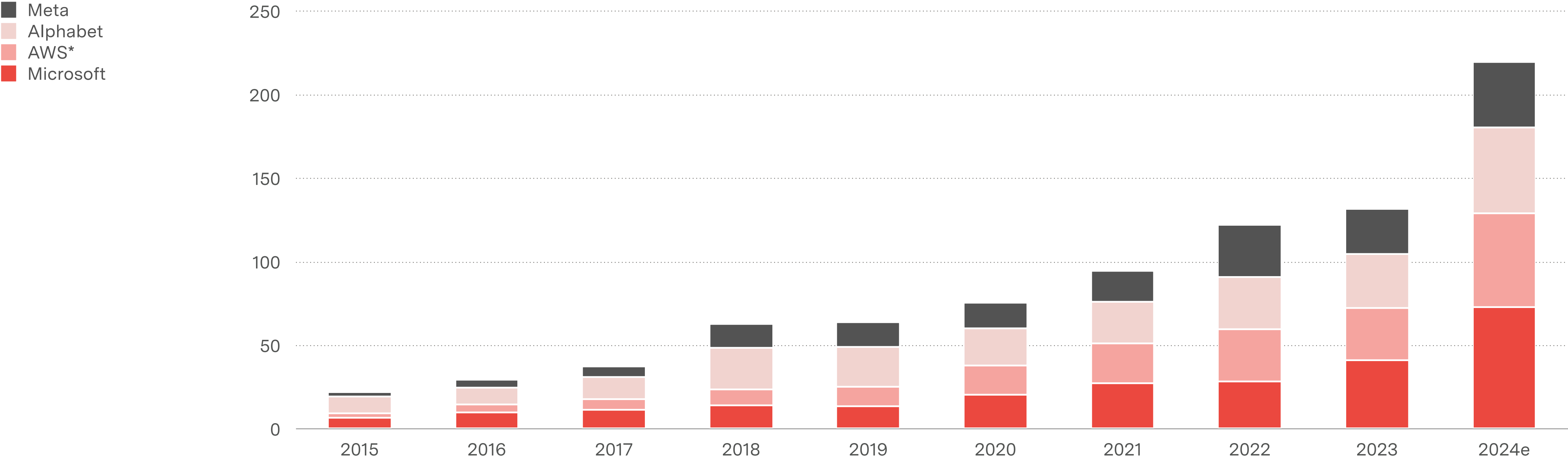


Source: Nvidia, Intel

The capex surge

~\$220bn of capex at the big four in 2024, up \$90bn from 2023, and all expect more growth in 2025

Capex (\$bn)

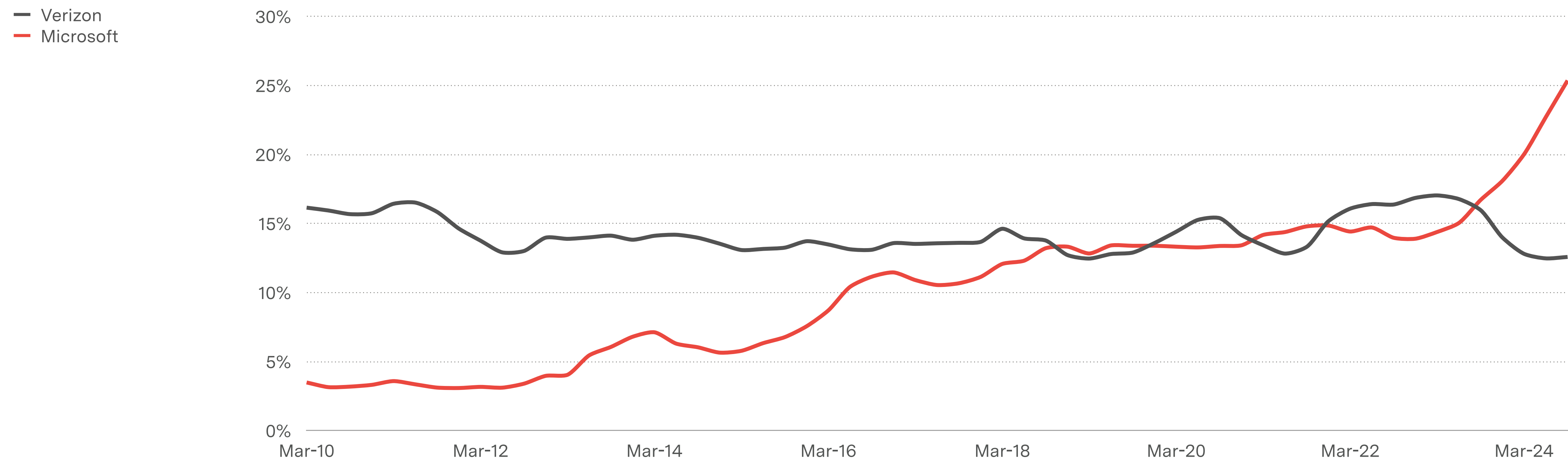


Source: Companies, company guidance. Includes capital leases
* Amazon forecasts \$75bn total capex, 'the majority' for cloud

From the edge to the centre

Remember when telcos built the infrastructure and software had no assets?

Capex/sales (TTM)



Source: Microsoft, Verizon. Includes capital leases

Here come the bankers

A flood of capital creates opportunities for capitalists

media.kkr.com

KKR

BACK TO PRESS RELEASES

KKR And Energy Capital Partners Announce \$50 Billion Strategic Partnership To Support AI Growth Through Investments In Data Centers And Power Generation

October 30, 2024

Strategic partnership with available capital and scale ready to meet the urgent need to fund data center, power, and grid infrastructure in the U.S. and globally

Scaling of AI and cloud infrastructure in the U.S. expected to cost at least \$1 trillion by 2030


NEW YORK & SUMMIT, N.J.-(BUSINESS WIRE)- KKR, a leading global investment firm, and Energy Capital Partners (“ECP”), the largest private owner of power generation and renewables in the

Reader Available

The Information

Exclusive

Andreessen Horowitz Is Building a Stash of More Than 20,000 GPUs to Win AI Deals



Ben Horowitz, left, and Marc Andreessen. Photos by Getty

ft.com

FINANCIAL TIMES

My Account

Artificial intelligence + Add to myFT

Wall Street frenzy creates \$11bn debt market for AI groups buying Nvidia chips

Huge loans for ‘neocloud’ groups raise concern over chipmaker’s dominance of artificial intelligence market

Wall Street’s largest financial institutions have loaned more than \$11bn to a niche group of tech companies based on their possession of the world’s hottest commodity: Nvidia’s artificial intelligence chips.

Blackstone, Pimco, Carlyle and BlackRock are among those that have created a lucrative new debt market over the past year by lending to “neocloud” companies, which provide cloud computing to tech groups building AI products.

Neocloud groups such as CoreWeave, Crusoe and Lambda Labs have acquired tens of thousands of Nvidia’s high-performance computer chips.

And everything is still changing under our feet

All the science and engineering questions are still moving



The last time software had marginal cost

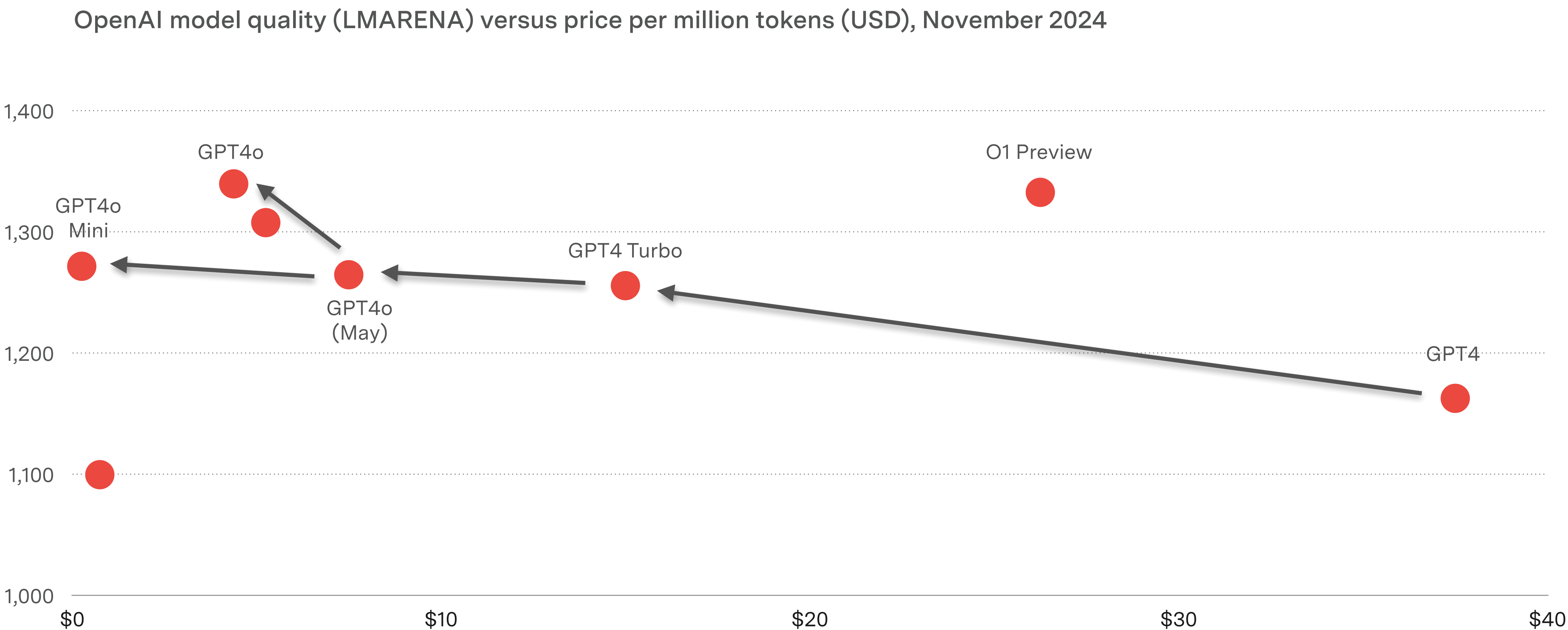
The consumer internet model of 'launch free, go viral, work out revenue later' doesn't work with today's LLM cost model

Source: IBM



Huge efficiency gains

Many more models, many more specs to measure, and all converging on a commodity



Source: Companies, LMARENA

The ‘feeds & speeds’ phase of the market

Many more models, many more specs to measure, and all converging on a commodity

Model quality (LMARENA) versus price per million tokens (USD), November 2024



Source: Companies, LMARENA

Better or cheaper - plus open source

Best, or 90% as good at 5% of the price

Model quality (LMARENA) versus price per million tokens (USD), November 2024

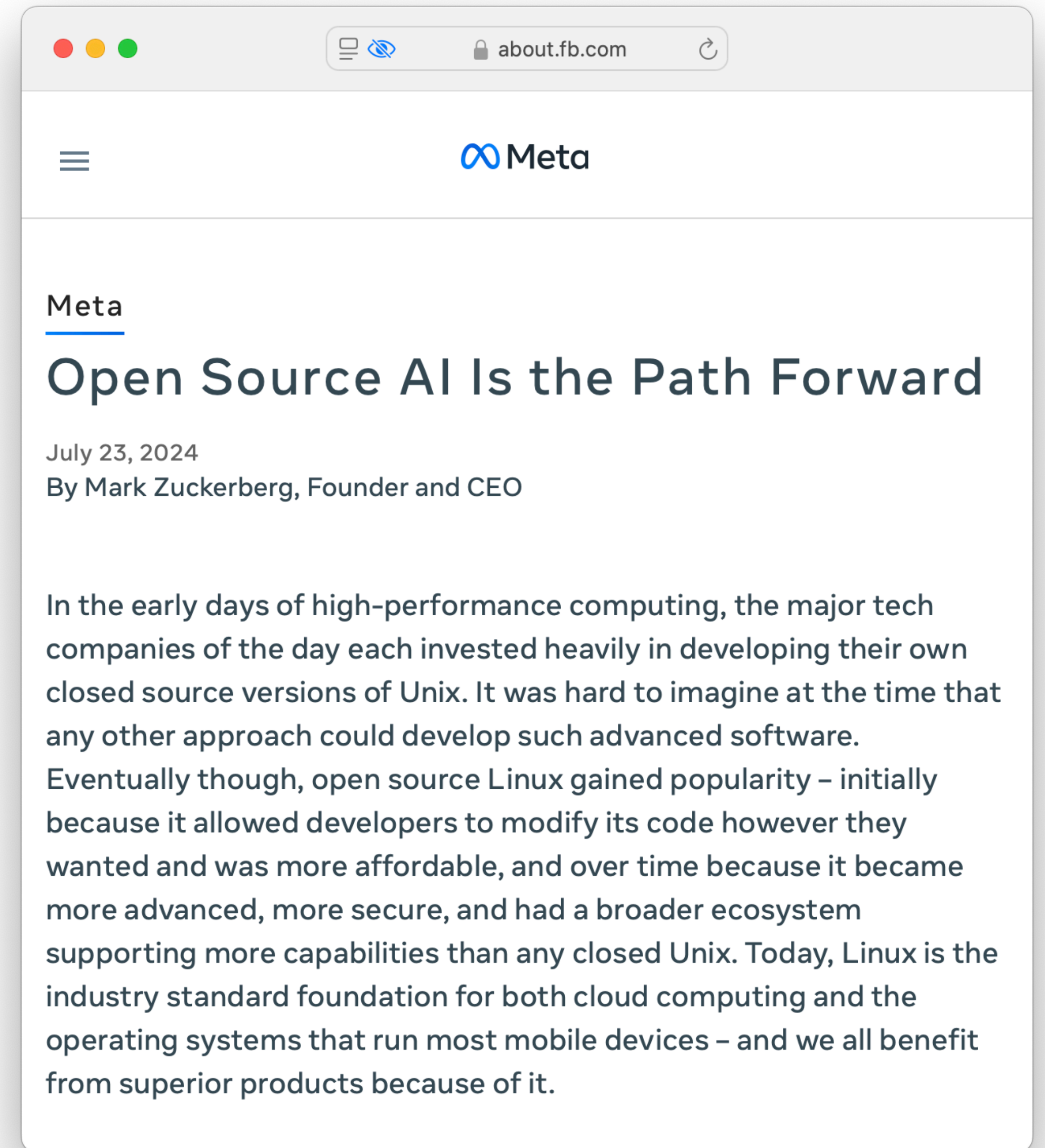


Source: Companies, LMARENA

“Everyone in tech is giving
someone else’s business
model away for free”

Meta’s open source

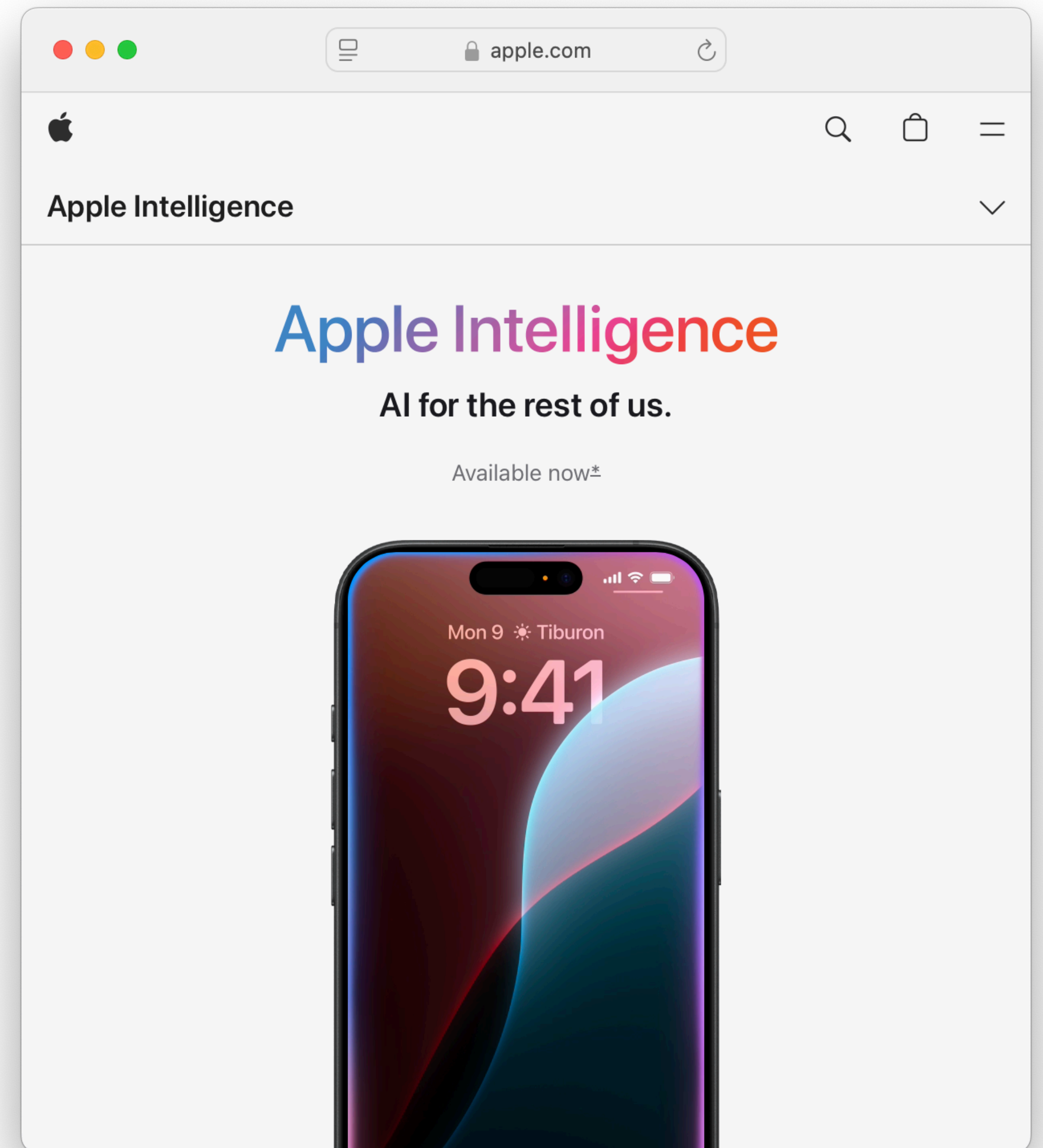
Turn models into commodity infrastructure!



“Everyone in tech is giving
someone else’s business
model away for free”

Apple’s edge computing

Turn models into just another API call!



The great model boom of 2023-2024

Better, faster, cheaper - pick two



Best model

Best cheaper
model

Best model
that fits on
the edge

And more
and better
every few
weeks

“If anything in this life is certain, if history has taught us anything, it is that you can kill anyone.”

Michael Corleone

“If anything in this life is certain, if history has taught us anything, it is that ~~you can kill anyone.~~”

Semiconductors are cyclical
Commodity tech goes to marginal cost

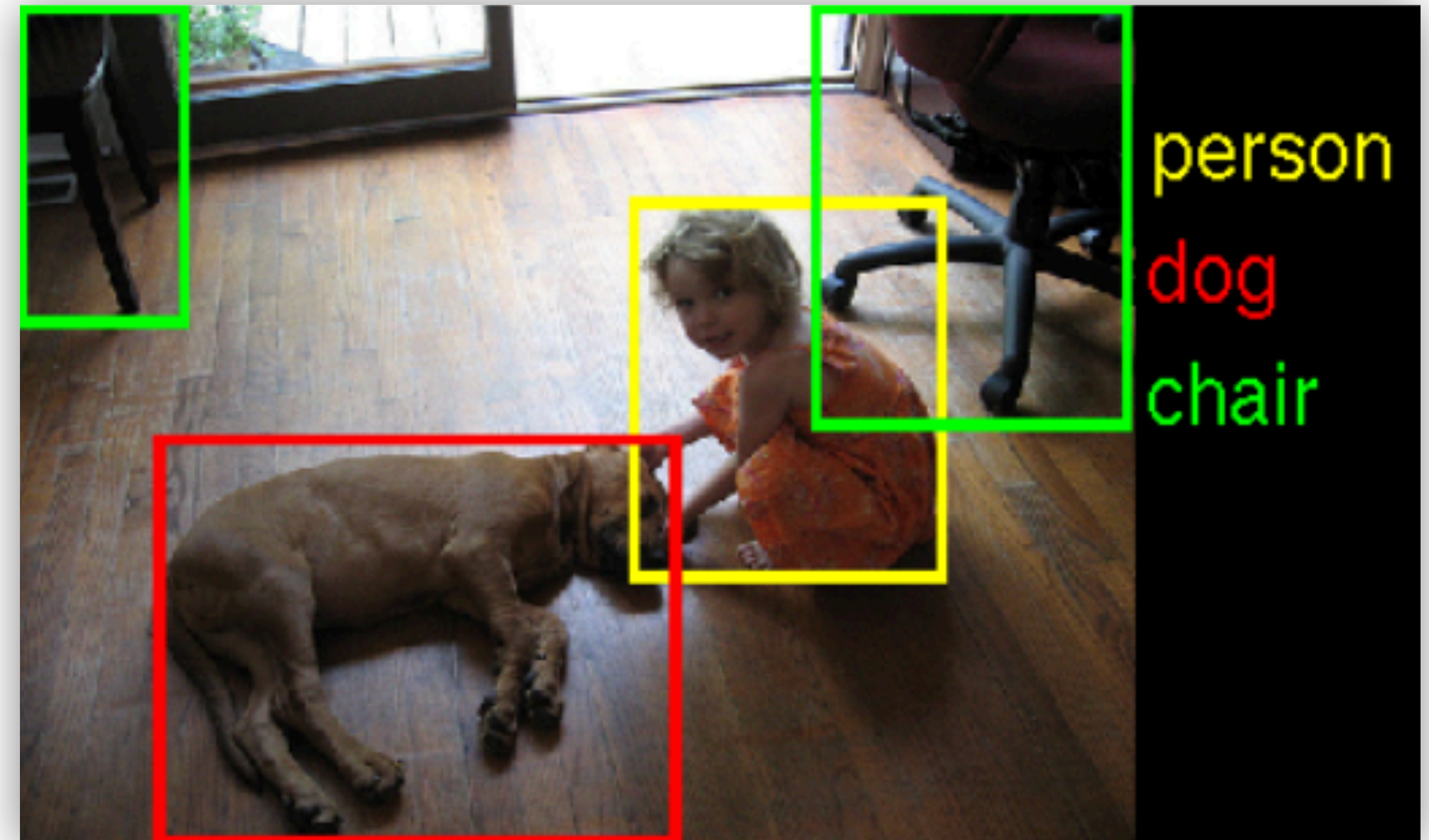
“If anything in this life is certain, if history has taught us anything, it is that ~~you can kill anyone.~~”

Semiconductors are cyclical
Commodity tech goes to marginal cost
And every new tech produces a bubble

How is this useful?

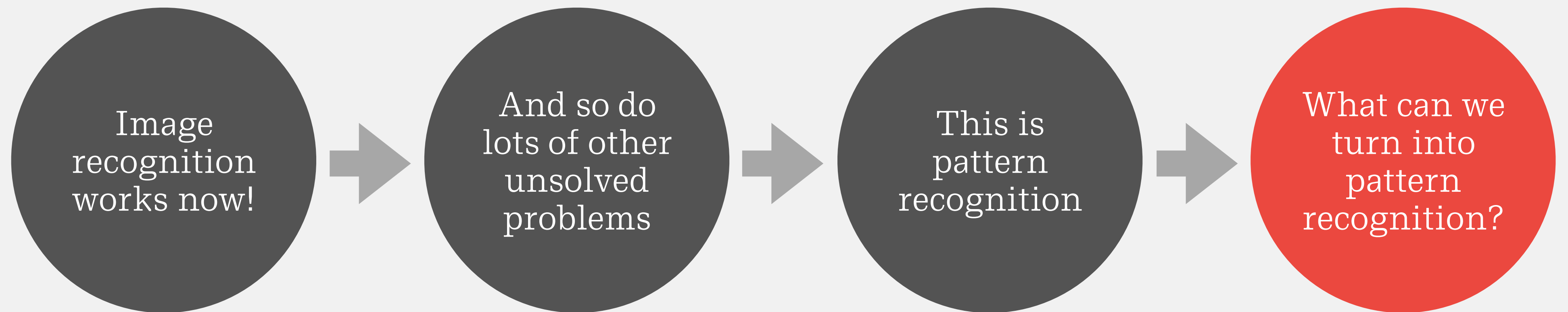
2013: how is Machine Learning useful?

“That’s clever... but so what?”



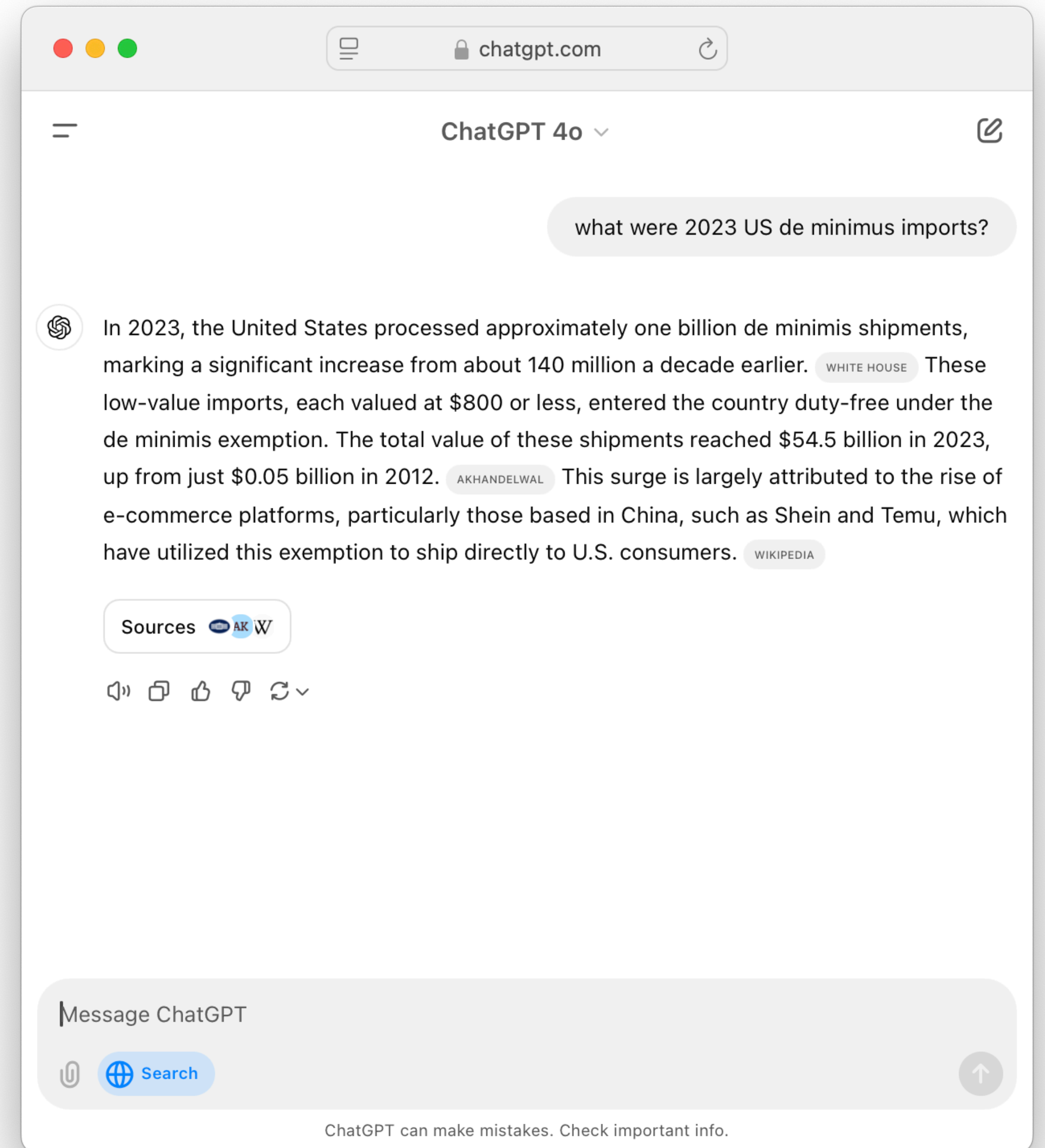
2013: how is Machine Learning useful?

What's the right level of abstraction to understand this?



2023: why is Generative Machine Learning useful?

“That’s clever... but so what?”

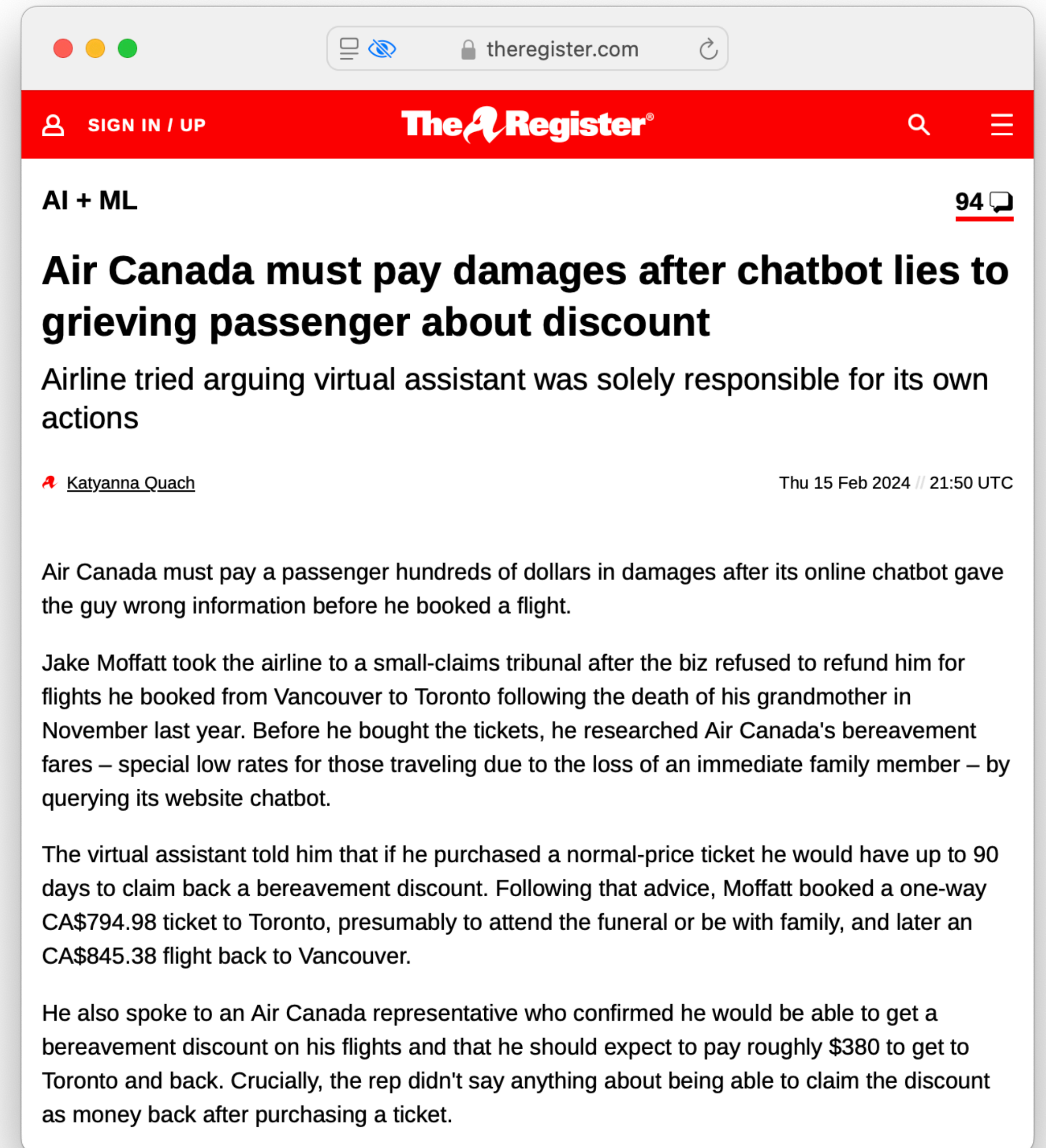


And what *can't* it do?

~~‘Answer this question’~~

‘What do answers to questions like this tend to look like?’

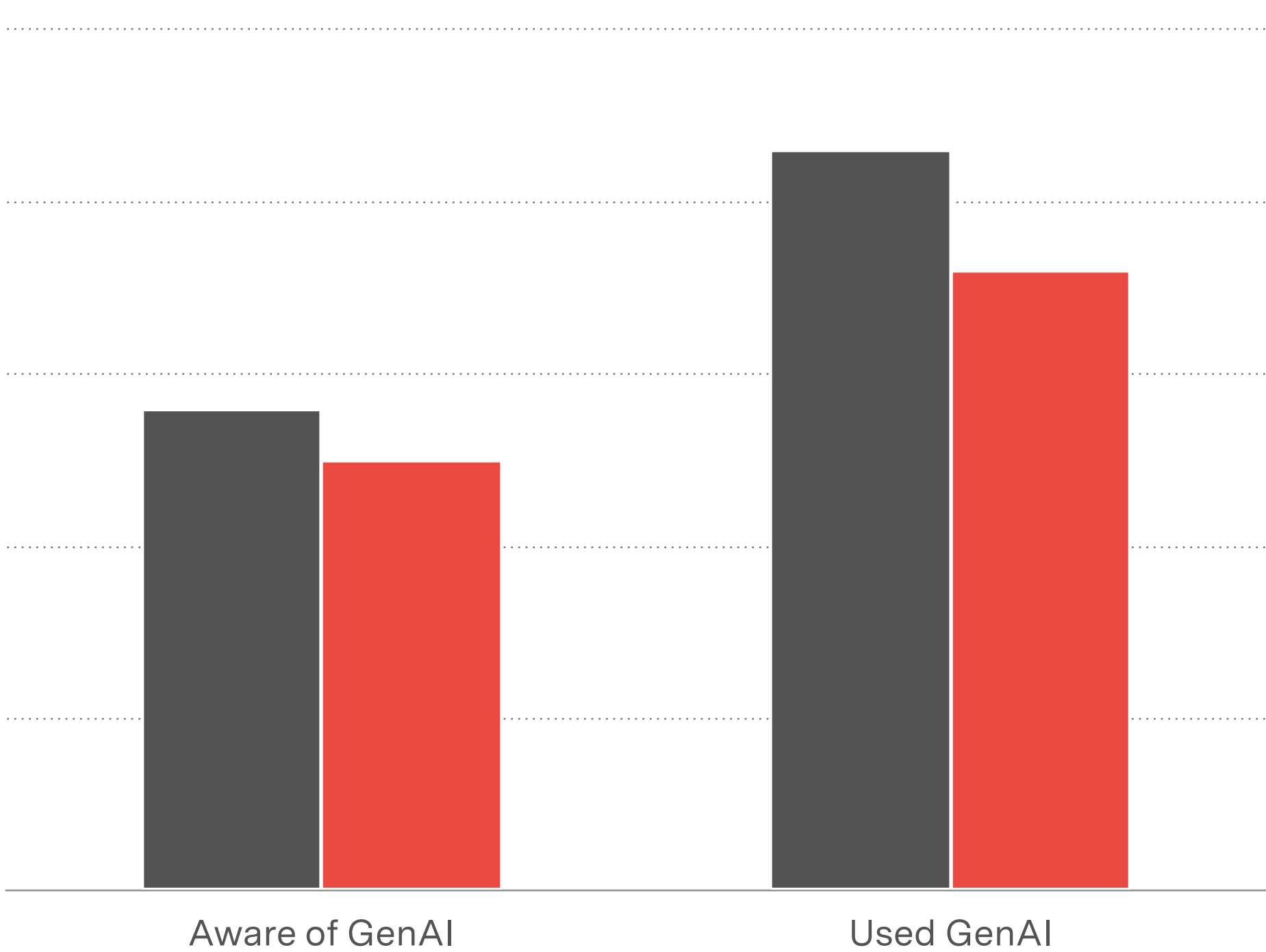
‘What would the average person probably say?’



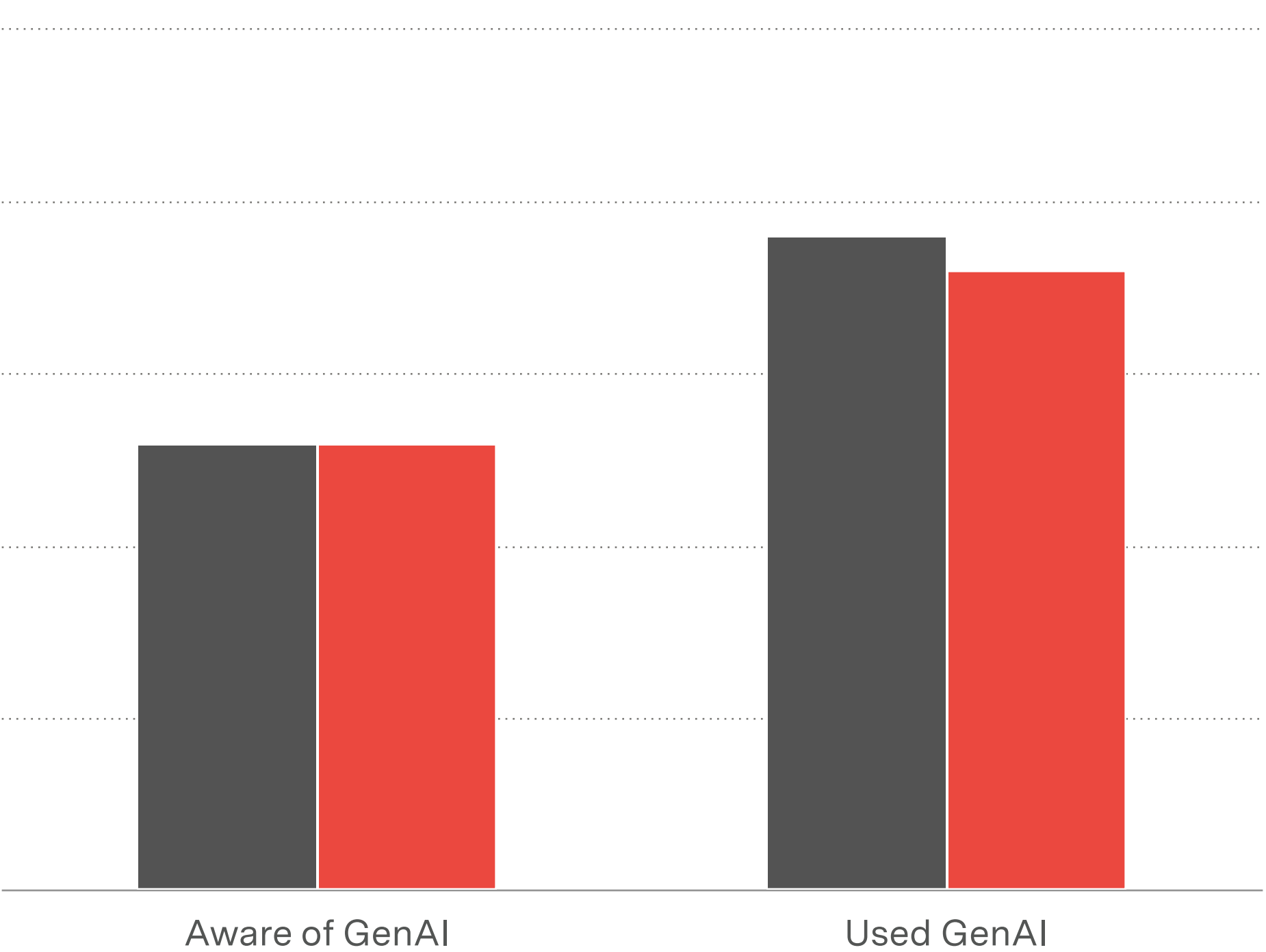
Error awareness is still limited

And the apparent fluency of text output conceals the nature of the model behind

Those who agree that “Generative AI always produces factually accurate answers”



Those who agree that “Generative AI responses are unbiased”



Handling ‘errors’ in a probabilistic system

Is this a science problem, or a use case and design problem?



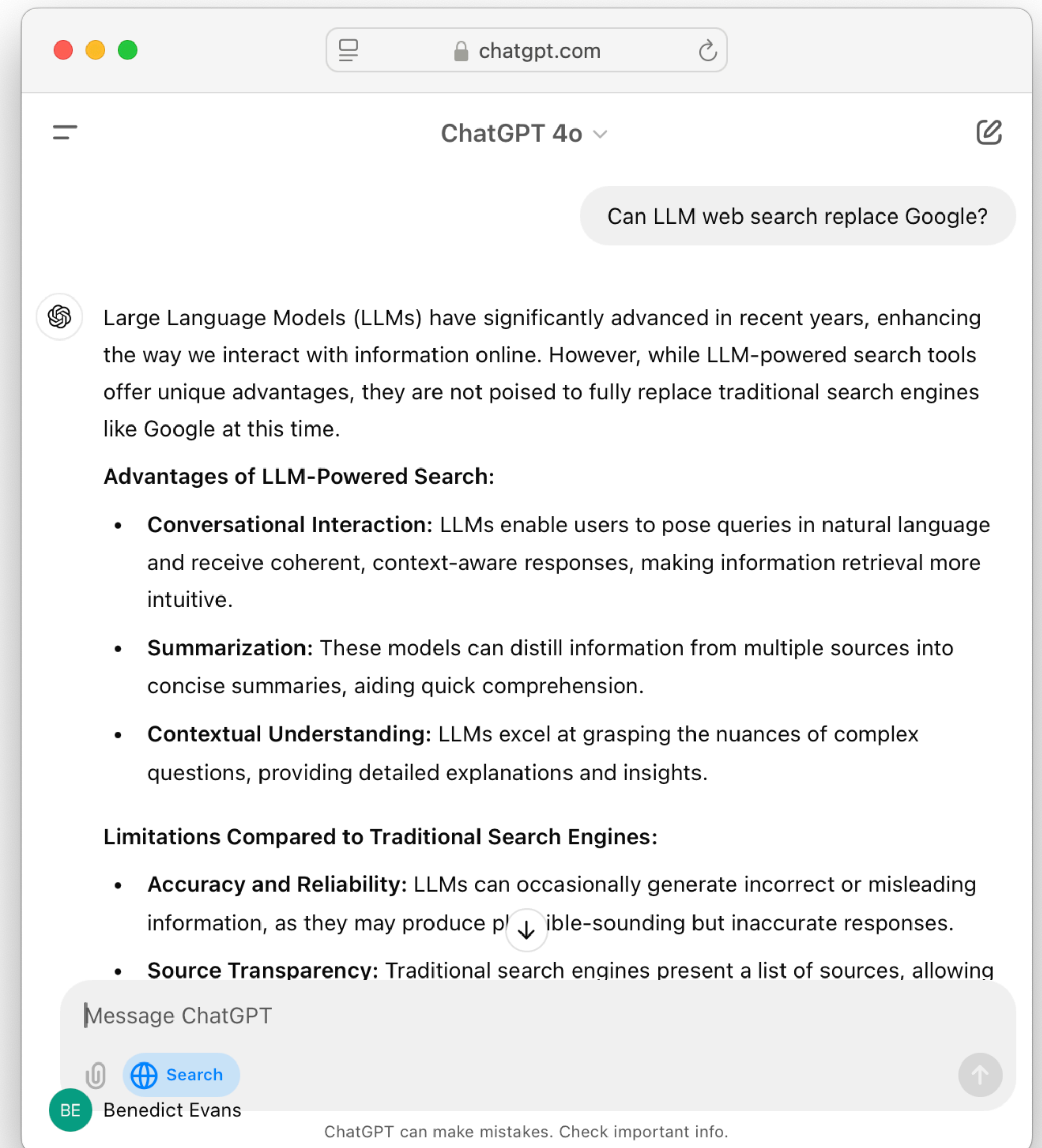
Can you use this for general search?

Do you want navigation, or just the answer?

Does it matter if it's wrong? Can you tell?

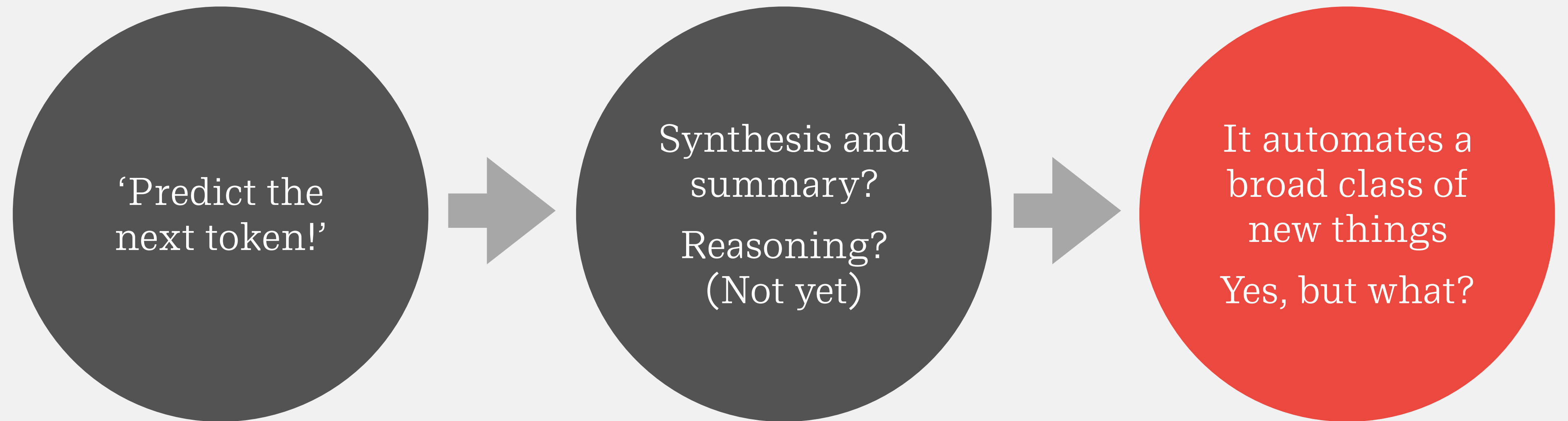
How much preprocessing and filtering does this need?

(Alphabet had \$56bn of FCF in the last 12 months, so it's worth finding out)



2024: how are LLMs useful?

What's the right level of abstraction to understand this?



AI gives you infinite interns

Can we predict the effects from first principles?

Useful - but imagine trying to do this exercise for 'the internet' in 1995

Cost arbitrage?

Labour versus
automation

Content?

Internet made
distribution 'free'
- do LLMs make
(some) creation
'free'?

Language?

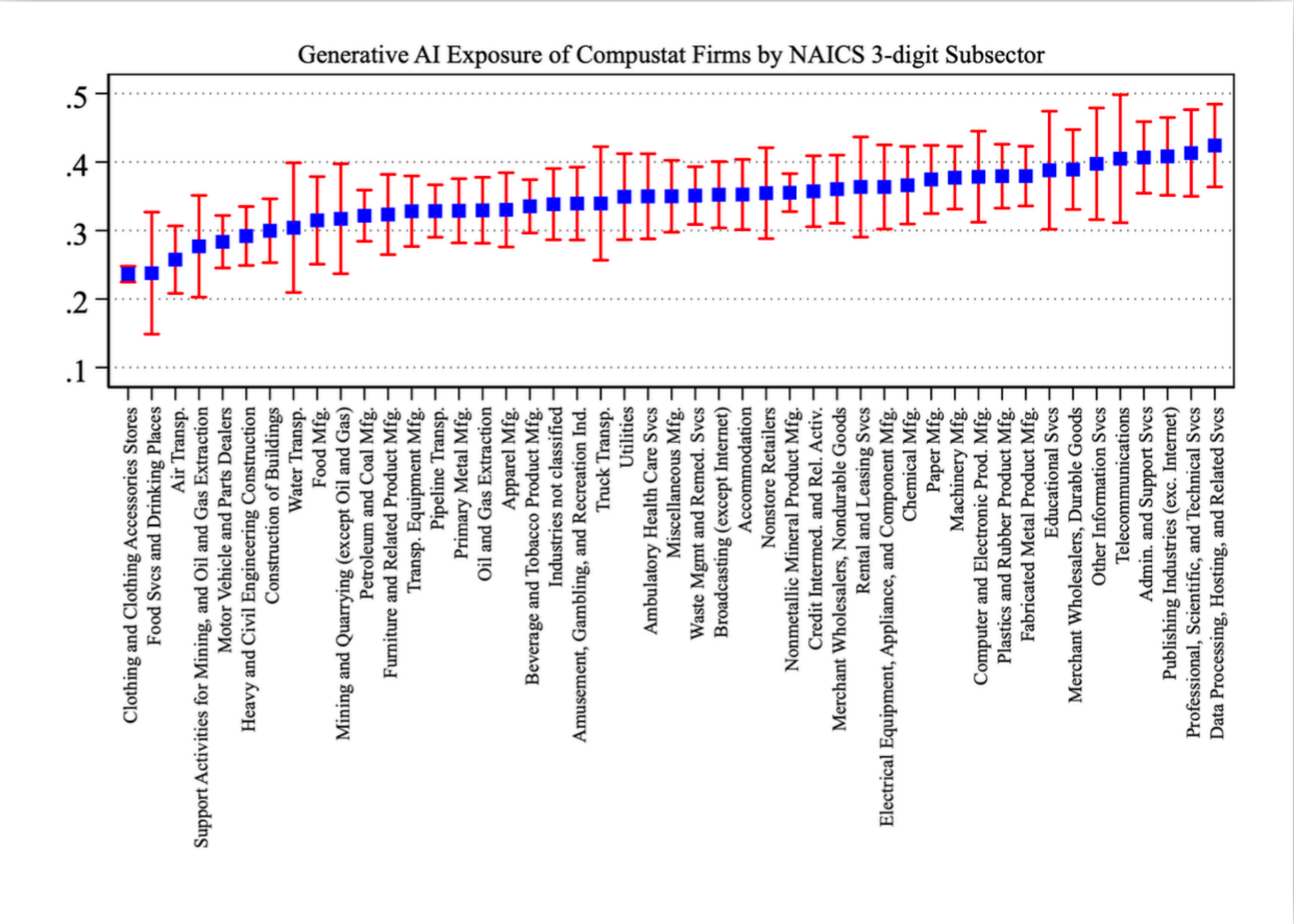
How does seamless
translation change
the web? Science?
Pop culture?

Can we predict the effects top-down?

Imagine doing this for ‘the internet’ in 1995

Or ‘mobile internet’ in 2005

What would you have got right, and wrong?



“The jury is still out out on whether Copilot is useful enough to justify the cost”

Chevron CIO, July 2024

Do you have that use case?

"VisiCalc took 20 hours of work for some people and turned it out in 15 minutes" - Dan Bricklin

Now, imagine a lawyer seeing it: "that's very clever, but I don't have that use case"



“I have that use case!”

Immediately, obviously useful for some professions and some workflows...

Coding

Errors easy to see
20-30% efficiency
gains

Marketing

Errors easy to see
No ‘wrong’
answers

Customer support

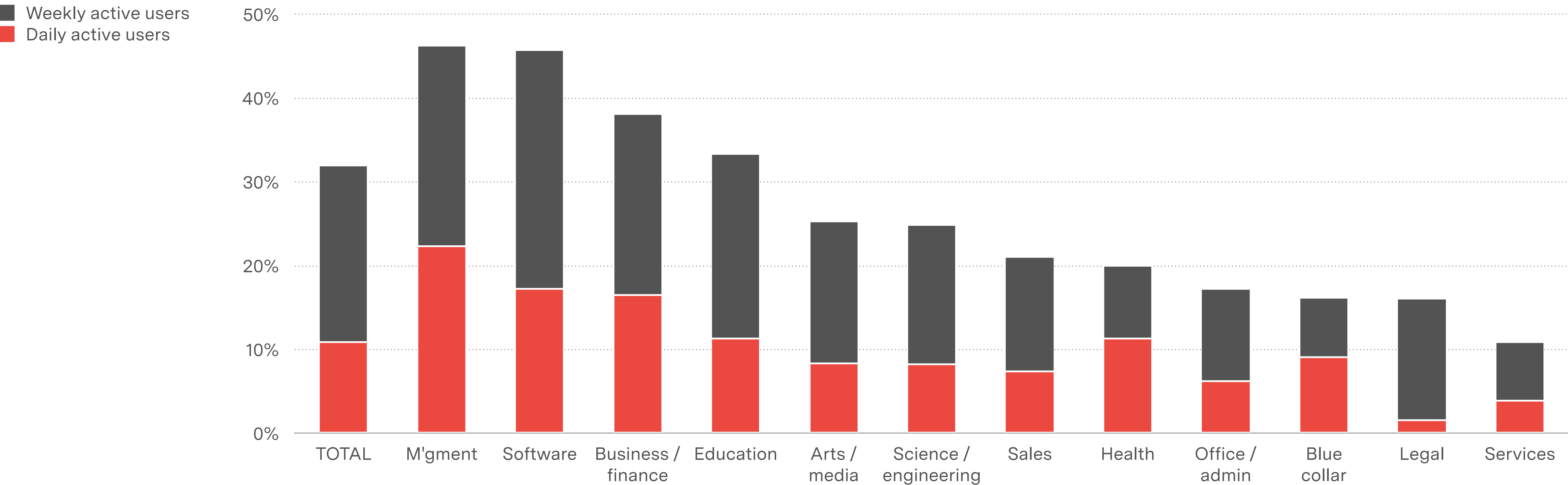
(But use with
caution)

Early adopters

“I have that use case!”

Glass half-empty or glass half-full? 10% adoption already is great - but why do 90% not find it useful?

Employed people, “Do you use generative AI?” response by occupation, USA, August 2024



Source: Blick, Blandin & Deming

“You’ve got to start with the customer experience and work backwards to the technology”

Steve Jobs

How do we deploy this?

How do we always deploy new technologies?

Standard patterns for deploying and using new technologies



Is this an Accenture question?

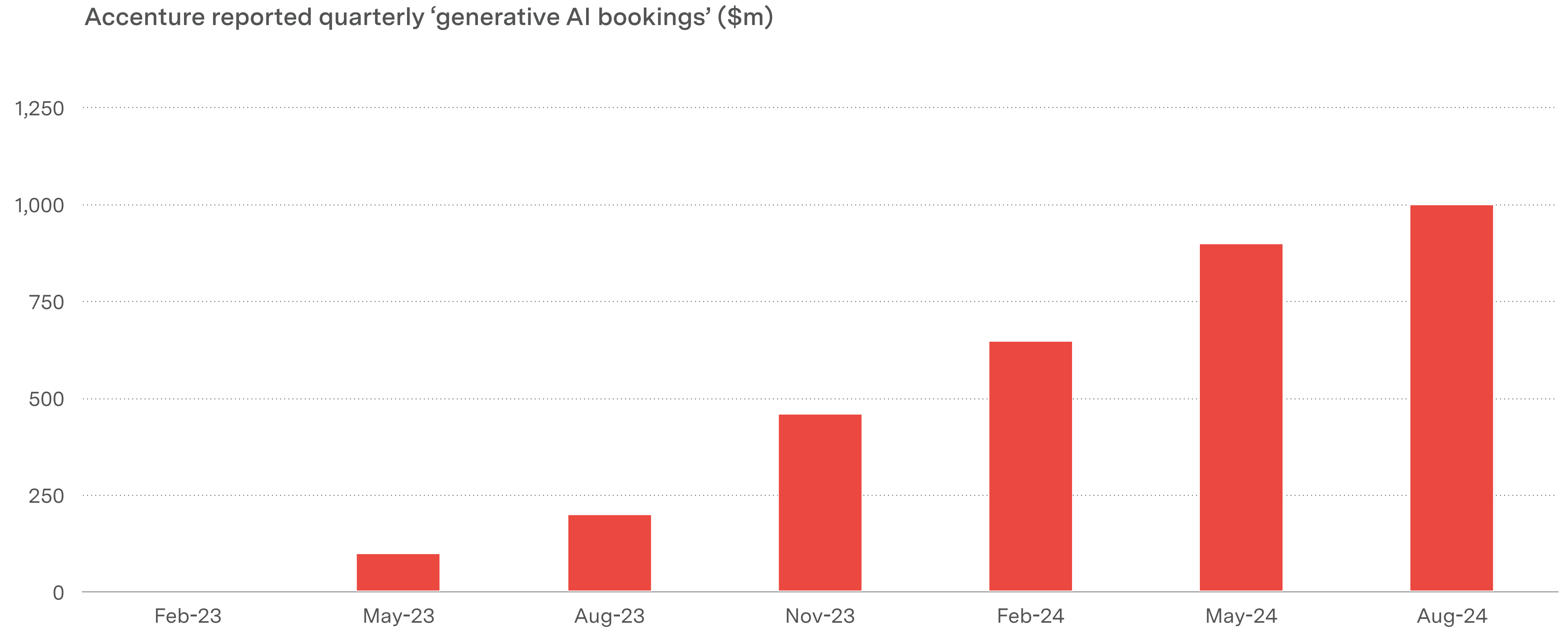
Or a Bain/BCG/McKinsey question?

A CIO question? CFO? CEO?

Top line or bottom line innovation?

“What’s our AI strategy?”

Accenture now claims more generative AI revenue than OpenAI - but almost all of this is still pilots



Standard tech procurement questions

How do enterprises always deploy new technology? What questions do they always ask?

Buy versus
build?

Single vendor
or multi-
vendor?

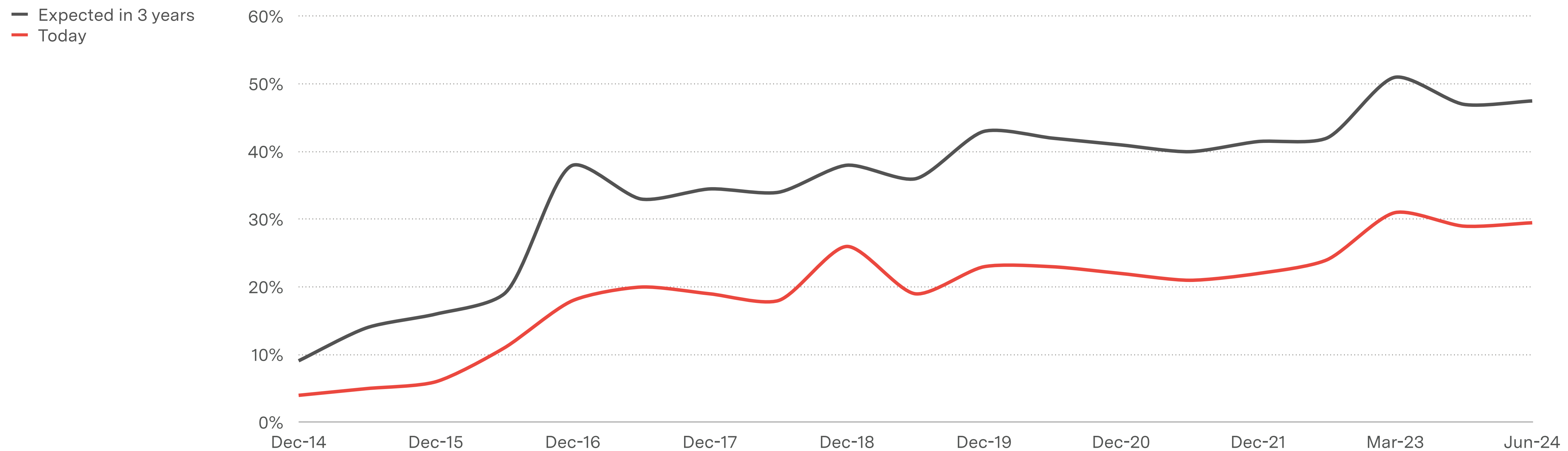
Which use
cases first?
Whose
budget?

Opex or
capex?
What's the
EPS impact?

The future can take a long time

Cloud is old and boring - but still only 30% of workflows

Enterprise workloads in public cloud

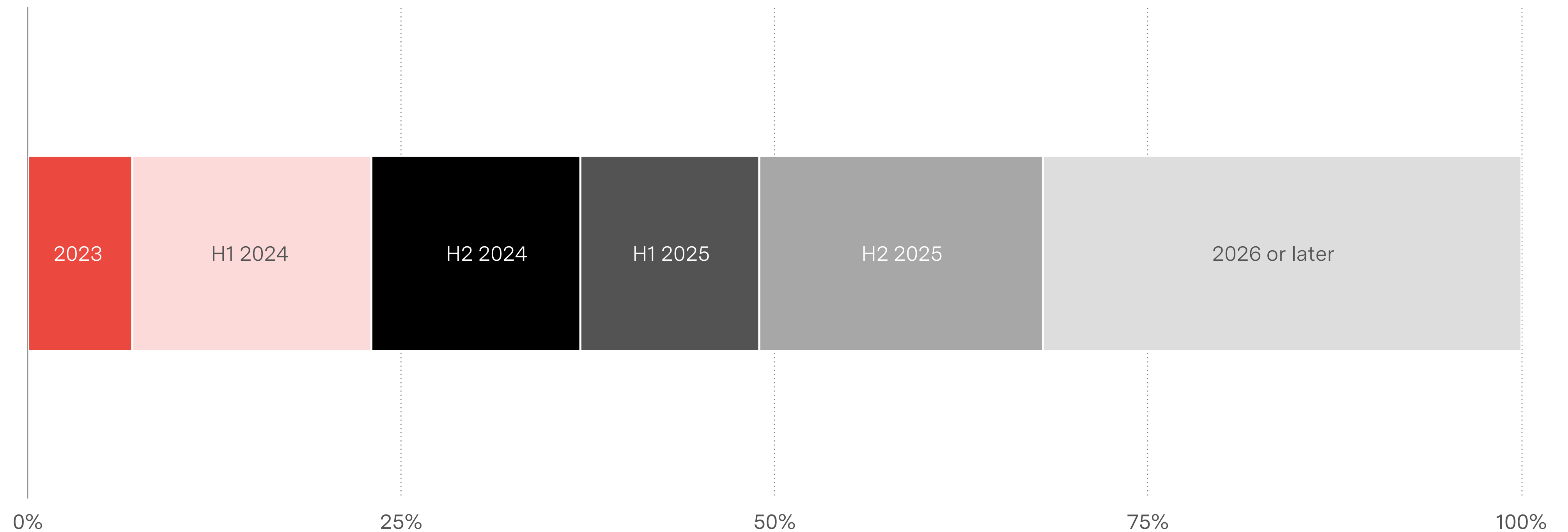


Source: Goldman Sachs CIO Survey

The future can take a long time

A quarter of CIOs have launched *something* - but half don't plan anything for at least a year

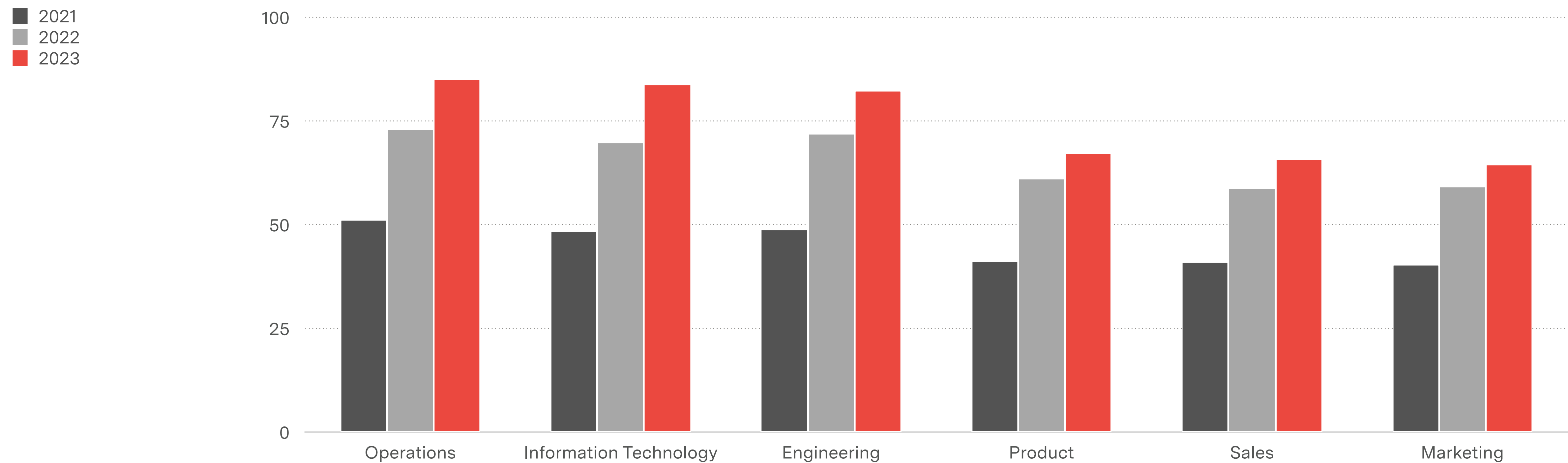
CIO expected timing for first LLM projects in production, August 2024



But eventually, new platforms mean new tools

SaaS enabled a huge expansion in automation, unbundling workflows out of SAP, Excel or email

Average number of SaaS apps used per department, large enterprises

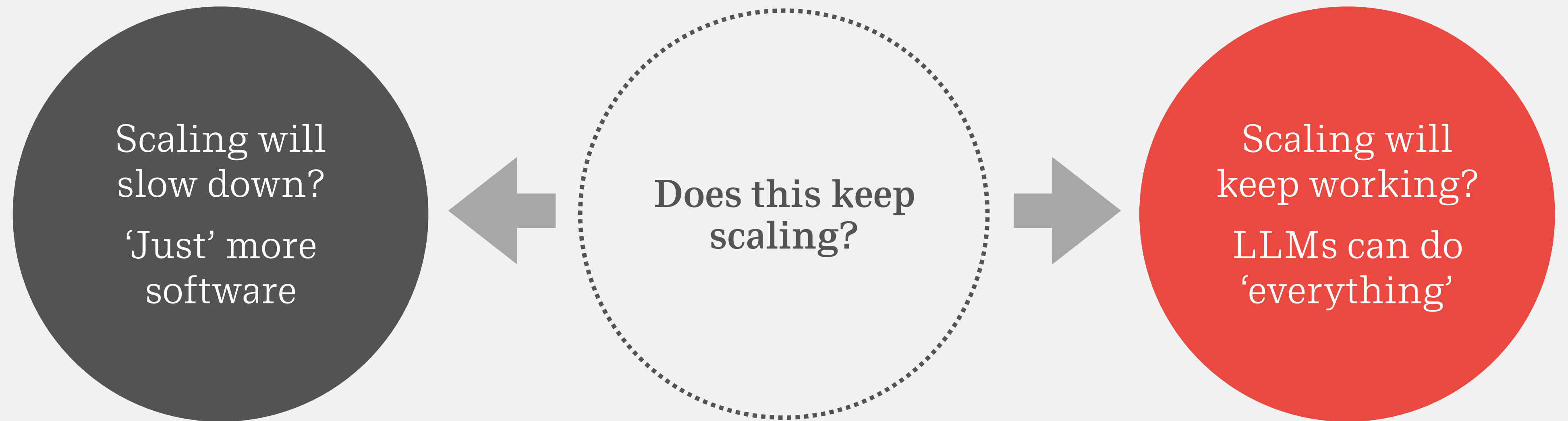


“There are two ways to make money.
You can bundle, or you can unbundle”

Jim Barksdale

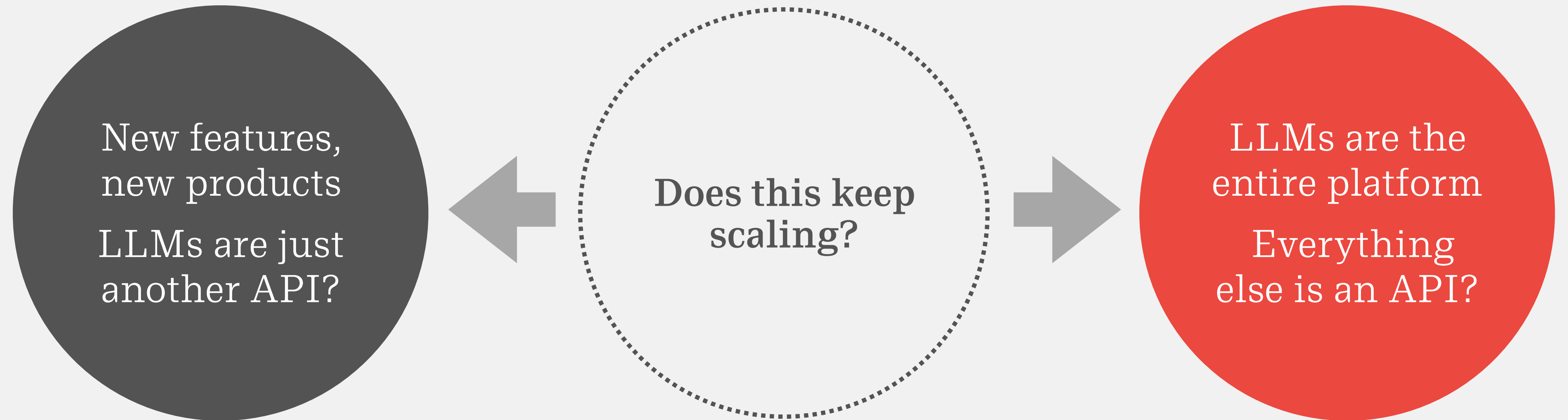
Unless the LLM can do the whole thing?

If the models get good enough, maybe we will need much less software?



Are LLMs Infra? APIs? Platforms? The new UX?

Will we use logical systems to control LLMs, or use LLMs to control logical systems?



“It's not the customer's job to know what they want”

Steve Jobs

Do LLMs break our use-case-discovery models?

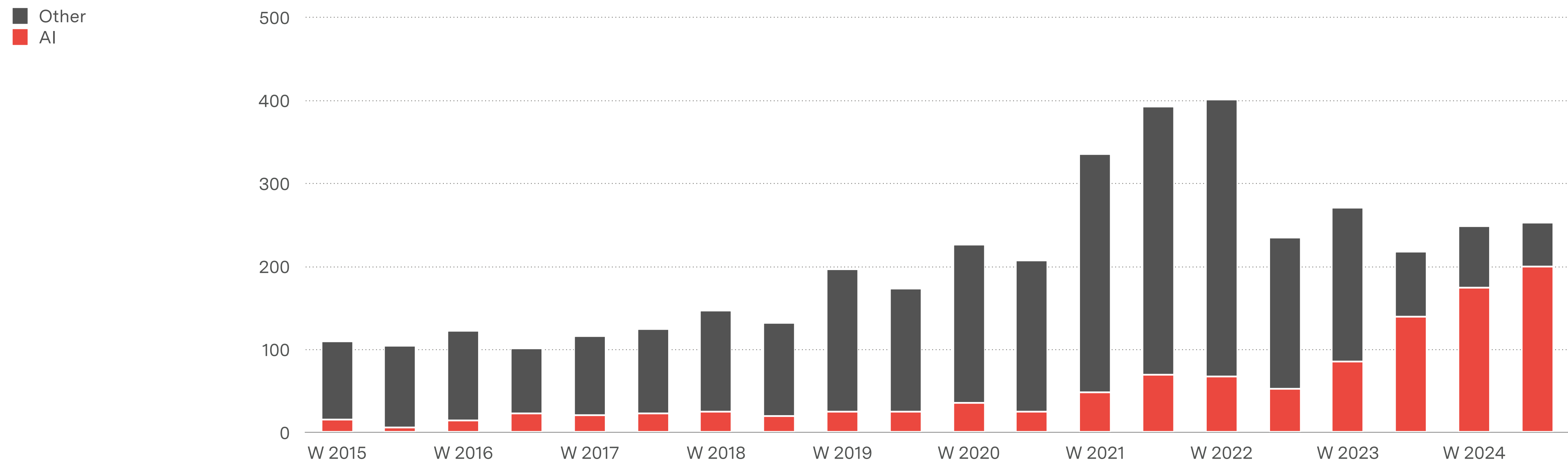
How do entrepreneurs invent new use cases and forms of self-expression if everything has the same UX?



New use cases, new unbundlers

How many AI startup are really a bet on unbundling both Oracle and ChatGPT?

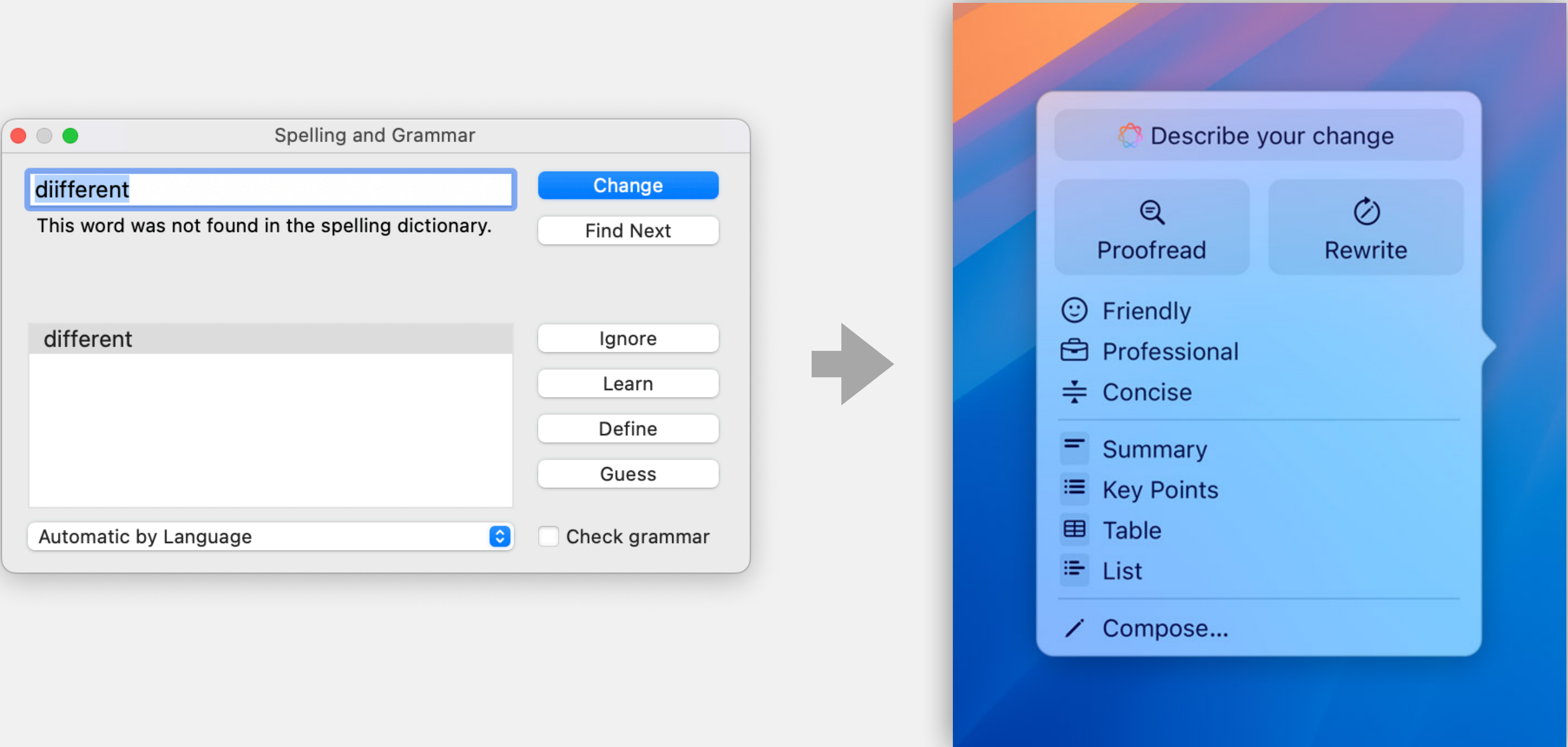
Y Combinator startups by field



Source: Y Combinator

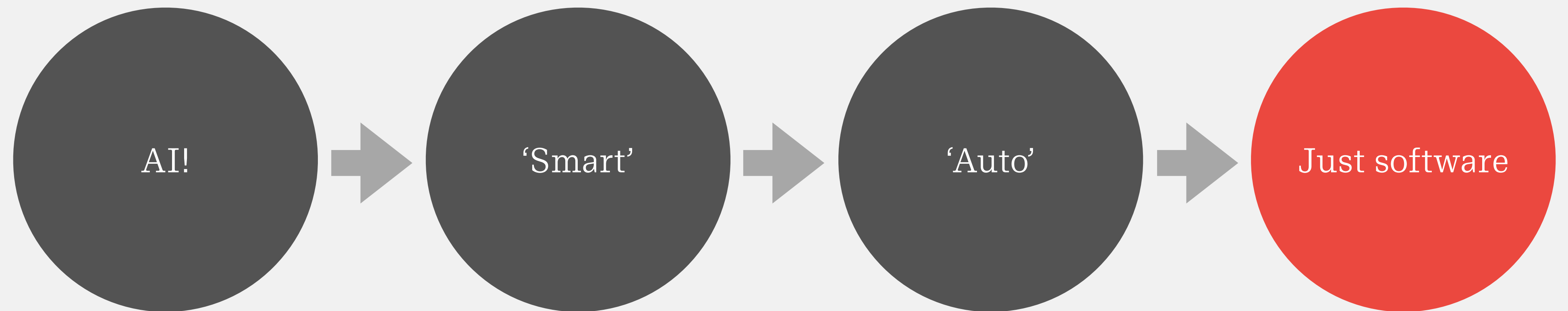
Sometimes, it really is just a feature

“AI is whatever doesn’t work yet”



‘AI’ tends to turn into ‘automatic’

As technology matures, it disappears



Three models for LLM product?

How far do LLMs enable new features and how much do they just replace apps entirely?

New features

Rewrite my email

Summarise the
reviews

New tools

Read 500 10ks
and tell me...

Generalised AI?

Buy me a house

“In from three to eight years we will have a machine with the general intelligence of an average human being”

Marvin Minsky, 1970

“‘Intelligence’ is whatever machines haven't done yet”

Larry Tesler, 1970

Perhaps, all AI questions have one of two answers

Will this just be like everything else? No-one really knows



Meanwhile...

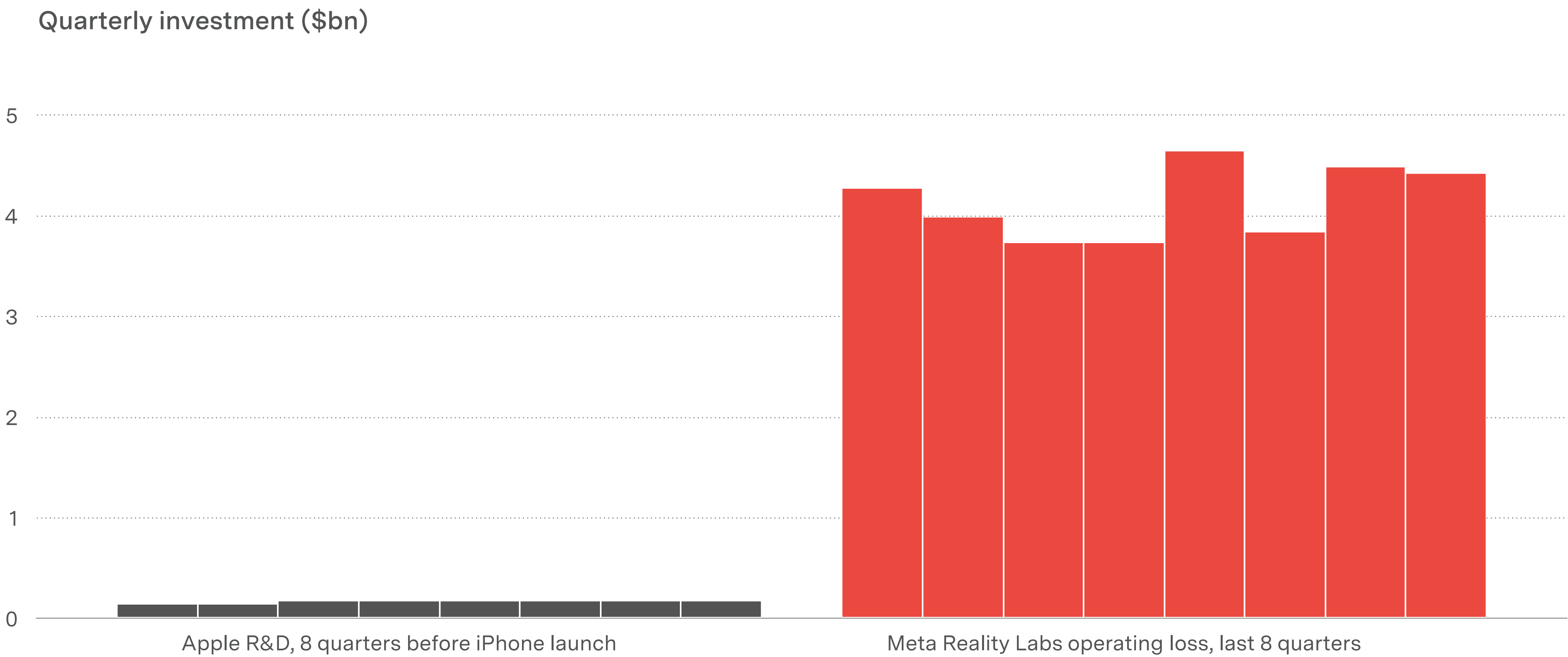
From vision to deployment

What's already big, what's being built, and what comes next?



Meta is still metaversing

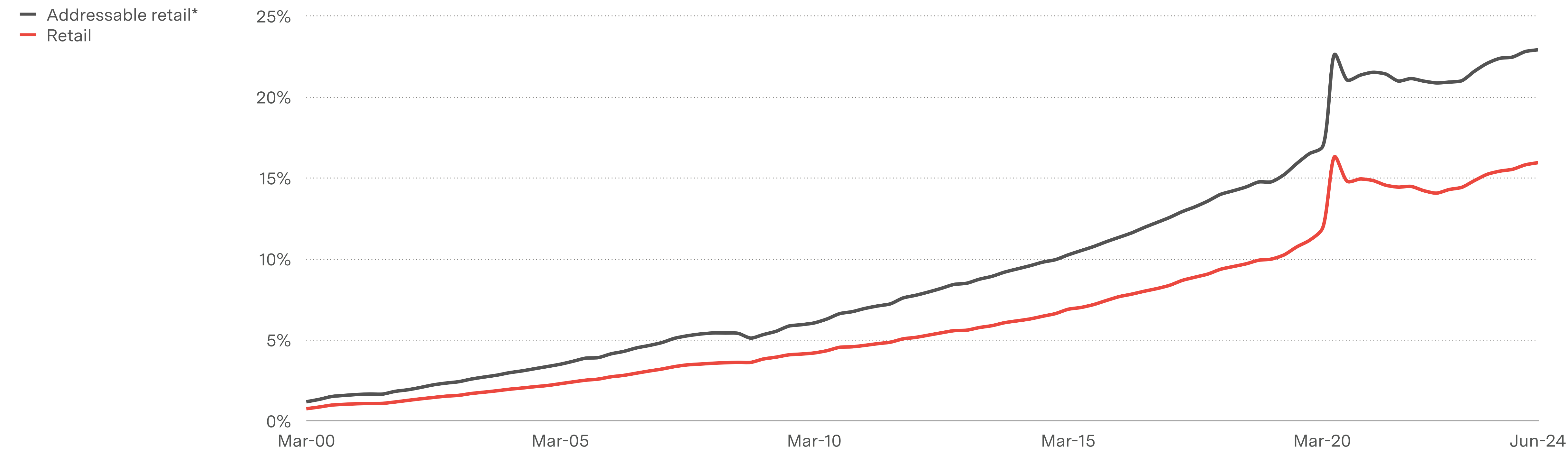
Meta still believes in VR & AR - it's invested at least \$60bn so far, and \$17.4bn in the last 12 months



E-commerce is still there

The most boring chart in tech, with a brief exception - but this is trillions of dollars of global value

E-commerce as % US retail

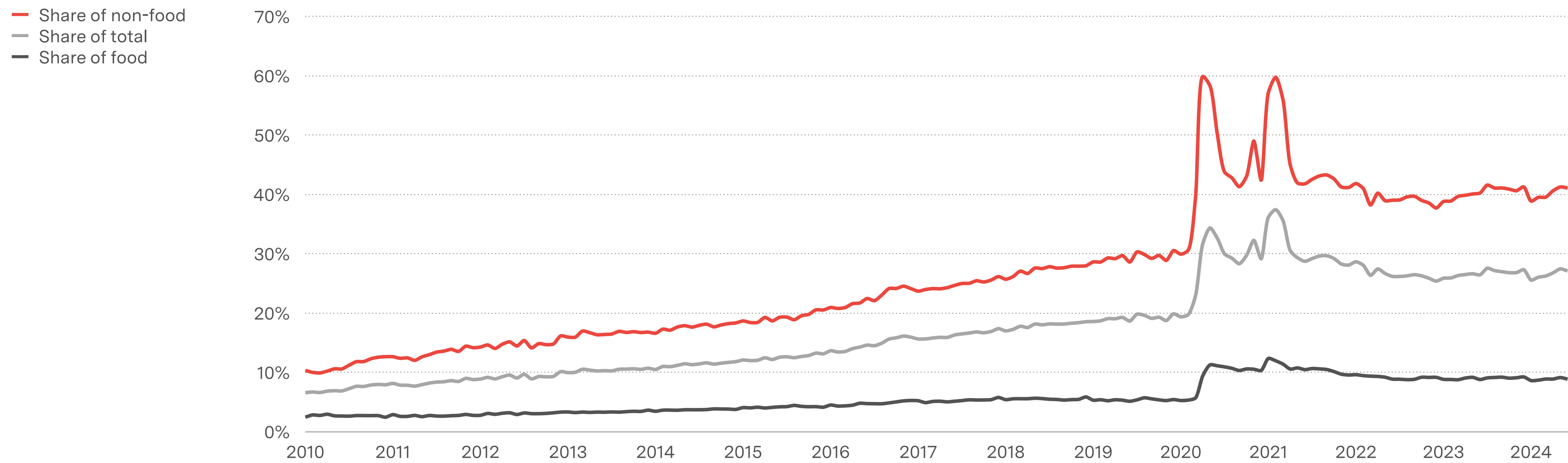


Source: US Census. Seasonally adjusted
* Excluding cars, car parts & fuel

‘Back to the trend line’

But 40% of UK non-food is now online

E-commerce as % UK retail

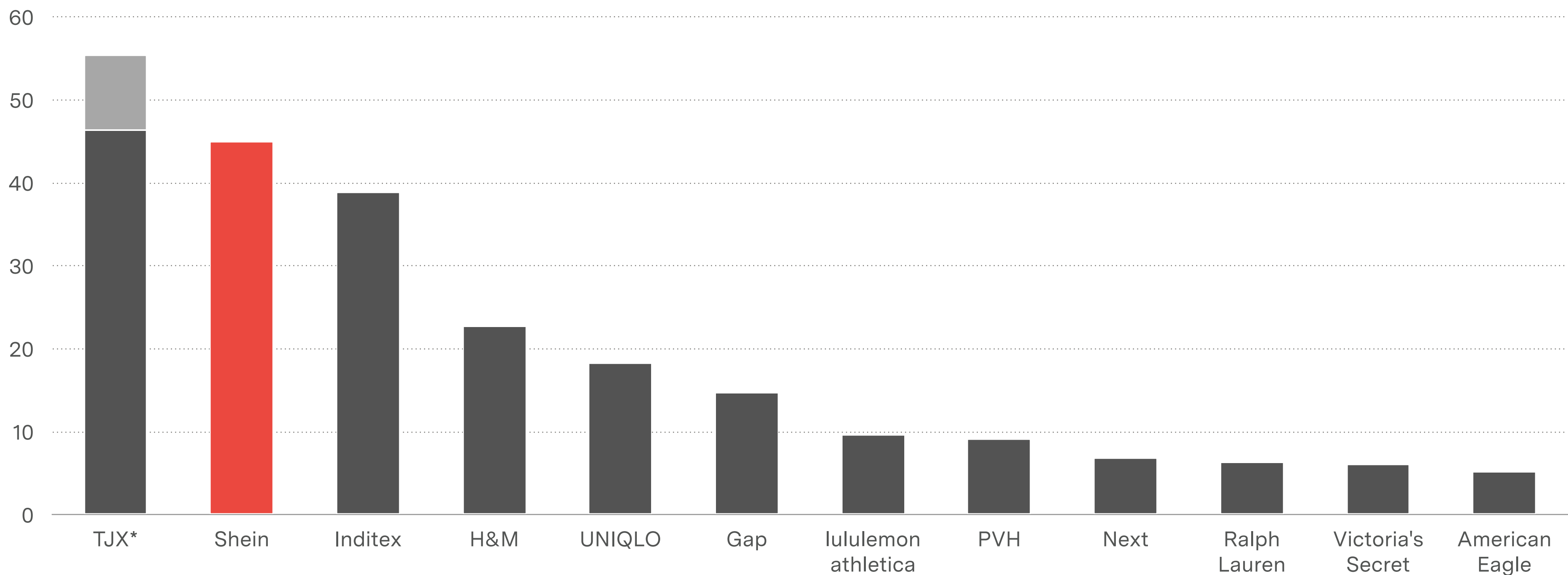


Source: ONS

New channels, new winners

Shein is on track to be the largest pure-play apparel retailer on earth

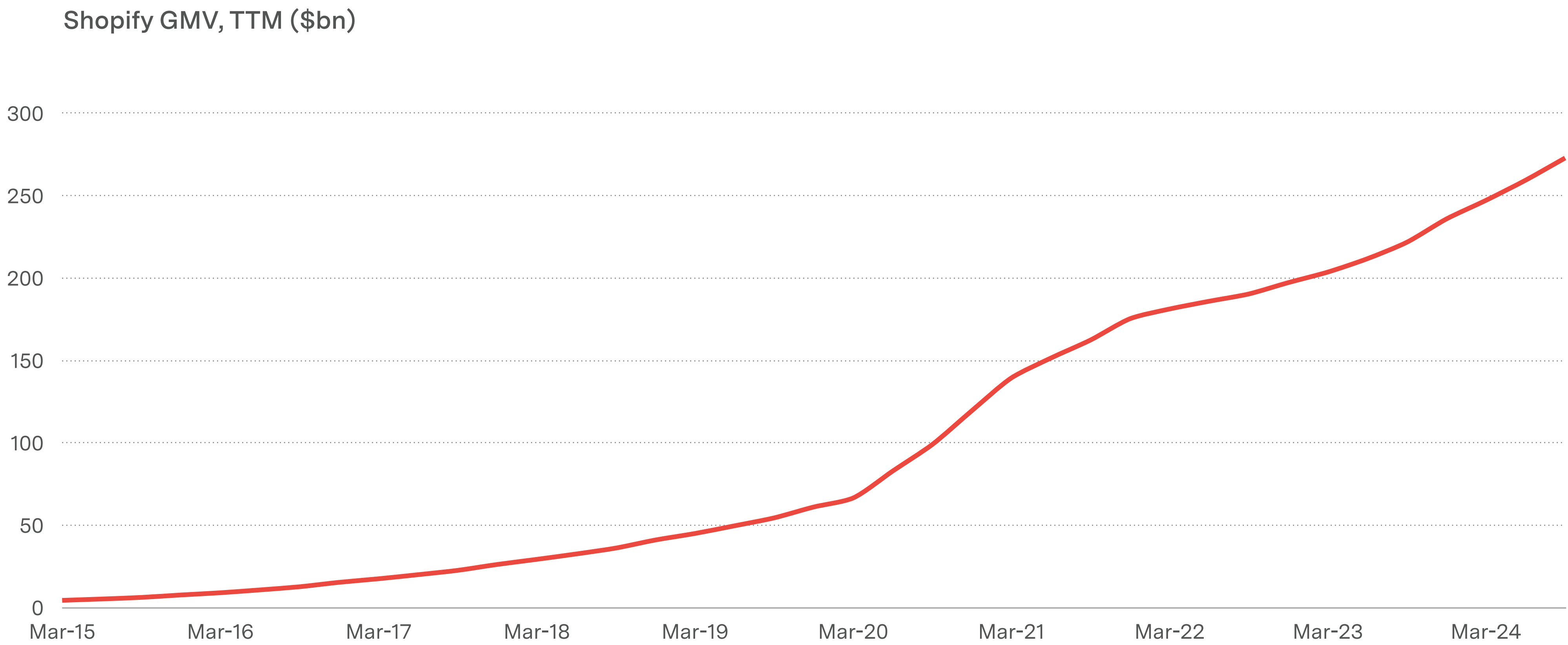
Largest global pure-play apparel retailers by revenue, 2023 (\$bn)



Source: Companies, press reports of GMV for Shein.
* TJX reports \$54bn including \$9bn of home goods. NB Amazon has est. \$60-80bn apparel sales

Unbundling Amazon

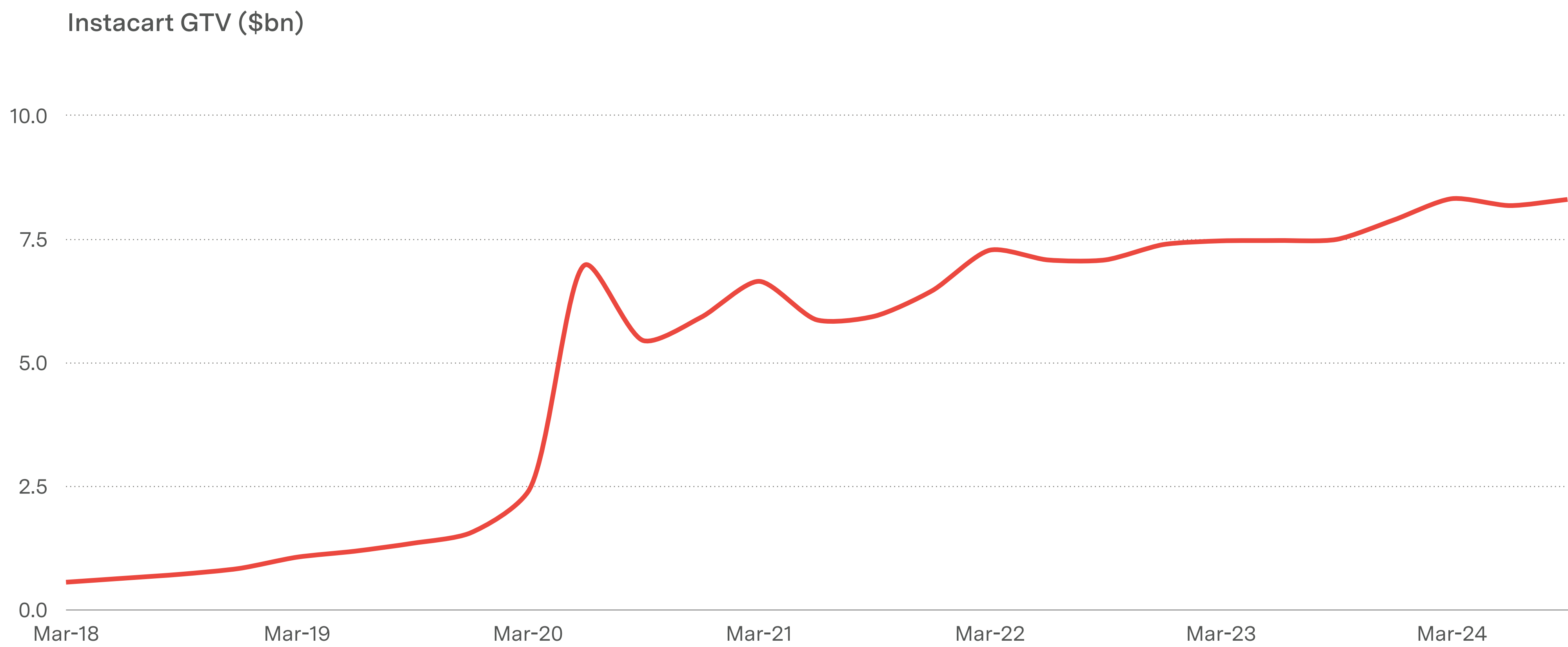
Shopify powered \$270bn of e-commerce in the last 12 months - 35% the size of Amazon GMV



Source: Shopify

Breaking habits

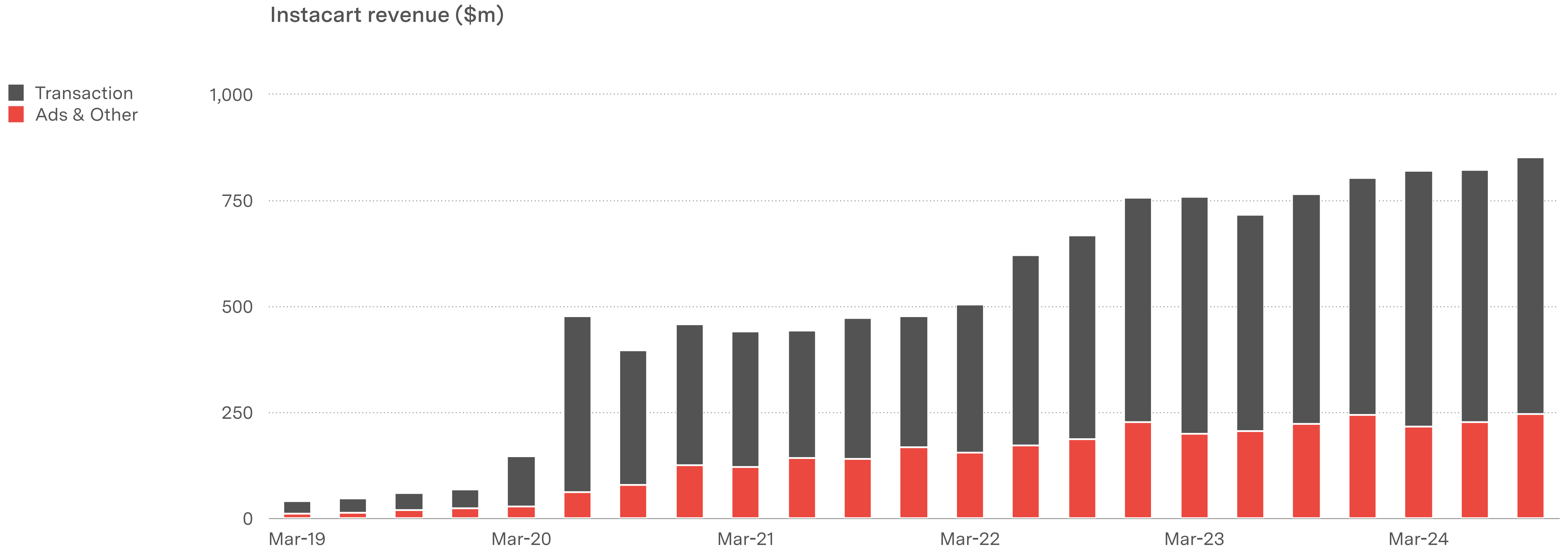
The pandemic drove a step change for grocery, and Instacart now has a \$33bn run-rate



Source: Instacart

A delivery business, or a \$1bn ad business?

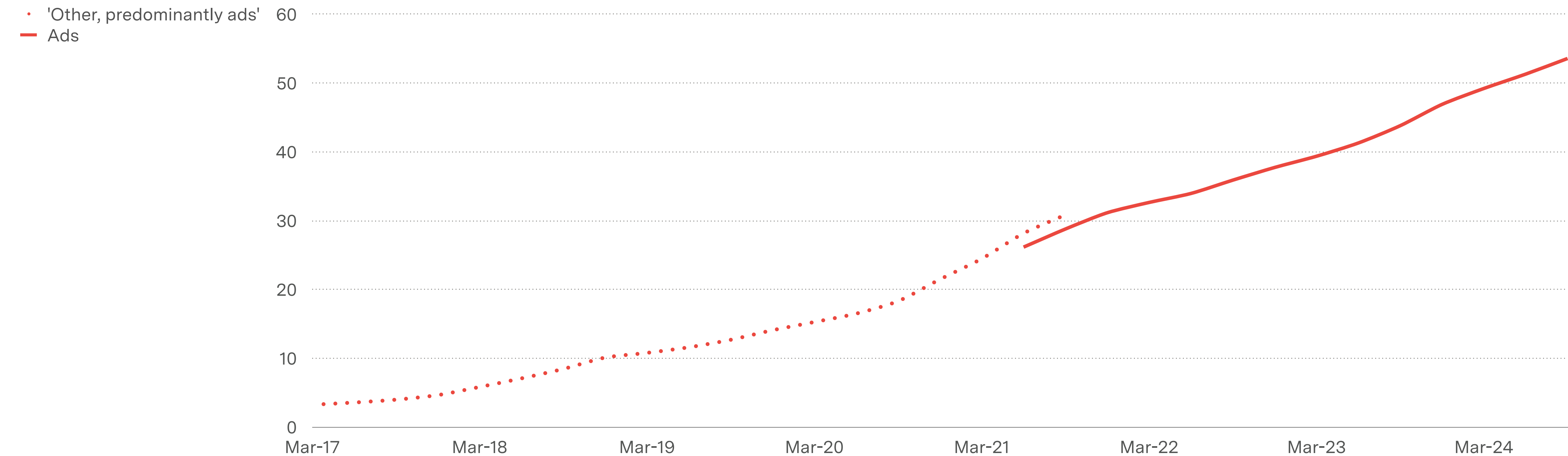
A quarter of Instacart revenue and almost all the FCF comes from advertising



The retail media boom

Amazon probably has more cashflow from advertising than from retail or AWS

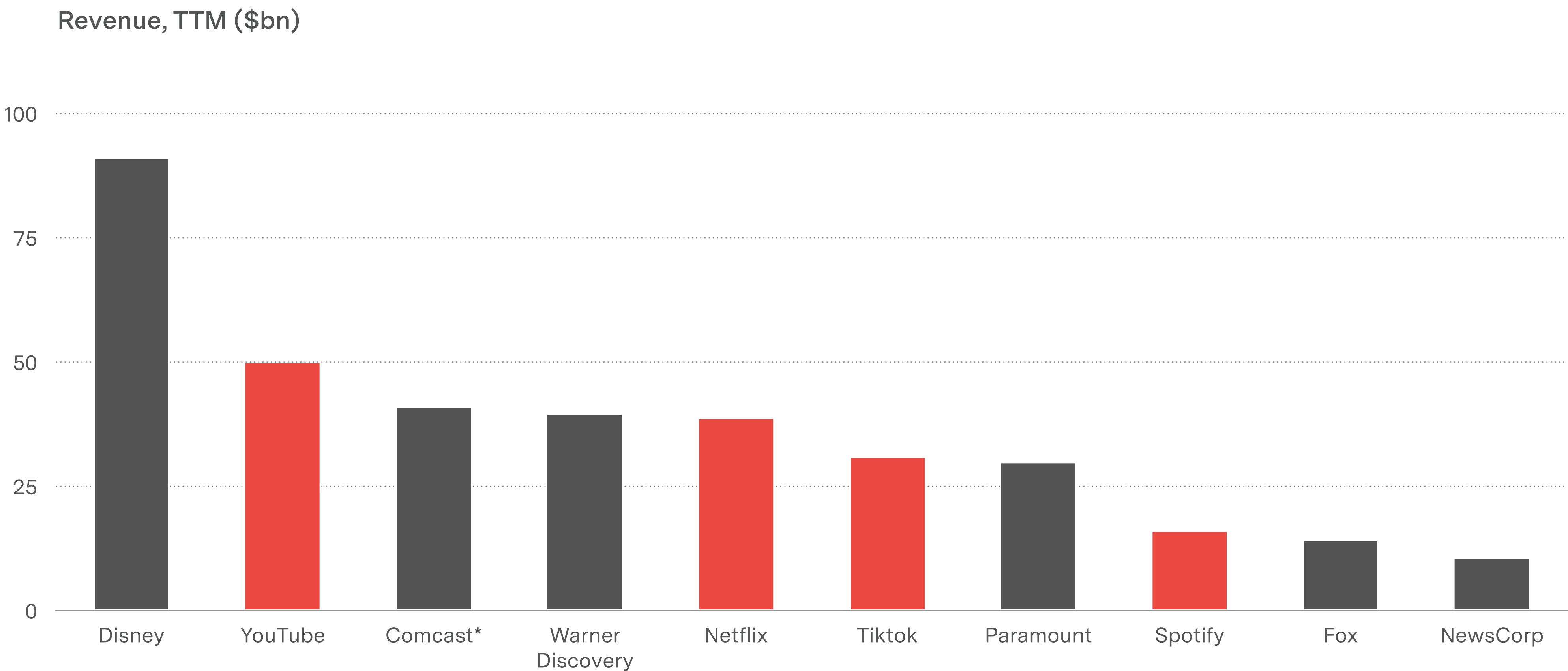
Amazon advertising revenue, TTM (\$bn)



Source: Amazon

Software eats media

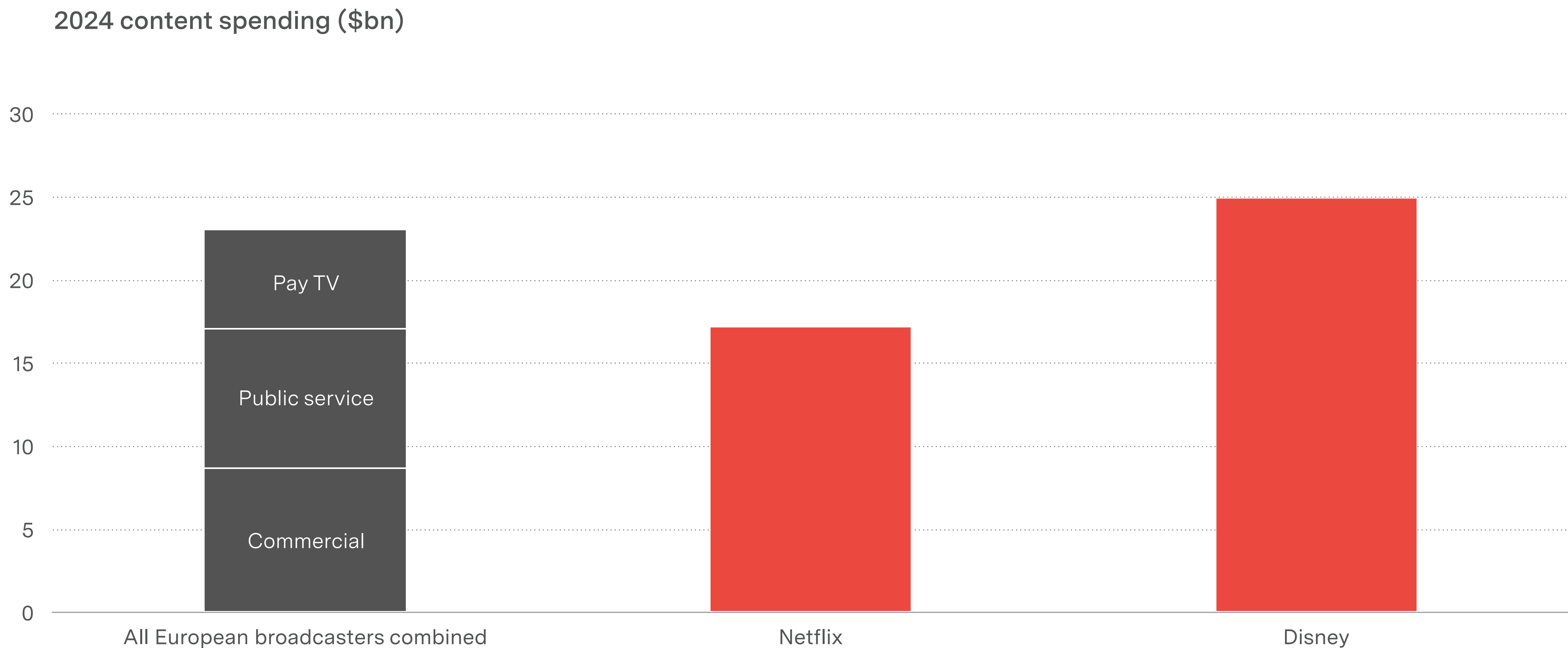
New channels, new models, new bundles



Source: Companies, press reports for TikTok
* Media division only

Big fish, little fish

Streaming means TV companies must now compete across the whole value chain

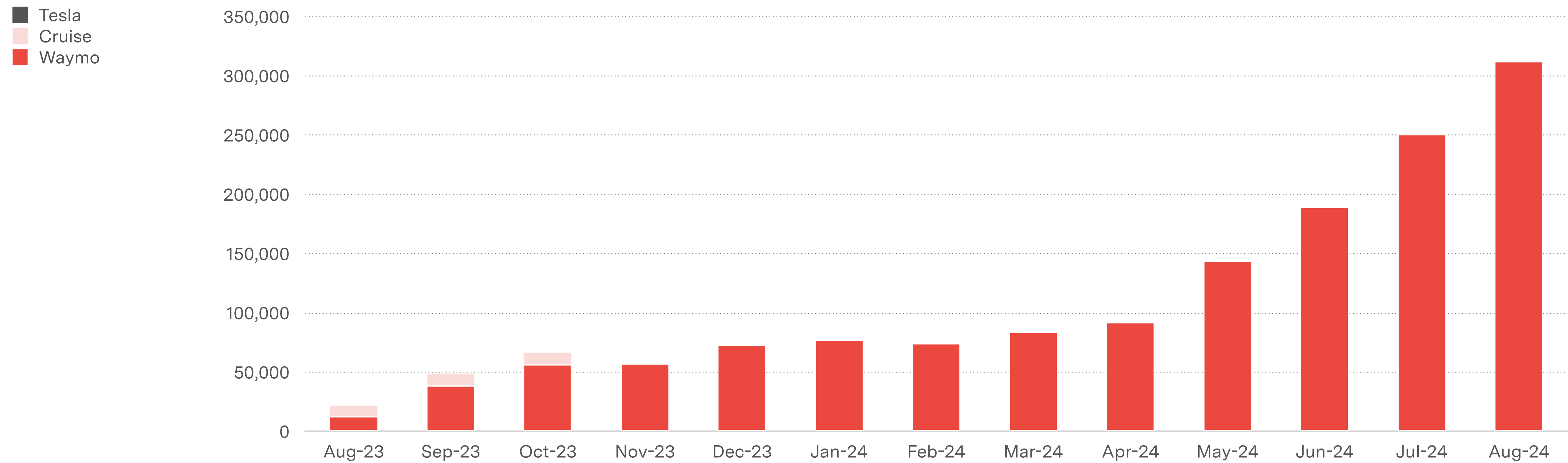


Source: Ampere, companies

Software eats cars?

After a decade of promises and tens of billions of dollars, autonomy might be starting to work

Monthly ‘robotaxi’ trips in California

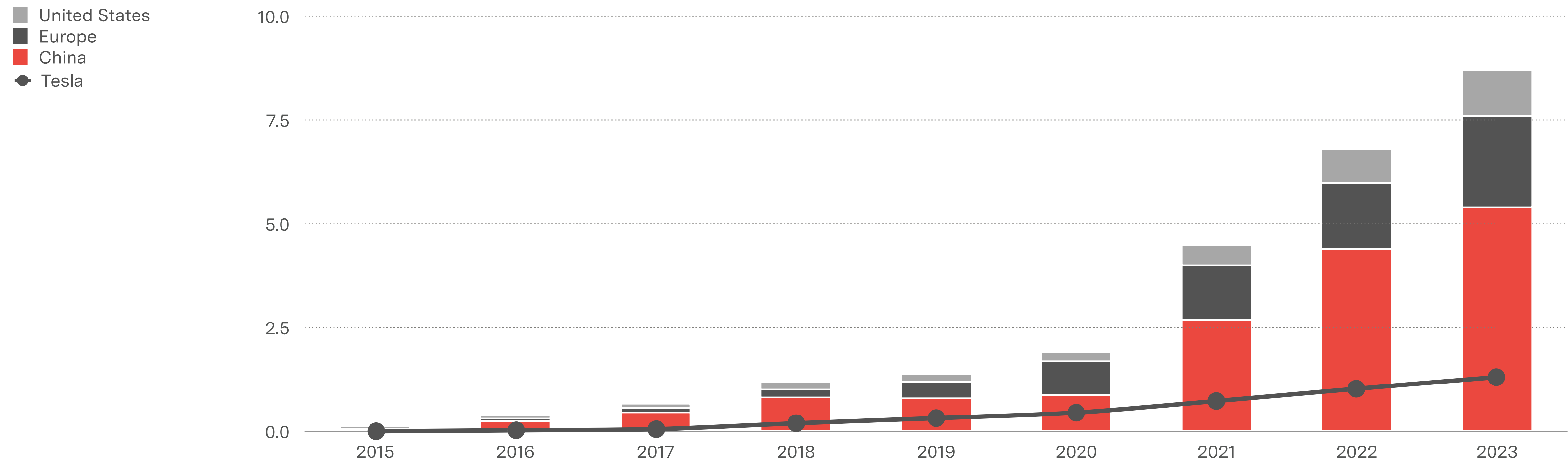


Source: California Public Utilities Commission

Cars become software?

BEVs are close to 10% of unit sales, but who will win? Will this work like Android?

BEV unit sales (m)



Source: IEA, Tesla

But what kinds of questions are these?

Tech changes the landscape for each industry, and then moves on

Will TV be re-aggregated? Does streaming mean a structurally lower profit pool?



Ask a TV analyst

Shein is the biggest apparel retailer on earth. Is that sustainable?



Ask an apparel analyst

Will EVs let the Chinese car industry do in the 2020s what the Japanese did in the 1980s?



Ask a car analyst!

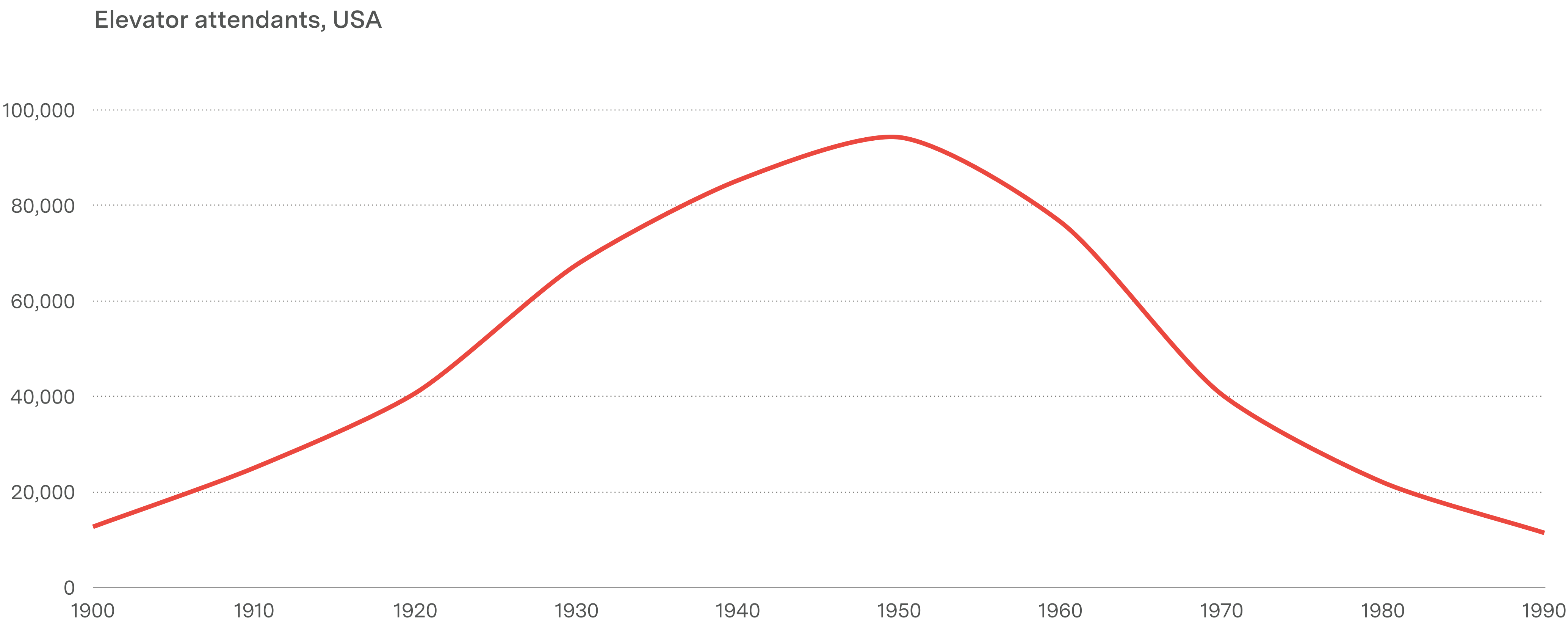
“‘Intelligence’ is whatever machines
haven't done yet”

Larry Tesler, 1970

‘Tech’ is whatever machines
haven’t done yet?

When automation works, it disappears

Otis launched the ‘Autotronic’ automatic elevator in 1950



Source: US Census

About

What matters in tech? What's going on, what might it mean, and what will happen next?

I've spent 20 years analysing mobile, media and technology, and worked in equity research, strategy, consulting and venture capital. I'm now an independent analyst. Mostly, that means trying to work out what questions to ask.

For more, see www.ben-evans.com

Thank you

Benedict Evans

November 2024

www.ben-evans.com
