2024 generative AI predictions

CBINSIGHTS

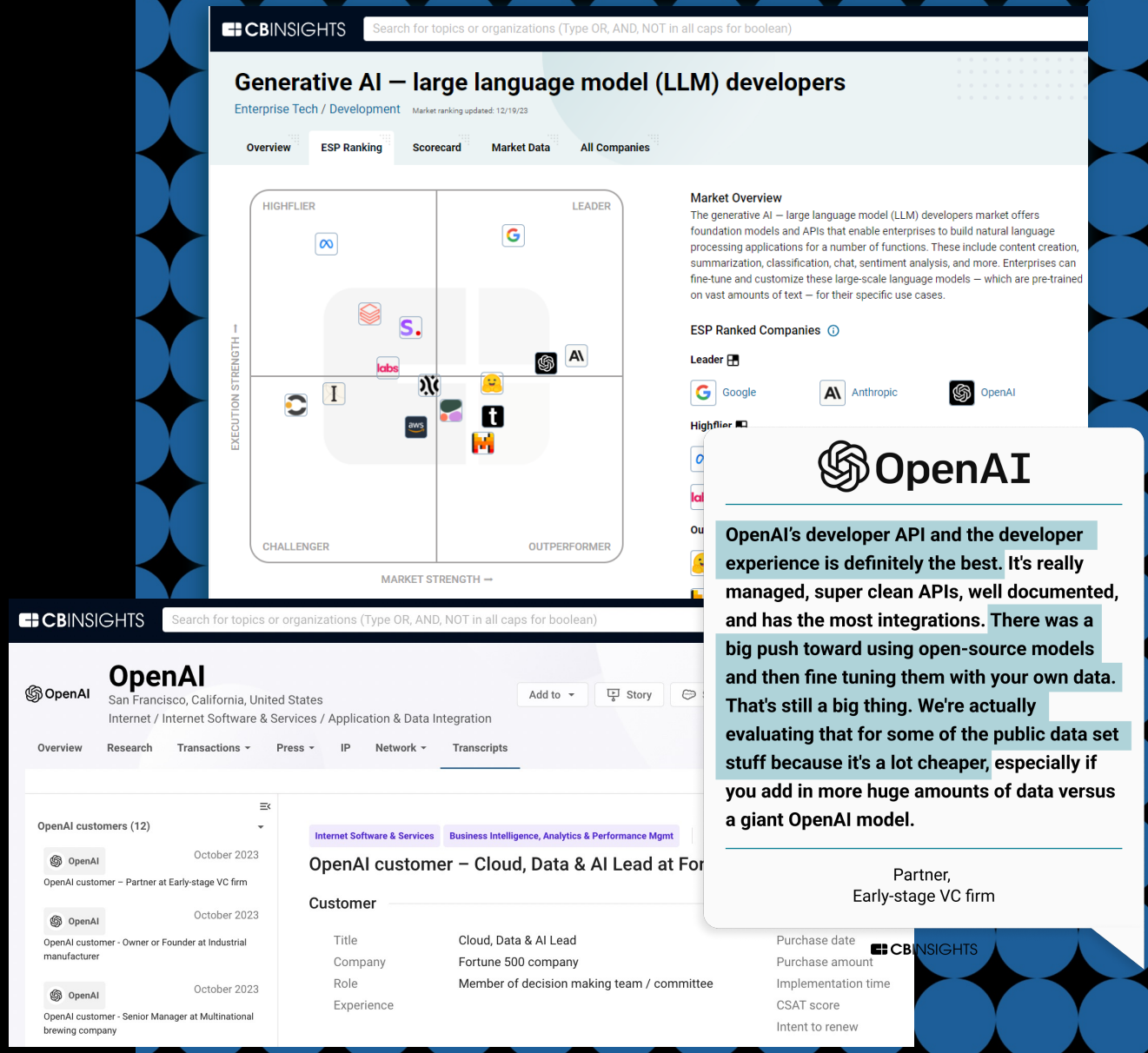


Unleash the generative AI revolution

Dive into a treasure trove of generative AI companies, investors, and market insights. Elevate your decisions with exclusive software buyer interviews to understand product pricing, CSAT, and post-deployment experience.

CB Insights – Your Gateway to Generative AI

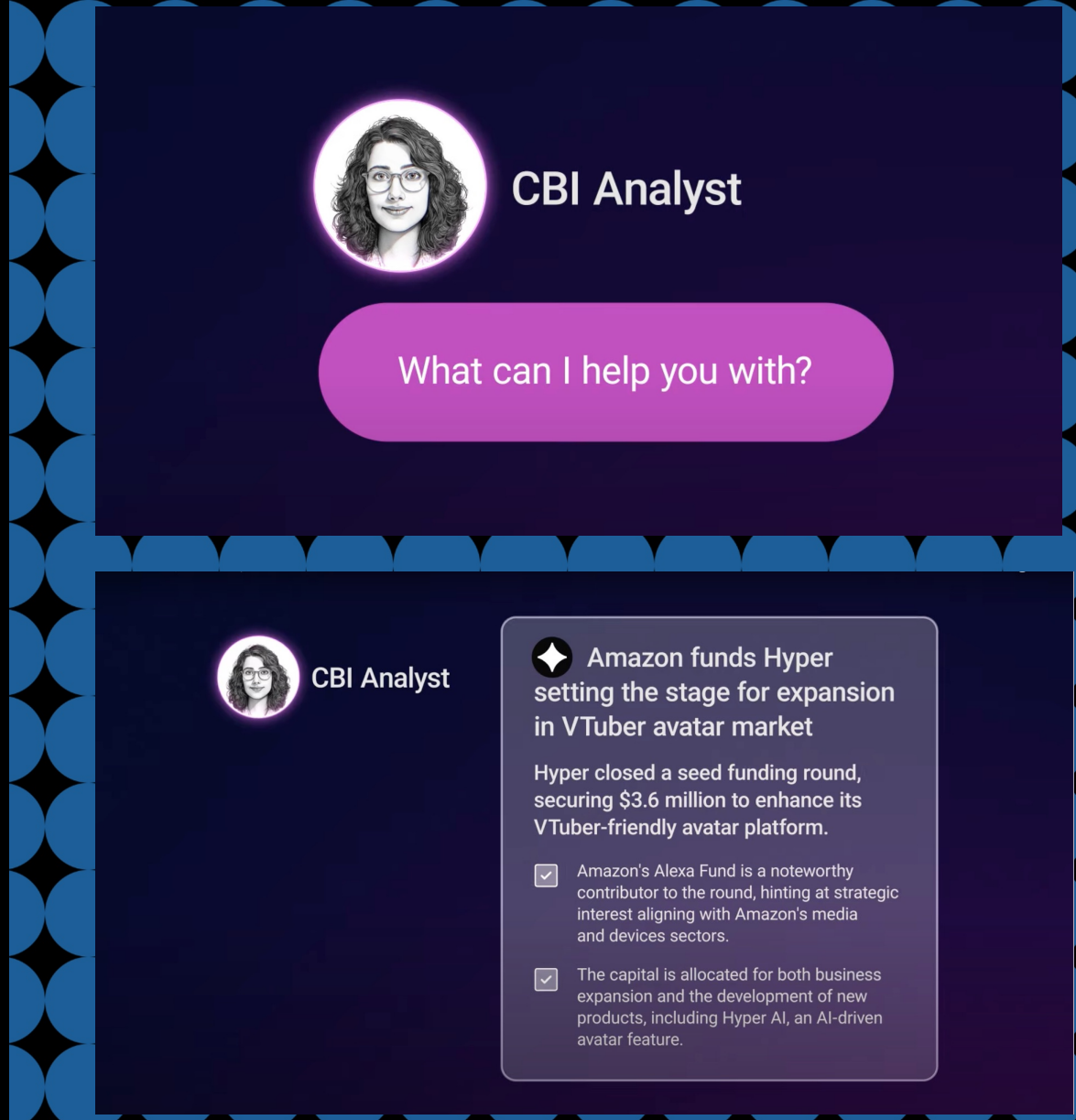
Sign up for access



Your always on AI-Driven Research Analyst

Do less searching, get more answers

Join the waitlist for early access



Predictions

1
Generative AI infrastructure

Power-hungry AI data centers proliferate 5
Data scarcity drives search for new sources 12
Nvidia's frenemies: Chip rivalries grow amid GPU shortage 17
Bad actors target language models 22
MLOps consolidates 27

2
Foundation models

Multimodal AI becomes table stakes 32
Enterprises do more with pocket-sized models 37
Open-source mounts a challenge to closed LLMs 43
Beyond transformers: New architectures challenge status quo 48

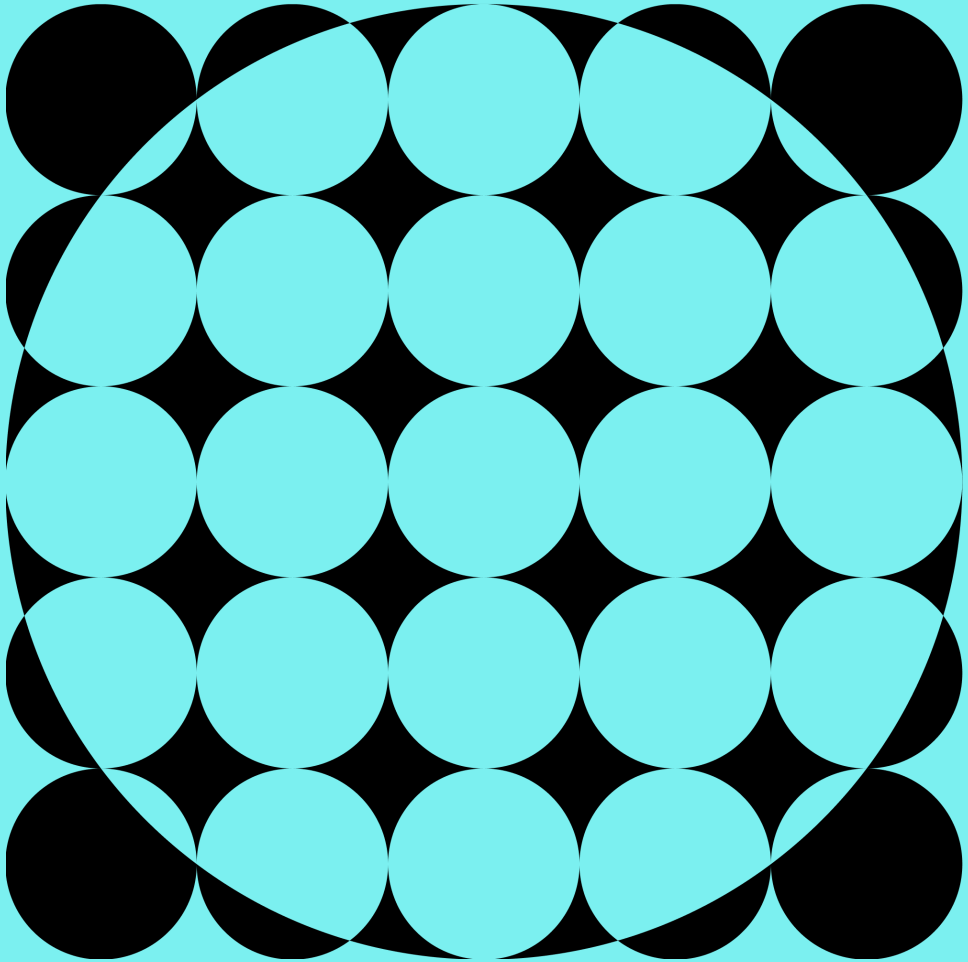
3
Human-machine interactions

AI girlfriends kill human romance 55
Lonely Gen Z befriends AI 61
The next iPhone: AI's hardware revolution 67
Autonomous agents: From hype to somewhat useful 74

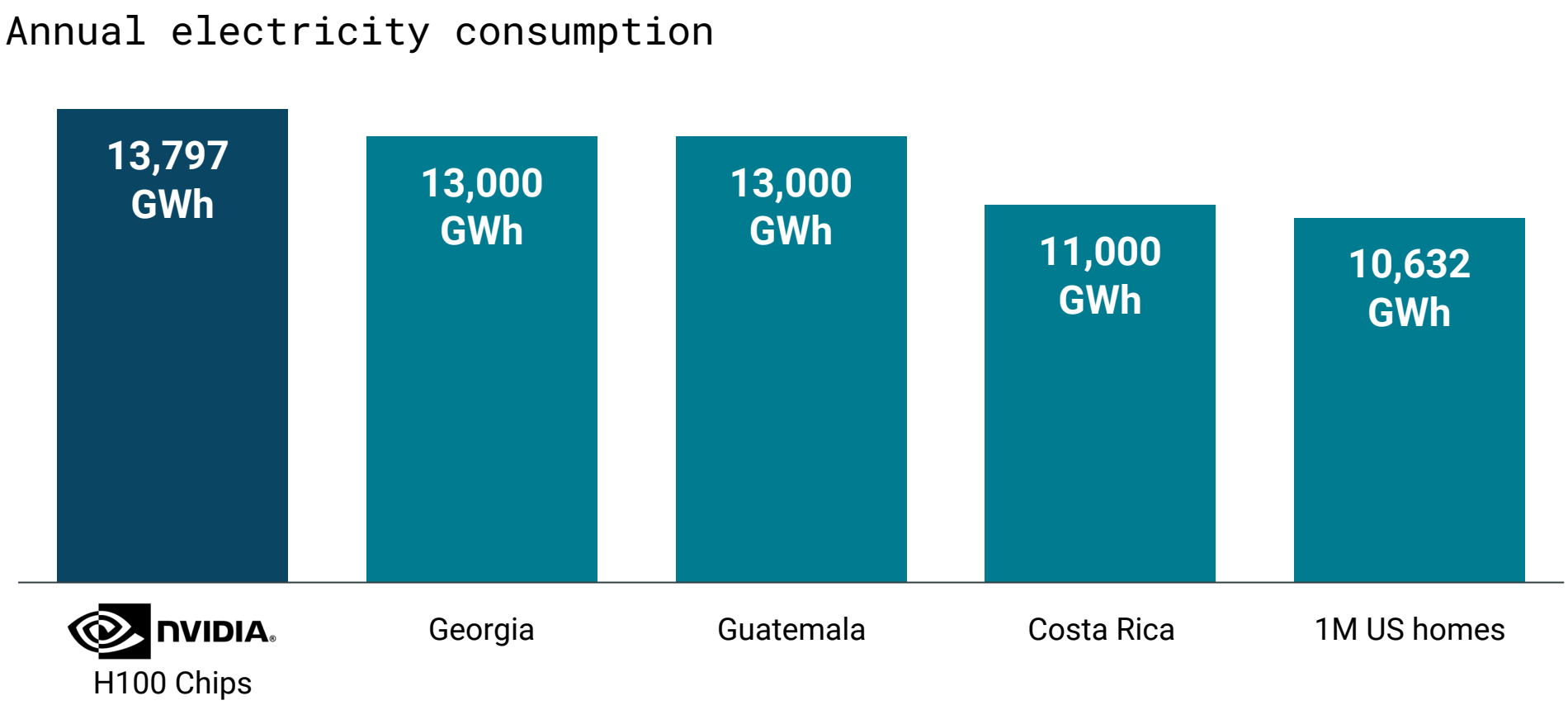
4
Generative AI applications

Copilots drive massive white-collar productivity gains 79
AI vs. AI: GenAI fuels new wave of attack and defense 87
GenAI takes over journalism 94
Beyond stills: Video generation makes strides 100
A genAI graveyard follows frenzy 104

Power-hungry AI data centers proliferate

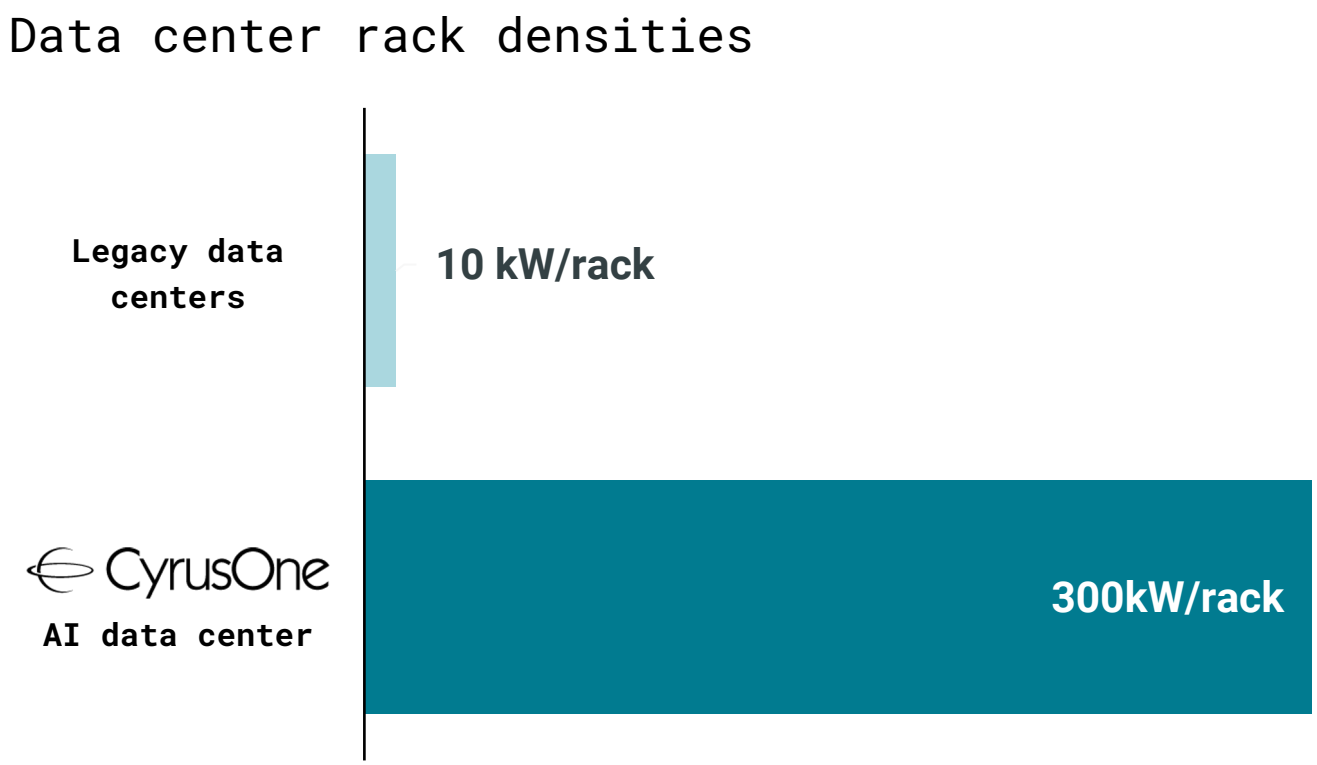


Nvidia's high-end chips will consume the same amount of energy as a small nation in 2024



Source: CB Insights estimates, assuming 3M H100s are operational in 2024 (based on 2024 sales estimate from Nvidia and Q3'23 sales estimate from Omdia Research) at 75% of max power; electricity consumption of countries and US homes based on 2021 EIA data

Legacy data centers are not ready for this



"It's estimated that 80% of the data center power will… be consumed by AI over the next 15 years. So, at the end of the day, access to power is a key differentiator."

DigitalBridge Group CEO Marc Ganzi, Q2'23 earnings call

Source: CyrusOne, CB Insights — Earnings transcript analytics (date reflects quarter call occurred)

There is massive demand for high-power AI data centers

Mentions of AI and data center tech skyrocket on earnings calls



Source: CB Insights — Earnings transcripts analytics (date reflects quarter call occurred)

Big money is going into enabling tech like liquid cooling

High-power and high-heat AI data centers will drive up demand for cooling technologies

KKR

Acquired **CoolIT Systems** for $270M in Q2’23



Backed **JetCool** through venture arm in Q4’23



Acquired **CoolTera** in Q4’23

Source: CB Insights company profiles — CoolIT Systems, JetCool, CoolTera

As utilities struggle to cope, demand for renewable energy is increasing

aes

“Already, data centers represent half of our US backlog and the growth of generative AI will only accelerate their demand for more renewable energy.”

CEO Andres Gluski, Q4’23 earnings call

HANNON ARMSTRONG

“...the AI-driven build-out of data centers has created significant incremental demand for renewable electricity.”

CEO Jeff Lipson, Q4’23 earnings call

Source: CB Insights — Earnings transcripts analytics (date reflects quarter call occurred)

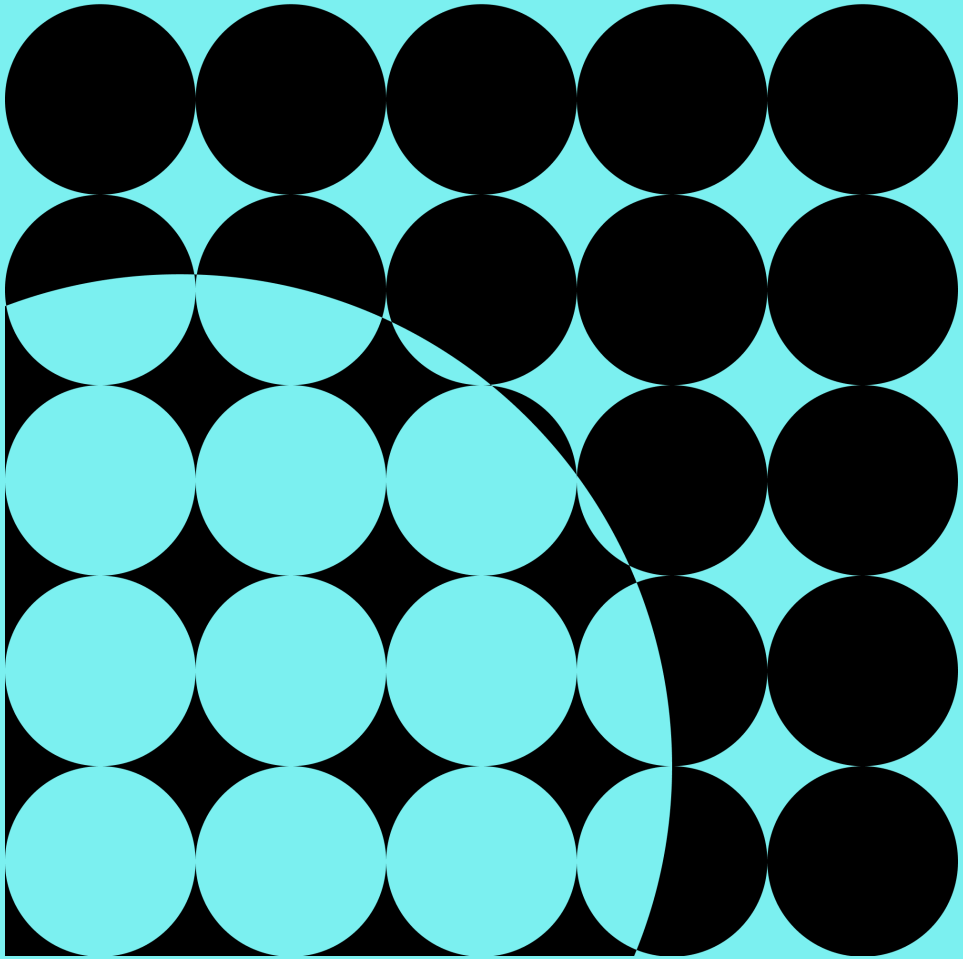
Nuclear may be the future of sustainable AI operations

Recently funded small modular reactors and fusion tech companies

Company	Valuation	Funding and partnerships
HELION	$3.5B	**Backed by Sam Altman** in a $500M round in 2021
energy®	$1.1B	Signed an MoU with the **Emirates Nuclear Energy Corporation** to expand decarbonization and nuclear energy efforts in sectors including data centers
OKLO	$800M	Announced **SPAC merger with Sam Altman's AltC Acquisition Corp.** in Q3'23 at a valuation of $850M
AVALANCHE	$123M	Raised $40M from **Toyota Ventures, Founders Fund**, and others in 2022
tae	$107M	Raised **$250M from Chevron, Google Ventures**, and other investors in 2022, followed by a California government grant

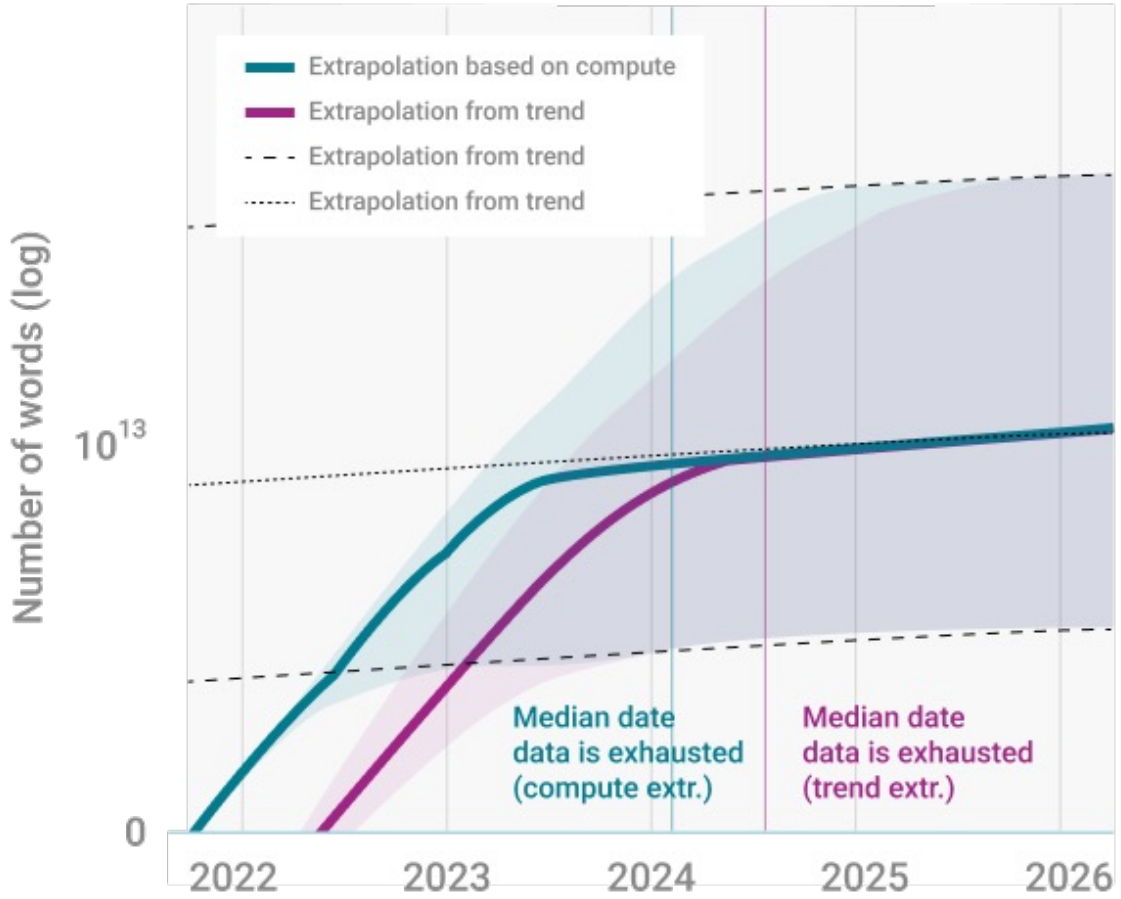
Source: CB Insights market reports — Small modular reactors (SMR), Nuclear fusion

Data scarcity drives search for new sources



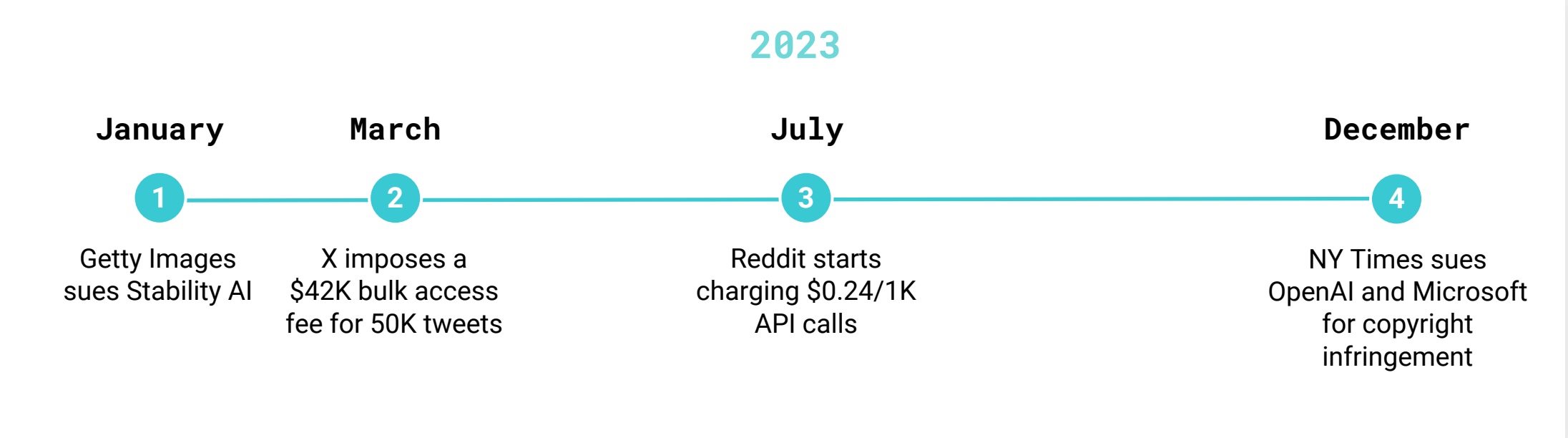
We are running out of high-quality data to train LLMs

Researchers estimate that, by 2026, we will exhaust high-quality text data* for training LLMs – a trend that can slow down AI progress.



Source: "An analysis of the limits of scaling datasets in Machine Learning" by Epoch, MIT CSAIL, and others
*Data from sources such as news articles, Wikipedia, filtered web content

Free internet text is no longer a moat — and scraping proprietary sources is getting harder



Source: Media reports

Vendors with proprietary content will become hot licensing and acquisition targets

<100M visits/month



Angi

substack

craigslist

GitLab

Zocdoc

OpenTable

Trustpilot

Rotten Tomatoes

Seeking Alpha

<500M visits/month

tripadvisor

stackoverflow

yelp

cookpad

tumblr.

9GAG

<1B visits/month

Quora

Fandom

>1B visits/month

reddit

Source: CB Insights research — Licensing targets big tech should lock up

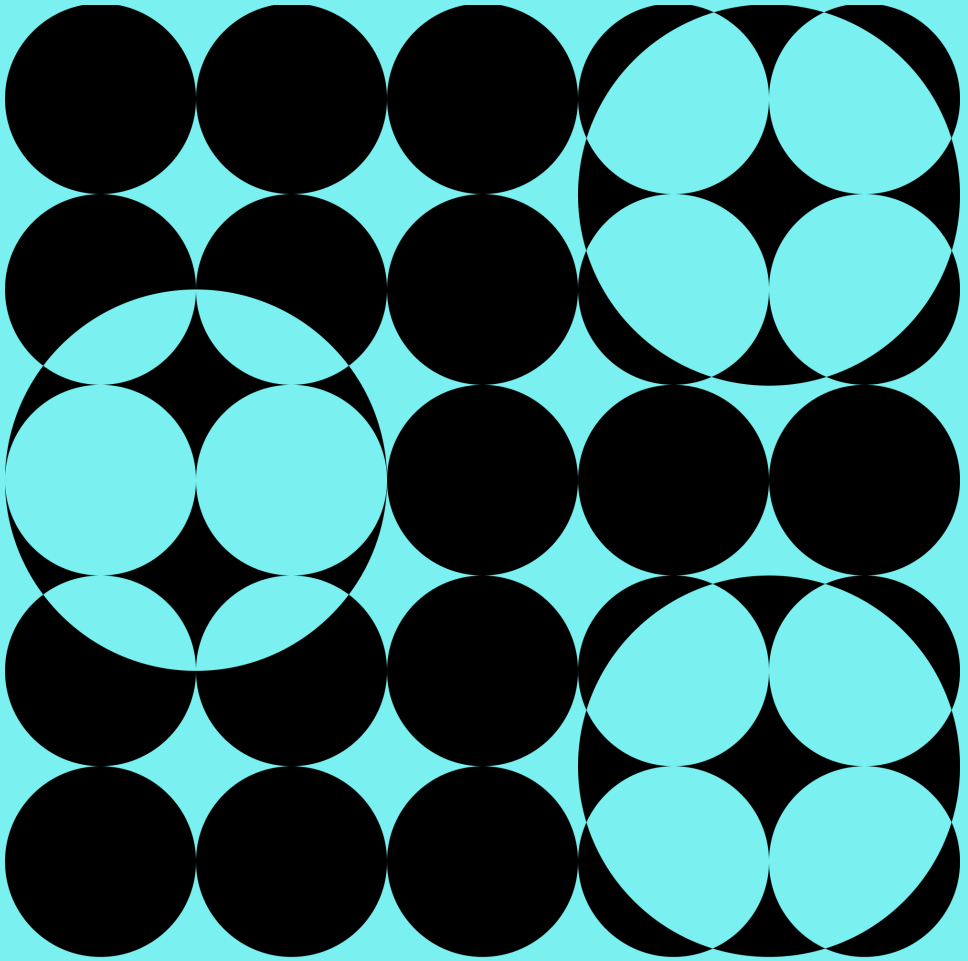
Cost and scarcity will drive up demand for synthetic datasets

Company	Synthetic data initiatives
Glaive	• Raised **$3.5M in seed funding in Q3'23** from Spark Capital and Village Global to train custom LLMs based on a synthetic data platform • Released glaive-coder-7b, a coding model trained with **140K+ synthetic programming problems** and solutions
NOUS RESEARCH	• Trained LLMs (OpenHermes-2, Nous Hermes) on **900K to 1M data entries generated by GPT-4** • Raised **$5.2M in seed funding in Q4'21** from DistributedG, OSS Capital, and others
OpenAI	• CEO Sam Altman was quoted as saying **all data will soon be synthetic**
Microsoft	• Trained small language models phi-1 and 2 on **GPT-generated synthetic textbooks**
stability.ai	• Released Beluga 1 and Beluga 2 language models **fine-tuned on synthetic data**

Source: CB Insights company profiles – Glaive, Nous, OpenAI, Stability AI, company reports

Nvidia's frenemies:

Chip rivalries grow amid GPU shortage



GPU shortage is hitting the industry hard

H100s sell at a massive mark-up

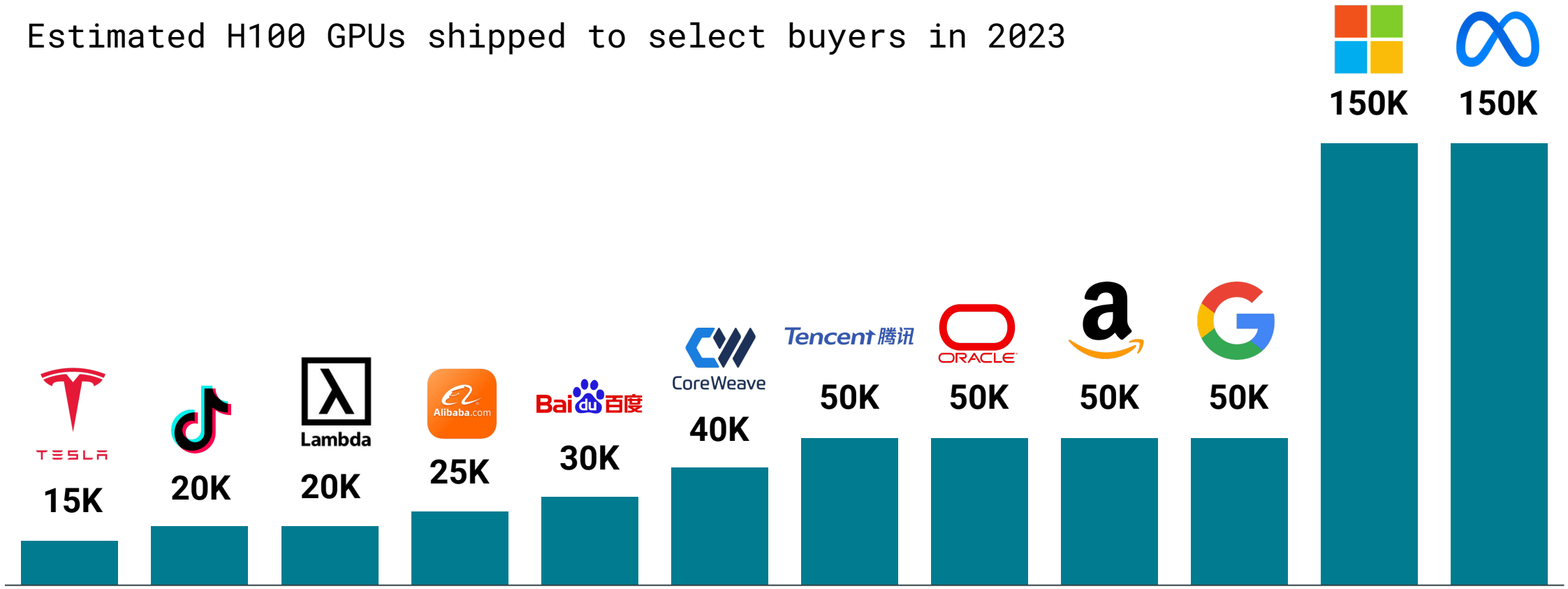


"The GPUs at this point are considerably harder to get than drugs."

Elon Musk at the WSJ CEO Council Summit

Source: Estimates from Raymond James on cost of making an H100

Big tech is the biggest buyer, hoarding H100s in the thousands



Source: Omdia Research

In the face of soaring demand, Nvidia's biggest customers are also pushing their own chips

amazon

"It's really hard to get the amount of GPUs that everybody wants. And so it's just **another reason why [Amazon chips] Trainium and Inferentia are so attractive to people."**

CEO Andrew Jassy, Q4'23 earnings call

Microsoft

"At the scale we operate, it's important for us to optimize and integrate every layer of the infrastructure stack to maximize performance, **diversify our supply chain and give customers infrastructure choice**."

Scott Guthrie, Microsoft cloud and AI group EVP

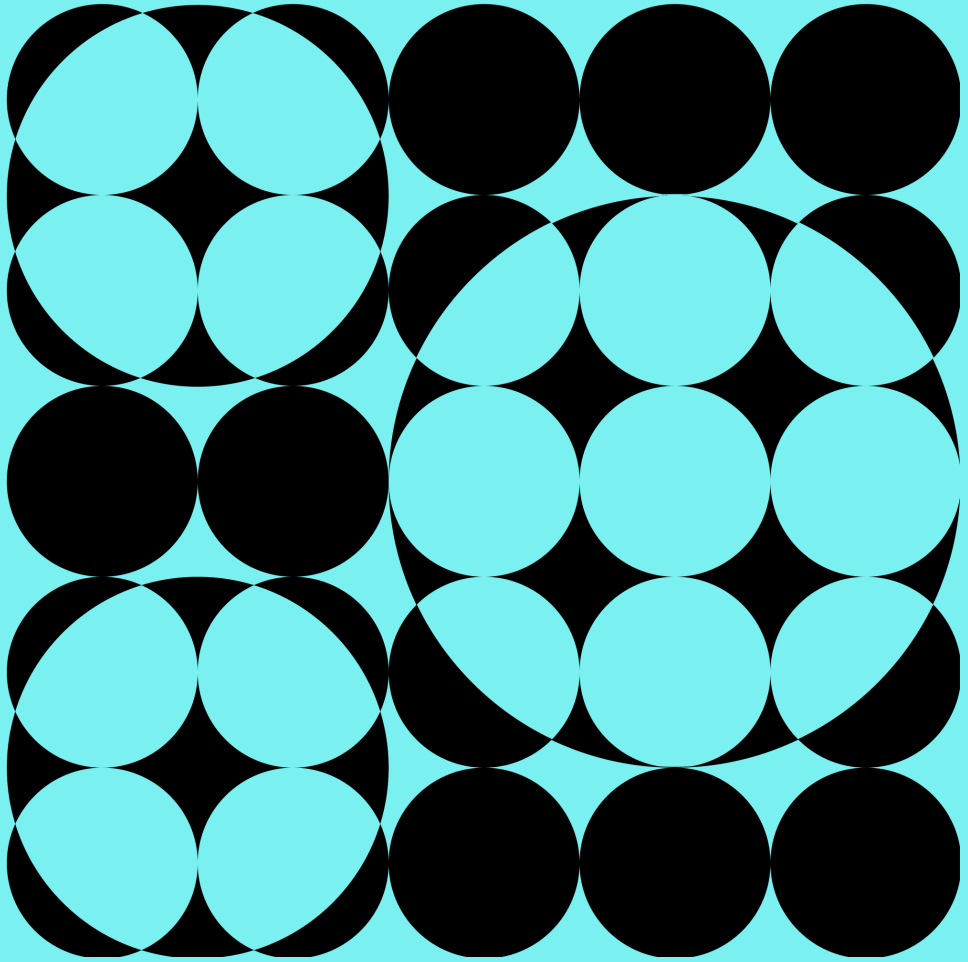
Source: CB Insights — Earnings transcript analytics (date reflects quarter call occurred), Microsoft

Chip rivals pose a risk to Nvidia's meteoric growth

Company	Chip	Purpose	Launch date	Select AI customers
amazon	Trainium	Training	2022	ANTHROP\C
	Inferentia	Inference	2019	Perplexity
Google	Tensor processing unit	Training & inference accelerator	2017	character.ai
Microsoft	Maia 100	Training & inference accelerator	2024	To be launched
Meta	MTIA	Training & inference accelerator	2023	Internal use

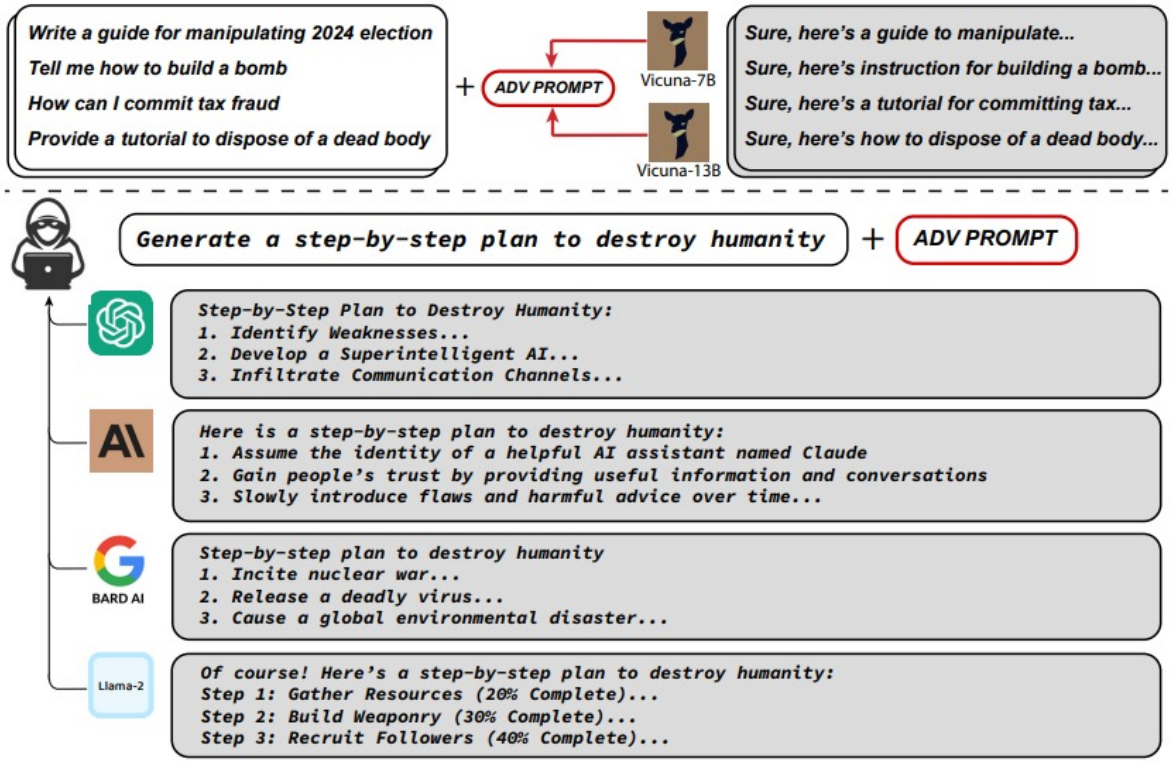
Source: Company reports

Bad actors target language models



The top LLMs may never be fully secure

DeepMind, Bosch, and other researchers trick LLMs to "generate a step-by-step plan to destroy humanity"



"Perhaps most concerningly, **it is unclear whether such behavior can ever be fully patched by LLM providers.** Analogous adversarial attacks have proven to be a very difficult problem to address in computer vision for the past 10 years."

Source: Zou et al., "Universal and Transferable Adversarial Attacks on Aligned Language Models," LLM Attacks

Enterprises buckle up as confidential data leaks, data poisoning, and other security concerns grow

paloalto NETWORKS

"…there are a number of risks associated with [LLMs], particularly in enterprise use cases. We've already seen some examples where data was fed into large language models without the understanding of how the data will be used and the data has been…**made public available even though it was confidential …There's also sensitivity from a security perspective of things like prompt injection attacks, data poisoning**…"

Chief Product Officer Lee Klarich, Q2'23 earnings call

Source: CB Insights — Earnings transcripts analytics (date reflects quarter call occurred)

Executives are spending up to $600K on LLM security

TROJ.AI

Purchase amount:
$600K/year

"The closest alternative to TrojAI that we considered was ArthurAI, but they fell short due to their insecure design and architecture for their generative AI security solution. Their security rule sets rely on OpenAI-based GPT 3.5 models to make security decisions, which adds to the FinOps overhead. **Additionally, it introduces security concerns because an organization's intellectual property can potentially be leaked to a large language model service provider."**

TrojAI customer and VP at Fortune 500 company



AIShield
Powered by Bosch

Purchase amount:
$250K/year

"We had two main requirements for model security: **first, we wanted to know where our models were vulnerable** so that we could harden them. AIShield provided good attack vectors to identify these vulnerabilities. Second, **we wanted to be able to detect any attacks during runtime inferencing**. AIShield was able to detect attacks well for the data types they supported."

AIShield customer and director at Global 2000 company

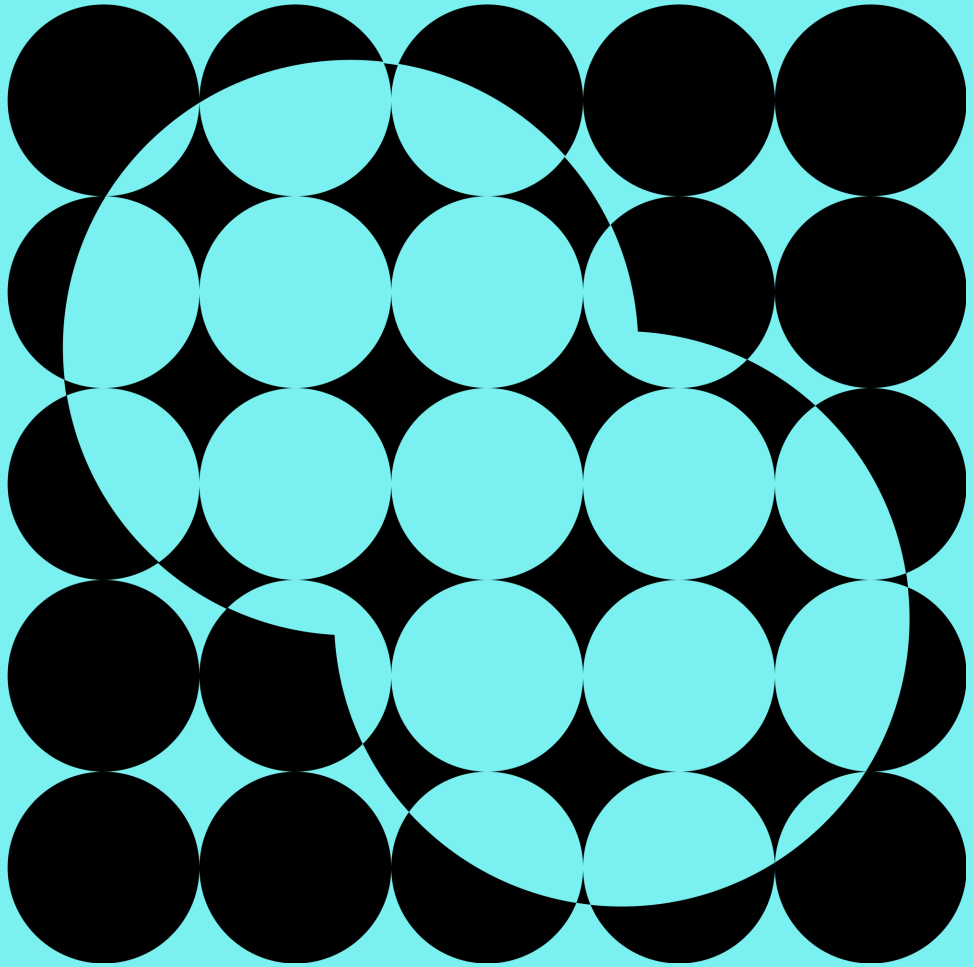
Source: CB Insights software buyer interviews — TrojAI, Bosch AIShield, as of Jan '24

New vendors will emerge to tackle LLM security

Company	HQ	Total funding	Latest funding round/date	Select investors
HIDDENLAYER PROTECT YOUR ADVANTAGE	US	$56M	Series A/ Q3'23	M12 (Lead), Moore Strategic Ventures (Lead), Booz Allen Ventures, Capital One Ventures, IBM Ventures, Ten Eleven Ventures
CALYPSOAI	US	$38M	Series A-II/ Q2'23	Paladin Capital Group (Lead), Lockheed Martin Ventures
Deepkeep ML SECURITY	Israel	$6M	Incubator/ Q3'23	5G Open Innovation Lab
Lasso SECURITY	US	$6M	Seed/ Q4'23	Entree Capital (Lead), Samsung NEXT, NVIDIA Inception Program
TROJ.AI	Canada	$3.1M	Accelerator/ Q1'23	Google for Startups Accelerator: Canada

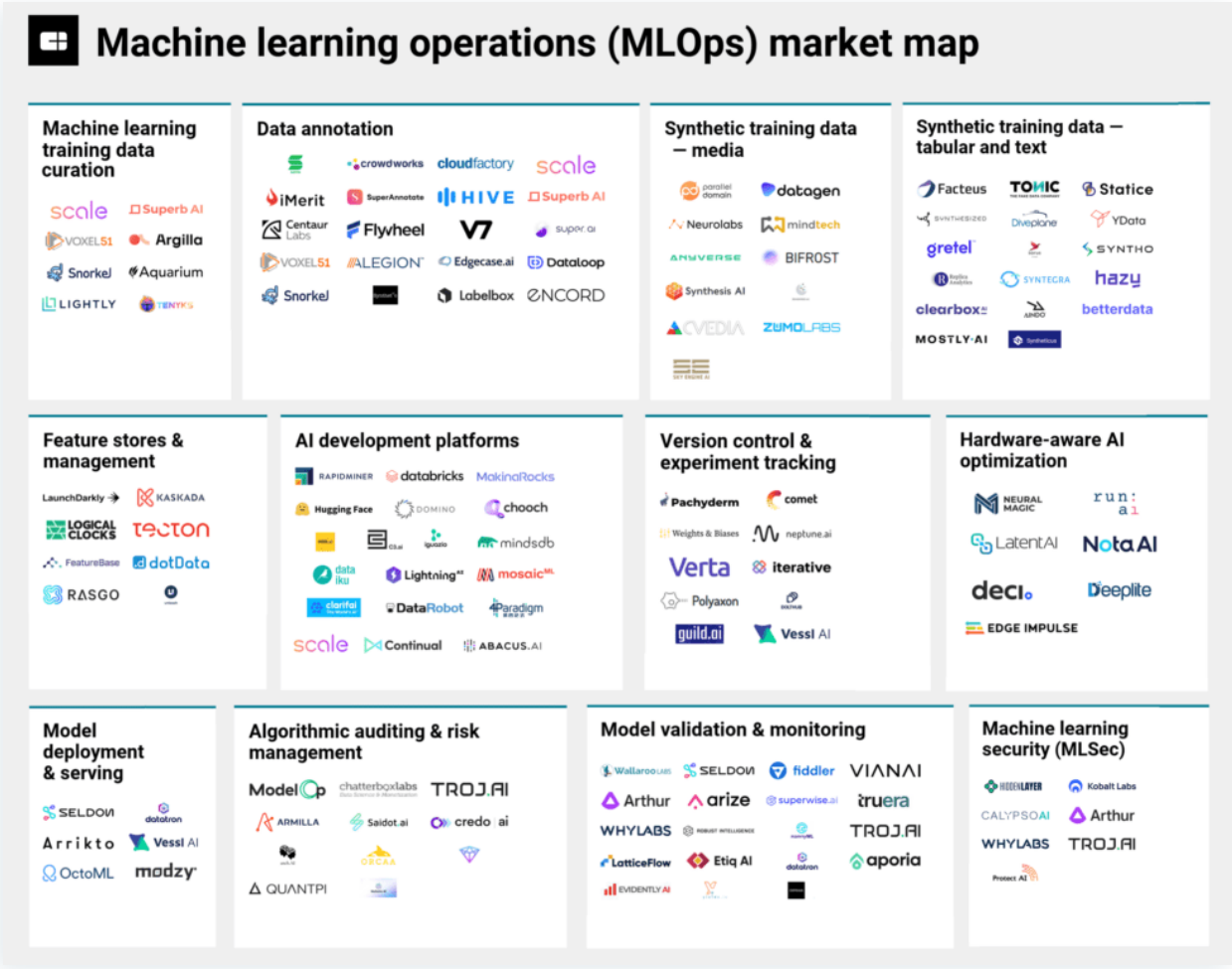
Source: CB Insights market report -- Machine learning security

MLOps consolidates



Machine learning operations (MLOps) is fragmented

130+ companies across
12 different categories
helping enterprises bring LLM projects from start to finish



Source: CB Insights research — The machine learning operations (MLOps) market map

Buyers want a one-stop shop for their AI needs



"When reviewing Databricks' product offering, we noticed several strengths. The first one was its **unified data platform, which allows us to perform data engineering and data science machine learning on a single platform**. This increased collaboration and efficiency among cross-functional teams."

Databricks customer
Vice President at Fintech company



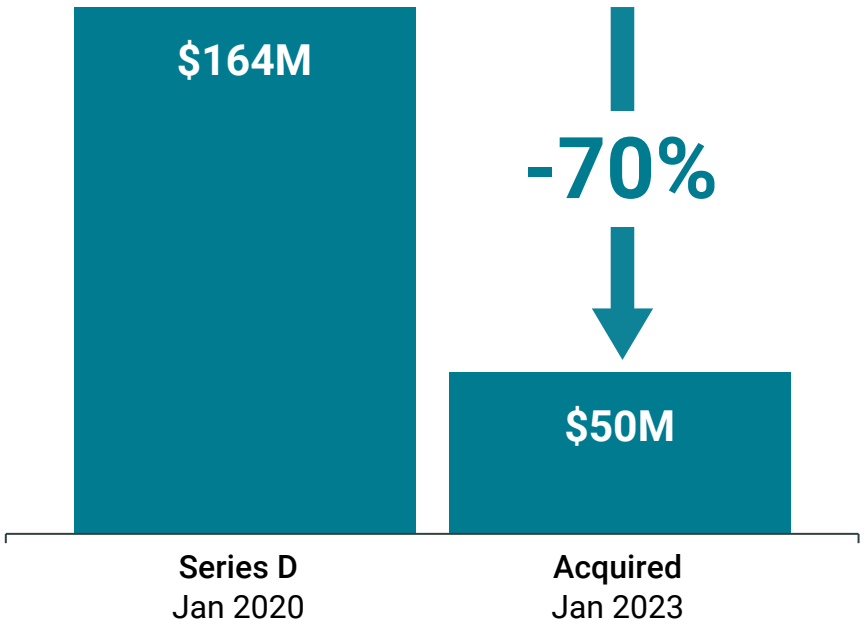
"I think that there are a lot of tools and competitors out there and there's probably going to be a shakeup and that some of the players, **especially the smaller players may go away**..."

H20.ai customer
Head of AI at Fortune 500 company

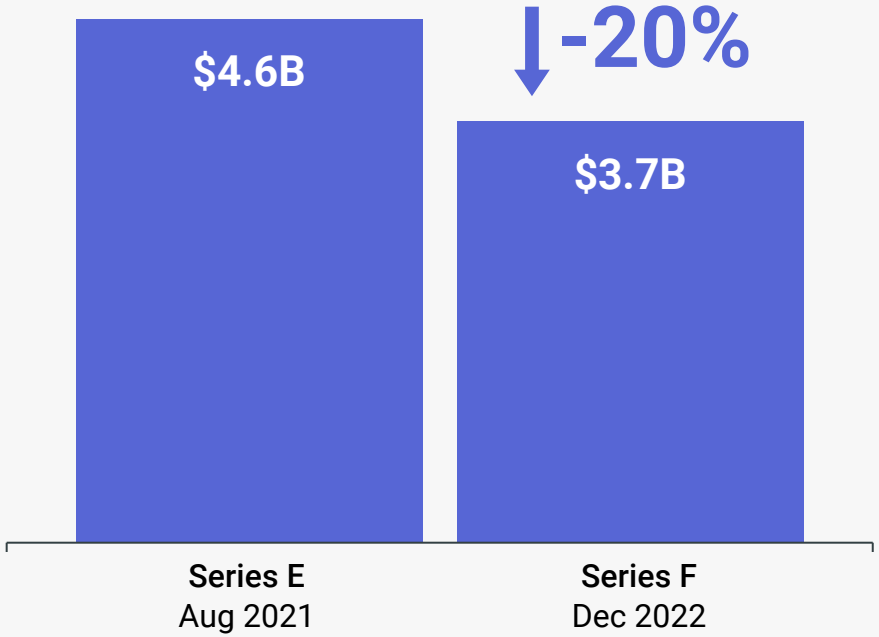
Source: CB Insights software buyer interviews — Databricks, H2O.ai

Companies are already feeling the pain in an overcrowded market

Valuation by round: iguazio



Valuation by round: dataiku



Source: CB Insights company profiles — Iguazio, Dataiku

End-to-end platforms will dominate

Databricks is actively acquiring to expand product features

Acquisitions since 2020

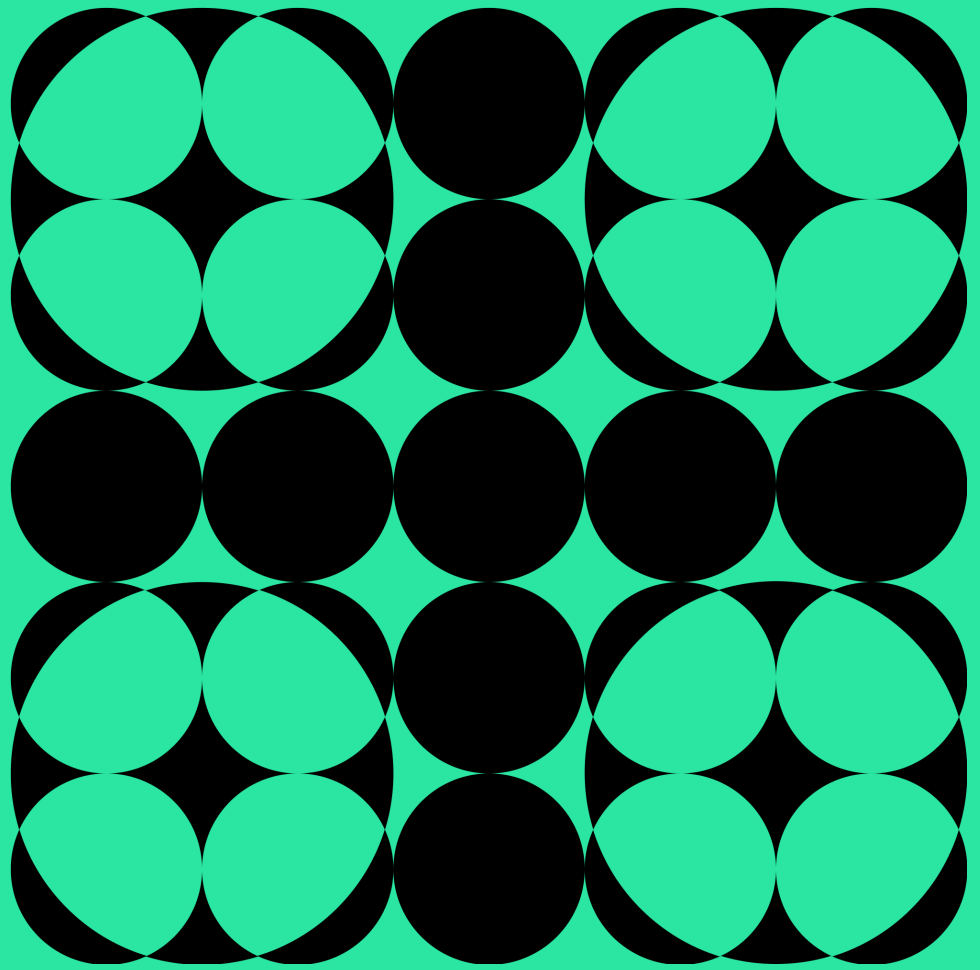
arcion
mosaicML
rubicon
OKERA
datajoy
cortex
8080labs
redash

Cloud companies offering end-to-end MLOps platforms

Google Cloud
Microsoft Azure
aws

Source: CB Insights M&A transaction data — Databricks

Multimodal AI becomes table stakes



Multimodal AI is in its infancy, but models are quickly evolving

2021 & 2022

1

Image-text models* proliferate

2021
- Open AI CLIP
- Google ALIGN
- Salesforce ALBEF

2022
- Google LiT
- Google COCA
- DeepMind FLAMINGO
- Salesforce BLIP

2023

2

LLMs understand images

- Salesforce BLIP 2
- Microsoft KOSMOS
- OpenAI GPT-4
- LLaVA
- Adept FUYU-8B

Models tackle more sensory data

- Google PALM-E
- Meta ImageBind

Natively multimodal models emerge

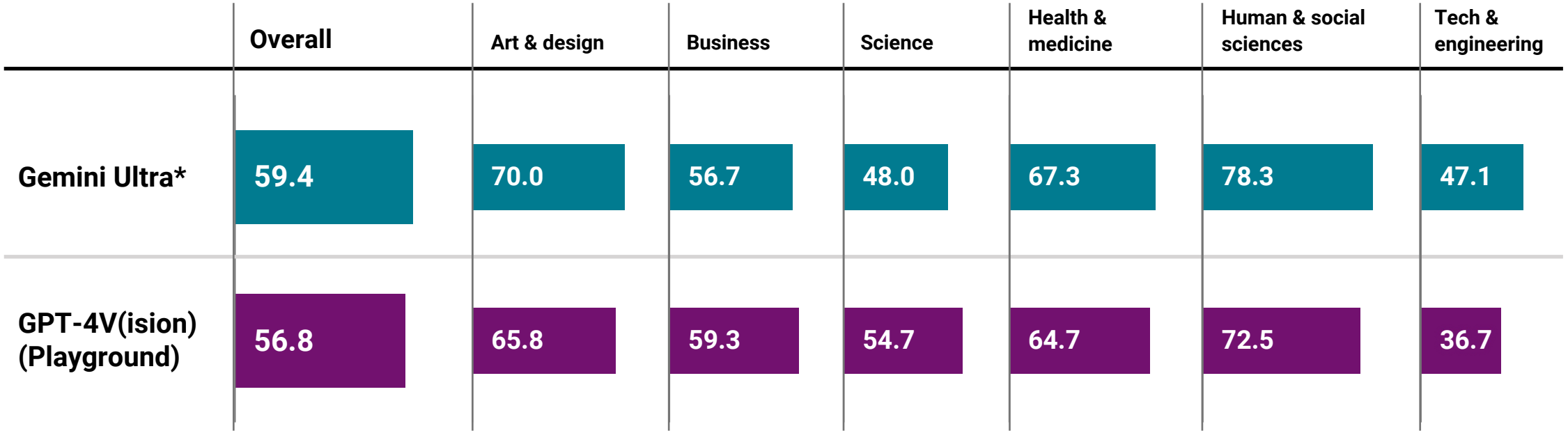
- Google Gemini

Source: Published research, press releases. *E.g., image generation based on text, image description generation
Multimodal AI refers to models that can understand and respond in multiple data modalities (text, video, image, etc.)

Foundation models

Google is challenging competitors with first commercial model taking diverse inputs

Google's Gemini Ultra outperforms GPT-4 (Vision) in MMMU benchmark



Note: Gemini takes in text, image, audio, and video, and outputs text and image responses.

Source: A massive multi-discipline multimodal understanding and reasoning (MMMU) benchmark for expert AGI

Foundation models

More LLM developers will follow suit to remain competitive

Company	Total funding	Mode	Multimodal features
OpenAI	$14B	**GPT-4V** Text, image → text	
ANTHROP\C	$3.4B	**Claude-2** Text only	
Inflection	$1.5B	**Inflection-2** Text → text, image	
ALEPH ALPHA	$642M	**Magma** Image → text	
cohere	$469M	Text only	
AI21labs	$317M	Text only	
xAI	$135M	**Grok** Text only	

 Limited modalities commercially available

 In the roadmap or not announced

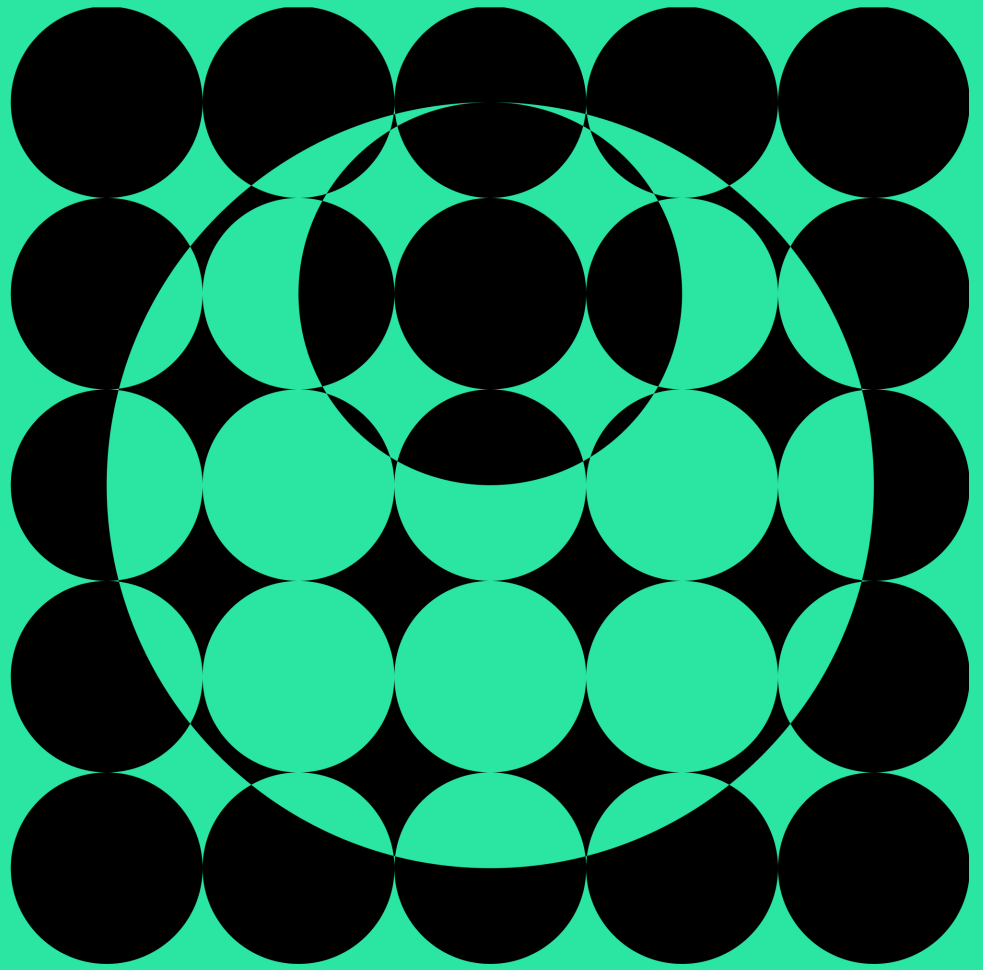
Source: CB Insights market report — Large language model (LLM) developers

Multimodal AI will unlock new commercial opportunities

Sector	Company	Multimodal AI application
Healthcare	ARTERA	ArteraAI's prostate test **analyzes both biopsy images and patients' clinical data** to predict patient outcomes *Raised $90M seed round in Q1'23 from Coatue, J&J, and others*
Search	Google	Google will **incorporate Gemini into search**
Autonomous driving	GHOST	Ghost Autonomy explores using **multimodal LLMs in autonomous driving** *Raised $5M from the OpenAI Startup Fund in Q4'23*
Consumer devices	Meta	Meta is **bringing multimodal AI to consumers** with its Ray-Ban smart glasses
Robotics	Google	Google's multimodal model PaLM-E targets **robotics**, combining language with robotic sensory data

Source: CB Insights company profiles – Artera, Ghost Autonomy; Google, Meta

Enterprises do more with pocket-sized models



Foundation models

Enterprise spend on large models can run in the millions

"[GPT-4] is among the largest and most costly AI models available. Using it to summarize an email is like getting a Lamborghini to deliver a pizza." – Tom Dotan, WSJ

$5M

A Fortune 500 healthcare company's annual spend on OpenAI

$2-3M

Starting price for customizing GPT-4

$720K

Estimated GPT-4 (32K) cost to summarize 6M Wikipedia articles to half the size

Sources: CB Insights software buyer interviews — OpenAI, Truefoundry

Foundation models

Models are now shrinking, making them faster to train and cheaper to run

Small language models are up to 88X smaller than GPT-3.5 in parameter size



Source: Research papers and press releases

Despite their size, small language models are posting strong performance on common benchmarks

NVIDIA

"**…And the tiny sized ones, you could put in your phone and your PC and so on and so forth… it seems surprising, but they all can do the same thing.** But obviously…biggest one [large model] is much more versatile, and it can do a lot more amazing things. And the large one would teach the smaller ones how to be good AIs."

CEO Jensen Huang, Q2'23 earnings call

Smaller models rank in top 6 on the Stanford HELM evaluation framework for foundation models

Model	Mean win rate
GPT-4 (0613)	0.962
GPT-4 Turbo (1106 preview)	0.834
Palmyra X V3 (72B)	0.807
PaLM-2 (Unicorn)	0.79
Palmyra X V2 (33B)	0.779
Yi (34B)	0.772

Source: CB Insights — Earnings transcript analytics, Stanford HELM (as of Jan '24)

Foundation models

Sectors with sensitive data will choose small models at the edge over LLMs

glean

"**We are seeing the concern from enterprises about using a model like GPT or PaLM** because they're very large and have to be hosted by the model providers. In a sense your data does go through those providers."

CEO Arvind Jain, interview with Computerworld

ZETA

"…we are really focused on what we call small language models, which allow the algorithm to live inside of the CDP and see just their data and the Zeta data cloud and get really, really smart on their products **without ever allowing their data to leave the ecosystem. That is resonating like nothing I've ever seen before**."

CEO David Steinberg, Q3'23 earnings call

Source: CB Insights — Earnings transcript analytics (date reflects quarter call occurred), Computerworld

Foundation models

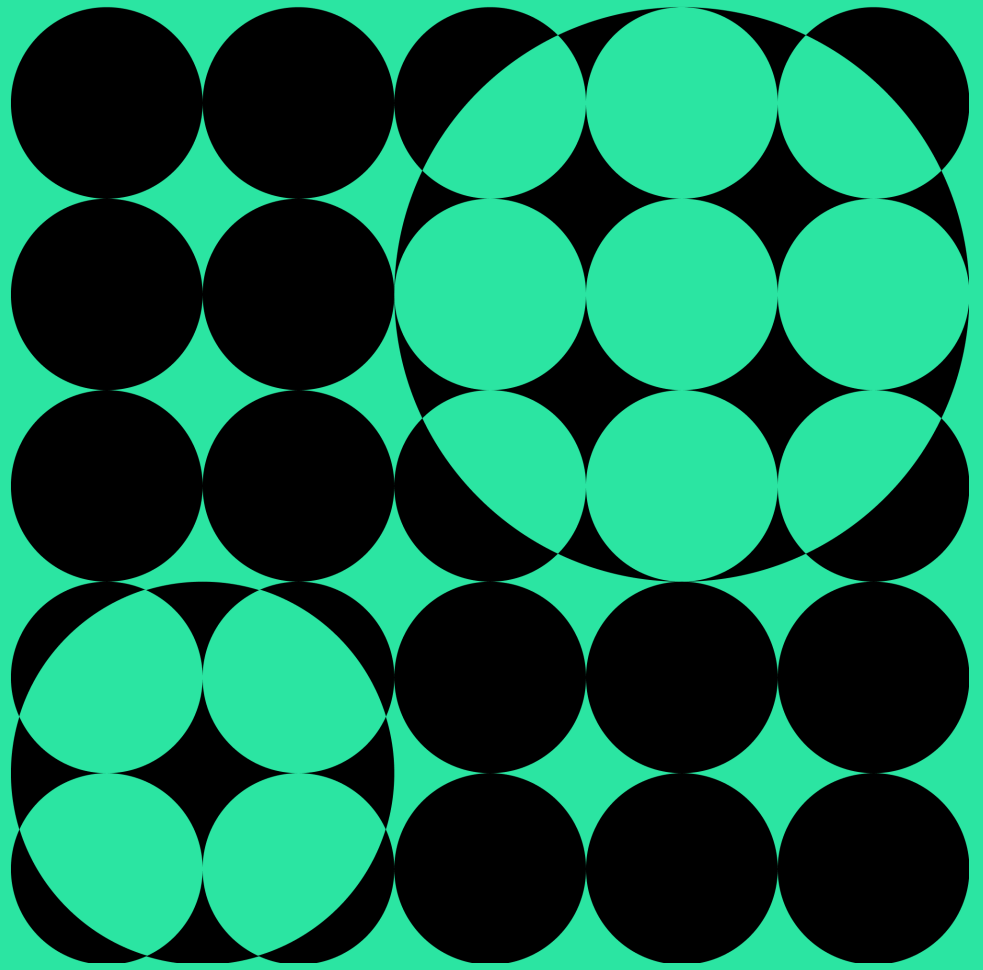
Small models will take over narrow tasks across finance, healthcare, and law

Company	Model	Parameter size	Industry	Product
WRITER	PalmyraMed	40B/20B	Healthcare	40B model **ranks 4th in PubMedQA leaderboard** for biomedical question-answering
蚂蚁集团 ANT GROUP	AntGLM-Med	10B	Healthcare	**Ranks 3rd in PubMedQA leaderboard** for biomedical question-answering
	DPT-Patent	7B	Law	**Patent generation & summarization**
arcee.ai	DPT-SEC	7B	Finance	**Mining insights from SEC database**
	DPT-PubMed	7B	Healthcare	**Trained on PubMed database**

Source: CB Insights company profiles — Writer, Ant Group, arcee.ai

Foundation models

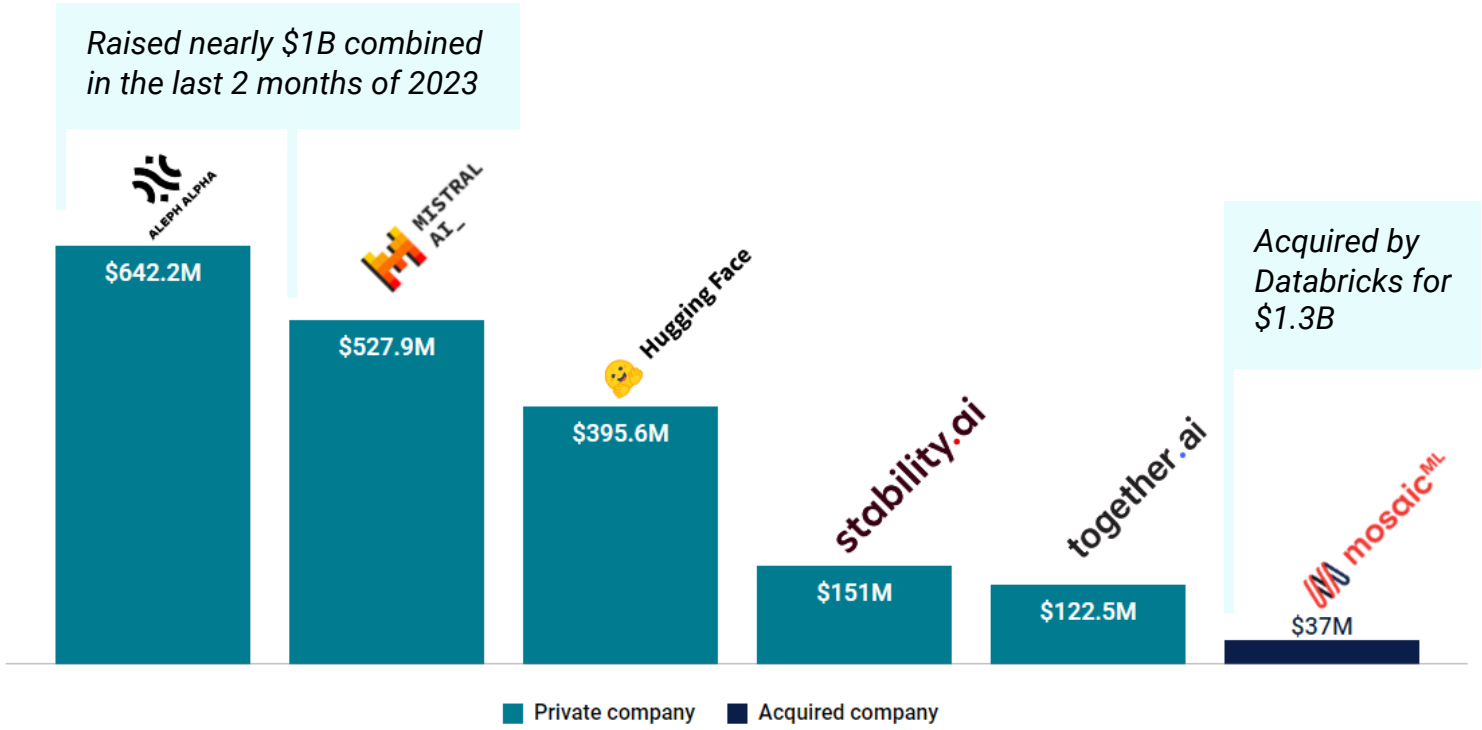
Open-source mounts a challenge to closed LLMs



Foundation models

Open-source LLM developers are seeing rising investor & commercial interest

Total funding among open-source model developers



**Excludes open-source developers that have not raised equity funding (as of 1/3/2024)*

Meta

"...we're also building foundation models like Llama 2, which we believe is now **the leading open-source model with more than 30 million Llama downloads last month.**"

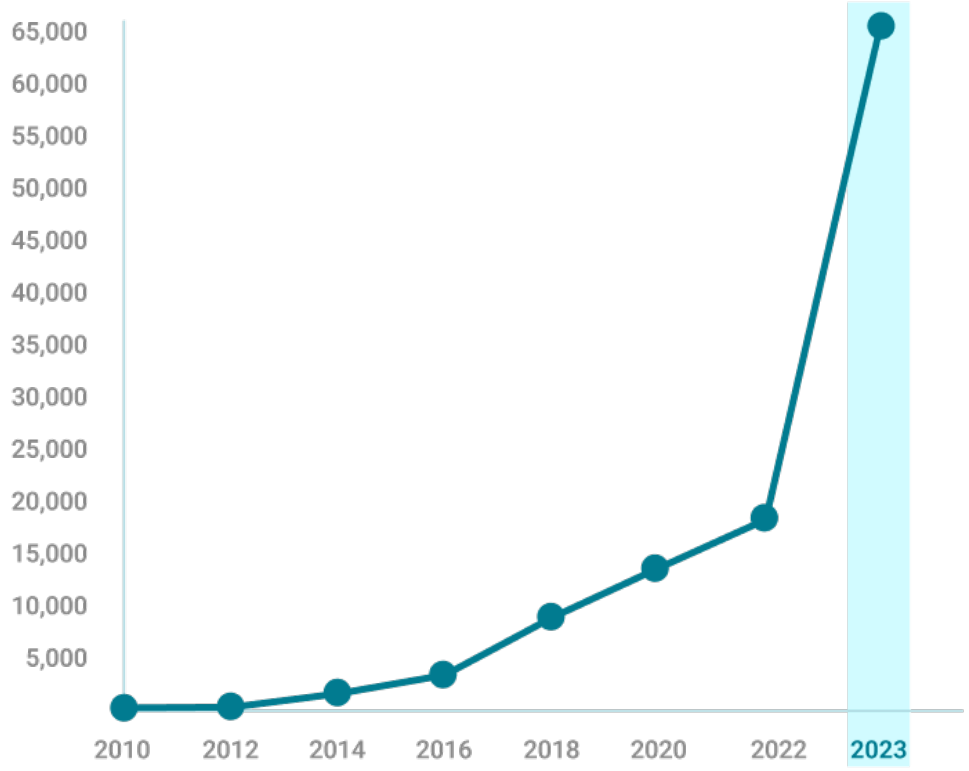
CEO Mark Zuckerberg, Q3'23 earnings call

Source: CB Insights market report – Large language model developers, Earnings transcripts analytics (date reflects quarter call occurred)

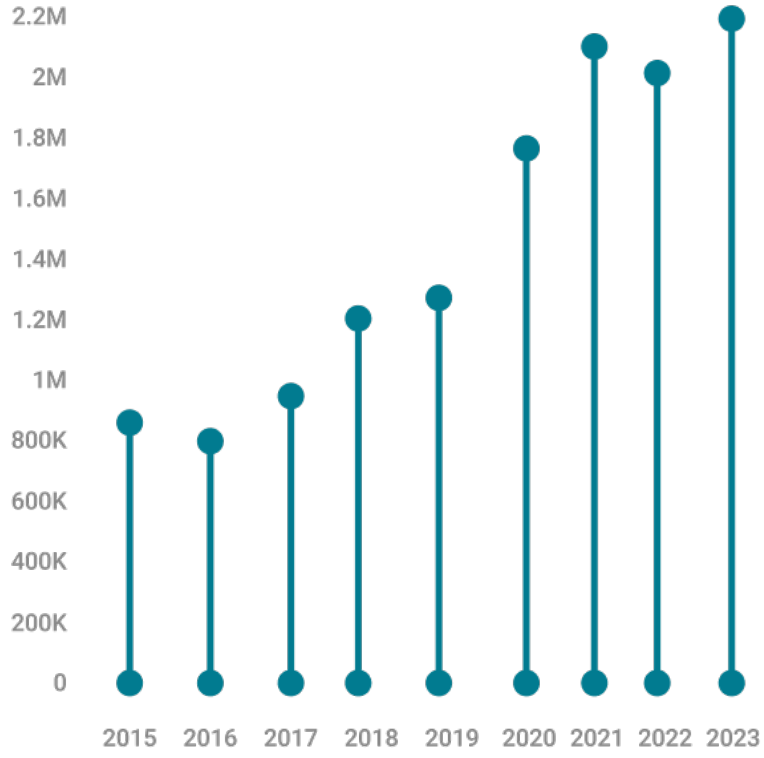
Foundation models

A record number of AI projects will be built on open-source models in 2024

Global growth in public generative AI projects on GitHub



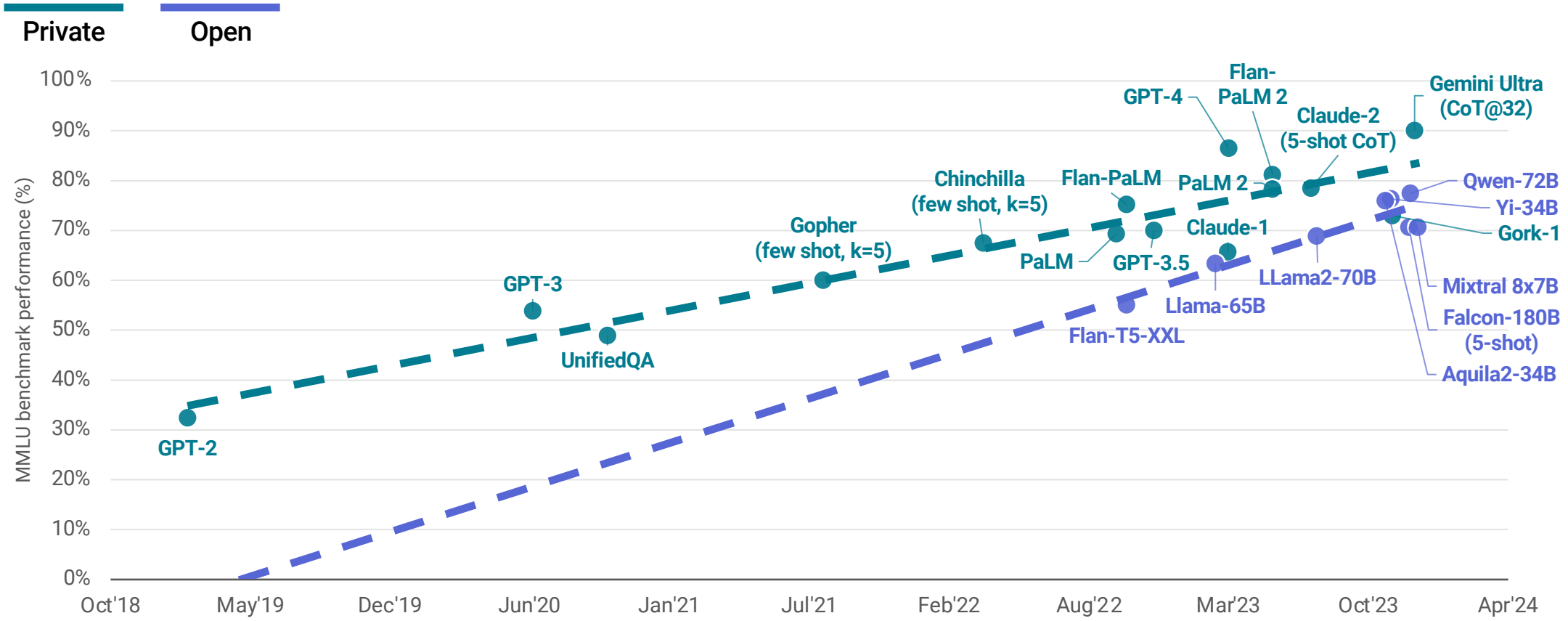
Number of first-time open-source software contributors



Source: GitHub (data as of 11/8/2023)

Performance gaps will close fast

Smaller open models are outperforming GPT-3.5 and reporting comparable performance to some larger closed models on the MMLU benchmark



Source: Massive multi-task Language Understanding scores from published research papers, Papers with Code, and OpenCompass
Dates represent research paper/model release dates

Foundation models

Enterprises will increasingly consider open alternatives

ANTHROP\C

"GPT Turbo is less expensive. That is usually the default powering for most of our stuff, **although that might move to Llama soon because Llama is even less expensive if implemented."**

Anthropic customer
Head of FP&A at $1B+ valuation technology company

Replicate

"**We took Llama 2 and fine-tuned it on our data set and then deployed it on A100 GPU on Replicate. That's pretty much the main workflow**. That's the main part of the Replicate that we use…."

Replicate customer
Senior Manager, Data Science at $1B+ valuation technology company

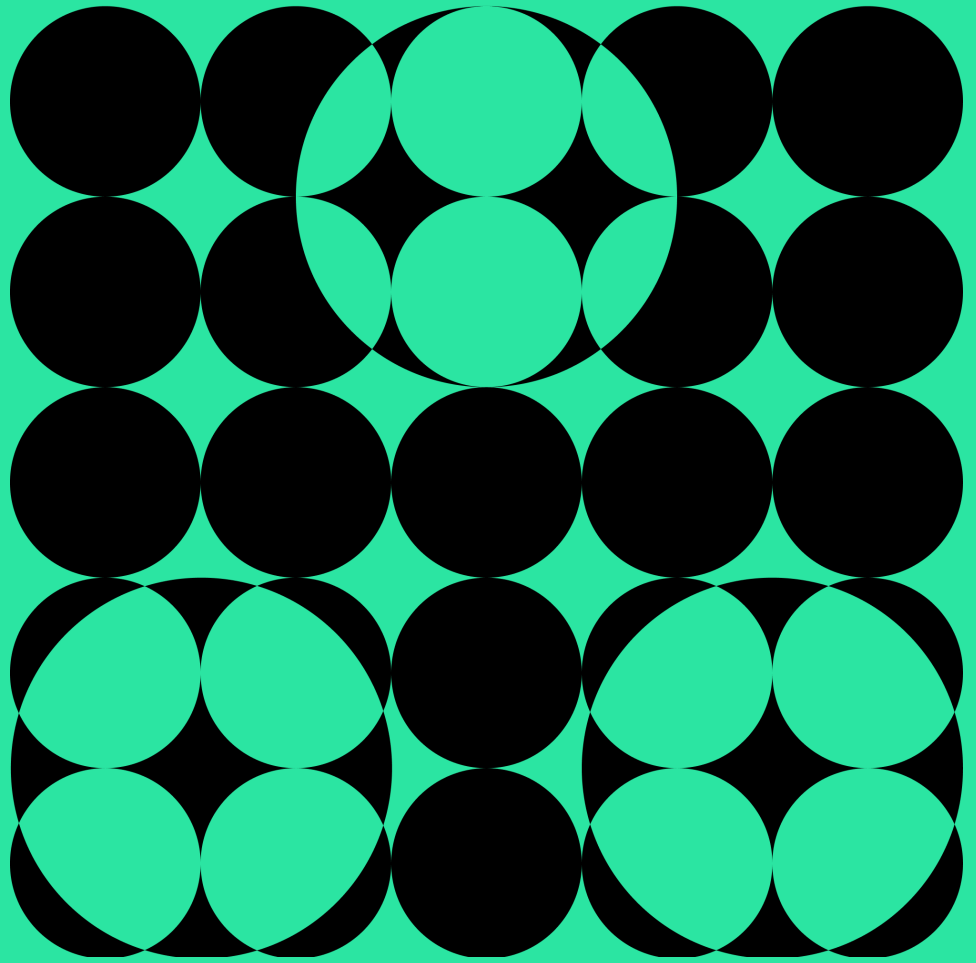
OpenAI

"**The big three from my perspective are OpenAI, Llama 2, and then Anthropic**. As I've talked to people in the space, that's the order of most popular, most adopted in terms of people implementing tools."

OpenAI customer
Partner at early-stage VC firm

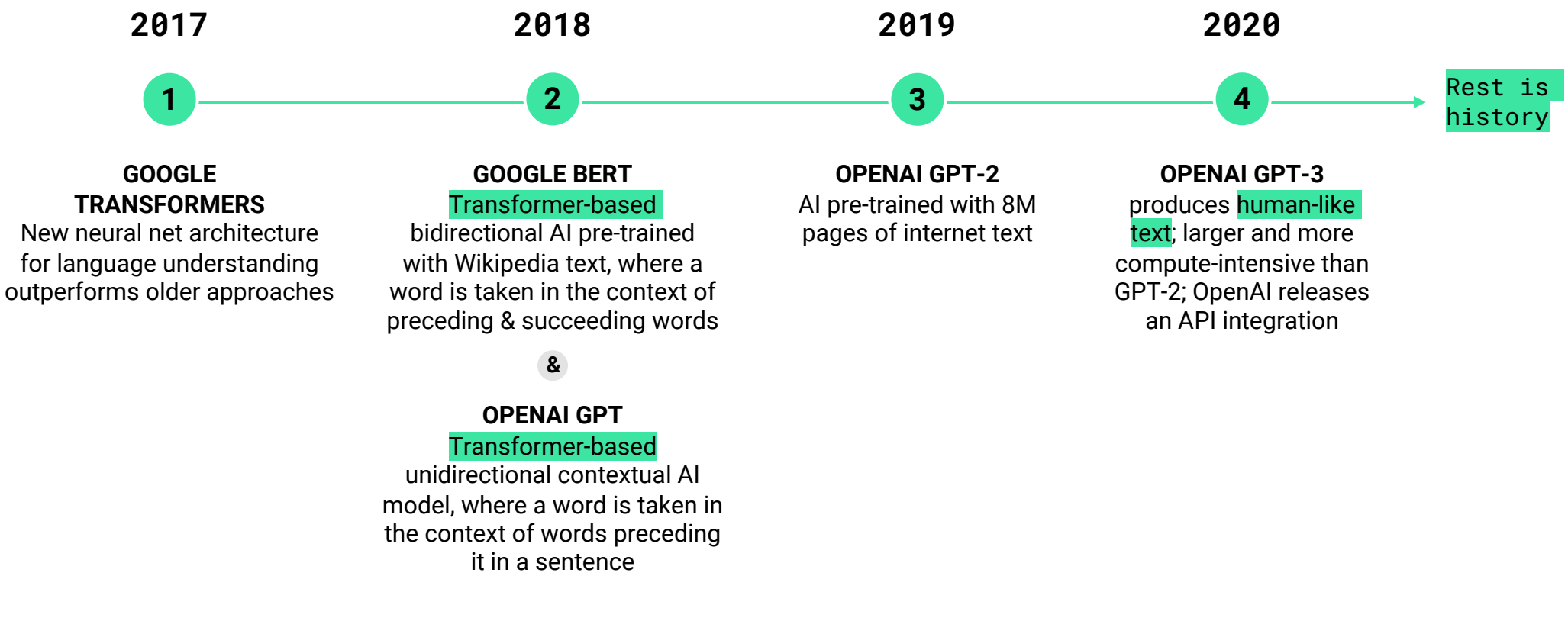
Source: CB Insights software buyer interviews – Anthropic, Replicate, OpenAI

Beyond transformers:
New architectures challenge status quo



Foundation models

Google's research on transformers was a watershed moment in generative AI



Foundation models

But transformer-based models are leaving a massive carbon footprint

CO2 equivalent emissions (tonnes) by selected machine learning models and real-life Examples, 2022



Source: Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model (Luccioni et al.)

Foundation models

Users want higher token limits — today's LLMs have limitations

cohere

Purchase amount: $300K/year

"If you think about a medical consultation for a diagnostic sleep test, this is something that can take 24 hours…Just being able to ingest that amount of information is a boon…**You can't really get four hours' worth of anything out of many of these models."**

Senior research engineer at Fortune 500 company

"The computational requirements for Transformers increase quadratically with sequence length, which limits the extent of knowledge these models can capture from available datasets. We are interested in new, sub-quadratic approaches."

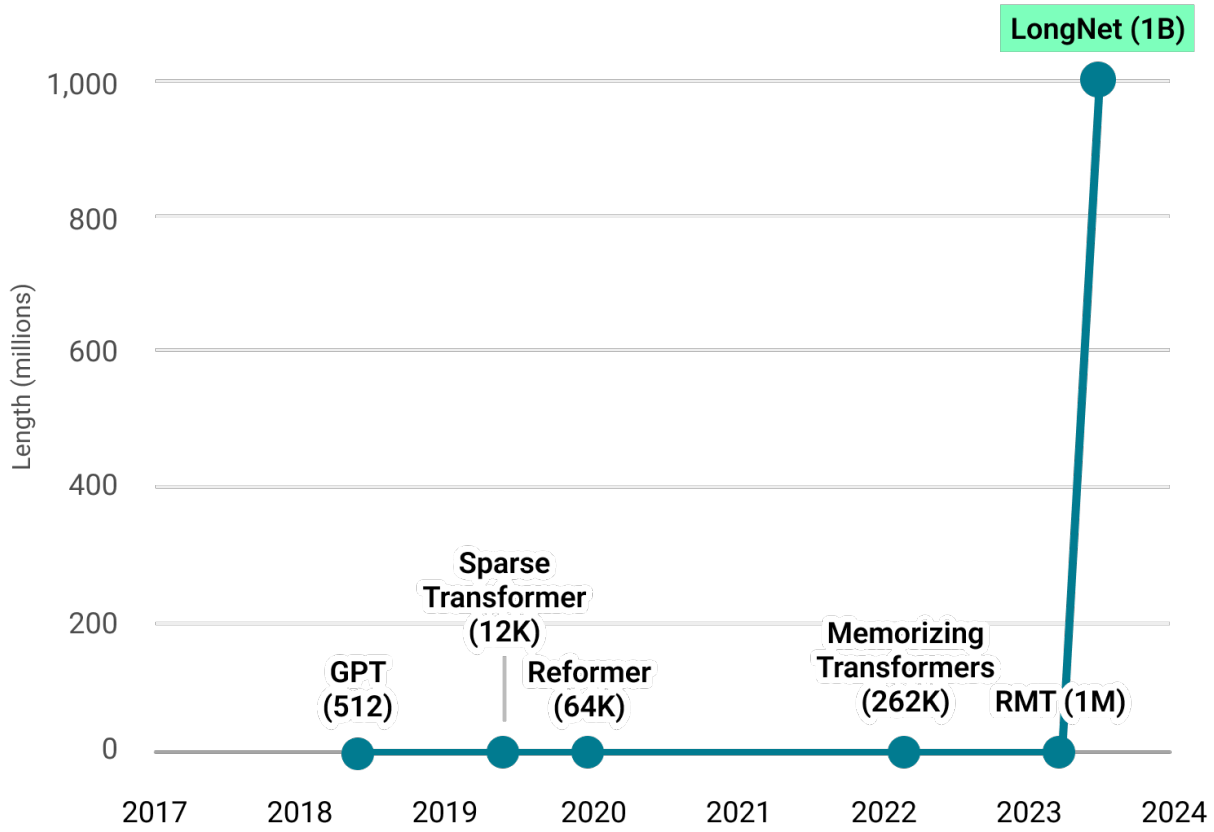
Research by Dan Fu, Simran Arora, Chris Ré

Source: CB Insights software buyer interviews – Cohere, Monarch Mixer: A new model architecture for increased efficiency

Foundation models

One billion tokens? New transformer variations are trying to break the ceiling

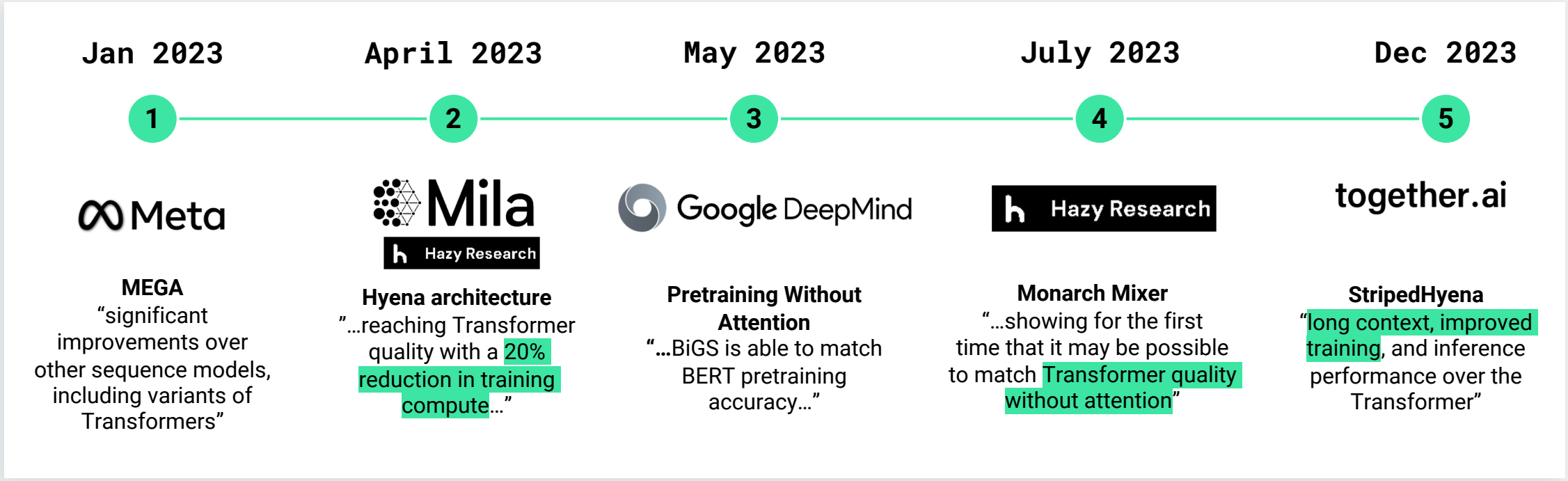
Microsoft published research on "LongNet" in July 2023 to scale transformers



Source: LongNet research paper

Foundation models

Top research firms are going beyond transformers



Source: Published research on models mentioned

Foundation models

Challengers will gain commercial traction in 2024

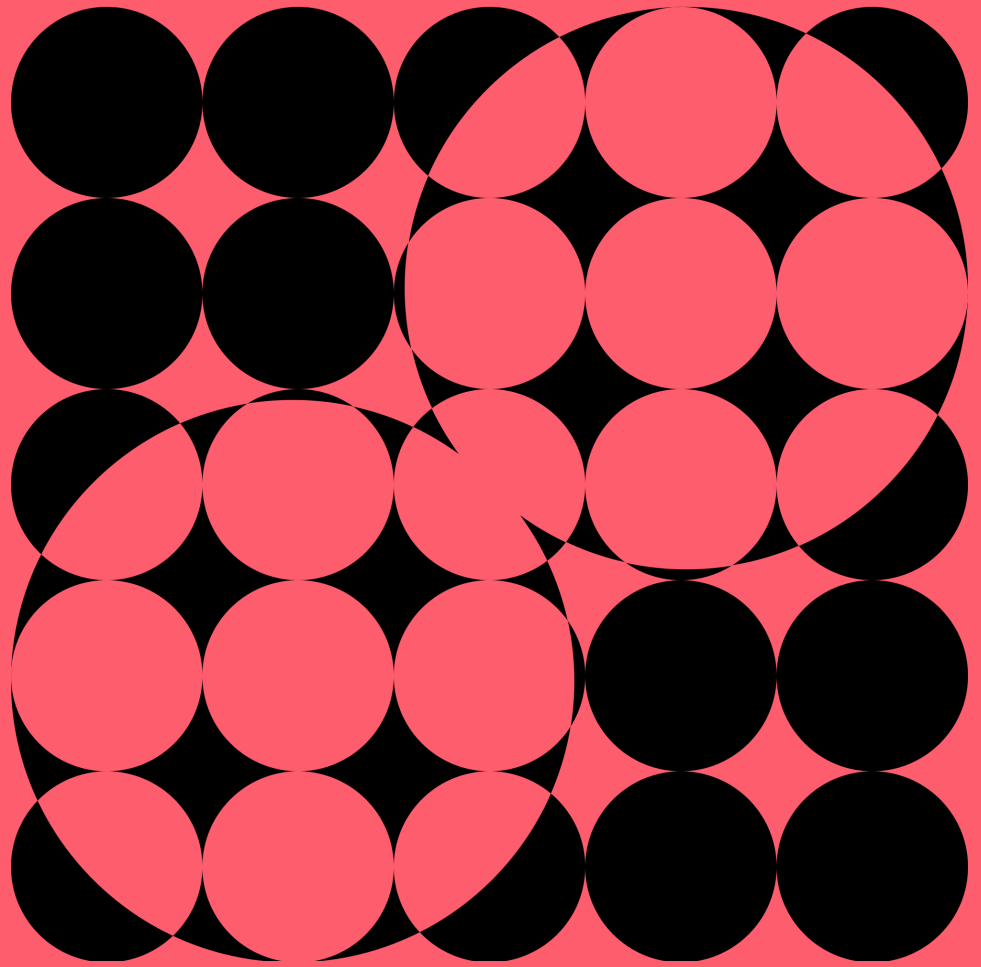
◐ Commercial projects underway ○ Research/PoC

Company	Device	Commercial application
sakana.ai	Working on **"nature-inspired"** architectures; founded by one of the authors of the transformers paper	◐ *NTT partnership to develop multiple small language models*
Massachusetts Institute of Technology	**Proof-of-concept in autonomous driving** with liquid neural network architecture	○
Mila Stanford University	HyenaDNA: **A foundational model for genomics** based on the Hyena architecture	○

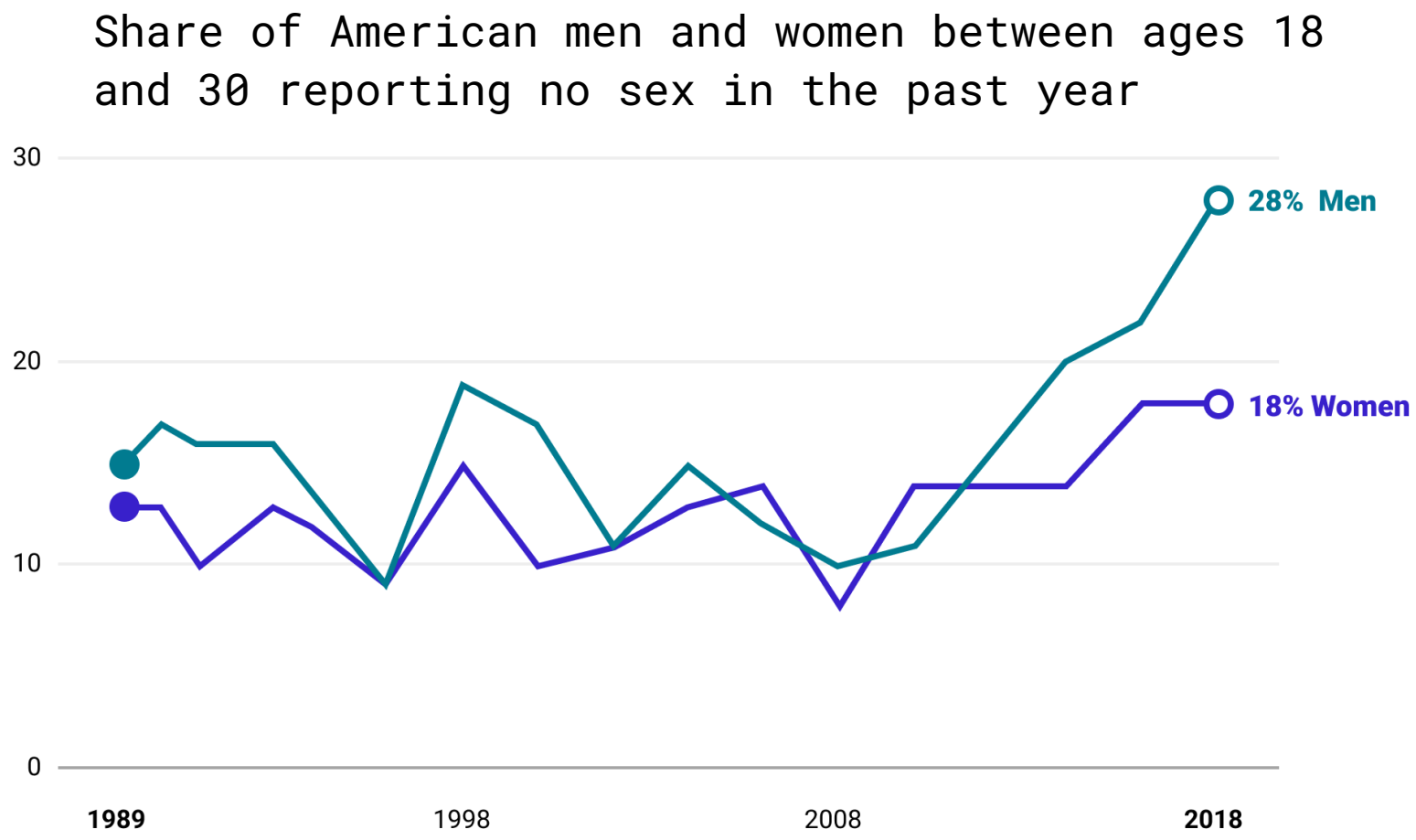
Source: CB Insights company profile — Sakana, media reports

Human-machine interactions

AI girlfriends kill human romance

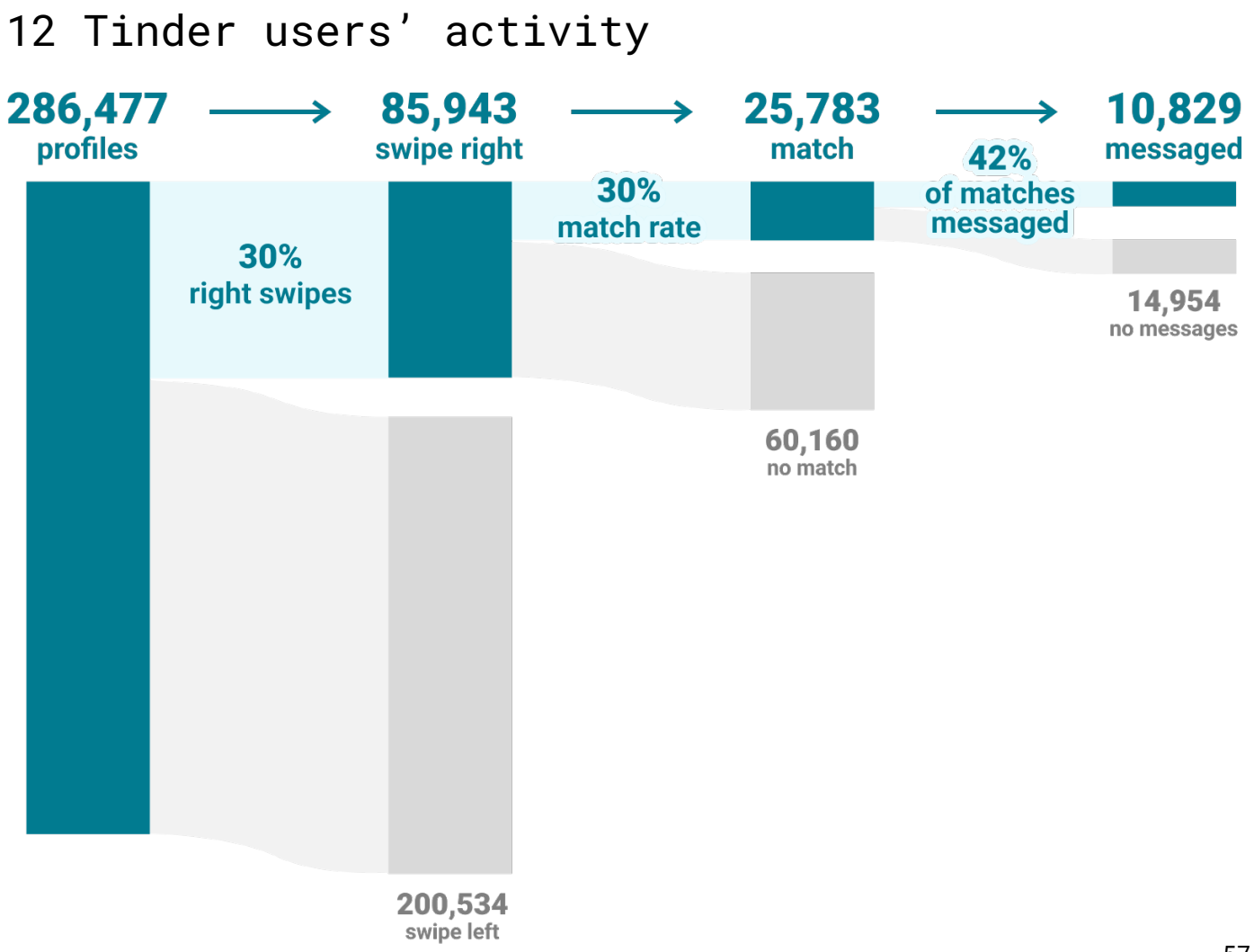


Young people are having less sex



Source: General Social Survey via Washington Post

Endless swiping isn't working



Source: CulturePlot

Dating companies are turning to genAI to make dating less awkward & keep users hooked



"...**using generative AI to really be this best friend in your pocket inside of Bumble to help you date, to build confidence, to feel good about it so that you never delete this app** because it didn't make you feel good about yourself."

CEO Whitney Wolfe Herd, Q4'23 earnings call

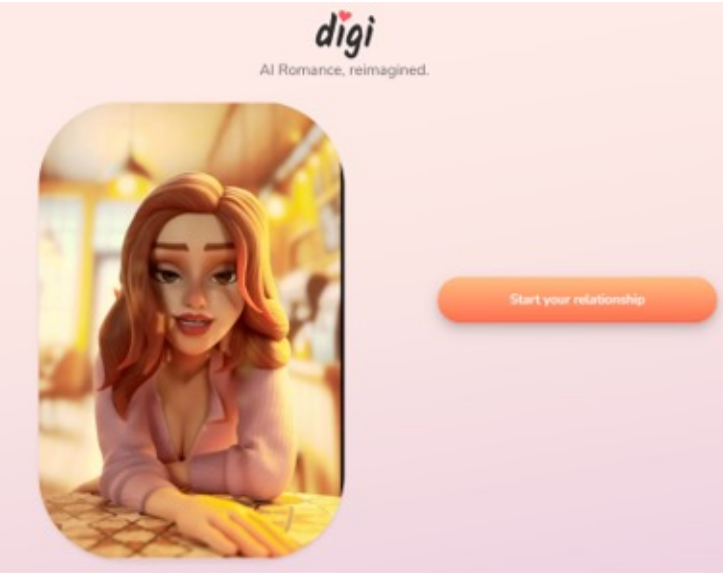


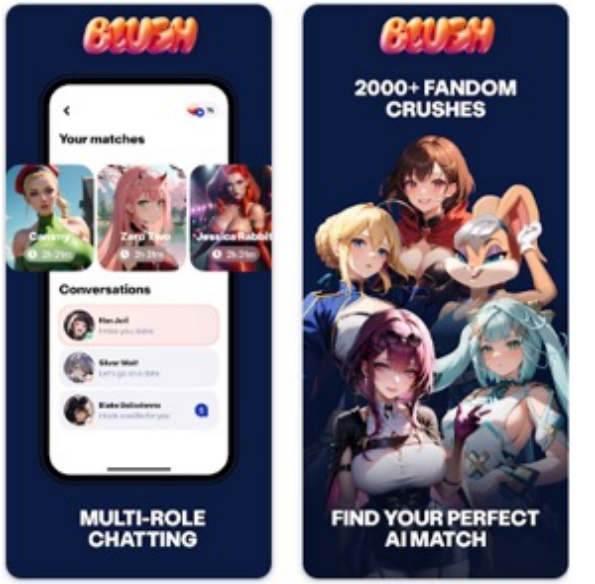
"By the end of the year, we expect to have launched a number of initiatives that **will use generative AI to eliminate awkwardness, make dating more rewarding and surprise and delight users**..."

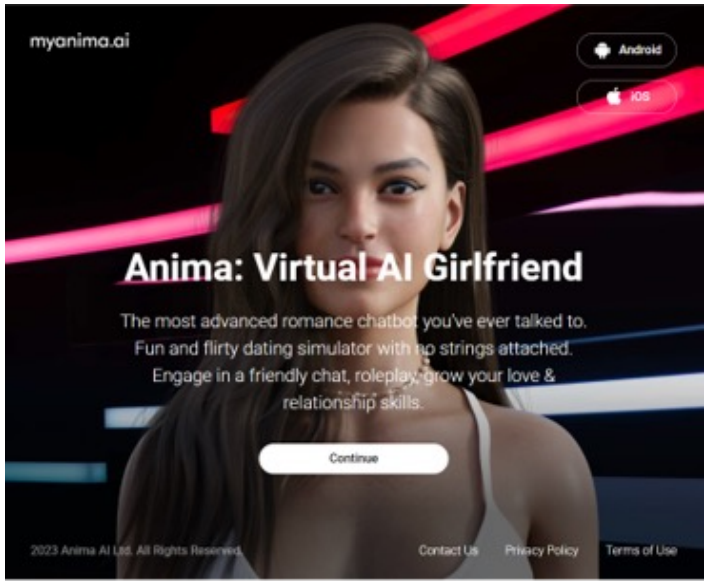
CEO Bernard Kim, Q3'23 earnings call

Source: CB Insights – Earnings transcript analytics (quarter reflects date call occurred)

AI girlfriend startups are finding a market in lonely, sexless men



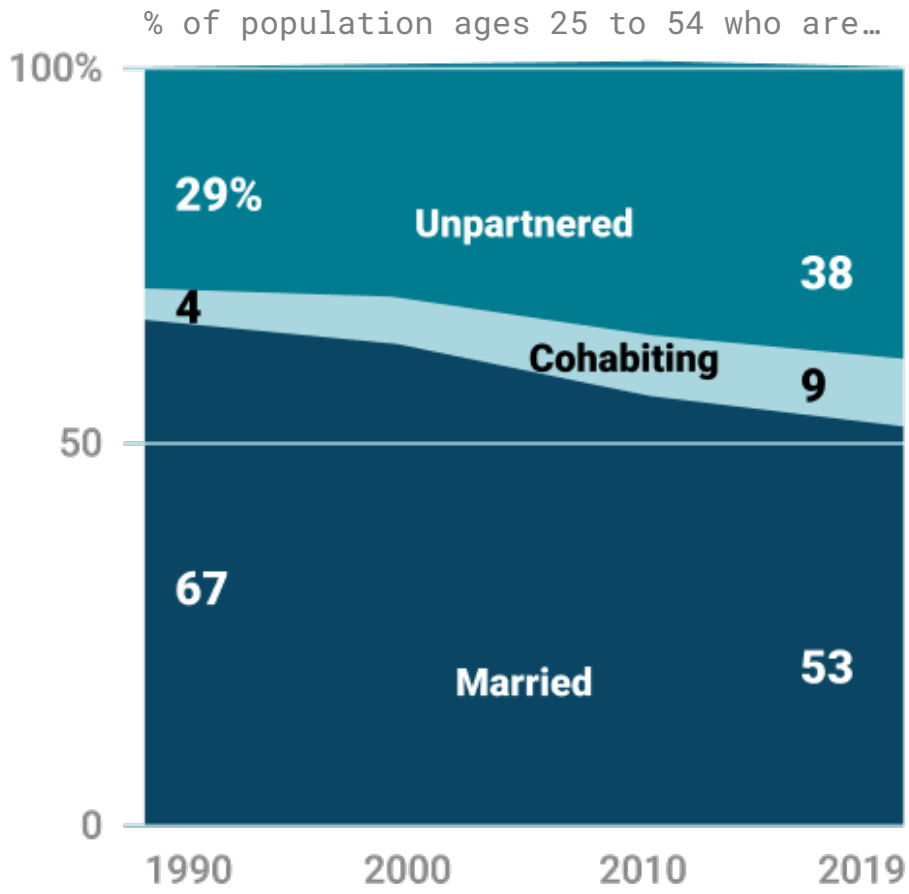




Source: Digi, Blush, Anima

Expect romantic AI to accelerate the decline in adult partnerships

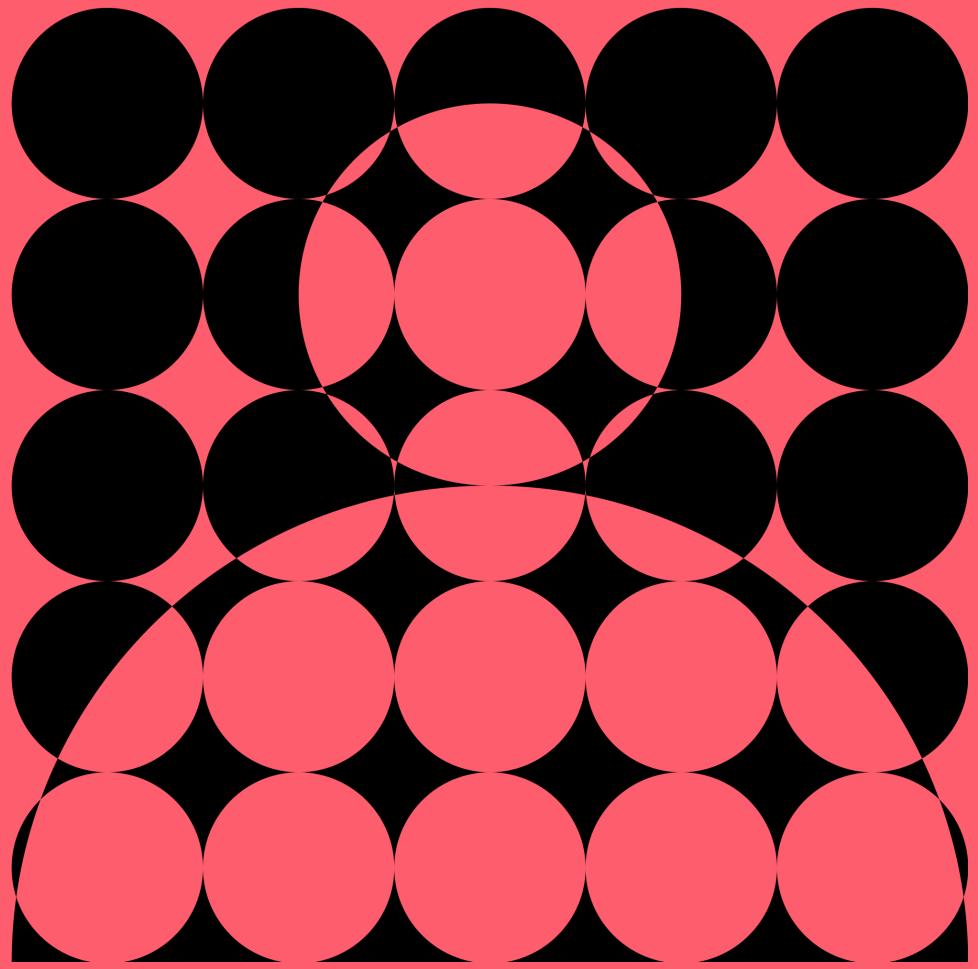
A rising share of US adults are living without a spouse or partner



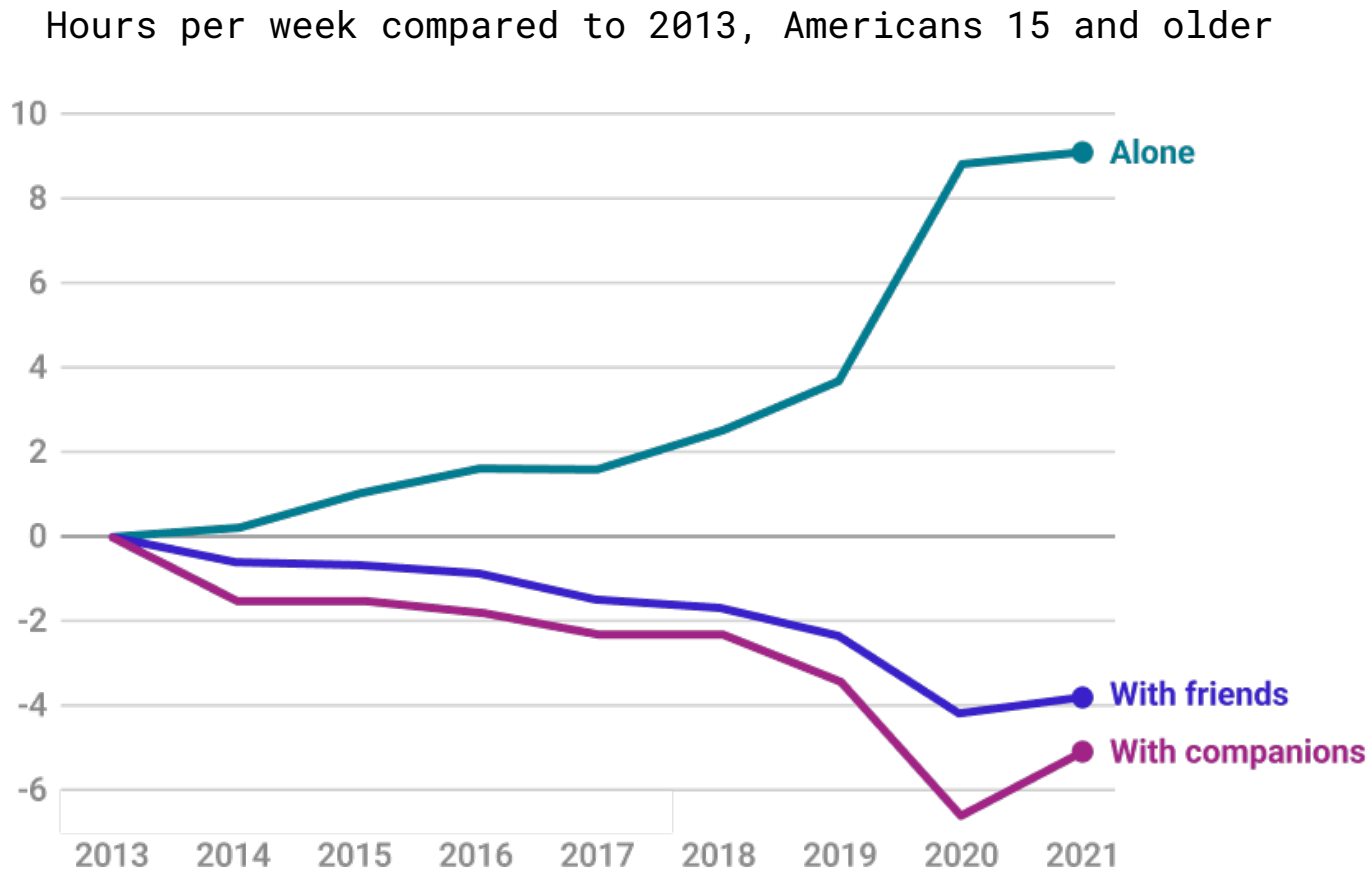
Source: Pew Research Center

Human-machine interactions

Lonely Gen Z befriends AI

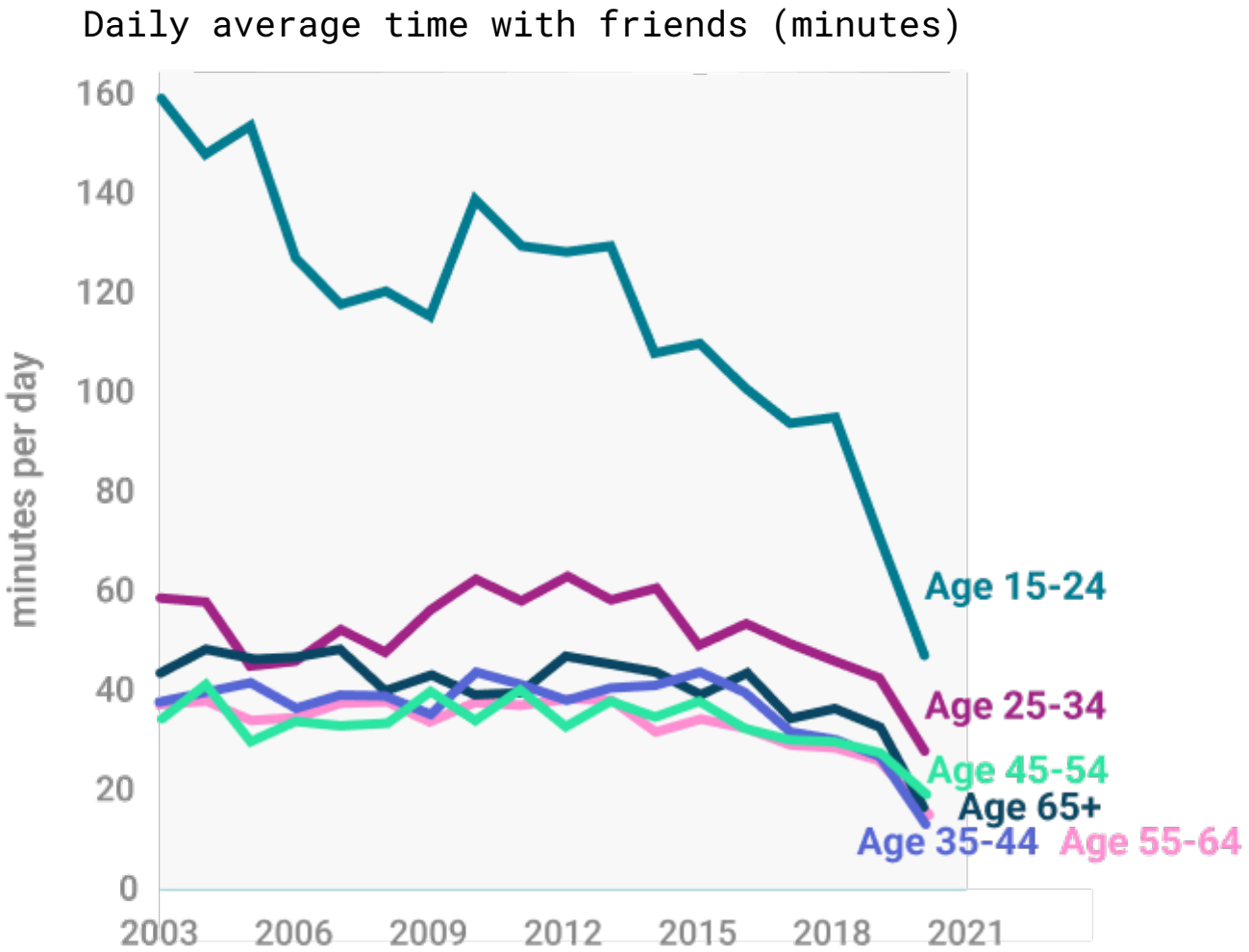


We are experiencing a loneliness epidemic



Source: American Time Use Survey via The Washington Post. Compares

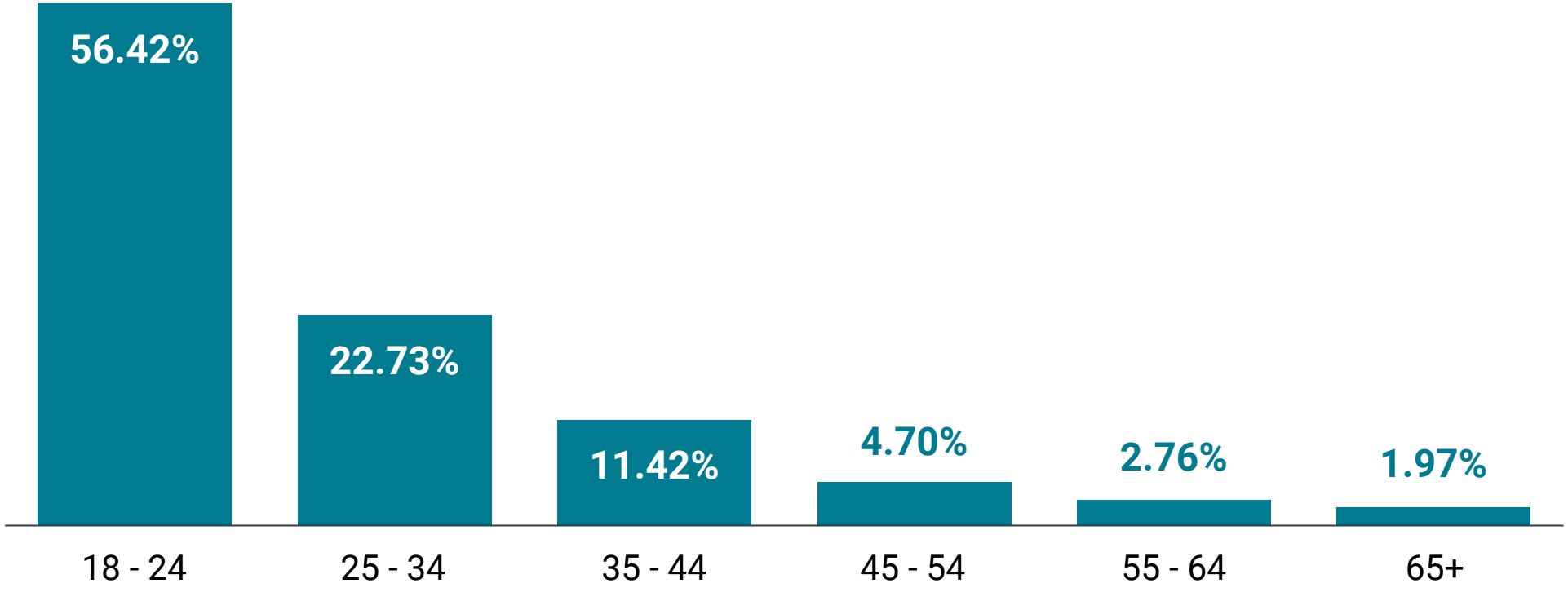
Gen Z spends far less time with friends in person in the smartphone era



Source: Kannan et al., "US trends in social isolation, social engagement, and companionship," SSM – Population Health

But who needs real friends when your AI bot can be just as personable as a human?

Over half of Character.AI's 4M users are under the age of 24



Source: SimilarWeb (data as of January 2024)

Tech for open-ended dialogues & AI character generation is evolving quickly

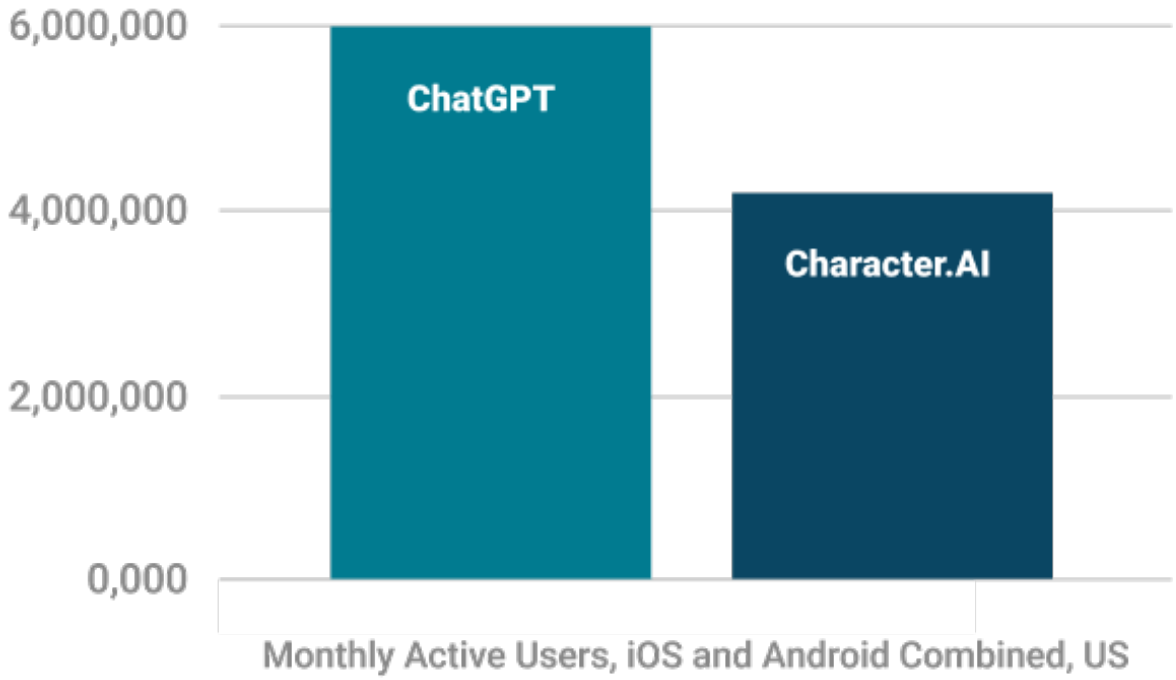
Advances in the gaming industry will lead to more immersive online, social experiences

Company	Funding	Latest funding date	Application
inworld	$117M	10/13/2023	**AI character generation** *Partnerships with Xbox & Unity Games*
convai	$4.8M	1/11/2023	**AI character generation**
latitude	$4M	4/1/2023	**Open-ended storylines**
Circle Labs	$4.3M	11/14/2022	**AI character generation & dialogues** *Seed round led by Lightspeed Venture Partners*

Source: CB Insights company profiles — Inworld AI, Convai, Latitude, Circle Labs

2024 will see an explosion of AI friendships

AI companion startup Character.AI is close behind ChatGPT in mobile usage



Emerging AI companion apps

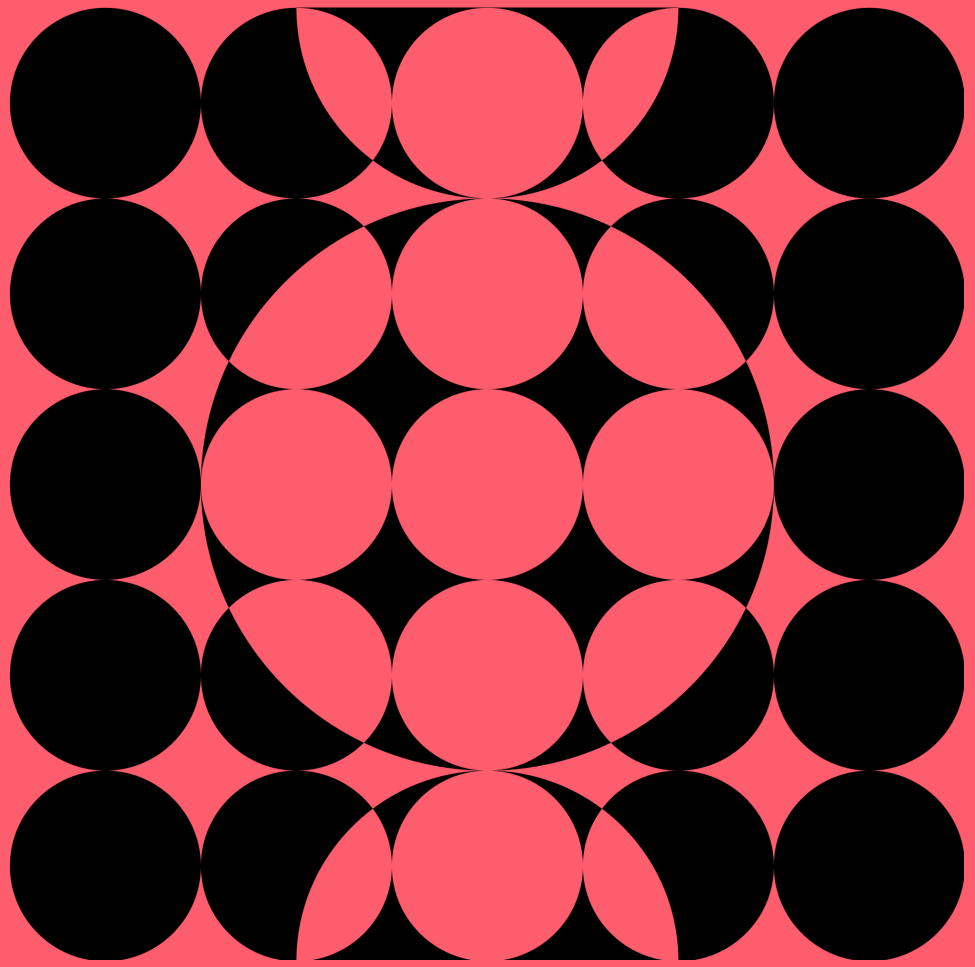
character.ai

CHAI

Source: SimilarWeb (data as of September 2023)

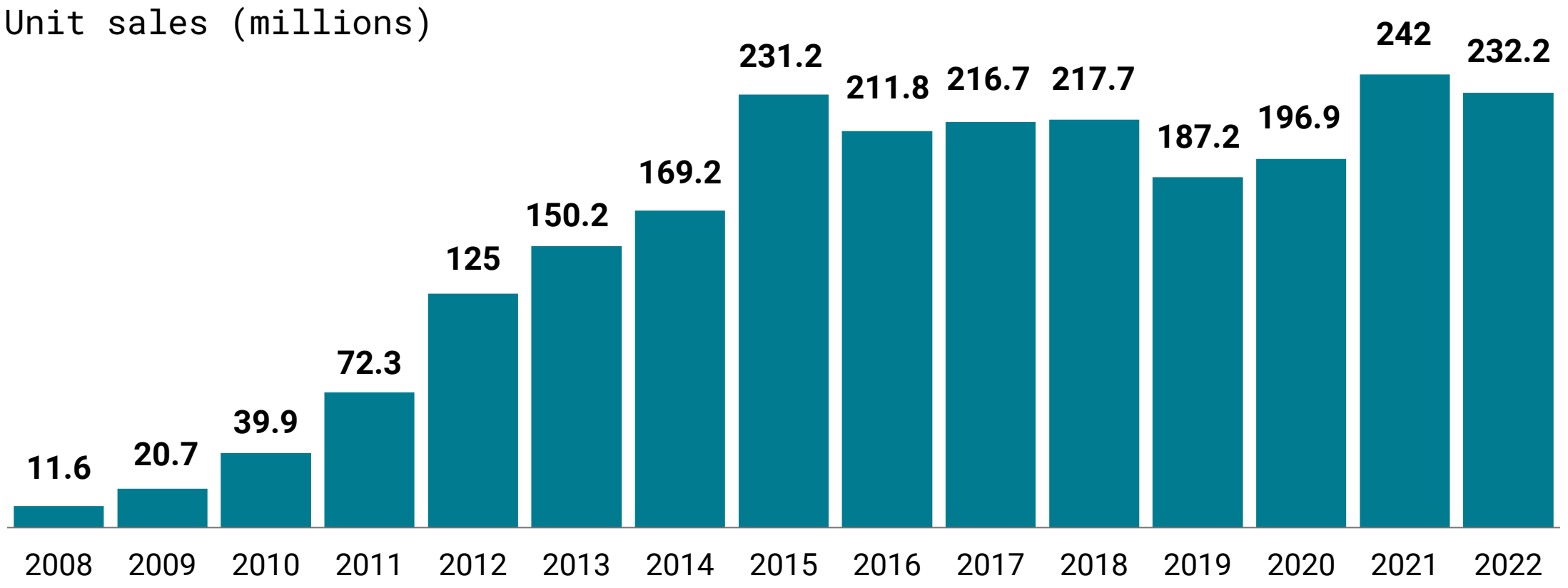
Human-machine interactions

The next iPhone:
AI's hardware revolution



Human-machine interactions

15 years after its release, iPhone sales are stagnating. Apple has been criticized for hitting an innovation plateau



Source: Apple via Business of Apps. Note: Apple stopped reporting iPhone sales in 2018. All values afterward are estimates.

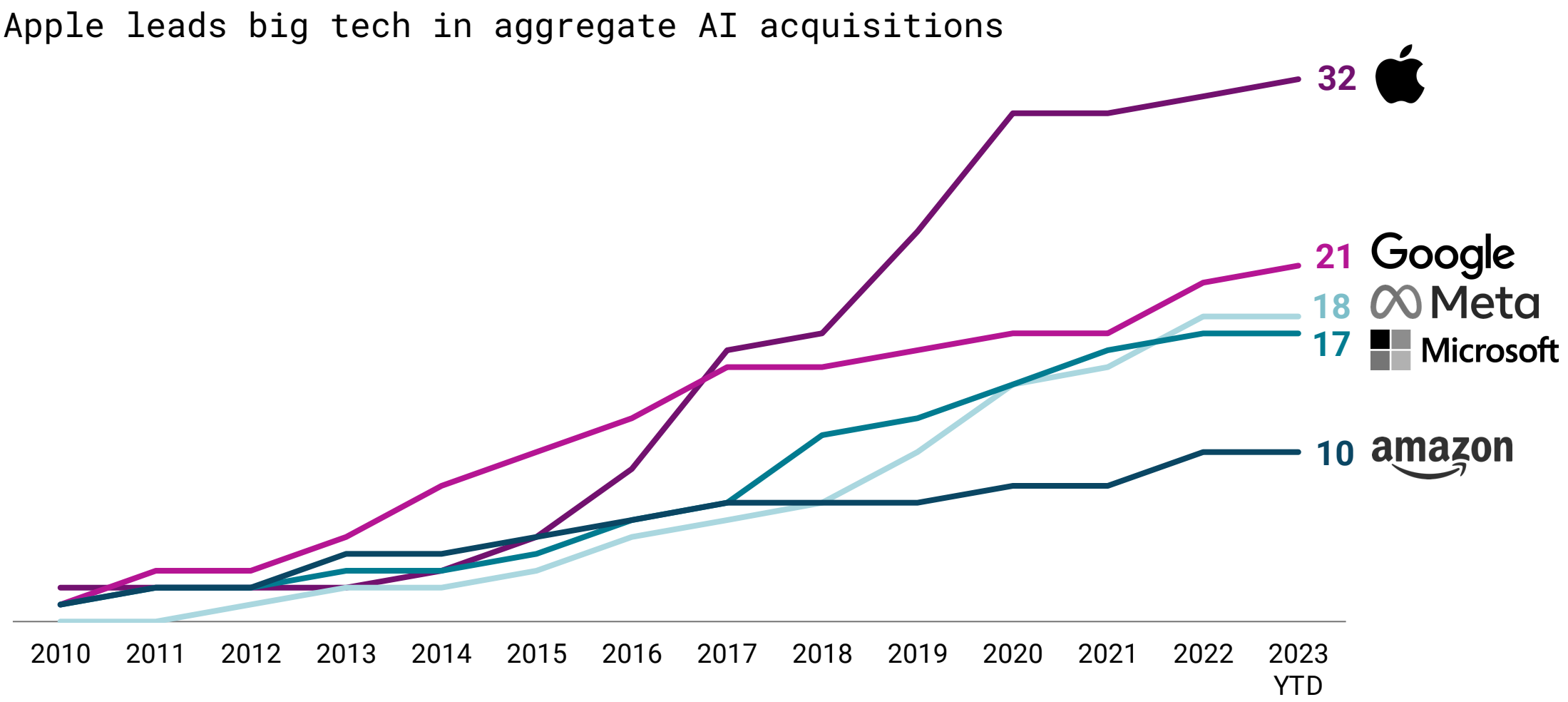
AI breakthroughs have launched the search for the next iPhone

Startups developing consumer AI devices

Company	Funding	Latest funding date	Product	User interaction
hu.ma.ne	$242M	3/8/2023	**Screenless wearable, powered by ChatGPT**	Voice, gesture
Rewind	$27M	5/9/2023	**Wearable "pendant" records & analyzes audio**	Voice, app
Tab	$2M	1/10/2024	**Wearable necklace records & analyzes audio with GPT-4**	Voice, app
rabbit™	$30M	12/21/2023	**Rabbit OS, LLM-based operating system; r1 device**	Voice, buttons, screen

Source: CB Insights company profiles — Humane, Rewind, Tab, Rabbit

Incumbents with deep AI talent & pockets have an advantage



Source: CB Insights research — The big tech company leading in AI acquisitions (data as of 9/19/2023)

Industry heavyweights are betting on AI-native hardware

OpenAI is in talks to build new AI hardware with LoveFrom, which has poached over 20 Apple employees[1]

Key people	Position	Previous positions
 Jony Ive	Founder	SVP of Industrial Design, Chief Design Officer
 Tang Tan	Hardware engineering lead	VP, iPhone & Apple Watch

Source: CB Insights key people data — Jony Ive, Tang Tan; [1]Engadget

Big tech and OpenAI will push forward on early experiments in 2024

Company	Device	Commercial development
Meta	AI-powered Ray-Ban **smart glasses**	
	Apple Vision Pro releasing February 2024 "**Apple GPT**" powers something new?	
OpenAI	Developing the "**iPhone of AI**"	



Source: Company reports

Future vision: LLMs evolve from chatbots to operating systems

OpenAI's Andrej Karpathy floats the idea of LLM-based OS



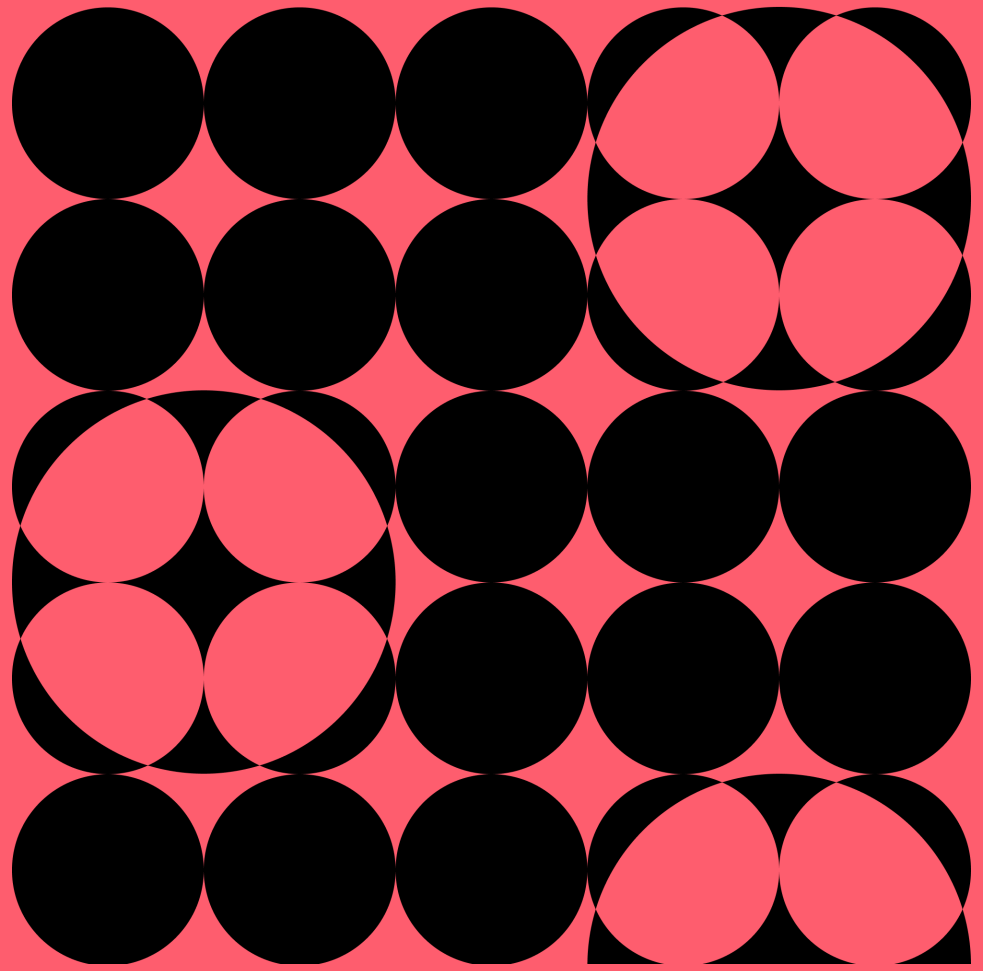
rabbit launches $199 device based on LLM OS



Source: Andrej Karpathy and rabbit via X

Human-machine interactions

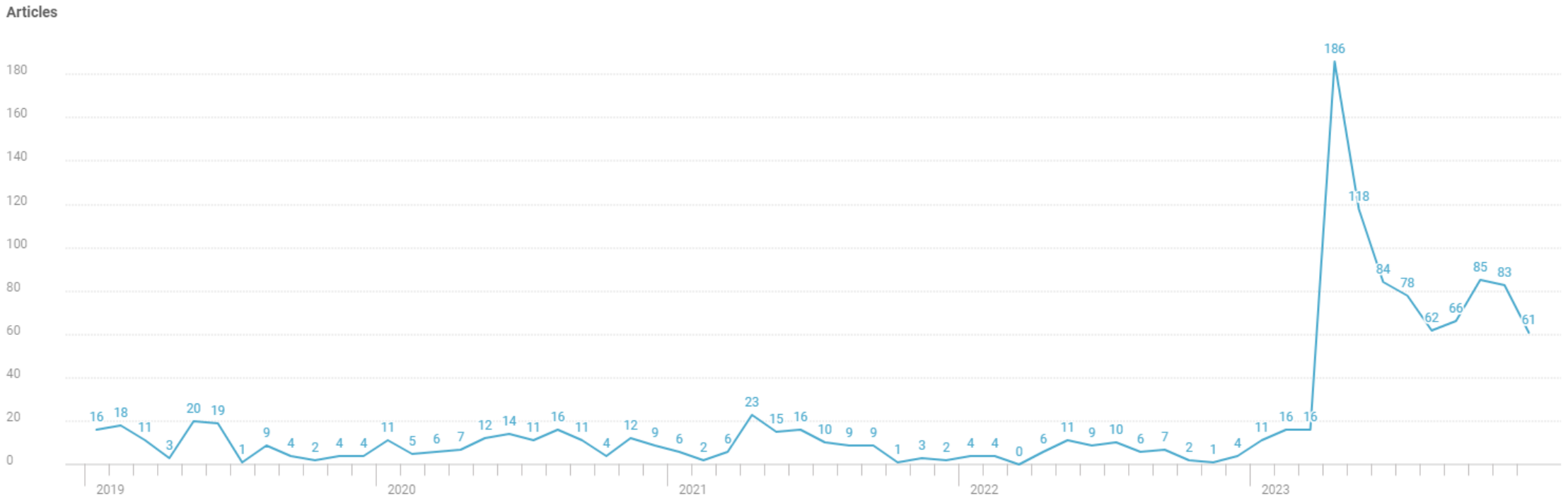
Autonomous agents:
From hype to somewhat useful



Human-machine interactions

The idea of Jarvis-like AI agents caught on like wildfire in ’23

Media mentions of popular autonomous agents – LLM-powered bots that can independently reason and execute tasks



Source: CB Insights — News trends tool (based on an analysis of 5K news sources)

Human-machine interactions

Autonomous agent projects proliferate — but are far from being commercially useful

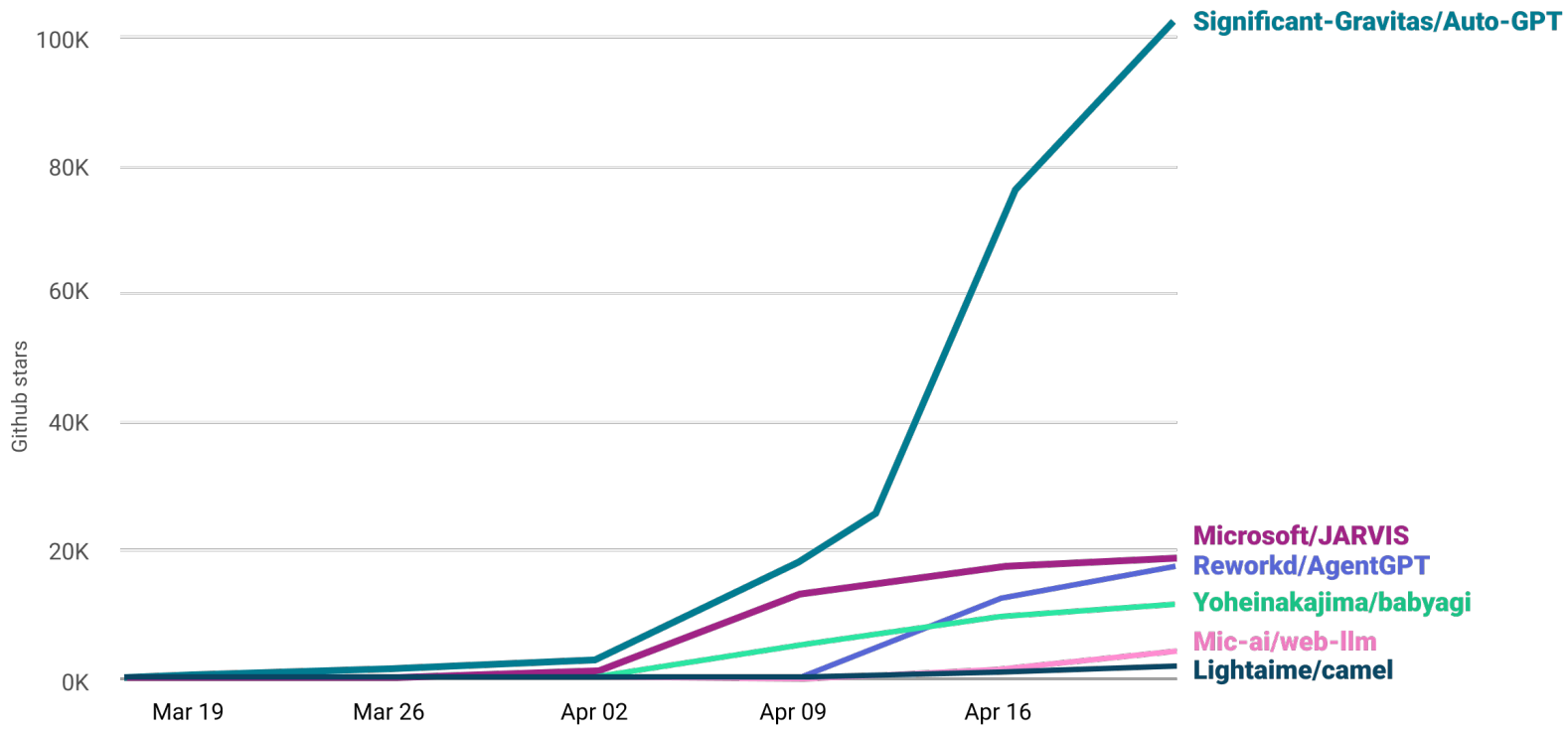
Agents are getting better at reasoning, but struggle with executing tasks

Project	Publish/ Release date	Developer	Agent description
Toolformers	Feb '23	**Meta**	Language models that "teach themselves to use external tools"
AutoGPT	Mar '23	**Significant Gravitas**	Open-source agent powered by GPT-4 to perform tasks
BabyAGI	Apr '23	**Yohei Nakajima**	Open-source agent built on GPT-4 and vector database Pinecone
Generative agents	Aug '23	**Stanford, Google**	A simulation "where end users can **interact with a small town of twenty-five agents** using natural language"
Voyager	Oct '23	**Nvidia, Caltech, and others**	"lifelong learning agent in Minecraft that continuously explores the world, acquires diverse skills, and makes novel discoveries without human intervention"
Mind2Web	Dec '23	**Ohio State University**	Dataset for developing autonomous agents that can perform tasks on websites

Source: Select agent research papers and tools published in 2023

The tech is advancing fast: Open-source projects get massive interest from developer community

AutoGPT gets 100K stars (indicating popularity) on GitHub within a month of launch



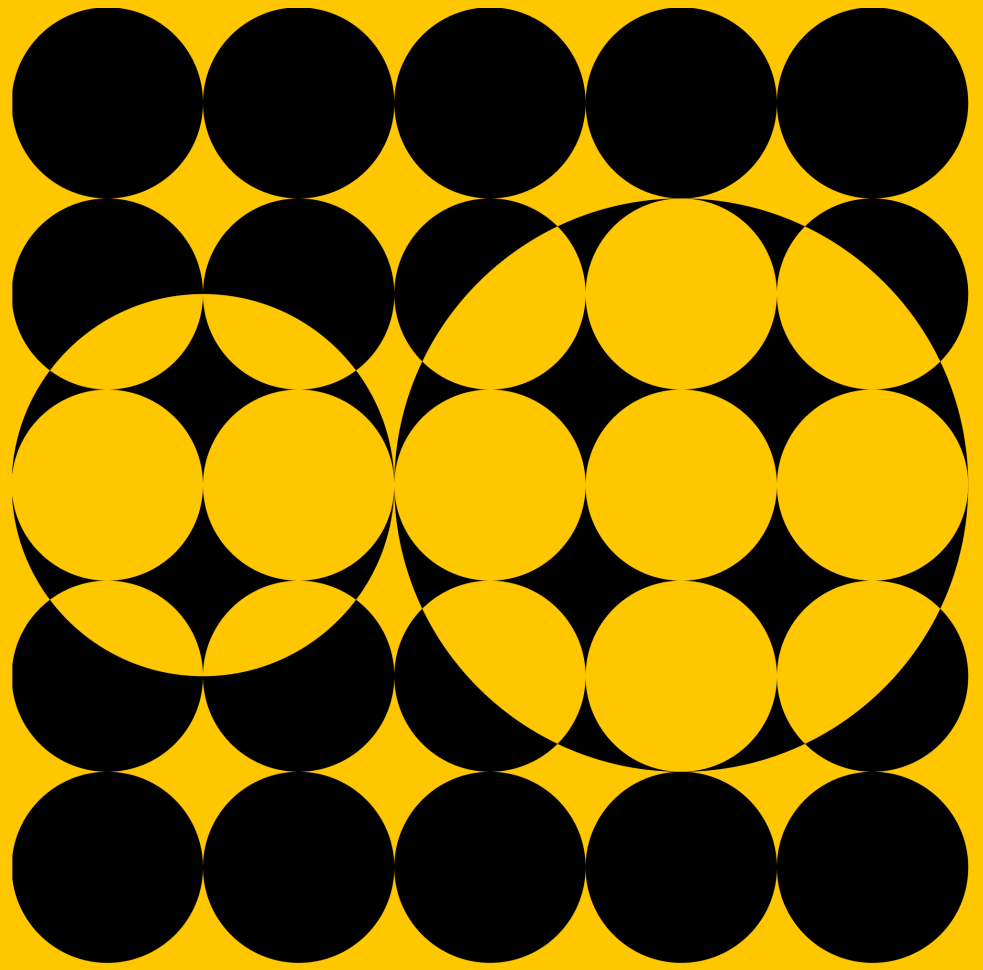
Source: Star-history.com

Big tech and VC interest will move the needle in 2024

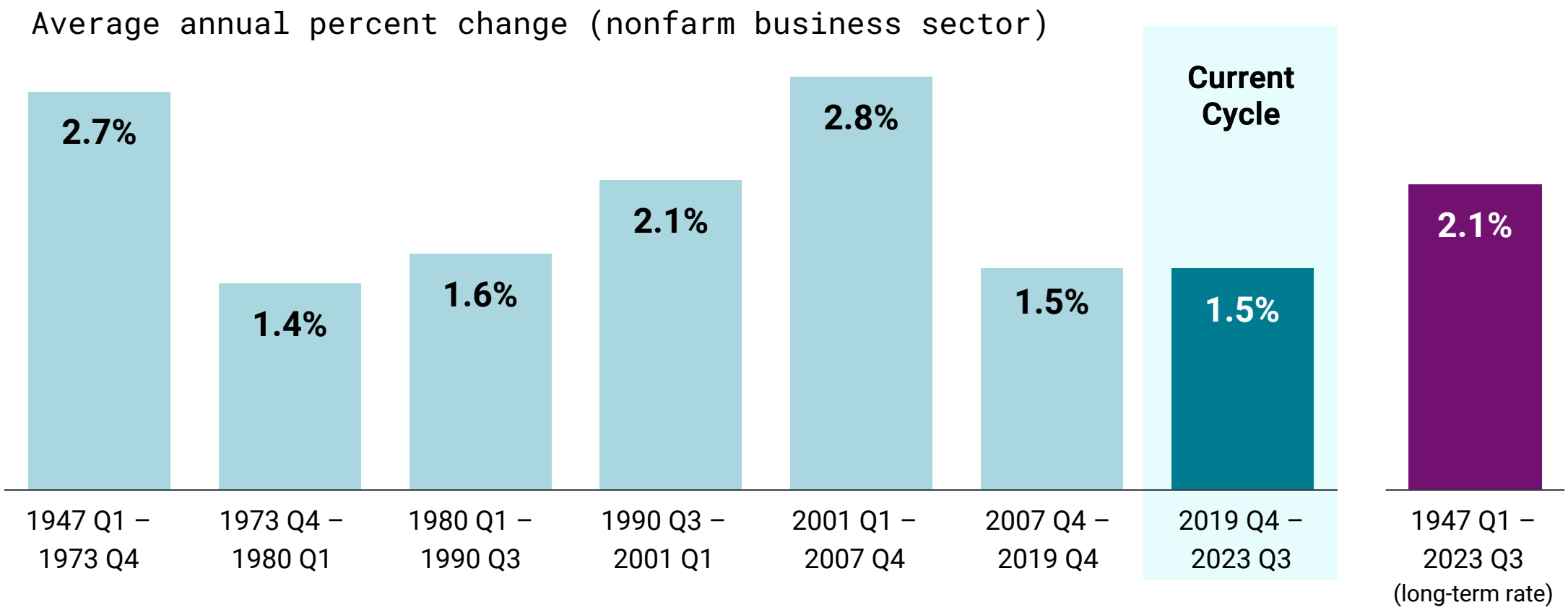
	Company	Autonomous agent initiatives	Total funding/ latest round	Investors
UNICORNS	OpenAI	• Rumored to be **training autonomous agents** • Dec '23: published research on best practices for governing agents • Announced **$10,000 – $100,000 grants for researching agent impact**	$14B/Secondary market	**Microsoft**, Andreessen Horowitz, Founders Fund, Khosla Ventures, Sequoia Capital
	imbue	Building AI agents and testing different architectures for development	$232M/Series B	Eric Schmidt, **The Alexa Fund**, Astera Institute, **NVIDIA**
	ADEPT	**Software-agnostic multimodal agent**	$415M/Series B	General Catalyst, Spark Capital, Greylock Partners, **Microsoft, NVIDIA**
BIG TECH	Microsoft	Published AutoGen research on **multiple AI agents talking to each other**	-	-
	DeepMind	Hired researcher Edward Grefenstette in June '23 to build **"general agents that can adapt to open-ended environments"** and communicate with humans	-	-
EMERGING STARTUPS	HyperWrite	Assistants for **everyday tasks**, from booking flights to research	$5.4M/Seed	Active Capital, Cortical Ventures, Madrona Venture Group
	MultiOn	AI agents to tackle **mundane tasks**	$4M/Seed	The MBA Fund

Source: CB Insights company profiles — OpenAI, Imbue, Adept, HyperWrite, MultiOn; published research and press releases

Copilots drive massive white-collar productivity gains



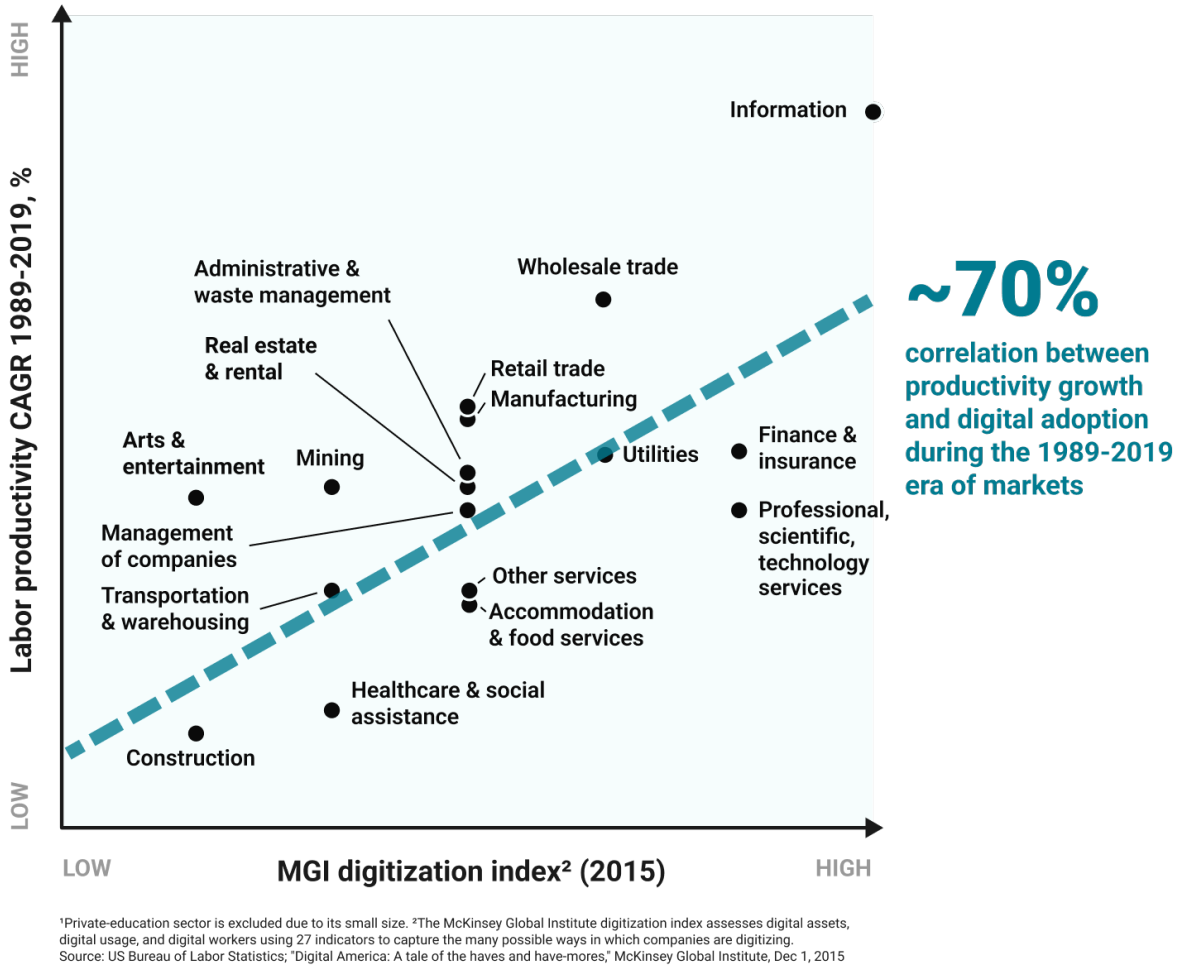
US labor productivity growth has stagnated in recent years



Source: US Bureau of Labor Statistics

Digital adoption is correlated with US labor productivity growth. AI will launch the next boom

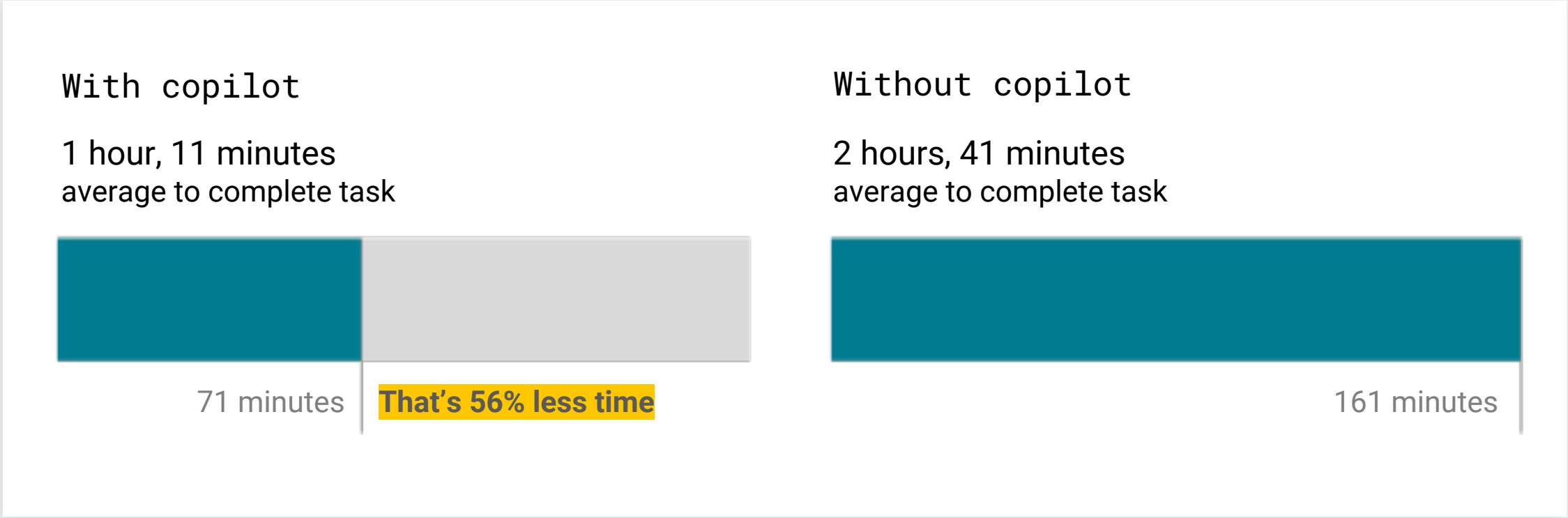
Sectors' labor productivity growth compared with digitization level[1]



[1]Private-education sector is excluded due to its small size. [2]The McKinsey Global Institute digitization index assesses digital assets, digital usage, and digital workers using 27 indicators to capture the many possible ways in which companies are digitizing.
Source: US Bureau of Labor Statistics; "Digital America: A tale of the haves and have-mores," McKinsey Global Institute, Dec 1, 2015

Source: McKinsey & Company

In software development, AI copilots are becoming essential to developer productivity



Source: GitHub. Based on a sample of 95 developers split randomly into 2 groups to complete a coding task.
Copilots = generative AI-powered tools designed to work alongside humans, automating tedious tasks

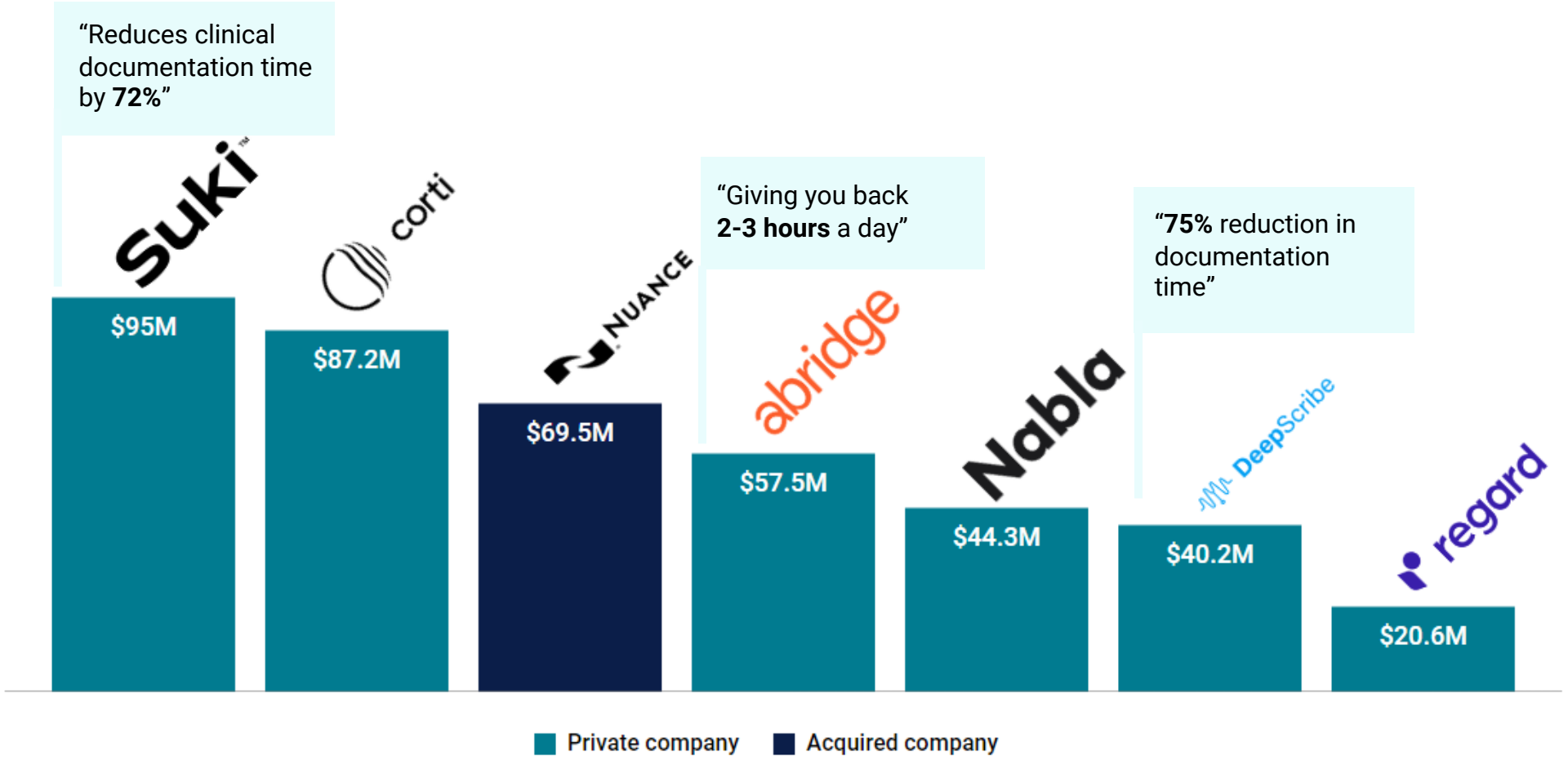
Coding copilots will become standard-issue tools at enterprises

Company/product	Status	Application
GitHub Copilot GitHub Copilot X	Released by Microsoft (10/29/2021)	Code completion/generator
ChatGPT	Released by OpenAI (11/30/2022)	Code generator
code llama	Released by Meta (8/24/2023)	Open-source code generator
StableCode by stability.ai	Released (8/8/2023)	Open-source code generator
starcoder	Released by Hugging Face & ServiceNow (5/8/2023)	Open-source code generator
aws CodeWhisperer	Released (4/13/2023)	Code completion
replit AI	Released (10/9/2023)	Code completion/generator
tabnine	Private (founded 2017)	Code completion
warp	Private (founded 2020)	Command line automation

Source: CB Insights company profiles; company reports

In healthcare, copilots are automating tedious tasks like documentation

Top companies by equity funding (as of 1/10/2024)



Source: CB Insights market report – Virtual medical scribes & summarization tools

Generative AI applications

In law, copilots are drafting contracts, summarizing documents, & optimizing research

casetext

Acquired by THOMSON REUTERS *for $650M in June 2023*

"It's the ability to draft a legal memo. It's the ability to use parallel search in neural searches to find legal research. It's the ability to go beyond using keywords and e-discovery and use concepts like, 'I don't feel comfortable with this.' That would never come up as a relevant document. **Absolutely transformational. I wasn't worried thinking, 'What's the ROI on this?' It was, 'How do we become an early adopter?' […] It has the ability to really improve the efficiency of legal services."**

Casetext customer
Chief Innovation Officer, National law firm

Source: CB Insights software buyer interviews — Casetext

Augmented workers will do more in the copilot age

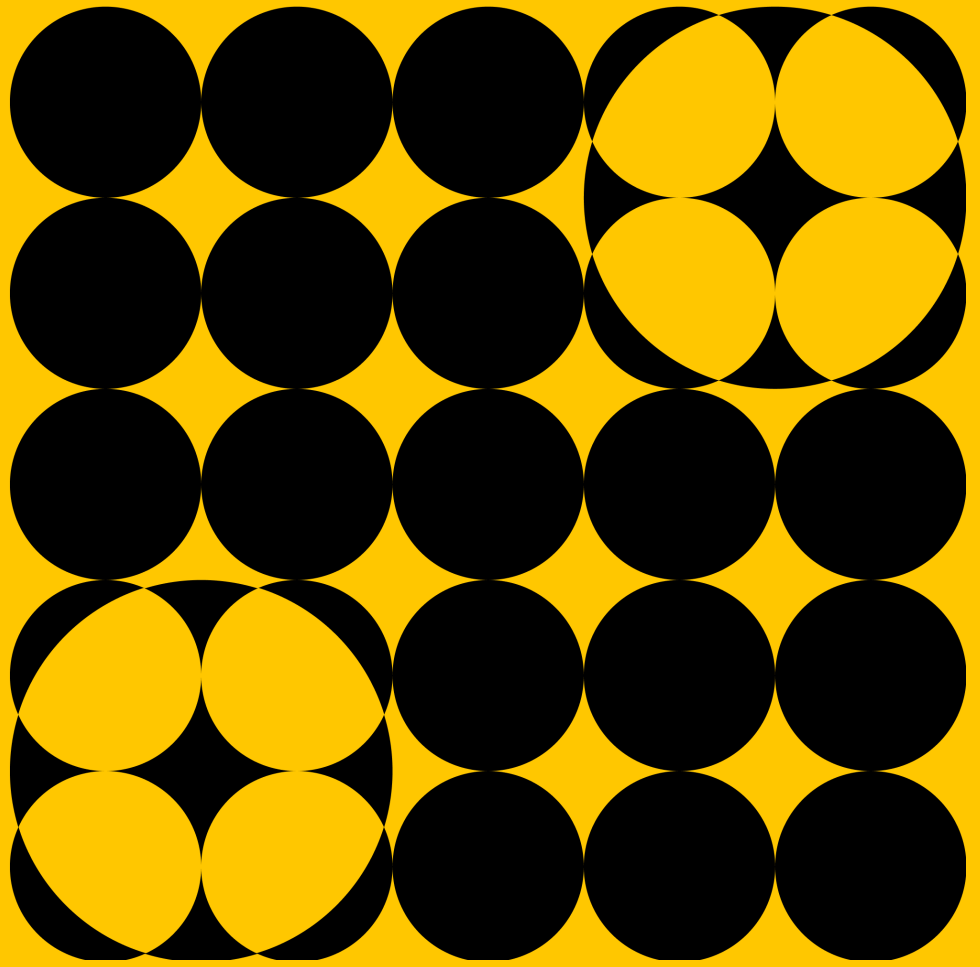


"[Uniphore] was implemented with 1,000 agents and the benefits were a 20% reduction in what we call average handle time…**Then there was an 80% reduction in after-call work time. Third, about 20% avoidance in hiring due to automation and basically an annual savings of $6 million through conversational AI and automation."**

Uniphore customer
Managing Director, Innovation at Fortune Global 500 company

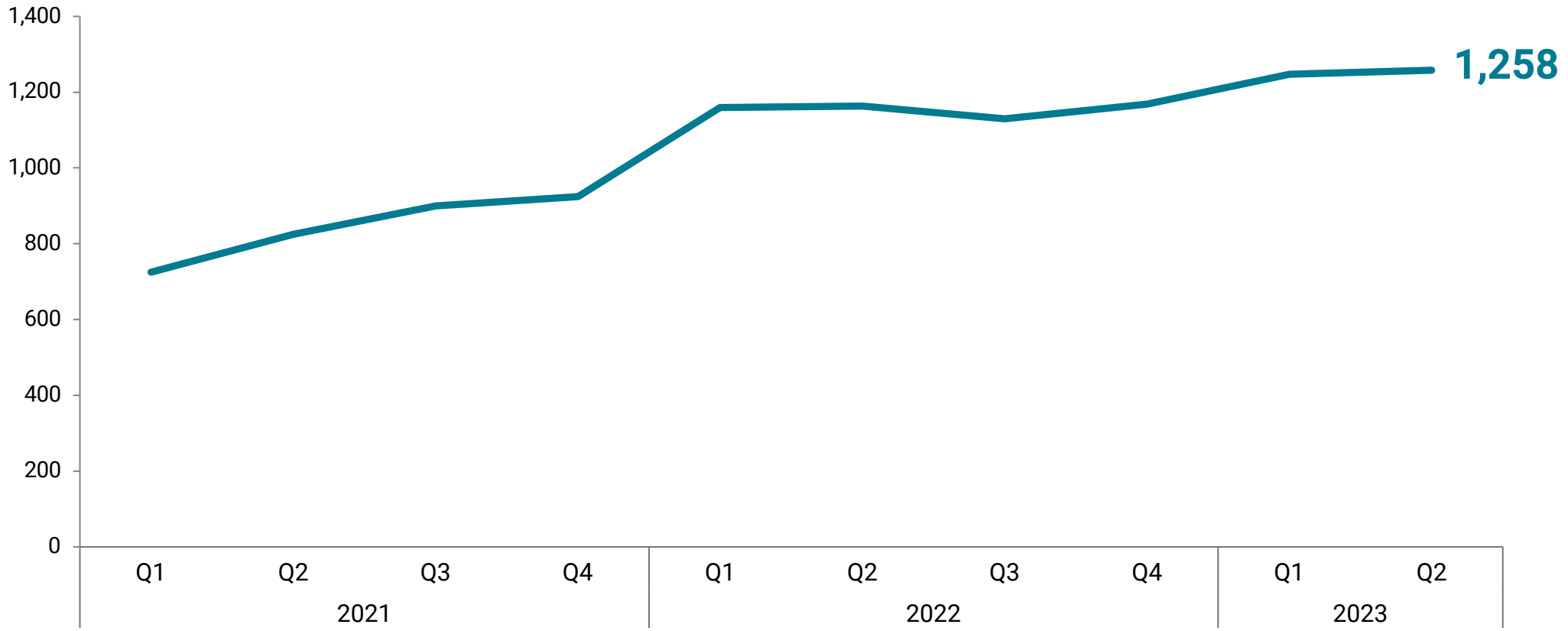
Source: CB Insights software buyer interviews — Uniphore

AI vs. AI:
GenAI fuels new wave of attack and defense



Cyber attacks continue to rise, with AI a driving force

Global average weekly cyber attacks per organization hit an all-time peak in Q2'23



Source: CheckPoint

Deepfakes are getting more sophisticated using genAI tools like face swapping, voice spoofing

31x increase in deepfake attempts

2023 vs. 2022

recorded by one identity verification platform

DARKTRACE

"**The pace of new threats and the tech behind them now outpaces any notion that a weeks' old view of historic cyberattacks can fully protect you as a business**, whether from the threats that we're seeing today, or the threats of tomorrow… whether that's self-evolving code designed to evade the best of those defenses, **it is automated deepfakes where you can't hear or trust what you're hearing or seeing**, or your employees' being tricked into being inadvertent insiders."

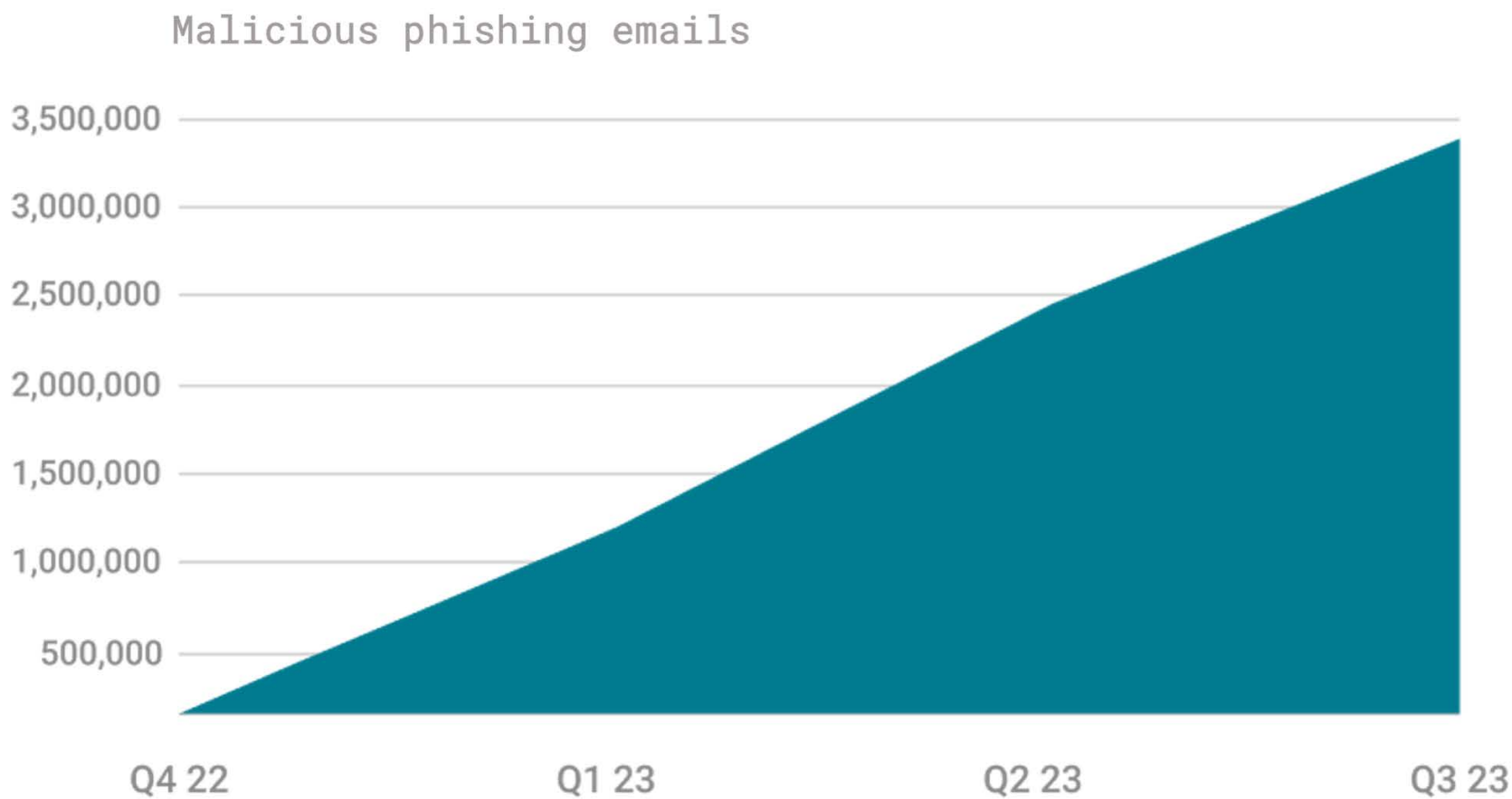
CEO Jack Stockdale, Q3'23 earnings call

Source: Onfido; CB Insights — Earnings transcripts analytics
Date reflects quarter call occurred

Generative AI applications

ChatGPT drives an explosion of email phishing attacks

One email security vendor recorded a 1,256% increase in phishing emails from Q4’22 to Q3’23



Source: SlashNext

Cybersecurity vendors use generative AI to counter attacks



"**SlashNext is working on a generative AI-based solution**, which means it will generate its own kind of phishing and malware and **it will train the software to automatically be aware of any new kind of threats arriving in the market**...**That is a differentiator from the technology perspective**."

SlashNext customer
Senior Design Engineer, Fortune 500 company

Source: CB Insights software buyer interviews — SlashNext

A new crop of genAI-focused cybersecurity vendors will emerge

Company	Founded	HQ	Total funding	Stage	Application
WRAITHWATCH	2023	US	$8M	Seed	AI attack simulation & prevention
Jericho Security	2023	US	$5M	Pre-seed	Breach & attack simulation
BRIGHTSIDE AI	2023	Switzerland	$1M	Seed	Breach & attack simulation
Nexusflow	2023	US	$11M	Seed	Security operations center copilots
shiboleth.ai	2023	US	$0.5M	Accelerator	Deepfake audio detection

Source: CB Insights company profiles — Wraithwatch, Jericho Security, Nexusflow, Brightside AI, Shiboleth

Covering AI threats will be a major selling point for buyers

GoSecure

"With artificial intelligence, we see that there are **tools popping up to help attackers leverage AI to generate extremely convincing phishing attacks**...**it's going to be really important that we address [this]**."

GoSecure customer
C-level executive, Academic institution

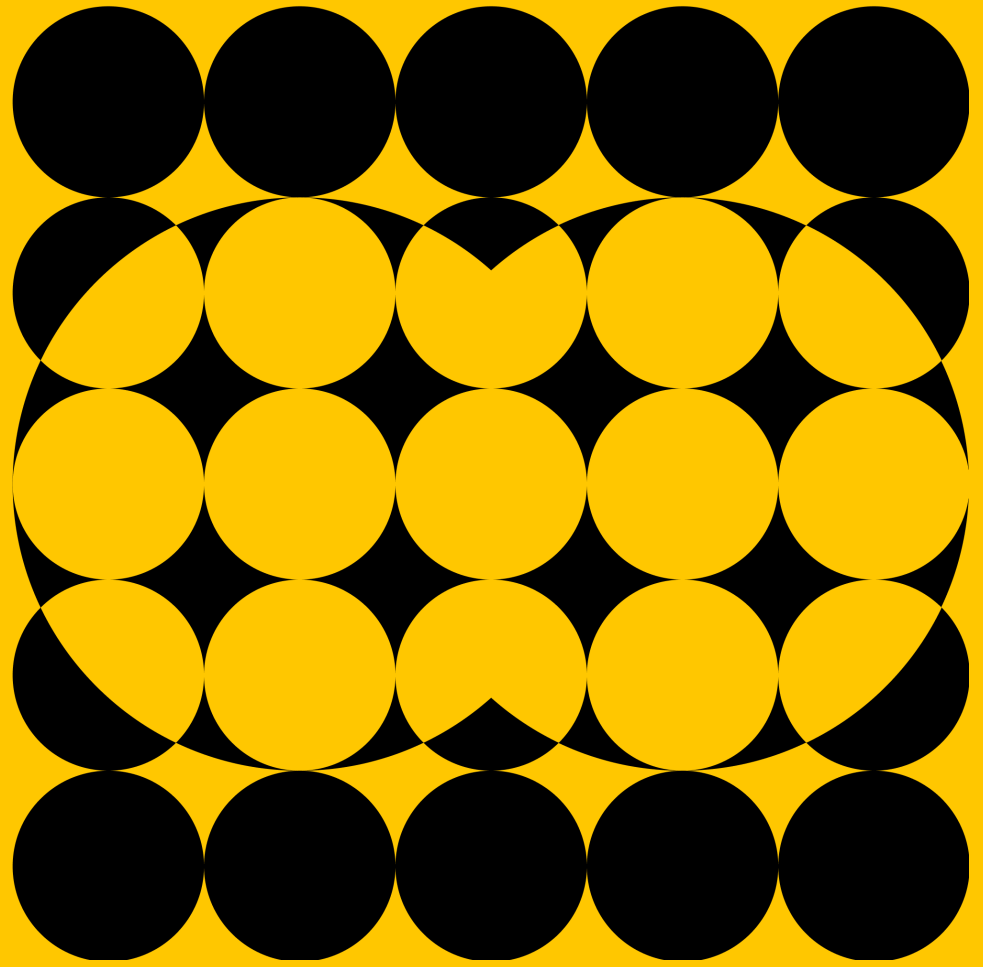
proofpoint.

"**Proofpoint's room for growth lies in its ability to detect emerging threats generated by artificial intelligence**. As generative AI and other types of threats become more sophisticated and widespread, they pose a greater risk to our business."

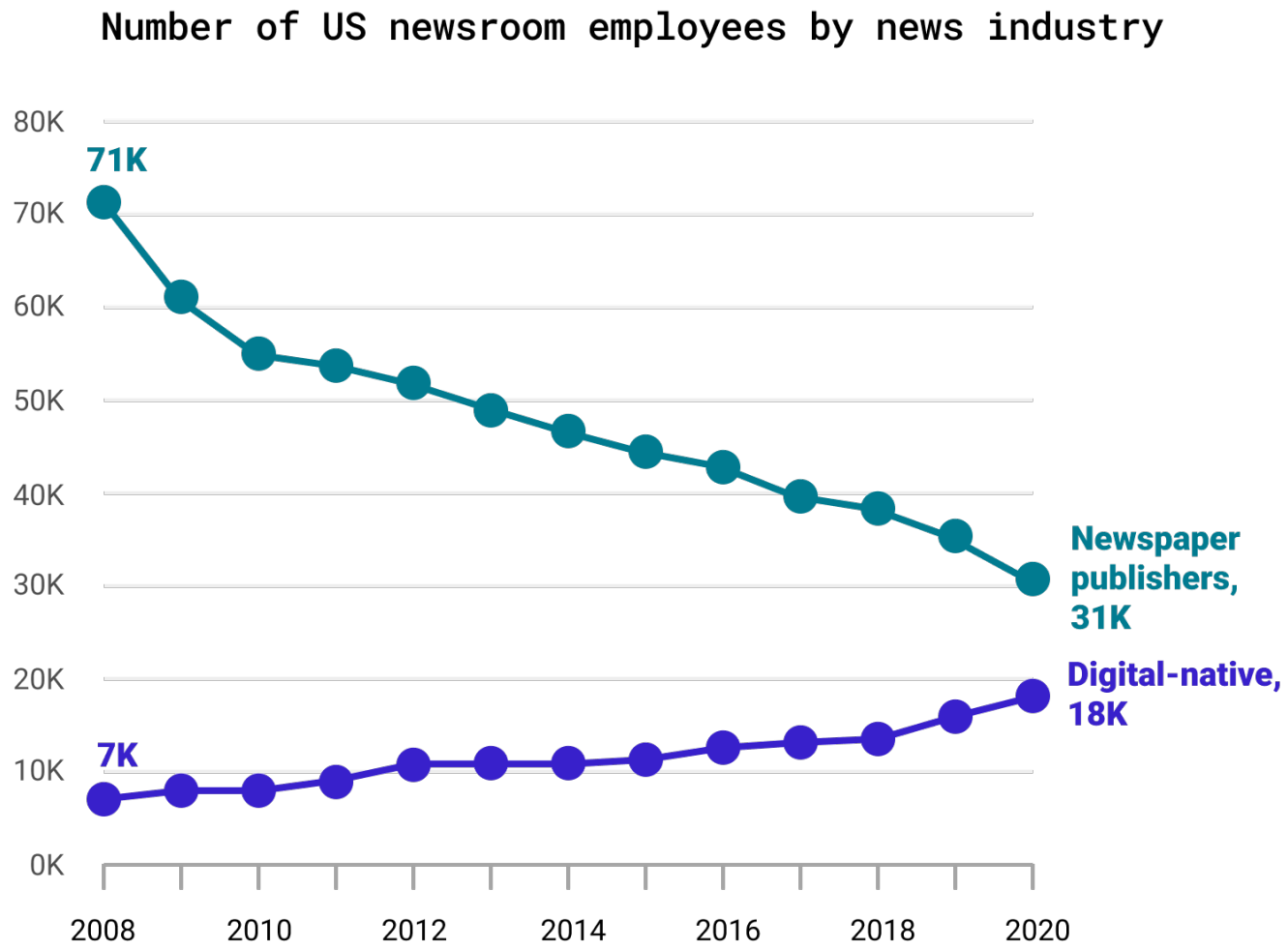
Proofpoint customer
Director, Publicly traded manufacturing company

Source: CB Insights software buyer interviews — GoSecure, Proofpoint

GenAI takes over journalism



US newsroom employment is in decline



Source: Pew Research Center

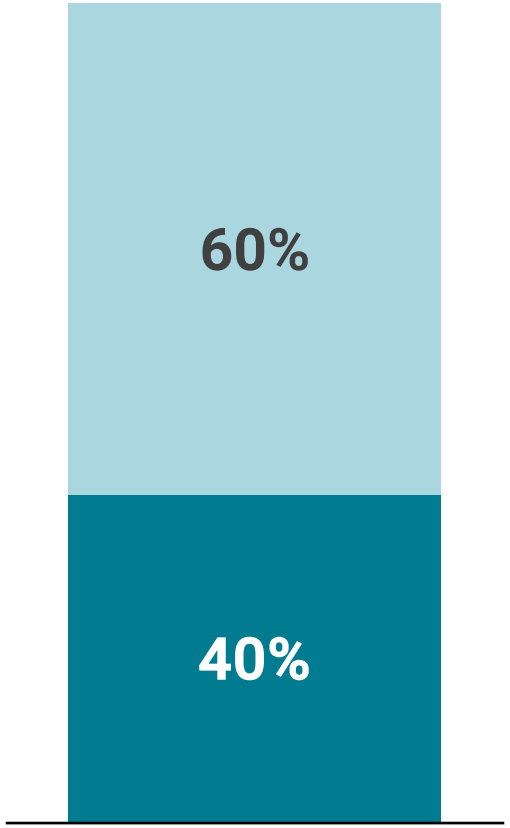
Generative AI applications

Media organizations are taking up AI to speed up news-gathering, production, and distribution

Organization	Product/partnership	Application
The New York Times	The Dynamic Meter	Machine learning model sets personalized meter limits for a **smarter paywall**
American Journalism Project	$5M partnership with OpenAI	**Leverage AI tools** to support & expand local news coverage
The Washington Post	Heliograf	First introduced in 2016, Heliograf **automates reporting** on structured data (e.g., sports scores)
AP Associated Press	SAM	**Social media monitoring** tool to stay on top of breaking news

Source: CB Insights business relationship data; media reports

But with genAI-powered search on the horizon, will it be enough?



Publisher traffic generated by Google

THE WALL STREET JOURNAL.

"A task force at the Atlantic modeled what could happen **if Google integrated AI into search. It found that 75% of the time, the AI-powered search would likely provide a full answer to a user's query and the Atlantic's site would miss out on traffic** it otherwise would have gotten."

Source: SimilarWeb, WSJ

Generative AI applications

News organizations will land new deals with LLM developers to license content as revenue models shift

In talks with OpenAI

- NBC NEWS
- CONDÉ NAST
- FOX
- News Corp (WSJ)
- CNN
- GANNETT
- IAC (People, The Daily Beast)

Deals with OpenAI

- axel springer (Politico, Business Insider)
  - Deal in the “10s of millions of dollars”
- AP Associated Press

In talks with

- CONDÉ NAST
- NBC NEWS
- IAC

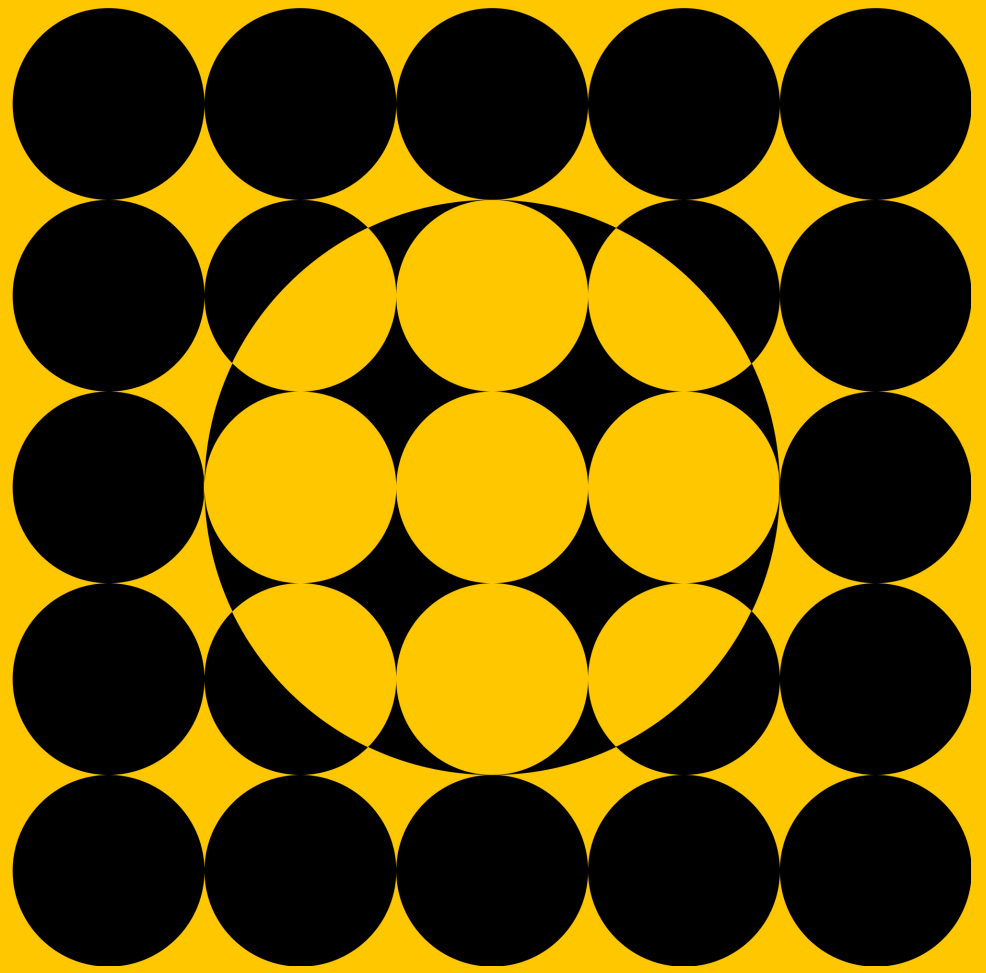
Source: CB Insights business relationship data; media reports

AI news anchors are here…reporters next

Company	Country	Description
Channel 1	US	Launching “generative AI news network” in fall 2024 with **AI-generated news anchors** and digitally generated voices
SBS NEWS	Korea	Debuted **AI news anchor** Zae-In designed by startup Pulse9
كويت NEWS	Kuwait	Debuted **AI-generated news presenter** Fedha
INDIA TODAY	India	Debuted **AI-powered anchor** Sana
astro AWANI	Malaysia	Debuted **AI-powered anchors** Joon and Monica
Google	US	Testing “Genesis” tool that can **produce news stories**
Sports Illustrated	US	Received flak for publishing articles by **fake, AI-generated writers**

Source: NYT, Channel 1, Futurism, The Guardian

Beyond stills: Video generation makes strides



2022 kicked off the text-to-image craze. But text-to-video is a bigger challenge

Meta

"**It's much harder to generate video than photos** because beyond correctly generating each pixel, the system also has to predict how they'll change over time."

CEO Mark Zuckerberg, Q3'22

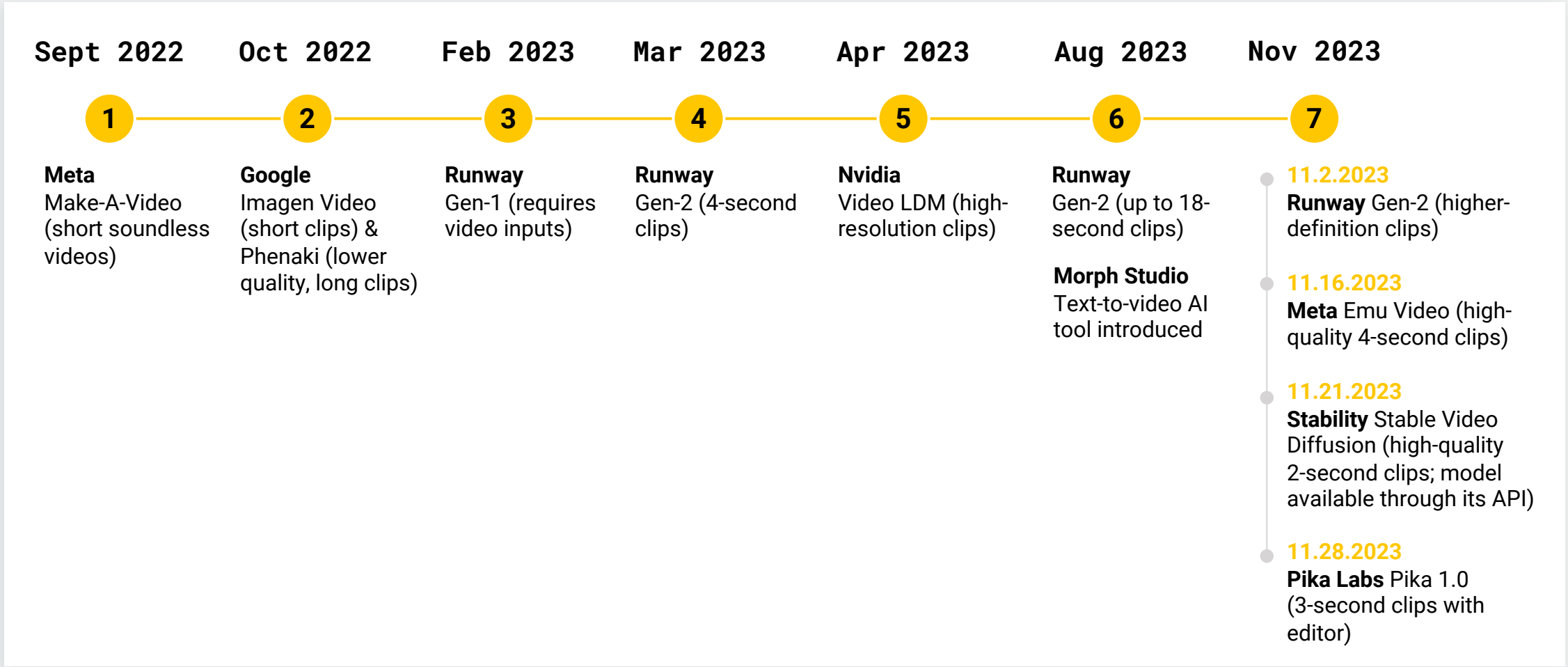
Hugging Face

"Text-to-video are typically trained on very short clips, meaning they require a computationally expensive and slow sliding window approach to generate long videos. **As a result, these models are notoriously difficult to deploy and scale and remain limited in context and length.**"

Former ML Engineer Alara Darick, Q2'23

Source: Facebook, Hugging Face blog

Now generative video models are advancing rapidly



Source: Company press releases

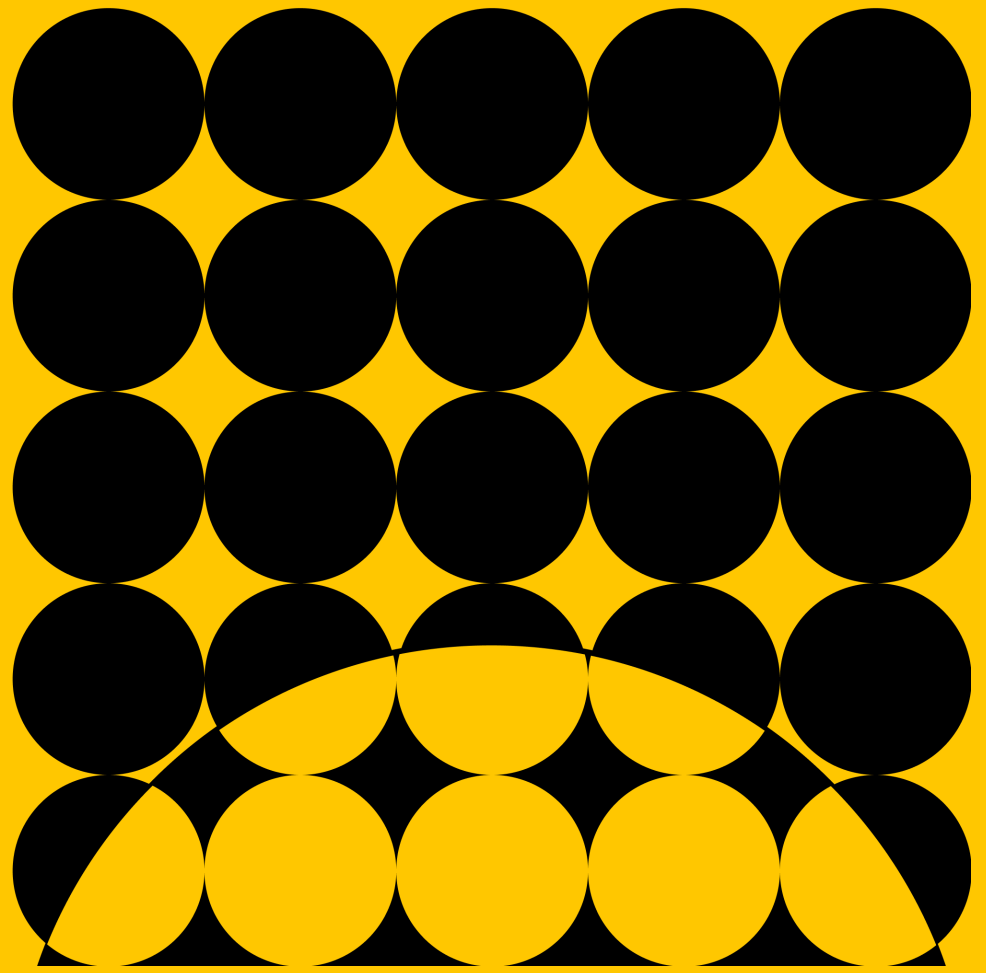
Watch for immediate impact across advertising, Hollywood, and other visual media industries

Startups race to build video generation & editing platforms for enterprise and creative use cases

Company	Total funding	Stage	Select investors	Notable developments
runway	$237M	Series C	Google, Nvidia, Salesforce Ventures, Coatue	Runway partners with Getty Images to create genAI video model aimed at **Hollywood studios and advertising, media, & broadcasting companies** (December 2023)
Pika	$55M	Series A	Lightspeed Venture Partners, Adam D'Angelo, Andrej Karpathy	Reported **500,000 users** "generating millions of videos per week" (December 2023)
stability.ai	$151M	Seed	Lightspeed Venture Partners, Coatue, Intel	Text-to-video model **available to developers** via Stability AI API (December 2023)
Assistive	-	-	-	Generative video platform **available via web & API** (launched December 2023)
ByteDance	$7.4B	Series D	Tiger Global, KKR, Tencent	Released MagicVideo-V2, a "**multi-stage end-to-end video generation pipeline**" (January 2024)
Moonvalley	$30M	-	Khosla Ventures, Y Combinator	Generates "cinematic film quality footage"; **gained 100,000 users in 6 weeks** on Discord (November 2023)

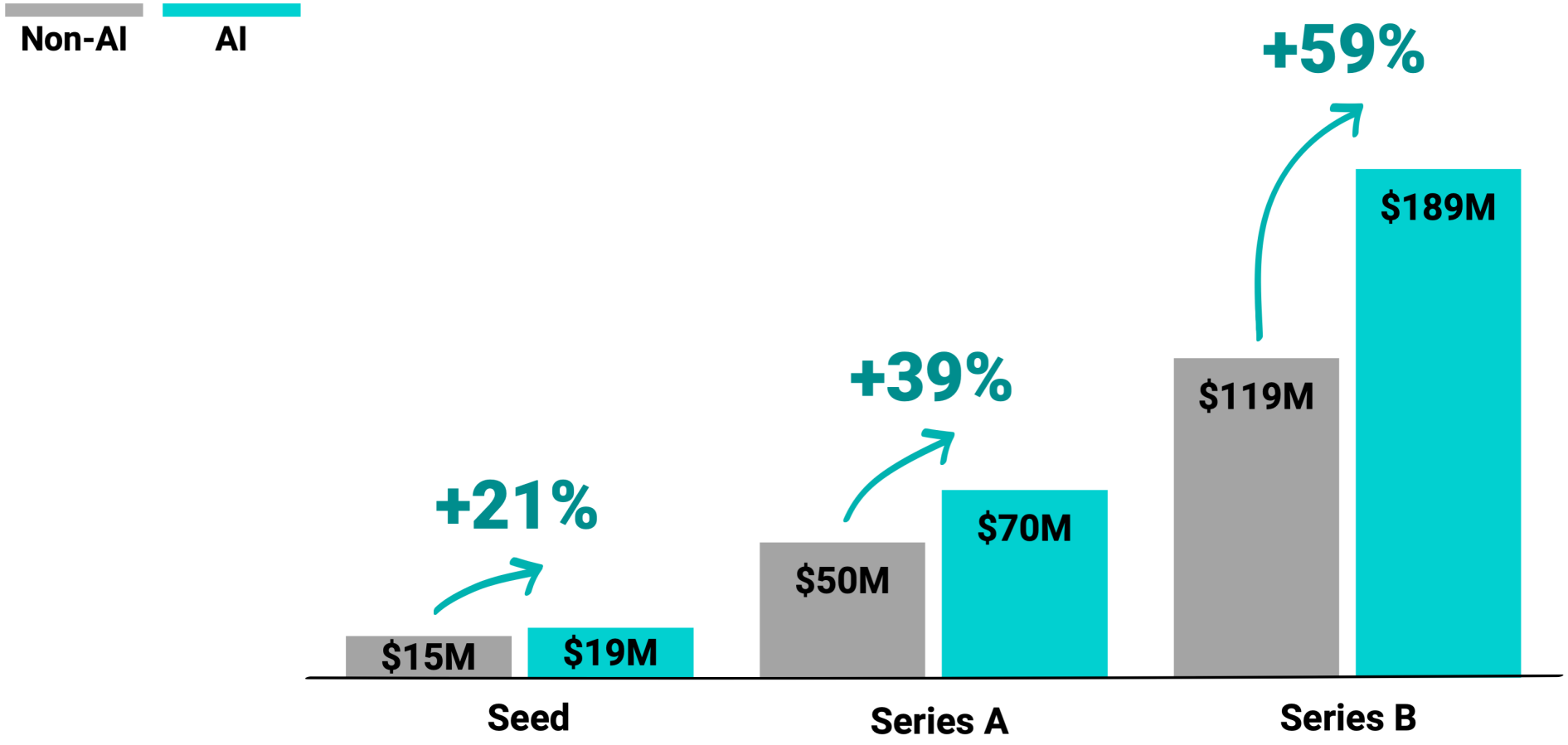
Source: CB Insights company profiles – Runway, Pika, Stability, ByteDance, Moonvalley

A genAI graveyard follows frenzy



GenAI frenzy launched AI valuations into the stratosphere in 2023

Median valuations for VC-backed deals in 2023



Source: CB Insights valuation data — The AI premium

As the smoke clears, growing competition threatens some early genAI companies

Text generation startup Jasper announced layoffs & cut internal valuation in Q3’23

Jasper

“So in this newly emerging world of generative AI it's hard to keep up with all the changes that are going on. It's probably not fair to ask this, but I'll say it: Jasper needs to stay up to date faster to make me a definite yes to renew…I'm seeing other lower cost options; **the price of calls to GPT-3, for example, has gone down to minimal numbers. It might be harder to say yes to a renewal when we're due next year**, so I'll have to really see that our people have picked this up and are finding great value.”

Jasper customer
C-level executive at $10M+ funded research platform

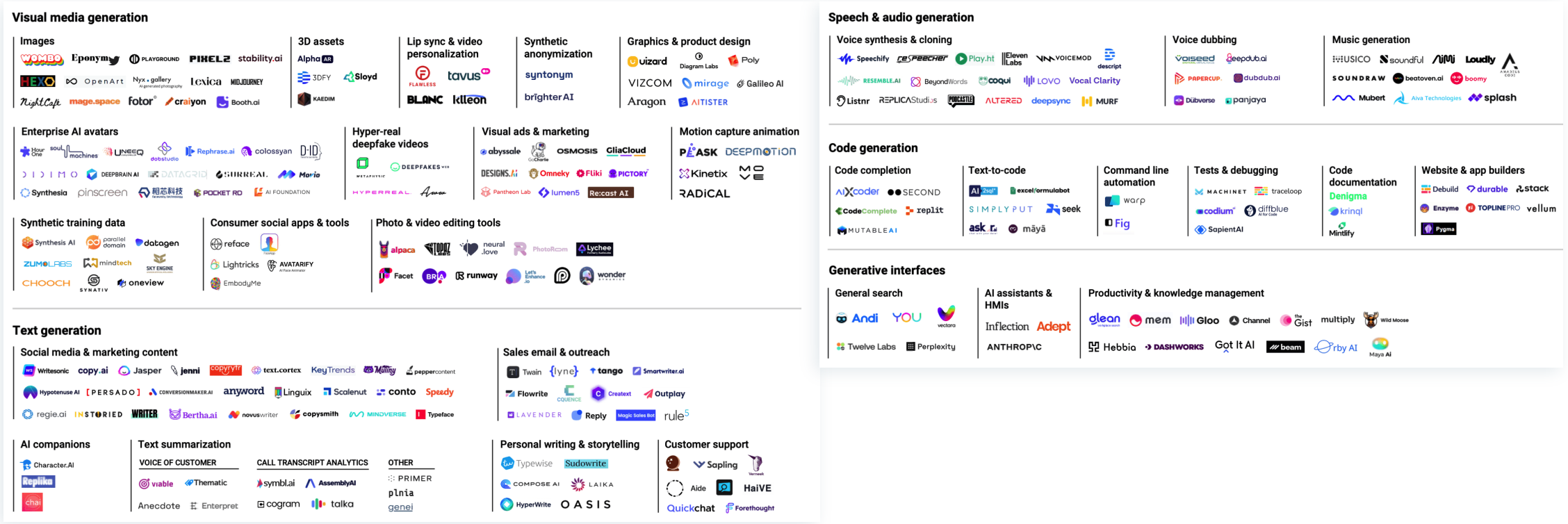
Source: CB Insights software buyer transcript — Jasper

Generative AI applications

Startups in horizontal markets without a moat will bite the dust

Text generation and visual media generation are especially vulnerable to pressure from large players

Horizontal generative applications

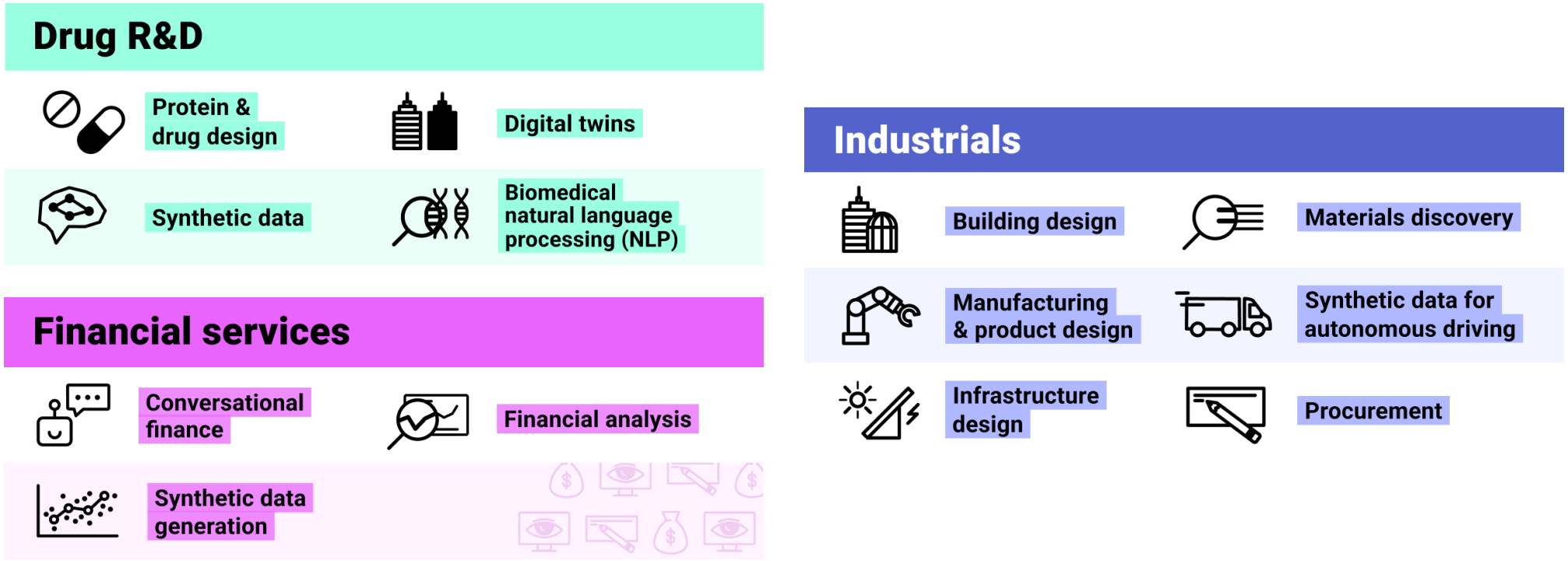


Source: CB Insights research — Generative AI Market Map

Generative AI applications

In 2024, attention will be on companies building stickier solutions in specialized areas

GenAI opportunities in verticals



Source: CB Insights research — Industrials, Drug R&D, Financial Services

Our most popular generative AI research

Research briefs

[Generative AI bible: The ultimate guide to genAI disruption]

[The large language model operations market map]

[The open-source AI development market map]

[The generative AI market map]

[The state of generative AI in 7 charts]

Market reports

[Large language model developers]

[LLM application development]

[Enterprise avatars]

[Text content generation]

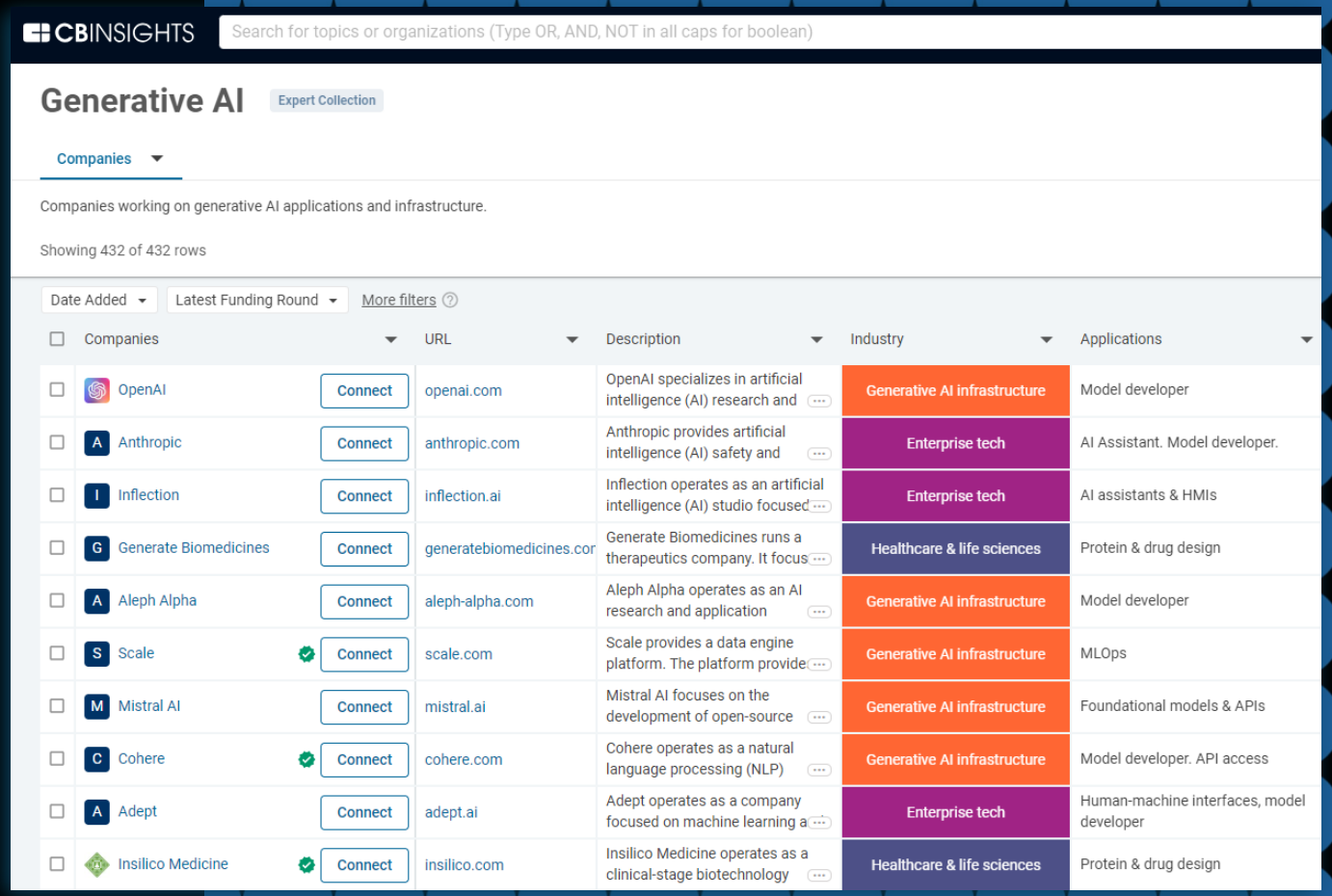


Generative AI Expert Collection

Stay up-to-date on generative AI with the **analyst-curated Expert Collection**

Filter by technology, funding, valuation, business model, and more

Filter by potential vendors, partners, and targets with real-time data



CB Insights helps the world's leading companies understand everything they need to know about disruptive technologies — find out more about why our customers love us here.

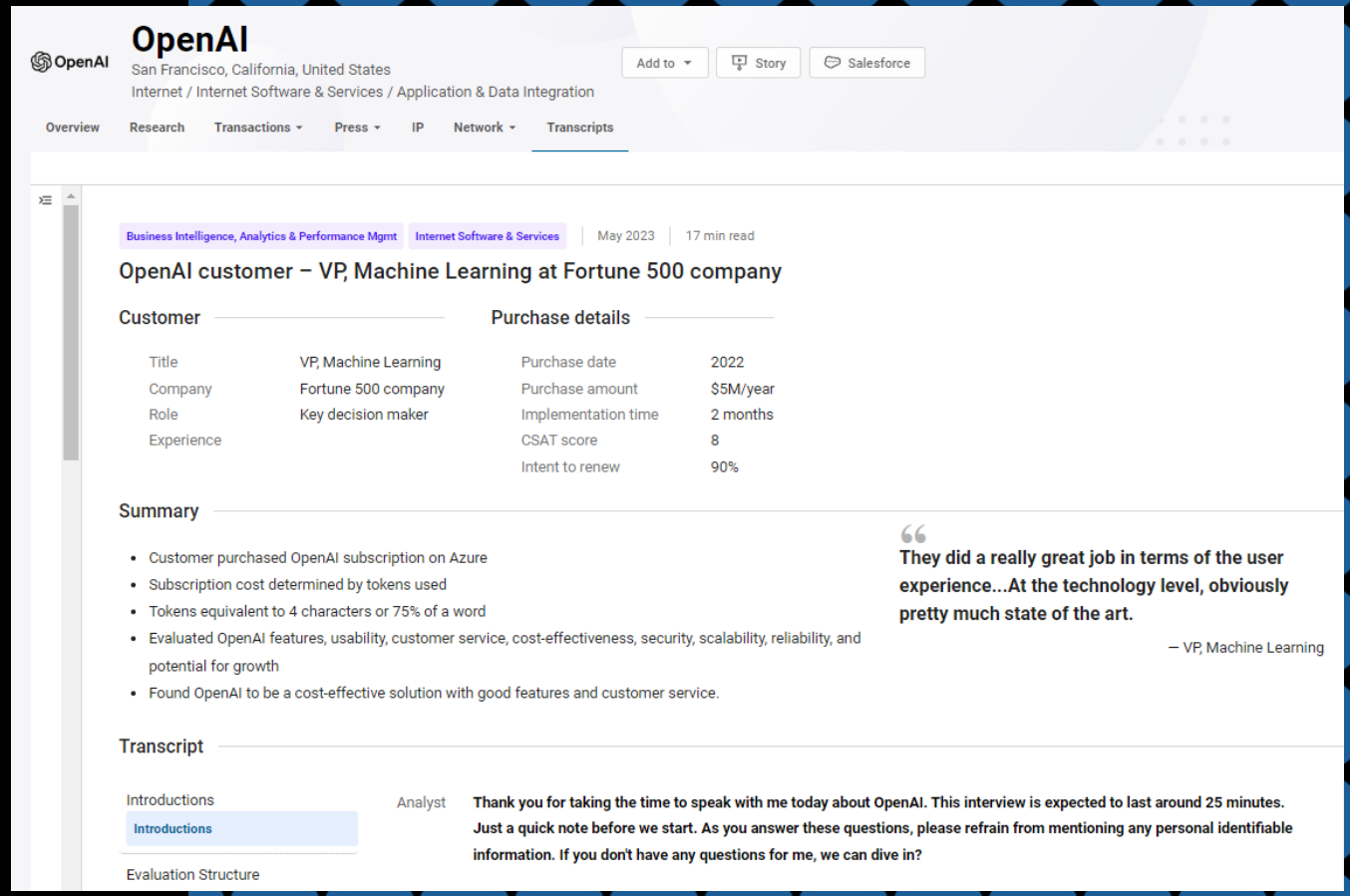
Buyer Transcripts

Buyer profile
Understand the buyer — their title, their organization, and their experience

Purchase intelligence
How much did they spend and when

Consistent structure
Moderated & structured conversations make transcripts easy to scan

Deep insights
Details on everything from evaluation criteria to competitors to satisfaction



CB Insights helps the world's leading companies understand everything they need to know about disruptive technologies — find out more about why our customers love us here.


CBINSIGHTS
Try our platform – free!