PitchBook

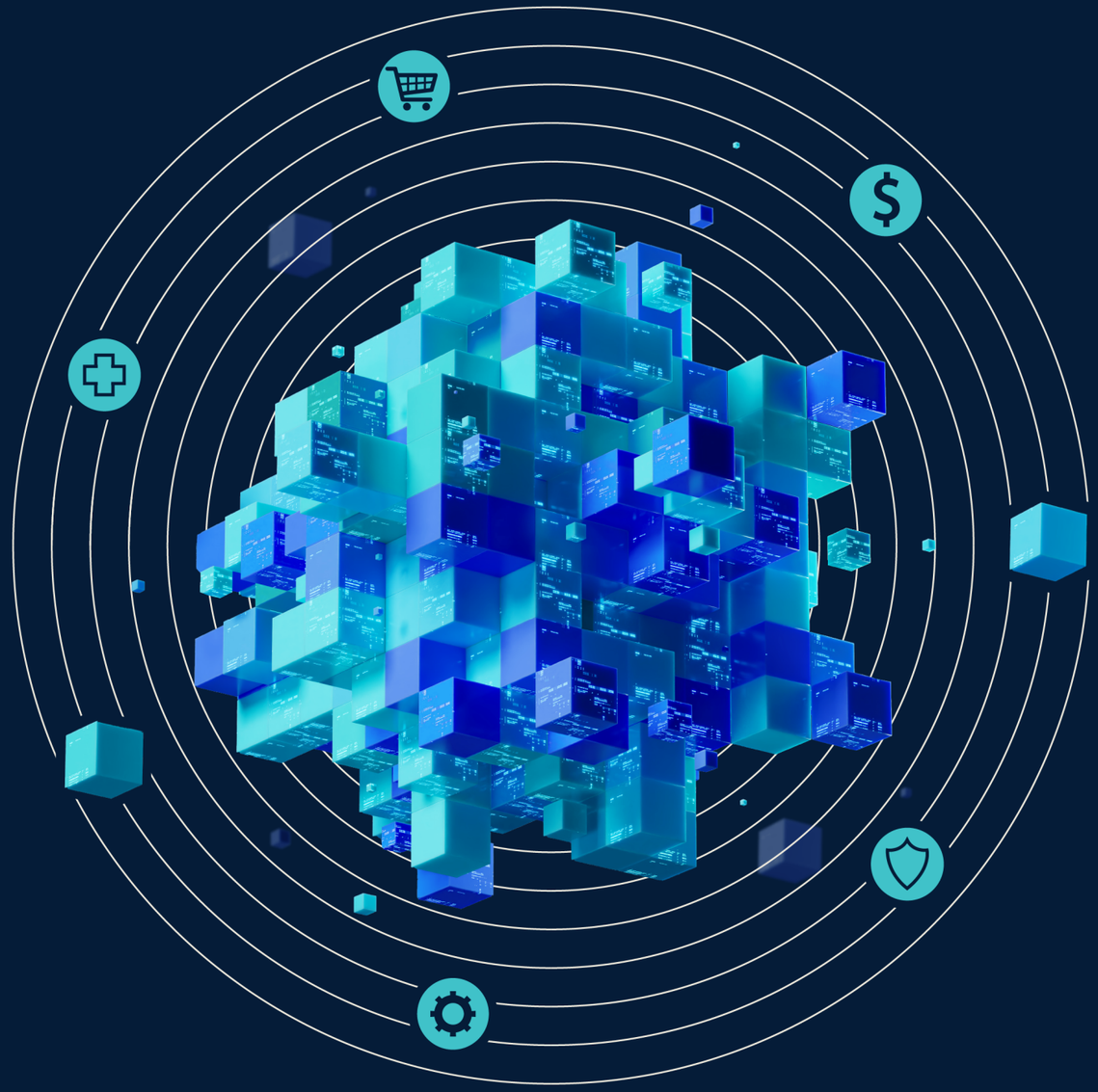
EMERGING TECH RESEARCH

Artificial Intelligence & Machine Learning Overview

Industry and taxonomy update with latest VC activity

2024

Published on April 18, 2024



Contents

This report is an annual overview of the AI & ML industry. For a more granular perspective on the AI & ML industry, please see our latest quarterly edition.

Institutional Research Group

Analysis

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

Data

Matthew Nacionales Senior Data Analyst

Publishing

Report designed by **Jenna O'Malley** and **Julia Midkiff**

Published on April 18, 2024

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

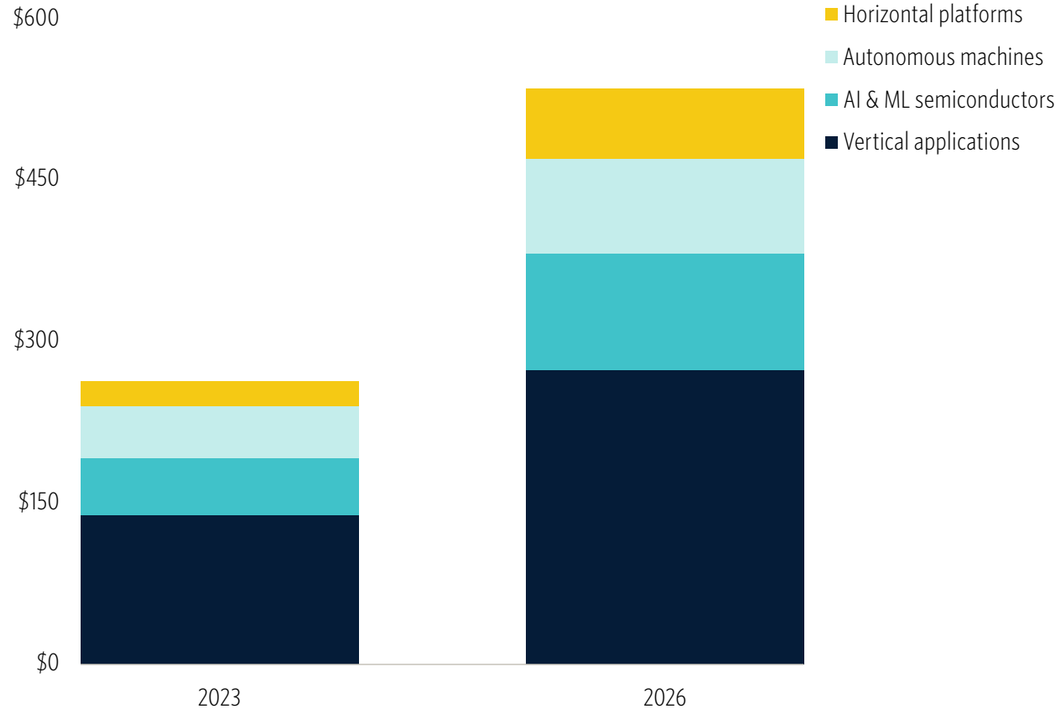
Vertical overview

Innovation in AI justifies reimagination of both the IT hardware landscape and the addressable market of software. AI leaders and private equity firms find AI datacenters to be among the leading investment opportunities. The combination of advanced hardware and intelligent software makes it possible for AI to subsume conventional software by delivering intelligent outcomes while avoiding traditional delivery mechanisms like graphical user interfaces or individual workstations. AI can be delivered via conversational devices or predictive digital experiences, or can abstract processes into a third-party cloud. We are beginning to see how AI can supplant complex workflows of people-based processes that involve basic communications, repetitive tasks, and Microsoft Office functions while enabling more strategic thinking. This process can extend across segments of AI including robotics and consumer electronics.

This report summarizes opportunities across the four segments defined in our analyst-curated vertical: horizontal platforms, vertical applications, AI & machine learning (ML) semiconductors, and autonomous machines. In this update, we add graphics processing unit (GPU) cloud as a category to the AI & ML semiconductors segment given the recent development of alternative cloud startups. Our generative AI (GenAI) taxonomy stands separately as a bottom-up framework of foundation model use cases. This report shows the opportunities in each top-down category, enabling comparison of hardware and software.

We estimate the AI & ML market reached $262.5 billion in 2023, led by the vertical applications segment, along with significant contributions from semiconductors and autonomous machines. We added new market research into foundation model application spending, a new category that reached only $1.5 billion last year but will be the fastest-growing category in AI, forecast to grow at a 119.7% CAGR to 2026. This trend promises to make horizontal platforms a fast-growing segment of the market, even as it has fallen short of promises to create independently large businesses for direct access to innovative AI algorithms.

AI & ML market size estimate ($B) by segment*



Source: PitchBook Emerging Tech Research • Geography: Global • *As of March 31, 2024

AI & ML landscape

1. Horizontal platforms
2. Vertical applications
3. AI & ML 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 · ThoughtSpot · DRIVENETS · SingleStore · DEVO · Icertis · BigPanda · harness · workato

AI core

DataRobot · THEODORA · data iku · 达闼 · scale · neo4j · DATASTAX · sparkcognition · Vercel · Weights & Biases

Foundation models

OpenAI · ANTHROPIC · databricks · Inflection · MISTRAL AI_ · ALEPH ALPHA · cohere · ADEPT · Hugging Face · AI21labs

Computer vision

trax · MEGVII 旷视 · BLACK SESAME TECHNOLOGIES · ffock 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 · Ovia · motive · 斑马智行 Powered by AliOS · G7 · CAMBRIDGE MOBILE TELEMATICS · PATEO · SMARTDRIVE

3 AI & ML 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

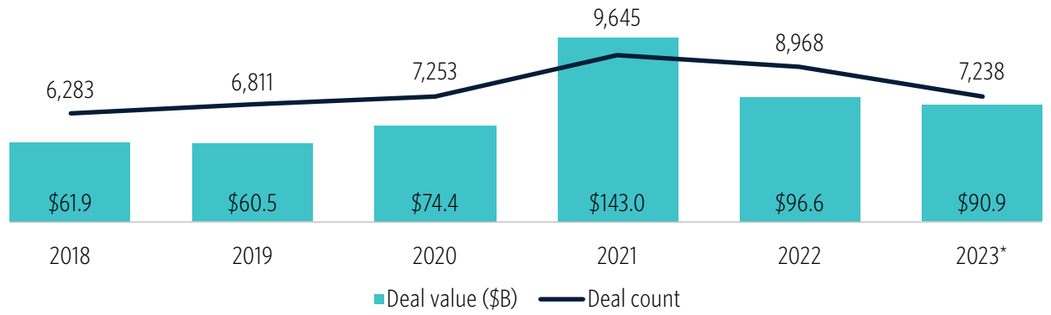
ANDURIL · nuro · Shield AI · Skydio · Geek+ · AGILE ROBOTS · GreyOrange · EPIRUS · COWAROBOT 酷哇机器人 · opentrons

VC activity

VC activity has remained consistent throughout the VC downturn since Q3 2022, leaving aside the tech giant megadeals that commit upfront funds for future cloud purchases. We tracked $22.3 billion invested in Q4 2023, including an outlier $2.0 billion deal size for Anthropic, led by Alphabet. Median valuations have remained above the market midpoint, while equity sold has fallen lower, demonstrating founder-friendly conditions. Deal count has remained much lower than predownturn levels, even given the excitement about GenAI 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, demonstrating the value of basic innovation.

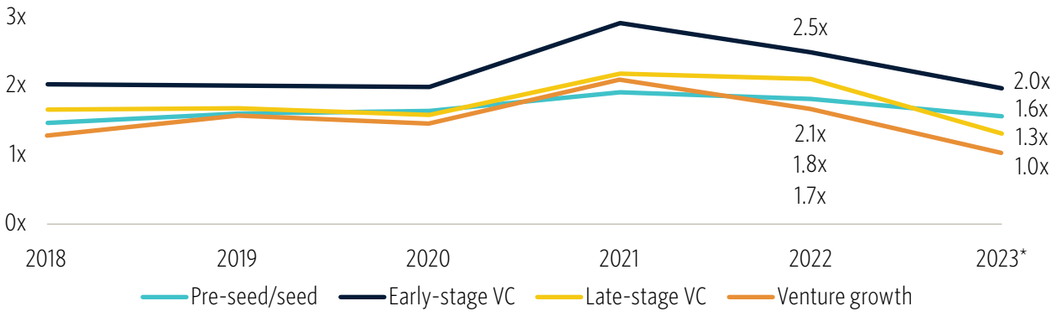
Q3 2023's bump in M&A did not sustain into Q4, as both deal value and count fell. We tracked only $2.7 billion in disclosed VC exit value in Q4, which will likely settle as the lowest quarter since Q1 2019. Tech giants remained dormant in M&A given their focus on partnerships with leading large language model (LLM) startups. Exceptions include NVIDIA's pending acquisition of Run:ai, AMD's acquisition of nod.ai in ML operations (MLOps), IBM's acquisition of Manta in database management, and ServiceNow's acquisition of UltimateSuite in predictive analytics. The recent successful listing for AI connectivity hardware company Astera Labs will encourage other companies riding the GenAI tailwind to go public. More broadly, impressive results for AI companies in public markets should encourage further listings, although our review of the IPO pipeline for 2024 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

Median AI & ML VC pre-money valuation step-up by stage



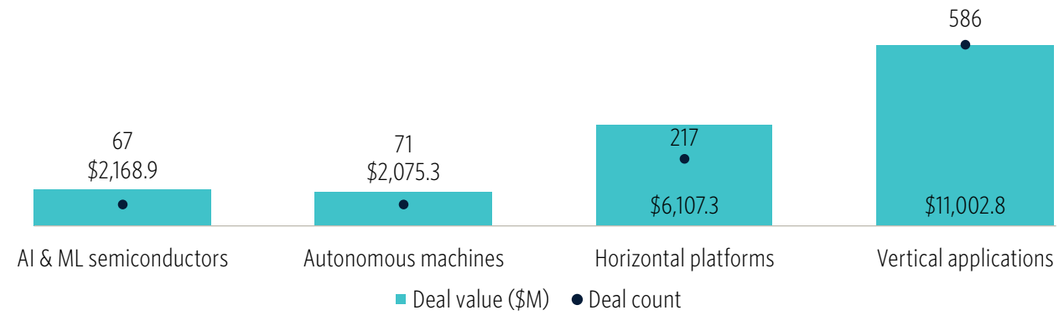
Source: PitchBook • Geography: Global • *As of December 31, 2023

Key AI & ML VC exits by exit value in Q4 2023*

Company	Close date (2023)	Segment	Category	Exit value ($M)	MOIC	Exit type	Acquirers(s)
Clearpath Robotics	October 2	Autonomous machines	Intelligent robotics, industrial robots	$454.5	$454.5	Acquisition	Rockwell Automation
LeddarTech	December 22	AI & ML semiconductors	Intelligent sensors & devices	$259.0	$0.0	Public listing	Prospector Capital
Hivestack	December 12	Vertical applications	Consumer AI, ad targeting	$100.0	$100.0	Acquisition	Perion Network
Ponder	October 24	Horizontal platforms	AI automation platforms, database management	$46.0	$46.0	Acquisition	Snowflake
Tessian	December 19	Vertical applications	AI in IT, information security automation	N/A	N/A	Buyout	Proofpoint, Thoma Bravo
UltimateSuite	December 18	Horizontal platforms	AI automation platforms, predictive analytics	N/A	N/A	Acquisition	ServiceNow
Heap	December 7	Horizontal platforms	AI automation platforms, predictive analytics	N/A	N/A	Acquisition	ContentSquare
WireWheel	December 4	Vertical applications	AI in IT, information security automation	N/A	N/A	Acquisition	Osano
Nod.AI	October 19	Horizontal platforms	AI core, model development tools	N/A	N/A	Acquisition	Advanced Micro Devices
Sisu Data	October 16	Horizontal platforms	AI automation platforms, predictive analytics	N/A	N/A	Acquisition	Snowflake

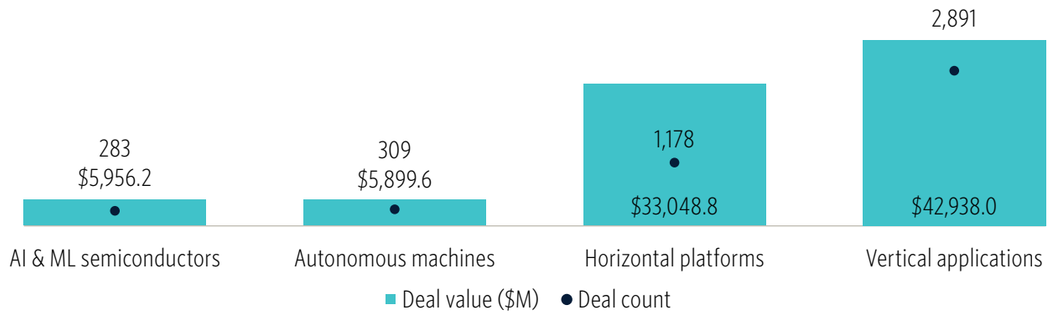
Source: PitchBook • Geography: Global • *As of December 31, 2023

Q4 2023 AI & ML VC deal value ($B) by segment*



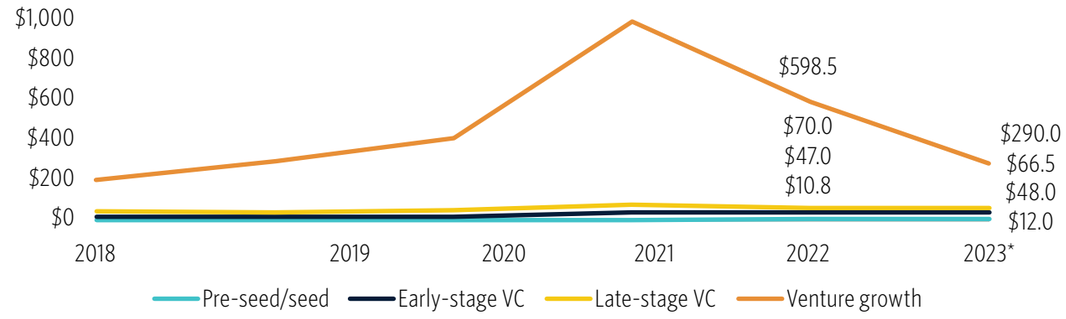
Source: PitchBook • Geography: Global • *As of December 31, 2023

TTM AI & ML VC deal value ($B) by segment*



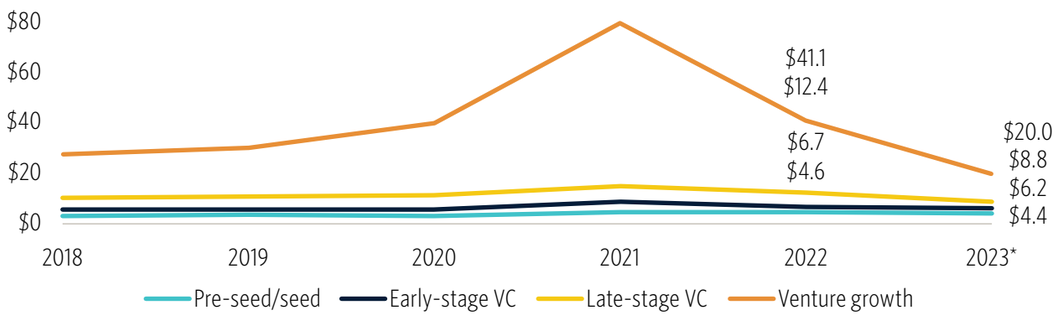
Source: PitchBook • Geography: Global • *As of December 31, 2023

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



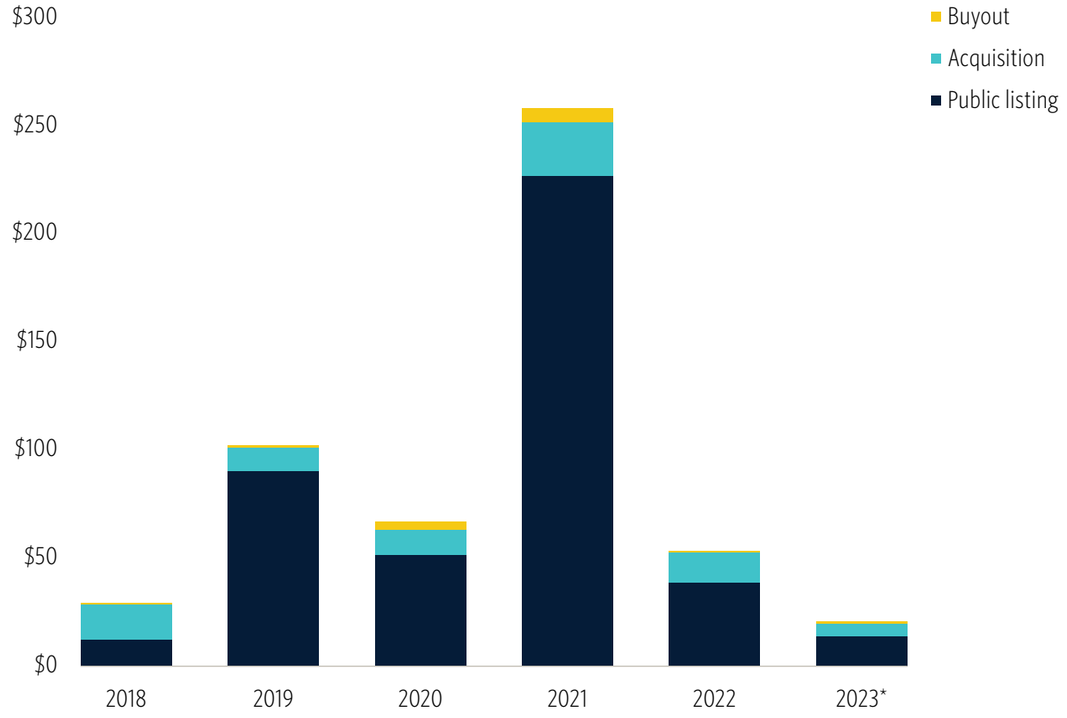
Source: PitchBook • Geography: Global • *As of December 31, 2023

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



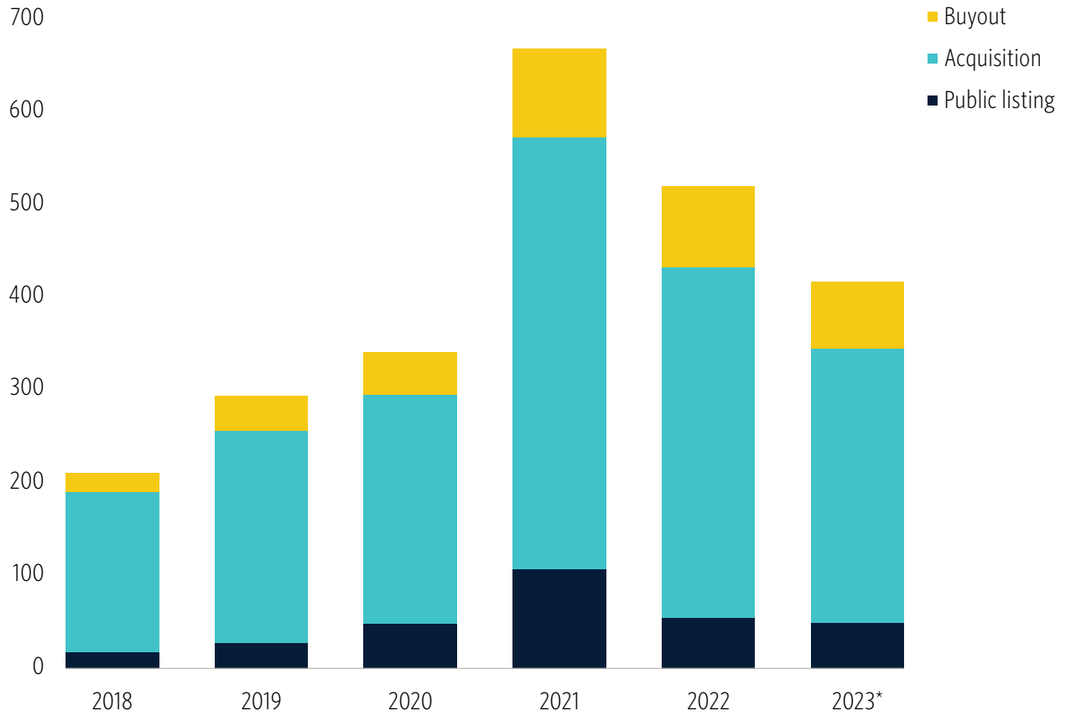
Source: PitchBook • Geography: Global • *As of December 31, 2023

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



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

AI & ML VC exit count by type



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

Segment overview

Horizontal platforms

Recent progress by research startups and tech giants suggests foundation model competition will become more intense after OpenAI took an early lead in the field.

Vertical applications

Custom AI models show promise in applications including coding agents, legal automation, and industrial analytics.

Autonomous machines

Newly open-sourced robotics datasets encourage development of foundation models for self-programming machines.

AI & ML semiconductors

GenAI hardware constraints encourage innovation in customer-sided inference semiconductors and silicon photonics.

Horizontal platforms

Overview

Horizontal platforms empower end users to build and deploy AI & ML algorithms across a variety of use cases. These platforms directly apply scientific advances in AI & ML research to commercial applications. Companies in this segment have differentiated AI & ML approaches and are built with AI & ML from the ground up—this is also referred to as AI-first. Furthermore, some horizontal platforms are used to improve AI & ML algorithms but do not use AI & ML themselves.

Subsegments include:

- **AI core:** Building blocks of AI & ML deployments, including developer tools needed to build and deploy models to production. Categories within this subsegment include AI as a service (AIaaS), AI & ML developer tools, AI platform as a service (PaaS), automated ML (autoML), cognitive computing, data preparation platforms, quantum AI, and TinyML.

- **Computer vision:** The use of AI & ML to analyze visual data and make meaningful predictions about both the physical world and digital images. The technology can be used across use cases to label and make predictions about visual data. Key products utilizing computer vision across a range of verticals include AI-enabled augmented reality, computer vision as a service, facial recognition, geospatial analysis, and visual data labeling software.

- **Foundation models:** Foundation model companies train custom neural networks using generalizable techniques including transformers, diffusion models, and custom agent architectures. Companies in this space incur substantial computing costs to compete on academic benchmarks and commonly have artificial general intelligence (AGI) as a long-term goal.

- **Natural language technology (NLT):** NLT uses computational linguistic techniques to learn from communications data and make predictions about the structure and content of language. Categories within this subsegment include conversational AI, neural machine translation, natural language generation, natural language processing (NLP), and natural language understanding.

- **AI automation platforms:** Software and services that enable enterprises across all verticals to leverage AI to automate critical business processes via predictive analytics. Categories include AI-first applications of the following products: AI for IT operations (AIOps), business intelligence, contract lifecycle management automation, database management, decision intelligence, and intelligent process automation.

Industry drivers

Costs of foundation model training coming down: GPT-4 was highly costly to train as of its launch in 2023, with estimates ranging around $100 million for a single training run. Recently, Databricks trained a state-of-the-art open-source foundation model for $10 million, though with a fraction of the parameters of GPT-4. The model is not meant to compete with GPT-4 on question answering, but can perform well relative to other open-source models for enterprise tasks.

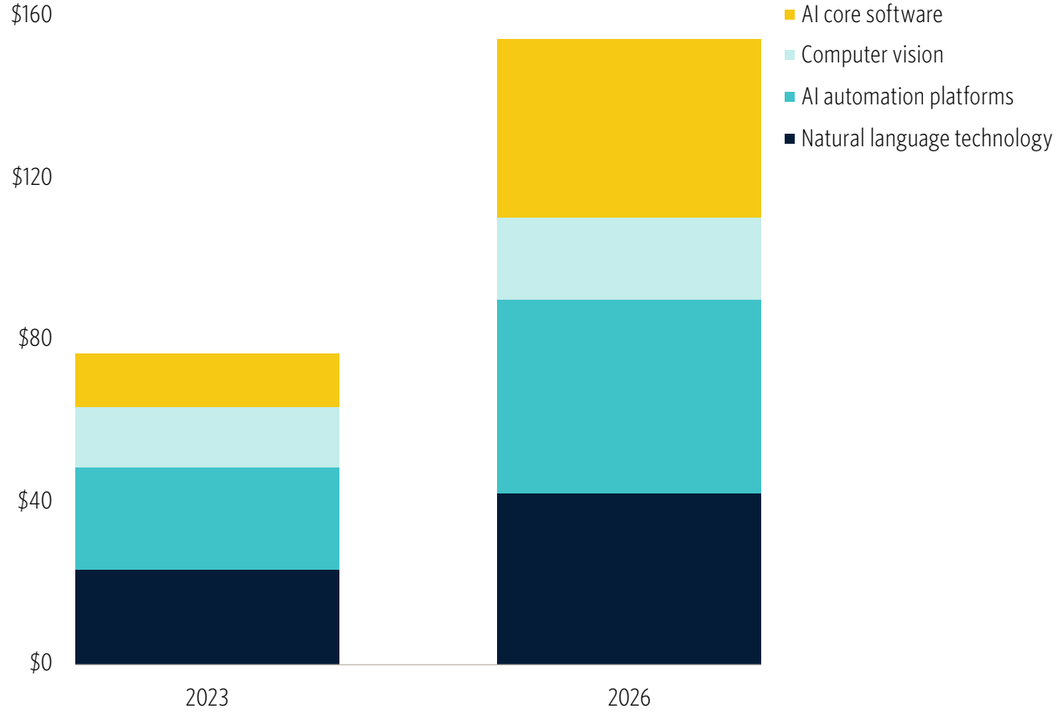
Most developers are now using AI: Five years ago, only 26% of developers had experience using AI, according to an Evans Data survey. Now, 65.2% of developers use AI in their development work.[1] This applies to both code generation and application development, with developers drawing on open-source AI models along with operations tools to build custom applications. AI open-source packages are among the most popular on GitHub and in data science libraries.

1: "Global Development Insights Webinar: AI and Machine Learning Across the Development Landscape," Evans Data Corporation, March 26, 2024.

Market size

We estimate the horizontal platforms market comprises a $60.6 billion market across the software categories in our taxonomy. This includes estimates for spending on horizontal computer vision and natural language software technology, which we estimate reached $14.8 billion and $23.6 billion, respectively, in 2023. AI business intelligence is currently a $16.1 billion market, including ML models built into business analytics software. AI core software that enables custom model development will grow faster than the business intelligence niche, with both AI lifecycle platforms like AWS SageMaker and API-based foundation model inference contributing to a $12.9 billion market in 2023 that we forecast to grow at a 50.6% growth rate out to 2026. We have not forecast this outlier growth in the past, yet vendor guidance suggests that horizontal platform usage will increase more rapidly than in the past as successful pilot projects are scaled up. Intelligent process automation will also benefit from enhanced AI agents, growing at 32.3% rate over the forecast period.

Horizontal platforms market size estimate ($B)*



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

Business model

Horizontal platforms typically contain a variety of computational costs, including the following components for a typical hyperscaler LLM deployment:

Infrastructure costs:

- **Storage:** Data is typically stored in a cloud server and carries fixed costs per gigabyte based on the scale of the data. Embedding models to create vectors from user data is much cheaper than LLMs and can be used as an initial step to do retrieval-augmented generation or fine-tuning.
- **Network:** Data moved into and out of the cloud environment is charged on a volume-based rate.
- **Compute:** Model building includes separate costs for building, training, and deployment. These charges can be assigned via spot or reserved pricing.

Operational costs: Conversation with LLMs incurs different rates for inputs and output tokens, with output tokens doubling the price of inputs. Model training carries higher charges per hour of computing used.

Security & compliance costs: AI & ML data must have an audit trail for various data privacy compliance frameworks.

Each of these components has separate costs for both training and deployment. As a result, we estimate that training a single model of GPT-4's scale can cost around $100 million in compute resources. Optional incremental costs include breach discovery and log management, maintenance and support costs, data encryption, and secure web gateways. Startups can unbundle hyperscaler platforms and provide data preprocessing and monitoring on top of the hyperscaler's platform.

Opportunities

Foundation models: We track 73 companies developing competitive foundational LLMs or competitive approaches to AGI. This list is headed by tech giants, along with Anthropic, OpenAI, and Mistral. Hyperscalers and tech giants have continued to align with startups, leaving Apple to partner with Google and develop internal models. Other private companies have shown that training foundation models can produce a competitive product offering, including Abacus.AI and Databricks, both of which recently released open-source LLMs. The challenge and costs of competing on technical merits has been demonstrated by the leadership changes at Inflection AI and Stability AI, both of which had their CEOs move on.

Recent progress by research startups and tech giants suggests foundation model competition will become more intense after the leadership of GPT-4 for a calendar year. The rapid release of the second tranche of Amazon's $4 billion commitment for Anthropic in Q1 2024 reinforces the view outlined in our Q4 2023 AI & ML Report that Amazon's support for Anthropic is a differentiator for the foundation model leader. Anthropic has established an enterprise customer base through collaboration with AWS and channel partners including Accenture and forecasts 325% revenue growth in 2024, standing out from other OpenAI competitors.[2] Concerns about Anthropic's model performance have been alleviated by the launch of Claude 3, which went to the top of research team LMSYS Org's chatbot leaderboard, which aggregates crowdsourced opinions of model quality,

2: "Anthropic Projects At Least $850 Million in Annualized Revenue Rate Next Year," The Information, Stephanie Palazzolo and Maria Heeter, December 26, 2023.

over GPT-4. The company can stand out in enterprise deployments because of its integration with the leading public cloud, AWS, and its focus on baked-in safety without the need for volatile reinforcement learning from human feedback.

Vector databases for GenAI: Vector databases store unstructured data as mathematical formulas with similarity scores that can be interpreted by LLMs. Vector databases yielded two of the top valuation step-ups among data preparation startups in 2023: Pinecone grew its valuation by 3.9x in its Series B, resulting in a $750.0 million post-money valuation, while Qdrant grew by 2.3x to a $142.8 million post-money valuation in its Series A. These two companies show the divergence in business models, as Pinecone is closed-source and coded in C++ and Python, while Qdrant is open-source and coded in Rust. We track vendors in the space in our LLM operations market map. The most popular database vendors with vector offerings include PostgreSQL (and supporting startups such as Neon), MongoDB, Redis, and Elasticsearch.[3] Other open-source projects gaining momentum include Milvus (commercialized by Zilliz), Qdrant, and LanceDB.

Startups can stand out by bringing serverless computing to vector databases. Most vector databases have been built on container architectures to deploy database instances to the cloud. Requiring a container for each database instance can increase costs and configuration complexity. To reduce computing costs, startups such as Eto with its open-source project LanceDB have developed serverless vector databases, thus lowering costs for data storage and embeddings creation and pushing market leader Pinecone to launch a serverless product in Q1 2024. Over time, vector databases may be used to create knowledge graphs that find deterministic connections between data points, along with simple semantic relationships. Knowledge graphs can be used to produce more reliable results. In 2023, vector databases began to surpass more conventional relational databases in developer momentum with only DuckDB trending faster in terms of GitHub stars than Milvus and Qdrant as of December 2023.[4] The raw star counts of these startup databases are surpassing large incumbents, including MongoDB and CockroachDB, suggesting that the databases have sufficient awareness to drive significant commercial outcomes.

No-code analytics: Business intelligence (BI) is approaching a $21.0 billion market size in 2024, growing at rate of over 10%. The market can expand via nontechnical user engagement with enterprise data warehouses via AI interfaces. In 2023, ThoughtSpot acquired Mode for $200.0 million. Mode enables BI analysts to configure charts with natural language, thereby disrupting the market for Tableau and Power BI. ThoughtSpot integrated the visualization capabilities into its existing AI search to enable data questions in a search bar. The company also offers a standout structured query language (SQL)-integrated developer environment. The acquisition value is the highest AI BI acquisition we tracked in 2023, though the company had raised $82.2 million previously.

In 2023, AI BI startup Numbers Station raised a $12.5 million Series A with a 1.9x valuation step-up. The company removes the need for data scientists to define data schemata before building models with GenAI models, allowing business users to get more reliable data served from data warehouses. The platform has recently become available via a cloud product. In Q1 2023, Engine AI reached a $36.9 million post-money valuation in a seed extension, focusing BI on the financial sector. These innovative BI engines show that startups can structure data queries in novel ways using AI. SQL automation remains a useful feature of GenAI models, yet we believe that more holistic solutions that use AI to understand siloed datasets are needed to bring data analytics to more enterprise users.

3: "DB-Engines Ranking of Vector DBMS," DB-Engines, March 2024.

4: "Open Source Database – Ranking," OSS Insight, January 30, 2024.

Neural search: Search is the broadest enterprise application of LLMs because any business user can benefit from document retrieval. In Q1 2024, GenAI startup Glean demonstrated that GenAI application leaders can apply foundation models to earn valuation premiums consistent with leading research labs like OpenAI. Glean is a neural search company that leverages GenAI to summarize unstructured data from across enterprise knowledge silos including software-as-a-service (SaaS) applications. The company grew its annual recurring revenue (ARR) 290% in 2023 to $39 million, according to CNBC reporting. Employee count is planned to double in 2024.[5] The 56.4x ARR multiple mirrors the roughly 50x multiple that OpenAI achieved in its recent secondary financing round. Some foundation model leaders have been valued at 70x to 100x ARR in early-stage rounds, yet maturing business models encourage tempered outlooks while remaining far above conventional SaaS companies. Search is a leading GenAI use case, along with customer support, coding assistance, and sales automation, and has also yielded outstanding results for public company Elasticsearch and startups Perplexity and Pryon.

Risks and considerations

Pure-play AI companies in public markets struggling to reaccelerate revenue growth: Midsized vendors face challenges growing revenue, with only one public enterprise software company, SentinelOne, on pace for over 30% growth in 2024. Our review of 2023's revenue results and consensus analyst estimates shows few public companies improving their revenue growth outside of large companies. nCino, OpenText, and Palantir are some of the few companies expected to accelerate revenue growth in 2024. Palantir was rewarded with 34% share price growth in Q1, making it an outlier among pure-play AI core companies. Revenue growth acceleration is more common among conglomerates like Alphabet, Microsoft, and IBM, suggesting that the benefits of AI deployment are accruing primarily to large companies.

Public cloud hosts offer comprehensive capabilities for AI & ML model building and deployment: Public cloud hosts can offer end-to-end LLM operations platforms. They also feature AI marketplaces, including products such as APIs, microservices, datasets, and prebuilt algorithms, that compete with horizontal platforms. Incumbents with offerings in this space include Alibaba, Amazon, Apple, Baidu, Microsoft, Google, Tencent, and IBM. In September 2023, AWS announced general availability of Amazon Bedrock for GenAI model development. The platform enables customers to build GenAI applications in private cloud instances with security and privacy controls and became a focus for the database leader at its re:Invent conference in December. Microsoft has kept pace with new AI services based on OpenAI models offered via Azure. Most cloud customers will opt to use this functionality over startup operations software.

AI ethics and regulation: LLMs compound the black-box nature of AI & ML because of their closed-sourced nature. Models can lack clarity on critical metrics including privacy, security, ethics, and transparency, thus limiting their utility in sensitive use cases. Active regulatory efforts are ongoing from the US federal government, the EU, China's internet regulator, and the UN High Commissioner for Civil Rights. Explainability is critical to understanding ethical issues in training data bias and discriminatory model parameters but is unavailable without a ground-up focus during the requirements stage. Startups can future-proof their applications by building on explainable open-source models like research lab AI2's OLMo model.

5: "AI Startup Glean Lures Citigroup as Investor at $2.2 Billion Valuation After Revenue Quadruples," CNBC, Jordan Novet, February 27, 2024.

Vertical applications

Overview

Vertical applications in AI & ML address specific problems within industries and are not always AI-first. Many startups in this category design a solution to an industry problem using software and integrate AI & ML to optimize some part of their product. These solutions typically differentiate based on the quality of the dataset used to train the industry-specific model and the industry expertise of the data scientists identifying decision-making areas that can be enhanced by AI & ML models. As a result, many of these startups help automate specific functions within their industry but have limited ability to cross-apply their AI & ML to other industries.

Subsegments include:

- **AI & ML in financial services:** Includes technologies that embed AI & ML into existing financial services via advanced analytics, process automation, robo-advisors, and self-learning programs. Product categories include financial chatbots, intelligent banking, lending analytics, payment optimization, predictive underwriting, and robo-advisors.

- **AI in healthcare:** Includes technologies that leverage AI & ML to improve medicine and the provision of care. Product categories include AI-based drug discovery, clinical decision support, genetic analytics, healthcare administration, and personal health.

- **Consumer AI:** Includes technologies that use AI & ML to enhance B2C business models. Product categories include AI in media & entertainment, AI & ML advertising technology, Metaverse, e-commerce recommendation engines, education technology, intelligent price optimization, and smart retail.

- **Industrial AI & ML:** Includes technologies that automate industrial processes and unlock industrial data to find new efficiencies. Product categories include crop maximization, energy grid automation, geospatial analysis, heavy industry automation, Internet of Things (IoT) predictive analytics, supply chain optimization, and telecommunications optimization.

- **AI in IT:** Includes enterprise software tools that optimize specific functions typically administered by IT departments, including both back-end and front-end use cases. Product categories include government tech (govtech), human resource automation, information security (infosec) automation, IT infrastructure management, legal automation, property technology, sales & marketing automation, and software development tools.

- **AI in transportation:** This category includes intelligent fleet management and driver assistance. It excludes autonomous vehicles.

Industry drivers

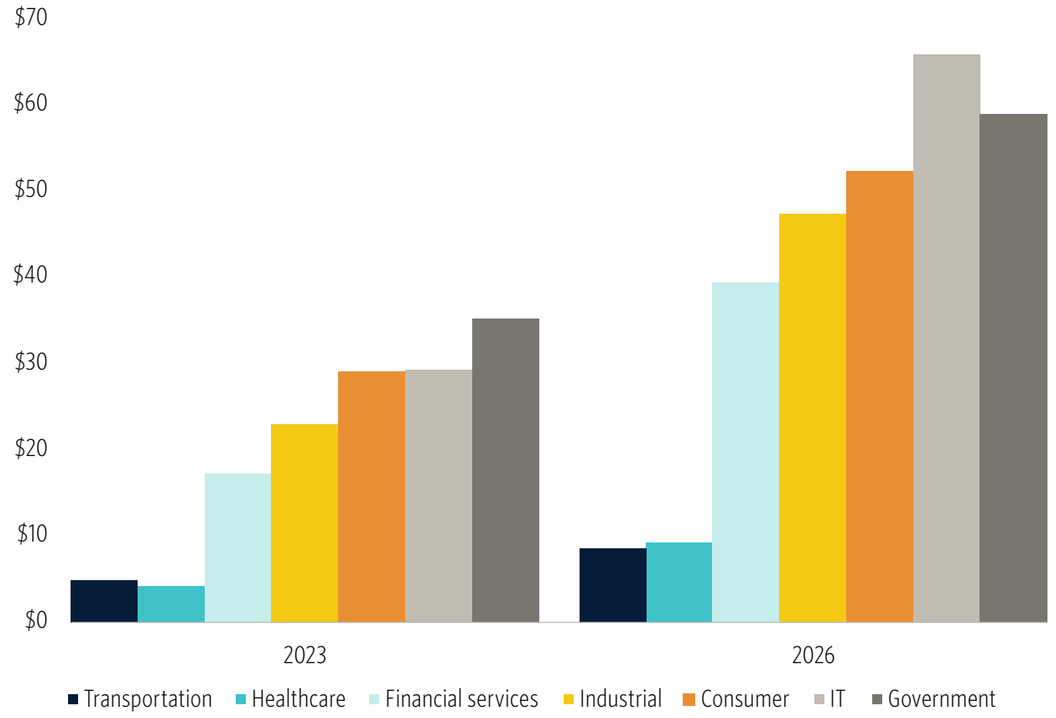
Growing availability of industrial datasets: Startups can build preliminary models based on data and frameworks from sources including Google Dataset Search, Kaggle, UCI Machine Learning Repository, Microsoft COCO, and GitHub. In Q4, Alphabet subsidiary DeepMind collaborated with academic researchers to open-source the Open X-Embodiment dataset for robotics. The project has already resulted in a new robotics foundation model with RT-X. 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. Proprietary datasets and synthetic data remain important for these new paradigms.

Enterprise customers are achieving return on investment (ROI) with AI solutions: An IDC survey sponsored by Microsoft found that organizations earn a return on AI investments within 14 months and a 3.5x ROI on average.[6] This usage occurs across industries, encouraging evaluation of specialized AI solutions in all use cases.

Market size

We estimate the vertical applications market reached $138.1 billion in 2023, significantly exceeding our prior forecast, and will grow at a 25.6% CAGR out to 2026, resulting in a $273.4 billion market. We added in data on government spending that is generally addressed by IT applications and NLP and reached $35.2 billion in 2023, making it the largest category in the segment. IT on its own is growing faster and will exceed government spending by 2026. GenAI is elevating media & entertainment as an AI category, with $19.9 billion spent on use cases like media optimization, augmented content, and customer service in 2023 as gaming and streaming services increasingly rely on AI to deliver entertainment. Healthcare and transportation remain smaller markets around $5.0 billion each, though customers take advantage of ML models available in existing cloud deployments and conventional software applications rather than procuring AI-centric software.

Vertical applications market size estimate ($B)*



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

6: "The Business Opportunity of AI," IDC, Ritu Jyoti and David Schubmehl, November 2023.

Business model

The business models of enterprise AI vendors can vary widely and have considerable influence on their cash flow profiles. Software AI startups employ several key business models, including:

- **SaaS:** A monthly subscription that bundles data engineering, model building, and deployment. Additional setup and maintenance fees may be layered on top of a base package. Applications can be delivered as a cloud service through channel partners. In most cases, managed services are required via the vendor or a third-party consulting firm.

- **Codevelopment agreements:** An AI-first vendor partners with an industry specialist to codevelop an AI solution to be used by the industry specialist as a customer and licensed to the industry specialist for distribution. Often, this includes an upfront payment to the AI-first vendor to develop the initial models. This arrangement can be necessary to supply AI-first startups with the requisite data to enter a new industry. It is a low-margin business model compared with intellectual property (IP) ownership.

- **Joint venture:** An AI-first vendor and enterprise create a joint venture to develop vertical applications and market them. Typically, the customer is given the choice to buy out the vendor at a future date.

In each case, AI-first vendors bear costs related to data engineering, model building, and maintenance, thereby creating a cost structure that exceeds those of SaaS business models. Consulting firms also arrange outcome-based or upfront capital expenses, though these are not commonly employed by startups. In each case, startups must require customers to consent to data contribution to improve their models without a royalty fee for the customer.

Opportunities

Coding agents: New AI models must be trained to automate the entire software development lifecycle. Most developers are likely using GitHub Copilot, with 82.6% of developers reporting using AI for writing code.[7] Among other coding assistants, Codeium, Cursor, and Tabnine stand out. Each of these has limitations in training data and so cannot automate the entire application development process including infrastructure code, while having limited ability to learn from existing customer codebases. At the GeekWire Summit in October, GitHub conceded the insufficiency of its Copilot project to integrate with all phases of the software development lifecycle (SDLC), including debugging, testing, and deployment. This market gap spurred startup innovation in Q4, with SDLC automations startups Factory.ai raising a $5.0 million seed round and Sweep raising a $2.0 million round. GitHub has since announced its Enterprise Copilot, which integrates with user codebases, at its Universe conference.

At the early stage, in Q1 2024, Magic reached a $500.0 million valuation in its Series B led by specialist AI investor Nat Friedman—the largest round we have tracked him leading. Friedman tested the company's custom coding foundation model and found that it outperformed other coding models on the market. Since then, Cognition Labs launched the Devin coding agent and asked for a $2 billion valuation soon after founding. The startup has made progress embedding reasoning and long-term planning in its agent and led the field in a test of resolving real-world GitHub issues. These approaches, in addition to a pending foundation model from well-funded Poolside AI, suggest that custom code models can better supplement developers than assistants from Microsoft and OpenAI.

Financial copilots: The finance industry is a leading adopter of innovative LLM techniques, encouraging startups to integrate LLMs into custom applications. While many pre-LLM AI

7: "2023 Developer Survey," Stack Overflow, May 2023.

companies struggle to maintain their valuations in the current market, AlphaSense became a unicorn in February 2023 and further grew its valuation in a September extension. AlphaSense leverages open-source LLMs for summarization of corporate disclosures. It recently launched a product feature called Smart Summaries to summarize unstructured data across earnings calls, sell-side research, and expert interviews. Using open-source models instead of closed-source APIs allows customization of LLMs using pre-existing datasets and integration with existing products rather than chat interfaces like ChatGPT. This approach shows that OpenAI's APIs are not the only way to build products with GenAI in finance.

At the early stage, Cognaize increased its valuation by 6.2x during 2023 in a Series A led by seed investor Argonautic Ventures, resulting in a $93.0 million post-money valuation. The startup's financial insights platform called Melody applies deep learning to common financial documents using a multimodal model that includes computer vision, NLP, optical character recognition (OCR), and tabular data inputs that are recursively annotated by human-in-the-loop financial analysts from customer data. The company's customers include two of the top global ratings agencies and top European banks. Financial applications have also been leading use cases for general-purpose LLM startups like LlamaIndex and LangChain.

Industrial GenAI: Industrial data analytics benefit from the existing manual processes in industrial settings and the end-to-end capabilities needed to gain insight into physical processes near the point of origin. Prominent industrial data analytics startup Seeq benefits from a partnership with hyperscaler AWS to deliver GenAI for time series analysis. The partnership allows process experts to build time series models without knowing the Scikit or Python programming languages. The startup is uniquely positioned to build these solutions given the integrations the company features with both on-premises process historians and relational databases along with industrial clouds.

Data visualizations can be generated using the latest multimodal models. C3.ai has launched industry-specific GenAI dashboards that can be customized by users across its target industries, including aerospace, manufacturing, and oil & gas. Buyers are hesitant to adopt these solutions en masse but see their benefit in retaining younger workers and reducing hiring needs.

Further, GenAI models are being applied to time series data. Nixtla, the developer of time series foundation model TimeGPT, found a suitable application of time series models in financial forecasting, wherein 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.[8] Use of these models can help industrial startups stand out by adding qualitative context to time series data and enabling more rapid and accurate fitting of models to real-world data.

Legal automation: The legal software market is large enough to support some of the most outstanding VC exits in the AI ecosystem. We have identified distinct markets across online legal content, e-discovery and forensics software, and enterprise legal management. Online legal content includes legal search repositories such as LexisNexis and Thomson Reuters' Westlaw. E-discovery helps lawyers conduct investigations as part of litigation. Enterprise legal management allows corporate legal departments to manage contracts and insource legal work. We estimate the market size across these use cases reached $12.4 billion in 2023. These markets are relatively mature, and we expect only 5.3% annual growth out to 2026. Even so, the serviceable markets have generated large outcomes in e-discovery, from companies such as CS Disco, and in online legal content, from startups such as Casetext.

8: "A Decoder-Only Foundation Model for Time-Series Forecasting," Google Research, Rajat Sen and Yichen Zhou, February 2, 2024.

Legal AI startups have raised significant capital over the past two years and have fresh momentum from GenAI. Legal foundation model startup Harvey scaled to $10.0 million in revenue soon after founding via partnerships with law firms and professional services firms including PricewaterhouseCoopers. In Q4, the company raised an $80.0 million Series B at a 4.2x valuation step-up less than a year after founding, led by AI specialist VCs Elad Gil and Kleiner Perkins. The company partnered with OpenAI to develop a custom legal foundation model and is expanding from the legal domain to other professional services.

Risks and considerations

Hyperscaler vertical solutions: Industry-specific cloud services are a strategic focus of large cloud service providers such as AWS, Google, and Microsoft. AWS offers 14 industry cloud solutions with specific storage and analytics capabilities for large industries. Supply chain has been a focus with new capabilities announced at the company's flagship conference. Contact center analytics is also a priority given the success of the Amazon Connect contact center service. Microsoft Azure offers 13 industry clouds with built-in AI based on the company's OpenAI partnership.

Excessive fixed costs: Leading GenAI companies commonly allocate 60% or more of their VC funding to computing resources. State-of-the-art generative models like Mistral AI's Mixtral can cost at least $400,000 to train, based on current GPU pricing, and are often retrained on a regular basis. Additionally, images, audio, or video data incur storage costs. Lastly, transferring trained models across cloud environments incurs transfer costs, which can flow to costs of goods sold (COGS), if model training is required for each customer engagement. Investors should question how startups utilize cloud resources in their model training and inference.

Overfit models: With large datasets, a model can be trained extensively to the point where it has too many features to make predictions on novel datasets. As a result, some models trained on quality datasets may not have "model-market fit," even if the software solution itself appeals to customers. Startups should engage in early stopping for neural network training and utilize small datasets to avoid this problem.

Privacy concerns: The retail and advertising industries are another AI & ML end market that is rooted heavily in the use of personal data. With recent lapses in protection and governance of personal data, more concerns will likely emerge surrounding data science products in consumer-centric use cases. Regulatory changes such as the EU's AI Act have far-reaching effects on how AI & ML-focused companies need to operate, especially if other jurisdictions follow suit. Adherence to these higher standards will likely add operational costs and could slow implementation in some geographies. Some AI developers are facing lawsuits for their usage of internet data. This can also affect customer behavior without the intervention of the government, as potential customers may choose an alternative because of privacy concerns.

Autonomous machines

Overview

Autonomous machines can perform tasks in human-present environments without explicit human control. These machines synthesize ML, computer vision, and datasets of the physical world such as navigation. The segment requires the design of complex hardware with software “brains” and is therefore valued differently than AI & ML software and semiconductors. Current applications of AI in robotics lack “common sense” to learn from their environments. This segment does not include optimization of existing robots, including those used in manufacturing and supply chains.

Subsegments include:

- **Autonomous vehicle software & design:** Software and hardware solutions that enable self-driving or driver assistance capabilities for cars, trucks, and other on-road vehicles.
- **Intelligent robotics & drone design:** Robotic systems and uncrewed aerial vehicles that can operate without human input. AI & ML can be used for the learning, control, and adaptation of robots. Common applications of autonomous robotics include picking up, carrying, and putting down objects.
- **Intelligent robotics & drone software platforms:** Operating systems for autonomous robots. These platforms can include fleet management and predictive maintenance.

Industry drivers

- **Regulatory momentum:** The National Highway Traffic Safety Administration is taking steps to enable wider deployment of self-driving vehicles. It is proposing the AV STEP program, which would provide an alternative regulatory pathway for companies to deploy self-driving cars without steering wheels.
- **Labor shortages:** Factories and supply chain warehouses face gaps in their workforces, thus encouraging adoption of autonomous solutions.
- **E-commerce growth driving autonomous robotics implementation:** E-commerce grew to a record 22.0% of retail in 2023 after not taking share in 2022, encouraging automation of warehouses to serve elevated demand.[9]

Market size

We estimate that the autonomous machines market reached $48.6 billion in 2023 and forecast it to grow at a 21.4% CAGR to $88.8 billion in 2026. This estimate includes commercial and delivery drones, autonomous vehicles, and autonomous robots. Most of this estimate stems from semiautonomous passenger car sales. While the industrial robotics market is larger on its own, at over $100 billion, we focus on the smaller opportunity for autonomous robots within the

9: “US E-Commerce Sales Penetration Hits New High in 2023,” Digital Commerce 360, Abbas Haleem, February 26, 2024.

manufacturing, healthcare, and supply chain industries. Autonomous driving is expected to grow at a 22.6% CAGR through 2026, which is higher than for robotics, given the challenge of displacing existing industrial robotics in manufacturing and warehouse settings.

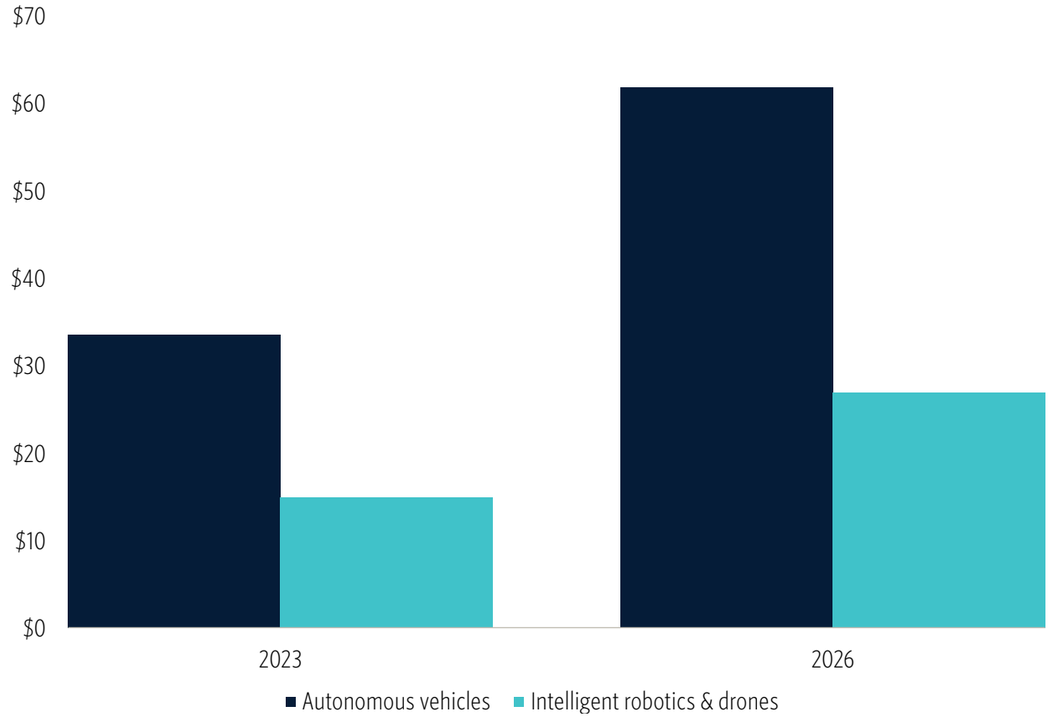
Business model

Beyond traditional hardware sales & service business models, autonomous machine suppliers are increasingly delivering robotics as a service to operate with multiple types of hardware. AI & ML algorithms trained for ML can be deployed on an edge device via software development kits and paid for via recurring revenue. This model is becoming more prevalent among incumbents and is table stakes for startups.

Opportunities

Robotics foundation models: Transformer models have been shown to improve performance on multimodal tasks including image labeling and visual reasoning. In Q4, Alphabet subsidiary DeepMind collaborated with academic researchers from 33 labs to open-source the Open X-Embodiment dataset for robotics, stitching together motion patterns and video data from 22 robot types. The project has already resulted in a new robotics foundation model with RT-1-X and an open-source model called Octo. In Q1 2024, leading robotic AI startup Covariant announced its foundation model RFM-1, an 8 billion-parameter transformer model trained on multimodal data from Covariant's real-world robotic systems operating in warehouses, including images, videos, sensor data, and language descriptions. RFM-1 can simulate realistic future video frames

Autonomous machines market size estimate ($B)*



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

based on initial conditions and planned robot actions, enabling accurate physics predictions. It can also map natural language instructions directly to robot actions, allowing intuitive human-robot collaboration. The company previously raised a $75.0 million Series C extension led by Index Ventures.

These innovations can make improvement on factory floors but also in human-present contexts. At NVIDIA's GTC conference, the company announced a foundation model called GR00T that can manipulate humanoid robots using LLMs for robotic vision. The company's software is being trialed by robotics leaders including Yaskawa, Solomon, PickNik Robotics, READY Robotics, Franka Robotics, and Universal Robots. This innovation competes against efforts from Tesla and a partnership between OpenAI and Figure. Autonomous mobile robots will likely maintain a larger market share in the medium term than humanoid robots, yet conversational interfaces will be critical for reprogramming.

AI for autonomous driving: Vehicle ML algorithms struggle to deal with edge cases including unfamiliar objects and scenarios. The first wave of startups collected much of their data from the perspective of the cars' sensors, adding real-world data over time through simulation and testing. Some edge cases, including climate, human behavior, stray objects, and the built environment, have separate datasets that can be used to generate their own models. A leading example of this is Tesla, which built a data engine to enable continuous learning from its fleet cameras. The company's data stack continuously labels fleet data to identify triggers in real-world events.

Self-driving software startups are continuing to scale through a downturn for self-driving funding. In Q1 2024, automotive AI leader Applied Intuition reached a $6.0 billion valuation with a 1.6x valuation step-up in a Series E led by existing investors and Porsche Ventures. The company benefits from the large existing market for semiautonomous passenger cars. Kodiak Robotics has innovated in sensor fusion for vision data and places a local computer on autonomous vehicles to carry out data analytics, thus removing the need for a central database. The company raised a $75.0 million round in Q3 2023. This trend benefits robotics companies more generally, as edge deployment can remove the need for high-bandwidth connectivity.

Autonomous drones: Autonomous drones can unlock new use cases in field monitoring, building monitoring, and extreme environments. Construction monitoring is an area of high spending growth, with additional connectivity becoming available in rural areas. Autonomous drones can form a central part of its growth story. We estimate that the commercial drone market reached $4.3 billion in 2023 and forecast it to grow at a 26.9% CAGR to $8.3 billion in 2026.

Drones proved uniquely capable of drawing large deal sizes in 2023. Skydio surpassed a $2.0 billion valuation in its Series E with a 1.9x valuation step-up. The company previously pivoted from consumer to enterprise drone manufacturing and has become the largest US drone manufacturer while growing its customer base to over 1,000 in a range of industries. In the defense sector, Shield AI also surpassed a $2.0 billion valuation in raising a $500.0 million Series F co-led by Riot Ventures and US Innovative Technology Fund. Shield AI develops an autopilot platform that applies across drones and fighter jets. Earlier-stage startups Exyn Technologies, Embention, vHive, PINC, and PRENAV have sensor-enabled drones that address a range of use cases, from supply chain to subterranean mine inspection.

Risks and considerations

Products killed before commercialization: The lag between product announcements and launch dates can last over a year. During this time, technical roadblocks can emerge, and cost overruns can reduce the economics of already low-margin hardware products. Robotics teams should have extensive experience with go-to-market strategies to overcome these commercialization hurdles.

Robotics research runs into roadblocks: Robotics has lagged language processing as an area of AI research, even leaving aside the hurdles of autonomous driving. Reinforcement learning has failed to move robotics into the real world due to overfitting of models on their training data, leading to a lack of generalization. Imitation learning was offered up as a potential solution by robotics research labs, yet has also faced limited success in creating general-purpose software. Now, transformer models are uniting robotics research with other AI fields, but it is unclear if next-token prediction will work well for dynamic real-world environments.

Regulatory risk: In autonomous vehicles and mobile robotics, uncertainty around regulatory frameworks may limit the growth of emerging technologies. In California, the Department of Motor Vehicles suspended self-driving company Cruise's permits, citing public safety, which led to a restructuring of the business. In October 2023, a Cruise autonomous vehicle struck a pedestrian who had previously been hit by another driver, adding to a series of safety risks. US congressional oversight of drone usage also factors into Federal Aviation Administration (FAA) rulemaking for delivery drones, with new drone standards required as part of the FAA's Reauthorization Act in February 2024.

AI & ML semiconductors

Overview

AI & ML model training and inference require hardware with maximum computational efficiency and customized processing for AI calculations. AI requires a high volume of contemporaneous calculations that run in parallel and benefit from specialized chips. Using general-purpose chips can cost thousands of times more than using an AI chip for these purposes. As a result, a segment has been developed for both the design and software-based optimization of computing hardware, including both semiconductors and sensors.

Subsegments include:

AI chips: These types of computer chips attain high efficiency and speed for AI-specific calculations. AI chip optimizations include parallel calculations, low-precision calculations to reduce transistor count, AI & ML algorithm compression, and using novel programming languages built specifically to efficiently translate AI computer code for execution on an AI chip. AI chips are faster and more efficient than central processing units (CPUs) by orders of magnitude for the training and inference of AI algorithms. Technologies within this category include:

- GPUs
- Field-programmable gate arrays (FPGAs)
- AI-specific application-specific integrated circuits (ASICs)
- Neural network processors
- Silicon photonics

GPU cloud: Startups that make advanced AI processors available via cloud services, including both datacenter hosting and third-party hardware rental.

Edge AI: Compression algorithms that optimize AI & ML models for deployment within various semiconductor environments and edge devices. Technologies in this category include:

- Post-training neural network binarization and quantization algorithms
- Quantization aware training
- Ensemble-boosting algorithms
- Edge AI APIs and software development kits (SDKs)

Intelligent sensors & devices: Devices that are optimized to run AI & ML models. This segment contains technologies including:

- **Sensors:** Devices that measure specific parameters in real-world conditions such as environmental conditions, motion, images, and chemical levels.
- **Sensor systems:** Assemblages of sensors and chipsets that empower specific IoT use cases such as smart buildings, industrial IoT, and connected vehicles.

Industry drivers

Expensive computation: The amount of compute being used in large AI training runs has been increasing exponentially. NVIDIA estimates that the compute required to train leading transformer models is increasing by 215x every two years.[10] The need to use NVIDIA GPUs due to their efficiency and configurability has pushed prices to around $40,000 per H100 chip.

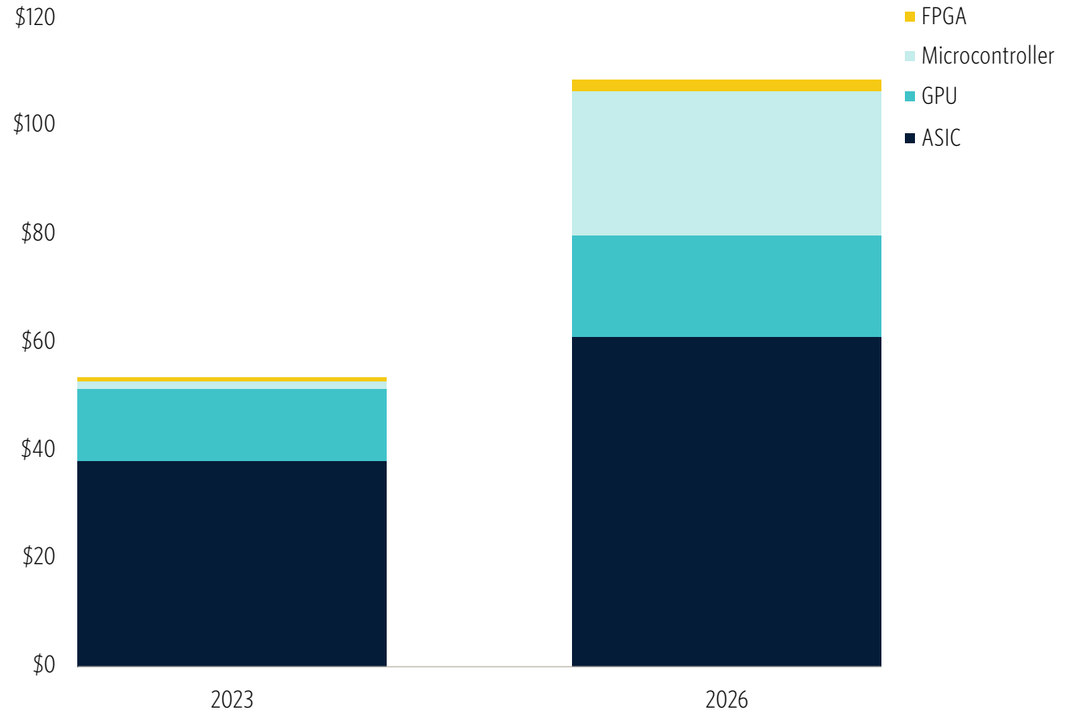
Research finding that deep learning improves at scale: The more compute that is used, along with optimal model architecture and large data training sets, the more accurate LLMs become on a wide variety of question answering tasks. This result has been empirically demonstrated by AI researchers and means that additional breakthroughs in accuracy will scale in tandem with computing power.

Demand shifting from discrete to integrated processors: AI chips have focused on the datacenter but are increasingly shifting to the edge to support emerging use cases including smartphones, tablet computers, wearable devices, and connected automobiles. This is leading to integrated processors growing faster than discrete processors.

Market size

We estimate that the AI & ML semiconductor market accelerated to reach $53.7 billion in 2023, achieving 26.2% growth over 2022. Over half of this total can be attributed to mobile phone ASICs, which will decline as a percentage going forward. We have revised our growth expectations up and expect the market to grow at a 23.6% CAGR from 2024 to 2026, resulting in an $88.8 billion

AI & ML semiconductors market size estimate ($B)*



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

10: "Company Overview," NVIDIA, February 21, 2024.

market. Our estimate includes all AI semiconductor varieties, such as field-programmable gate arrays, GPUs, microcontrollers, and ASICs. To be included in this estimate, these chips' primary function must be enabling AI processing. We forecast the highest-growth segments in this market to be field-programmable gate arrays and microcontrollers, although we anticipate all segments will grow at CAGRs over 12.0% from 2024 to 2026, based on emerging edge AI use cases.

Business model

AI semiconductors are currently deployed through unit sales and usage-based pricing from cloud hosts. For example, AWS sells usage-based instances of NVIDIA GPUs for AI & ML workloads. The usage-based pricing of GPU instances enables around a one-year payback for AWS for the underlying hardware, assuming constant usage of these GPUs and given current pricing of around $40,000 for a standard NVIDIA Tesla H100. AI & ML developers can thus gain access to GPUs without high capital expenditure.

Opportunities

GenAI semiconductors: Chip startups can create verticalized computing platforms for specific model types. Startups have struggled to stay ahead of NVIDIA's CUDA software when compiling AI models to hardware given the range of different frameworks. GenAI offers an opportunity to innovate around a single family of transformer models for a range of tasks and optimize hardware. Already, chip startups competed favorably with NVIDIA for BERT (bidirectional encoder representations from transformers) LLMs. SambaNova Systems disclosed significant speedups over NVIDIA hardware for BERT training and inference, with SambaNova particularly shining in inference latency reduction with a 5.8x latency reduction.[11] In Q3 2023, the company launched a new generation of chips configured for midsized LLMs, predicting a world where customers will be running a variety of fine-tuned models in parallel. This focus on enterprise customers can create an adjacent market to NVIDIA's dominance of hyperscaler cloud computing.

In the inference semiconductor niche, Groq is achieving scale with a customized language processing unit (LPU) inference accelerator that has been developed since the company was founded in 2016. The LPU is fabricated at legacy 14 nanometer nodes that, when combined with custom accelerator software, can achieve 10x improvements in both LLM output speed and power savings over NVIDIA hardware. Since recent announcements about record speeds on LLM tokens-per-second and a partnership with Aramco Digital, the company has achieved letters of interest (LOIs) and memorandums of understanding (MOUs) for 240,000 of its language processing units (LPUs) over the past five weeks, per our briefing with management.[12] The company forecasts 1.5 million shipments for 2025, factoring in a recent partnership with Norway-based edge datacenter host Earth Wind & Power, reaching a similar volume to NVIDIA's shipments, though with more competitive pricing per unit for alternative cloud providers. The company also counts 81,000 developers among its community, having built 21,000 applications on the company's architecture.[13] We have long believed that startups can stand out in AI inference rather than training, and this rapid scaling demonstrates that complementary hardware to NVIDIA GPUs will be needed for the rapidly growing inference market.

11: "Breakthrough Efficiency in NLP Model Deployment," SambaNova Systems, September 20, 2020.
12: Telephone interview with Jonathan Ross, CEO of Groq, by Brendan Burke, April 15, 2024.
13: Ibid.

GPU cloud: 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% to 70% cost savings on GPU hours for advanced NVIDIA A100s and offering unique access to the latest H100 chips.[14] Leading startup GPU cloud provider Lambda Labs has built the largest cluster of H100 chips of all public clouds, exceeding Google and Oracle.[15] 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 offered exclusively 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 nonlinearly if startup models become popular.

At the early stage, new cloud providers are coming online. In Q4, RunPod raised an $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 working primarily with AI startups, and the company's business model involves rental of GPUs from crypto miners. 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.

14: "Cloud GPU Comparison," Jolt, n.d., accessed February 8, 2024.
15: "State of AI Report Compute Index," Air Street Capital and Zeta Alpha, n.d., accessed February 13, 2024.

Ambient AI devices: Recent AI-enabled device releases show that small semiconductors can support edge usage of LLMs. In November 2023, Apple-alumni-founded device startup Humane launched its Ai Pin, which integrates a Qualcomm Snapdragon microprocessor to carry out speech recognition on a miniature form factor, in partnership with Microsoft and OpenAI for cloud inference. In October 2023, Qualcomm launched the Snapdragon 8 Gen 3, a premium processor designed to bring GenAI to the next wave of flagship smartphones. This processor is expected to be available in flagship Android devices. In addition to these, Qualcomm is set to introduce its latest midrange processor, the Snapdragon 7 Gen 3 for efficiency and cost-effectiveness. Though initial reviews of the device have been negative, we believe devices such as Ai Pin just scratch the surface of how new AI chips can be embedded in ubiquitous products to create connected real-world experiences.

At the early stage, Rewind and Tab have launched AI-focused necklaces that can integrate AI into daily life. Rewind records conversational data for AI search and features a wearable pendant that records audio for local processing of models including GPT-4 on iPhones. The always-on nature of its audio recording means that all the conversations that an individual has can be processed with AI, making the AI ambient in a sense. The company raised its valuation 4.7x to $350.0 million in its Series A in July 2023—less than a year after its seed round—with the CEO disclosing that the company has since been valued at $500.0 million. Recently, pre-seed wearable Tab has launched for pre-orders on social media and offers conversational life assistance based on LLMs.

Silicon photonics: This technology converts binary values into photons, which can more rapidly transfer data than digital bits and with less heat loss. Photonics can continue to improve AI training efficiency in the long run, thus creating the opportunity for large standalone chipmakers to be created in the emerging space. Photonics startups are achieving superior results on training speed for leading AI model types. Startup Lightmatter is optimizing its hardware for leading neural

network types including ResNet for computer vision, BERT for NLP, and Meta AI's deep learning recommendation model for digital personalization. The company claims to reduce training times by 10x compared with NVIDIA's A100, though these test results have not been verified by a standards body.[16] This research focus demonstrates that photonics will be used for AI training in the future, as opposed to more common current use cases including interconnect and switches.

Interconnect requirements for GenAI processors encourage the adoption of photonic processors, thereby accelerating the maturation of this emerging field. In Q1 2024, Celestial AI achieved unicorn status with a $175.0 million Series C led by US Innovative Technology Fund. The round also featured participation from AMD Ventures, demonstrating alignment with a leading GPU maker. Celestial AI has developed an advanced optical interconnect technology called the "Photonic Fabric," which enables high-bandwidth, low-latency connectivity between compute and memory resources. The lead investor previously invested in GPU cloud startup Lambda and may have an inside view into the interconnect requirements of cutting-edge datacenters. Connectivity presents challenges for LLM builders when routing high volumes of requests to large models. Photonics can play a central role in alleviating physical constraints over the next three years.

Risks and considerations

Chip development is expensive: The cost to design a leading semiconductor at a 2-nanometer node can reach $750.0 million, including software, physical design, testing, and tape-out. Chip companies are accustomed to regularly releasing new chips, thus resulting in an astronomical burn rate for startups.

Datacenter startups not reaching scale: Startups focusing on datacenter training remain reliant on contracts with research laboratories for specific tasks. Multiple advanced datacenter startups list Argonne National Laboratory as a leading client, despite the federal entity's low budget for procurement each year relative to hyperscalers. These research-based contracts continue to take the place of significant revenue-generating contracts with hyperscalers, thereby limiting revenue growth for startups. Pilot projects completed with hyperscalers in 2021 have not evolved into publicly disclosed commercial agreements, thus leaving startups behind on their commercial milestones for future fundraises, while the datacenter AI market decelerates from its high-growth period.

Semiconductor startups may not scale without a commercially viable architecture: Some advanced semiconductor startups lack the ability to manufacture a cost-competitive chip due to a lack of commercially viable architecture. Chip designs with the potential to commercialize and compete with products on the market today must feature the following characteristics:

- Pricing of under $0.70 per giga floating point operations per second (gigaFLOPS).
- Sufficient funding to bring a chip to production.
- Interconnected architectures to optimize data flow between processing and memory.
- Compatibility for training and inference software, because some chips require separate software for both processes, which has knock-on effects for software design.
- A fabrication node at 4 nanometers or below.
- Integrated developer kits.
- Validation from cloud hosts.

16: "'It's Going To Run a Lot of the Backbone of the Internet': Lightmatter CEO Nick Harris on Its Energy-Efficient AI Chips," Tech Brew, Hayden Field, July 23, 2021.

Appendix

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.

Top strategic acquirers of AI & ML companies since 2019*

Investor	Deal count
Accenture	21
Apple	21
Microsoft	14
Cisco Systems	13
Meta Platforms	13
ServiceNow	11
International Business Machines	10
DataRobot	9
Intel	9
Snowflake	8

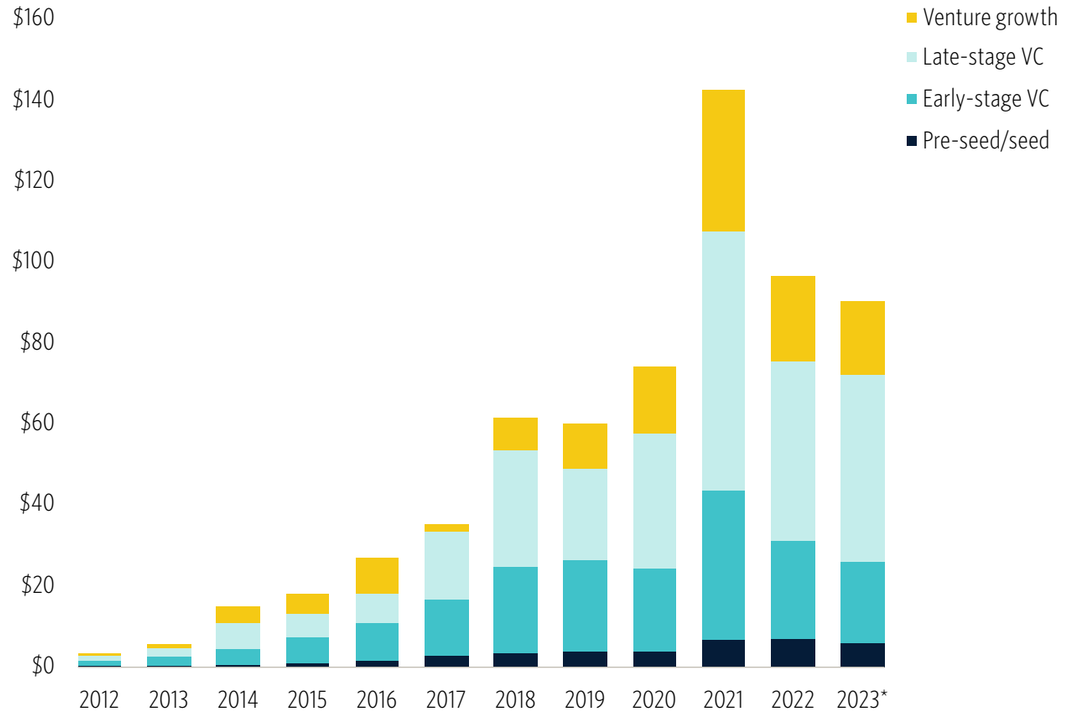
Source: PitchBook • Geography: Global • *As of December 31, 2023

Top VC investors in AI & ML companies since 2019*

Investor	Deal count
Alumni Ventures	409
Sequoia Capital	387
500 Global	358
Andreessen Horowitz	333
Accel	330
HongShan	319
Khosla Ventures	314
Right Side Capital Management	309
New Enterprise Associates	309
Enterprise Ireland	295

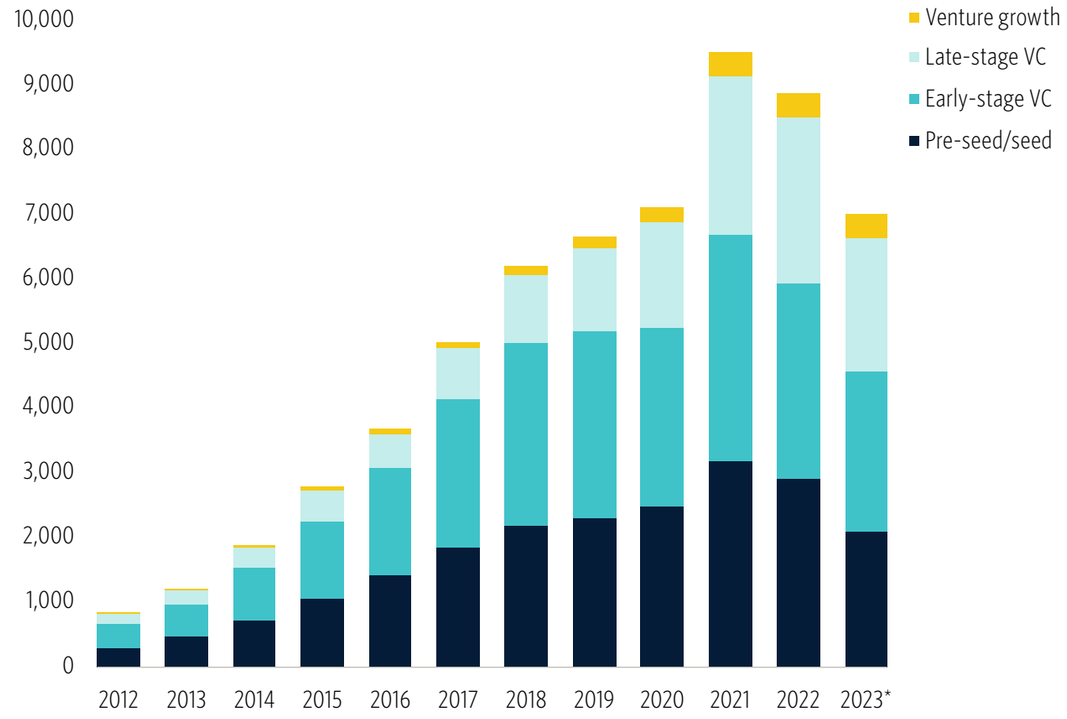
Source: PitchBook • Geography: Global • *As of December 31, 2023

AI & ML VC deal value ($B) by stage



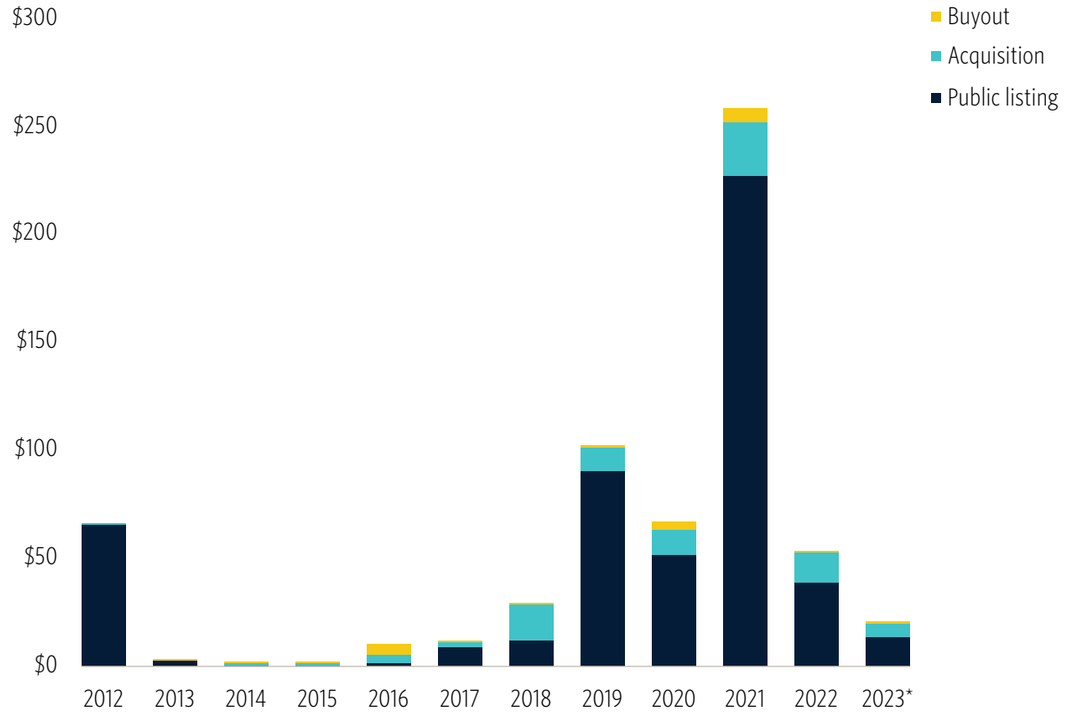
Source: PitchBook • Geography: Global • *As of December 31, 2023

AI & ML VC deal count by stage



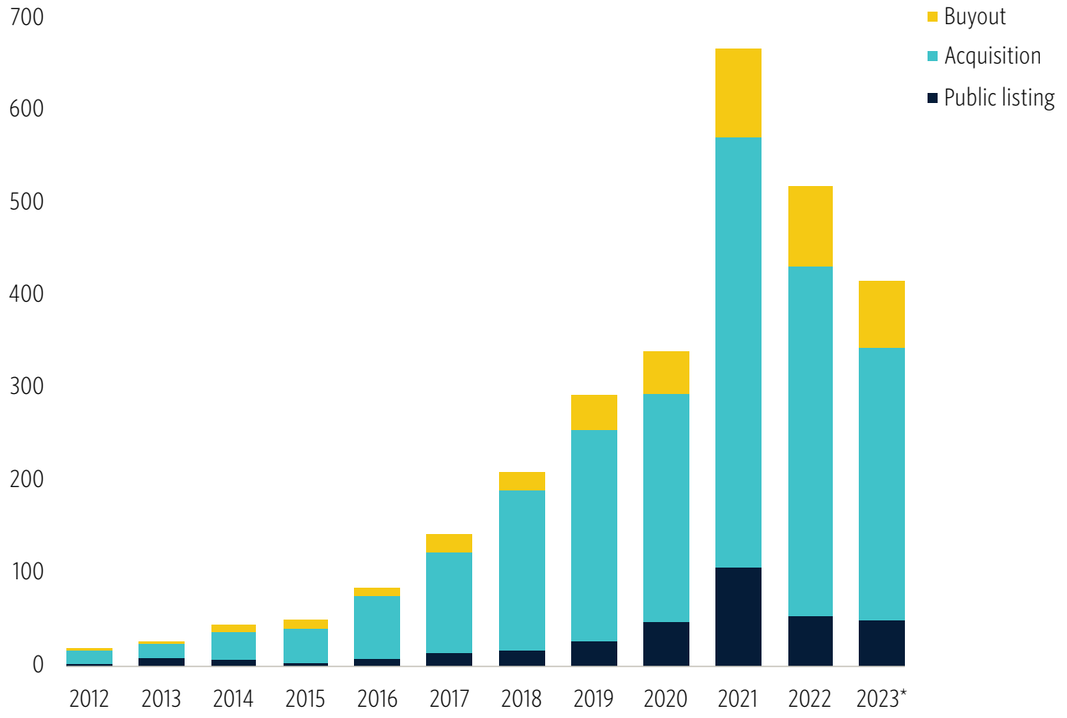
Source: PitchBook • Geography: Global • *As of December 31, 2023

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



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

AI & ML VC exit count by type



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

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.

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
Data Analytics
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