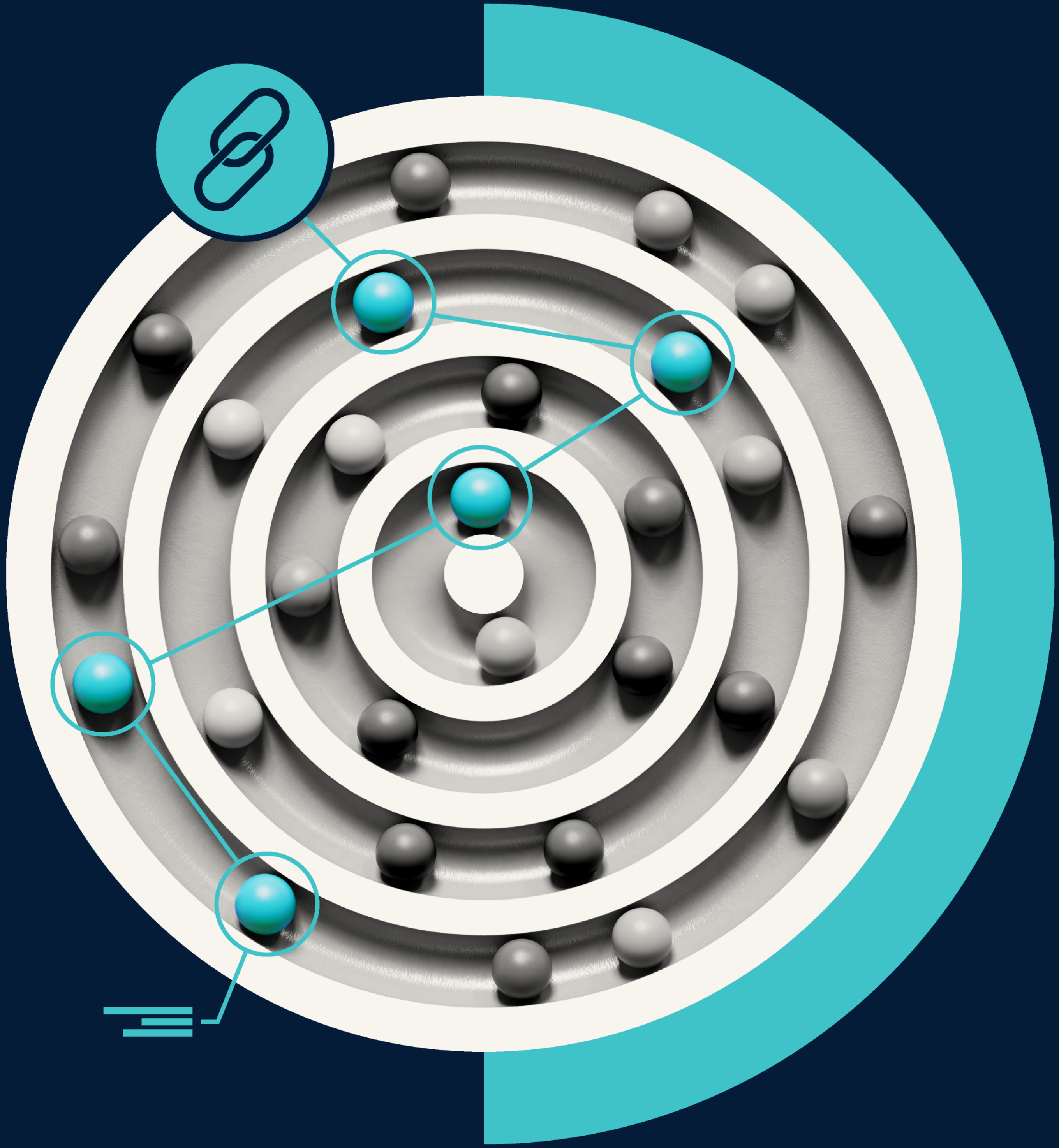
PitchBook

EMERGING TECH RESEARCH

Artificial Intelligence & Machine Learning Report

VC trends and emerging opportunities

Q4 2023



Contents

Institutional Research Group

Analysis

Brendan Burke Senior Analyst, Emerging Technology
brendan.burke@pitchbook.com
pbinstitutionalresearch@pitchbook.com

Data

Matthew Nacionales Data Analyst

Publishing

Report designed by **Megan Woodard** and **Julia Midkiff**

Published on February 27, 2024

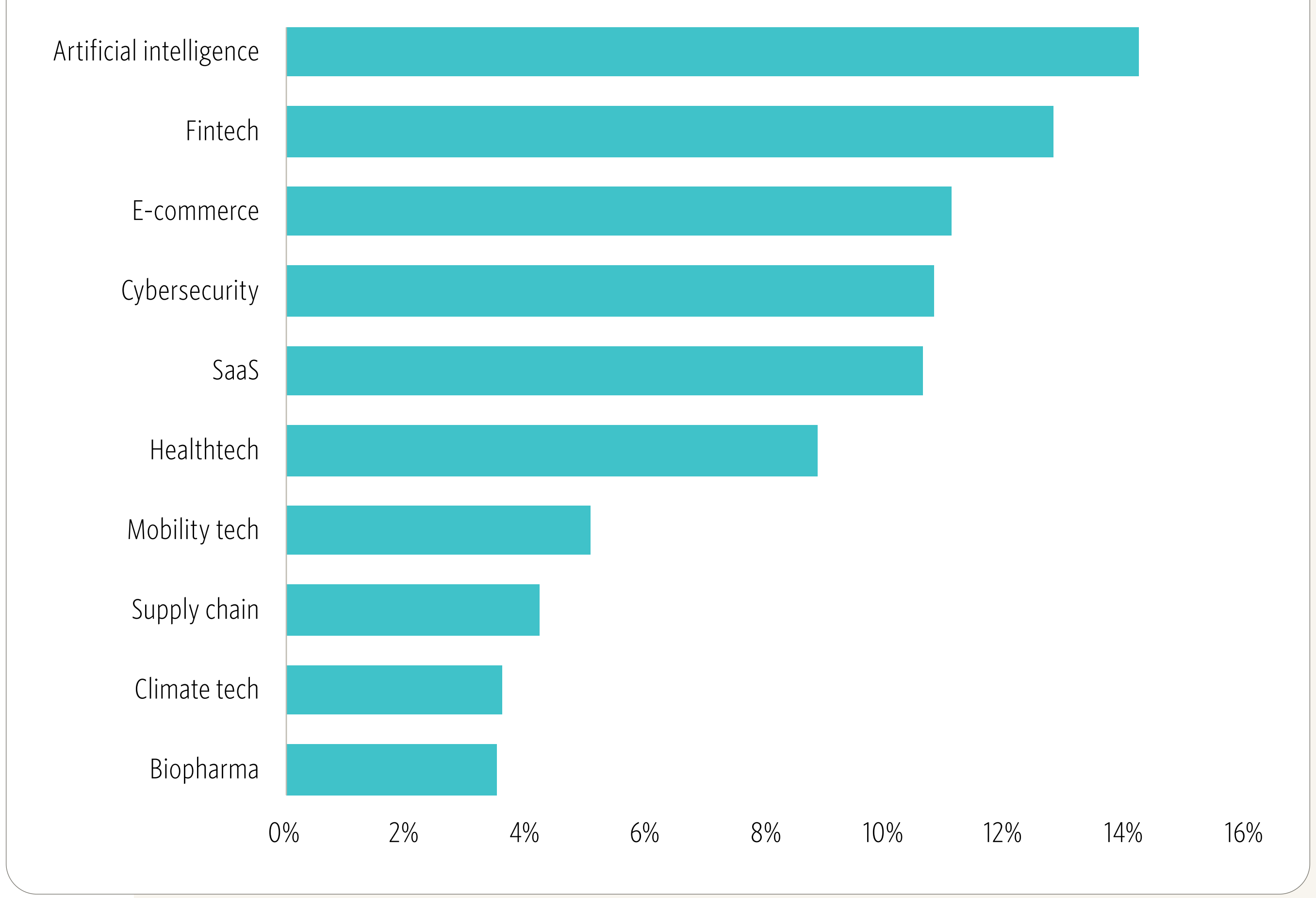
For previous updates as well as our complete AI & ML research, please see the designated analyst workspace on the PitchBook Platform.

Vertical update

The financial forecasts from our Q3 2023 AI & ML public valuation guide and comp sheet are beginning to materialize in blowout earnings guidance for a range of horizontal software giants and traction for AI service providers, emboldening the ecosystem of startups supporting tech leaders. Thus far in Q1 2024, standout earnings releases with AI exposure included Nvidia, Microsoft, ServiceNow, Arm, Meta, Palantir, SK Hynix, IBM, and Qualcomm. These results demonstrate enterprise efforts to integrate Nvidia GPUs into their datacenters, centralize data for AI retrieval, and deploy AI-integrated applications in IT and revenue operations departments. Not all public companies benefit equally, yet leadership in innovation is translating to outstanding results. These financial results are diffusing to a handful of startups offering models to cloud giants including Anthropic and OpenAI, both of whom have robust forecasts for revenue growth in 2024.

These results build on heightened expectations from Q4 that improved valuation expectations for all AI unicorns, not just foundation model startups. The Morningstar PitchBook Global Unicorn Industry Vertical Indexes mark unicorn valuations to a proprietary model via a combination of comparable public and private company valuations and precedent VC deals. Both outstanding unicorn deals and comparable company appreciation led the AI index to outperform all other verticals in Q4 with a 14.1% total return, even as cybersecurity led the field for the year. Public comparable company valuations in semiconductors and pure-play AI core software drove unicorn valuations higher, as laid out in our Q4 2023 AI & ML Public Comp Sheet and Valuation guide. Valuation growth for unicorns including Anthropic, AI21 Labs, Neuralink, and Shield AI also contributed to this outperformance. Q1 2024 share price gains should lead to continued leadership for the AI index.

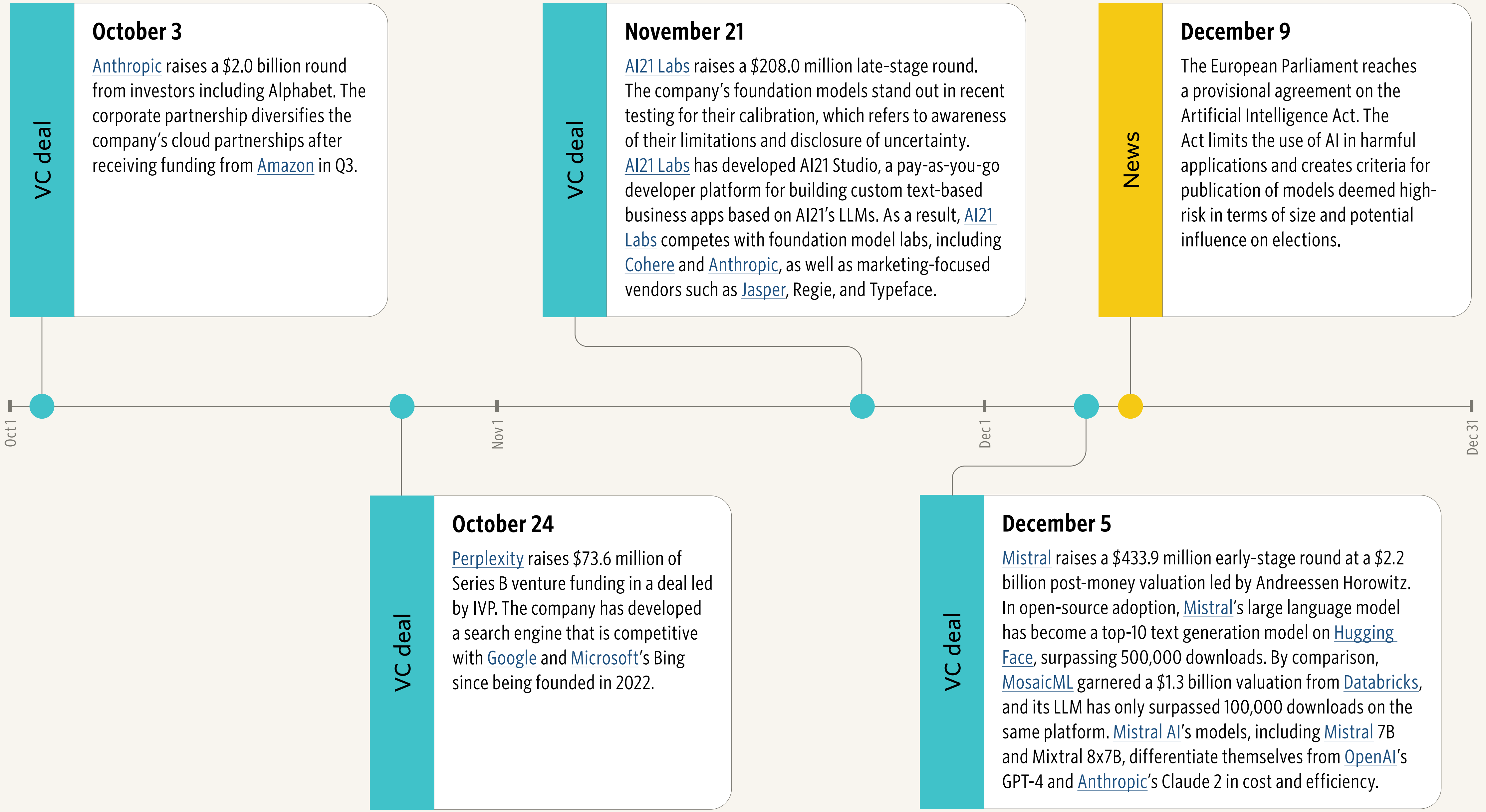
Q4 2023 Morningstar/PitchBook unicorn industry vertical indexes total return*



Source: Morningstar and PitchBook • Geography: Global • *As of December 31, 2023

Note: The Agtech index was excluded for scale.

Q4 2023 timeline



VC deal

October 3

Anthropic raises a $2.0 billion round from investors including Alphabet. The corporate partnership diversifies the company's cloud partnerships after receiving funding from Amazon in Q3.

VC deal

October 24

Perplexity raises $73.6 million of Series B venture funding in a deal led by IVP. The company has developed a search engine that is competitive with Google and Microsoft's Bing since being founded in 2022.

VC deal

November 21

AI21 Labs raises a $208.0 million late-stage round. The company's foundation models stand out in recent testing for their calibration, which refers to awareness of their limitations and disclosure of uncertainty. AI21 Labs has developed AI21 Studio, a pay-as-you-go developer platform for building custom text-based business apps based on AI21's LLMs. As a result, AI21 Labs competes with foundation model labs, including Cohere and Anthropic, as well as marketing-focused vendors such as Jasper, Regie, and Typeface.

VC deal

December 5

Mistral raises a $433.9 million early-stage round at a $2.2 billion post-money valuation led by Andreessen Horowitz. In open-source adoption, Mistral's large language model has become a top-10 text generation model on Hugging Face, surpassing 500,000 downloads. By comparison, MosaicML garnered a $1.3 billion valuation from Databricks, and its LLM has only surpassed 100,000 downloads on the same platform. Mistral AI's models, including Mistral 7B and Mixtral 8x7B, differentiate themselves from OpenAI's GPT-4 and Anthropic's Claude 2 in cost and efficiency.

News

December 9

The European Parliament reaches a provisional agreement on the Artificial Intelligence Act. The Act limits the use of AI in harmful applications and creates criteria for publication of models deemed high-risk in terms of size and potential influence on elections.

Q4 VC deal activity

1,665
total deals

$22.3B
total VC raised

11.0%
deal value growth QoQ

2023 YTD summary

7,238
total deals

$90.9B
total VC raised

-4.7%
deal value growth YoY

AI & ML landscape

1 Horizontal platforms

2 Vertical applications

3 Semiconductors

4 Autonomous machines

AI & ML VC ecosystem market map

This market map is an overview of venture-backed or growth-stage companies that have received venture capital or other notable private investments. Click to view the full map on the PitchBook Platform.

1 Horizontal platforms

AI automation platforms

- Automation Anywhere
- 明略科技
- ThoughtSpot.
- DriveNets
- SingleStore
- Devo
- Icertis
- BigPanda
- harness
- workato

AI core

- DataRobot
- THEODORA
- dataiku
- 达闼
- scale
- neo4j
- DataStax
- sparkcognition
- Vercel
- Weights & Biases

Foundation models

- OpenAI
- ANTHROP\C
- databricks
- Inflection
- Mistral AI_
- Aleph Alpha
- cohere
- Adept
- Hugging Face
- AI21labs

Computer vision

- trax
- MEGVII 旷视
- Black Sesame Technologies
- flock safety
- 依图 YITU
- oosto
- meero
- XREAL
- SCANDIT
- Eagle Eye Networks

Natural language technology

- Dataminr
- attentive
- AlphaSense
- uniphore
- verbit
- Rokid
- talkdesk
- grammarly
- patsnap
- kore.ai

2 Vertical applications

Consumer

- ByteDance 字节跳动
- 便利蜂
- FAIRE
- Contentsquare
- 今日头条
- Jellysmack
- animoca brands
- Niantic
- GoStudent
- Improbable

Financial services

- stripe
- AVANT
- JDT 京东科技
- NAVAN
- ramp
- Mission Lane
- lentra
- Tradeshift
- 量化派 Quant Group
- Liquidity Certainty is a science

Healthcare

- Tempus
- commure
- 平安医保科技
- freenome
- Human Longevity
- XtalPi
- Generate:Biomedicines
- NOOM
- Scientia Technologies

Industrial

- eagleview
- Relativity
- indigo
- Plenty
- FBN
- Bowery
- Plume
- KoBold Metals
- Bright Machines
- Uptake

IT

- Tricentis
- FIGURE
- securonix
- onetrust
- Arctic Wolf
- Perfect Day
- checkr
- Highspot
- transmit security
- GONG

Transportation

- metropolis
- momentive
- T3出行
- via
- motive
- 斑马智行 Powered by AliOS
- G7
- Cambridge Mobile Telematics
- PATEO
- SmartDrive

3 Semiconductors

Datacenter

- CoreWeave
- Crusoe
- Kao Data
- Lambda
- Genesis Cloud
- salad
- Cloudalize
- colovore
- Denvr dataworks
- FluidStack

Edge AI software

- 摩尔线程 Moore Threads
- SiMa.ai
- celona
- auradine
- Clobotics
- black ore
- deci
- Armada
- Edge Impulse

Intelligent sensors & devices

- DREAME
- Terminus 特斯联
- Verkada
- ambiq
- Leia Inc.
- Aqara
- humane
- Density
- sense

Processor design

- 紫光展锐 UNISOC
- Horizon Robotics
- SambaNova
- ESWIN 奕斯伟计算
- intellifusion 云天励飞
- 壁仞科技 Biren Technology
- BITMAIN
- Enflame 燧原科技
- cerebras
- GRAPHCORE

4 Autonomous machines

Autonomous vehicles

- cruise
- Waymo
- AIWAYS
- WeRide 文远知行
- momenta
- pony.ai
- 阿维塔 AVATR
- 赢彻科技
- Applied Intuition

Intelligent robotics

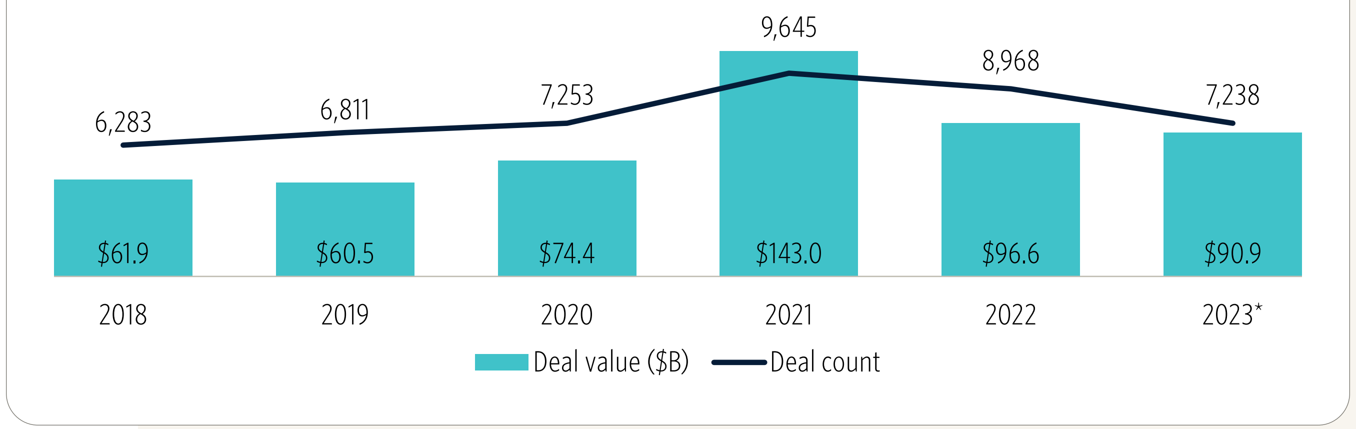
- Anduril
- nuro
- Shield AI
- Skydio
- Geek+
- Agile Robots
- GreyOrange
- EPIRUS
- COWAROBOT 酷哇机器人
- opentrons

VC activity

VC activity has remained consistent throughout the VC downturn over the past six quarters, leaving aside the tech giant megadeals that commit upfront funds for future cloud purchases. We tracked $22.3 billion invested in Q4, including an outlier $2.0 billion deal size for Anthropic, led by Alphabet. Median valuations remain above the market midpoint while equity sold falls lower, demonstrating founder-friendly conditions. Deal count has remained sharply lower than pre-downturn levels, even given generative AI (GenAI) excitement over the past five quarters. These totals would assuredly be much lower without tech giants' involvement in GenAI deals, with GenAI leaders raising $6.0 billion in Q4 across only 194 deals. The momentum in horizontal platforms led the segment to set a VC record in 2023 with $33.0 billion raised, even as vertical applications regressed to 2020's level.

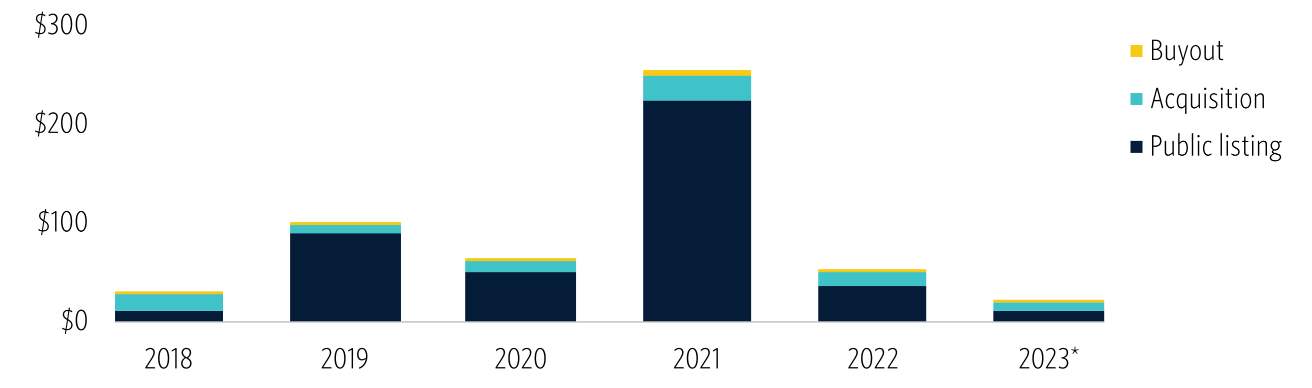
Q3's bump in VC exit M&A did not sustain in Q4, as both deal value and count fell. We only tracked $2.7 billion in disclosed deal value in Q4, which will likely settle as the lowest quarter since Q1 2019. Tech giants remained dormant in M&A given their focus on partnerships with leading LLM startups. Exceptions included AMD's acquisition of Nod.AI in machine learning operations (MLOps), IBM's acquisition of Manta in database management, and ServiceNow's acquisition of UltimateSuite in predictive analytics. An upcoming IPO for semiconductor startup Astera Labs promises to revive deal value in Q1 or Q2 2024. Impressive results for AI companies in public markets should encourage further listings, although our review of the IPO pipeline for this year does not heavily feature AI companies given their ability to stay private.

AI & ML VC deal activity



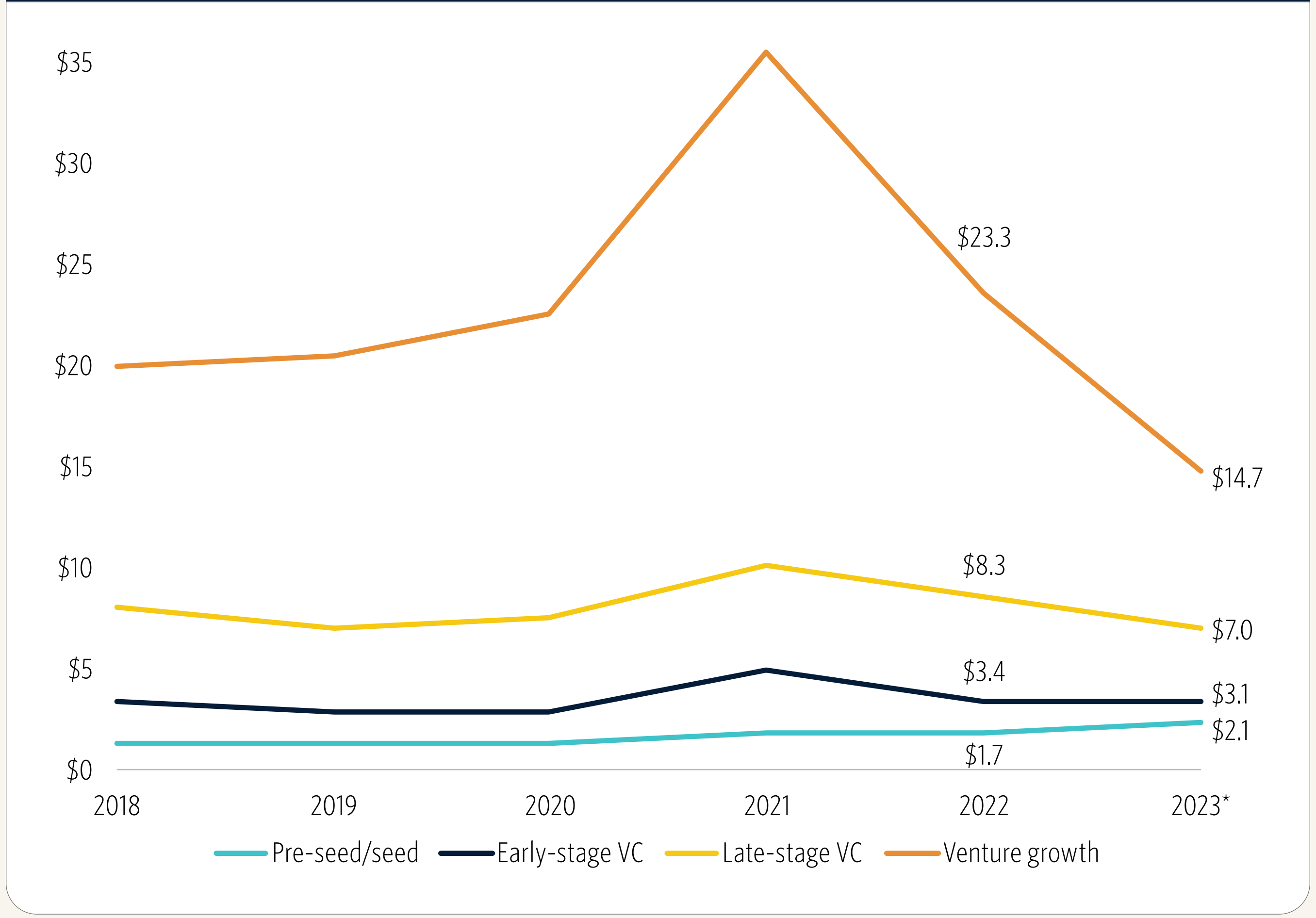
Source: PitchBook • Geography: Global • *As of December 31, 2023

AI & ML VC exit value ($B) by type



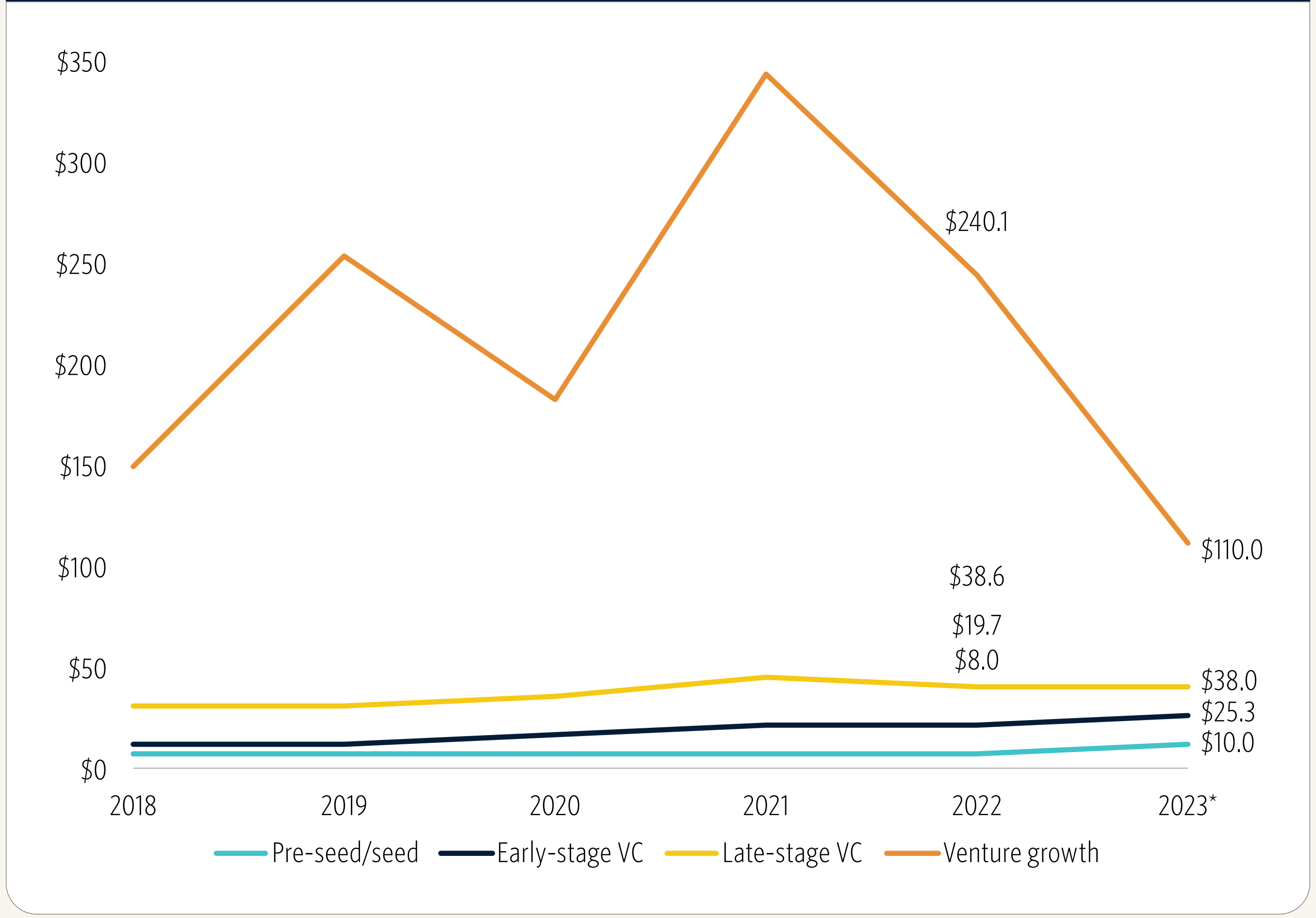
Source: PitchBook • Geography: Global • *As of December 31, 2023

Median AI & ML VC deal value ($M) by stage



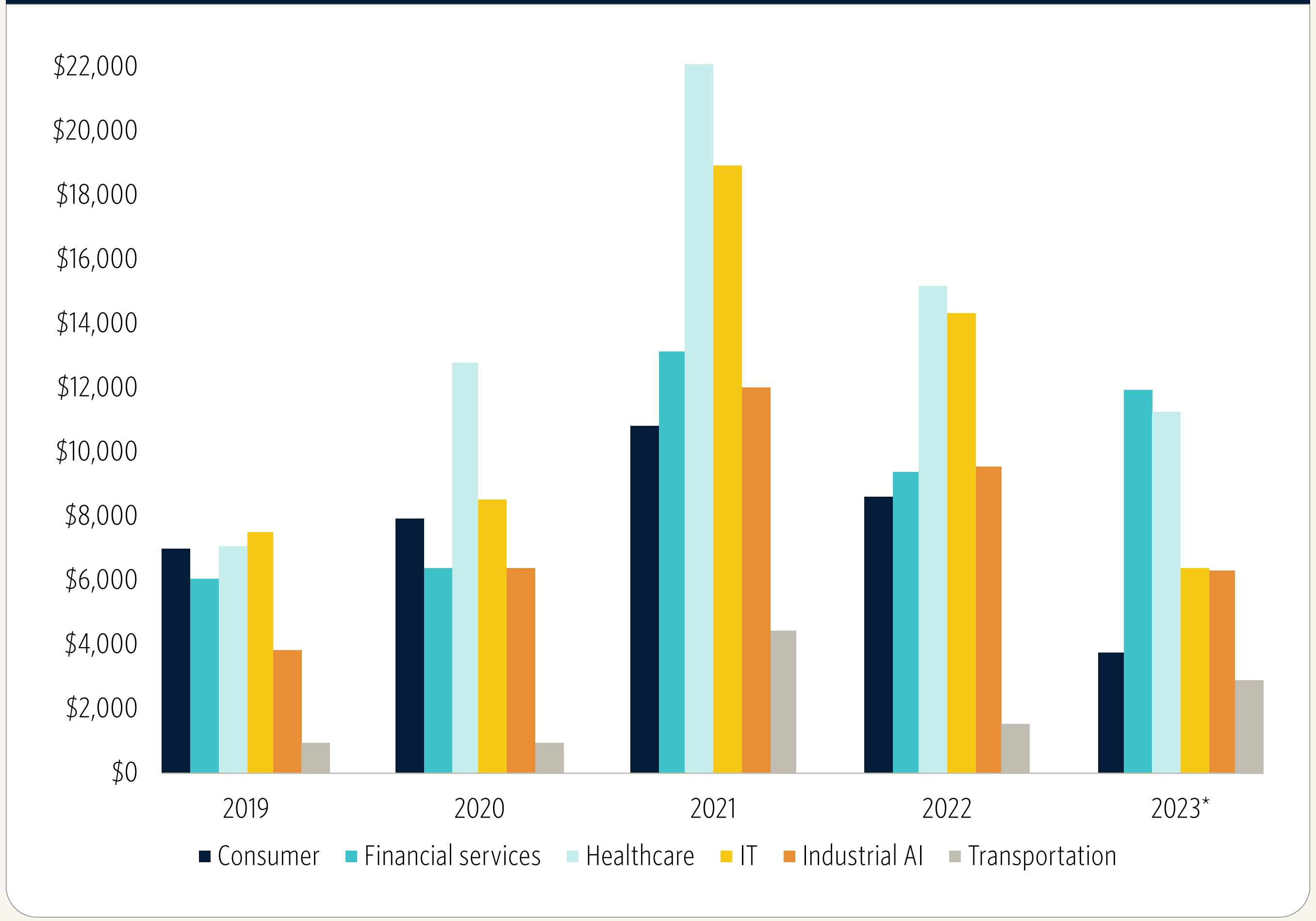
Source: PitchBook • Geography: Global • *As of December 31, 2023

Median AI & ML VC pre-money valuation ($M) by stage



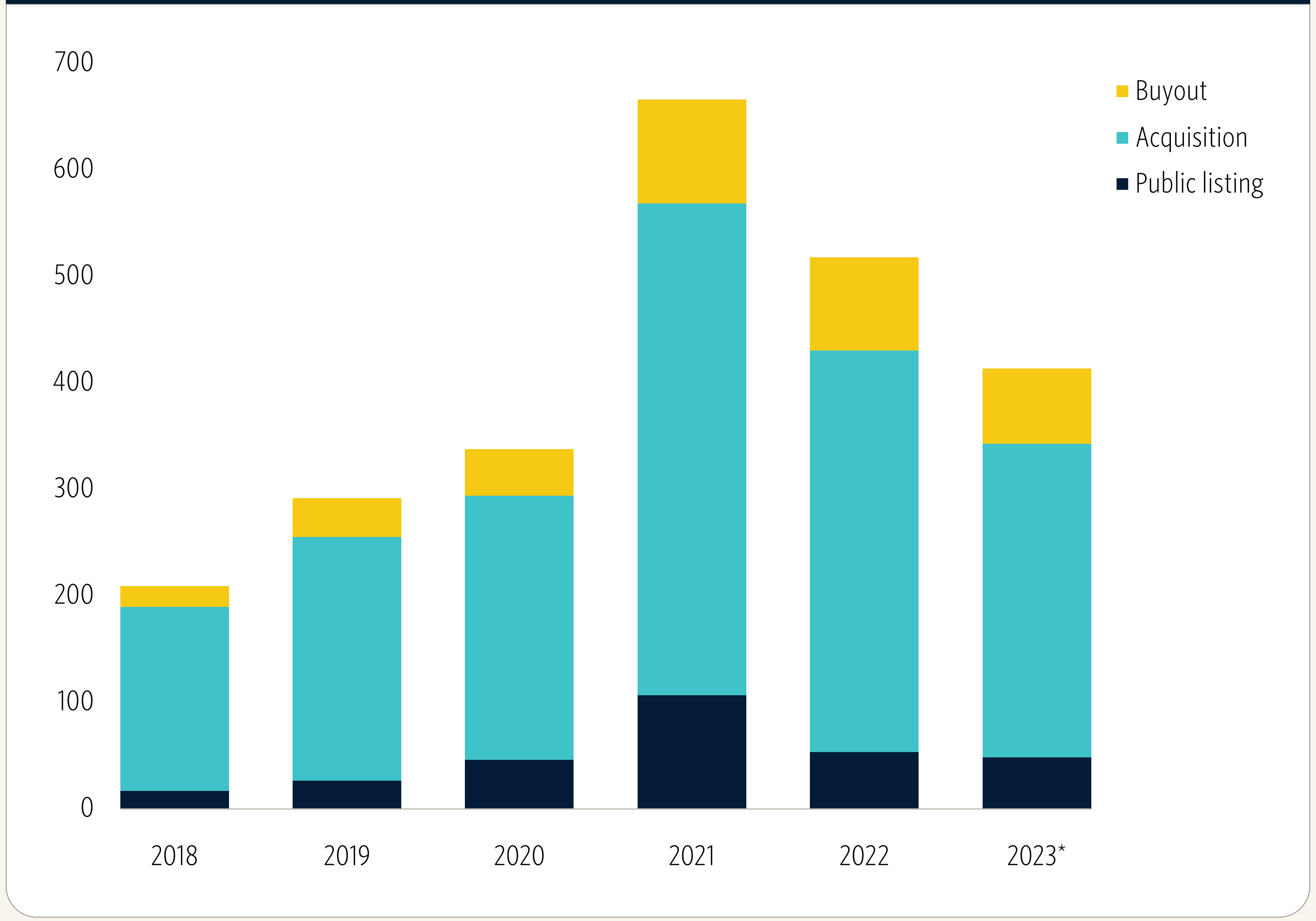
Source: PitchBook • Geography: Global • *As of December 31, 2023

AI & ML VC deal value ($M) by vertical application



Source: PitchBook • Geography: Global • *As of December 31, 2023

AI & ML VC exit count by type



Source: PitchBook • Geography: Global • *As of December 31, 2023

Key Q4 AI & ML early-stage VC deals by deal value ($M)*

Company	Close date (2023)	Subsegment, category	Deal value ($M)	Lead investor(s)	Valuation step-up (post to pre)
Aleph Alpha	November 6	AI core, foundation models	$486.2	Bosch Ventures, Ipai, Schwarz Unternehmenskommunikation	N/A
Mistral AI	December 5	AI core, foundation models	$433.9	Andreessen Horowitz, General Catalyst, Lightspeed Venture Partners	N/A
01.AI	December 14	AI in IT, productivity optimization	$200.0	N/A	N/A
Didi Autonomous Driving	October 13	Autonomous vehicles, autonomous vehicle software	$149.0	GAC Capital	N/A
xAI	December 5	N/A	$134.7	N/A	N/A
together.ai	November 2	AI core, foundation models	$102.5	Emergence (Financial Services), Kleiner Perkins, Nvidia	4.6x
Stability AI	October 1	AI core, foundation models	$86.0	Intel, Sound Ventures, Millennia Capital	N/A
Harvey	December 19	AI in IT, legal automation	$80.0	Elad Gil, Kleiner Perkins	4.2x
Perplexity	November 6	Consumer AI, media & entertainment	$73.6	IVP	3.6x
Hayden AI	December 21	Autonomous vehicles, autonomous vehicle software	$53.0	Drawdown Fund	1.3x

Source: PitchBook • Geography: Global • *As of December 31, 2023

Key Q4 AI & ML late-stage VC deals by deal value ($M)*

Company	Close date (2023)	Subsegment, category	Deal value ($M)	Lead investor(s)	Valuation step-up (post to pre)
Anthropic	October 27	AI core, foundation models	$2,000.0	Alphabet	N/A
Metropolis	October 5	Transportation, ADAs	$1,700.0	3L Capital, Eldridge Industries	N/A
Ramp	December 6	AI in financial services, wealth management	$339.0	Sands Capital, Thrive Capital	0.7x
Neuralink	November 22	Intelligent sensors & devices	$323.2	Founders Fund	1.5x
Verkada	October 9	Intelligent sensors & devices	$305.0	Alkeon Capital Management, Linse Capital	1.9x
Lambda	October 5	Chips, GPU cloud	$300.0	US Innovative Technology Fund (USIT)	7.4x
AI21 Labs	November 21	AI core, foundation models	$208.0	Pitango Venture Capital, SCB 10X, Samsung NEXT Ventures, Walden Catalyst, b2venture	2.1x
Lightmatter	December 19	Chips	$155.0	Aliya Capital Partners, GV, SIP Global Partners, Viking Global Investors	1.6x
Crusoe	December 5	Chips, GPU cloud	$143.7	N/A	N/A
Pony.ai	October 24	Autonomous vehicles, autonomous vehicle software	$100.0	NEOM Investment Fund	1.0x

Source: PitchBook • Geography: Global • *As of December 31, 2023

Key Q4 AI & ML VC exits by exit value ($M)*

Company	Close date (2023)	Subsegment, category	Exit value ($M)	Post-money valuation ($M)	Exit type	Acquirers(s)
Clearpath Robotics	October 2	Intelligent robotics, industrial robots	$454.5	$454.5	Acquisition	Rockwell Automation
LeddarTech	December 22	Intelligent sensors & devices	$259.0	N/A	Public listing	Prospector Capital
Hivestack	December 12	Consumer AI, ad targeting	$100.0	$100.0	Acquisition	Perion Network
Ponder	October 24	AI automation platforms, database management	$46.0	$46.0	Acquisition	Snowflake
Tessian	December 19	AI in IT, information security automation	N/A	N/A	Buyout	Proofpoint, Thoma Bravo
UltimateSuite	December 18	AI automation platforms, predictive analytics	N/A	N/A	Acquisition	ServiceNow
Heap	December 7	AI automation platforms, predictive analytics	N/A	N/A	Acquisition	ContentSquare
WireWheel	December 4	AI in IT, information security automation	N/A	N/A	Acquisition	Osano
Nod.AI	October 19	AI core, model development tools	N/A	N/A	Acquisition	Advanced Micro Devices
Sisu Data	October 16	AI automation platforms, predictive analytics	N/A	N/A	Acquisition	Snowflake

Source: PitchBook • Geography: Global • *As of December 31, 2023

Top strategic acquirers of AI & ML companies since 2017*

Investor	Deal count	Investor type
Accenture	21	Corporation
Apple	21	Corporation
Microsoft	14	Corporation
Cisco Systems	13	Corporation
Meta	13	Corporation
ServiceNow	11	Corporation
International Business Machines	10	Corporation
DataRobot	9	VC-backed company
Intel	9	Corporation
Snowflake	8	Corporation

Source: PitchBook • Geography: Global • *As of December 31, 2023

Top VC investors in AI & ML companies since 2017*

Investor	Deal count	Pre-seed/ seed	Early-stage VC	Late-stage VC	Venture growth	Investor type
Alumni Ventures	400	163	125	97	15	VC
Sequoia Capital	307	82	117	78	30	VC
HongShan	275	30	171	59	15	VC
500 Global	260	123	78	50	9	VC
SOSV	254	130	61	53	10	VC
Andreessen Horowitz	248	56	93	69	30	VC
Right Side Capital Management	248	119	111	17	1	VC
Enterprise Ireland	236	90	87	53	6	VC
Accel	235	49	86	84	16	VC
Antler	234	206	28	0	0	VC

Source: PitchBook • Geography: Global • *As of December 31, 2023

Top VC- and PE-backed AI & ML companies by total VC raised to date*

Company	VC ($M) raised to date	Segment	Subsegment	IPO probability	M&A probability	No exit probability
OpenAI	$10,310.0	Horizontal platforms	AI core	34%	64%	2%
Anthropic	$7,254.0	Horizontal platforms	AI core	93%	4%	3%
Waymo	$5,500.0	Autonomous machines	Autonomous vehicles	40%	58%	2%
Databricks	$4,181.9	Horizontal platforms	AI core	91%	7%	2%
Relativity	$2,383.5	Vertical applications	Industrial AI	78%	6%	16%
Anduril	$2,315.1	Autonomous machines	Intelligent robotics	68%	30%	2%
Nuro	$2,132.0	Autonomous machines	Intelligent robotics	74%	24%	2%
JD Digits	$2,127.9	Vertical applications	Financial services	N/A	N/A	N/A
Horizon Robotics	$1,920.0	AI & ML semiconductors	Processor design	N/A	N/A	N/A
Inflection	$1,565.0	Horizontal platforms	AI core	27%	71%	2%

Source: PitchBook • Geography: Global • *As of December 31, 2023
Note: Probability data is based on PitchBook VC Exit Predictor methodology.

Emerging opportunities

AI datacenters

Today’s datacenter investments in generative AI will take until 2027 to pay off with SaaS-style margins, assuming continued high growth.

Local LLMs

Open-source models serving projects by Ollama, Llama.cpp, and Oobabooga have escalated rapidly in developer adoption since Q2 2023.

Domain-specific foundation models

Newly open-sourced datasets encourage development of new model architectures.

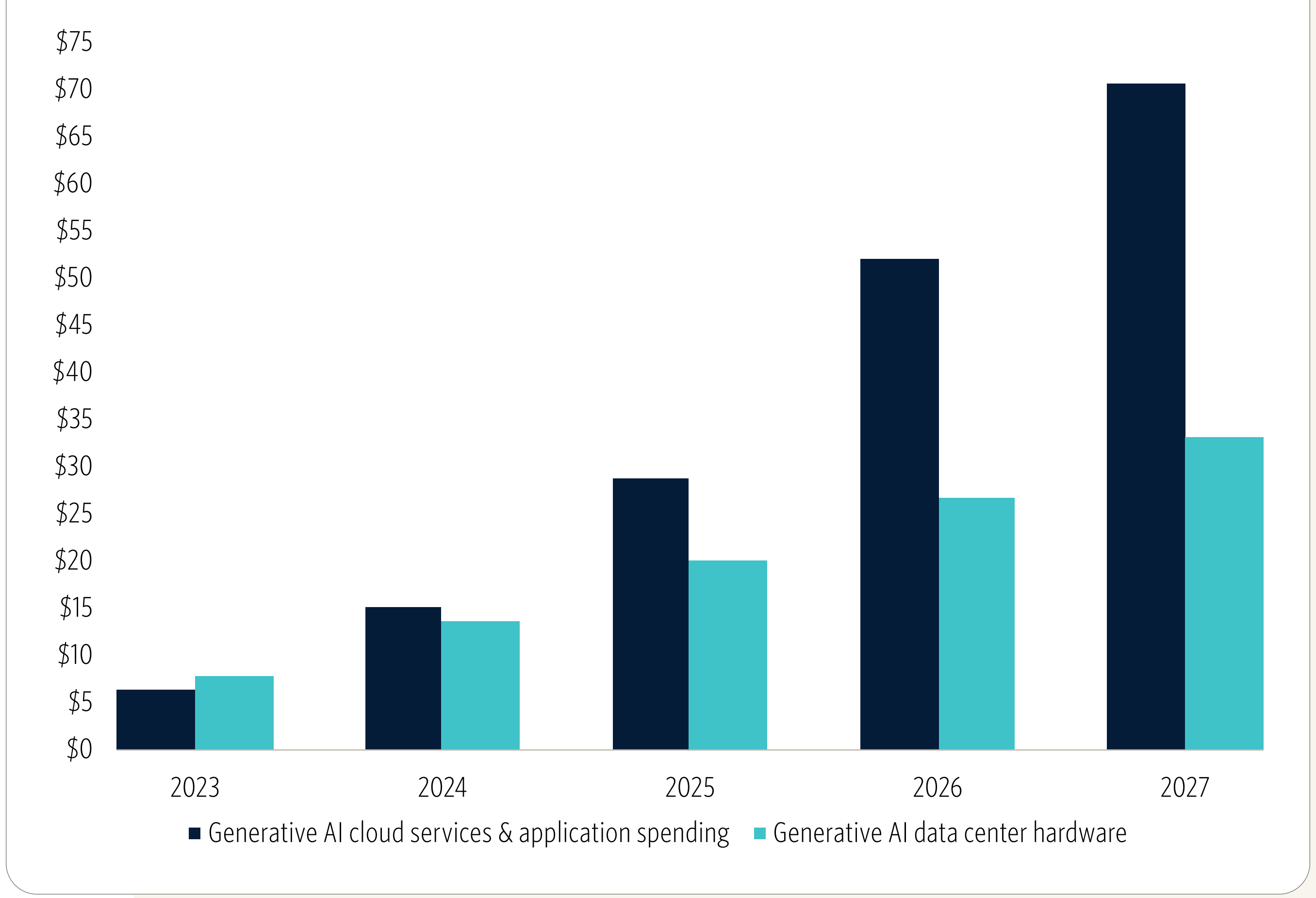
AI datacenters

Overview

AI datacenters distinguish themselves from traditional datacenters through their specialized infrastructure and operations designed to meet the rigorous demands of AI workloads. They are equipped with substantial computing resources, including high-performance servers, storage, and networking, as well as specialized accelerators, to efficiently process AI tasks. Due to the intense power and cooling needs required for high-performance GPUs, AI datacenters necessitate enhanced power connections and innovative cooling methods to manage the significantly higher power consumption and heat output. These centers feature a higher power density, often exceeding 60 kilowatts per rack, far surpassing traditional setups. Their location can be more flexible, allowing for cost savings on real estate, and they embody a greater level of complexity and planning to ensure reliability and mitigate potential failures.

Spending on hardware for AI datacenters currently exceeds the revenue for software applications built on top of them and may take three years to pay off capital investments and five years to generate high SaaS-style margins. According to IDC data, $8.0 billion was spent on generative AI processors, storage, and networking in 2023, producing $2.1 billion in cloud revenue and $4.5 billion in application sales.[1] We assume the vast majority of hardware spending came from cloud datacenter owners such as Microsoft, Amazon, and Google. While cloud and application spending will exceed datacenter hardware spending in 2024, the cost of goods sold will keep generative AI investments cash-flow-negative for several years, with high margins not likely to be generated until 2027. This low-margin period will challenge some lesser-funded datacenter providers to retain customers and continue offering GPUs at low cost.

Comparison of generative AI hardware and software spending ($B)



Source: IDC • Geography: Global

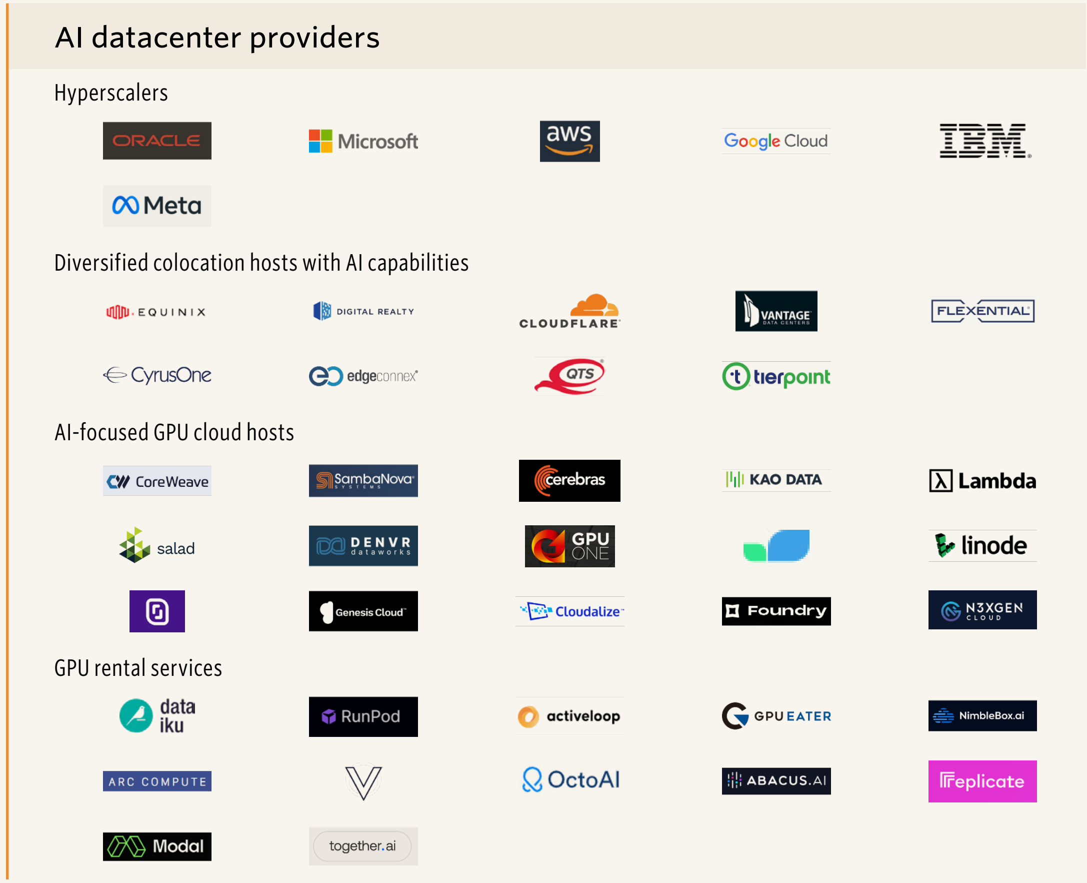
[1] “Worldwide Core IT Spending for GenAI Forecast, 2023–2027: GenAI Is Triggering Hyper-Expansion of AI Spending” IDC, December 20, 2023.

Market direction

Startup cloud providers are undercutting hyperscalers on cost and standing out in supply of advanced semiconductors. According to hourly on-demand pricing, startups are offering 50%-70% cost savings on GPU hours for advanced Nvidia A100s and offering unique access to the latest H100 chips.[2] Leading startup GPU cloud provider Lambda has built the largest cluster of H100 chips of all public clouds, exceeding Google and Oracle.[3] Nvidia's H100 chips are specially designed for transformer model training, encouraging use by leading training labs. Nvidia has since announced the H200 with enhanced memory that is exclusively offered by some AI specialists including Lambda. Nvidia is developing its own cloud services via colocation partners instead of building its own cloud datacenters. Enabling developers to cheaply build on independent GPU clouds offers the possibility to scale revenue non-linearly if startup products gain scale. At present, startups remain core customers for AI cloud specialists, creating risk of customer churn.

While hyperscalers dominate cloud computing, an ecosystem of colocation providers for enterprises developing local services can produce promising investment opportunities. Specialty cloud providers have carved out a $4.6 billion market from the nearly $150 billion internet-as-a-service market, more than 90% of which accrues to US-based hyperscalers and China cloud giants.[4] This market segment is large enough to support large VC-backed companies. Specialty cloud providers stand out not only for AI chip availability but local presence, multicloud support, edge computing, and support for multiple types of legacy hardware. Additionally, Equinix has launched a private cloud service for Nvidia SuperPODs, further strengthening their years-long

GPU cloud hosts market map



AI datacenter providers

Hyperscalers

Diversified colocation hosts with AI capabilities

AI-focused GPU cloud hosts

GPU rental services

[2] "Cloud GPU Comparison," Jolt, n.d., accessed February 8, 2024.
[3] "State of AI Report Compute Index," Air Street Capital and Zeta Alpha, n.d., accessed February 13, 2024.
[4] "Market Share: Enterprise Public Cloud Services, Worldwide, 2022," Gartner, July 25, 2023.

collaboration. This partnership is part of Nvidia's broader efforts to integrate its AI technology into specific domains, including robotics and datacenter architecture. Startup vendors including CoreWeave, Lambda, Together.ai, and RunPod figure prominently in this landscape.

Trending startups

In Q4, GPU cloud startup Lambda achieved an outlier valuation step-up of 7.4x to a $1.3 billion pre-money valuation. The valuation is justified by the company's revenue, which the company forecast to reach $250.0 million in 2023.[5] The company expects further 100% growth in 2024. The company stands out not only for GPU availability via its GPU Cloud, founded in 2018, but also AI compiler software that leverages popular open-source model training frameworks with a simple wrapper. Lambda's client roster includes GenAI unicorns such as Imbue, Picsart, and Writer. Imbue chose to pursue independent partnerships with companies including Dell and Lambda instead of partnering with a hyperscaler. Because of the limited supply of GPUs, Lambda has become a target for growth-stage companies yet may face churn as the competitive landscape shifts in favor of large companies.

At the early stage, new cloud providers are coming online. In Q4, RunPod raised a $18.5 million seed round to bring a scalable GPU cloud to market with serverless software. The company was founded in 2022 and joined the AI Grant program run by leading AI investors Nat Friedman and Daniel Gross. The company is primarily working with AI startups, and we believe the company's business model involves rental of GPUs from cryptominers. The company offers the second-lowest hourly prices on the market for Nvidia 40 GB A100 and A600 instances while making H100s available. The company benefits from the current disillusionment with Docker containers with serverless deployment while being closely involved with the Silicon Valley startup community. GPU rental has proven to be a difficult business model for startups, as evidenced by early GenAI innovator Banana's decision to leave the market. Other alternatives to RunPod include Replicate and Modal Labs.

[5] "Lambda Labs Nears $300 Million Financing," The Information, Kate Clark, October 10, 2023.

Key Q4 recent AI datacenter VC deals ($M)*

Company	Close date (2023)	Deal value ($M)	Post-money valuation ($M)	Lead investor(s)
RunPod	December 4	$18.5	$68.5	N/A
Lambda	October 5	$300.0	$1,550.0	Mercato Partners
Kao Data	October 5	$60.3	$423.7	Legal & General Group, Infratil
Genesis Cloud	October 1	$20.5	N/A	N/A
CoreWeave	May 31	$421.0	$2,521.0	Magnetar Capital, Nvidia

Source: PitchBook • Geography: Global • *As of December 31, 2023

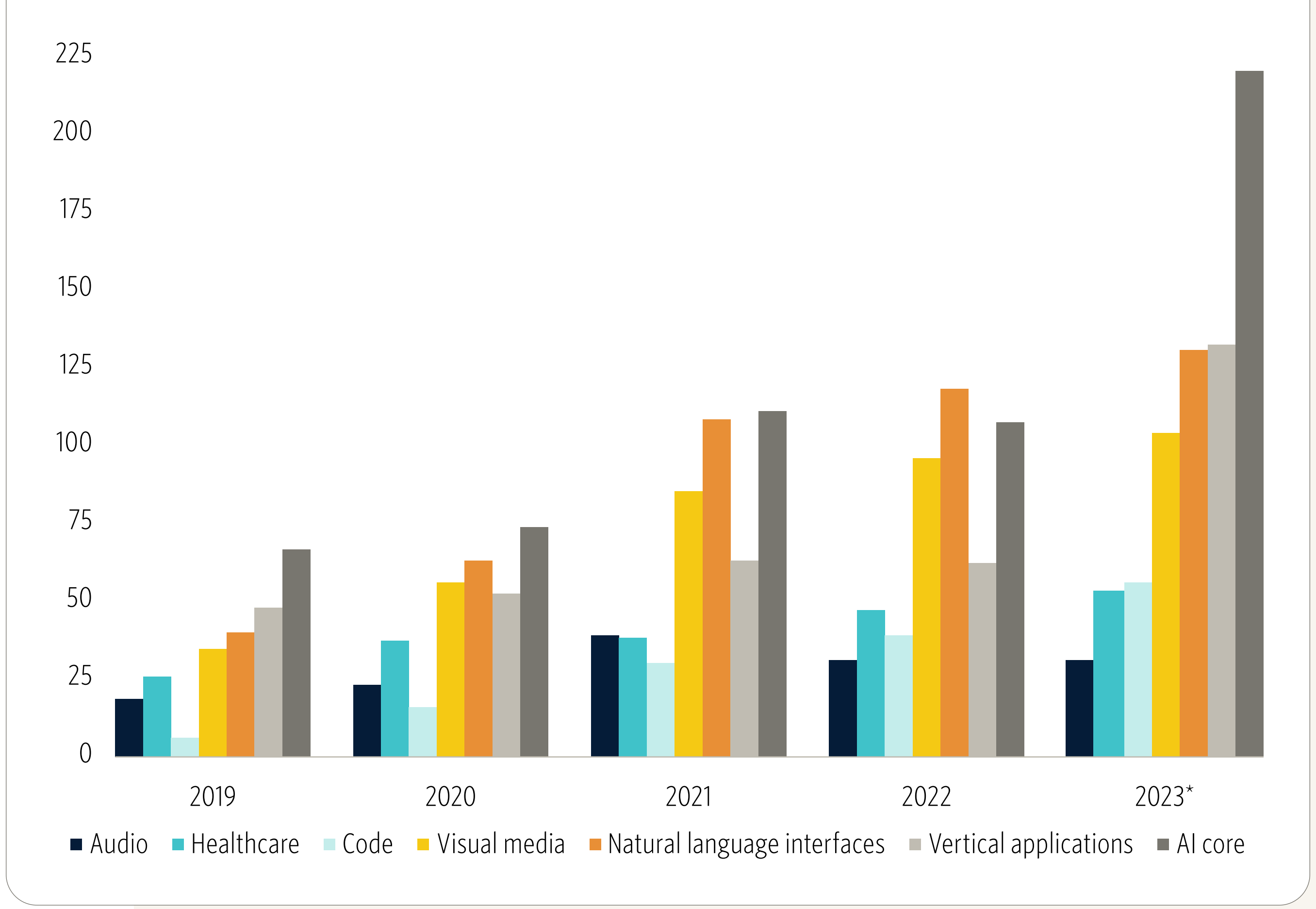
Domain-specific foundation models

Overview

Domain-specific foundation models (DFMs) leverage cutting-edge model architectures with unique datasets to identify novel intelligence not found in conventional large language models. Foundation models differentiate based on the usage of self-attention mechanisms to emphasize specific characteristics of a dataset. DFMs can differ from LLM architectures based on their datasets and balance between encoder and decoder layers. Particularly, multimodal datasets with both images and text can require customized attention mechanisms and fine-tuning strategies. Some pre-existing AI models may be relabeled as foundation models, but we believe investors can differentiate based on the underlying architecture and datasets involved.

Vertical applications have been secondary beneficiaries of the GenAI wave and are likely to produce some of the largest businesses through DFMs. In 2023, vertical application deal count grew more than any other segment in generative AI, with 112.9% growth to 132 deals—second only to AI core software. Vertical applications surpassed natural language interfaces, a segment that includes major use cases in customer support, sales & marketing, and search. Much of this growth owed to the emergence of new use cases that previously did not take advantage of GenAI, including corporate strategy, fintech, industrial, and legal. Early-stage startups lead the development of standalone businesses built on AI in these domains. If the field of GenAI progresses like supervised machine learning did, then large AI-native businesses will be created in each large end market.

Generative AI emerging space VC deal count by segment



Source: PitchBook • Geography: Global • *As of August 18, 2023

Note: A list of companies in each segment can be found in our Generative AI emerging space market map Q1 2024.

Market direction

Each vertical application encourages founders to train custom LLMs. Breakthroughs have been made across a range of industries as researchers find benefits to combining transformer architecture with quantitative data. Most of these research efforts focus on language or image data for knowledge search and content generation. Unique data collection and labeling methods enable DFMs in image-heavy domains including 3D images and retail along with language-focused domains like customer support, code, and legal. In healthcare, DFMs exist both for language (Hippocratic AI) and images (Hoppr). To complement these well-studied modalities, researchers and startups are developing time series foundation models that inherently incorporate industry-specific datasets of tabular data yet can apply across industries. Nixtla, the developer of time series foundation model Time GPT, found a suitable application of time series models in financial forecasting, where conventional language models have little ability to pattern match with new financial data. Google Research recently published research on a competitive model called TimesFM that also performs well across numerous unseen tabular datasets, with an initial focus on retail demand planning. Impressive results across these domains make it unlikely that a base model like OpenAI's will succeed commercially in every industry.

Domain-specific foundation models market map

3D image generation

meshcapade

Code

poolside Magic replit

Customer support

kore.ai PolyAI

Geospatial

IBM Jua

Healthcare

Hippocratic AI — Do No Harm — HOPPR

Legal

Harvey.

Materials

Orbital Materials

Retail

AI21labs vody

Time series

Archetype AI nixtla Alphabet

Robotics

NVIDIA.

Transformer architecture encourages aggregation and open access to unique datasets that can enable new foundation models. In Q4, Alphabet subsidiary DeepMind collaborated with academic researchers to open-source the Open-X Embodiment dataset for robotics. The project has already resulted in a new robotics foundation model with RT-X and a competitive open-source model Octo from academic researchers. IBM and NASA collaborated on a geospatial foundation model based on NASA's repository of Earth-satellite data. In healthcare, academic researchers leverage unique clinical datasets for biomedical image classification based on novel biomarker research with promising early results. Both proprietary datasets and synthetic data remain important for these new paradigms. Academic projects are likely to commercialize over time via startups.

Trending startups

Domain-specific models produce commercial traction for startups. Many AI core startups have not generated large customer bases outside of other AI startups and experimental developers, limiting their revenue growth. Bucking this trend, DFM startup Harvey has scaled to $10.0 million in revenue via partnerships with professional services firms including PwC. In Q4, the company raised an $80.0 million Series B at a 4.2x valuation step-up in less than a year, led by AI specialist VCs Elad Gil and Kleiner Perkins. The company is expanding from the legal domain to other professional services. Harvey customer PwC is developing an M&A-specific model based on internal documents from its M&A processes that can smooth document review. This business model resembles the custom LLM service offered by OpenAI yet may produce network effects within large professional services industries as architectural choices yield improved performance on individual tasks.

At the early stage, domain-specific model companies can raise large seed rounds to focus on high-value domains such as code. In Q3 2023, code model startup Poolside AI raised a $126.0 million startup to develop an end-to-end application development model that learns from both software development and production environments. The company disclosed model development efforts in collaboration with datacenter provider Scaleway that should produce a working model by mid-2024. Also, in Q3 Imbue raised a $212.0 million Series B that is focused on general-purpose agents, yet the company initially focused on code development models to differentiate from general purpose model labs. In more niche domains, startups are raising lower deal values to develop proofs of concept at a smaller scale. We believe some of those efforts may also be deserving of large funding rounds given their more immediate commercial applications.

Recent domain-specific foundation model VC deals ($M)*

Company	Close date	Domain	Deal value ($M)	Deal type	Lead investor(s)
Harvey	December 19	Legal (and other professional services)	$80.0	Series B	Elad Gil, Kleiner Perkins
HOPPR	November 26	Healthcare	$3.0	Late-stage VC	Health2047
Imbue	September 5	Code	$212.0	Series B	Astera Institute, Further Future Foundation, Nvidia
Archetype Ai	August 30	Time series	N/A	Seed	Venrock
Poolside AI	August 24	Code	$126.0	Seed	Felicis, Redpoint Ventures, Xavier Niel
Hippocratic AI	July 25	Healthcare	$15.0	Seed	Cincinnati Children's Hospital Medical Center, HonorHealth

Source: PitchBook • Geography: Global • *As of December 31, 2023

Local LLMs

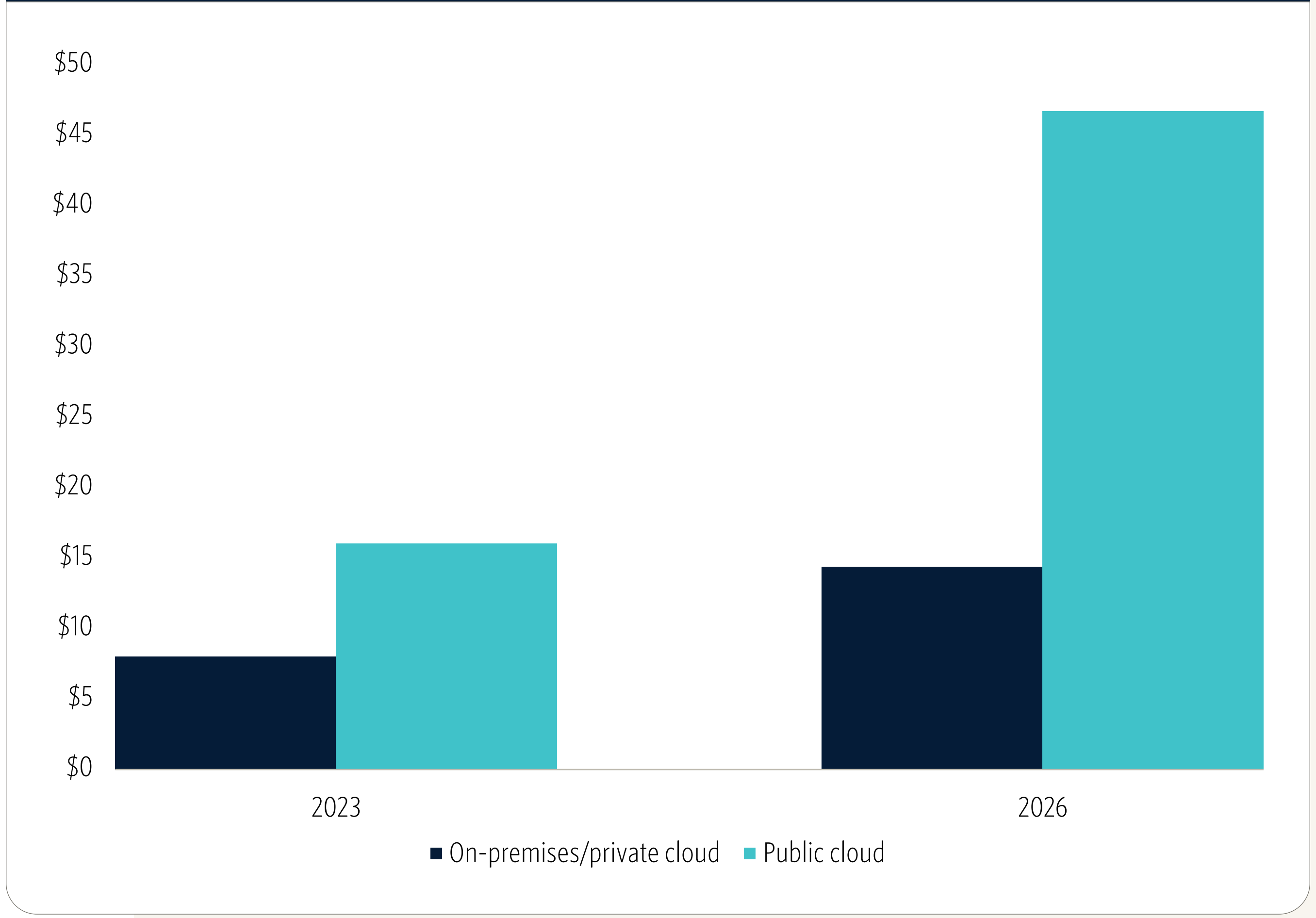
Overview

Local LLMs bring advanced computation from the cloud to distributed workstations, including CPUs and smartphones. They depend on small model architectures along with model-serving techniques that include optimization engines and request batching. Small models benefit from curated datasets that include synthetic samples of reasoning, some of which are generated by other AI models. This data-centric process reduces the number of parameters. To serve these small models locally, optimization engines convert model queries into python scripts and request-batching servers create a queue that can be managed by small hardware form factors such as laptops and phones. Local LLM frameworks can support multiple types of models including open-source models and API calls to closed-source models. Current frameworks focus on small instances of LLMs and fine-tuned open-source models including Meta's Llama, Mistral's Mixtral, and Microsoft's Phi.

On-premises workloads remain an important part of AI workloads. On-premises AI software services contributed 33.3% of AI software services spending in 2023, according to IDC, reaching $7.9 billion. While public cloud is growing much faster, on-premises spending is still on pace to double by 2027. This market size remains large enough for startup opportunities, particularly with the tailwind of generative AI.

Startups can address the local LLM opportunity by providing model-serving frameworks. Model serving addresses the need for hardware optimization in AI models. Nvidia stands out for the developer-friendliness of its CUDA software in serving models to the company's proprietary hardware via compiler software, encouraging adoption of its hardware in datacenters. Nvidia has achieved $1.0 billion annual recurring revenue (ARR) for its software, support, and services offerings for the datacenter hardware, which is not a high percentage of hardware sales given bundling of CUDA

AI lifecycle software market size estimate by deployment type ($B)*



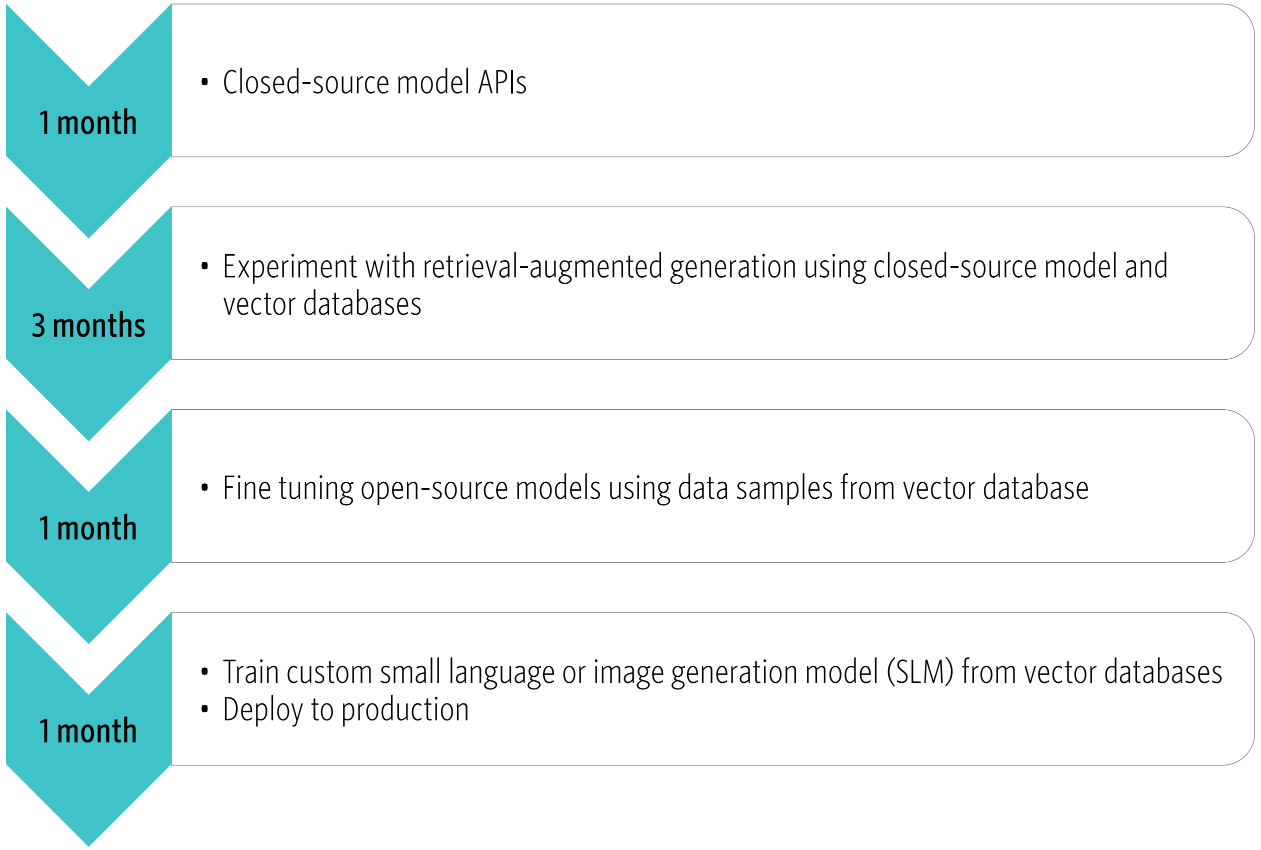
Source: IDC • Geography: Global • *As of November 27, 2023 software with GPU chips, yet it remains a large business that other hardware vendors would like to match.[6] Model-serving companies have generated large developer communities if not significant financial outcomes, led by Numenta's $352.8 million valuation from 2021.

Market direction

Deployment startups can benefit from partnerships with hardware manufacturers. Chip incumbent AMD has turned to hardware optimization startups to close its deficit in software to Nvidia, acquiring Nod.ai for hardware deployment and partnering with Lamini for training optimization. Nod.ai had previously been valued at $64.0 million, suggesting the acquisition may have been more than $100 million. These partnerships are not solely intended to address local deployment but also benefit AMD's PC processors, a market segment in which the company forms a duopoly with Intel. LG has partnered with VC-backed Upstage to develop on-device models. Open-source project llama.cpp optimizes models to run on Apple M2 chips, alleviating the need to buy GPUs to run LLMs. Apple previously acquired local machine learning model startup Xnor.ai for $200.0 million and may be similarly encouraged to find advanced methods to compress internally developed open-source models to run on its hardware.

Large enterprises are encouraged to consider small language models to run locally after experimenting. We have heard from leading AI consultants that sophisticated customers opt to train small language models and run them on self-hosted hardware once they develop sufficient databases. 25% of large enterprises currently self-host models, a figure we believe includes the most sophisticated users.[7] According to the consultants, power users are encouraged by the cost savings

Indicative internal LLM application development lifecycle for midsize to large enterprise



1 month
- Closed-source model APIs

3 months
- Experiment with retrieval-augmented generation using closed-source model and vector databases

1 month
- Fine tuning open-source models using data samples from vector database

1 month
- Train custom small language or image generation model (SLM) from vector databases
- Deploy to production

[6] "Investor Presentation Q3 FY24," Nvidia, November 27, 2023
[7] "State of AI: A 2023 Report on AI in Production," Retool, November 13, 2023.

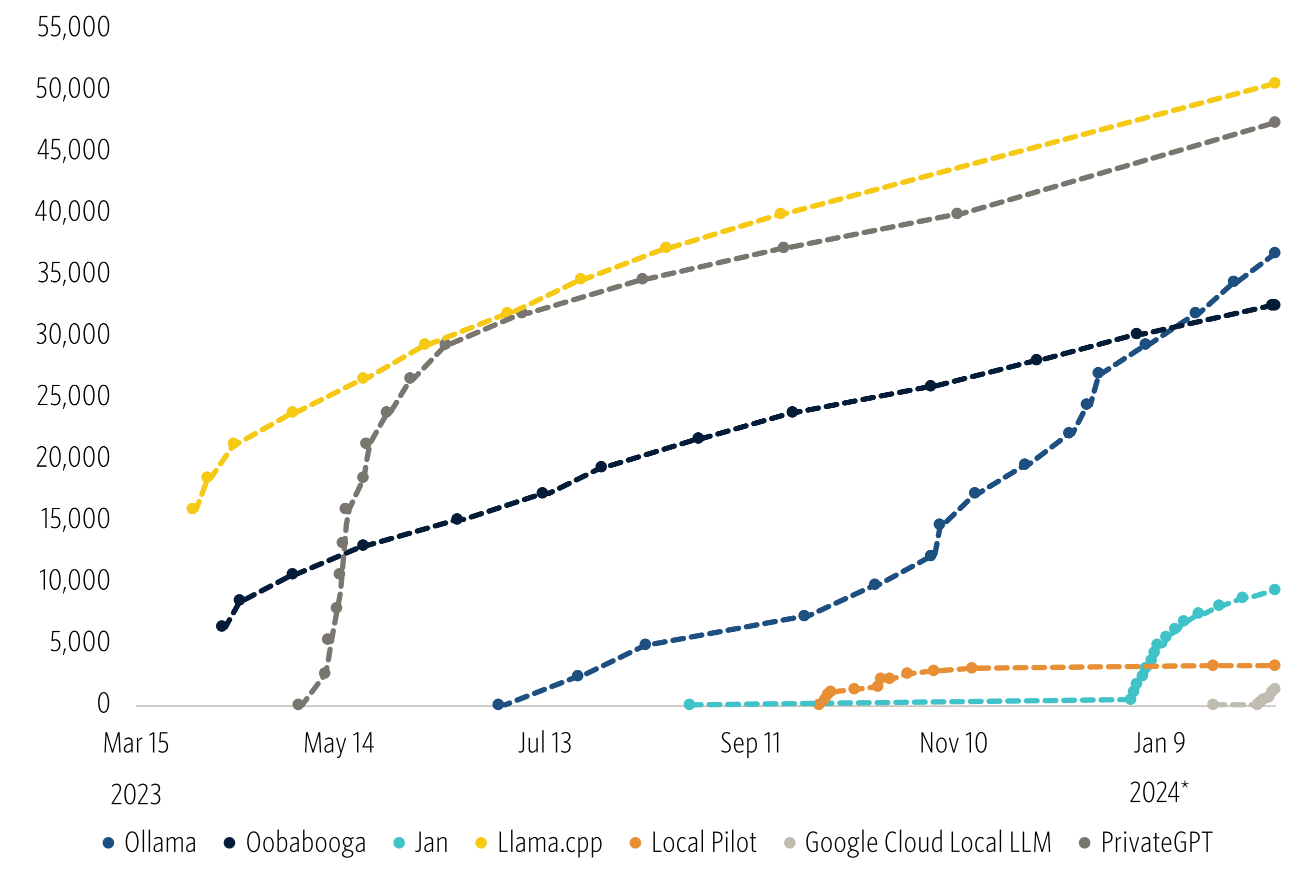
and accuracy in specific tasks offered by Microsoft's latest model Phi, which incorporates only 2 billion parameters—an infinitesimal fraction of those in leading models such as GPT-4. This technique can benefit from local LLMs. While independent developers have been leading proponents of local models, we believe that lengthy enterprise pilot projects are leading to custom training of local models.

Trending startups

We track hardware acceleration and edge deployment startups as a standalone category of our large language model operations (LLMOps) market map. The fastest-growing startups come from the open-source community and have not yet required VC funding. Open-source projects by Ollama, Llama.cpp, and Oobabooga have escalated rapidly in developer adoption since Q2 2023. Ollama is on a particularly hyperbolic trajectory, as it has grown in sync with new models from open-source startup Mistral. The project's local inference engine has become part of reference architectures from leading orchestration framework builders including LangChain and LlamaIndex. The project has raised seed funding from Index Ventures in the past and may be in line for a Series A.

In Q4, Armada raised a $40.0 million Series A led by blue-chip deep tech investors Founders Fund and Lux Capital to bring LLMs to the edge. The company is developing local datacenters built inside shipping containers that are intended to bring AI computing to remote regions and internet-of-things (IoT) applications. The team is led by former DataRobot executive Dan Wright, joined by big tech alums. The company directs its efforts to defense and industrial applications rather than large LLM use cases, making the company relatively underexposed to the generative AI wave. Early-stage startup Arcee.ai raised a $5.5 million round in Q1 2024 to adapt small language models (SLMs) for private cloud usage. The startup was founded by former Hugging Face engineers who have unique insights into the open-source model ecosystem.

GitHub star count for selected open-source local large language model repositories



Source: GitHub Star History • Geography: Global • *As of February 12, 2024

Select company highlights

ANTHROP\C

Overview

Anthropic was founded in 2021 by a group of seven former senior members of OpenAI. Several of the founders worked at Google Research before OpenAI. The founders split off from OpenAI with a desire to train safe AI systems that could mitigate the worst risks of superintelligence, including human extinction. To serve this mission, the company was established as a public-benefit corporation. Since its founding, Anthropic has developed proprietary AI systems and commercialized a series of large language models called Claude.

Claude leverages transformer architecture and a reinforcement learning process referred to as Constitutional AI. Constitutional AI tests the alignment of model outputs against a list of ethical rules and principles. In this way, the model self-improves by selecting outputs that conform to the company's ethics. The Constitutional AI process improves the safety of responses without extensive human intervention such as OpenAI uses via its outsourced reinforcement learning from human feedback approach. We have not tracked any patent for this technique. Because of this process, Anthropic does not shift its response types as often as ChatGPT, which faces complaints of censorship and model drift as the model changes over time.

Key company information

Founded
2021

Employees
300 in 8 offices globally

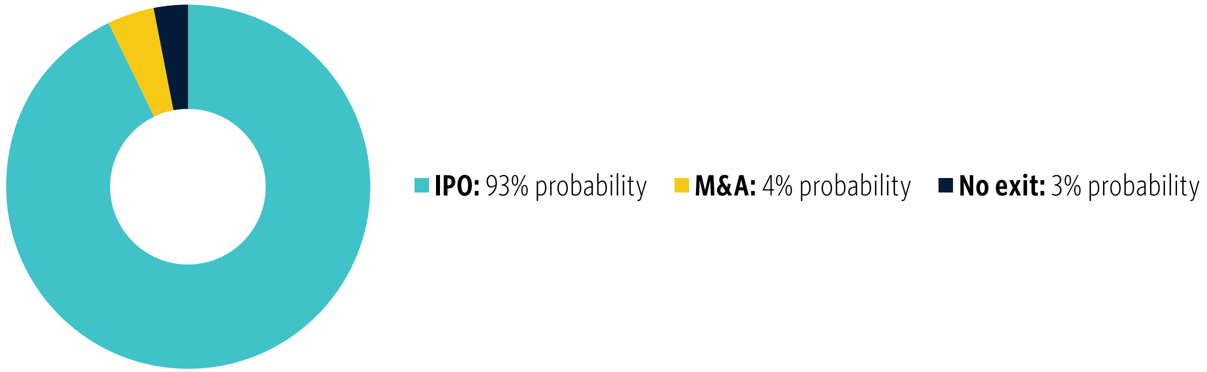
Total raised
$7.3B

Last financing valuation
$25.0B

Last financing
Raised $300.0M in a Series G

Lead investor(s)
Alphabet, Amazon, LG CNS, Calm Ventures, SK Telecom, Salesforce Ventures, Zoom Ventures, Spark Capital, Sam Bankman-Fried, Jaan Tallinn

Exit Predictor



Note: Probability data based on PitchBook VC Exit Predictor Methodology.

Leadership

The company is co-led by CEO & Co-founder Dario Amodei and President and Co-founder Daniela Amodei. The two are siblings and neither reports to the other. At OpenAI, Dario Amodei focused on research and Daniela Amodei on safety and policy. Dario Amodei formerly achieved a PhD in biophysics from Princeton and worked as a research scientist at Baidu and Google. Daniela Amodei comes from a non-technical background and formerly was a risk manager at Stripe and led human resources and business operations at OpenAI along with technical safety teams.

Other key co-founder executives include Head of Compute Engineering Tom Brown and Head of Policy Jack Clark. Brown led the engineering of OpenAI's GPT-3 and was formerly a startup CTO and Google researcher. Jack Clark was formerly a tech journalist and led communications at OpenAI.

Competitors

Anthropic competes in the foundation model category of AI against OpenAI, Microsoft, Google, Meta, and the open-source community. OpenAI, Meta, and Google are developing comparable LLMs. Recently, open-source foundation model startup Mistral matched Anthropic's performance despite starting two years later with fewer resources. Google has also made rapid progress with its Gemini Ultra model. OpenAI's GPT-5 is on pace to set a new bar for performance this year, which will require Claude to keep pace via successive model releases.

Anthropic partners with AWS and Google via cloud investments. The company raised a $4.0 billion investment led by Amazon that involved both cloud credits and equity investment. A month later, the company raised a $2.0 billion round led by Google. Based on these partnerships, AWS offers a managed Claude service on its Bedrock deployment platform. This partnership is a superpower for Anthropic given AWS's leading position in the cloud service provider market. Amazon has been leading in generative AI hiring while announcing new services at its Re:Invent conference, demonstrating the investment being made into the field.

Anthropic will be a primary beneficiary of existing AWS customers leveraging the hyperscaler's Bedrock LLM service. Additionally, significant AWS consulting relationships from Accenture, Deloitte, and Slalom will encourage the use of Anthropic models. We have already seen Accenture highlight Anthropic in an AWS reference architecture.[8] Accenture is the leading AI services firm by market share. Given that consultants present the first port of call for enterprise deployments, Anthropic can benefit from close relationships via AWS.

Outlook

Anthropic presents a legitimate alternative to OpenAI given the company's research talent, partnership with AWS, and focus on safety issues that will be a critical concern for enterprise customers. The company disclosed reaching $200.0 million in ARR in 2023. Enterprise accounts drive this financial total, with anchor customers including Notion, Quora, and DuckDuckGo. The company forecasts 325% growth in 2024 to $850.0 million ARR, demonstrating the traction the company's models have in enterprises and the potential for Anthropic to become a larger enterprise vendor than OpenAI. Successive financing rounds are likely to occur at a steep increase given recent discussions about a $18.4 billion valuation. Current investment opportunities include secondary shares from prior transactions along with a special-purpose vehicle (SPV) from Menlo Ventures.

[8] 8: "Accenture Creates a Knowledge Assist Solution Using Generative AI Services on AWS," AWS, October 5, 2023.

SELECT COMPANY HIGHLIGHTS: ANTHROPIC

Financing history

Series A

May 28, 2021

Total raised
$124.0M

Pre-money valuation
$550.0M

Lead investor(s)
Jaan Tallinn

Series B

April 29, 2022

Total raised
$580.0M

Pre-money valuation
$3.4B

Lead investor(s)
Sam Bankman-Fried

Series C

May 23, 2023

Total raised
$450.0M

Pre-money valuation
$4.6B

Lead investor(s)
Salesforce Ventures,
Zoom Ventures, Spark Capital

Late-stage VC

August 13, 2023

Total raised
$100.0M

Pre-money valuation
N/A

Lead investor(s)
LG CNS, Calm Ventures and
SK Telecom

Late-stage VC

September 25, 2023

Total raised
$4.0B

Pre-money valuation
N/A

Lead investor(s)
Amazon

Late-stage VC

October 27, 2023

Total raised
$2.0B

Pre-money valuation
$5.8B

Lead investor(s)
Andreessen Horowitz

Series G

February 1, 2021

Total raised
$1.0B

Pre-money valuation
$25.0B (estimated)

Lead investor(s)
Alphabet

SELECT COMPANY HIGHLIGHTS: DATABRICKS

databricks

Overview

Founded in 2013, Databricks offers a data science platform that includes AI-as-a-service (AIaaS) functionality with a suite of data science tools for data engineering, data warehousing, and ML algorithms. The company grew out of the open-source Apache Spark data science community and created an extensible product, called the Unified Data Analytics Platform, that can ingest data from enterprise silos and prepare it for cluster-based computing. Once an effective open-source product was in place, the company moved to a closed-source model and rapidly increased revenue beginning in 2016.

Databricks now offers a product suite on top of Apache Spark that includes an AI platform called Lakehouse AI, which is optimized for generative AI with data storage, a model training pipeline, and model-serving and monitoring; a data lake, which integrates unstructured data in a central database for analytics; and data security. The company has focused on the GenAI opportunity, training a custom LLM called Dolly that can run more efficiently than GPT-4 and achieve competitive results on question-answering tasks. Databricks recently acquired Einblick for data analysis using generative AI.

Key company information

Founded
2013

Employees
7,939 in 25 offices globally

December 2023
Leader in Gartner Magic Quadrant for Cloud Database Management Systems

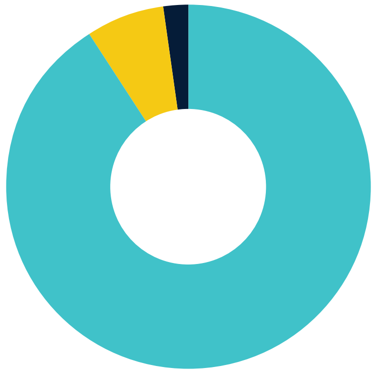
Total raised
$4.2B

Last financing valuation
$43.2B

Last financing
Raised $684.6M in a Series I

Lead investor(s)
Nvidia, T. Rowe Price, Counterpoint Global, Franklin Templeton Investments, Andreessen Horowitz, NEA

Exit Predictor



■ **IPO:** 91% probability ■ **M&A:** 7% probability ■ **No exit:** 2% probability

Note: Probability data based on PitchBook VC Exit Predictor Methodology.

Leadership

Databricks' management has high technical expertise capable of generating product-led growth and has increased its public-company experience. The company was founded by seven early contributors to Apache Spark, and all the co-founders remain at the company. CEO and Co-founder Ali Ghodsi earned a PhD in distributed computing and was vice president of engineering before being promoted to CEO at the start of the company's growth phase in 2016. The acquisition of MosaicML brought in the Vice President of Generative AI, Naveen Rao, who has credibility in the open-source AI community. The board features co-founders and VC investors, including 16z Co-founder Ben Horowitz, yet still lacks the public company leadership we believe is necessary to go public.

Competitors

Databricks stands out in the artificial intelligence & machine learning (AI & ML) market by supporting customized AI model training using open-source models atop its data lake, functionality that distinguishes it from legacy vendors such as SAP, Microsoft, and AWS while also being more AI-oriented than next-generation data-warehouse vendor Snowflake. Because of its ability to support batch processing for ML analytics and other high-value workloads, Databricks is capturing market share from incumbents such as Alteryx, IBM, Oracle, and SAS. Databricks runs on Microsoft Azure and AWS, although those hyperscalers would likely prefer to natively support their GenAI workflows. The company gains market credit for data governance via its Unity Catalog product—a critical problem for data scientists new to AI.

Outlook

Databricks has achieved the necessary scale and growth to realize high performance in public markets. The company neared $2.0 billion in revenue in 2023 with 50% revenue growth.[9] The company will need to defend itself against ML-optimized challengers that can carry out more efficient streaming data analysis and in-memory analytics within its database. Furthermore, it will benefit from analytics built on its database becoming essential to high-value industries, including IT and industrial. The company remains innovative and active in future-proofing its business, leading us to believe that, in the long run, Databricks will be valued more highly than relational database incumbents based on the growth opportunities of data lake architecture for streaming data.

[9] "Databricks revenue, growth, and valuation," Sacra, February 2, 2024.

SELECT COMPANY HIGHLIGHTS: DATABRICKS

Financing history

	Date	Total raised	Pre-money valuation	Lead investor(s)
Series A	September 24, 2013	$14.0M	$34.0M	Andreessen Horowitz
Series B	December 31, 2014	$33.0M	$214.0M	NEA
Series C	December 15, 2016	$60.0M	$453.0M	NEA
Series D	August 22, 2017	$140.0M	$800.0M	Andreessen Horowitz
Series E	February 5, 2019	$250.0M	$2.5B	Andreessen Horowitz
Series F	October 22, 2019	$400.0M	$5.8B	Andreessen Horowitz
Series G	February 1, 2021	$1.0B	$27.0B	Franklin Templeton Investments
Series H	Date undisclosed	$1.6B	$36.4B	Counterpoint Global
Series I	Date undisclosed	$684.6M	$4.2B	Nvidia, T. Rowe Price

About PitchBook Industry and Technology Research

Independent, objective, and timely market intel

As the private markets continue to grow in complexity and competition, it's essential for investors to understand the industries, sectors, and companies driving the asset class.

Our Industry and Technology Research provides detailed analysis of nascent tech sectors so you can better navigate the changing markets you operate in—and pursue new opportunities with confidence.

Contents are based on information from sources believed to be reliable, but accuracy and completeness cannot be guaranteed. Nothing herein should be construed as any past, current or future recommendation to buy or sell any security or an offer to sell, or a solicitation of an offer to buy any security. This material does not purport to contain all of the information that a prospective investor may wish to consider and is not to be relied upon as such or used in substitution for the exercise of independent judgment.

PitchBook Data, Inc.

John Gabbert Founder, CEO

Nizar Tarhuni Vice President, Institutional Research and Editorial

Paul Condra Head of Emerging Technology Research

Additional research

Eric Bellomo
eric.bellomo@pitchbook.com
Gaming
E-Commerce

Brendan Burke
brendan.burke@pitchbook.com
Internet of Things
Information Security
Artificial Intelligence & Machine Learning

Aaron DeGagne
aaron.degagne@pitchbook.com
Medtech
Digital Health

Alex Frederick
alex.frederick@pitchbook.com
Agtech
Foodtech

Jonathan Geurkink
jonathan.geurkink@pitchbook.com
Supply Chain Tech
Mobility Tech

Kazi Helal
kazi.helal@pitchbook.com
Biopharma
Pharmatech

Derek Hernandez
derek.hernandez@pitchbook.com
Enterprise SaaS
Infrastructure SaaS

Ali Javaheri
ali.javaheri@pitchbook.com
Emerging Spaces

Robert Le
robert.le@pitchbook.com
Insurtech
Crypto

John MacDonagh
john.macdonagh@pitchbook.com
Carbon & Emissions Tech
Clean Energy Tech

Rebecca Springer
rebecca.springer@pitchbook.com
Healthcare Services
Healthcare IT

Rudy Yang
rudy.yang@pitchbook.com
Enterprise Fintech
Retail Fintech