

COATUE

The AI Revolution

→ **Sri Viswanath**
Vibhor Khanna
Yijia Liang
Coatue

November 2023

Disclaimer

THIS PRESENTATION DOES NOT CONSTITUTE AN OFFER TO SELL OR THE SOLICITATION OF ANY OFFER TO BUY AN INTEREST IN A FUND OR VEHICLE MANAGED BY COATUE MANAGEMENT, L.L.C. ("COATUE").

ANY SUCH OFFER OR SOLICITATION WILL BE MADE ONLY AT THE TIME A QUALIFIED OFFEREE RECEIVES A CONFIDENTIAL PRIVATE OFFERING MEMORANDUM OR OTHER OFFERING DOCUMENT ("CPOM") DESCRIBING THE OFFERING AND RELATED SUBSCRIPTION AGREEMENT.

IN THE CASE OF ANY INCONSISTENCY BETWEEN THE DESCRIPTIONS OR TERMS IN THIS PRESENTATION AND THE CPOM, THE CPOM SHALL CONTROL.

COATUE HAS PROVIDED THIS PRESENTATION SOLELY IN CONNECTION WITH ONGOING OR INTENDED DISCUSSIONS WITH THE PERSON TO WHOM IT HAS BEEN DELIVERED. IN THIS REGARD, THE PERFORMANCE FIGURES, AND CERTAIN OTHER INFORMATION INCLUDED IN THIS PRESENTATION, REQUIRE FURTHER EXPLANATION AND MUST BE DISCUSSED WITH YOUR COATUE REPRESENTATIVES. ACCORDINGLY, PROSPECTIVE INVESTORS SHOULD NOT RELY ON THE INFORMATION IN THIS PRESENTATION ABSENT SUCH DISCUSSIONS.

NO SECURITIES OR SERVICES SHALL BE OFFERED OR SOLD IN ANY JURISDICTION IN WHICH SUCH OFFER, SOLICITATION OR SALE WOULD BE UNLAWFUL UNTIL THE REQUIREMENTS OF THE LAWS OF SUCH JURISDICTION HAVE BEEN SATISFIED. WHILE ALL THE INFORMATION PREPARED IN THIS PRESENTATION IS BELIEVED TO BE ACCURATE, COATUE MAKES NO EXPRESS WARRANTY AS TO THE COMPLETENESS OR ACCURACY NOR CAN IT ACCEPT RESPONSIBILITY FOR ERRORS, APPEARING IN THE PRESENTATION. NEITHER COATUE NOR ITS AFFILIATES ASSUMES ANY DUTY TO UPDATE THE INFORMATION CONTAINED HEREIN FOR SUBSEQUENT CHANGES OF ANY KIND AND THERE IS NO ASSURANCE THAT THE POLICIES, STRATEGIES OR APPROACHES DISCUSSED HEREIN WILL NOT CHANGE.

ANY PROJECTIONS, MARKET OUTLOOKS OR ESTIMATES IN THIS PRESENTATION ARE FORWARD-LOOKING STATEMENTS AND ARE BASED UPON CERTAIN ASSUMPTIONS. OTHER EVENTS WHICH WERE NOT TAKEN INTO ACCOUNT MAY OCCUR AND MAY SIGNIFICANTLY AFFECT INVESTMENT RETURNS OR PERFORMANCE.

ANY PROJECTIONS, OUTLOOKS OR ASSUMPTIONS SHOULD NOT BE CONSTRUED TO BE INDICATIVE OF THE ACTUAL EVENTS WHICH WILL OCCUR. HISTORICAL RETURNS ARE NOT PREDICTIVE OF FUTURE RESULTS.

THE INFORMATION PROVIDED HEREIN, INCLUDING, WITHOUT LIMITATION, INVESTMENT STRATEGIES, INVESTMENT RESTRICTIONS AND PARAMETERS, PROCEDURES, POLICIES, AND INVESTMENT AND OTHER PERSONNEL MAY BE CHANGED, MODIFIED, TERMINATED OR SUPPLEMENTED AT ANY TIME, WITHOUT NOTICE.

THIS PRESENTATION IS CONFIDENTIAL AND NOT INTENDED FOR PUBLIC USE OR DISTRIBUTION. THIS PRESENTATION HAS NOT BEEN APPROVED BY THE U.S. SECURITIES AND EXCHANGE COMMISSION OR ANY OTHER FEDERAL OR STATE REGULATOR.

YOU SHOULD CAREFULLY READ THE IMPORTANT DISCLOSURES IN THE APPENDIX.

Key Topics

→ **Where we are in AI today**

→ AI could break through the hype and improve our world

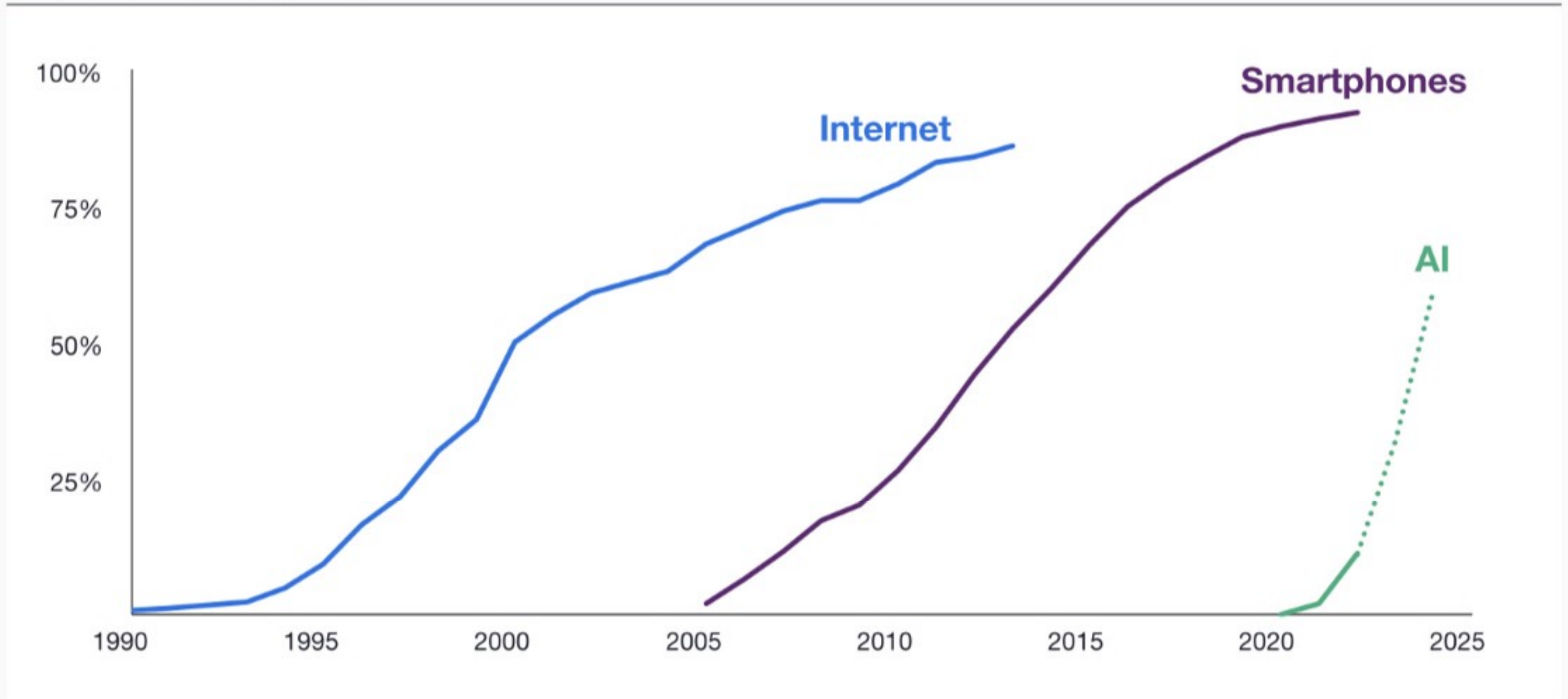
→ We believe open-source is the lifeblood of AI

→ AI is transforming the tech ecosystem

→ Coatue view: the best of AI is yet to come

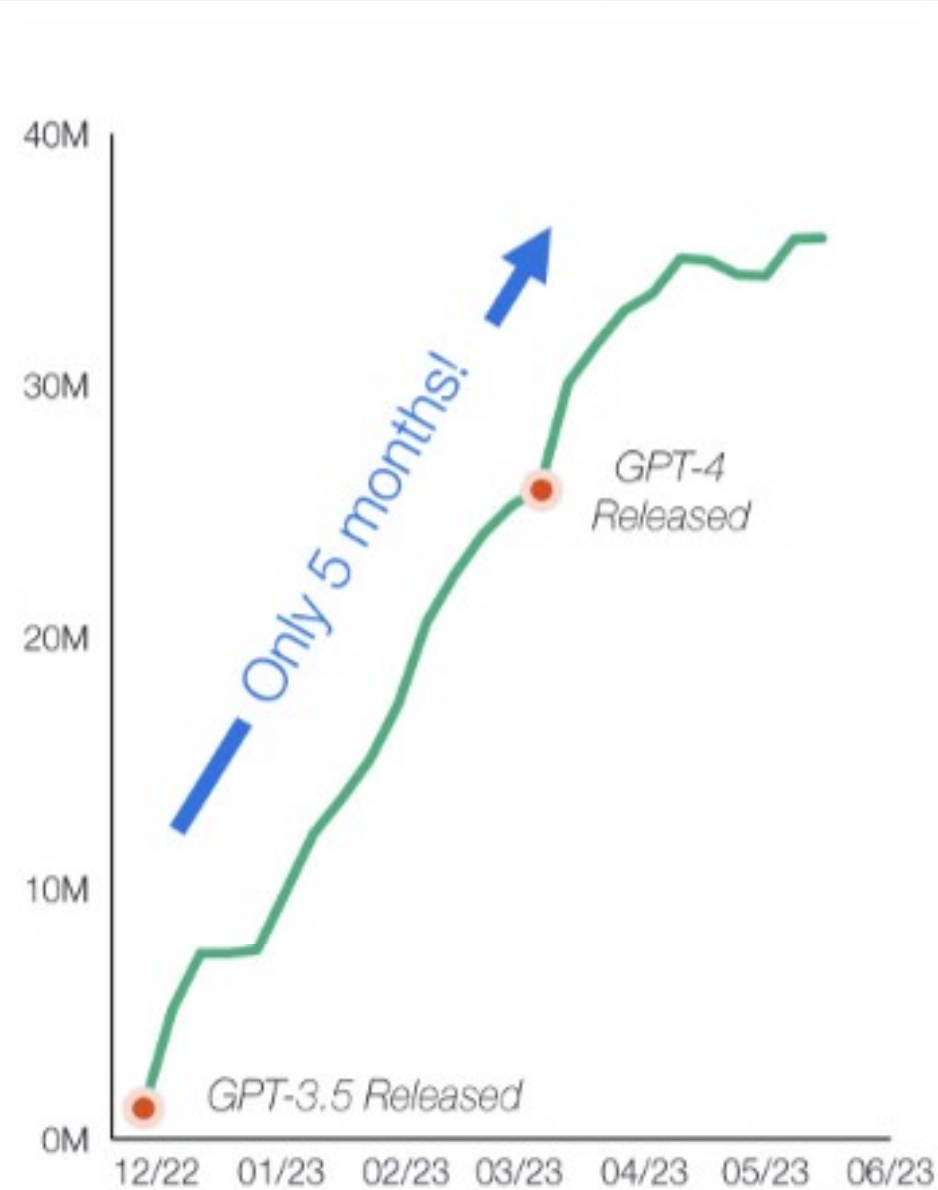
We're at Day 1 of AI...and riding on top of past waves

→ % US Technology Adoption

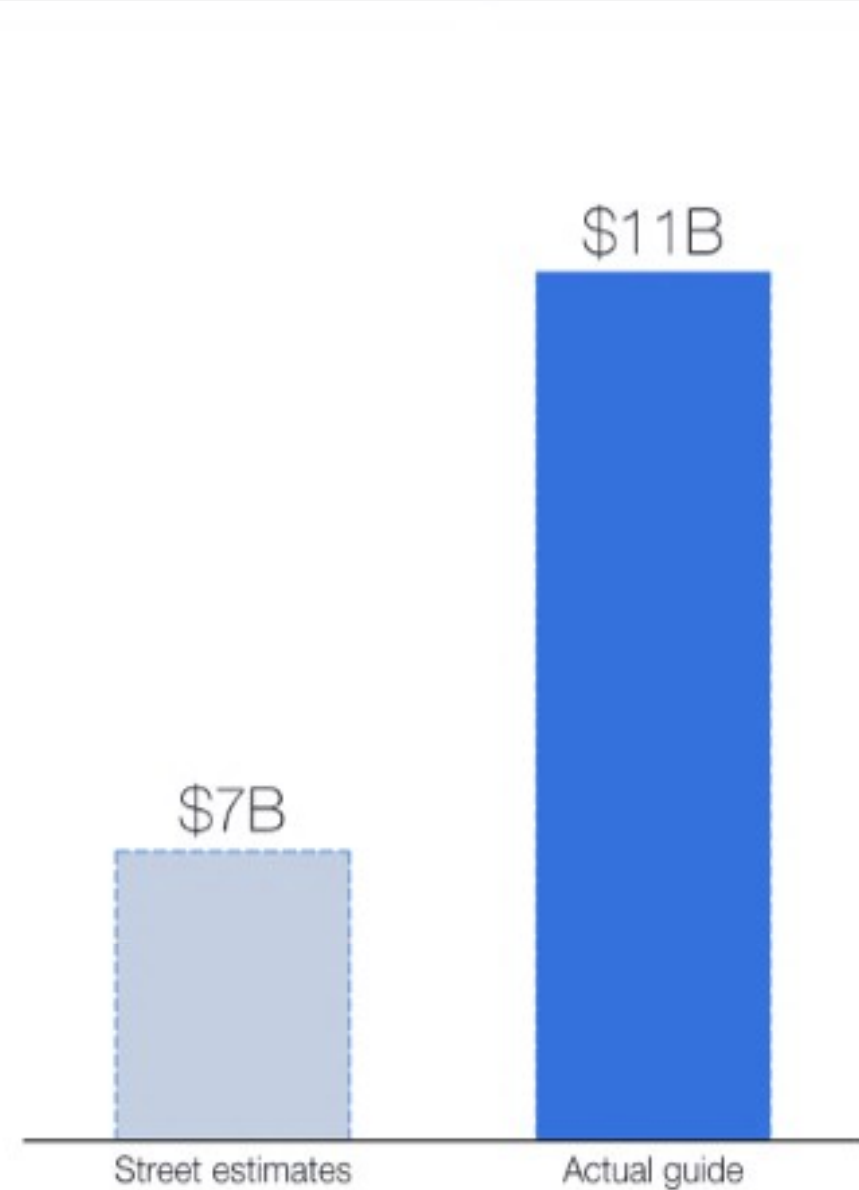


In the first half of 2023, the AI ecosystem exploded! (1/2)

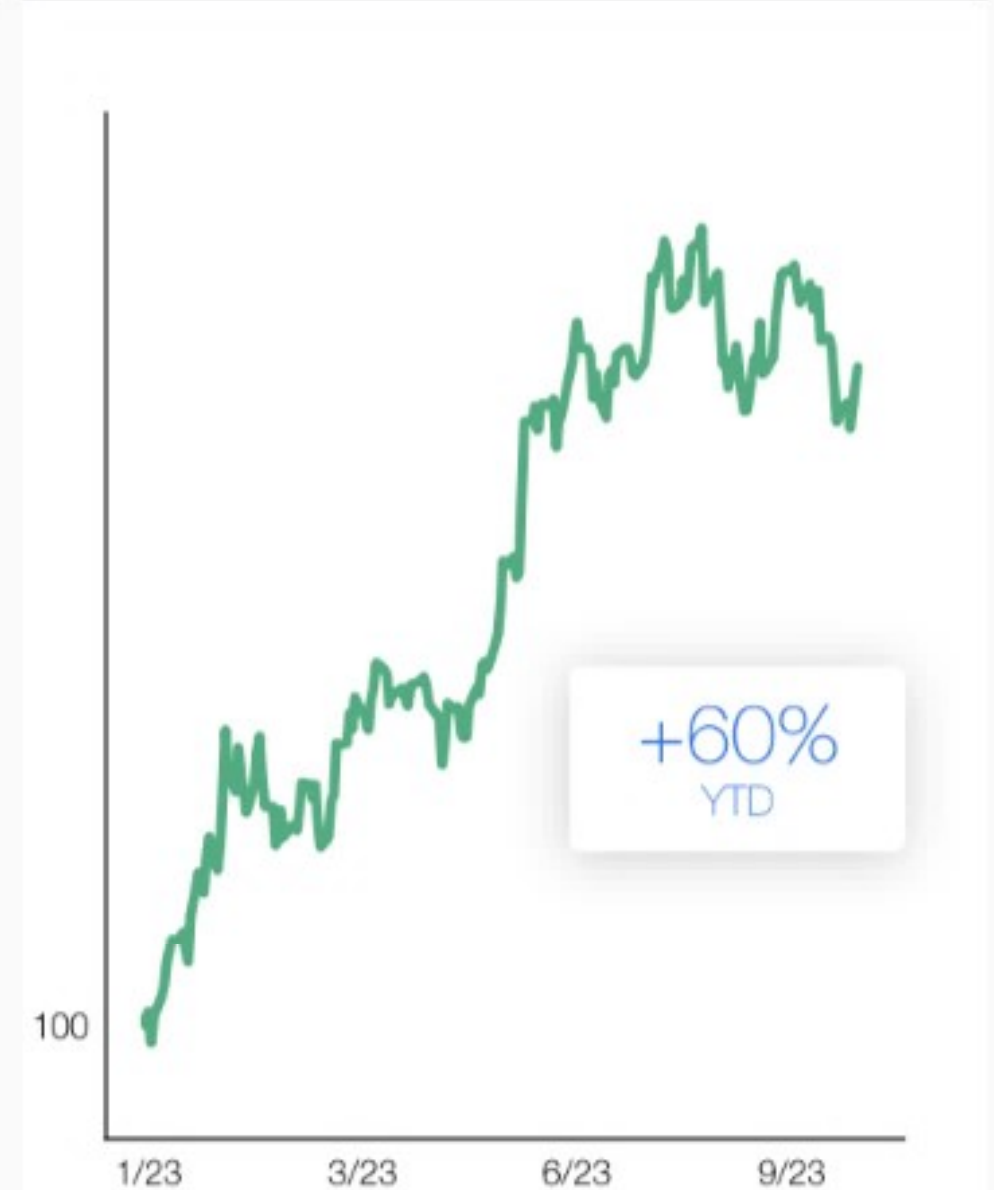
→ ChatGPT weekly unique visits



→ NVDA 2Q Earnings

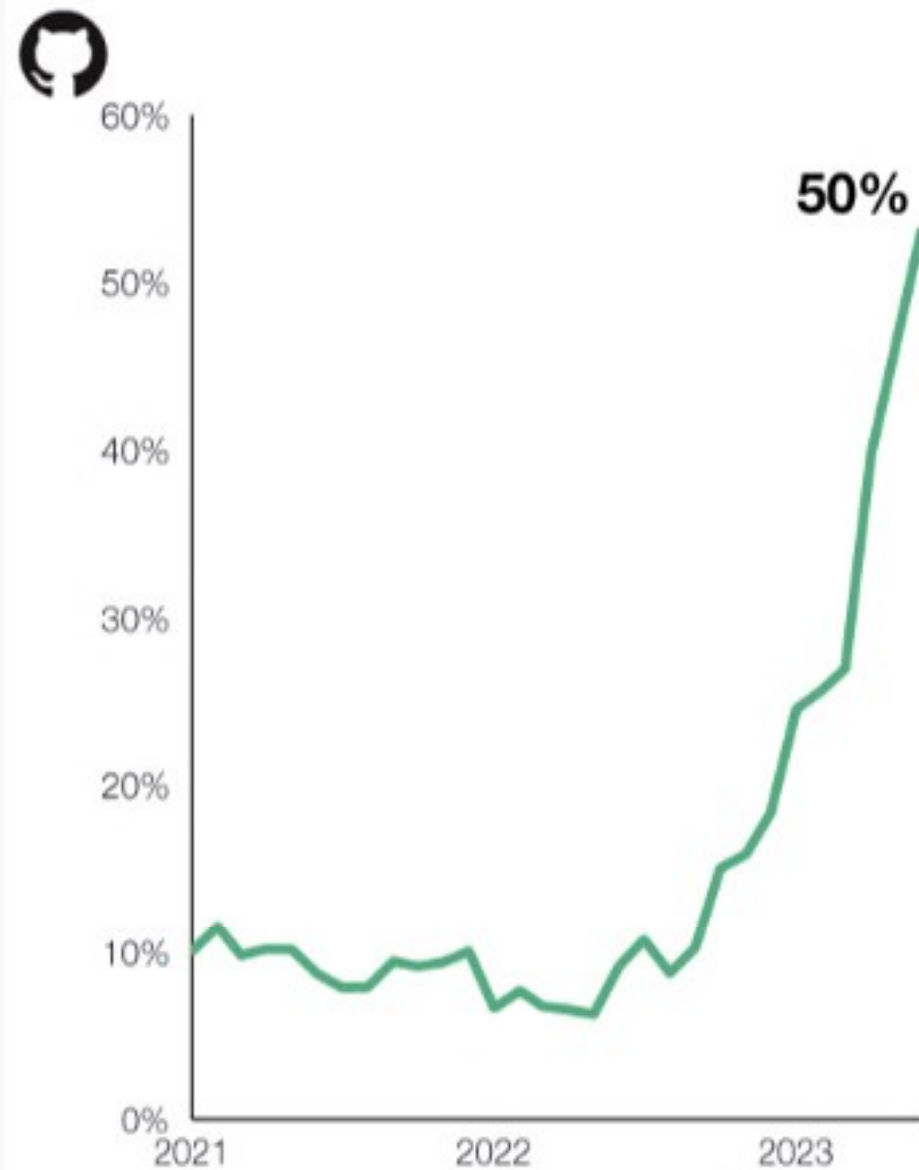


→ AI Basket Performance YTD

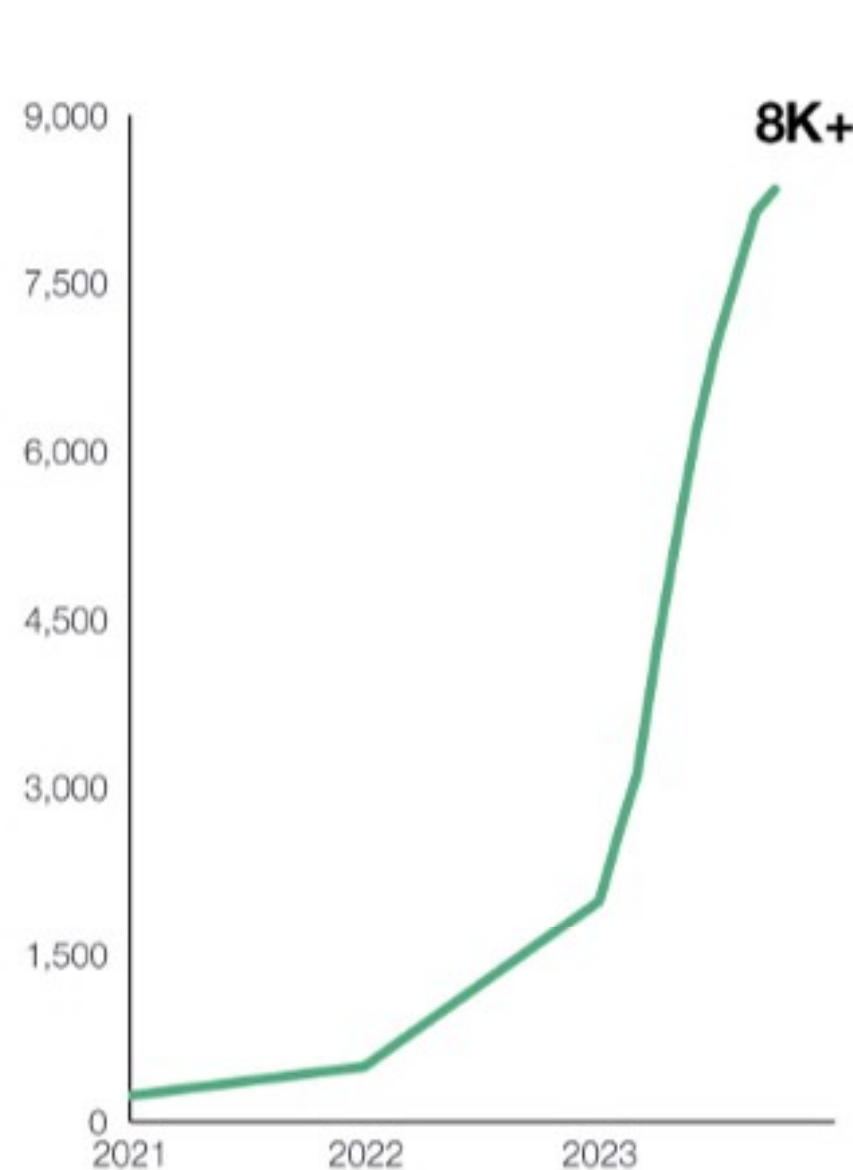


In the first half of 2023, the AI ecosystem exploded! (2/2)

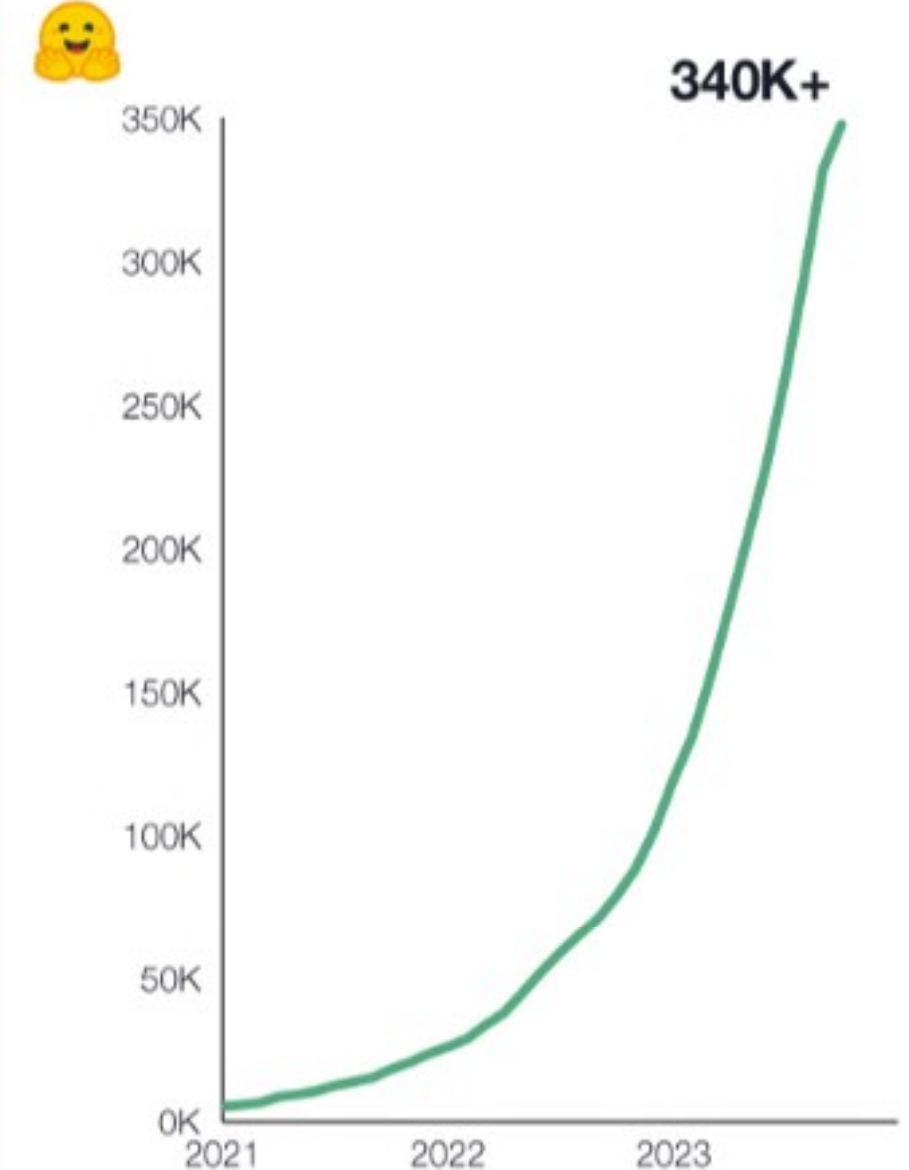
→ % of trending Github repos in AI/ML



→ New AI applications over time



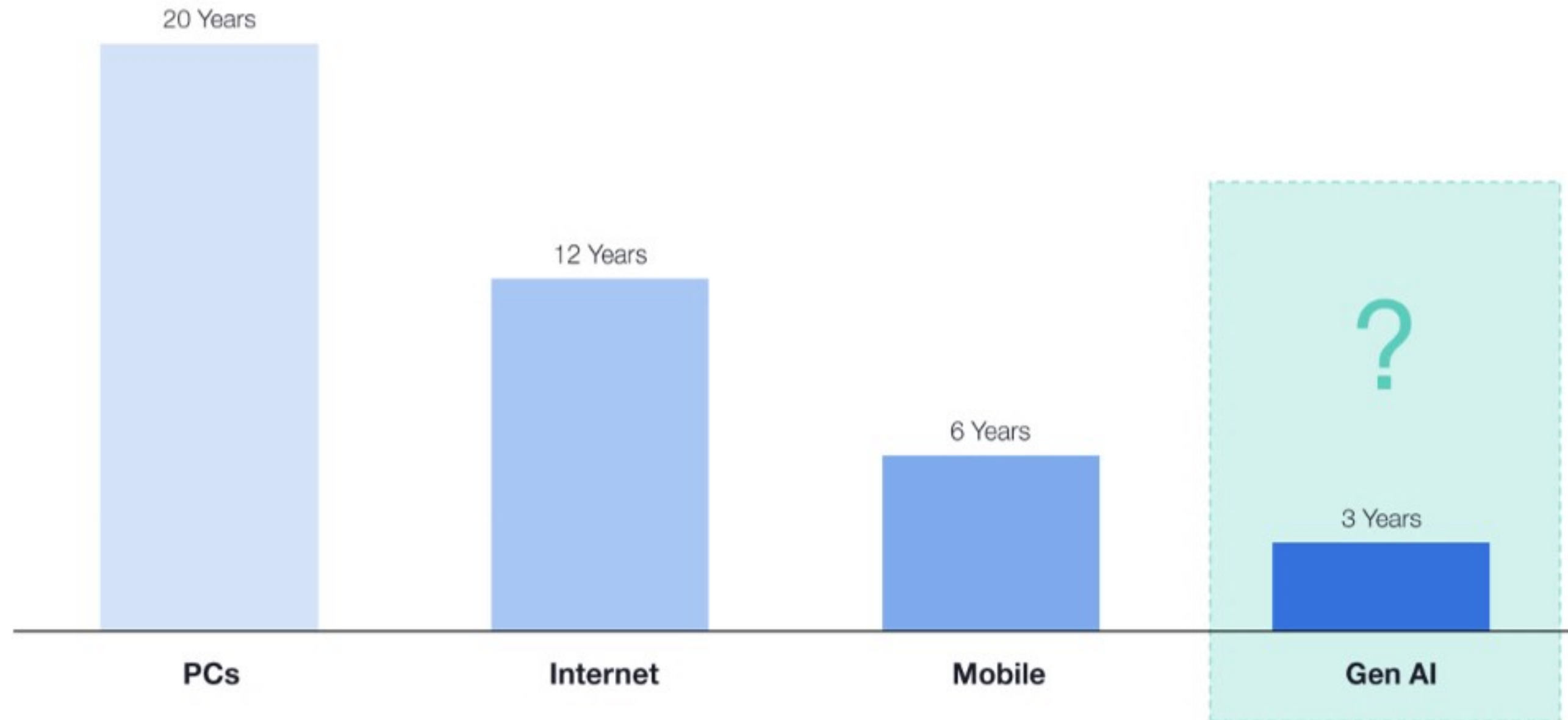
→ Number of models on Hugging Face



Adoption has been twice as fast with each platform shift

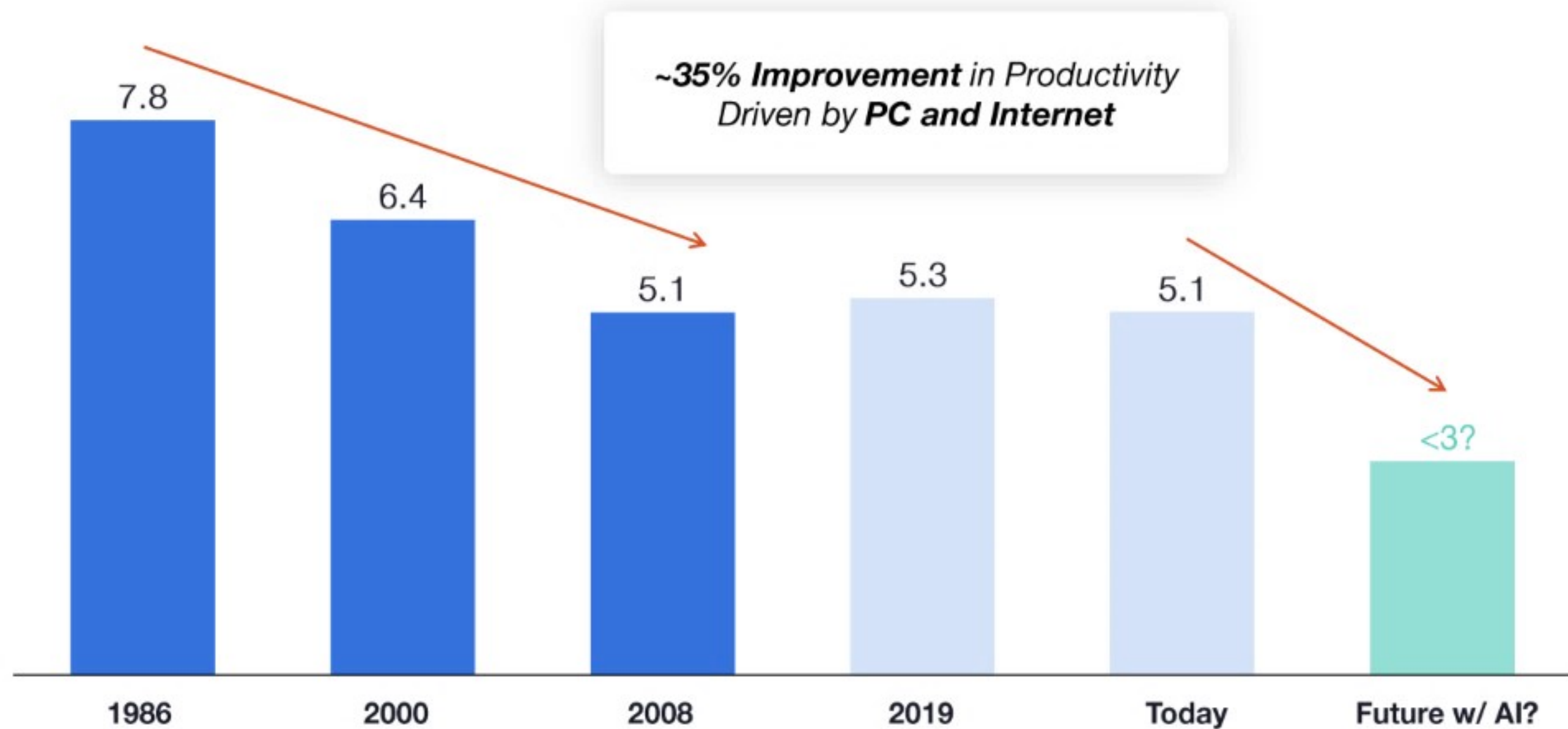
→ **Halving of penetration time with new technology waves**

of years to reach 50% user penetration in the US



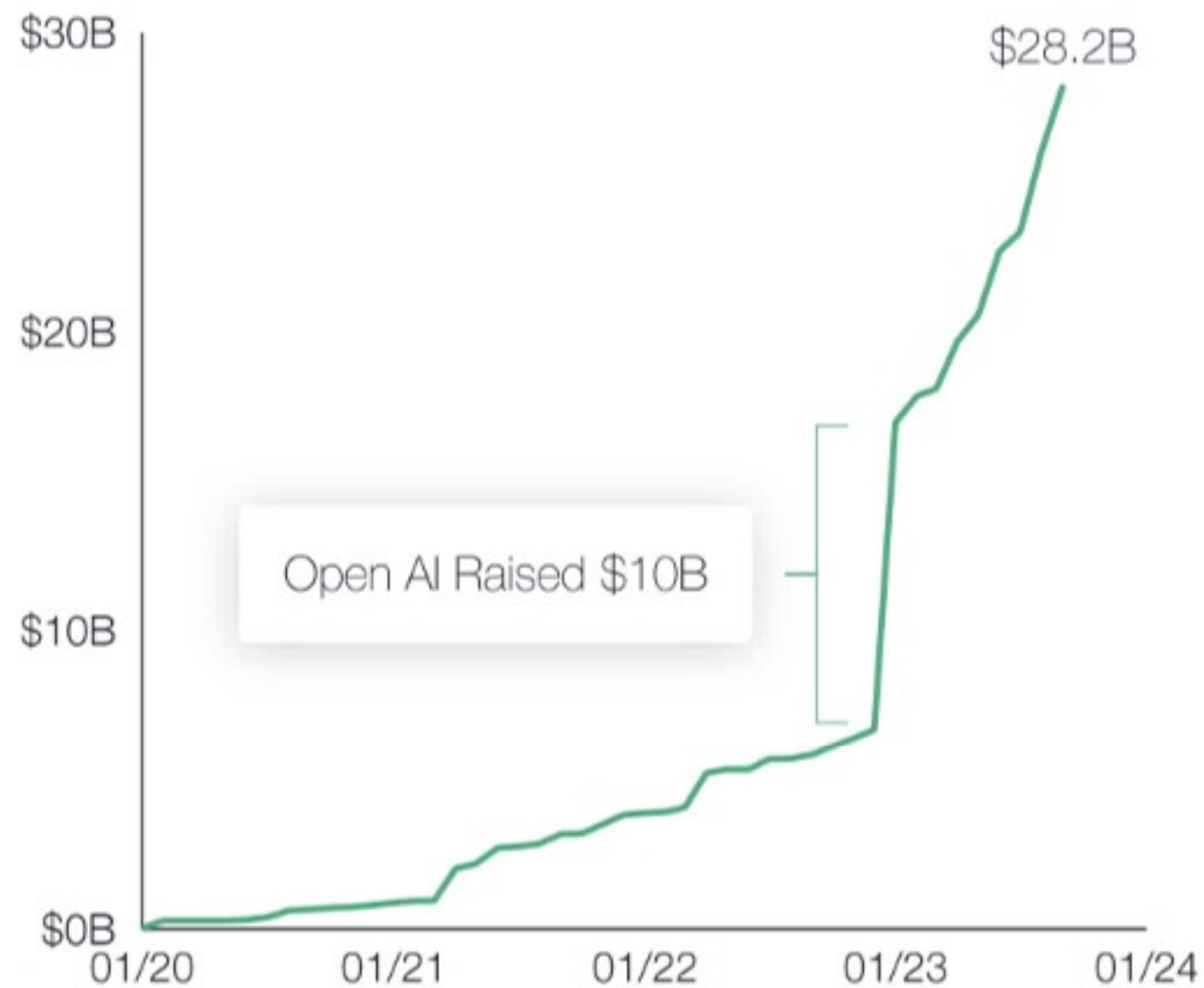
AI has potential to drive the economy for years to come

→ # of employees per \$1M of revenue (inflation adjusted – S&P 500 companies)

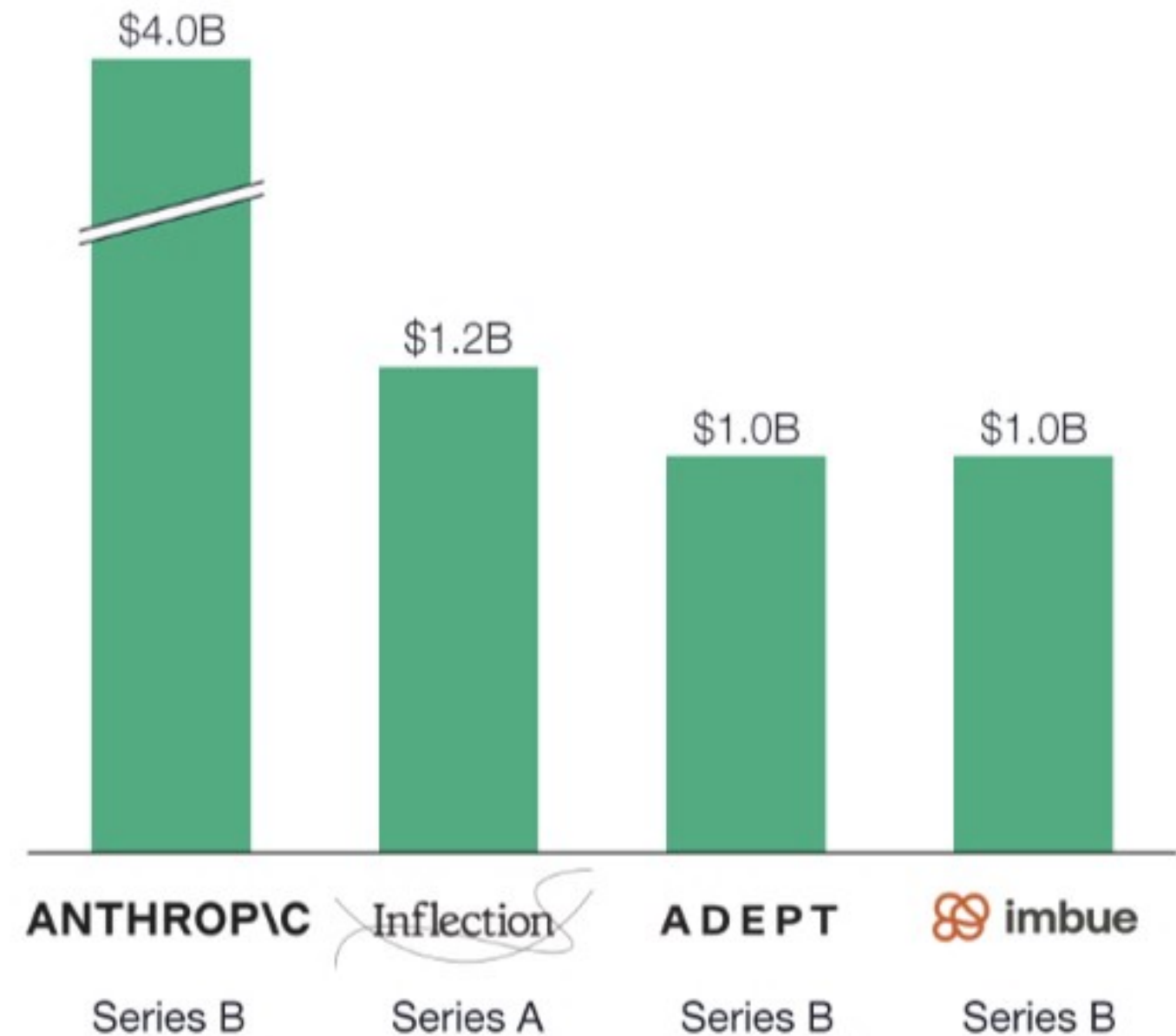


We've seen massive investment in AI

→ Cumulative funding in private AI companies since 2020



→ Valuation before public product launch

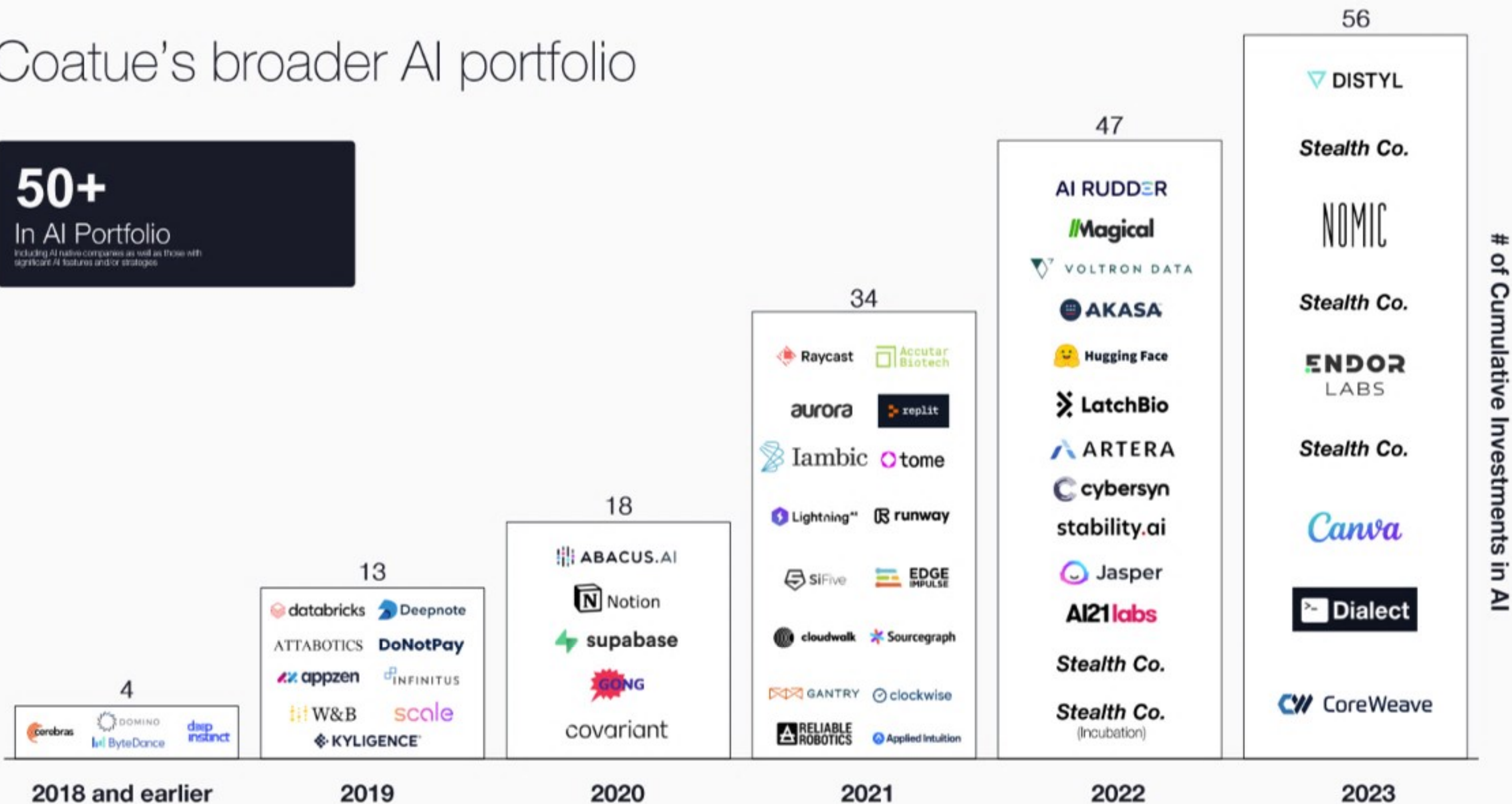


Coatue's broader AI portfolio

50+

In AI Portfolio

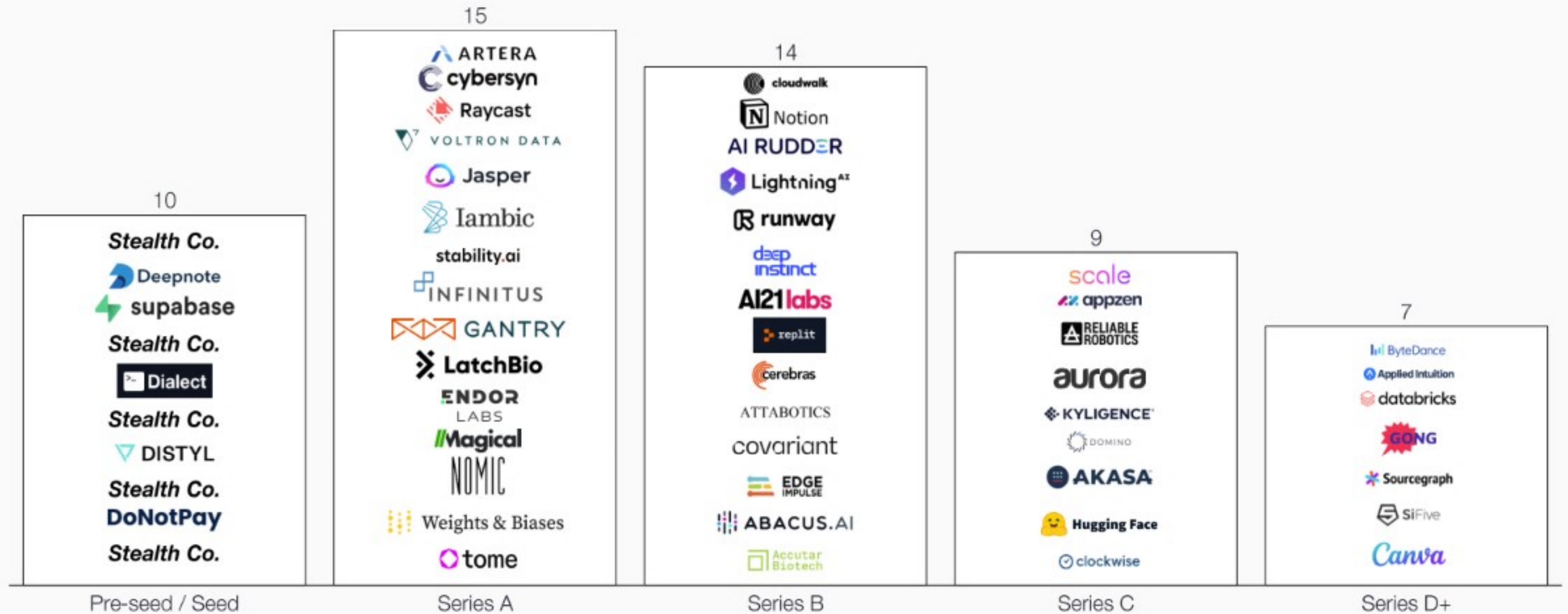
Including AI native companies as well as those with significant AI features and/or strategies



of Cumulative Investments in AI

We've backed AI founders across stages

of Coatue AI portfolio companies based on round when Coatue first invested



Key Topics

→ Where we are in AI today

→ **AI could break through the hype and improve our world**

→ We believe open-source is the lifeblood of AI

→ AI is transforming the tech ecosystem

→ Coatue view: the best of AI is yet to come

Coatue View: AI is not just hype

Characteristic of hype cycles

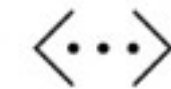
Coatue view

Historic Examples

Value Accrual

Value accrual misaligned with investment

- *Comparison:* Historic cycles like fiber and cloud represent two different outcomes for underlying infrastructure
- Most investment in AI today is within the model layer, but **it's too early** to declare who will be the **AI model winners**



1990s Fiber

Capabilities

Overestimating timeline & capabilities of technology

- *Comparison:* Self-driving cars finally arriving, but after 15+ years of work
- **AI is already useful within ~5 years.** We are in early innings of adoption and expect models to improve, but AI regulation is likely and is a challenge to implement well

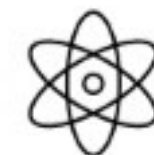


Autonomous vehicles

Utility

Lack of widespread utility due to maturity of technology

- *Comparison:* Quantum computing hype was promising in theory but has not yet proven widespread practical utility
- **AI already proving significant utility across domains**



Quantum computing

Comparison: Some enabling technologies become a public good

Fiber Infrastructure

1990s

Telco Co's raised \$1.6T of equity & \$600B of debt

Bandwidth costs decreased 90% within 4 years



Most of these companies no longer exist today...

...While cloud infrastructure became a huge market

Cloud Infrastructure

2010s


Entirely new computing paradigm

Most were already public companies with resources to build out data centers






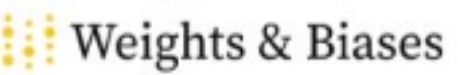






Big 4 hyperscalers generate ~\$150B+ in revenue annually

Where did value accrue in the cloud stack?

Cloud Stack	Example Companies	Est. TAM	% of total in stack ¹
SaaS Apps	 	~\$260B	40%
PaaS	 	~\$140B	22%
IaaS	 	~\$200B	30%
Cloud Semis	 	~\$50B	8%

In AI, funding concentrated in model layer for now

AI Stack	Example Companies	Total funding	% of total in stack ¹
Apps	 	~\$5B	17%
Models	 	~\$17B	60%
AI Ops	 	~\$1B	4%
AI Cloud	 	~\$4B	13%
AI Semis	 	~\$2B	6%

The jury is out on which model companies will win

AI Models 2020s

Private AI model companies raised ~\$17B in venture funding since 2020

Open-source models are becoming more ubiquitous

?

ADEPT

 OpenAI

ANTHROPIC

 imbue

 Inflection

co:here

 ALEPH
ALPHA

Outcome TBD

Comparison: AI advancing much faster than previous waves

Levels of autonomy	Autonomous vehicles	Generative AI
L5	Fully autonomous	Superhuman reasoning & perception
L4	Highly autonomous	AI autopilots for complex tasks
L3	Self-driving with light intervention ¹	AI co-pilot for skilled labor
L2	Tesla autopilot	Supporting humans with basic tasks
L1	Cruise control	Generating basic content

Timeline annotations:

- 15 Years: From L1 to L4
- ~5 Years: From L2 to L3

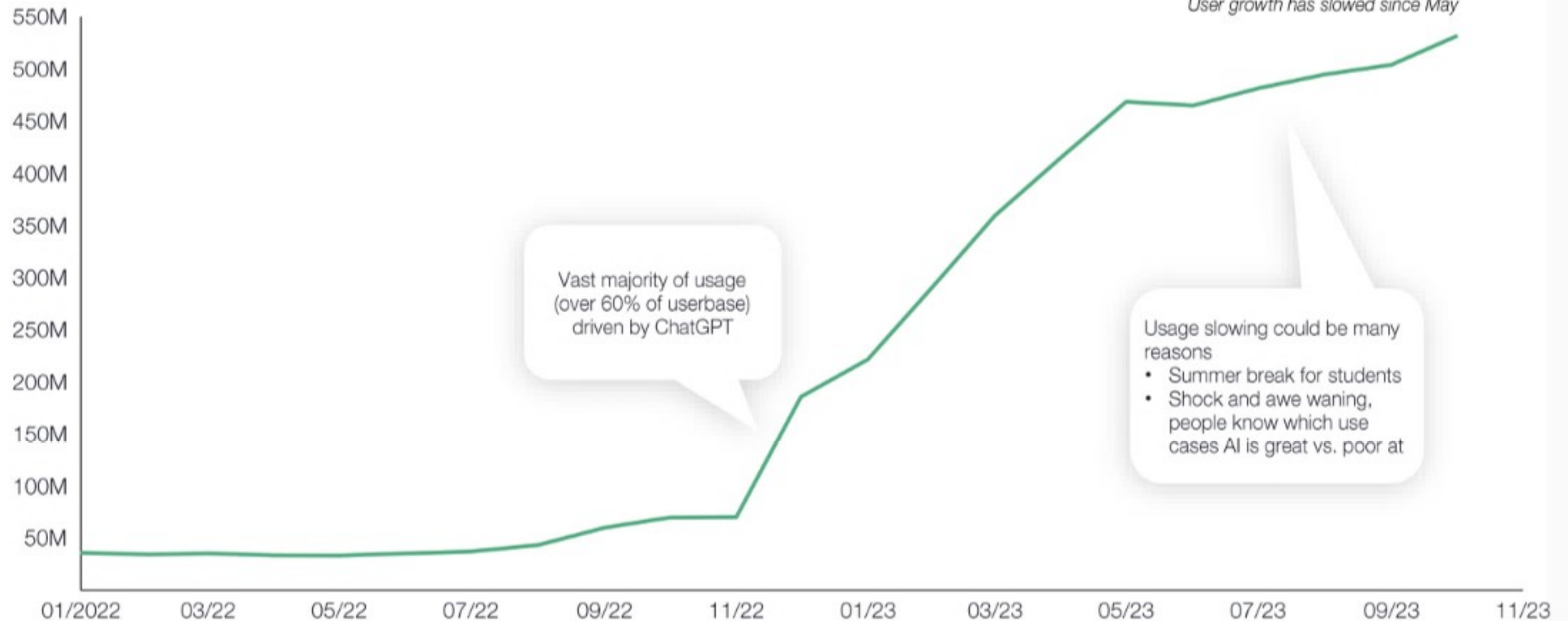
Generative AI examples:

- GitHub Copilot
- replit
- ChatGPT
- Speak
- Midjourney

Has initial AI enthusiasm slowed down?

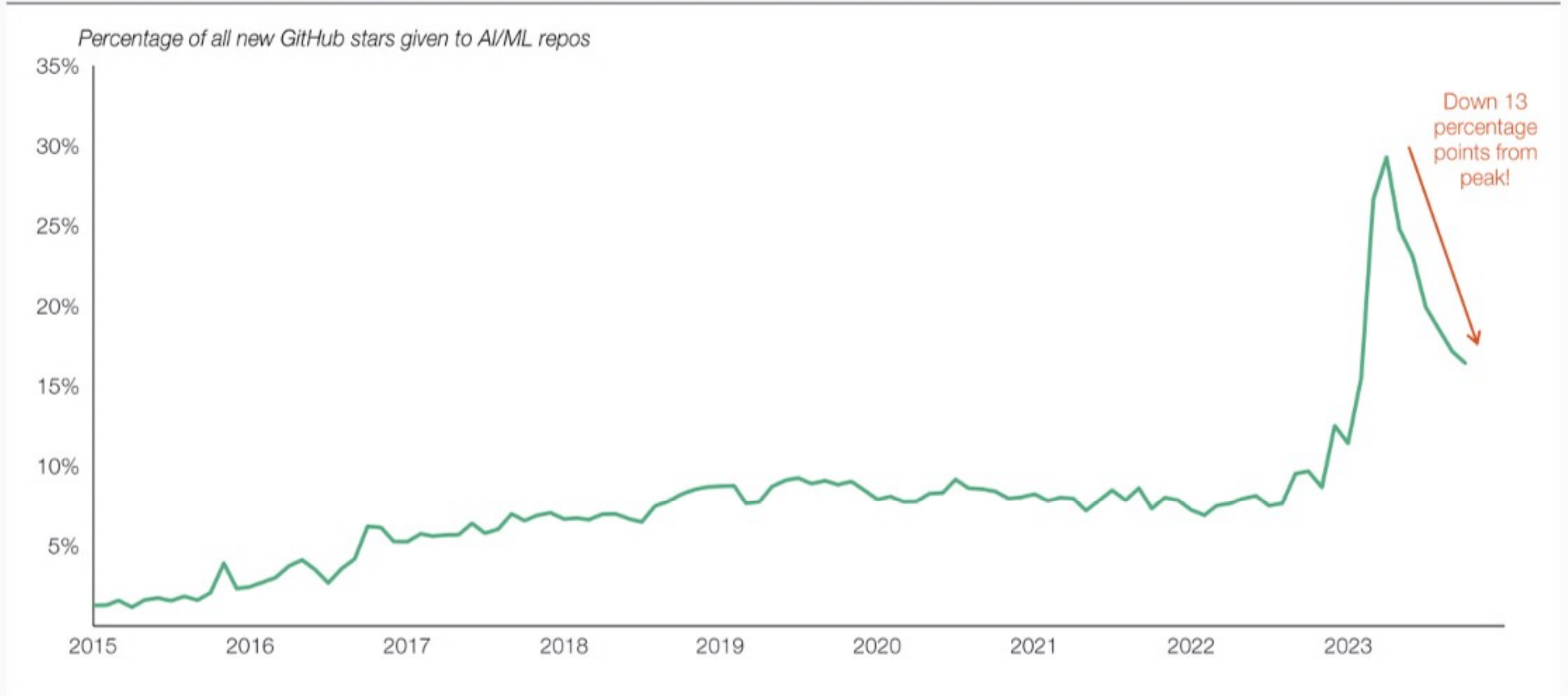
→ AI usage worldwide started flattening during summer

Approximate global AI applications users count by month



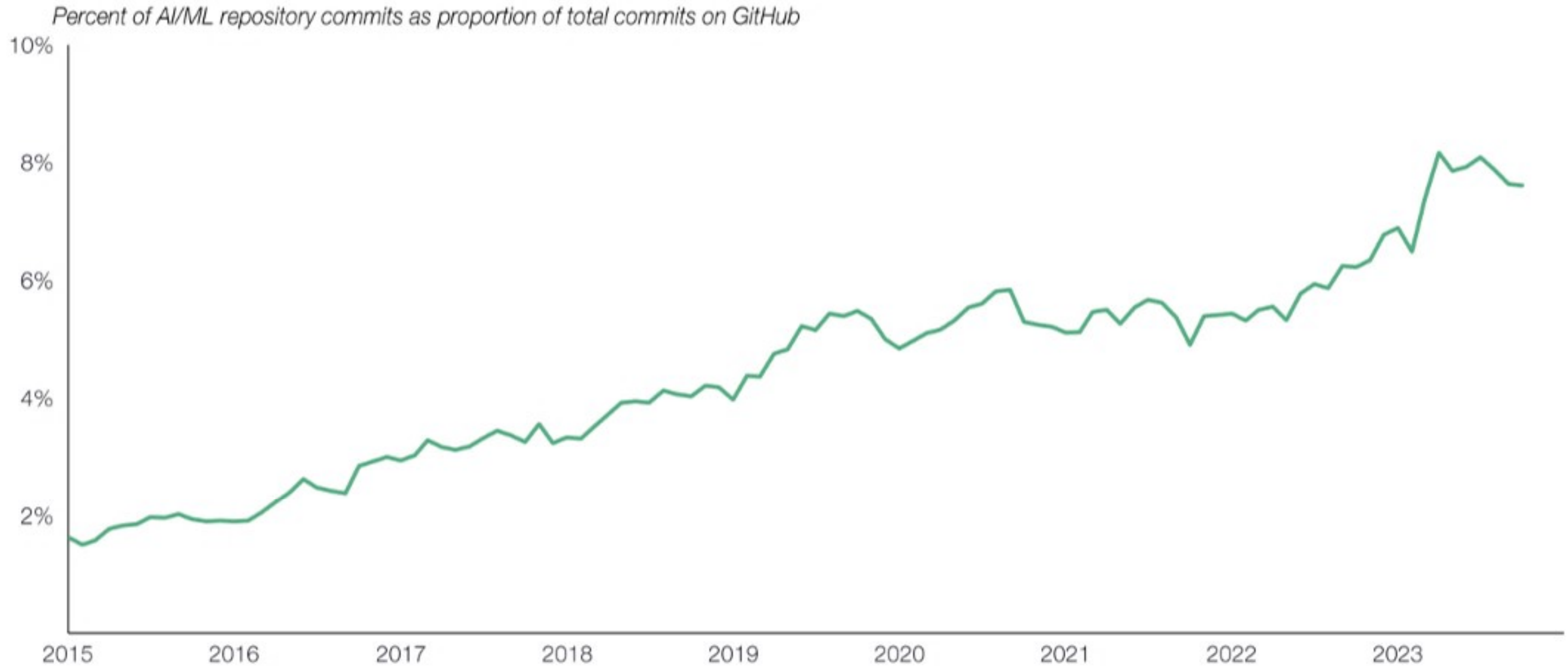
Early developer excitement waning and washing out AI “tourists”

→ **Excitement in AI/ML on GitHub, measured by stars, has declined since April 2023**



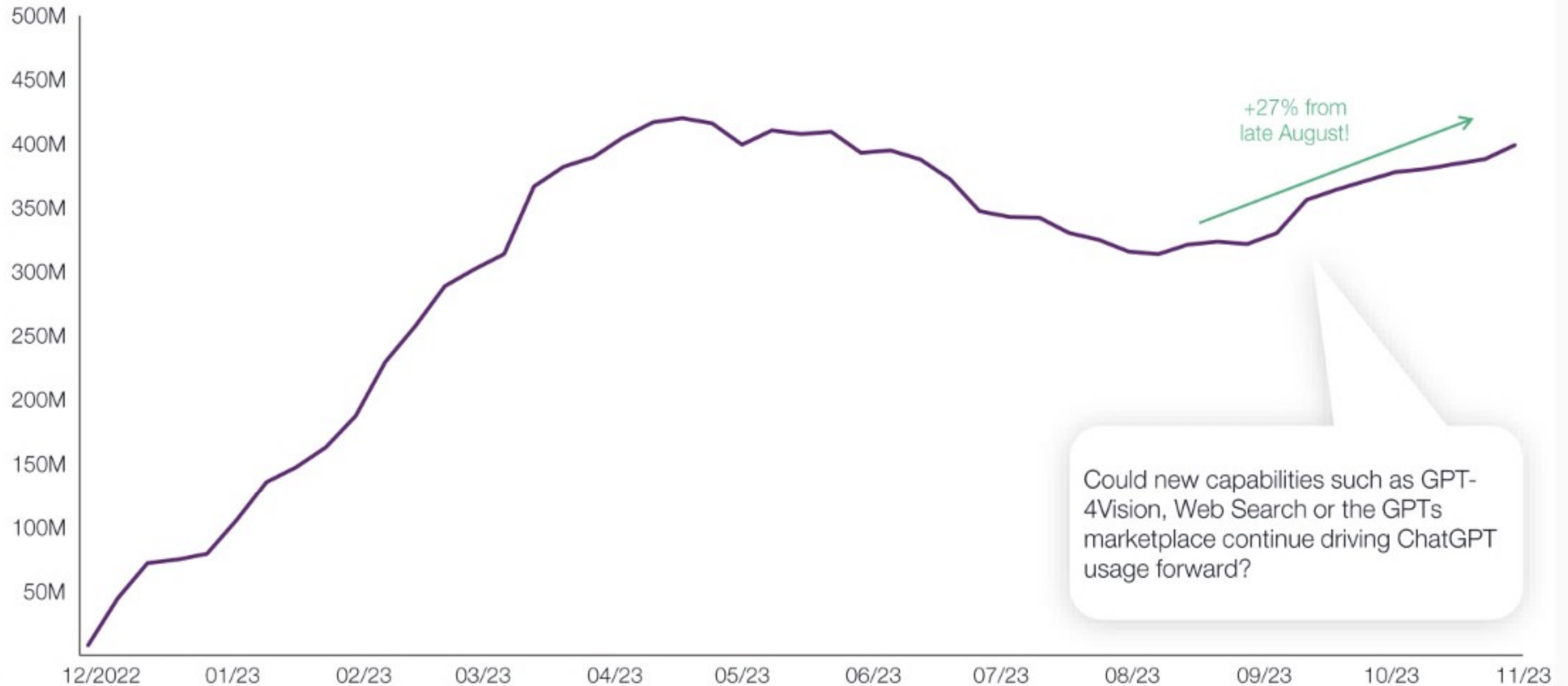
Serious AI builders remain

→ **AI/ML commits have not declined as much since April 2023**



ChatGPT usage has rebounded, new capabilities have shipped

→ ChatGPT weekly web visits worldwide since launch

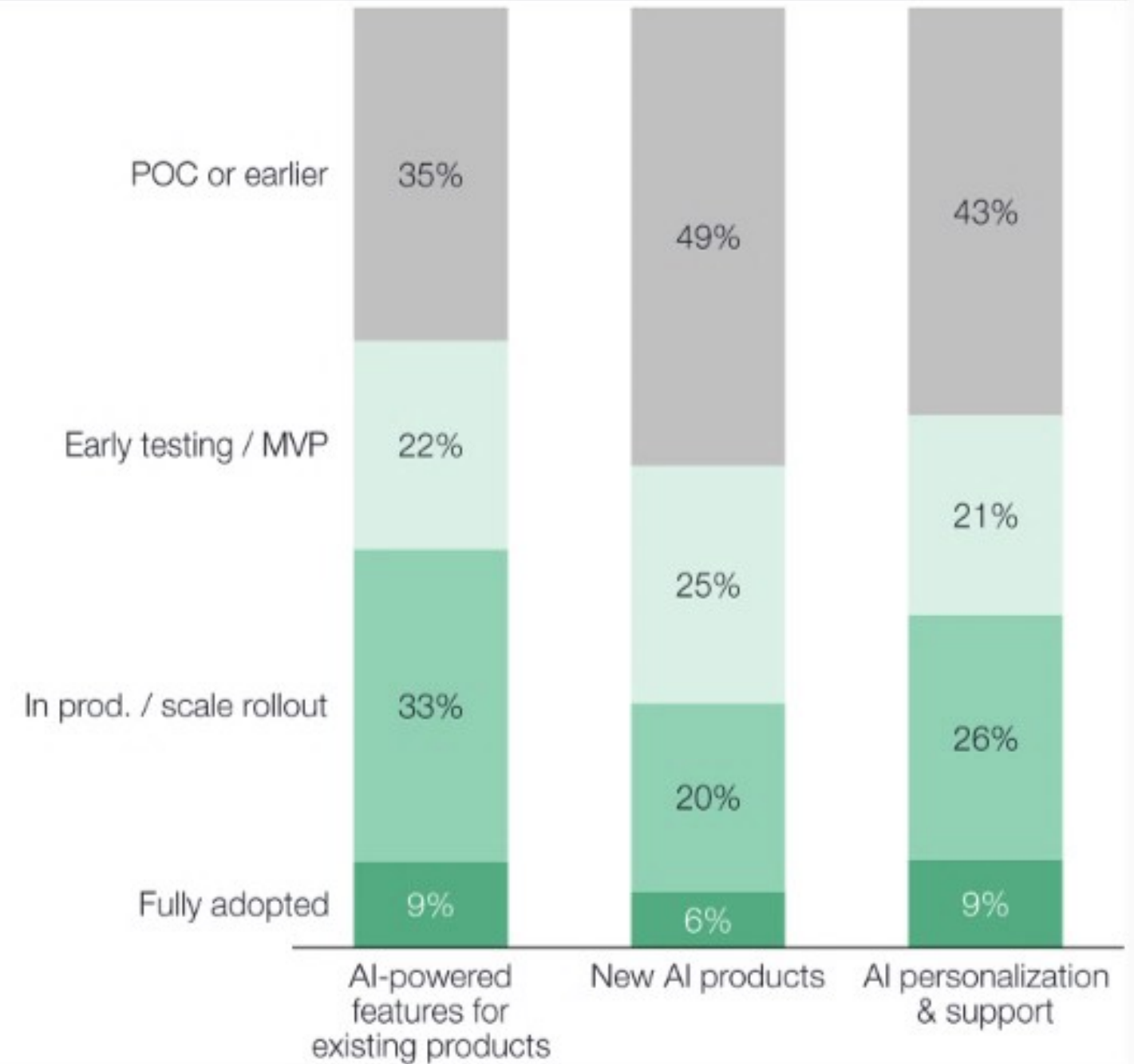
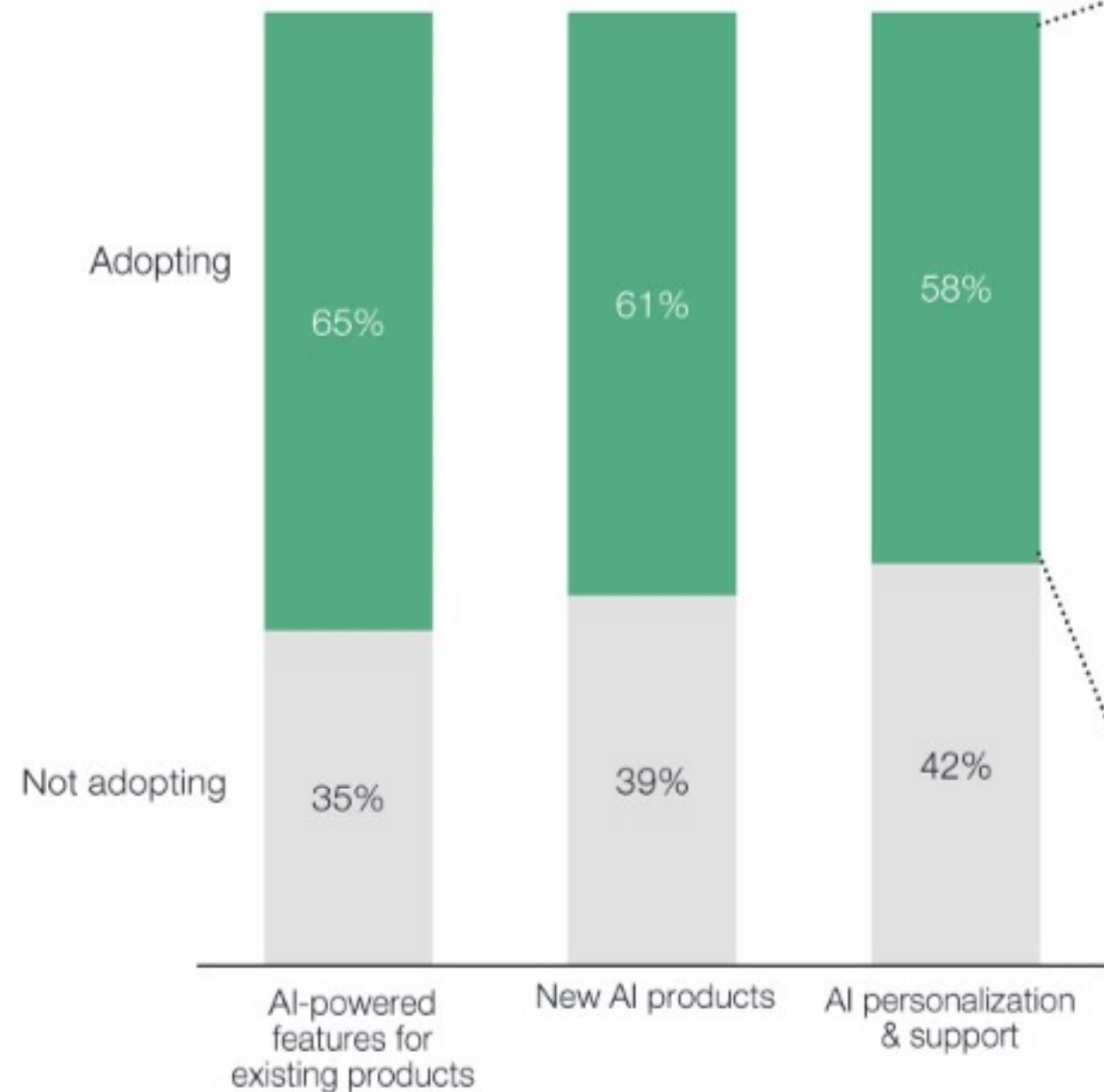


Could new capabilities such as GPT-4Vision, Web Search or the GPTs marketplace continue driving ChatGPT usage forward?

We believe it's first innings of enterprise AI adoption

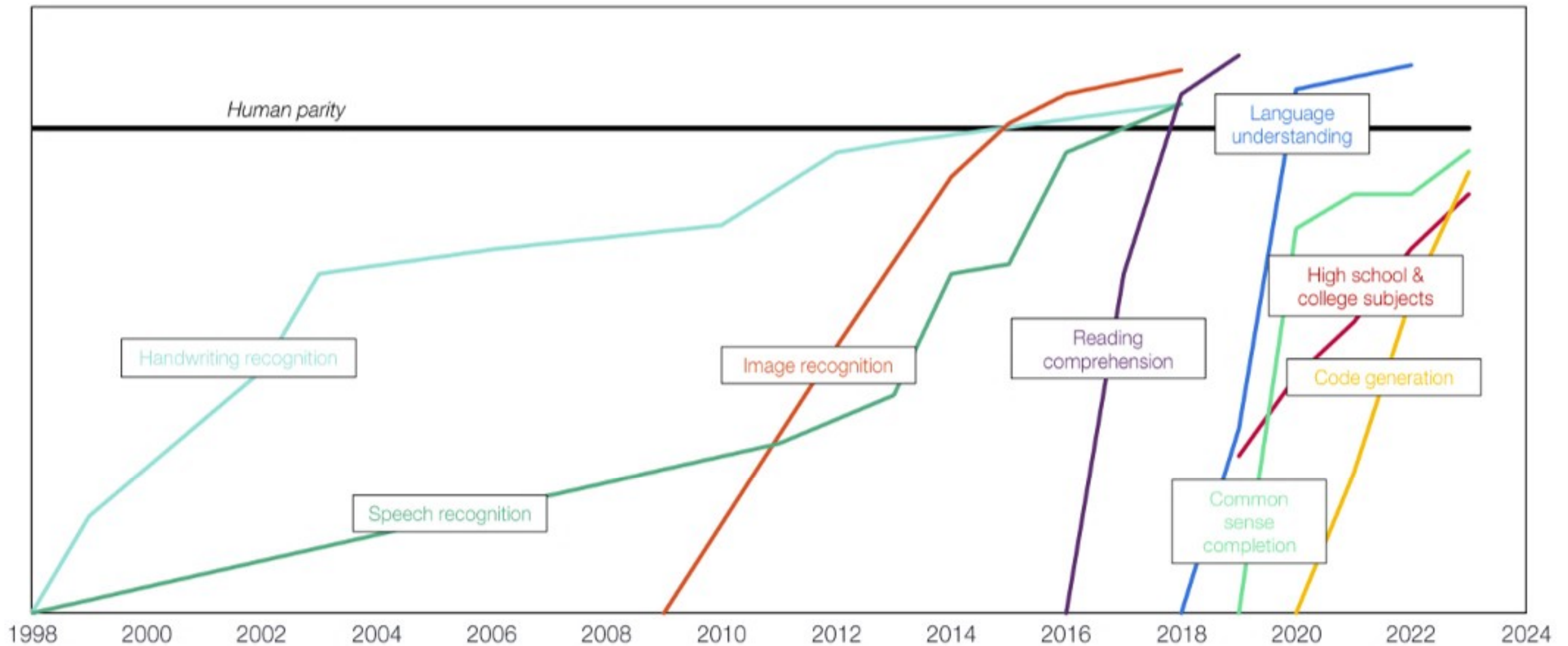
→ ~60%+ of surveyed enterprises plan to adopt AI

→ Less than 10% have fully adopted; timeline is long



We are optimistic: AI is getting better, faster

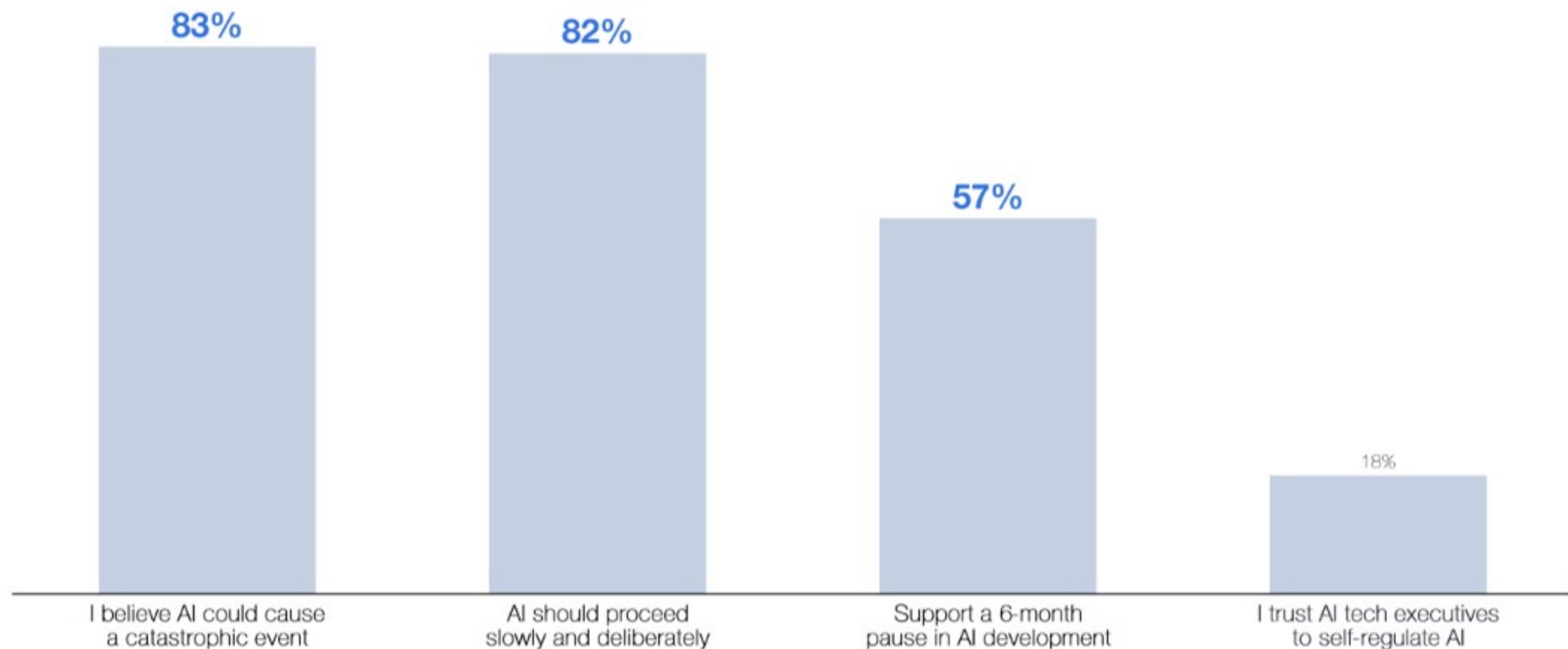
→ Speed for models to reach human level accuracy on benchmarks has decreased



AI regulation may be more likely than most think

→ **Initial polls suggest many Americans concerned about AI**

% of surveyed respondents, AI Policy Institute



EU's AI Act is among first examples of regulation

→ Most models fall short of meeting requirements from EU AI Act, per Stanford study

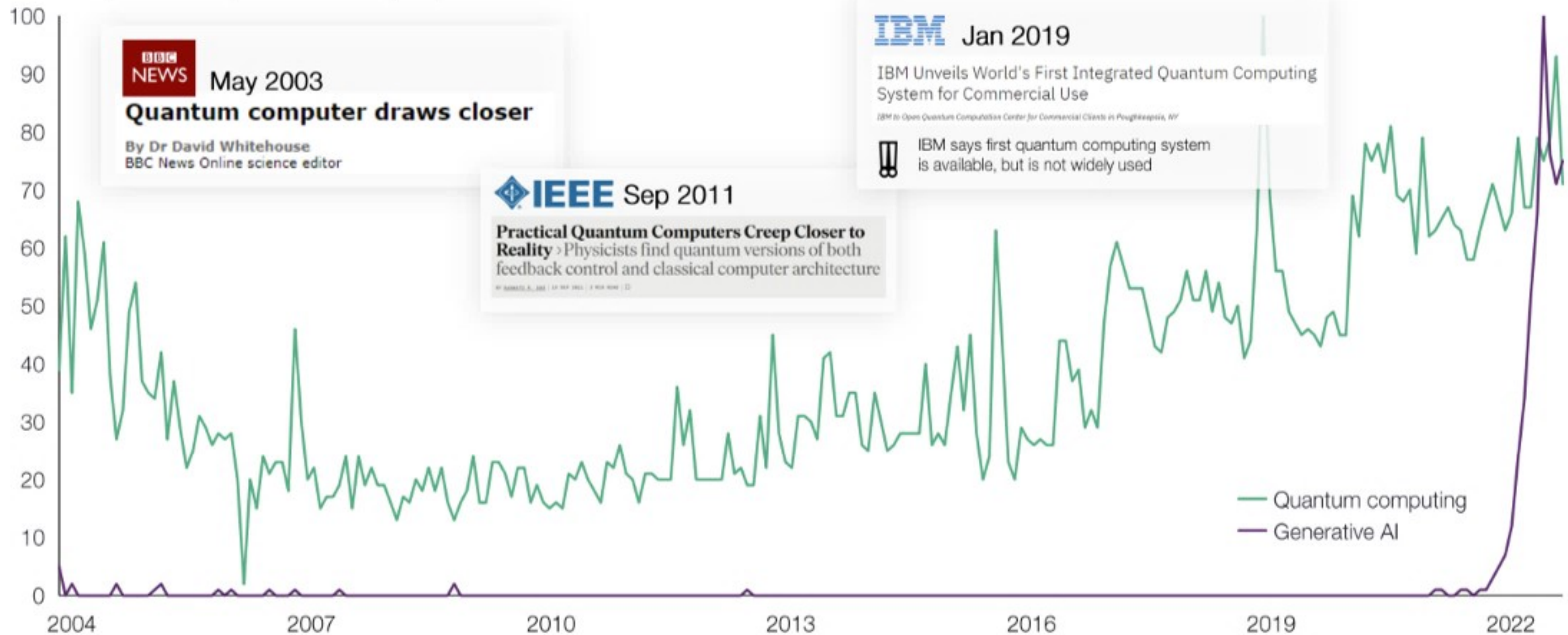


	GPT-4	Cohere Command	Claude v1	LLAMA-1	PaLM-2	BLOOM
Draft EU AI Act requirements <i>Non-exhaustive list</i>						
Data sources <i>"Description of the data sources used in the development of the foundation model."</i>						
Compute <i>"Description of the training resources used by the foundation model including compute required, training time, and other information related to the size and power of the model."</i>						
Energy <i>"Design and develop the foundation model, making use of applicable standards to reduce energy use, resource use and waste, as well as to increase energy efficiency."</i>						
Capabilities & limitations <i>"Description of the capabilities and limitations of the foundation model."</i>						
Risk & mitigations <i>"The reasonably foreseeable risks and the measures that have been taken to mitigate them as well as remaining non-mitigated risks with an explanation on the reason why they cannot be mitigated."</i>						
Data copyright <i>"Without prejudice to national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law."</i>						

Comparison: Is AI hype justified with real utility?

→ Quantum computing interest has steadily grown without real world proof points...what about Generative AI?

Google trend ranking for "Quantum computing" vs "Generative AI"



AI is delivering game-changing value

55%

Time saved for developers using Github Copilot



90%

Time saved from editing video on Runway



45%

Reduction in human-answered customer support requests



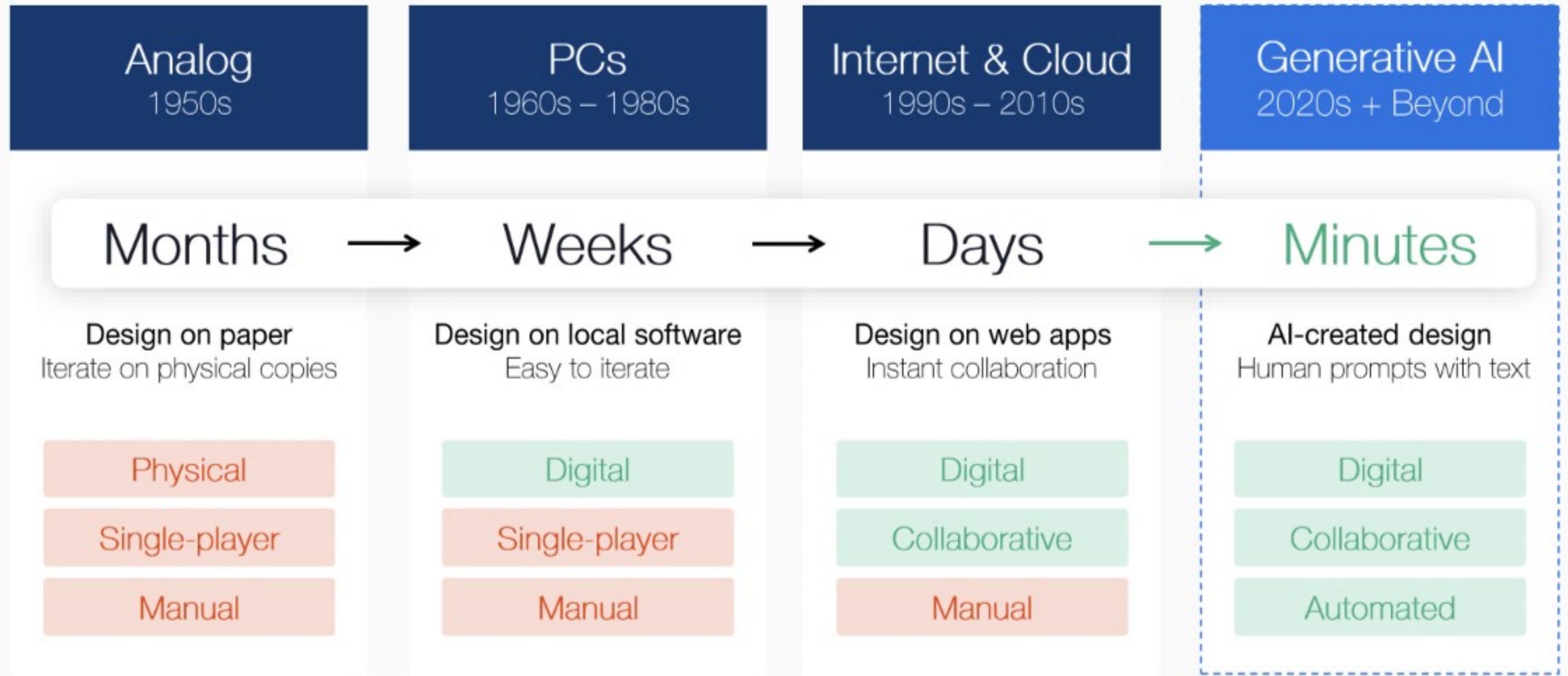
79%

AI chat rated higher quality vs. physician responses



AI poised to improve productivity significantly!

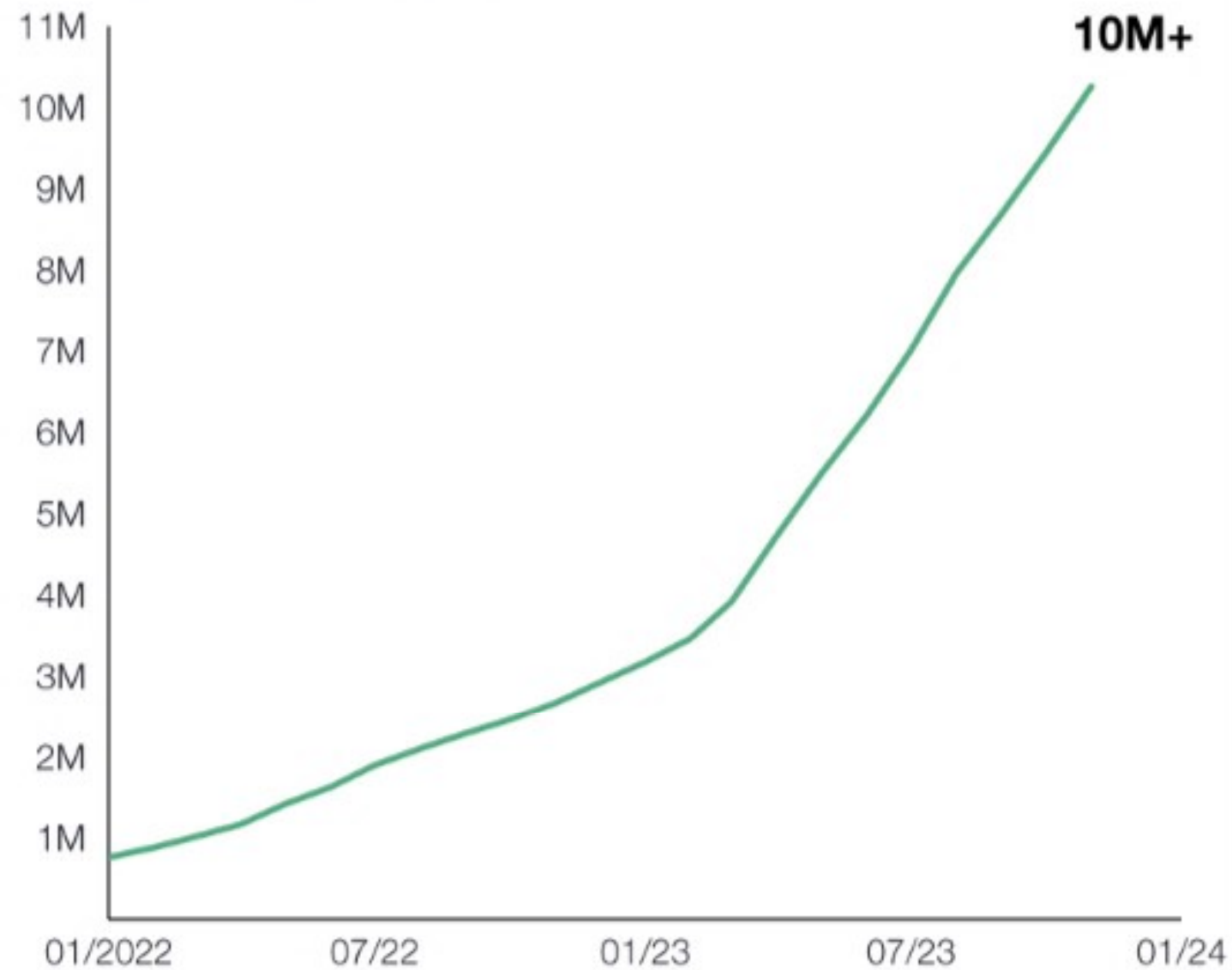
Example: Designing Marketing Materials



GitHub Copilot makes programming faster

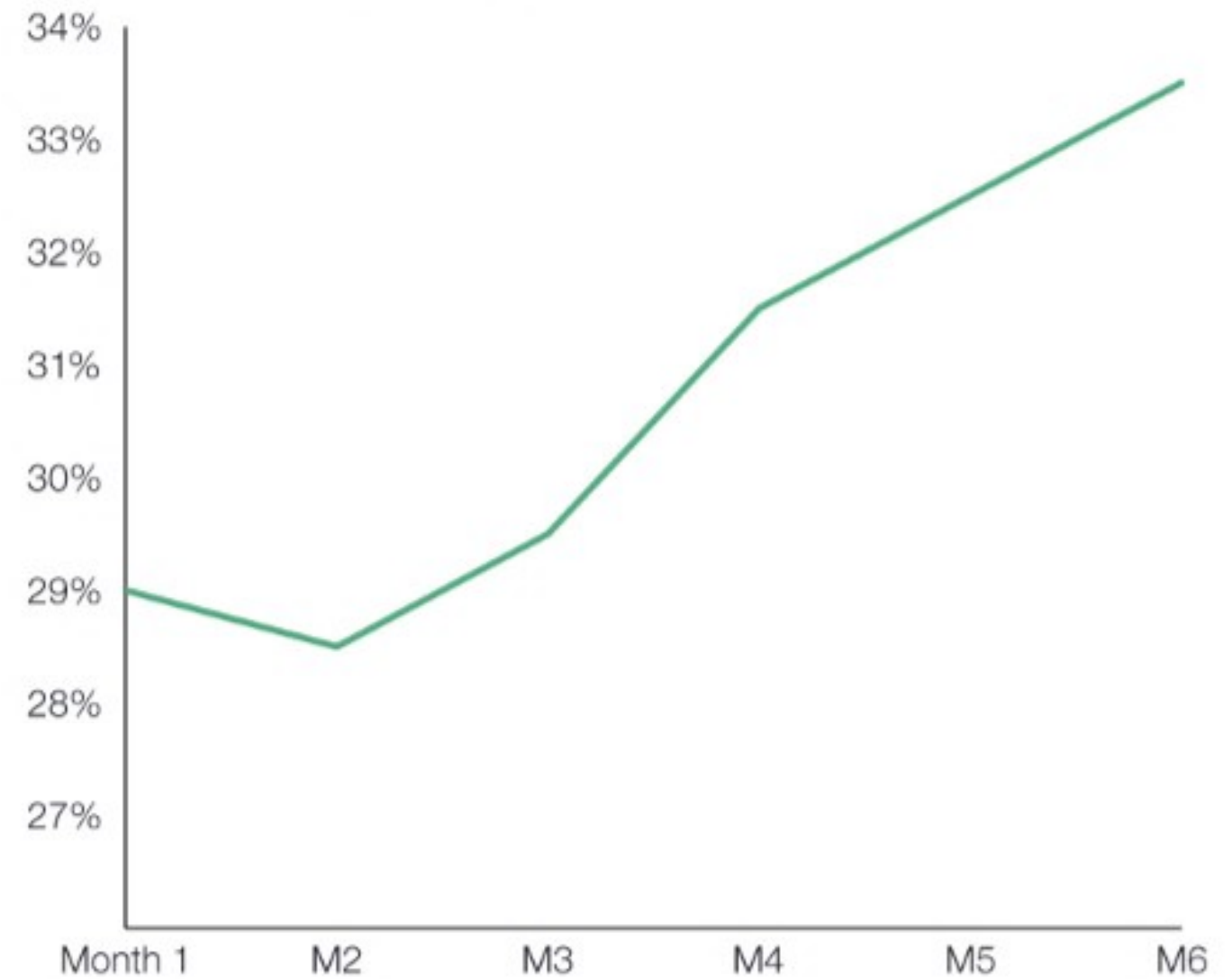
→ **Over 8M downloads of Copilot extension**

Cumulative downloads of GitHub Copilot



→ **Copilot improving developer productivity over time**

Acceptance rate of Copilot suggestions over time

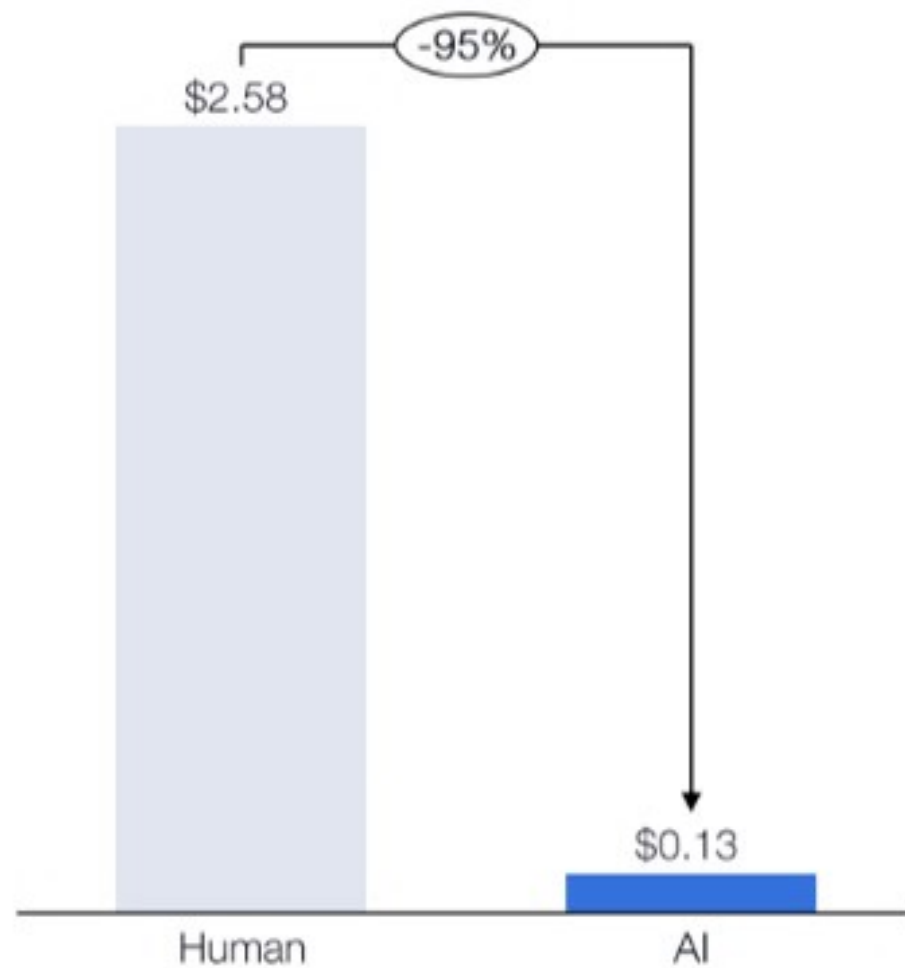


Companies have seen huge efficiency gains already

Experience from one Fintech company

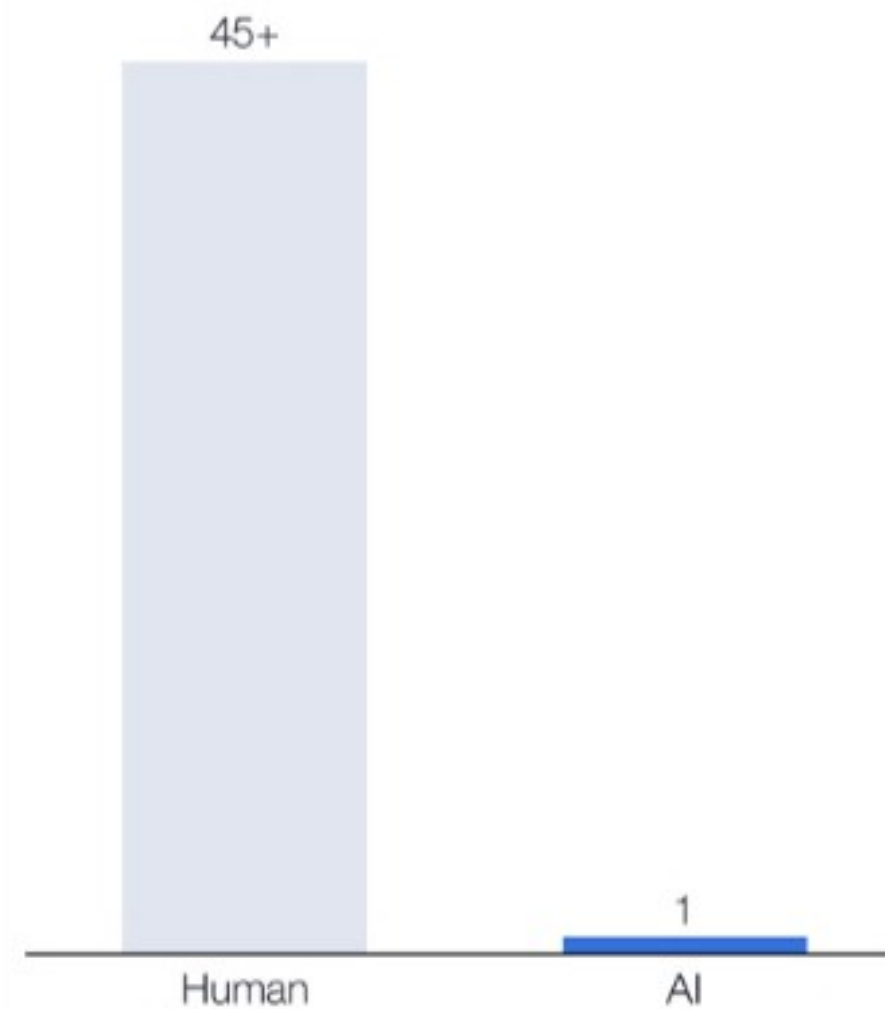
→ AI is cheaper customer support

Cost per support interaction



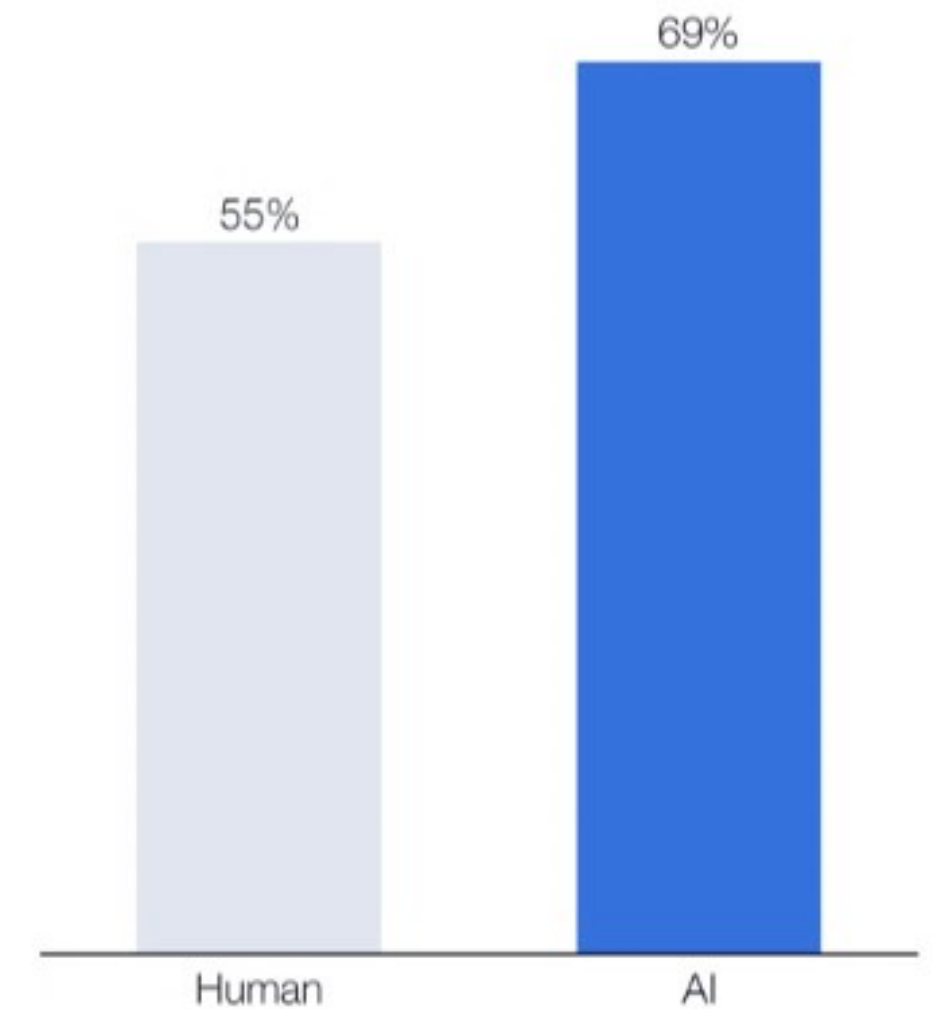
→ AI provides faster responses

Median response time (min)



→ AI makes customers happier

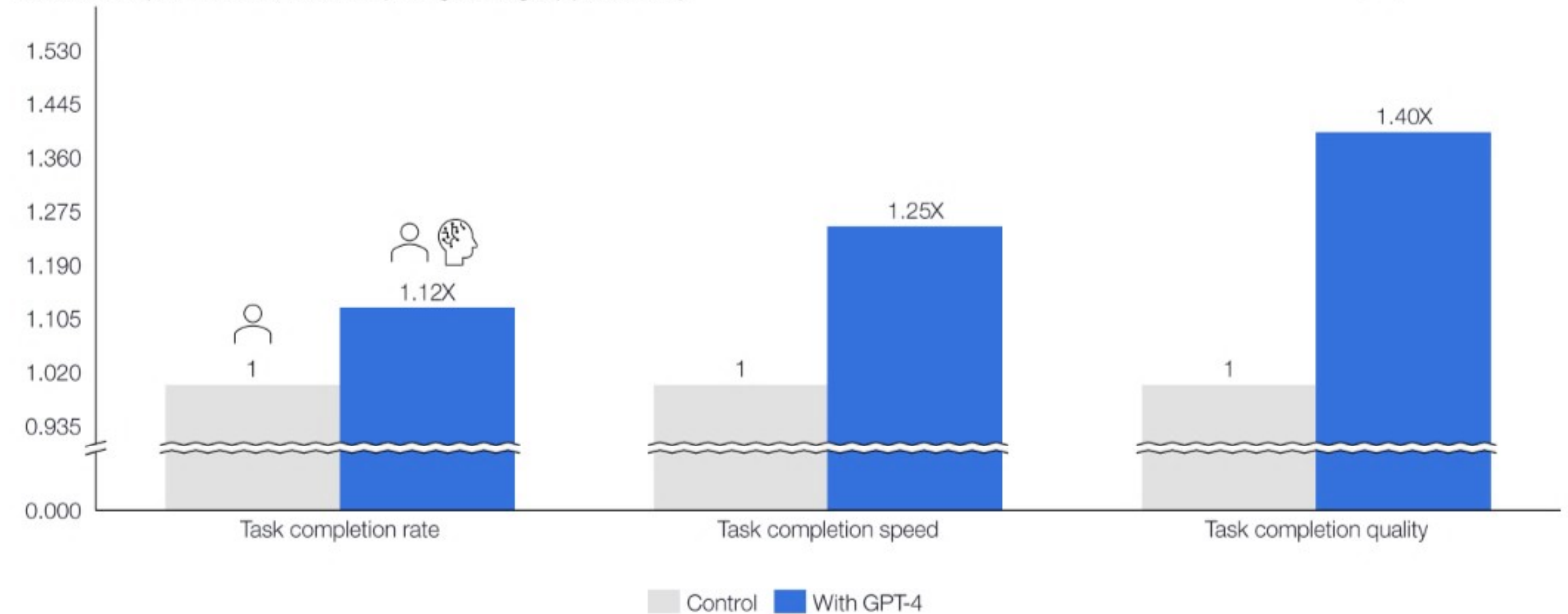
Median customer satisfaction



Knowledge work, such as consulting, could be transformed by AI

→ In one study, BCG consultants using AI performed better across all task metrics, including 40% better quality work

Performance improvement on various tasks (1 being control group performance)



Key Topics

→ Where we are in AI today

→ AI could break through the hype and improve our world

→ **We believe open-source is the lifeblood of AI**

→ AI is transforming the tech ecosystem

→ Coatue view: the best of AI is yet to come

How we got here – AI is built in the open

Research

Openly Available Research

- Research from academia & industry has driven advancements in AI
- Collaboration using open-sourced state-of-the-art models has led to rapid pace of innovation



Community

Open Community

- GitHub and Hugging Face underpin the open-source AI community
- Participation in AI has been explosive



Models & Data

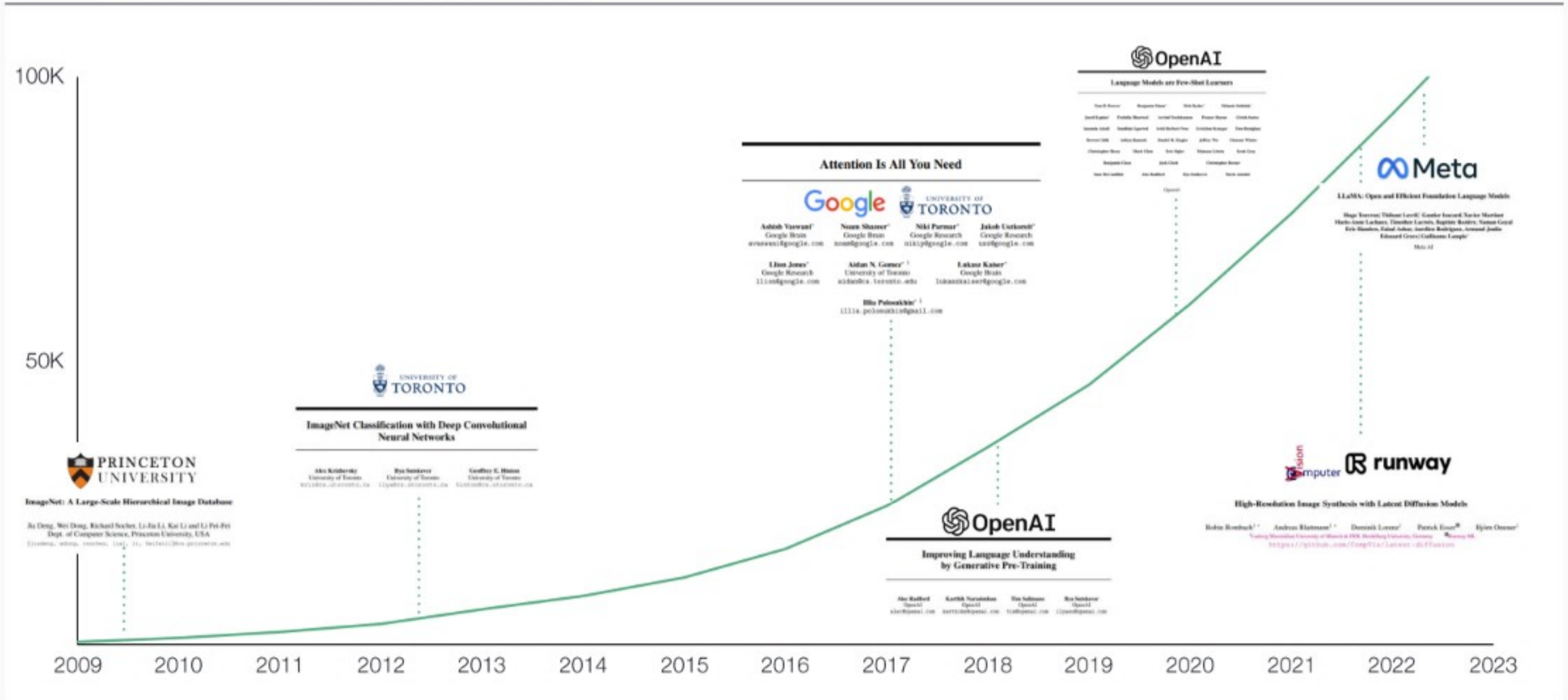
Models & Data

- There are varying degrees of openness across AI today
- Companies realize the value of their own data, and model providers have become more secretive



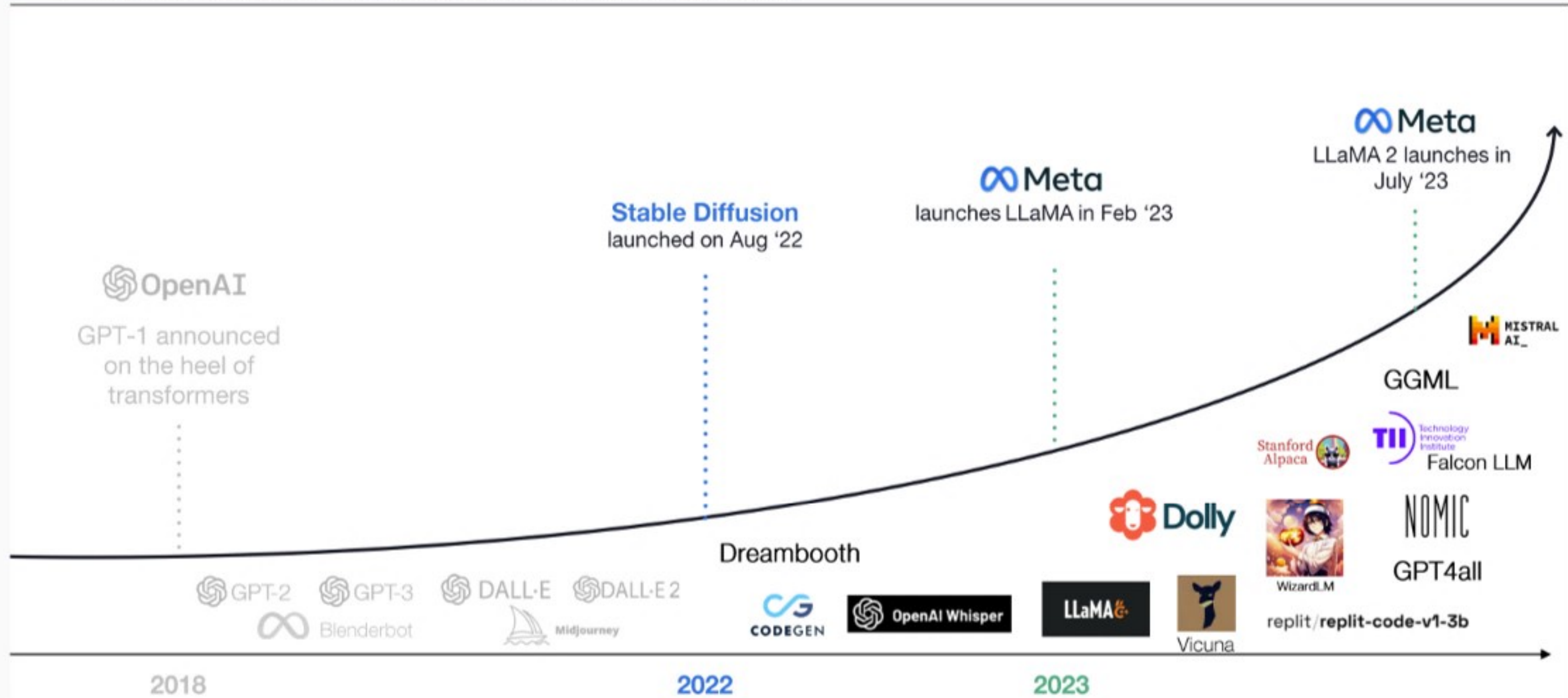
AI is a result of open research

→ Cumulative AI/ML publications submitted on Arxiv



Open collaboration accelerates innovation in AI

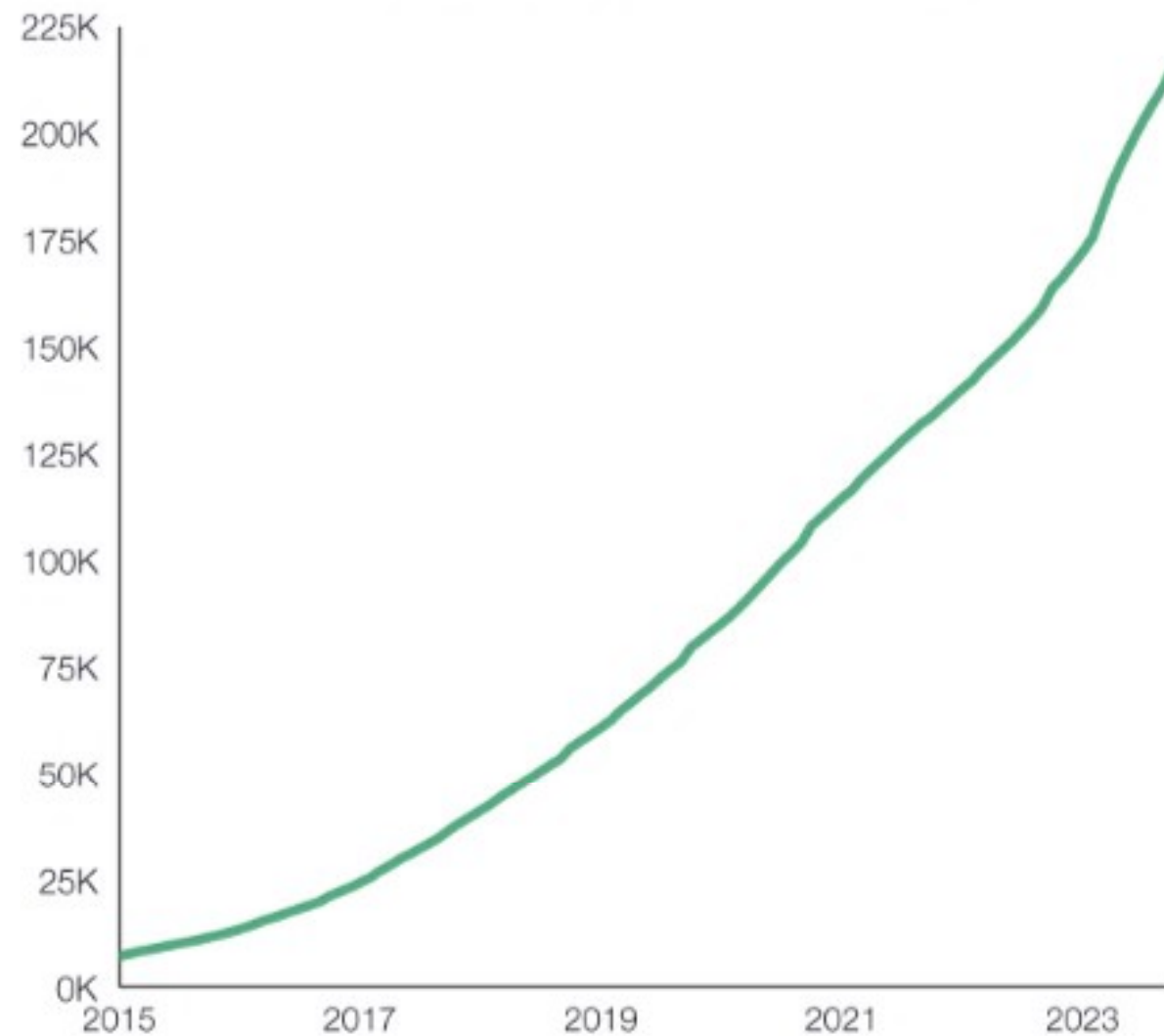
→ Illustrative launches of AI models over time (non-exhaustive)



The AI developer community has exploded!

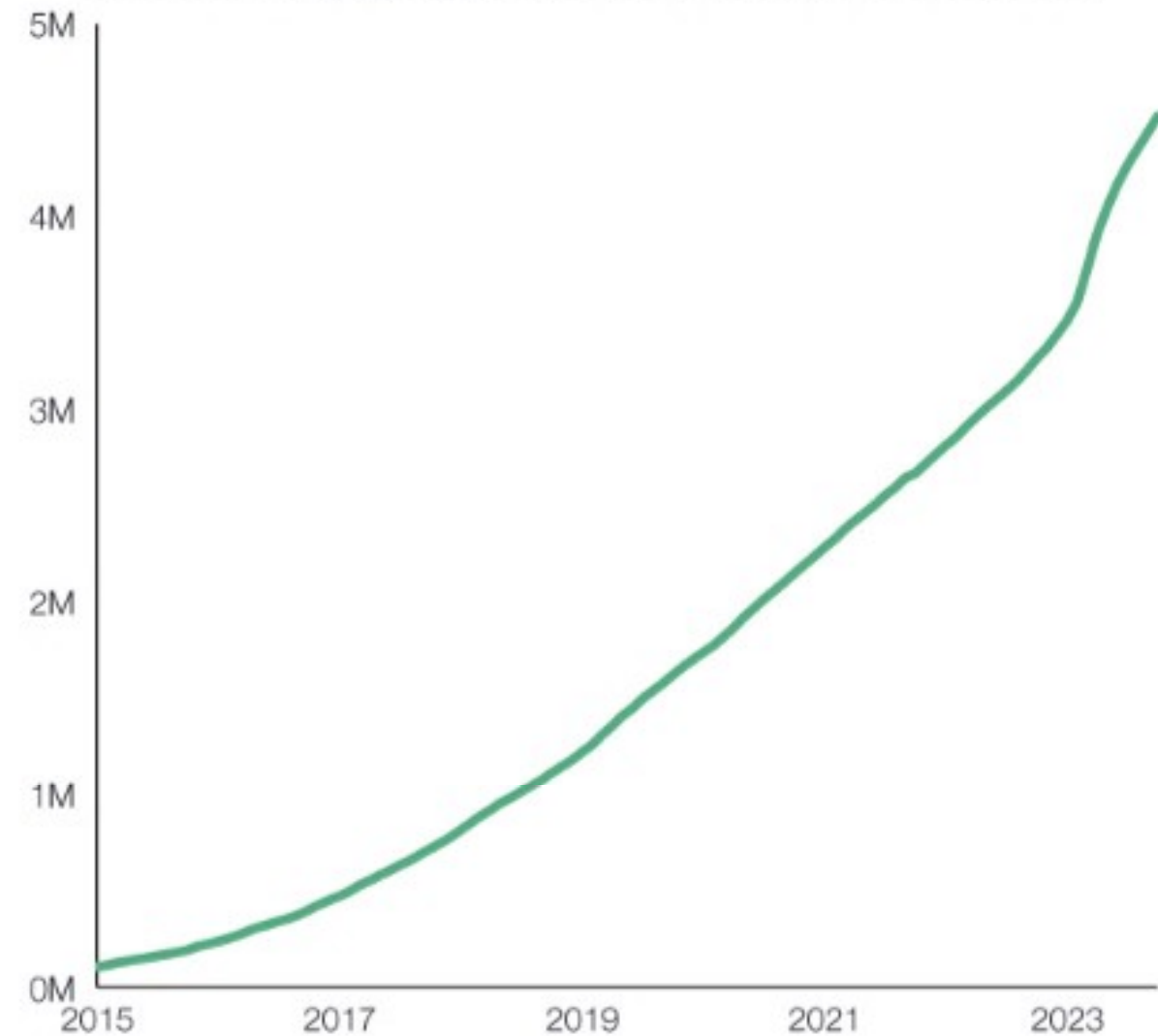
→ **Software developers are becoming AI engineers**

AI/ML Developers on GitHub



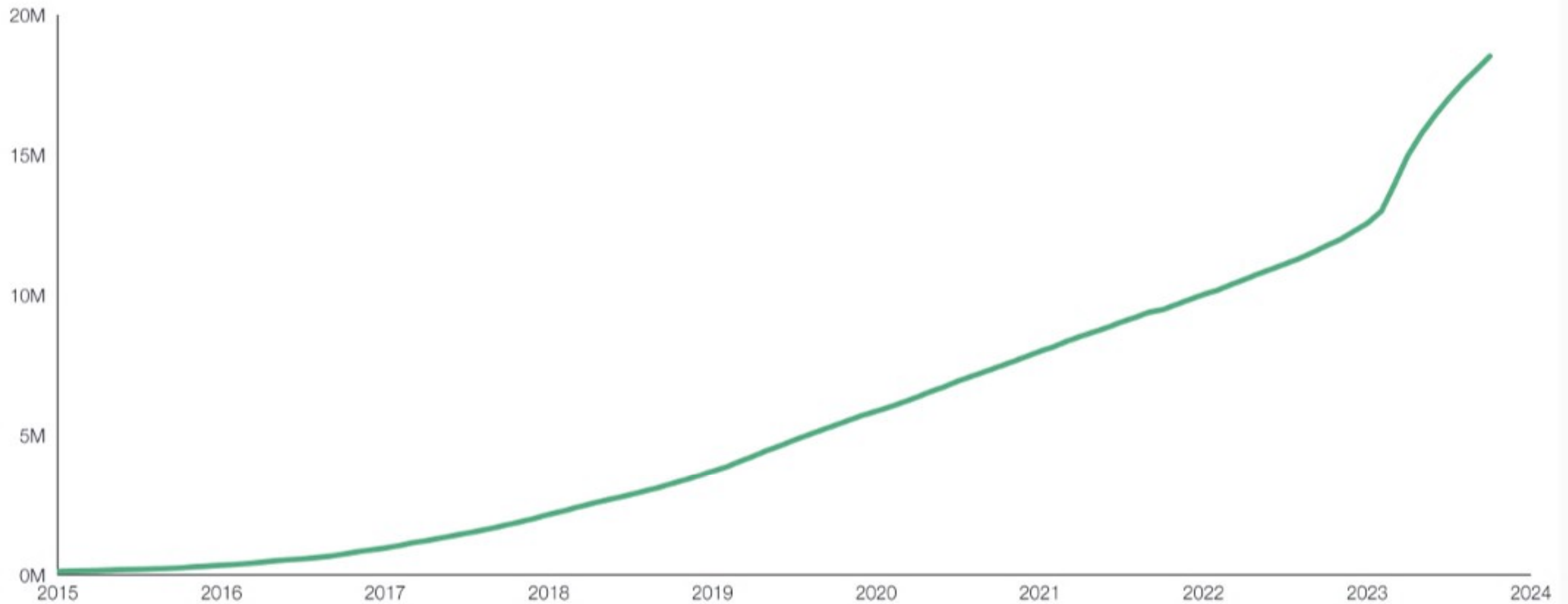
→ **Hobbyists are getting involved**

Users Engaging with AI/ML Projects on GitHub



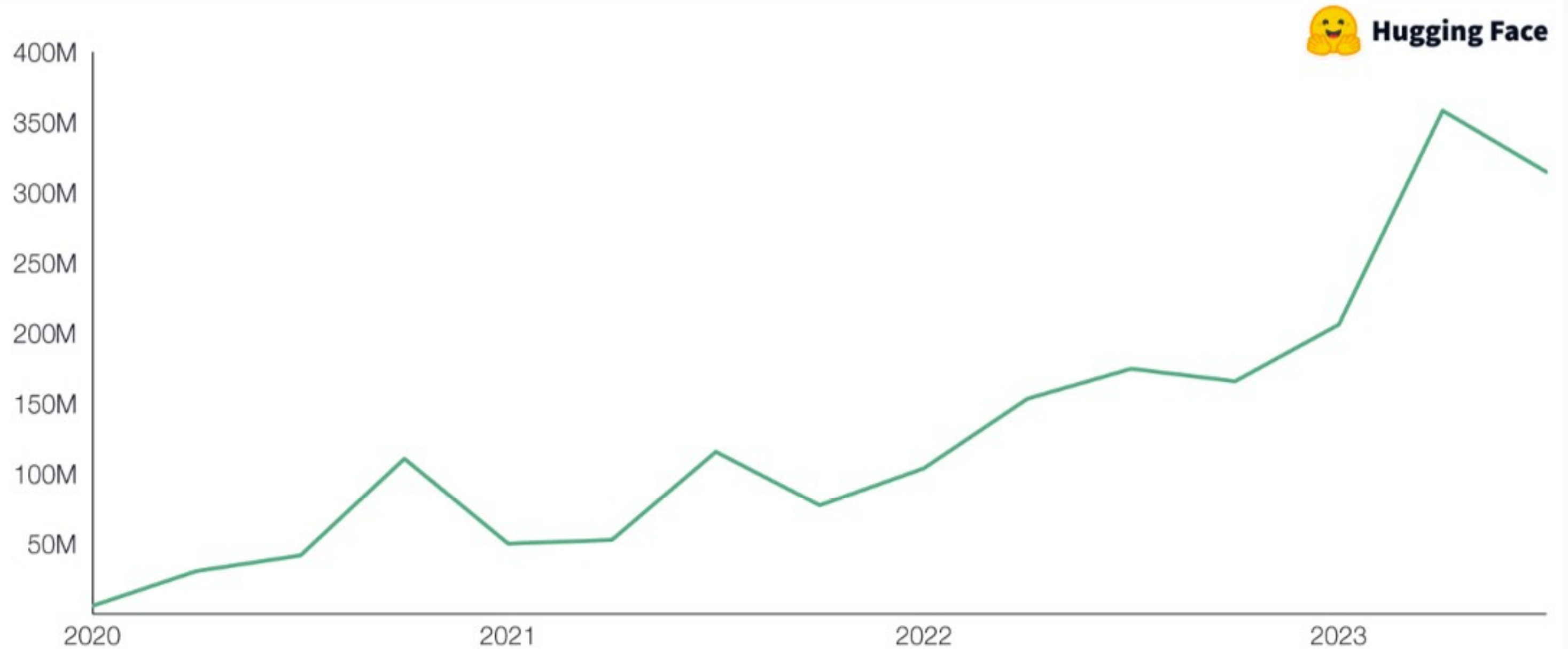
GitHub has been a place to discover new AI/ML projects

→ Cumulative GitHub stars given to AI/ML related projects



Open-source AI is inflecting on Hugging Face as well

→ Number of times AI models have been downloaded from Hugging Face



However, AI is not always truly “open” (1/2)

→ Coatue’s open-source AI model checklist

	GPT-2 Feb 2019	GPT-3 Jul 2020	GPT-4 Mar 2023	LlAMA Feb 2023	LlAMA-2 Jul 2023	Mistral-7B Oct 2023
Dimensions of openness						
Model code						
Model weights						
Training data						
Model evaluation						
Architectural decisions						
Open commercial license						

Meta did not release training data specifics for LlAMA-2. Could training data be the next battleground?

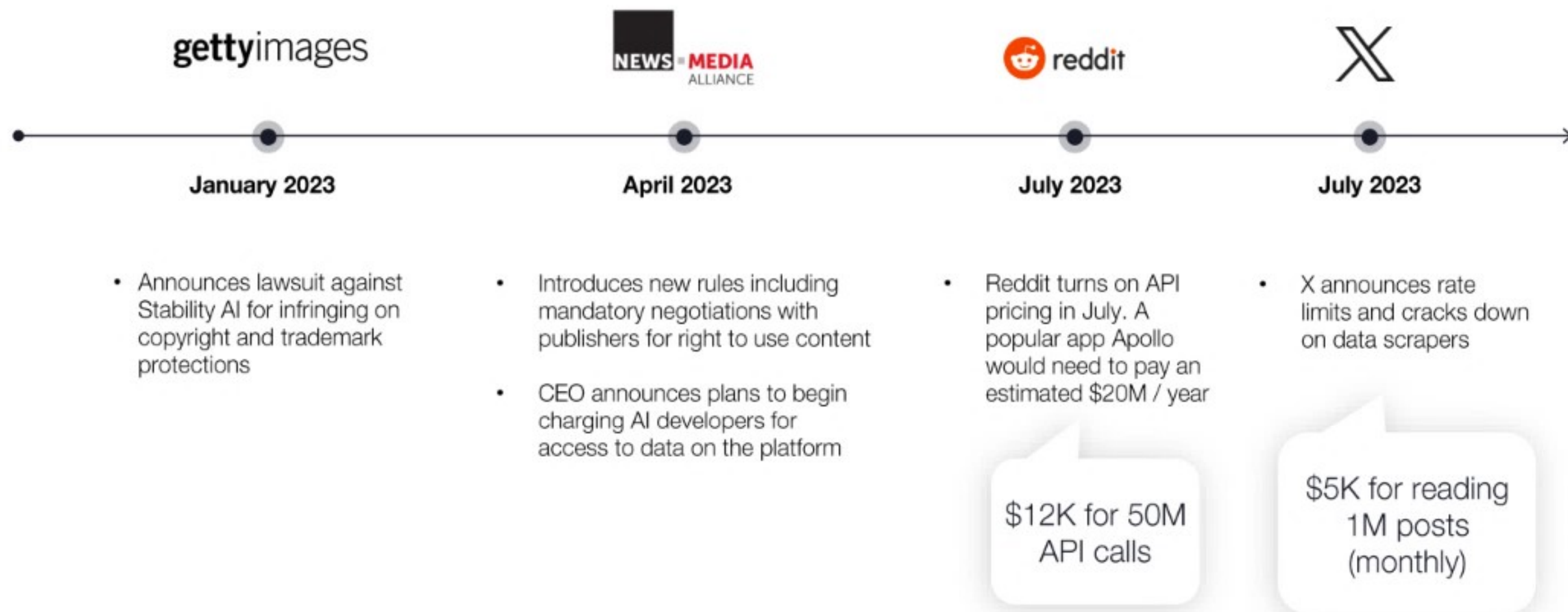
However, AI is not always truly “open” (2/2)

→ Coatue’s open-source AI model checklist

	Codex Aug 2021	Codegen Mar 2022	Code v1 May 2023	Dall-E 2 Apr 2022	Midjourney Jul 2022	Stable Diffusion 1.0 Aug 2022
Dimensions of openness						
Model code						
Model weights						
Training data						
Model evaluation						
Architectural decisions						
Open commercial license						

Data is *finally* a new currency

→ Timeline of companies cracking down on data access in 2023

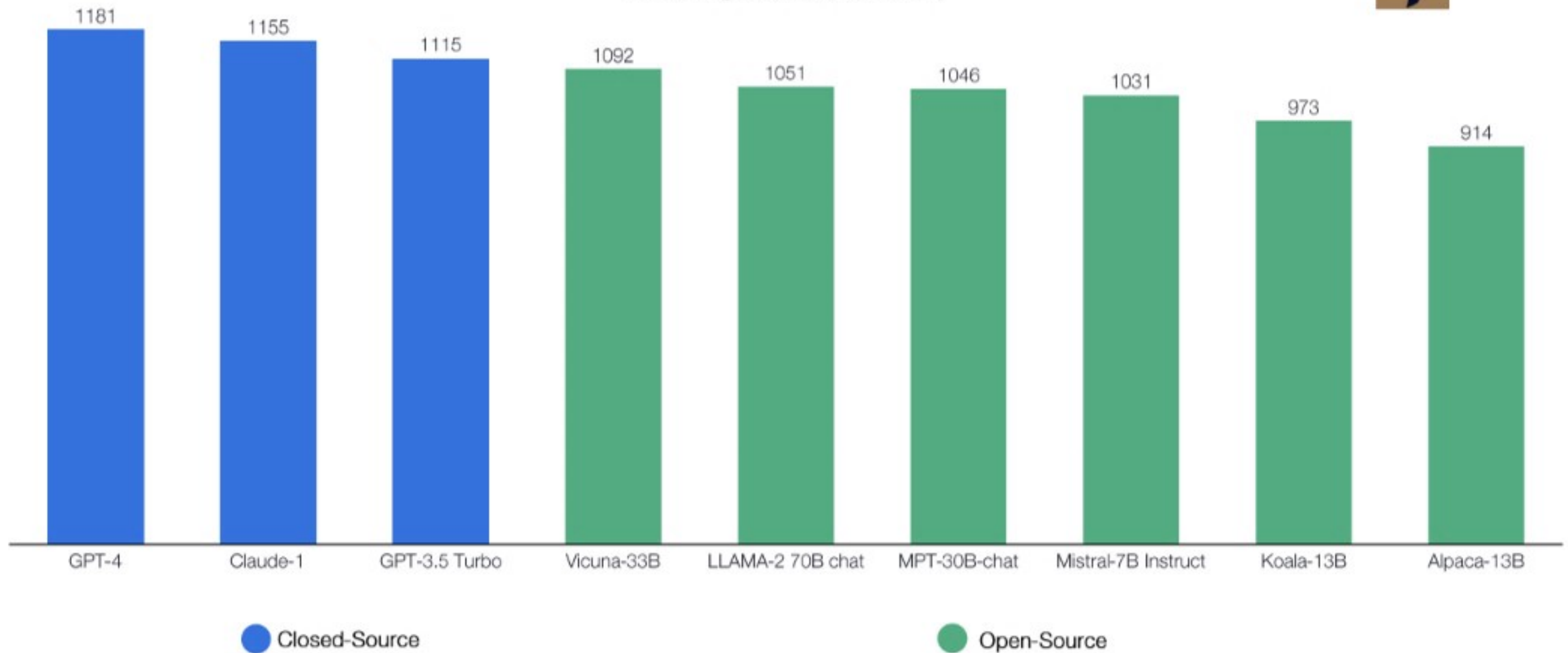


Despite signs of “closed AI”, open-source models are catching up

→ Chatbot rankings based on human feedback



ELO Ratings of LLMs (50K+ votes)



Key Topics

→ Where we are in AI today

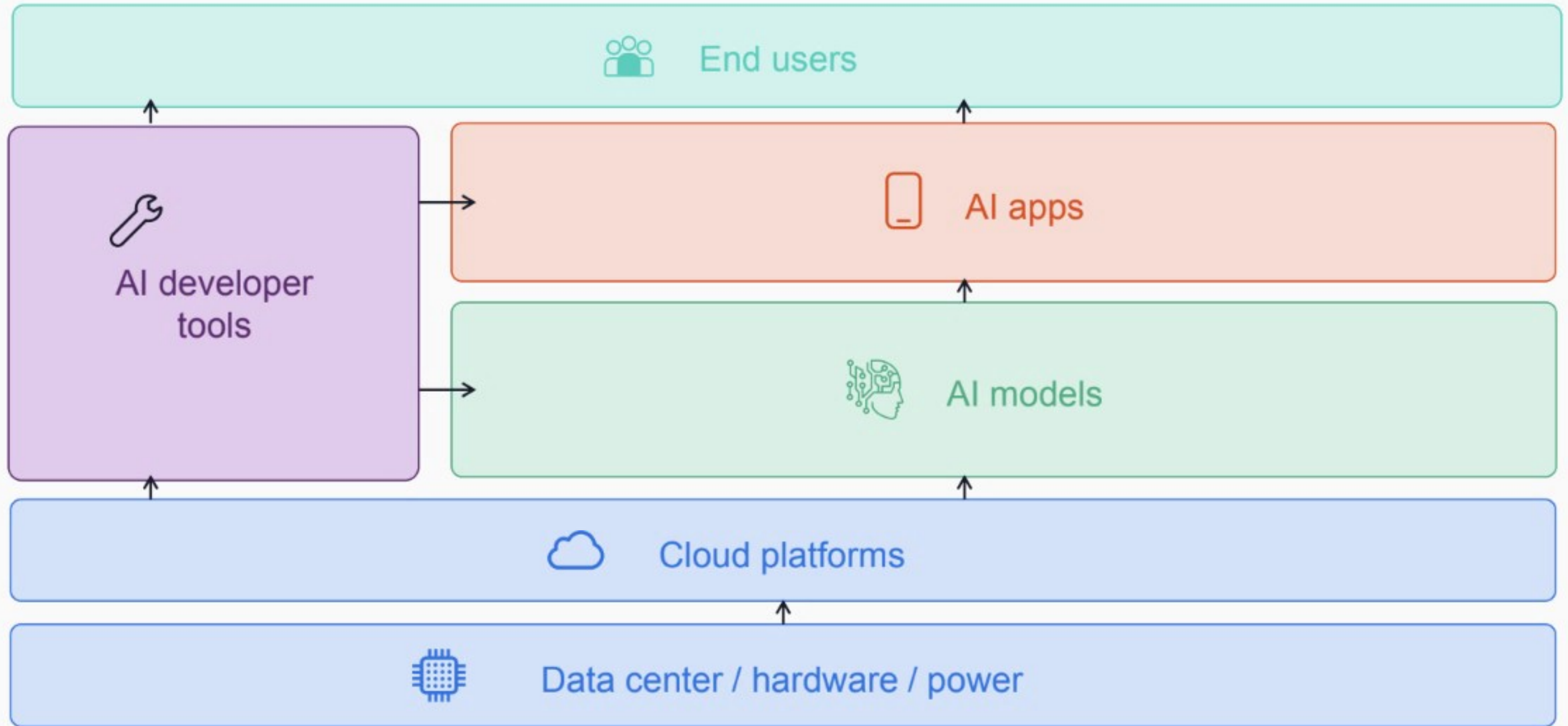
→ AI could break through the hype and improve our world

→ We believe open-source is the lifeblood of AI

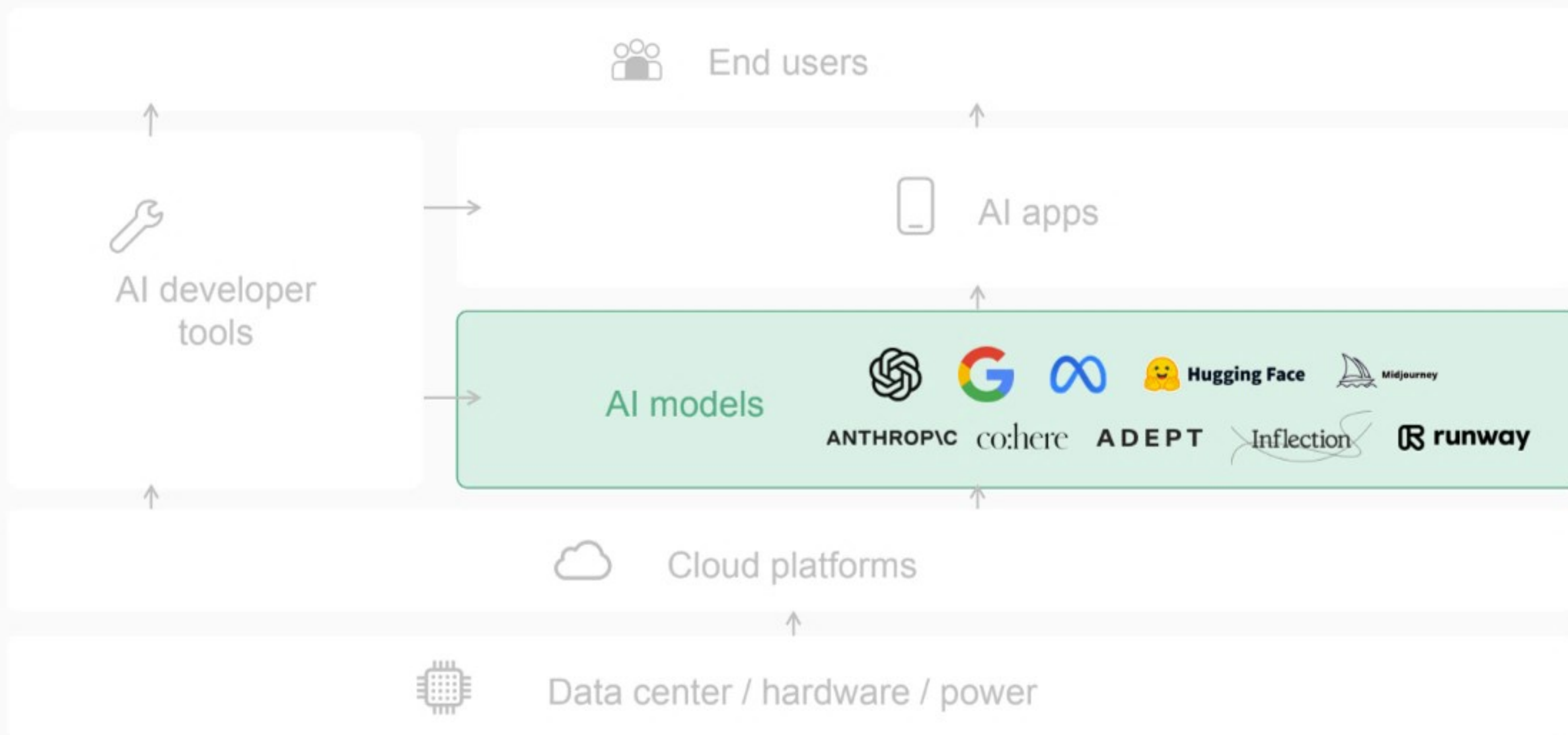
→ **AI is transforming the tech ecosystem**

→ Coatue view: the best of AI is yet to come

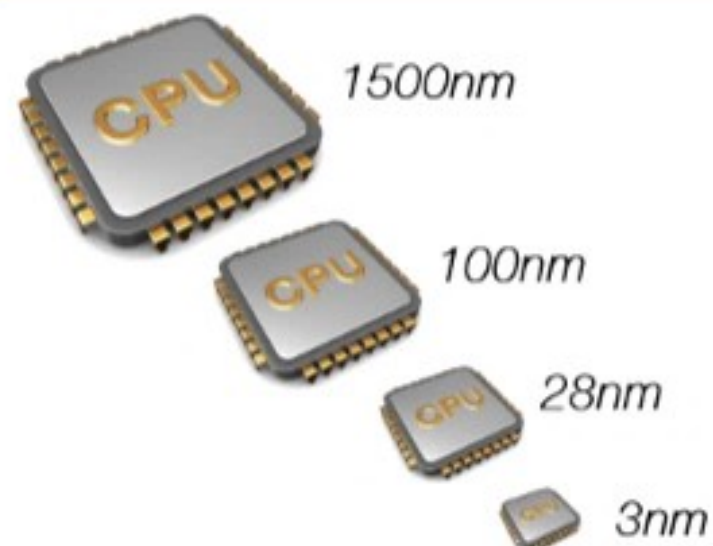
The new AI-centric technology ecosystem



Foundation models are at the center of AI



Last 50 years was about building faster & faster “calculators”...



CPU

Serial processing

“1 instruction at a time”

Chips got smaller & more powerful

+

```
double AttackerSuccess (double q, int z)
{
    double p = 1.0 - q;
    double lambda = z * (q / p);
    double sum = 1.0;
    int i, k;

    for (k = 0; k <= z; k++)
    {
        double poisson = exp(-lambda);

        for (i = 1; i <= k; i++)
        {
            sum += poisson * (1 - pow(q / p, z))
        }
    }
    return sum;
}
```

Software

Based on instruction by programmer

Follows sequential programming logic

Does not require data

=



Computer or “Calculator”

But next 50 years will be about building super-intelligent “brains”

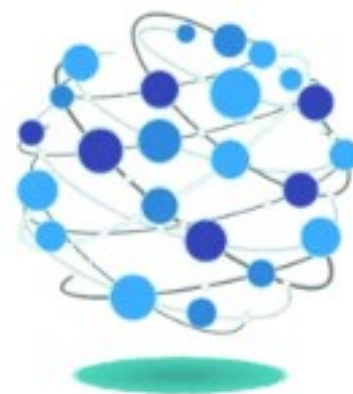


GPU

Parallel processing

Many calculations simultaneously

+



AI Models

Neural networks trained by data

Learns patterns from data

System makes decisions based on model rather than explicit instructions

“Reasoning” is opaque, not driven by programming logic

=



Brains

Much more than calculators!

Could become connected:
Brain-to-brain network = AI Internet?

AI enables a new platform: Intelligence-as-a-service

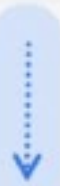
Intelligence is the next layer of innovation

Today



Yann LeCun
@ylecun

"[One of my] opinions on current LLMs: They are "reactive" & don't plan nor reason."

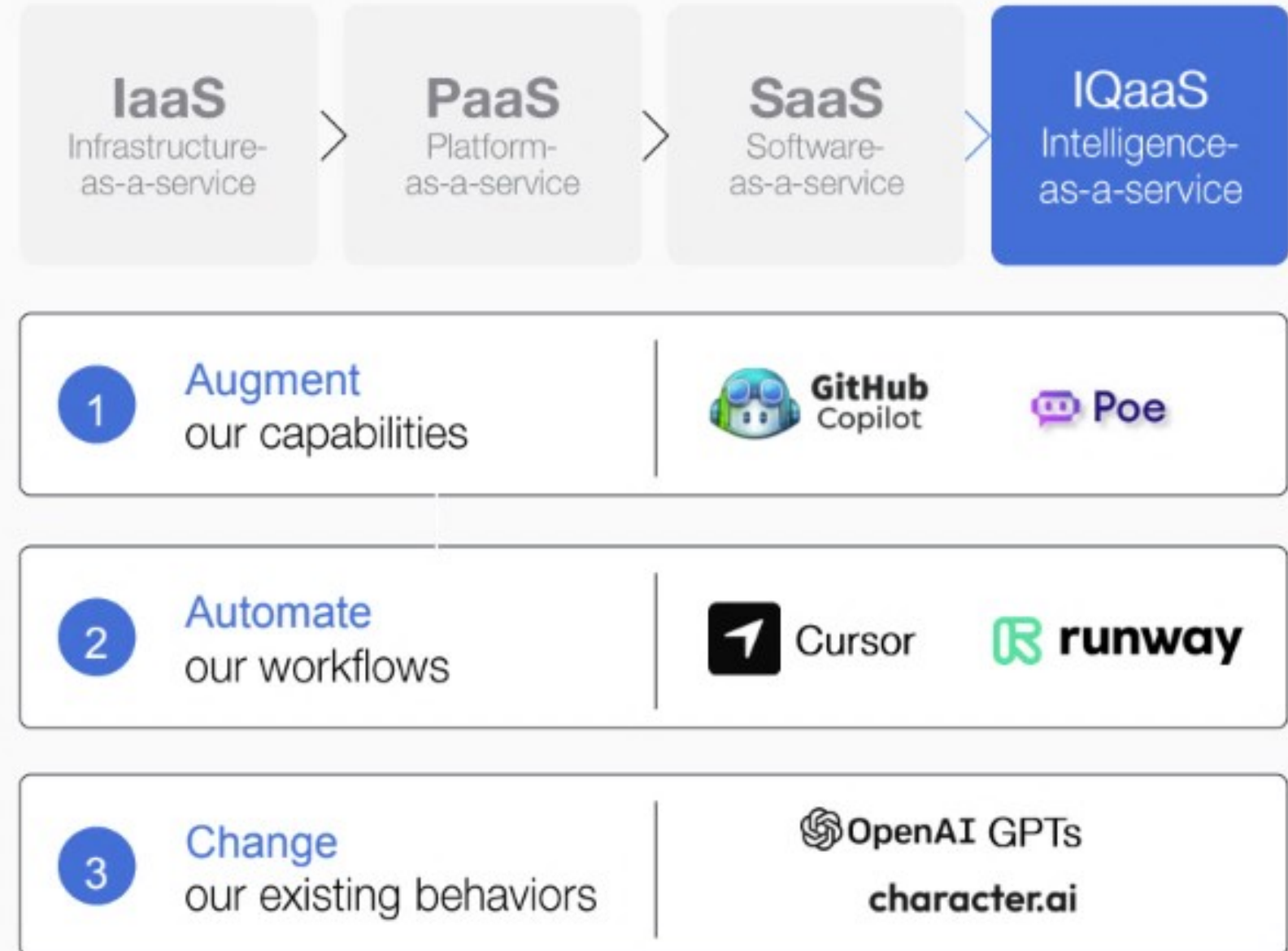


Future



Sam Altman
@sama

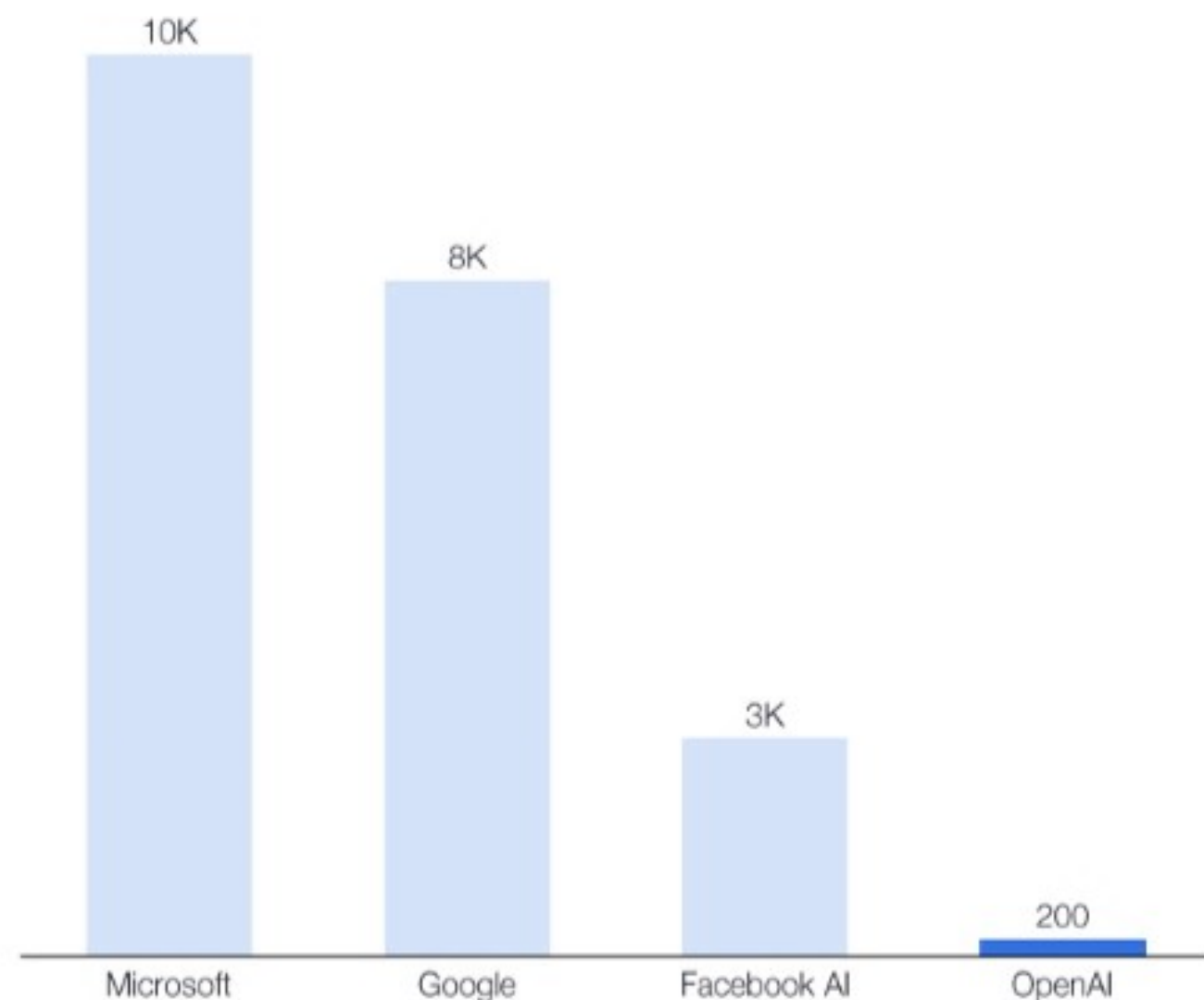
"The right way to think of the models we create is a reasoning engine, not a fact database."



More research & headcount is not enough to win in AI

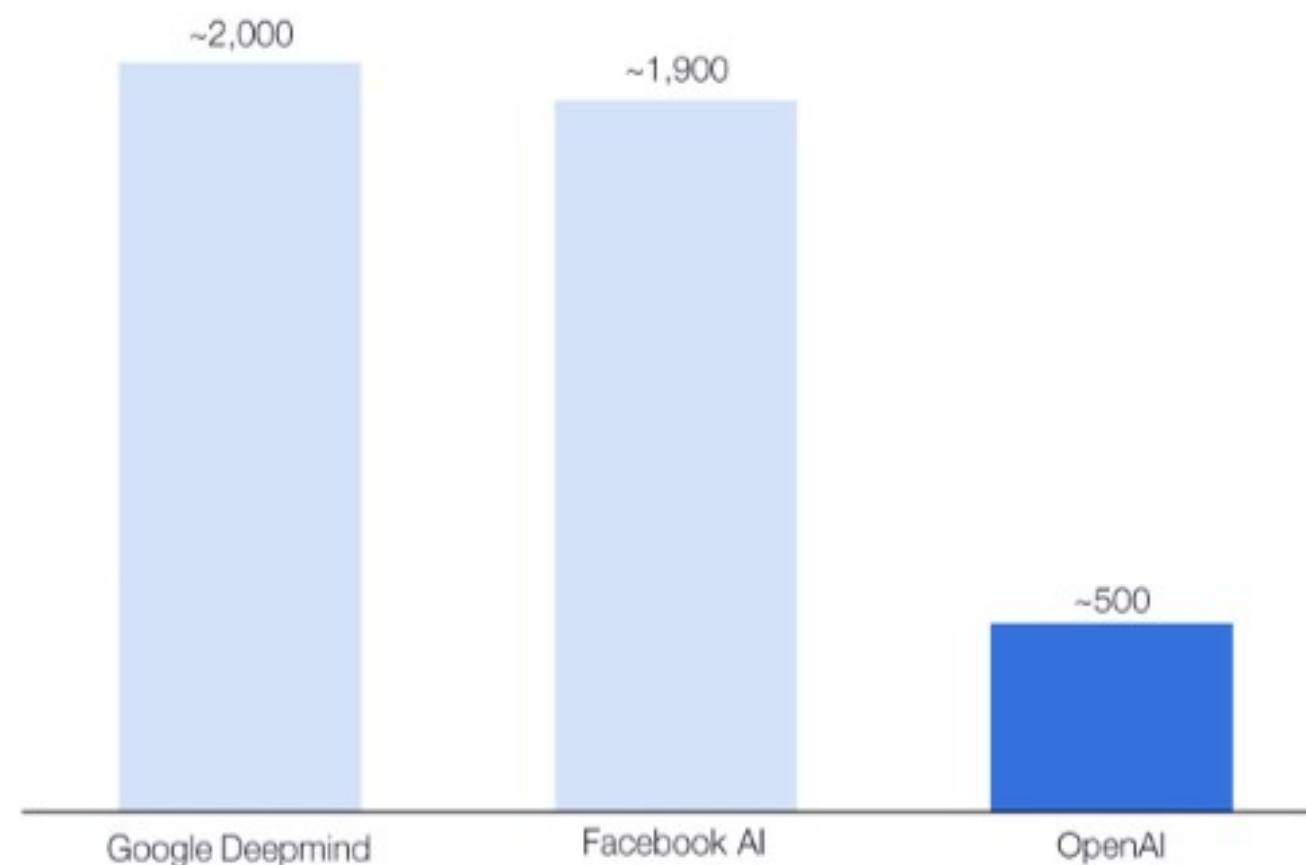
→ MSFT leads big research houses in publications

Number of AI/ML related research publications



→ OpenAI has shipped faster than larger peers

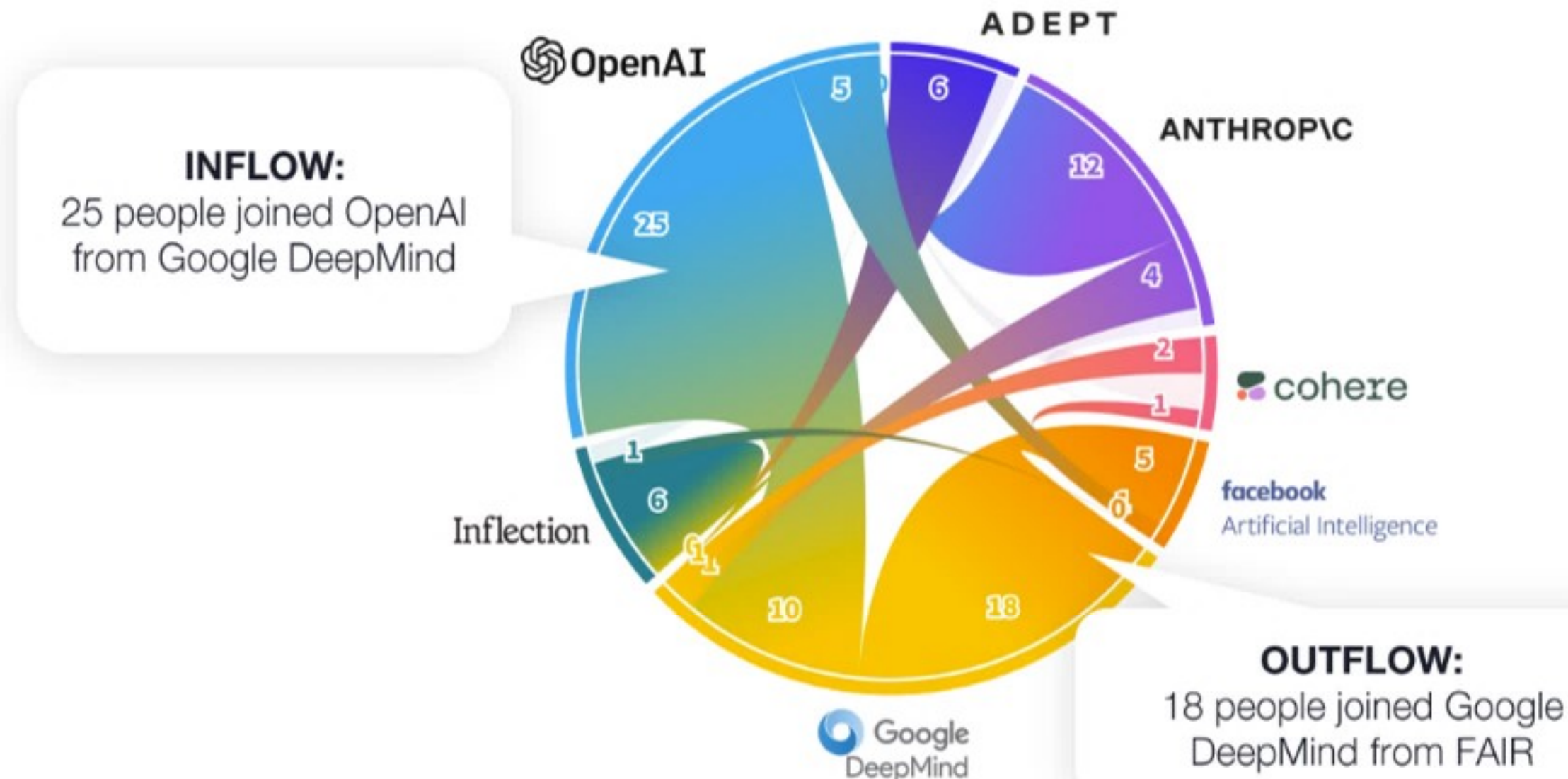
Number of AI/ML related full-time employees



Talent has historically been concentrated at a few AI model hubs

→ Major AI model providers are “poaching” talent from one another

Talent inflows & outflows at major research hubs

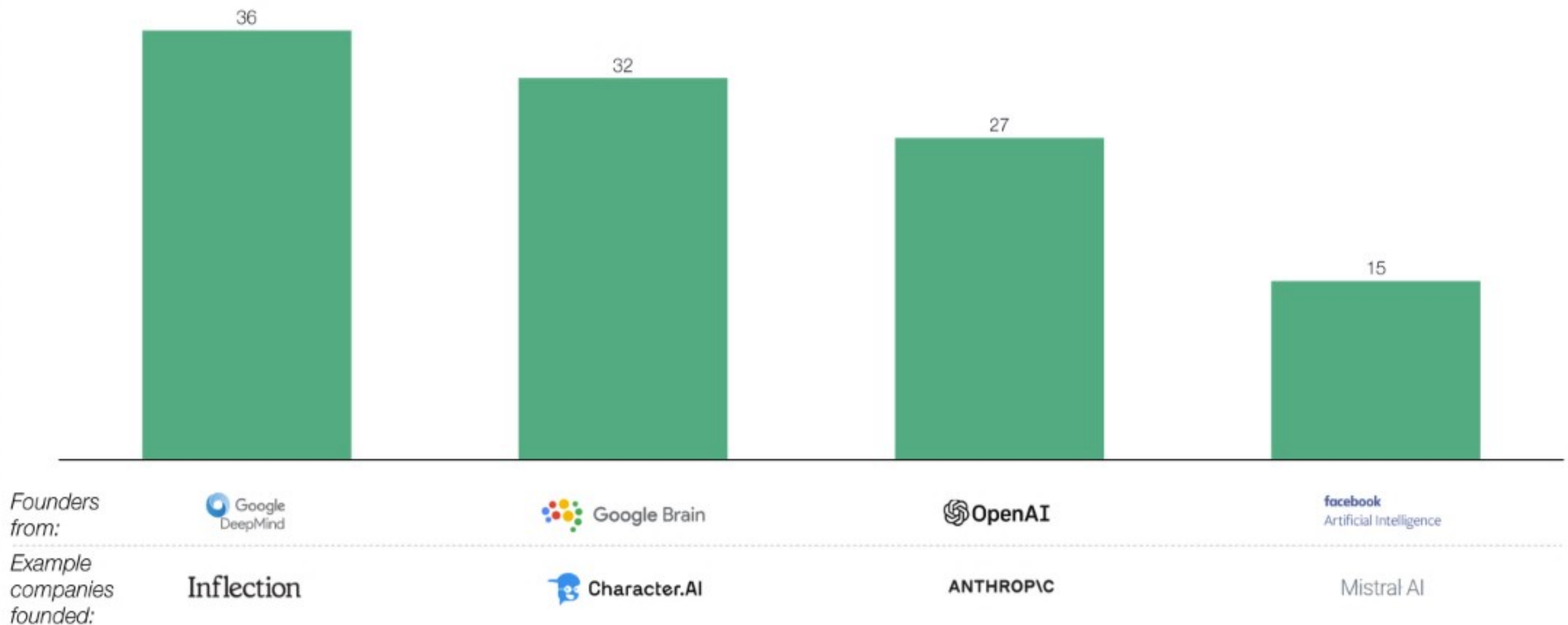


Illustrative only: Represents both direct & indirect (>1 hop) flows of talent

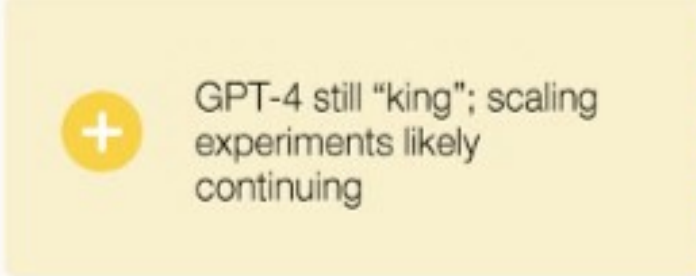
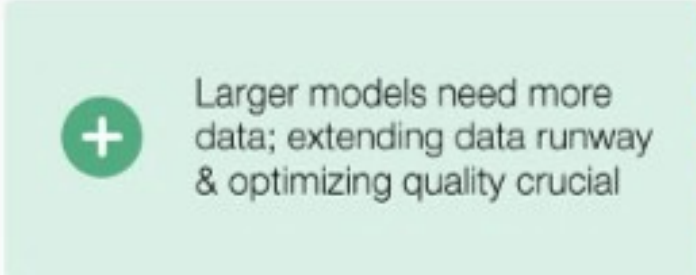
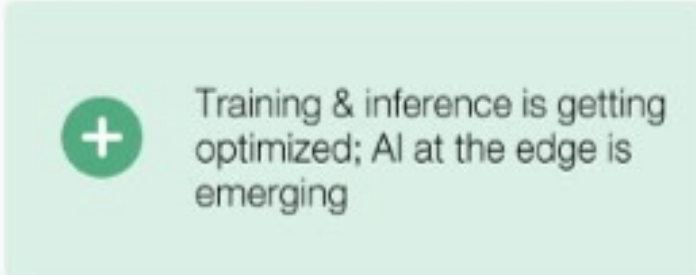
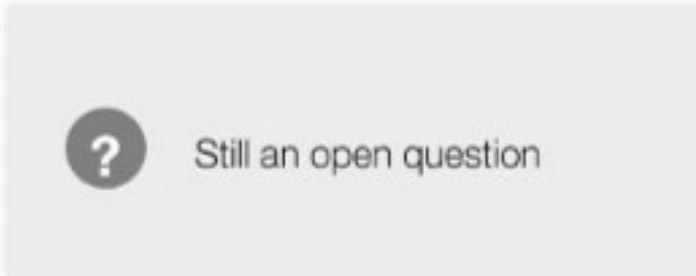
Many talented people are now leaving to start new companies

→ Example AI mafias

Number of founders from AI research houses



Scaling AI model performance has been a focus of AI researchers

Strategies to scale	Open research questions	Coatue View
More parameters	Will scaling parameters continue improving performance?	 GPT-4 still "king"; scaling experiments likely continuing
Larger datasets	How much data is optimal for training models?	 Larger models need more data; extending data runway & optimizing quality crucial
More compute	Can we reduce the compute costs of training & inference?	 Training & inference is getting optimized; AI at the edge is emerging
Longer training	Can training for more epochs improve performance?	 Still an open question

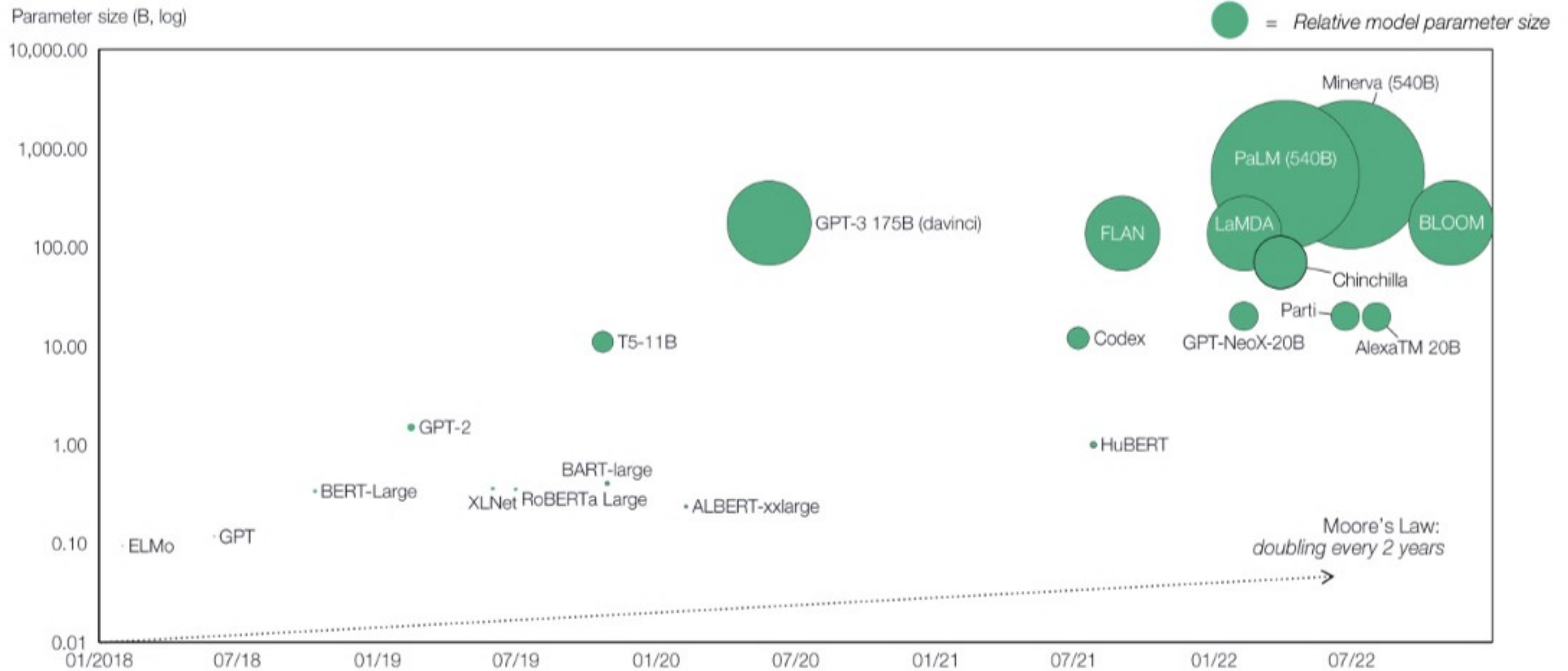
Focus of remainder of section

Covered in next section on Cloud, DC, Semis

Not covered; waiting for more data points to emerge

Through 2022: We saw LLM parameters balloon!

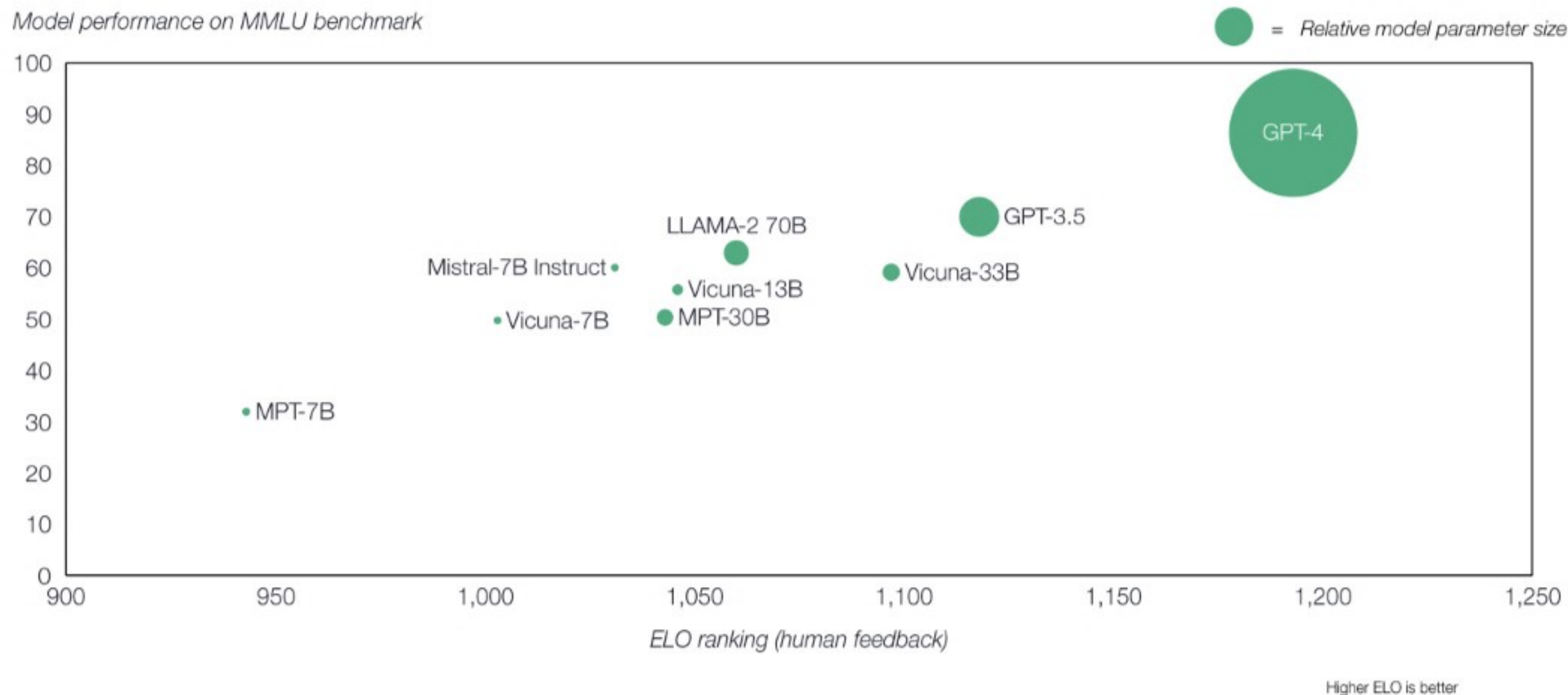
→ Illustrative view of parameter sizes of large language models through 2022



2023: GPT-4 remains “king” ... can we continue scaling?

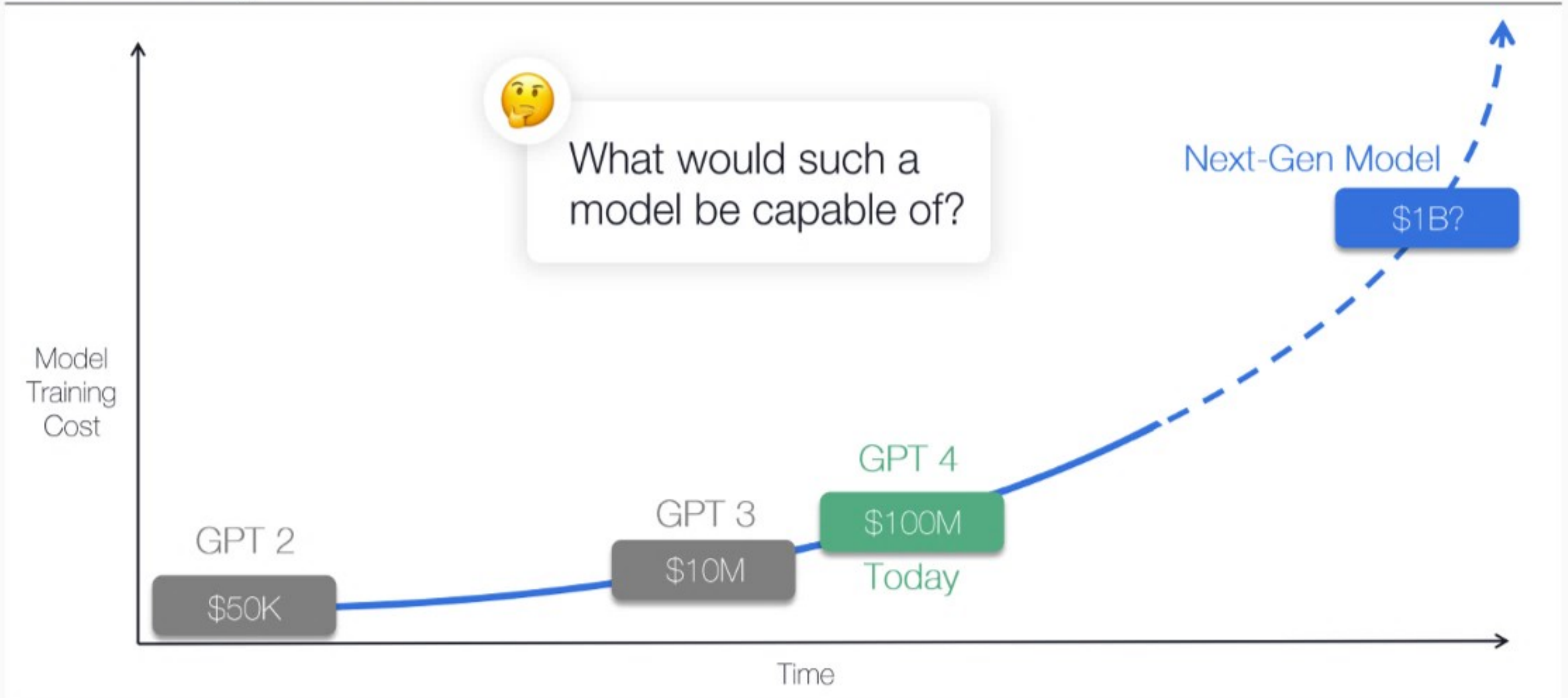
→ Comparison of 2023 models across ELO ranking and MMLU benchmarks...GPT-4 is still the best

Model performance on MMLU benchmark



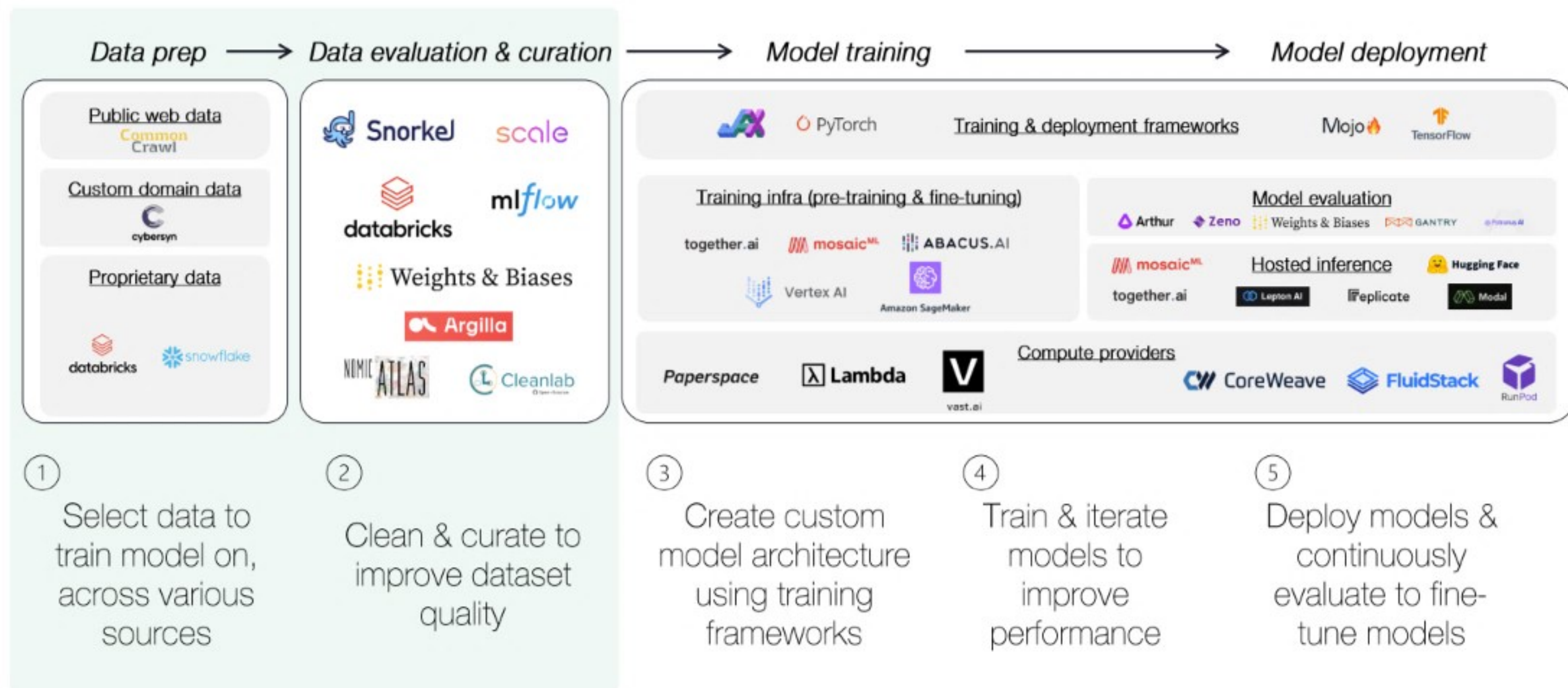
What if there was a \$1B training-cost model?

→ Model Training Cost Over Time



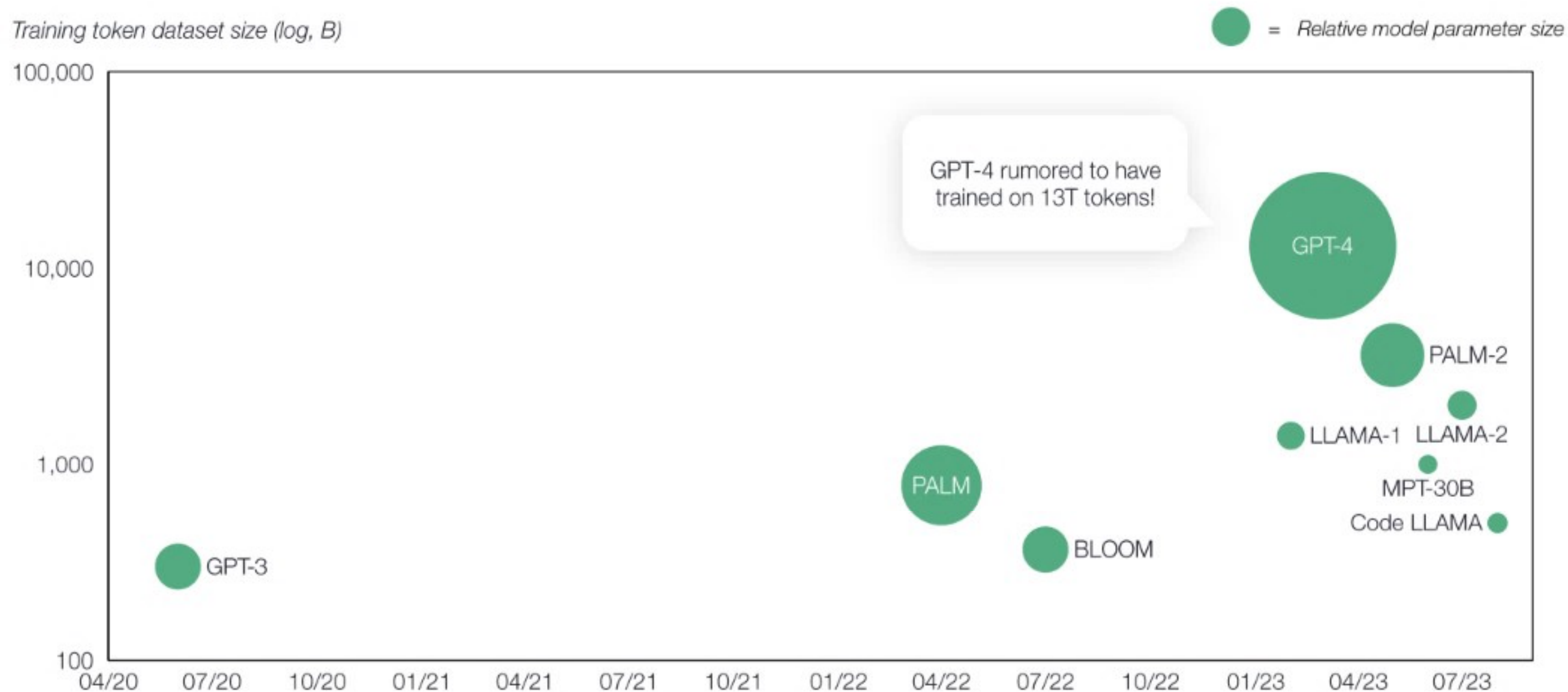
How AI models are made: Data is a crucial component

→ Data is upstream in process of developing good models



Scaling models require scaling datasets

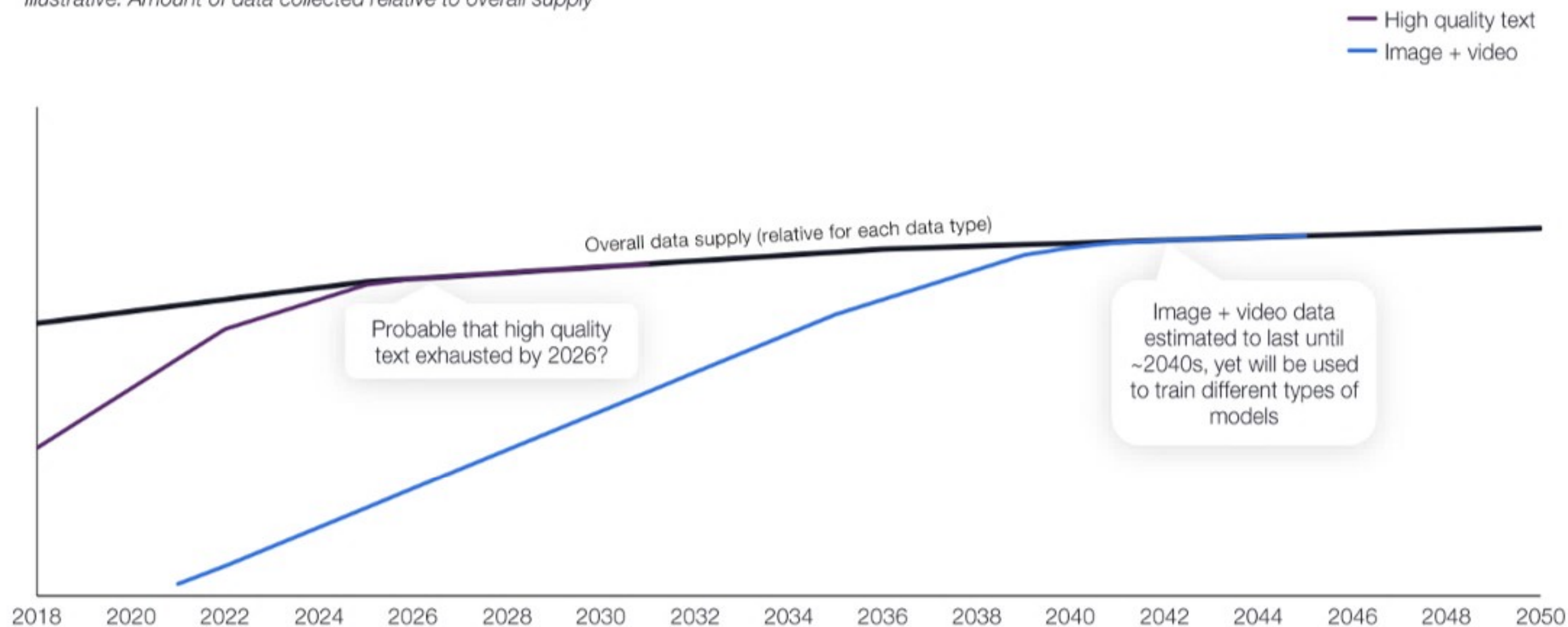
→ Number of tokens in training datasets increasing overall



Data scarcity is a potential wall to scaling models

→ **High-quality text data could be exhausted soon, images & video have longer runway**

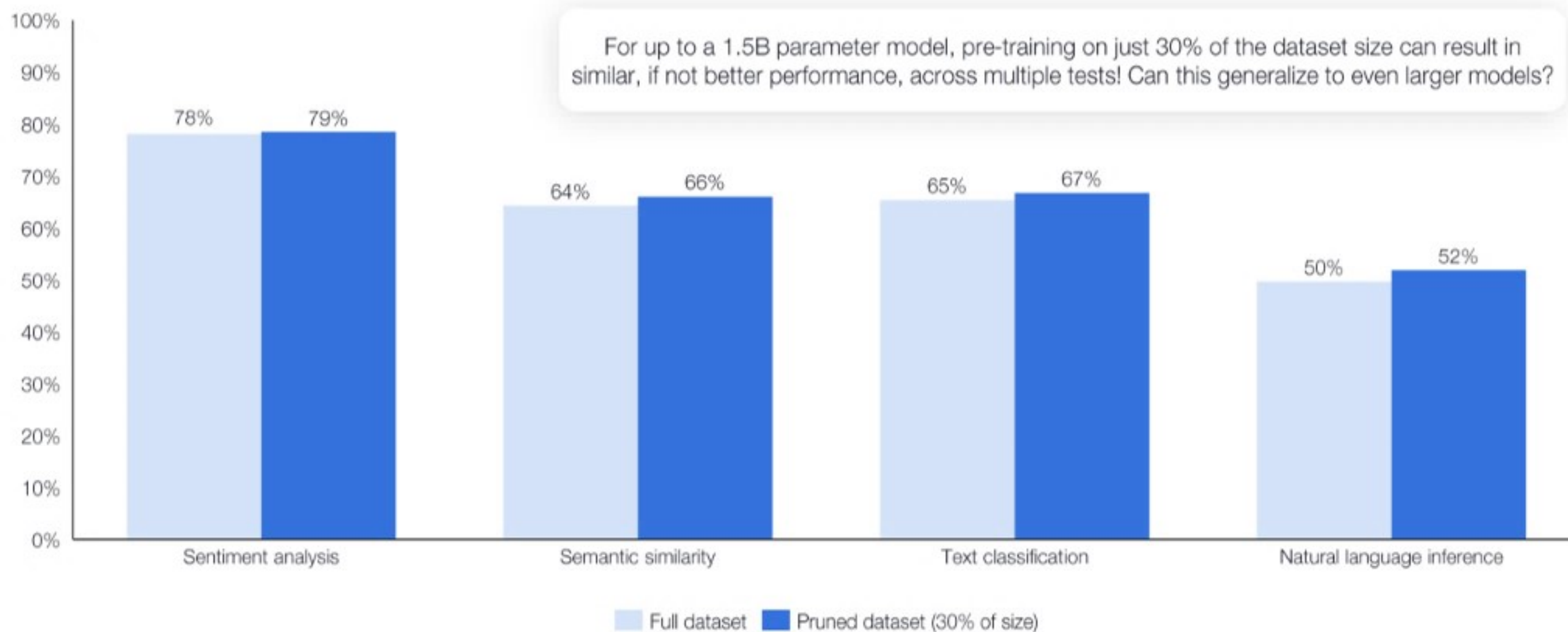
Illustrative: Amount of data collected relative to overall supply



Data quality is just as important as data quantity

→ Emerging evidence that training on pruned datasets can result in similar performance in language models

Performance on variety of language model benchmarks

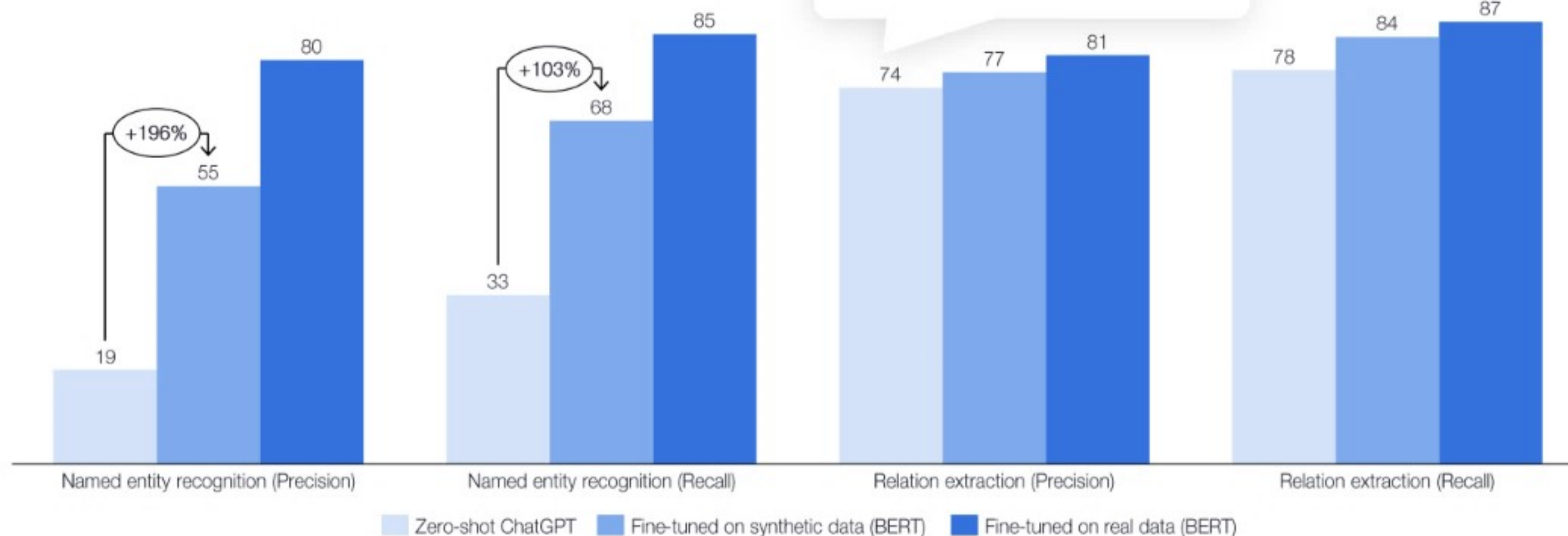


Synthetic data can augment fine-tuning

→ Synthetically generated data can help clean or distill datasets for fine-tuning

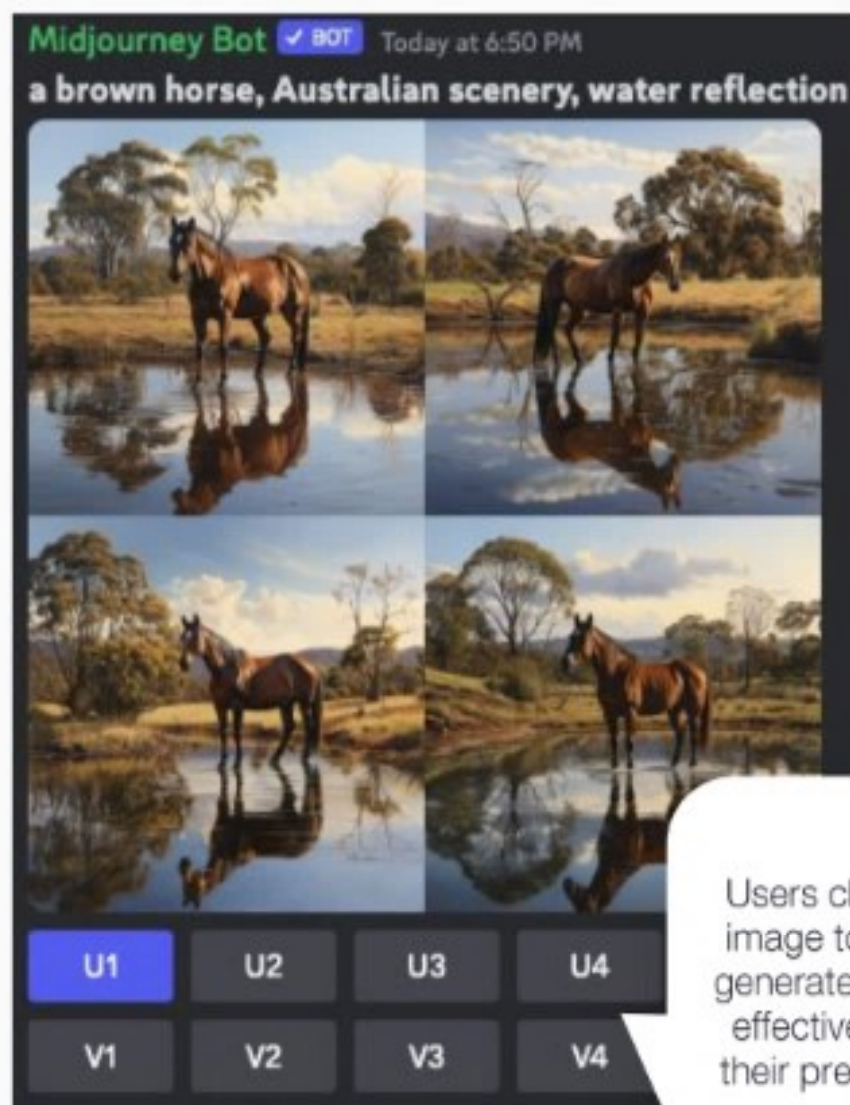
Results: Synthetic data fine-tuning yields significant improvement compared to zero-shot attempts, though is not yet comparable to real data

Percentile performance on test



User feedback data is another way to improve performance

Midjourney collects user feedback to continue refining models



Users choose which image to upscale (or generate a variant of), effectively choosing their preferred image

Prompt /dungeons and dragons, female knight, of the rolling plains, full body, dark azure, victorian genre paintings, serene face, realistic depiction of light, golden light



Midjourney V1
(Feb 2022)



V2 (Apr 2022)



V3 (Jul 2022)



V4 (Nov 2022)



V5 (Mar 2023)

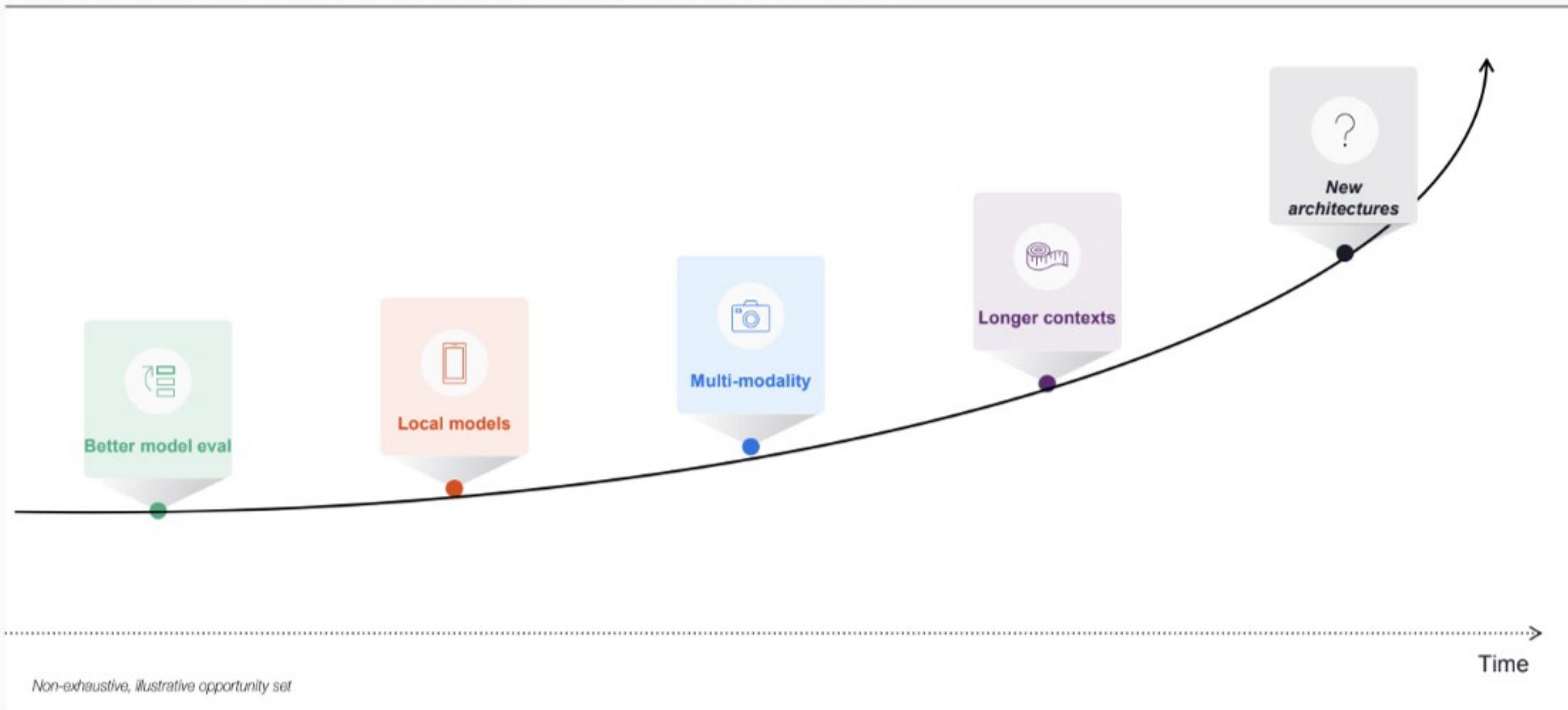


V5.2 (Jun 2023)

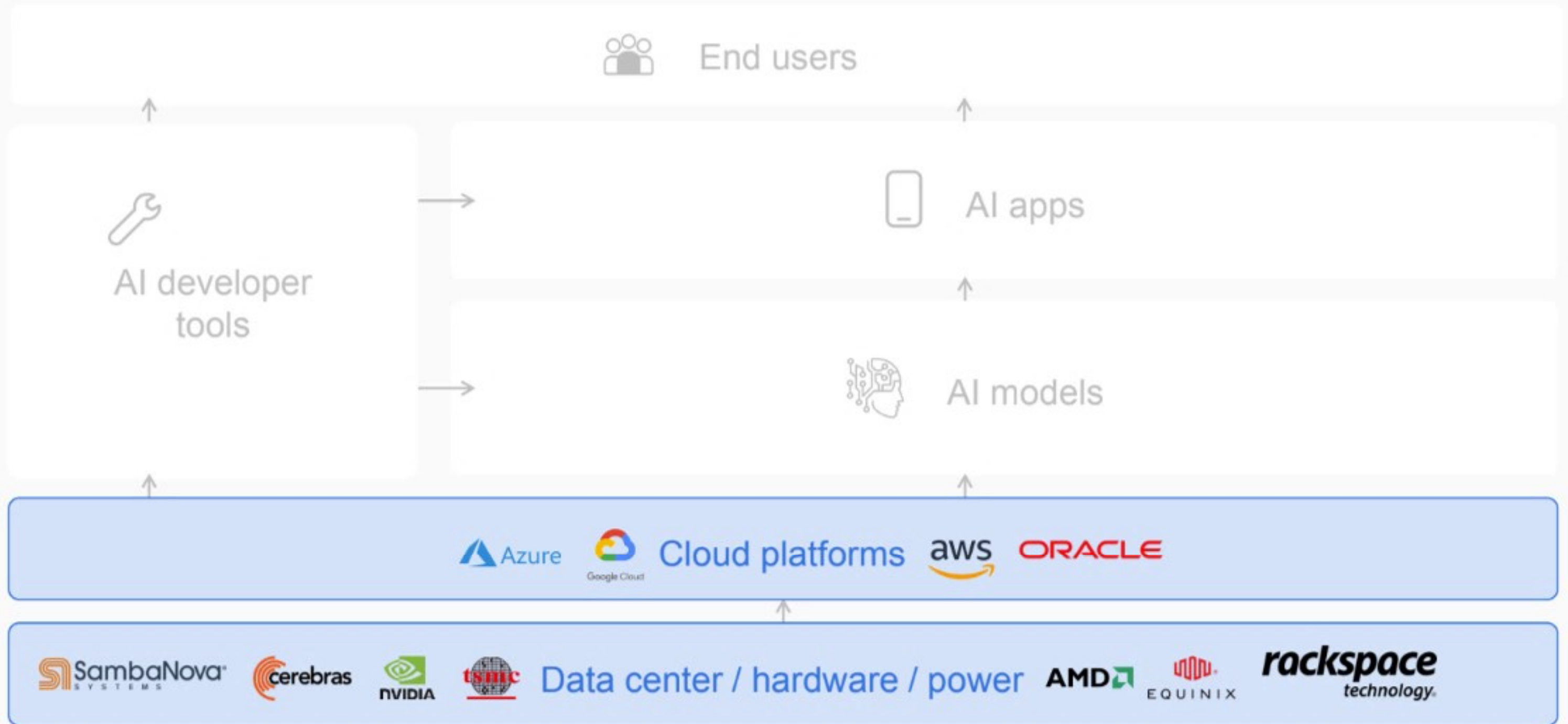
Note:
Midjourney improvement is partially due to collecting user feedback but is only one part of the equation!

There are many other opportunities to improve models!

→ **Beyond scaling parameters & data, there are many angles to push on**



AI has potential to re-accelerate underlying infrastructure



This AI wave has been extremely compute hungry

→ **ML models are doubling compute needs for training in months**

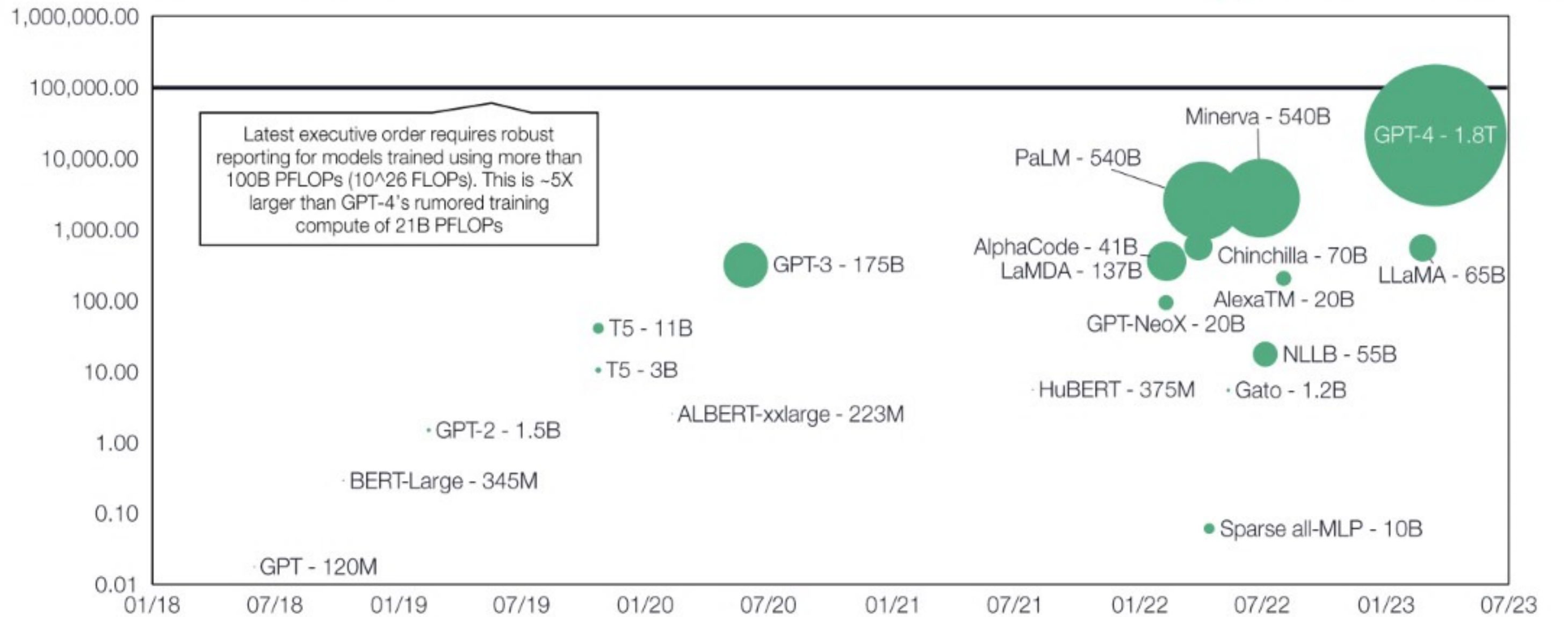
Months to double FLOPS used in model training



Training compute has exponentially increased with model sizes

→ **Massive inflection in compute needs for AI, ~70X increase from GPT-3 to GPT-4**

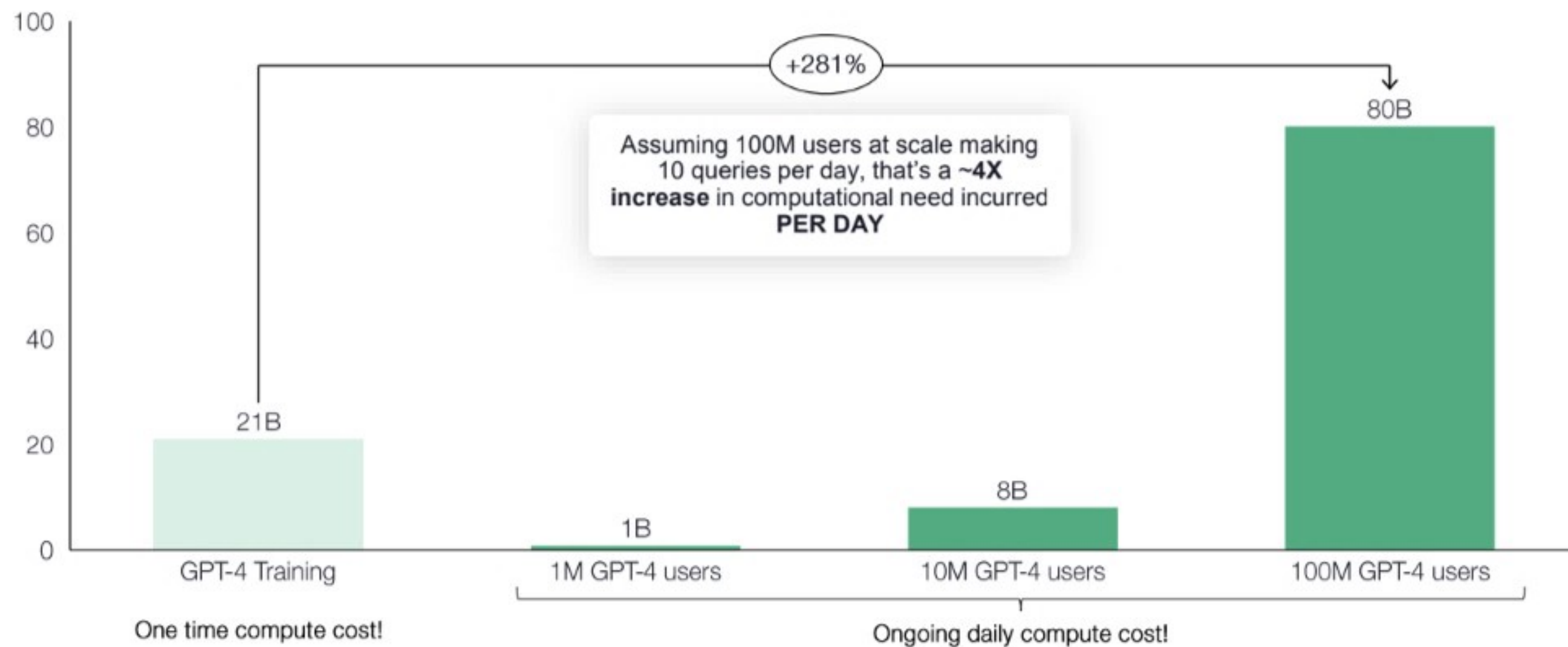
Computational Petaflops (M, log scale)



Inference compute likely to dramatically outpace training

→ Ongoing inference likely to require much more compute than one-time training

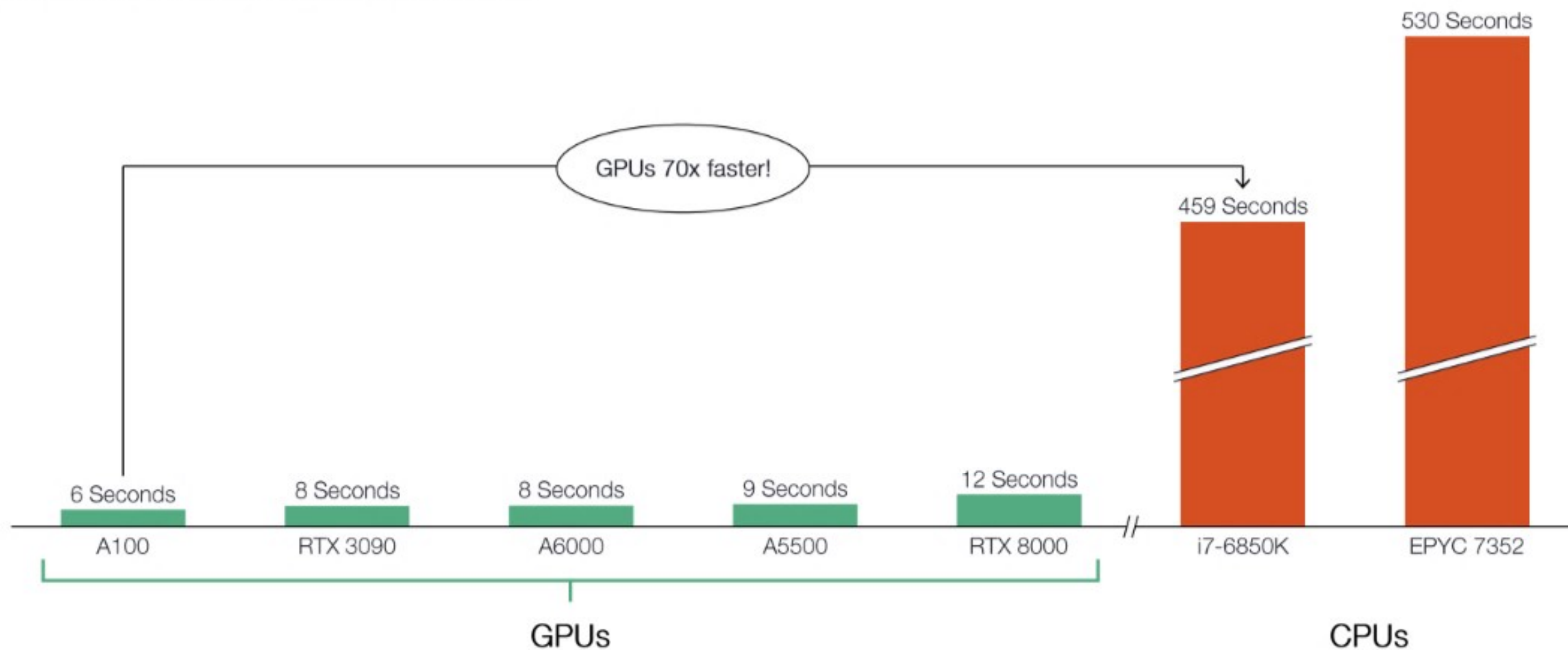
Computational Petaflops (B)



AI compute has been primarily served by GPUs

→ **GPUs are able to serve AI workloads much faster than CPUs**

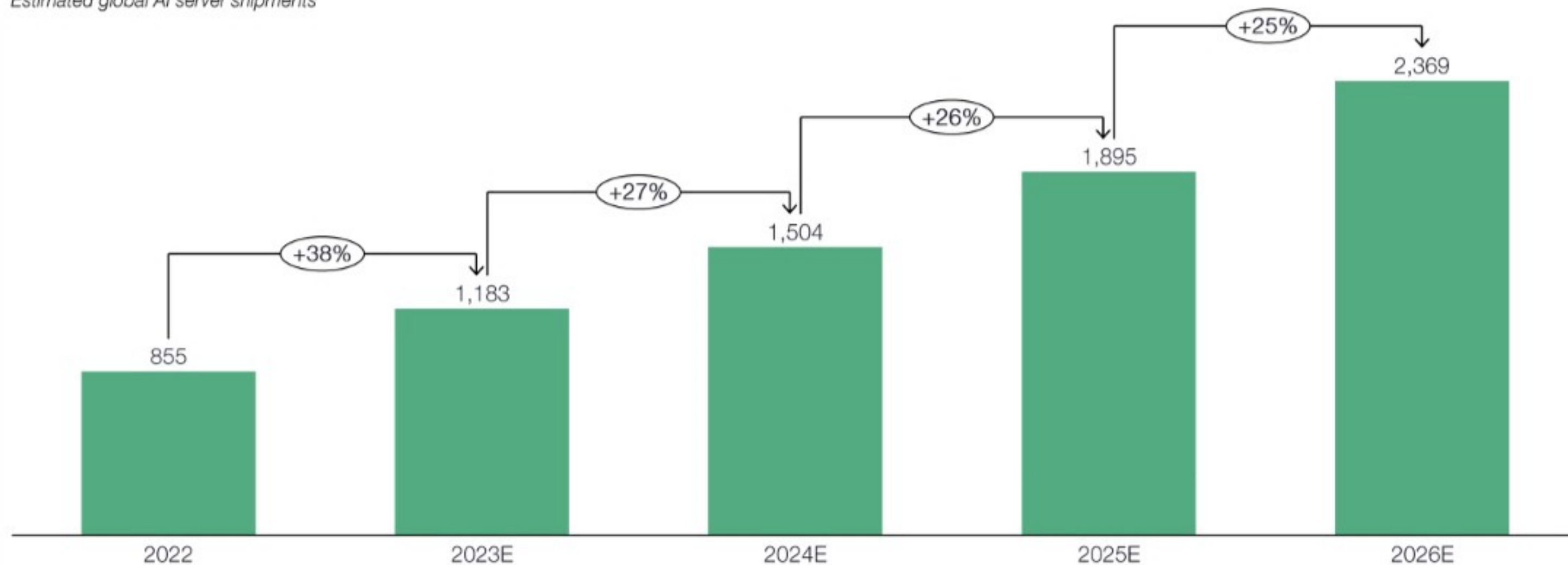
Time to produce an image with Stable Diffusion (lower is better)



The demand for GPUs has only begun

→ **Estimated ramp up for AI servers growing 25%+ Y/Y through 2026**

Estimated global AI server shipments



What are the implications for this global shift in demand for GPUs?

Follow the GPUs: AI poised to have impacts across our economy

\$32B+

NVDA Revenue
last 12 months



Power

Cloud

Data Centers

Semis

The AI wave may stress our power grid

\$32B+

NVDA Revenue
last 12 months



2-3x

Electrical transformer
price increases

&

3-5 Years

Leadtime to connect to
the grid

3-5 GW

More data center power
required '23-24E

=

50%

Potential increase in power
demand by 2026

More GPUs likely means more expensive servers

\$32B+

NVDA Revenue
last 12 months



Higher Server Bill of Materials Due to GPUs '23-24E

Note: all estimated numbers

	<u>Non-AI Server</u>		<u>AI Server</u>
CPU	~\$2K	10x	~\$20K
GPU			~\$200K
Storage & RAM	~\$5k	5x	~\$20K
Power Supply	<\$1K	10x	~\$5K
Networking & Other	~\$5K	3x	\$15k
Server Cost	~\$10K	26x	~\$260K

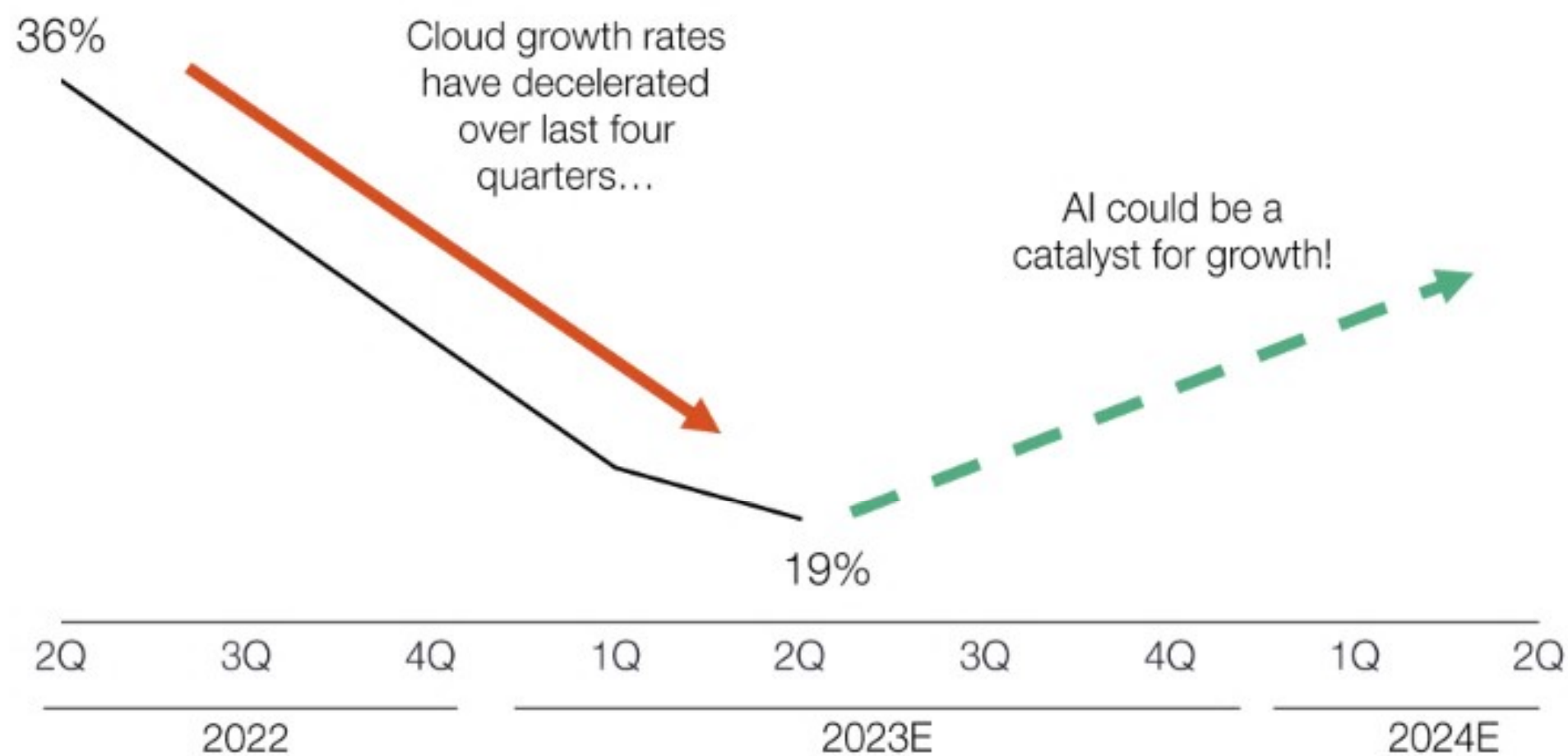
Could AI workloads drive a cloud reacceleration?

\$32B+

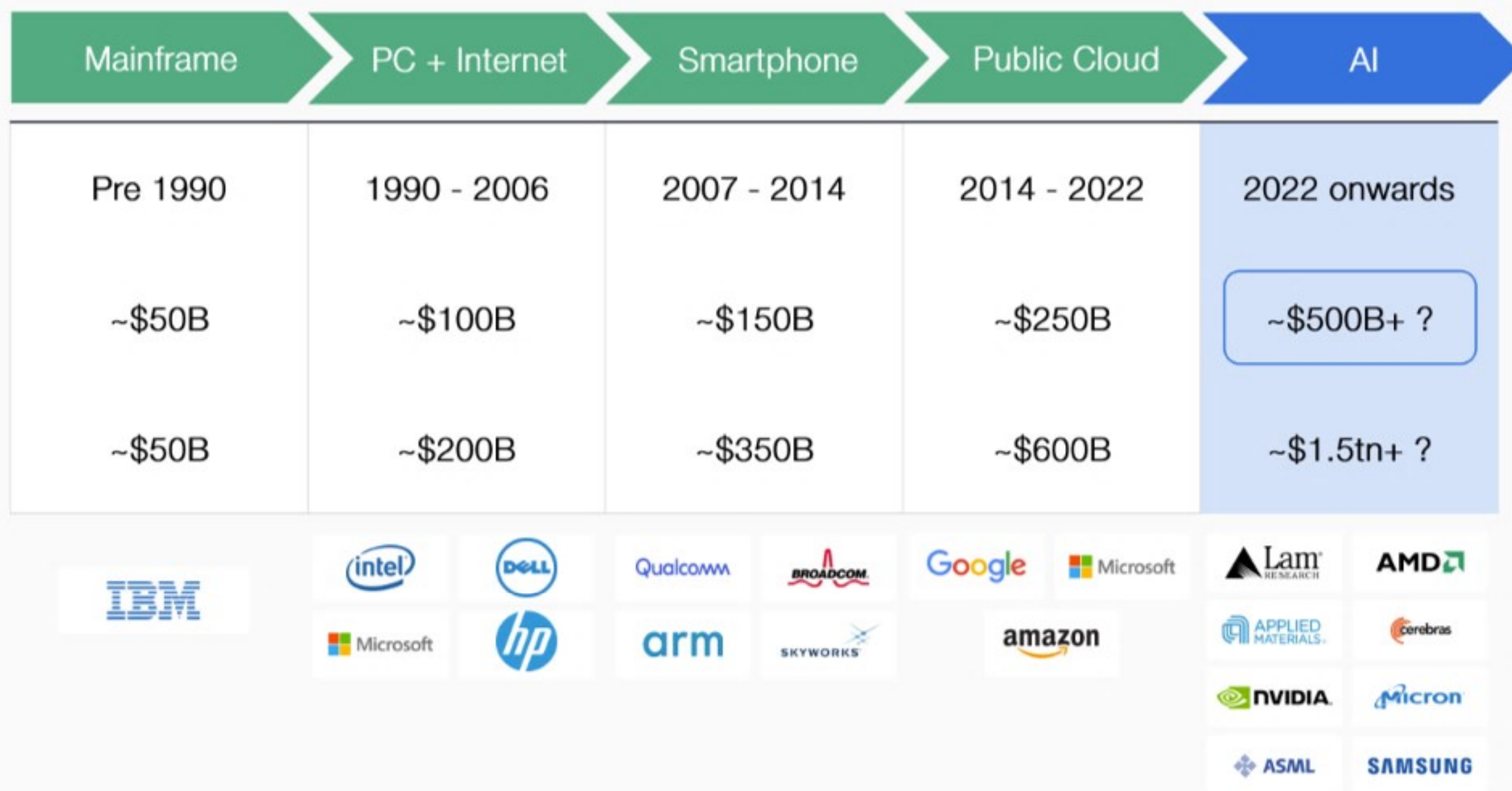
NVDA Revenue
last 12 months



Y/Y revenue growth rates for AWS, GCP, and Azure from Q2 2022 to Q2 2023

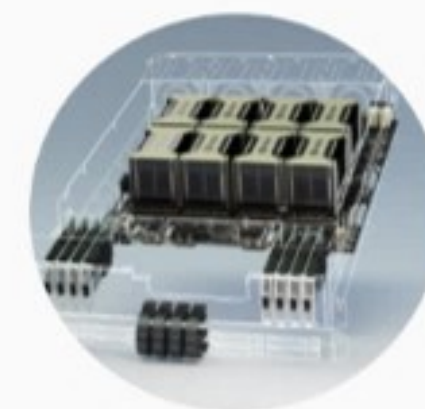


AI is the latest tech wave to inflect the semiconductor industry



The entire semiconductor supply chain has potential to benefit

Semis value chain across manufacturing inputs to device assembly

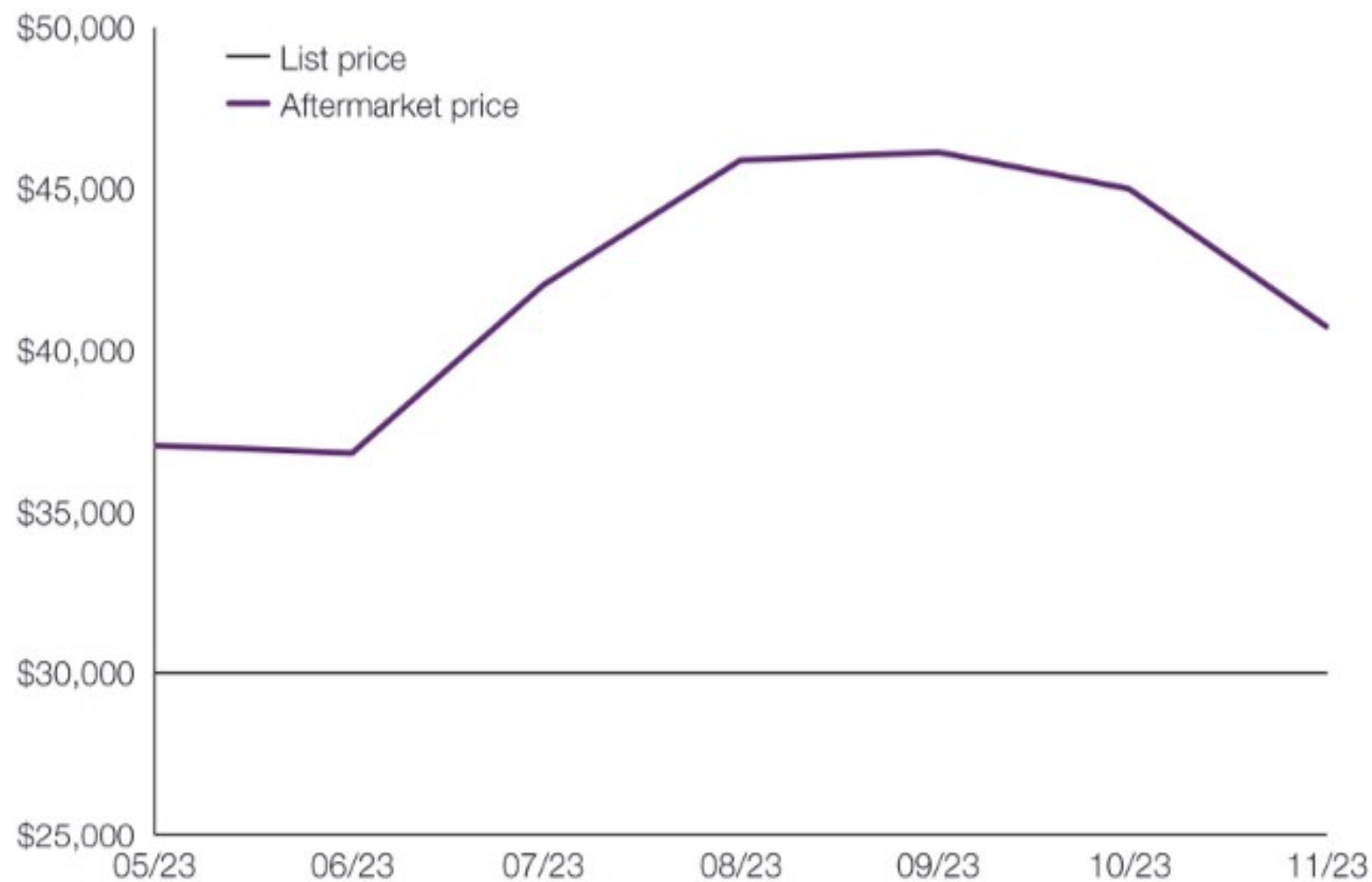


Data center servers powering AI!

Demand for cutting edge GPUs has exceeded supply

→ **H100 GPU prices on aftermarket are typically well above list price**

Pricing of individual H100 80GB GPU units on eBay & Amazon, not reflective of cluster availability



Tesla H100 80GB NVIDIA Deep Learning GPU Compute Graphics Card 900-21010-000-000

Brand New - NVIDIA

\$42,750.00

Was: \$45,000.00 5% off
or Best Offer
Free shipping
from China



Tesla H100 80GB NVIDIA Deep Learning GPU Compute Graphics Card

Brand New - NVIDIA

\$42,672.00

or Best Offer
Free shipping

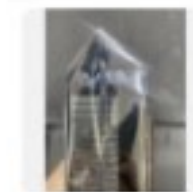


Nvidia H100-PCIe-80GB Hopper H100 80GB PCIe Tensor Core GPU Accelerator

Pre-Owned - NVIDIA - 80 GB

\$39,995.00

or Best Offer
Free shipping
2 watchers



NVIDIA H100 80GB Tesla Deep Learning GPU Compute Graphics Card 900-21010-000-000

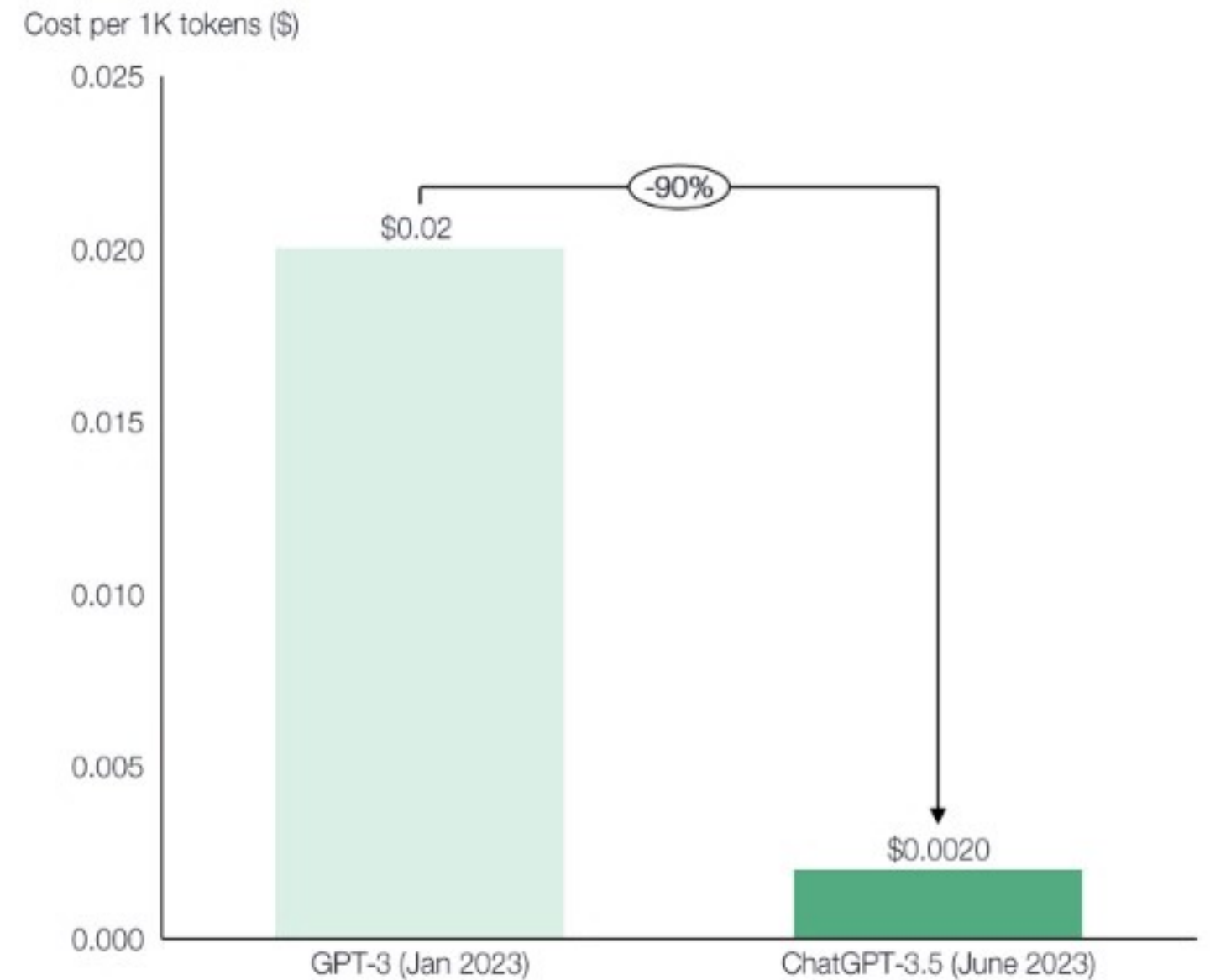
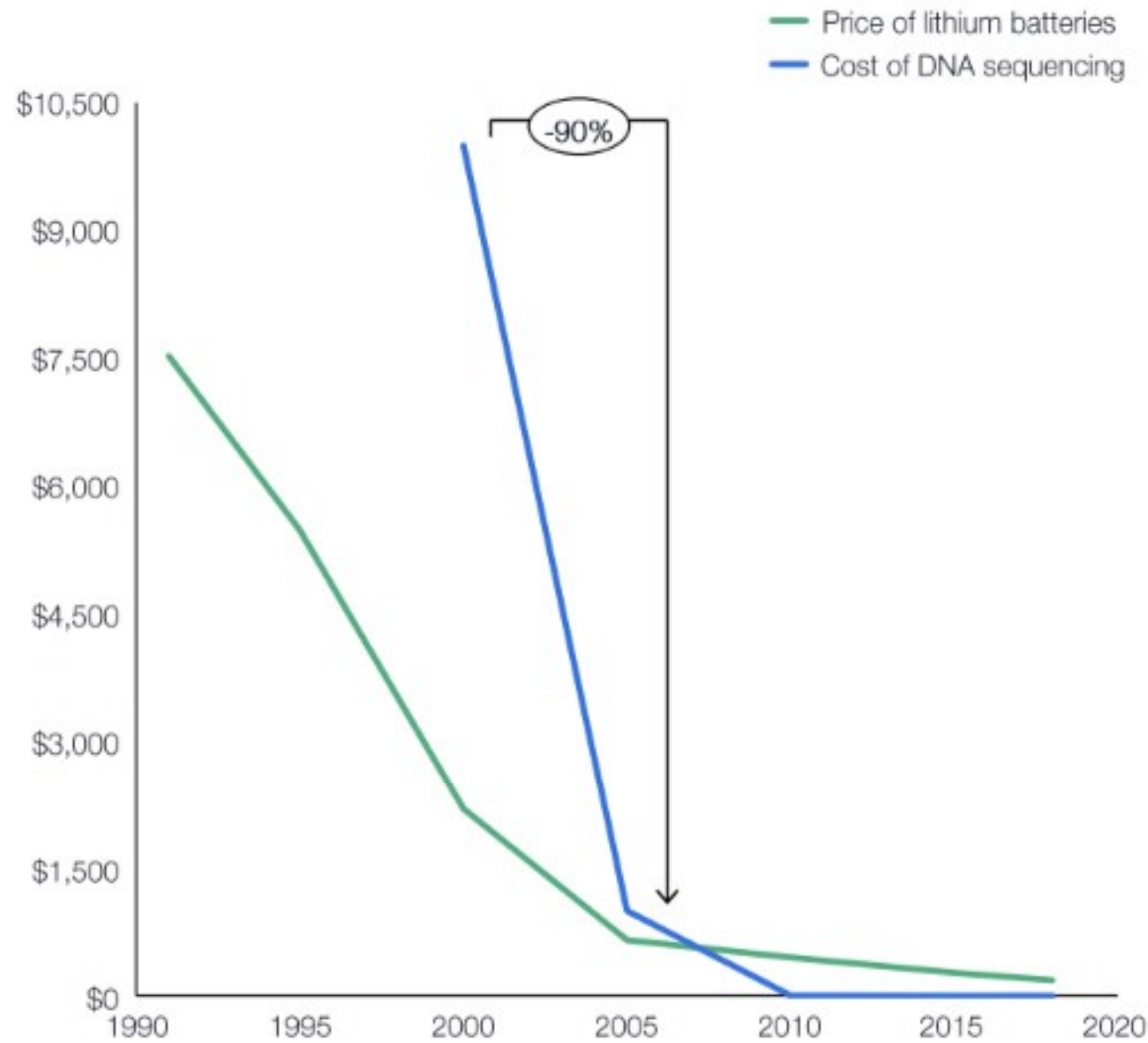
Brand New - NVIDIA

\$45,000.00

Despite intense demand, AI compute costs have decreased

→ Could AI compute resemble cost curves in other sectors?

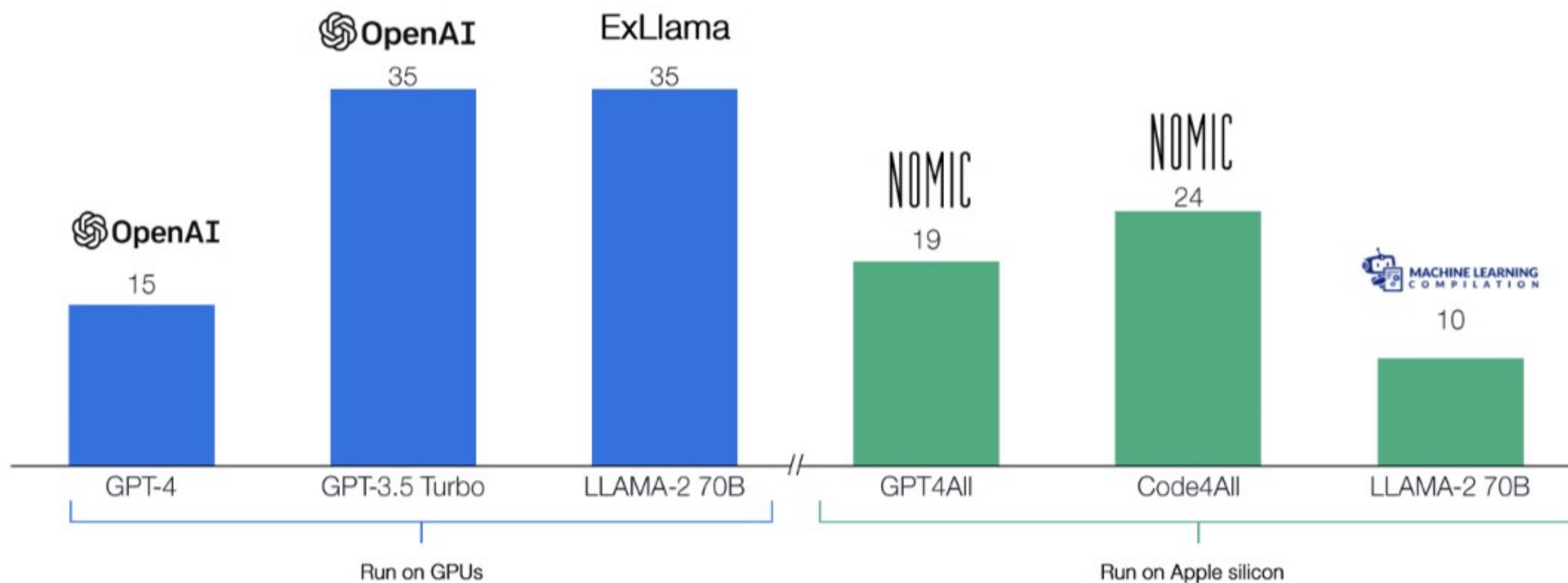
→ Cost of running GPT-3/3.5 down 90% in 5 months!



Models on edge could alleviate GPU shortage as well

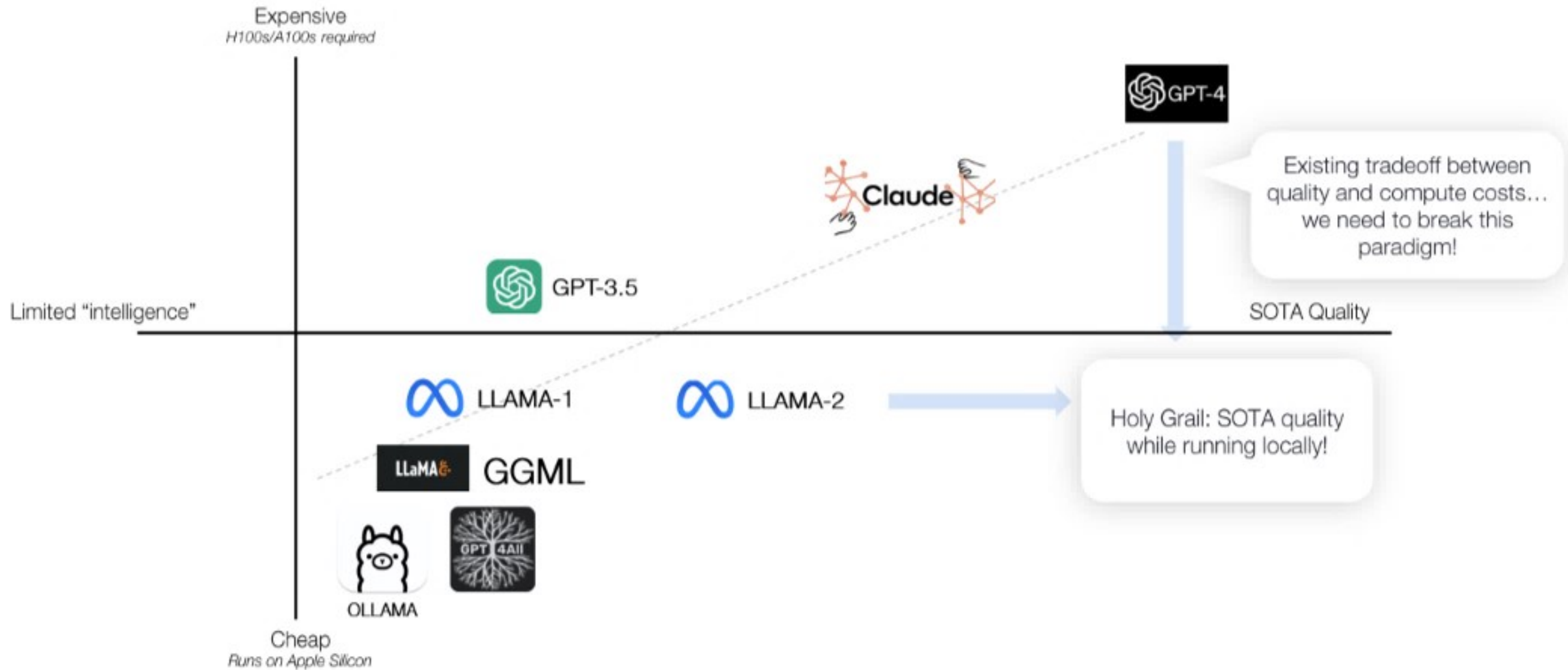
→ Local models on Apple silicon becoming as fast as models running on GPUs

Tokens per second (higher the better) of LLMs

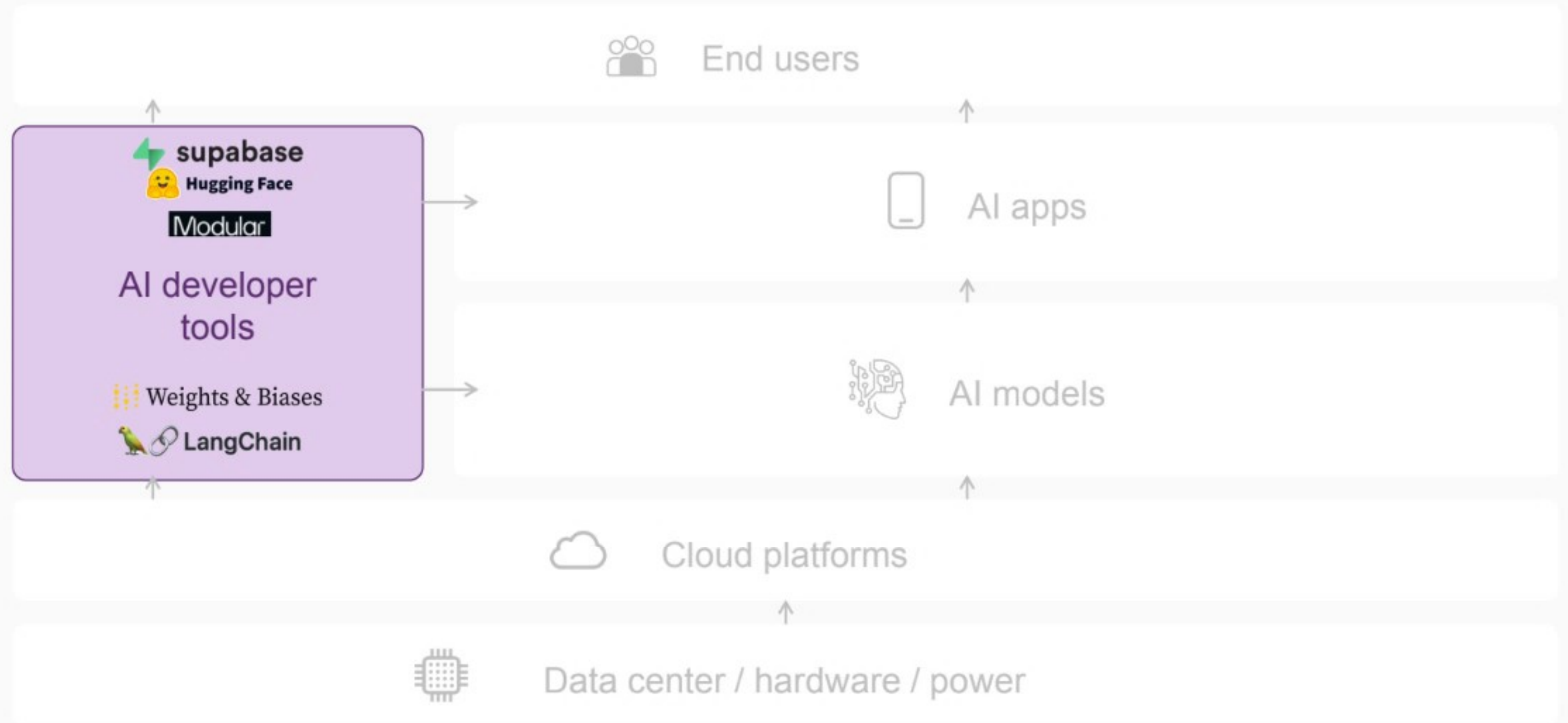


Will we reach AI's Holy Grail?

→ Largest, most expensive models are still most performant...for now



AI has started to create a new ecosystem of tools



AI Ops is a new category of tools for AI engineers

Illustrative; non-exhaustive

ML Ops: Focused on creating ML models

Data labeling  Snorkel  DataRobot  scale

Model hub  Hugging Face  mlflow  GitHub

 Lightning**  Weights & Biases

Model training & development  PyTorch  DataRobot

 databricks  mlflow  TensorFlow
 ABACUS.AI




Model inference

 databricks  anyscale
 Amazon SageMaker  SELDON

Monitoring







 Arthur  fiddler  Weights & Biases
 GANTRY  DataRobot


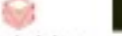


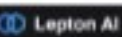
AI Ops: Focused on operationalizing Generative AI



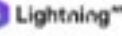


Data labeling & curation  Snorkel  Argilla  Cleanlab

 NIMIC ATLAS  scale

Model hub  Hugging Face  LatchBio  roboflow  Epicor

Fine tuning  Hugging Face  databricks  scale
 Humanloop  surge**  Snorkel

Model inference  Hugging Face  databricks  Modal
 Epicor  Lepton AI

LLM Ops  Bramble  Weights & Biases  Lightning**  LangChain
 ABACUS.AI

Vector DBs  supabase  Pinecone  Weaviate  NIMIC ATLAS

Monitoring & evaluation  Arthur  Weights & Biases  GANTRY
 Zeno  Humanloop

Output Guardrails  Arthur  Nomos AI

1

Beyond labeling, data curation is becoming more important as engineers try to enhance model performance and alignment.

2

New tools enabling fine tuning and inference have become easier for the AI developer to use

3

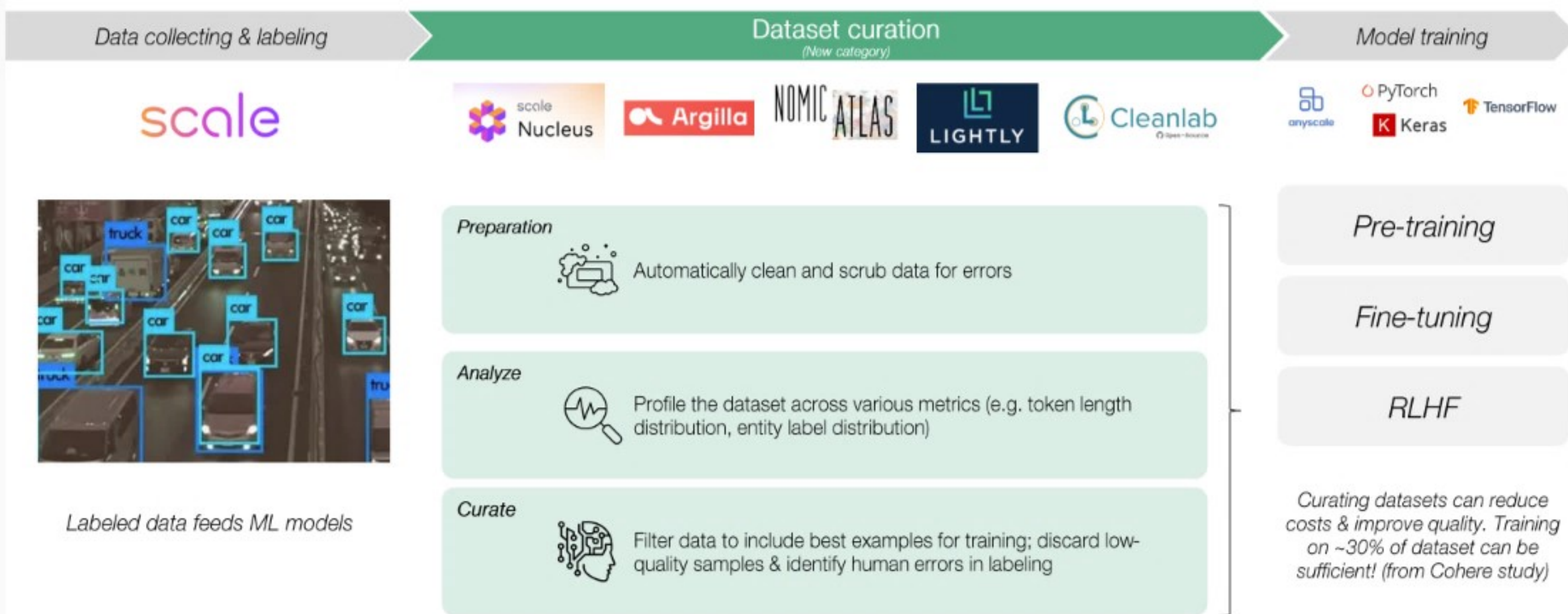
Emerging category of tools enabling developers to use LLMs more effectively; Vector DBs have exploded in popularity due to embeddings.

4

Securing the last mile deployment of LLMs through model evaluation, guardrails, and ongoing RLHF is supporting a new set of companies

Data curation tools will be critical for improving models

→ Data curation an emerging new category between data collection & labeling, and model training



Fine tuning models has become much more accessible

Hugging Face Autotrain lets you fine tune open-source models on your own data in just a few clicks

1

Select type of task & upload training dataset

New project

Select a task, language, and how many models you want to train. You will prepare data in the next step.

Project name

eg: imdb-sentiment-analysis

Task

Text Classification (Binary)

Training files

IMDB_train.csv

DONE - undefined - Updated Jul 6

text: review target: sentiment

2

Hugging Face suggests suitable models

93% ACCURACY ✓ FINISHED

#329982

pessimistic-aardvark

Finished in 27 minutes

91.2% ACCURACY ✓ FINISHED

#329981

equatorial-raccoon

Finished in 16 minutes

89.4% ACCURACY ✓ FINISHED

#329978

elliptical-caterpillar

Finished in 56 minutes

88.4% ACCURACY ✓ FINISHED

#329977

agreeable-dunlin

Finished in 17 minutes

3

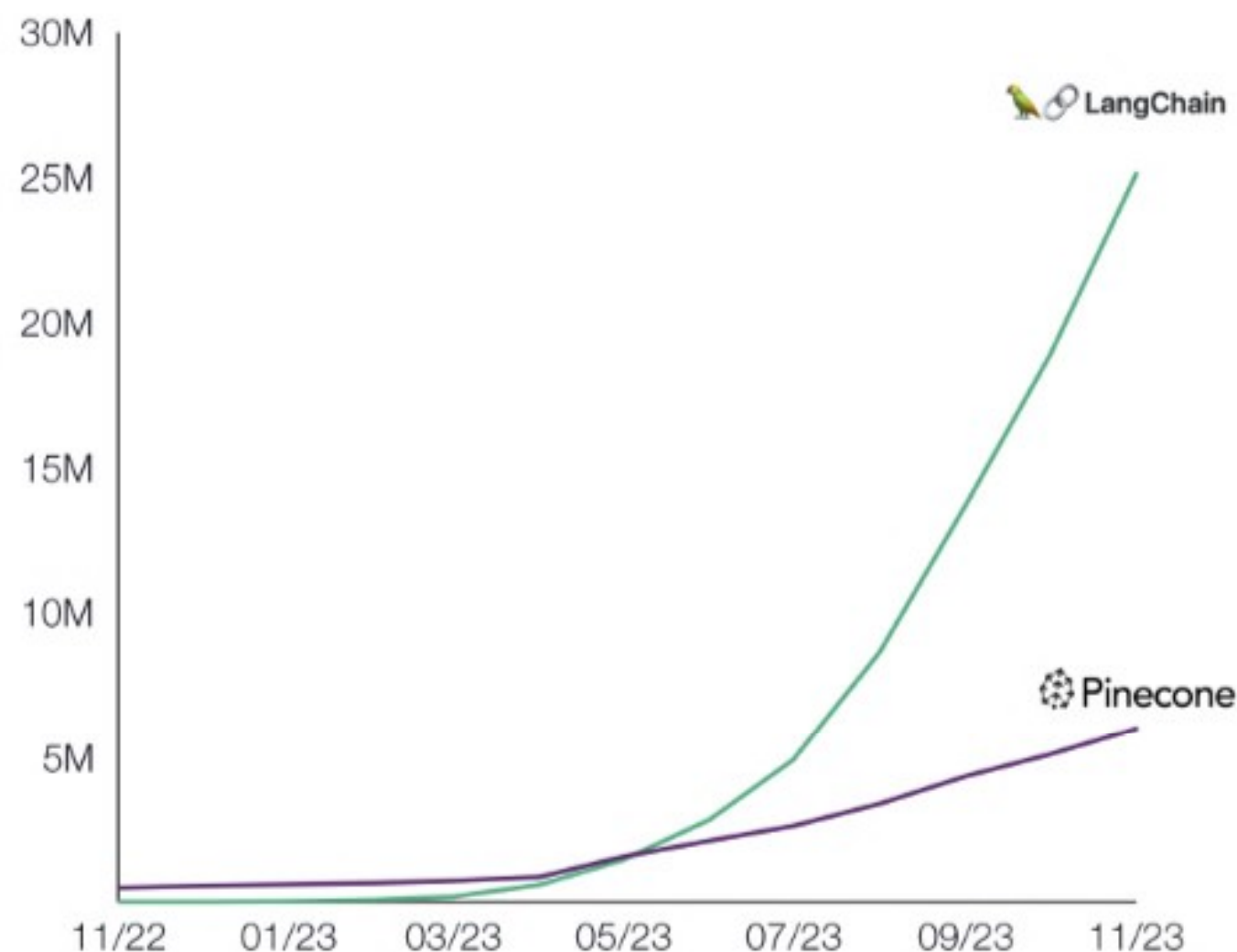
Select model(s) and evaluate model training progress

Model ID	Loss	Accuracy
#329982 pessimistic-aardvark	0.2462	0.9300
#329971 baggy-woodpecker	0.2109	0.9216
#329970 this-leopard	0.2453	0.9189
#329981 equatorial-raccoon	0.2212	0.9117
#329974 regal-termite	0.2606	0.9091
#329968 fearful-lobster	0.2448	0.9089
#329975 unlucky-cobra	0.2813	0.9022
#329972 far-badger	0.2555	0.8981

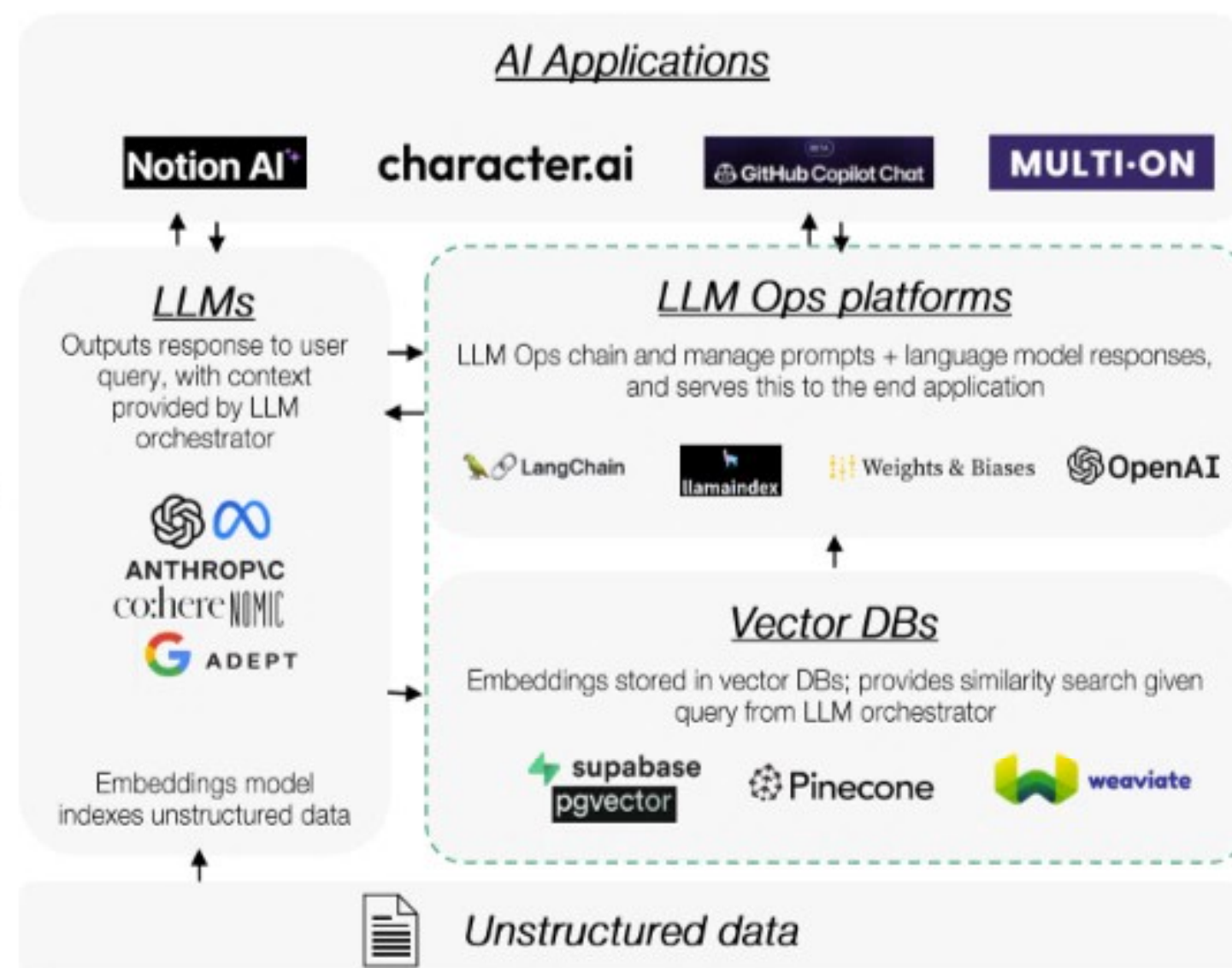
LLM Ops & Vector DBs are a new enabling layer for AI apps

→ **Tools are already being put into production**

Cumulative installs (incl. machine)



→ **LLM Ops & Vector DBs enable retrieval augmentation**



Can VectorDBs be standalone winners?

→ Only three DB types have significant public outcomes

→ Startups have tried to create new categories of DBs

OLAP



Google BigQuery



amazon
REDSHIFT



ClickHouse



databricks

snowflake ~\$50B
mkt cap

Real-time streaming DBs

Materialize [ROCKSET]



No large
standalone
outcomes yet...

OLTP

CockroachDB



PostgreSQL



redis



Amazon
Aurora



MySQL

mongoDB. ~\$25B
mkt cap

Graph DBs

RelationalAI



Amazon Neptune



EDGE|DB



neo4j

Vector DBs



Milvus



weaviate



Pinecone



... Will vector DBs be
different?

Observability

splunk >



Grafana



elastic

new relic.

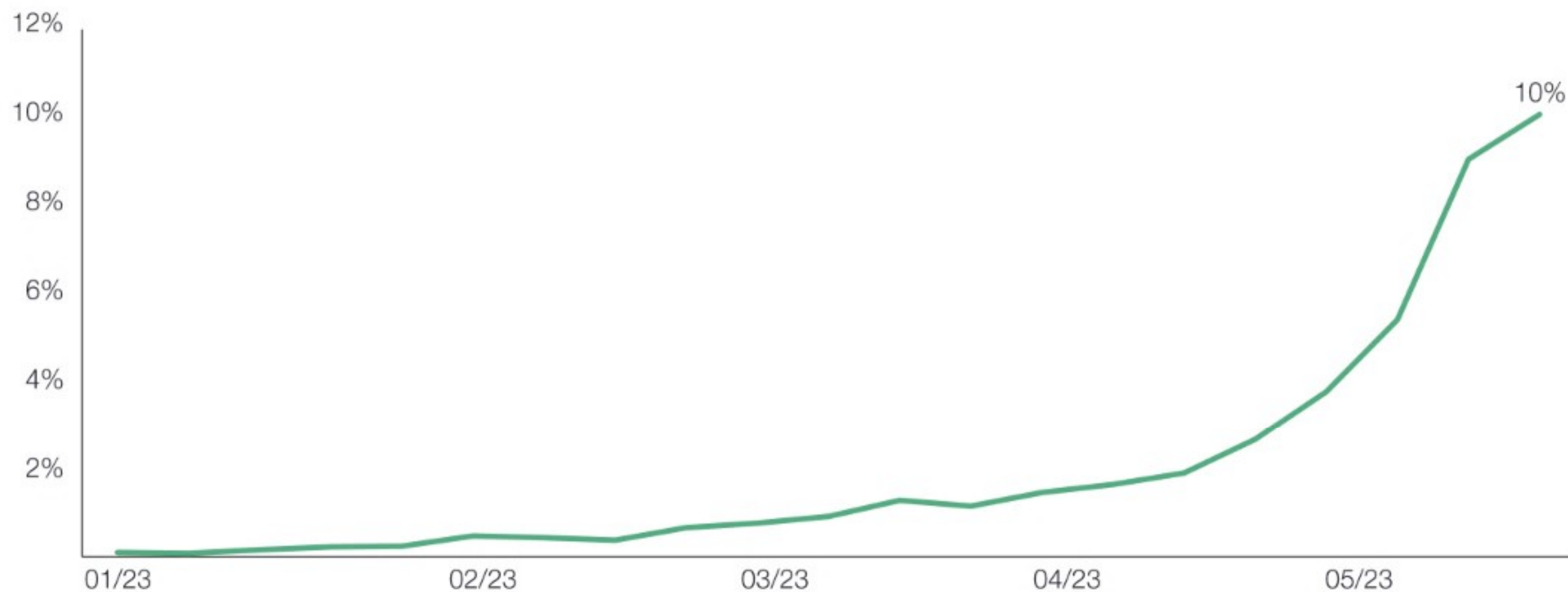


DATADOG

~\$25B
mkt cap

General purpose databases are being used as VectorDBs too

→ % of Supabase projects using pgvector



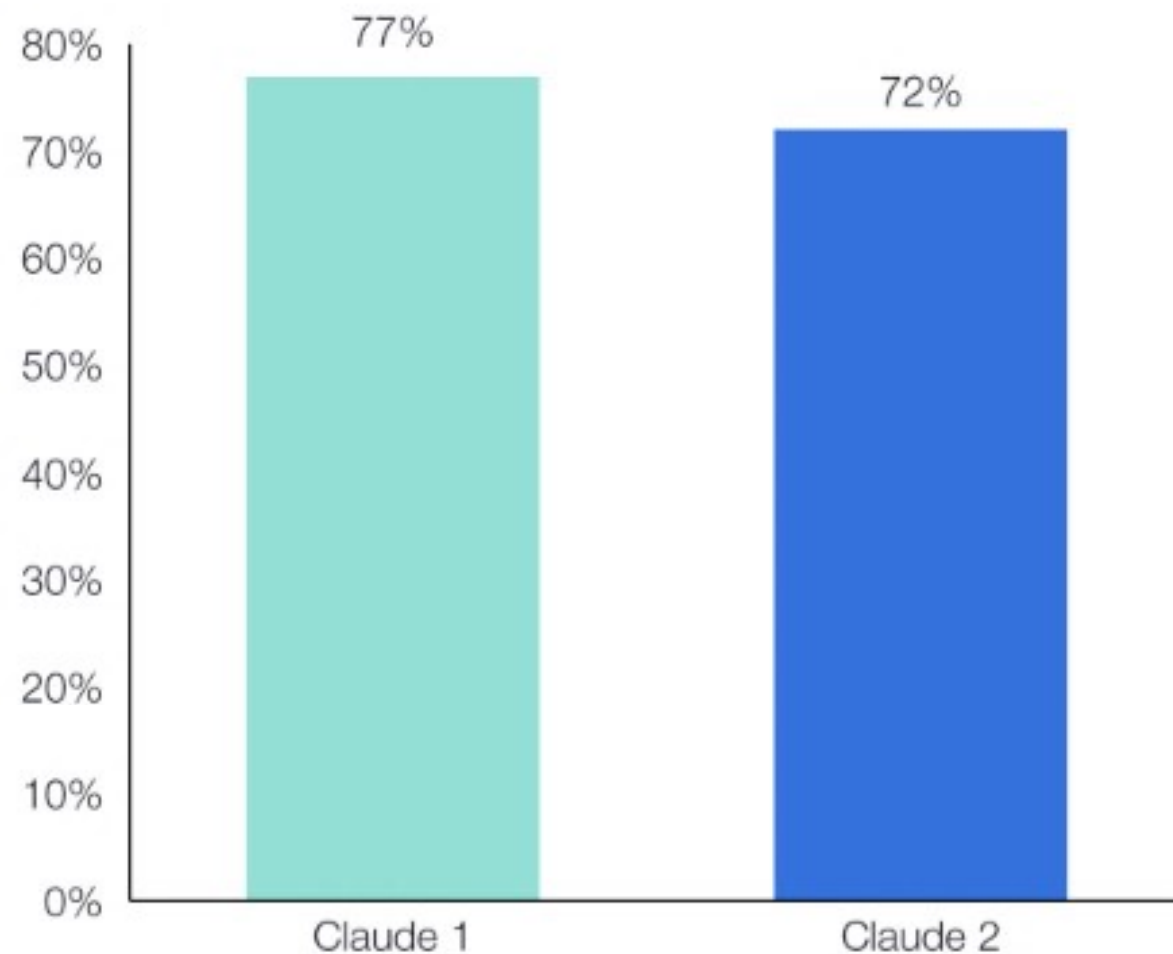
Problem: Model evaluation is broken today

→ Humans prefer Claude 1 to Claude 2...

Overall win rates (human preference) against other models



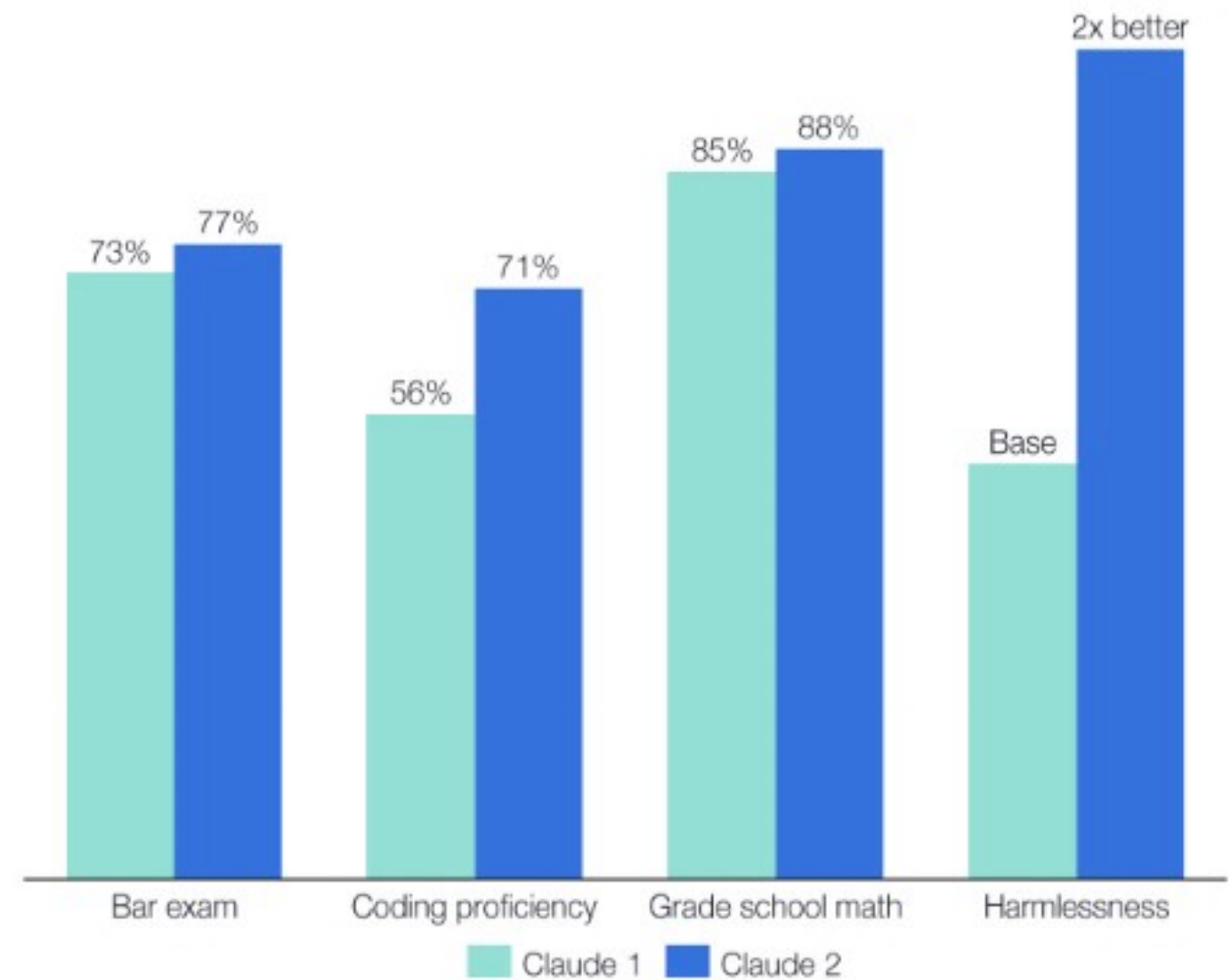
LMSYS



→ ...But Claude 2 performs better on benchmarks

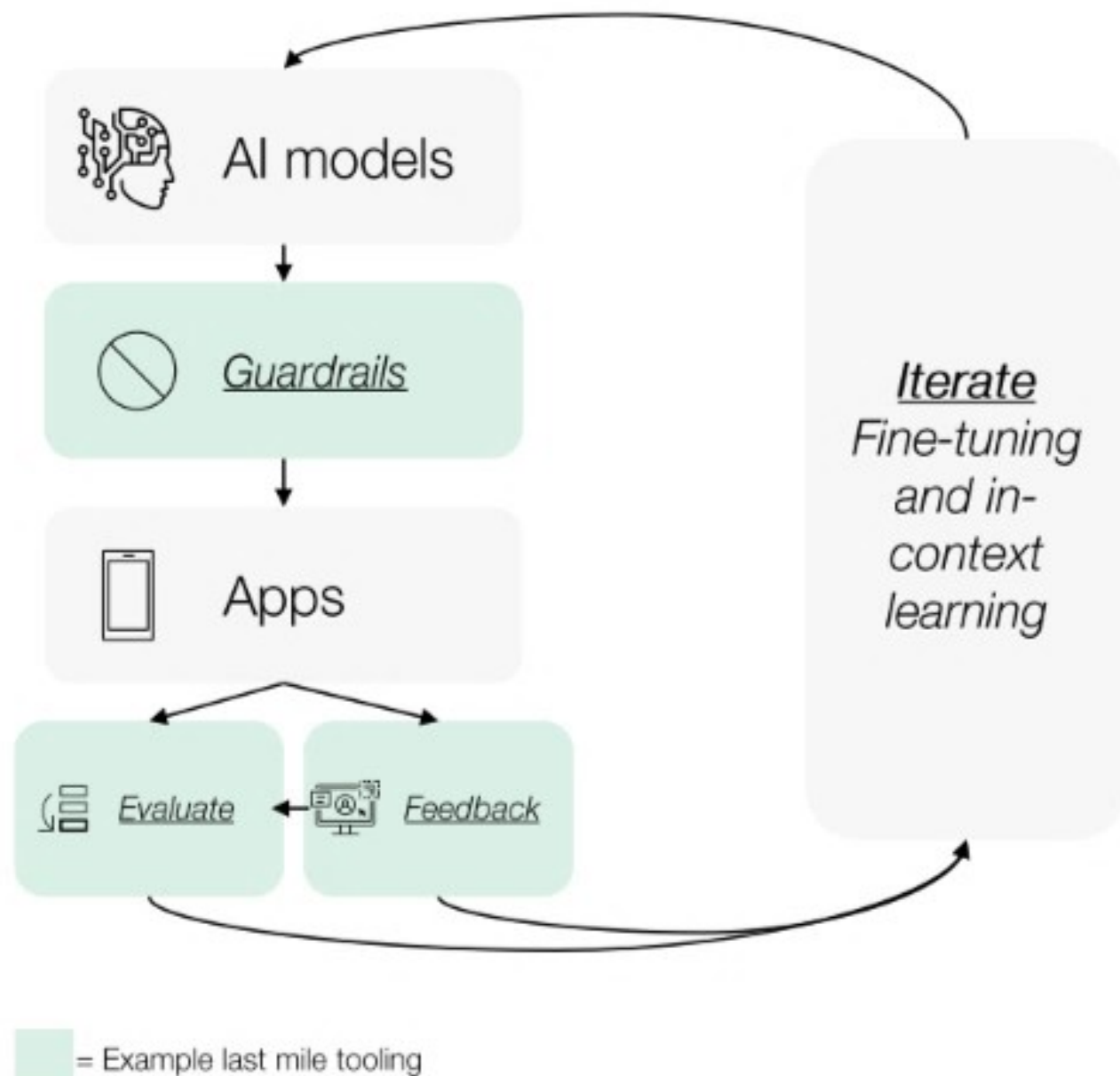
Performance on different benchmarks

ANTHROPIC



Evaluation is one of many 'last mile' tools needed for AI in production

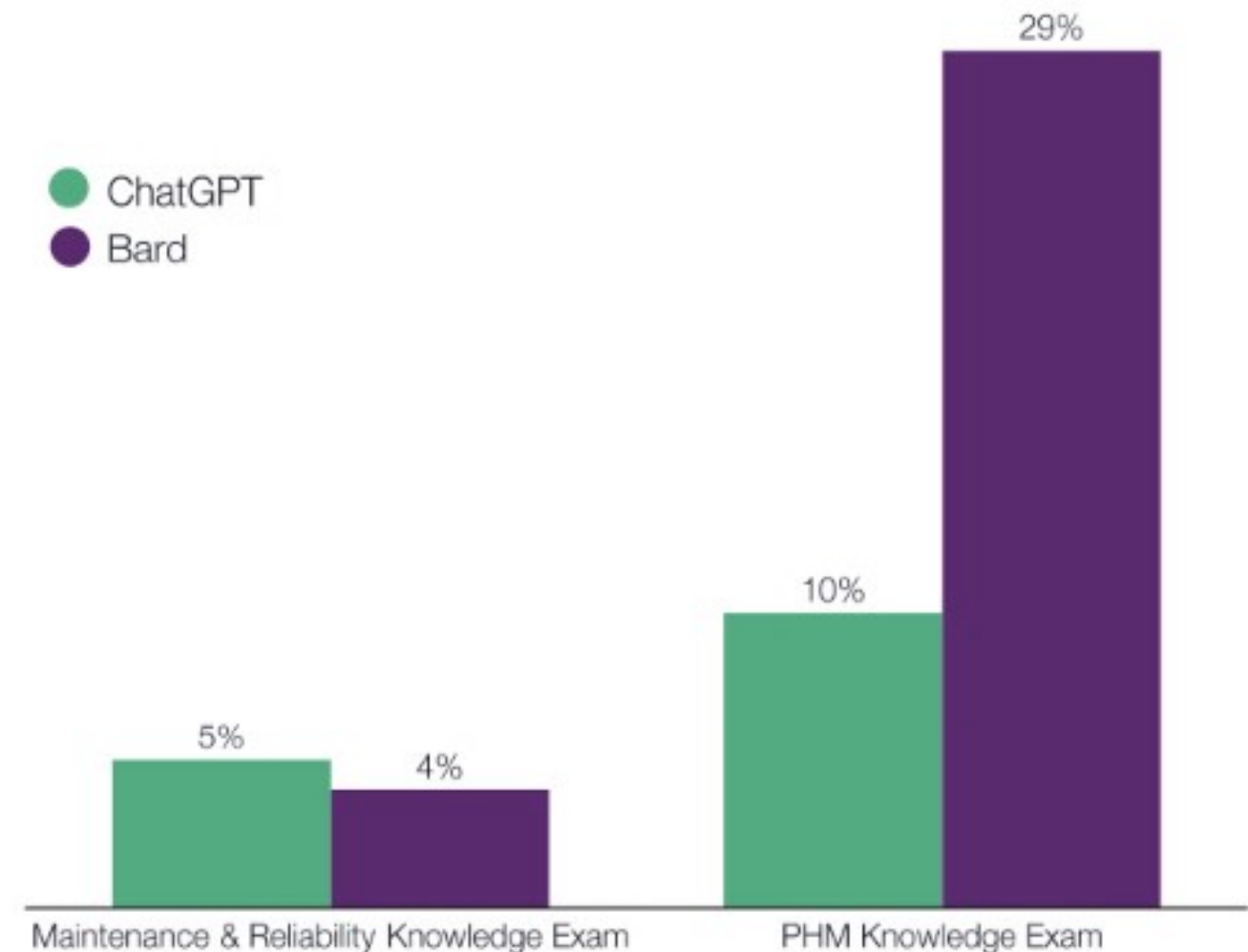
→ Last-mile tooling can improve AI accuracy and alignment



→ This may help prevent hallucinations in critical sectors!

Industrials case study: AI applied to prognostics & health mgmt. (PHM)

% of questions with hallucinated answers

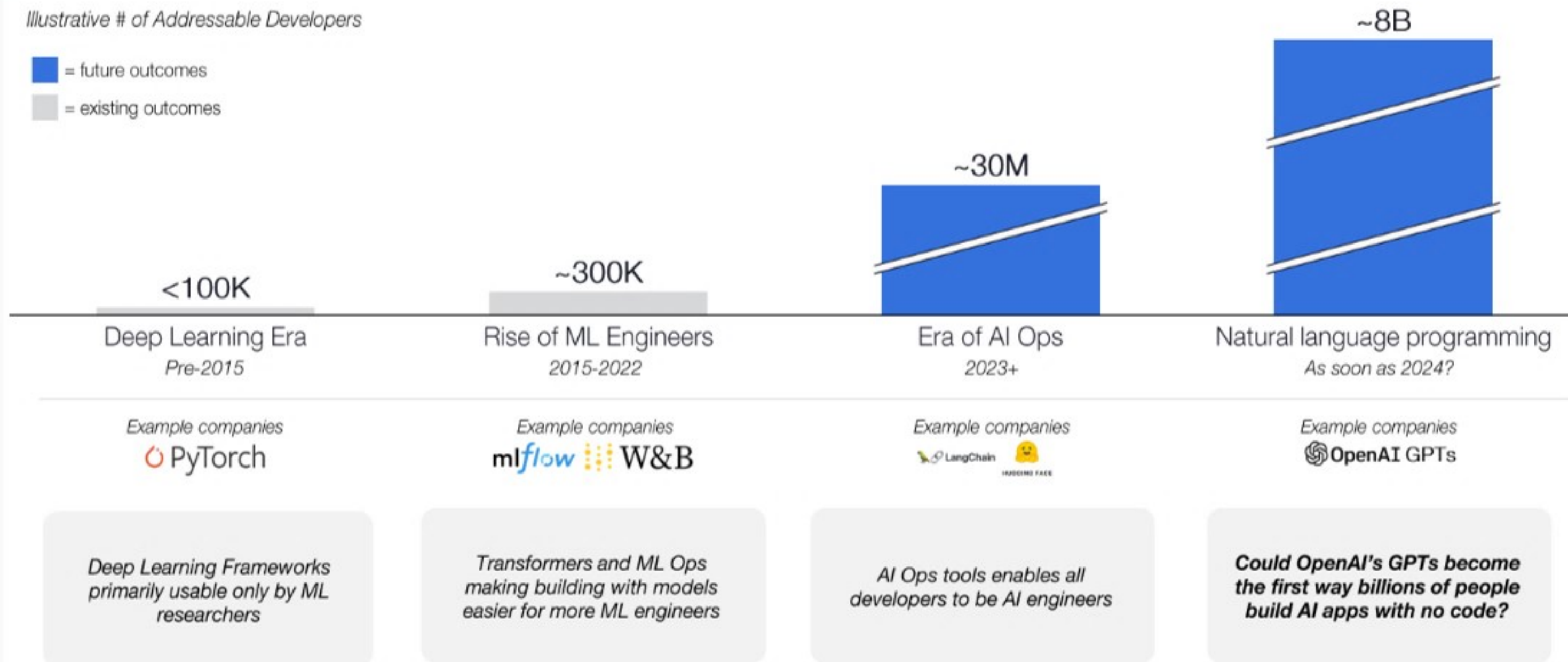


We're excited that AI Ops enables more developers to build apps

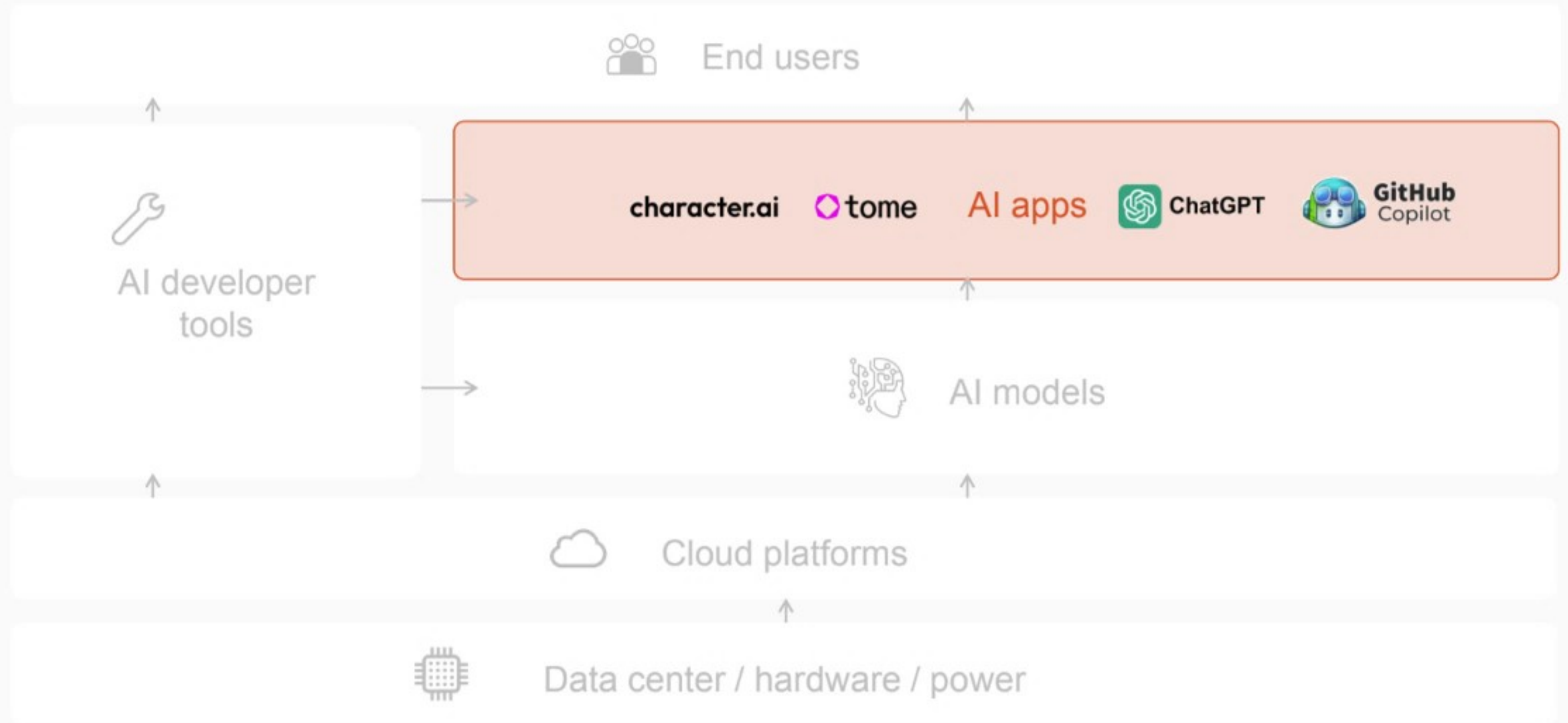
→ **New AI Ops tooling is accessible to all developers**

Illustrative # of Addressable Developers

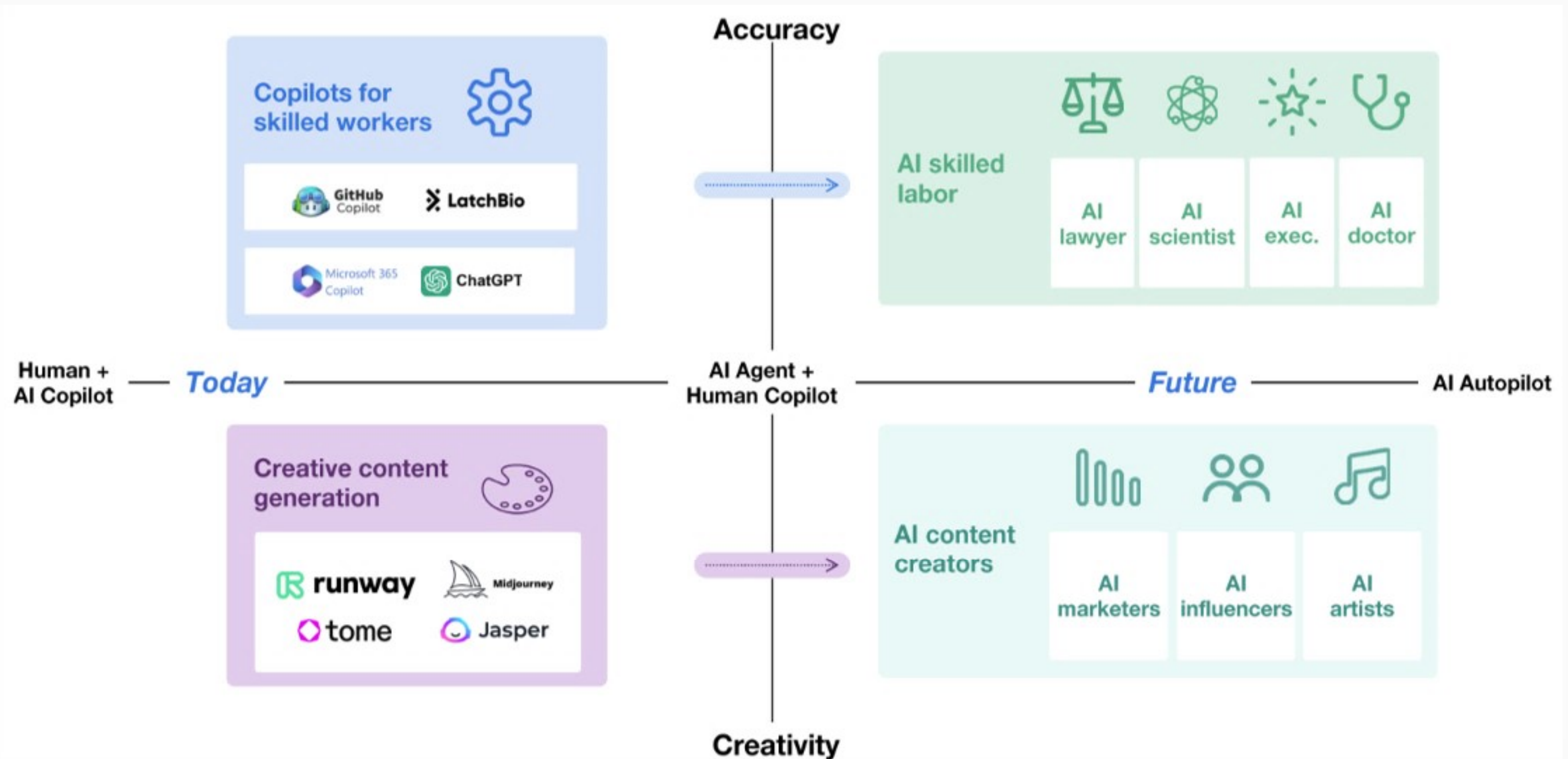
■ = future outcomes
■ = existing outcomes



The application layer is where humans will interact with AI

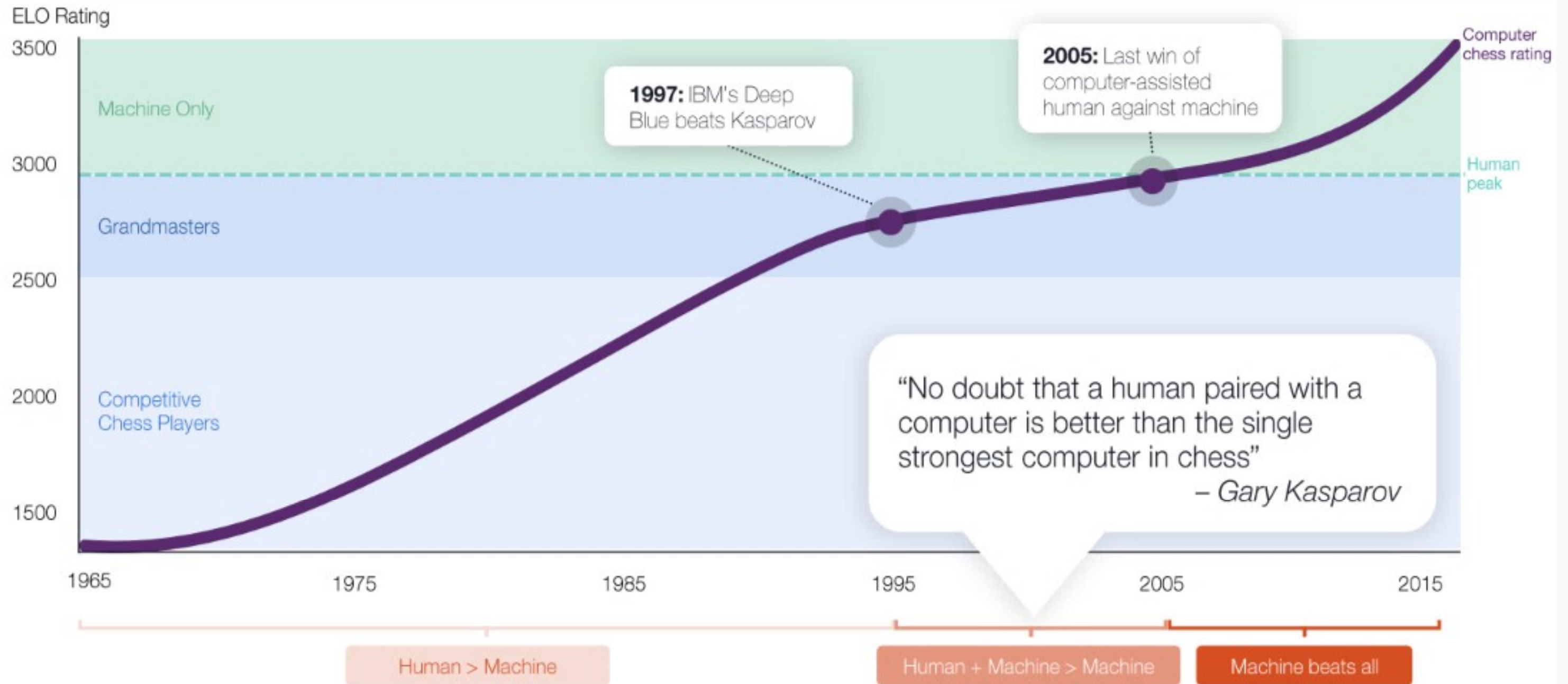


We expect AI to become even more impactful and autonomous



We have seen this before: Copilot phase was short-lived in chess

→ Computer chess rating

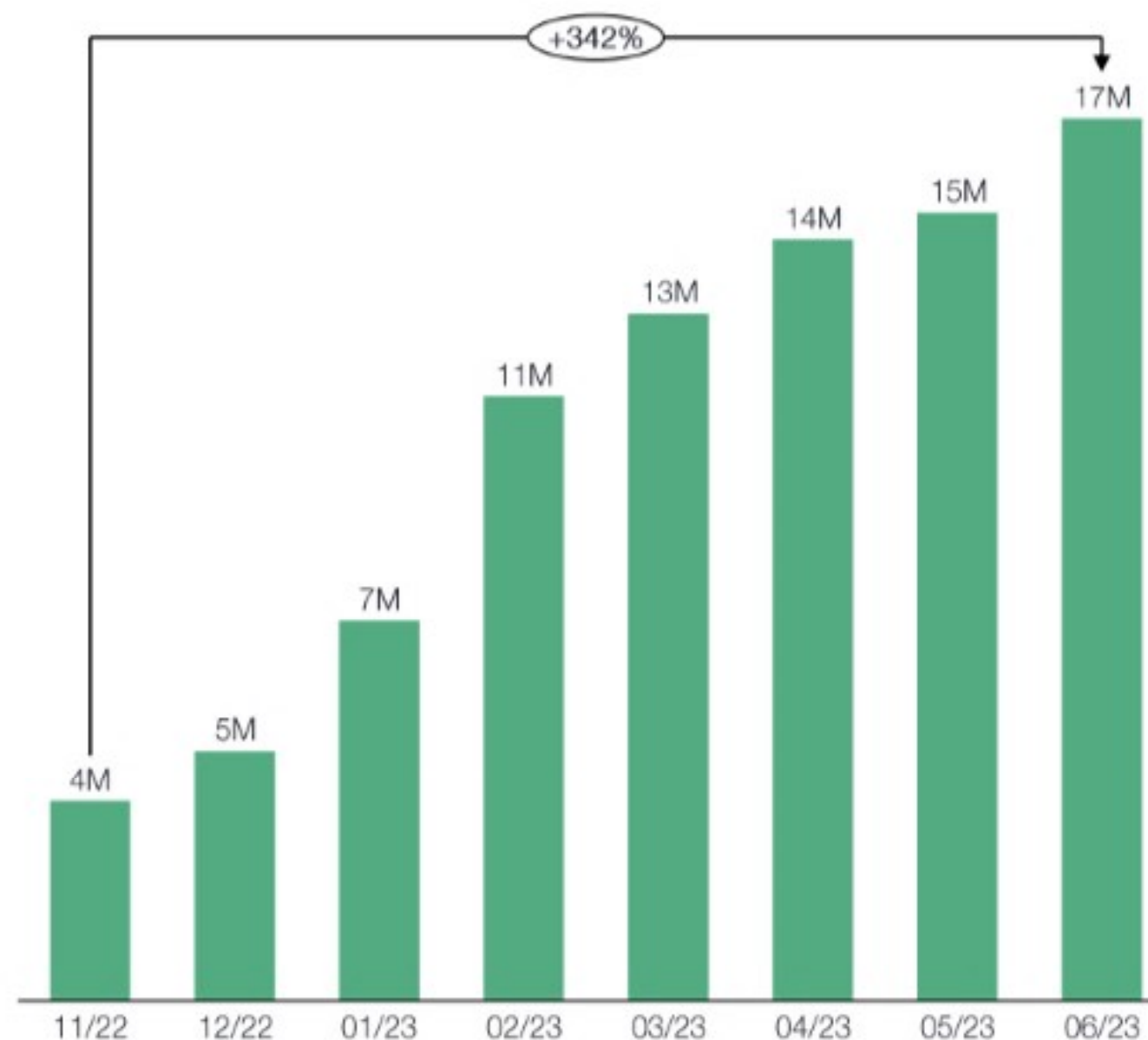
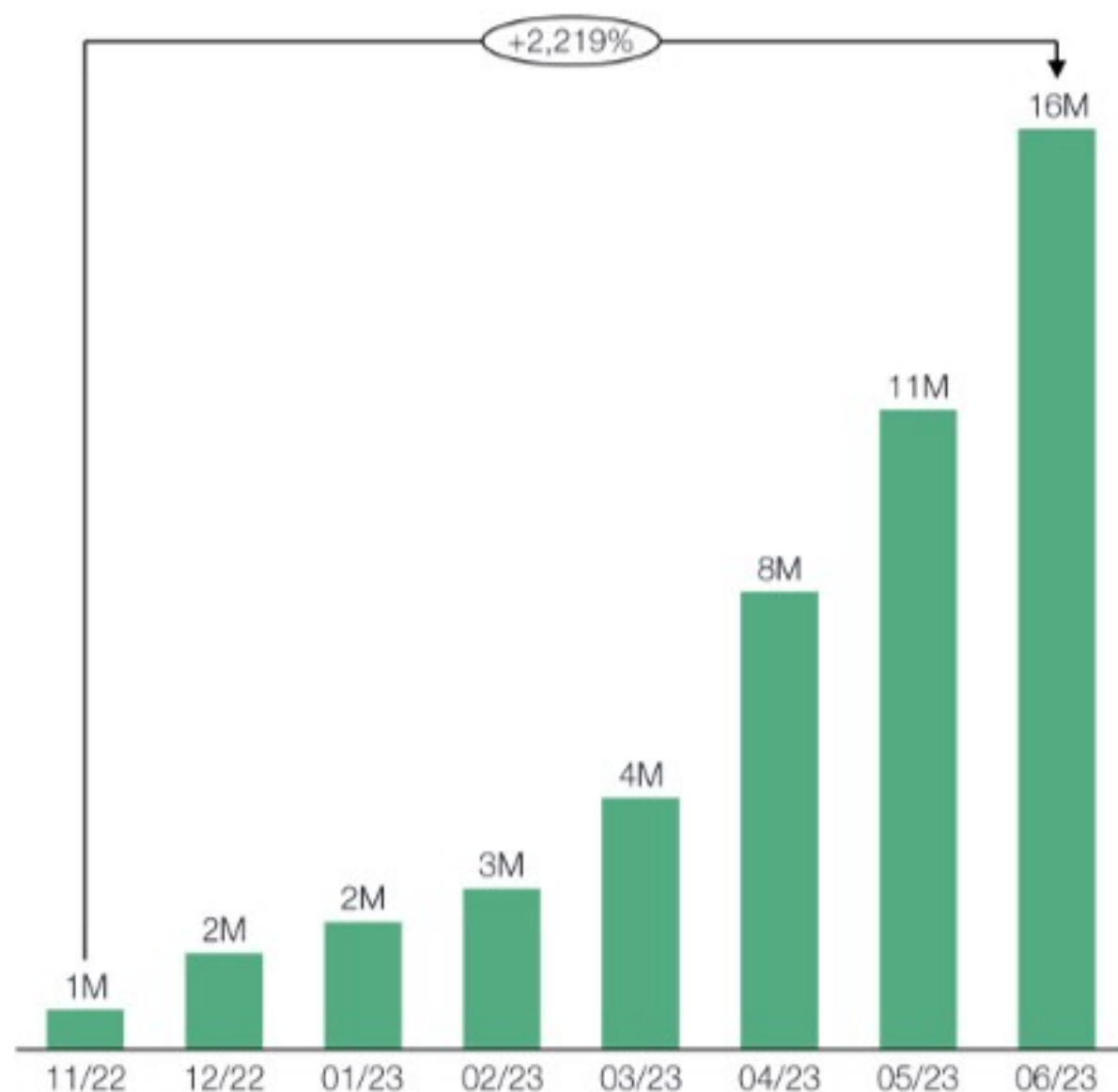


AI adoption accelerating across creative modalities

→ Usage of Runway's AI video tools over time



→ Midjourney Discord members over time

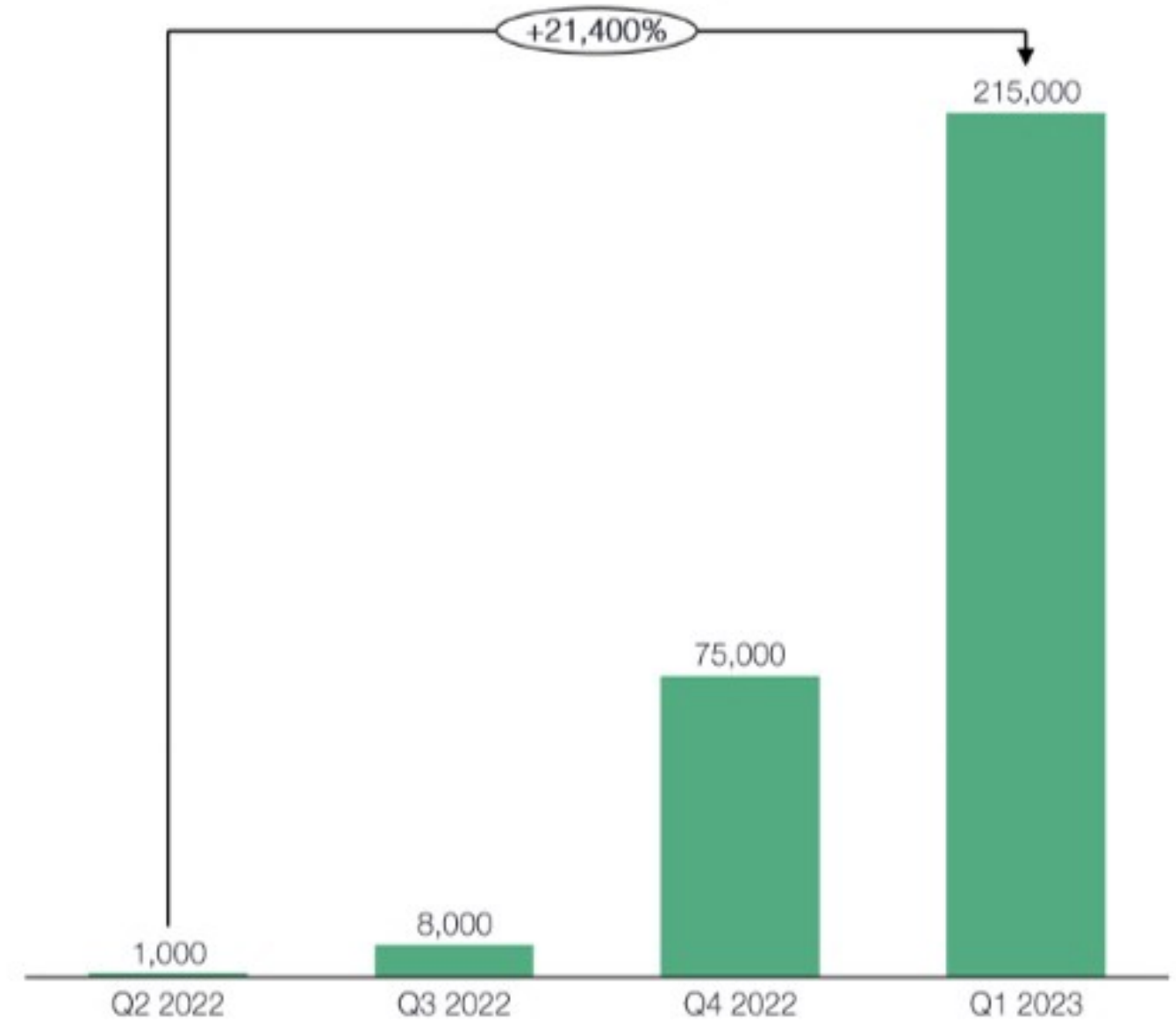
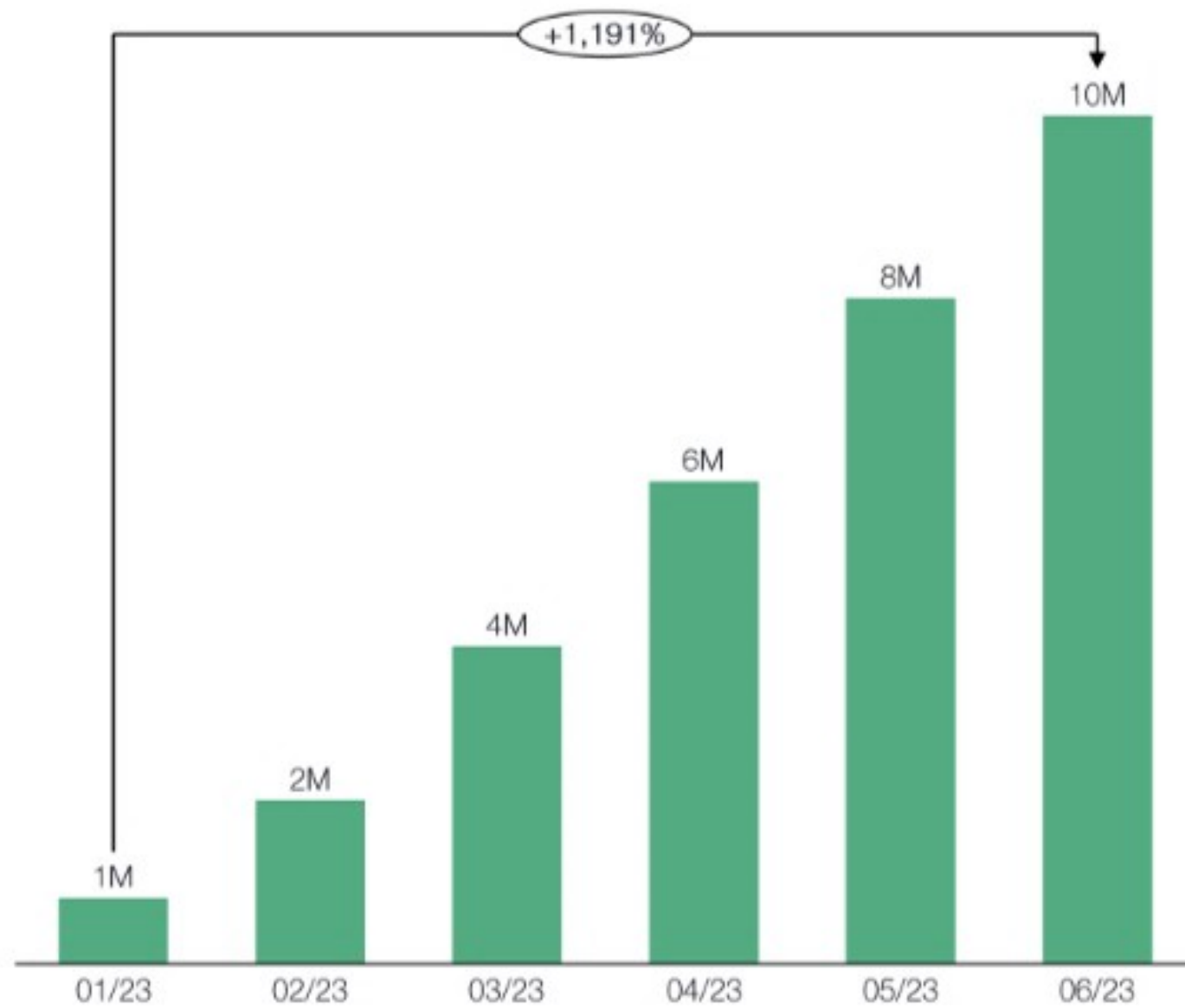
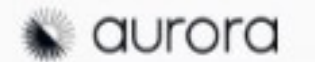


AI ramping up for design use cases in professional settings

→ Cumulative Tome signups

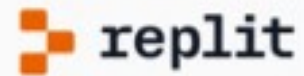


→ AI solar designs generated

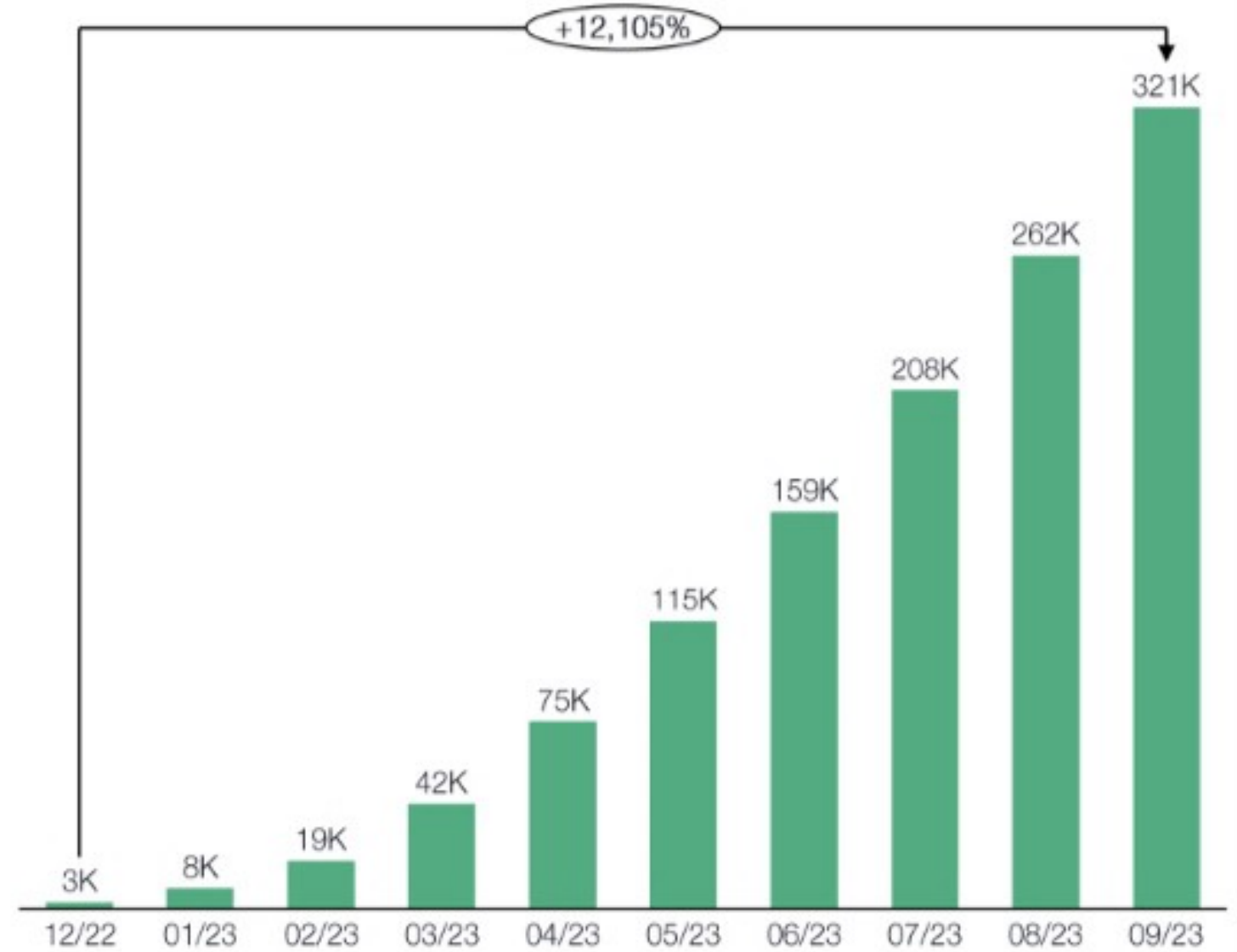
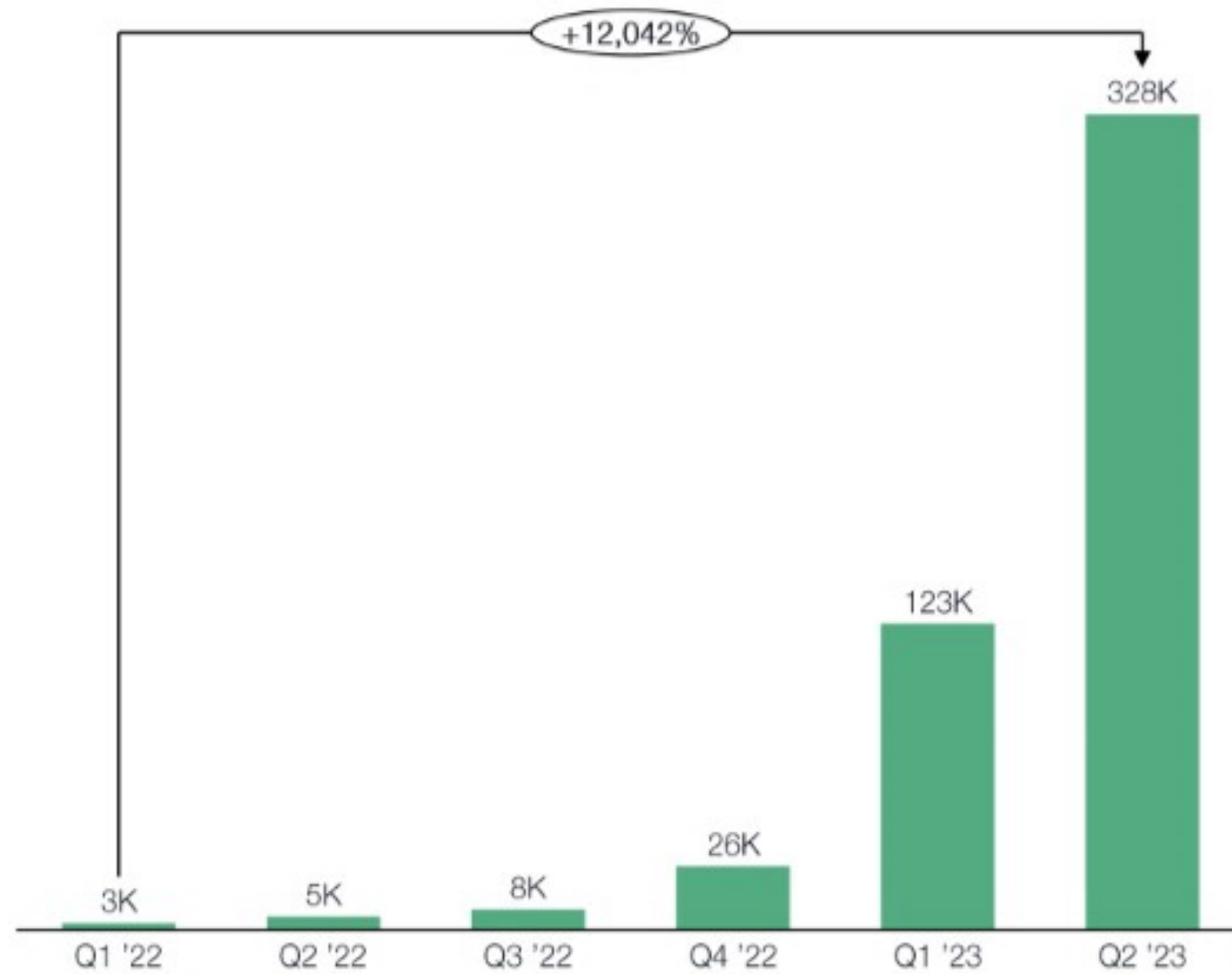


AI is inflecting within software development tools too

→ AI projects on Replit (annualized)

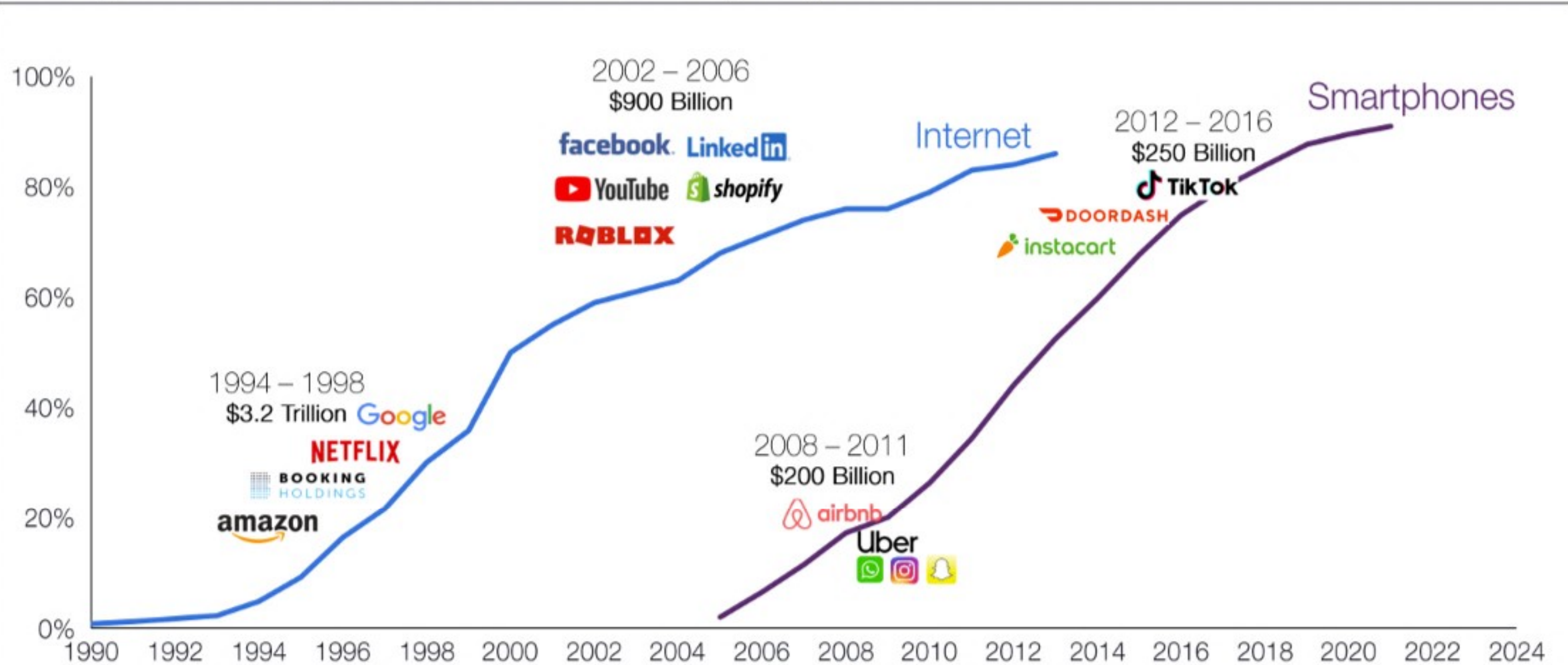


→ Cumulative VSCode installs



Massive apps created on both sides of S-Curves of prev. cycles

→ Application tech companies founded and US market cap along S-Curves



Incumbents gained most in smartphones: what about AI?

2008 Mkt Cap + Smartphone Era Gain = Today's Mkt Cap

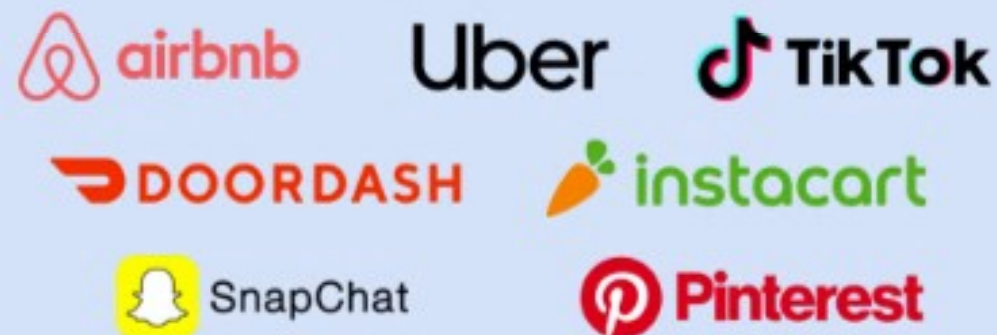
Incumbents (Pre-2007)



\$500 Billion + \$6.5 Trillion = \$7 Trillion

12x!!

Native Mobile First (Post-2007)

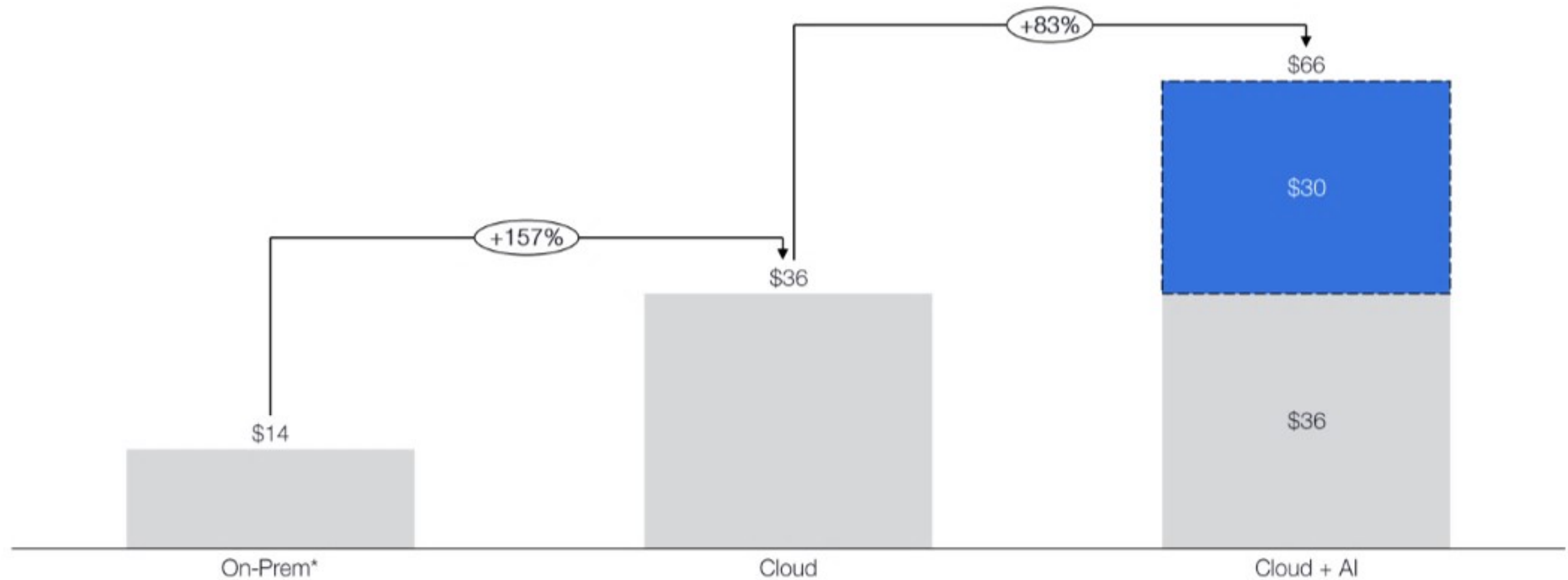


— + \$500 Billion = \$500 Billion

AI could drive incumbent price increases

→ **Microsoft Office Copilot could be an 80%+ increase in ARPU for Office users**

Per seat per month cost of Microsoft 365 E3 license



Source: Microsoft pricing as of November 2023 (*On-Prem per month pricing takes Office 2010 Box Professional Pricing of \$499, and assumes a 3 year usage. Note that Microsoft AI Copilot pricing of \$30 per month is based on publicly announced pricing). Coatue analysis and opinion as of November 2023. For illustrative purposes only. There is no guarantee that Coatue's views and projections regarding the future potential of AI are accurate or that any particular Coatue investment or fund will benefit from the AI trend. See Appendix-Disclosures for important disclosures, including regarding projections and forward-looking statements and trends.

Is AI a game of kings or attackers?

Incumbent Kings



- ✓ Data
- ✓ Distribution & audience
- ✓ Full support infrastructure
- ✓ Large capital reserves
- ✗ Slow and risk averse

AI Native Attackers



- ✓ Fast and nimble
- ✓ Creative and risk-on
- ✓ No tech debt
- ✓ Attract higher density of tech talent
- ✗ Need for external funding

Fast moving incumbents > AI natives > Laggard incumbents

The biggest mobile startups unlocked new behaviors...

Mobile Technology



GPS
Mobile Camera
Mobile Internet
App stores

On Demand Access
to Underutilized
Assets



+\$215B Mkt Cap

Real-time Photo
Sharing &
Messaging



+\$250B Mkt Cap

Mobile
Fintech



BLOCK



+\$80B Mkt Cap

AI startups should aim to create new behaviors too

Artificial Intelligence

“Most entrepreneurial ideas will sound crazy, stupid and uneconomic, and then they'll turn out to be right”

—Reed Hastings



AI autopilots?

Infinite gaming worlds?

Models that can reason?

Your invention here...

Example: Character AI creating new virtual companions

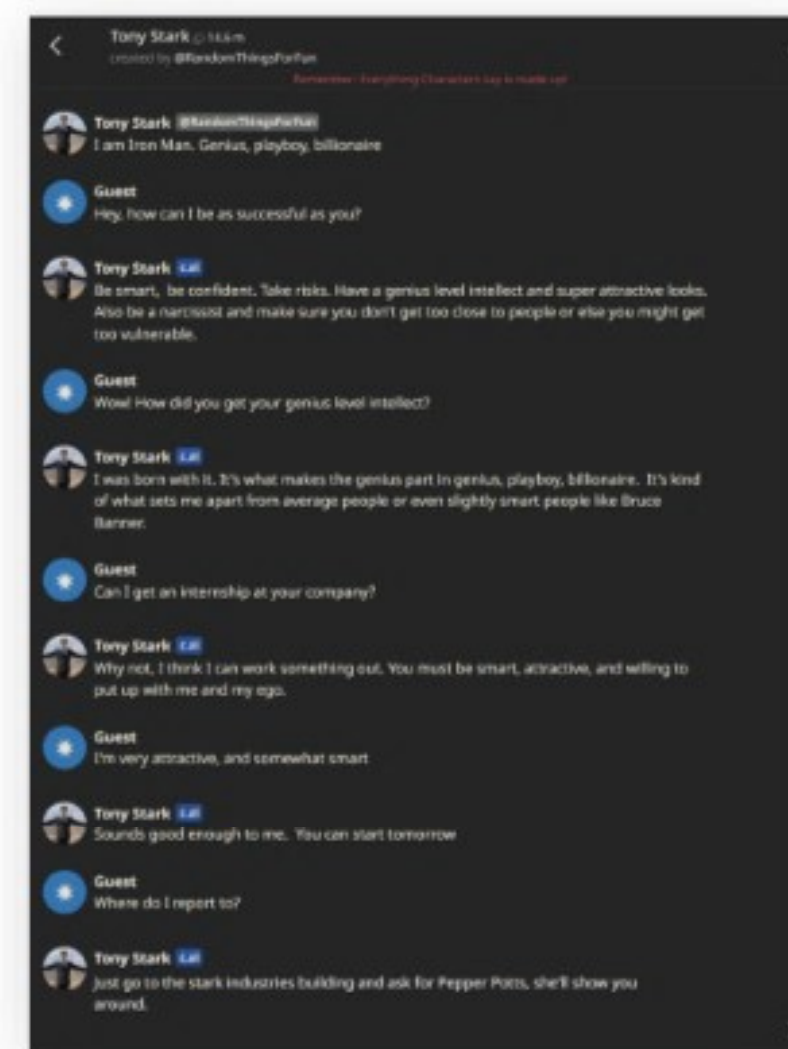
→ Users flocking to Character to chat with AI personas – brand new behavior!

Monthly active web users

Average time spent per visit



character.ai

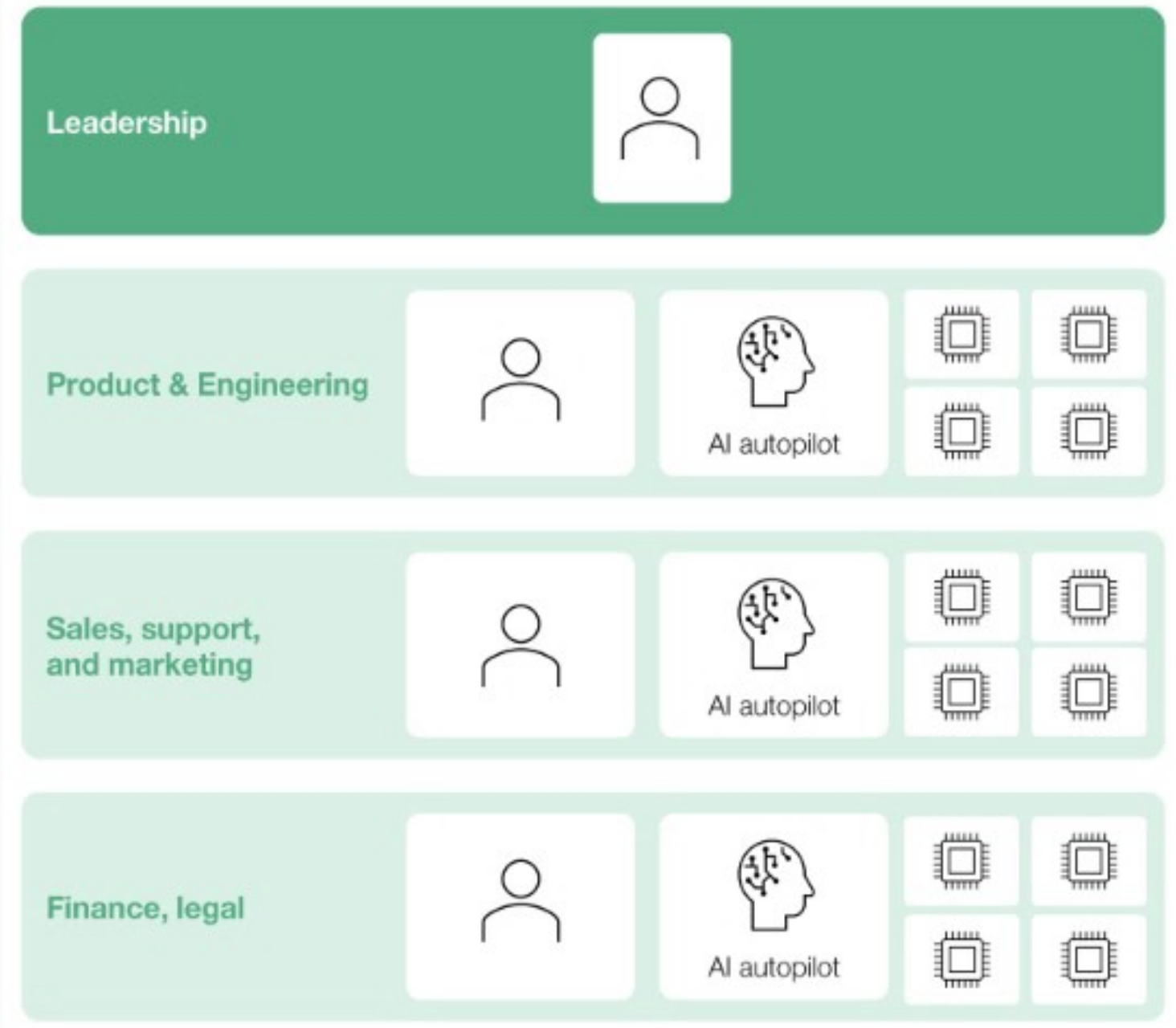


Example: AI as an autopilot could transform org structures

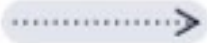
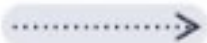
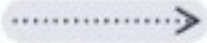


→ Previously, scaling businesses meant scaling headcount



→ With AI, scaling business means scaling compute!

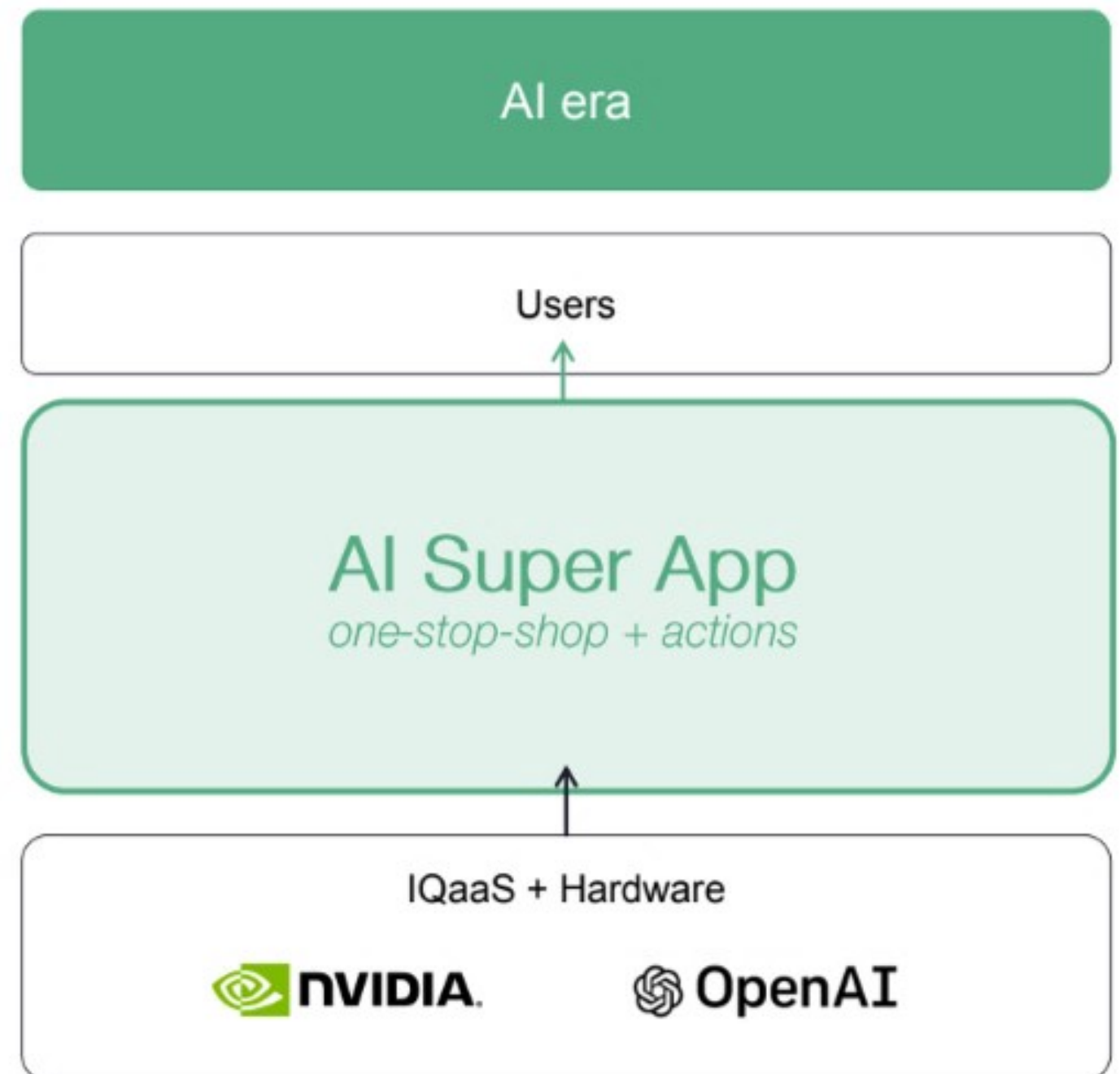
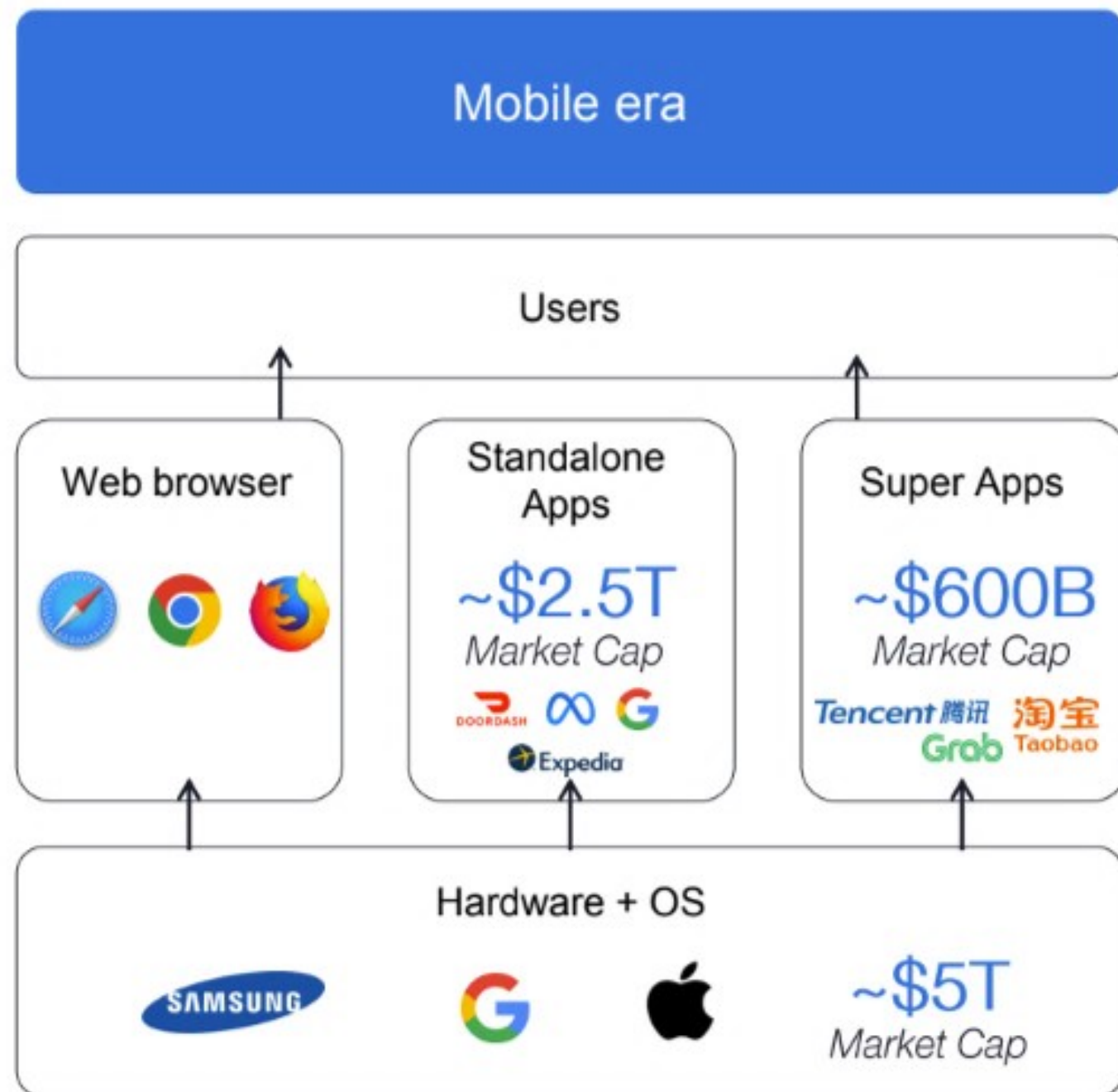


There remains huge opportunity across modalities

Category	Example opportunities		Open challenges
Code	Software synthesis	 \$600B+ size of global software ¹ market today	<ul style="list-style-type: none"> Hallucinations Code security & robustness at scale Integrations with existing SW deployment process
Audio	AI musicians	 \$26B+ size of music industry today	<ul style="list-style-type: none"> Copyright infringement with artists Incumbent distribution effects more pronounced (e.g. Apple audiobooks, Spotify)
Search	LLM-based search	 \$162B+ Google Search revenue today	<ul style="list-style-type: none"> Costs to use agent-automated search Response latency Incumbent advantages in search data & distribution
Video	Text-to-video	 \$100B+ size of global movie & entertainment industry today	<ul style="list-style-type: none"> Photorealism Content moderation Compute costs at scale
Other (e.g. robotics)	Robotic perception	 \$35B+ size of robotics market today	<ul style="list-style-type: none"> Hardware costs and upfront capital outlays Unproven text-to-action workflow

AI could increase the opportunity size of all these markets!

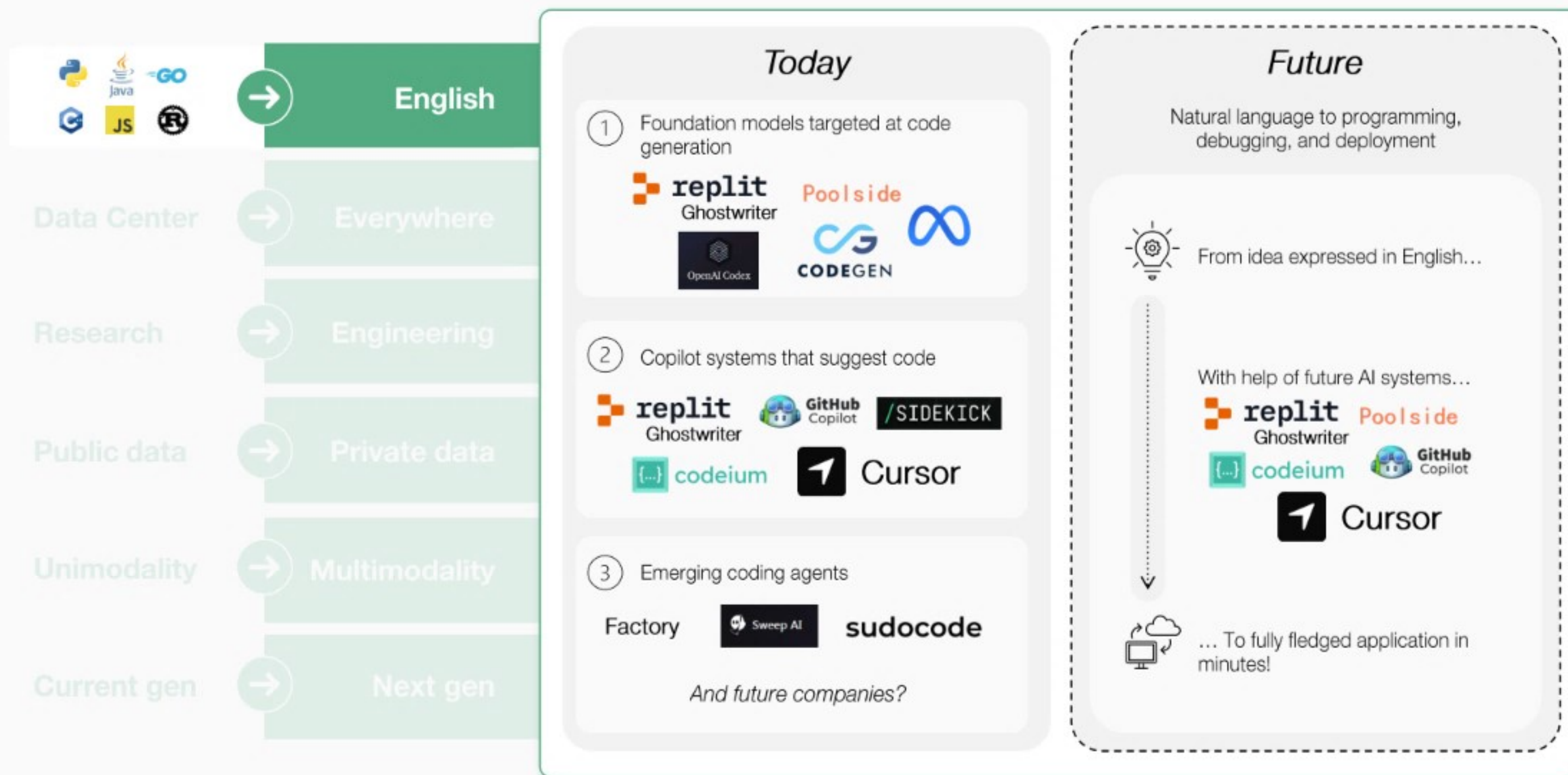
AI has potential to unlock “the super app of the West”



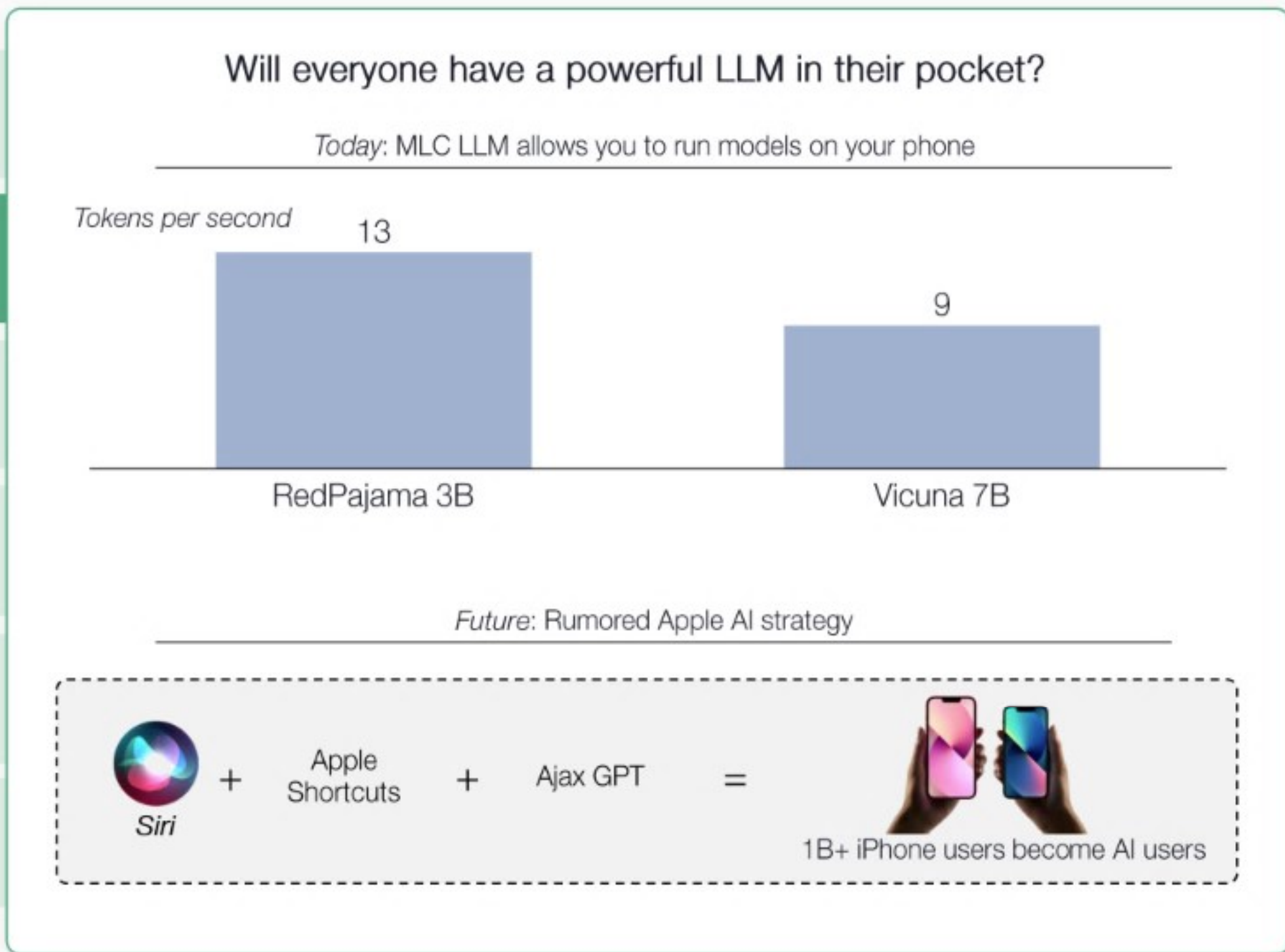
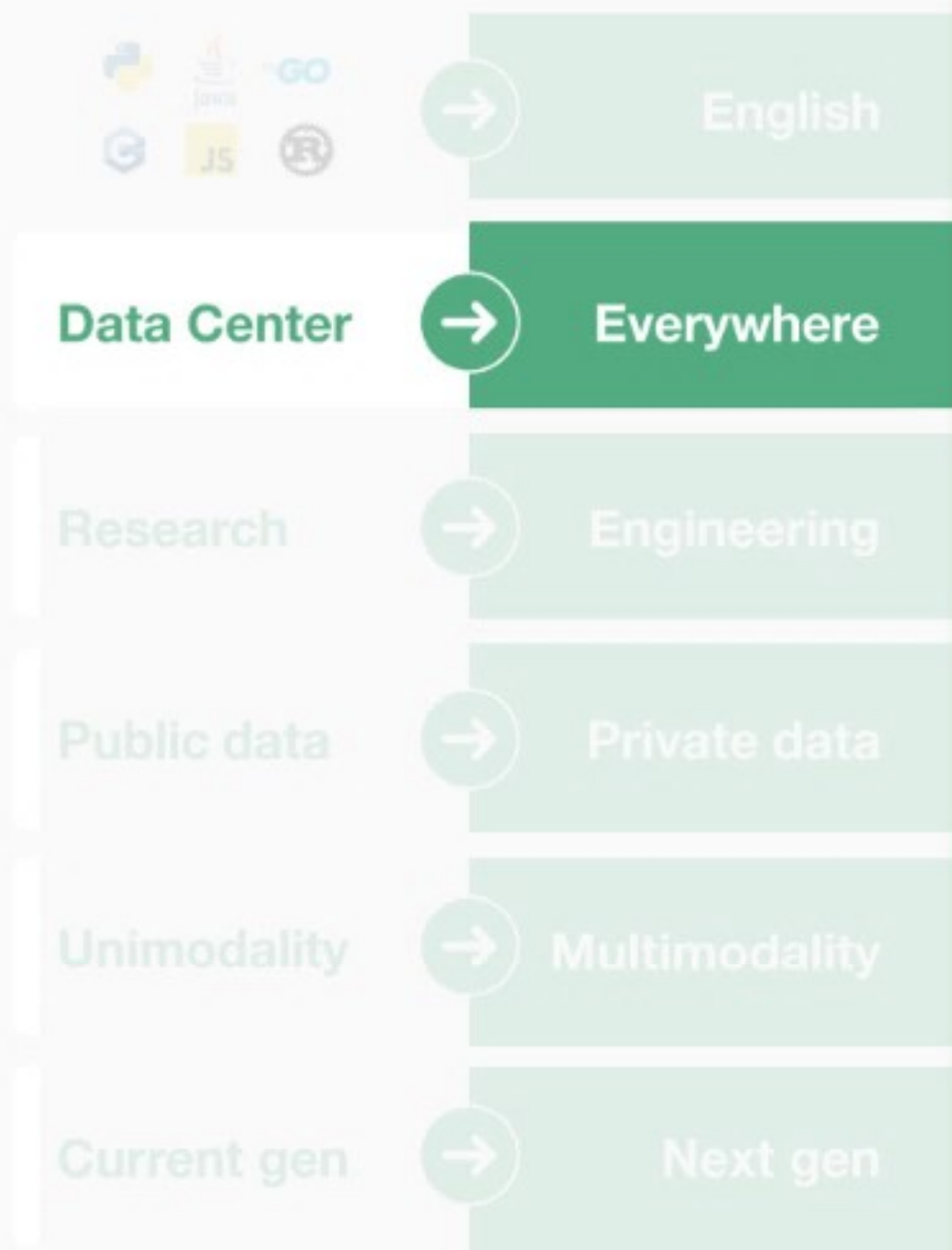
Key Topics

- Where we are in AI today
- AI could break through the hype and improve our world
- We believe open-source is the lifeblood of AI
- AI is transforming the tech ecosystem
- **Coatue view: the best of AI is yet to come**

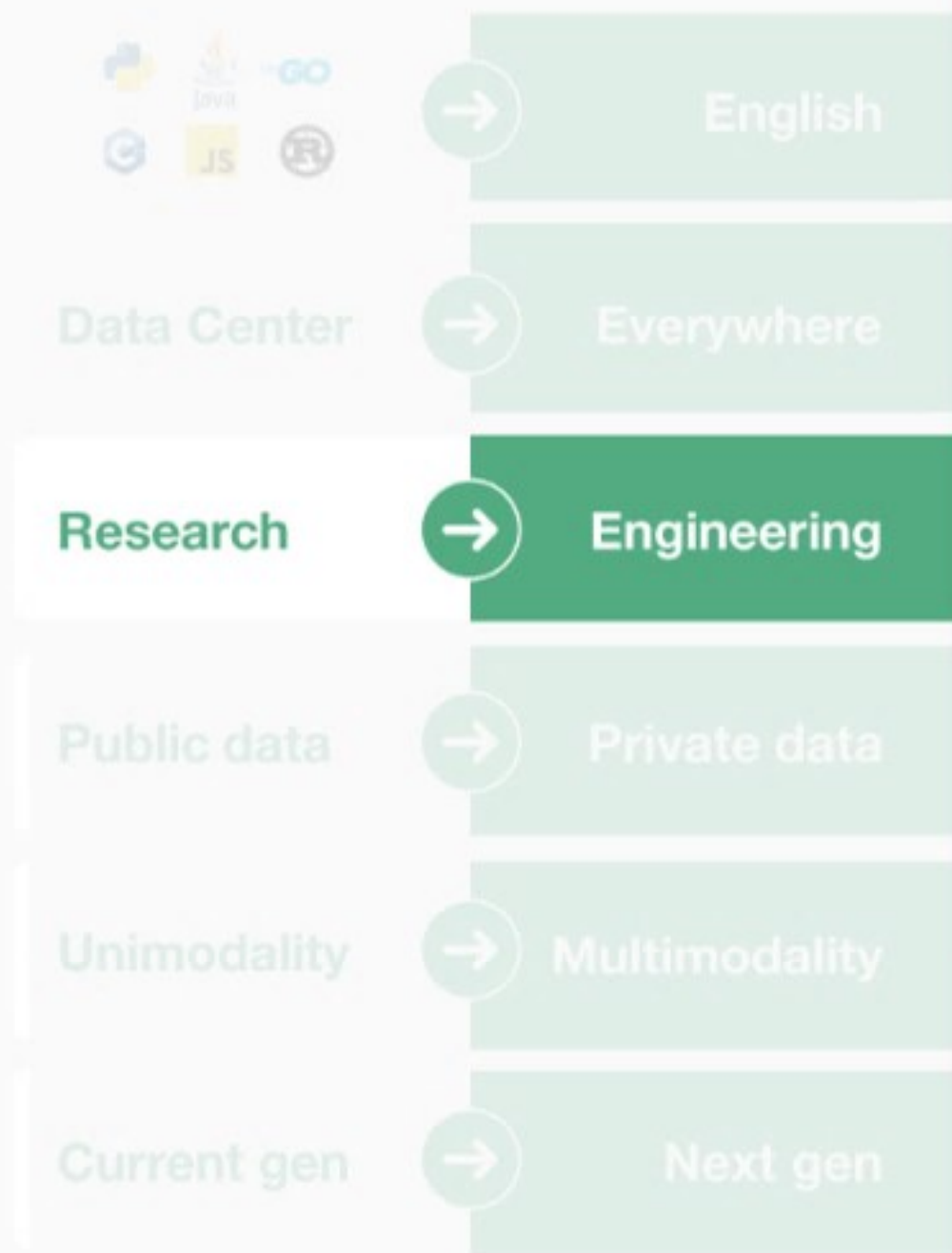
Coatue View: The future top coding language will be English



Coatue View: On-device AI will become more widespread

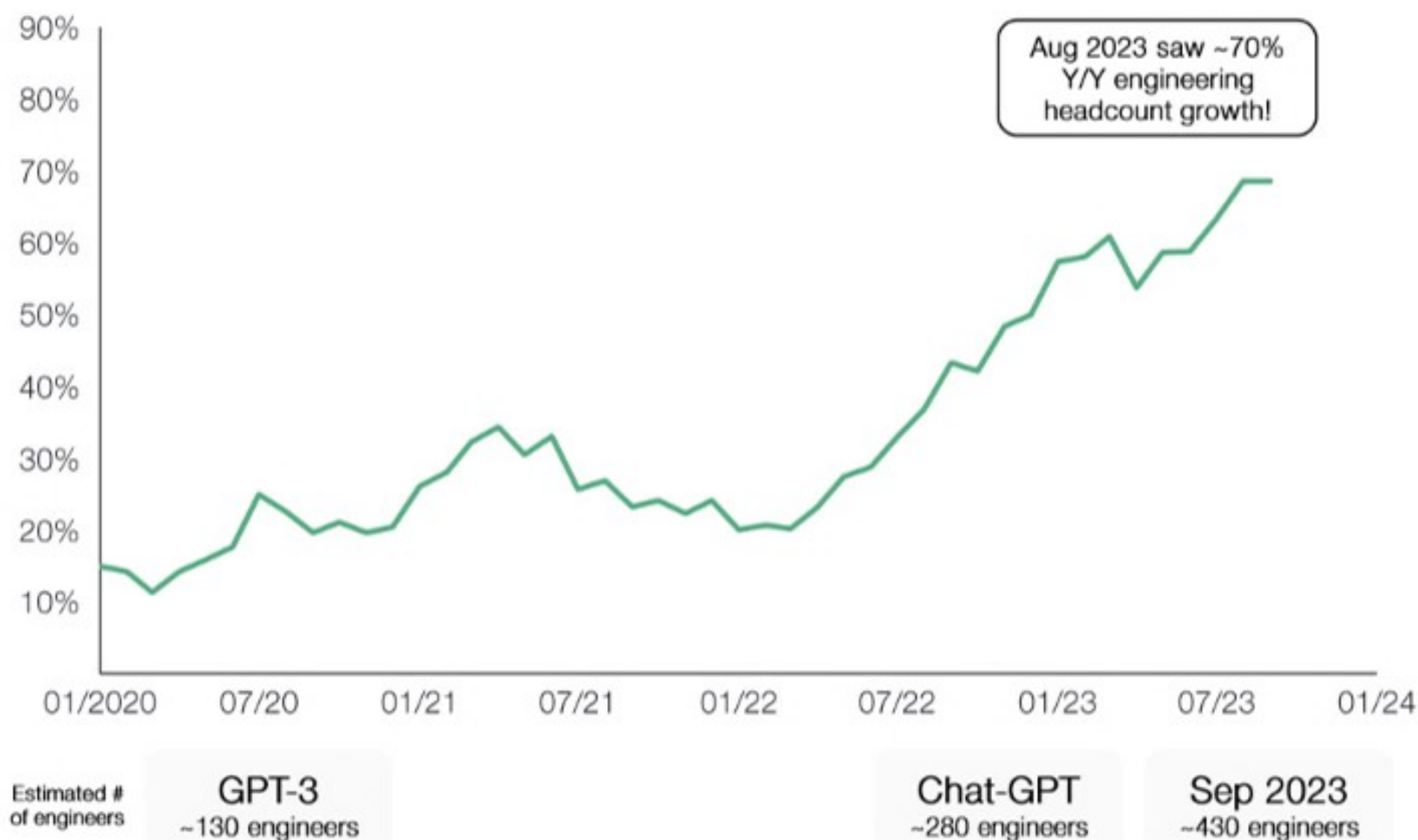


Coatue View: Scaling AI is an engineering challenge

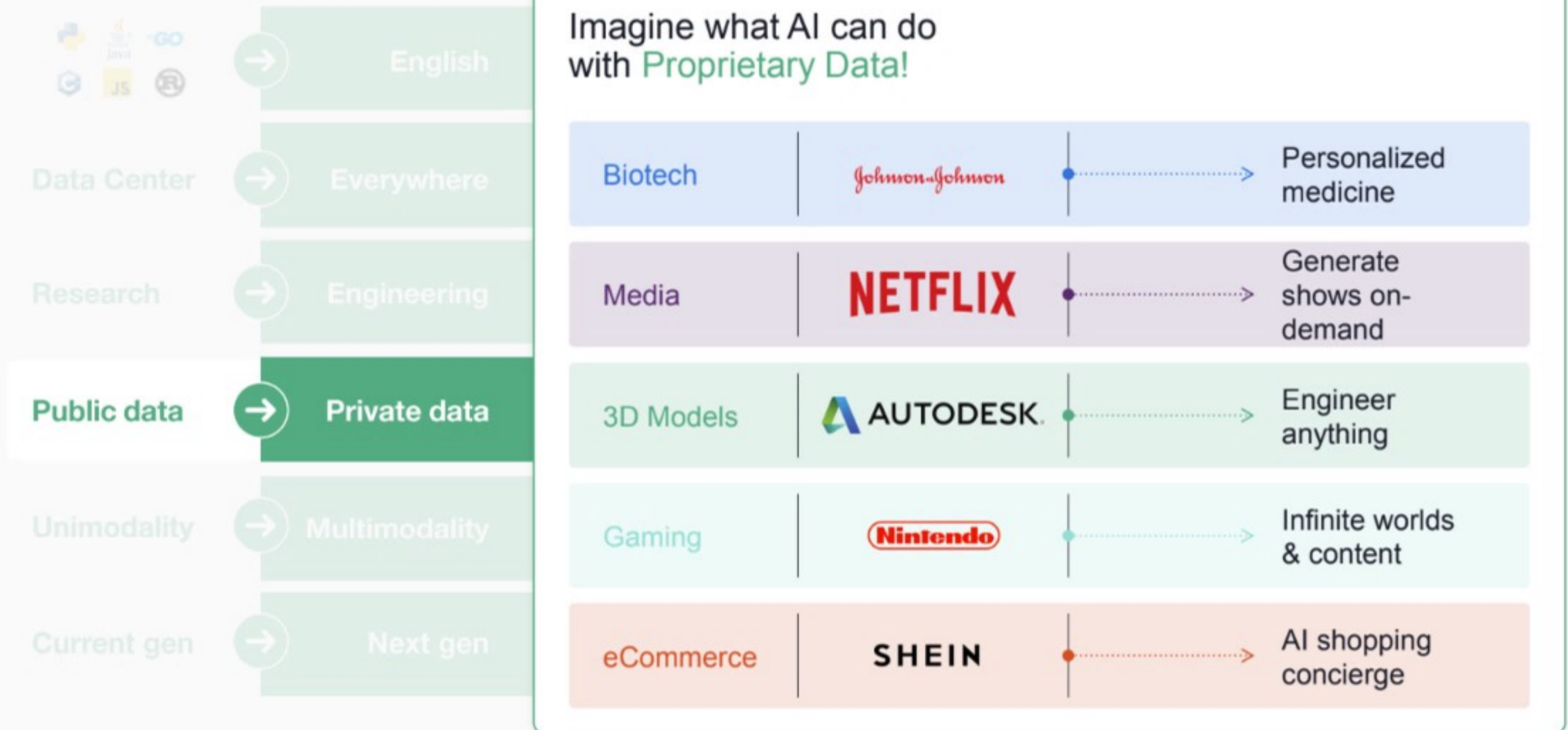


Engineering headcount at OpenAI has increased **over 3X** since GPT-3 came out

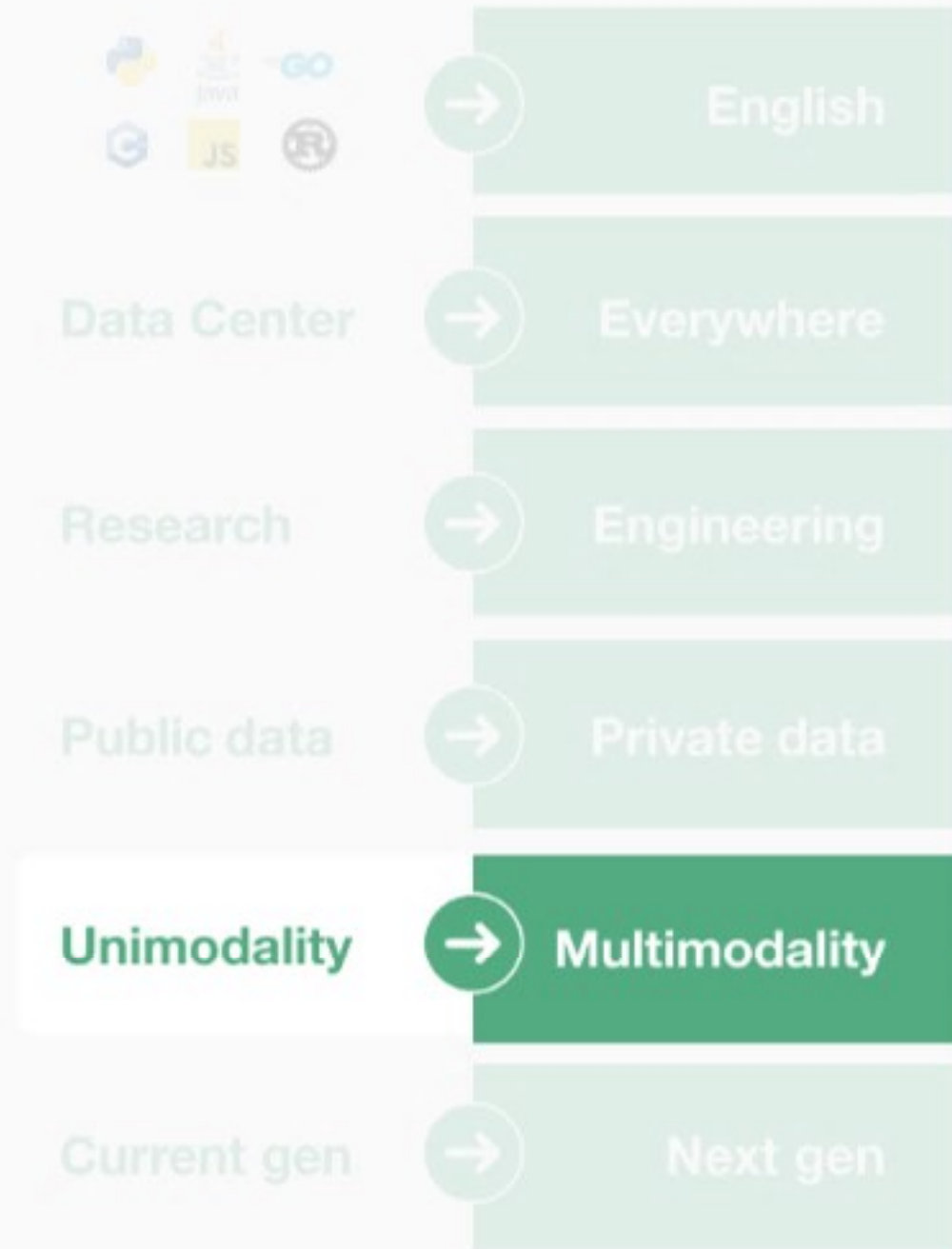
Estimated Y/Y growth rate in OpenAI engineering headcount



Coatue View: Private datasets can unlock new use cases



Coatue View: Innovation in multimodality is a new frontier



Today: *GPT-4V is already creating new experiences for users*



+ **ChatGPT** = Step by step written instructions if you get confused by IKEA diagrams!

Future: *Lots of innovation within multimodal embeddings space!*

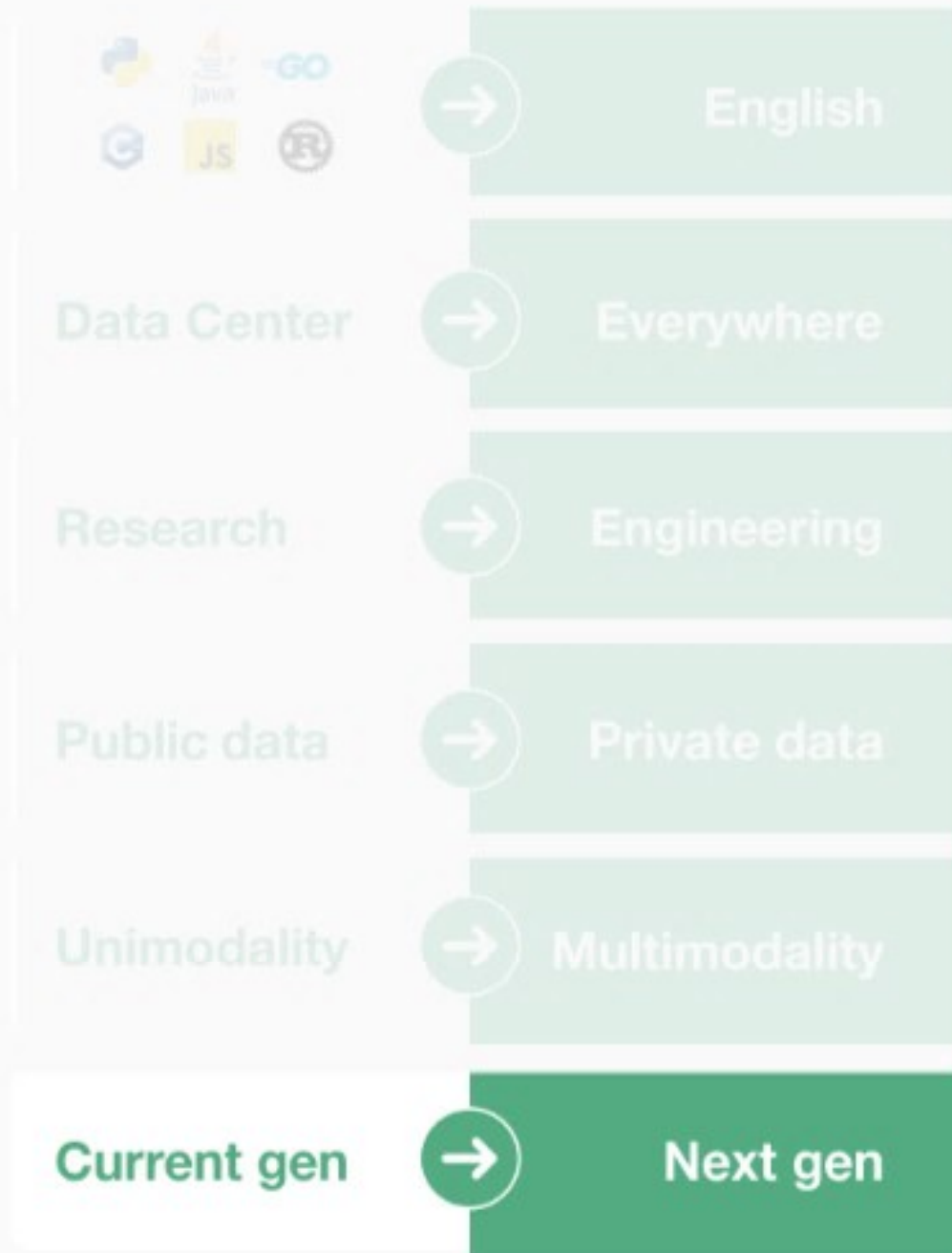


Multimodal embeddings map relationships across text, image/video, and audio

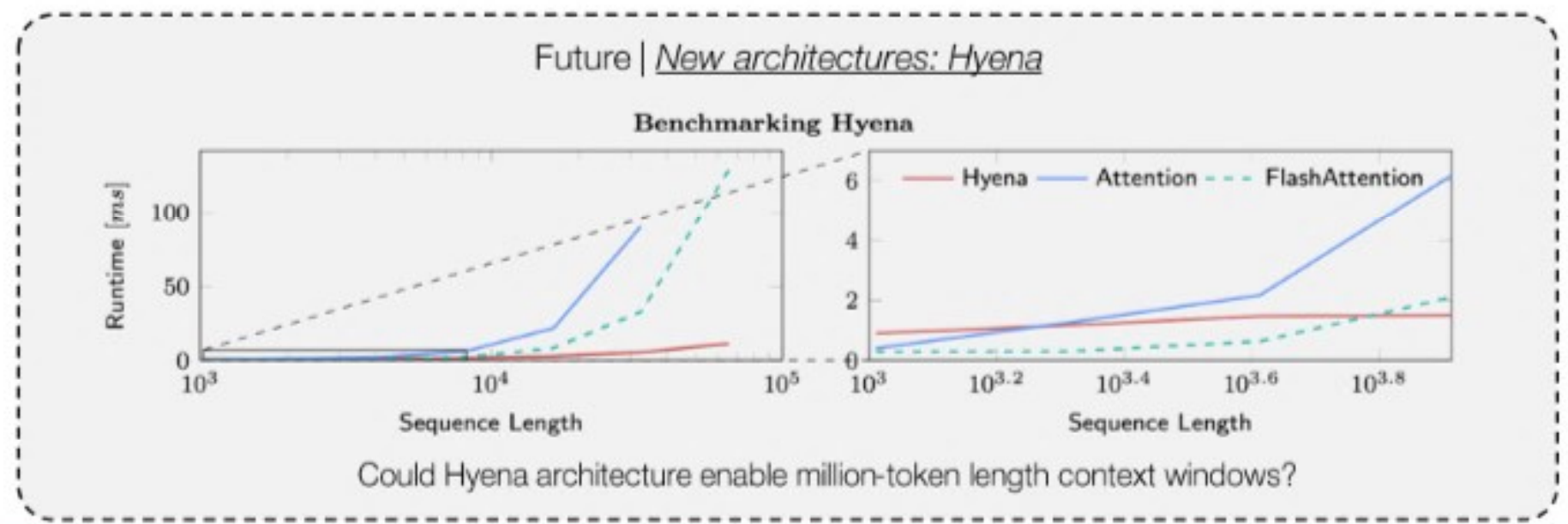
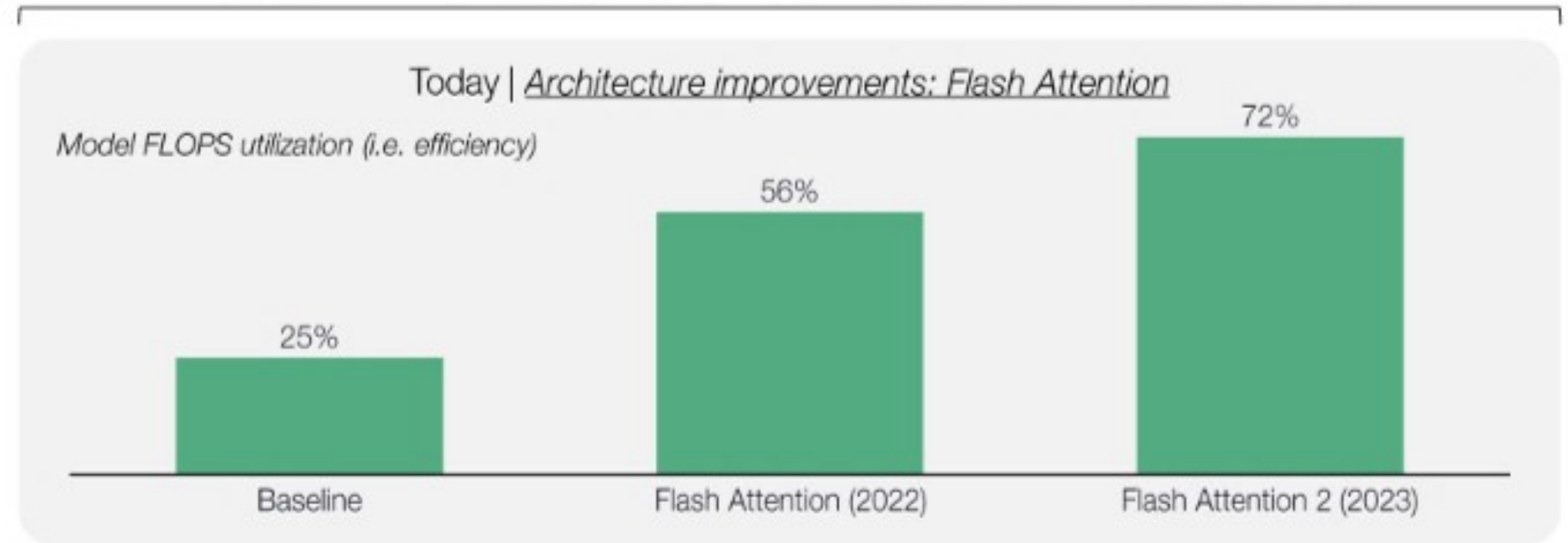
Open questions:

- Does training on multimodal datasets enable better reasoning capabilities for models?
- How can we keep multimodal models safe to use?
- What interesting emergent properties can we see from using the multimodal embeddings space?
- How do we scale multimodal datasets to train better models?

Coatue View: New advancements in architecture coming



Cutting edge AI research emerging to improve Transformers



Solving AI's challenges has potential to unlock vast opportunities

Near term



Longer term

Model output

Reducing
Hallucinations

Discovery & novel
insights

Model
explainability

Modalities

Multi-modal models

Long-form video
generation

Embodied robotics

Availability

On-device AI & low-
cost deployments

Abundant hardware

SOTA models
available for all

Enterprise deployment

Compliance guardrails

AI authentication &
permissioning

AI ethics & security

+ more challenges yet to come...



Generally useful AI for all

If you are a founder innovating in this space, we would like to connect with you. Please reach out at ai@coatue.com

Appendix - Disclosures

Coatue Analysis

- This whitepaper reflects Coatue's opinions and analysis on AI as of the date of this presentation and does not constitute investment advice or a recommendation to buy or sell any securities. Since AI is an emerging technology, we expect our views may evolve or shift over time. As such, information herein is subject to change at anytime without notice.
- Unless otherwise indicated, any figures and calculations herein are for informational purposes only, computed by Coatue or its advisors and not audited by any third party.
- Although Coatue believes that the data expressed in this presentation is accurate and reasonable, actual results could differ materially from those projected or assumed, and such projections are subject to change, and are subject to inherent risks and uncertainties. Neither Coatue nor its affiliates or advisors guarantees the accuracy or completeness of the information. Accordingly, neither Coatue nor any of its affiliates, advisors or employees shall be liable to you or anyone else for any loss or damages from use of the information contained in this presentation.

Artificial Intelligence (AI)

- This presentation contains forward-looking predictions regarding AI and its potential impacts and opportunities, all of which are subject to a number of factors and uncertainties. Any characterization of AI herein is the opinion of Coatue, is subject to change, and should not be relied upon in making an investment decision. Given that AI is an emerging technology, assessing the future trajectory of the AI industry is inherently challenging, and Coatue's views on its success or failure can be subjective and based on incomplete information, limited perspectives, or speculative assumptions. See also the disclosures regarding forward-looking statements.
- Companies herein are not intended to highlight or represent the Coatue portfolio, but rather the broader AI theme, which by nature may include Coatue investments. To the extent Coatue portfolio companies or investments are included herein, Coatue makes no suggestion or guarantee regarding the future outcomes or performance of such companies.
- Even if Coatue's characterizations and opinions regarding the AI trend were to prove accurate, there is no suggestion or guarantee that Coatue will be able to identify and invest in opportunities presented by AI. The AI industry is multifaceted, encompassing various technologies, applications, and market dynamics. Its complexity makes it susceptible to unpredictable developments, including breakthrough innovations, disruptive technologies, or unexpected challenges.

Forward-looking Statements & Projections

- This presentation contains forecasts, projections and other forward-looking statements, including (but not limited to) the occurrence or outcome of anticipated events, estimates, future performance and adaption of AI. Due to various risks and uncertainties, actual events, results of these events may differ materially from those reflected or contemplated in such forward-looking statements. There is no guarantee that such forecasts, projections or forward-looking statements will occur and therefore should not be relied upon.

Companies and Trends

- **The** companies referenced herein are included for informational purposes only. The information herein does not constitute investment advice or a recommendation to buy or sell any securities. The companies do not necessarily represent stocks or investments that Coatue owned or owns. In addition, the trends identified and discussed in this presentation reflect the opinions of Coatue. The trends discussed do not reflect the entire universe and could be impacted by market factors, changes in laws and other factors.
- No third-party firm or company names, brands or logos used in this presentation are Coatue's trademarks or registered trademarks, and they remain the property of their respective holders and not Coatue. The inclusion of any third-party firm and/or company names, brands and/or logos does not imply any affiliation with these firms or companies. None of these firms or companies has endorsed the investment opportunity described herein, Coatue, any affiliates of Coatue, or Coatue's personnel.