LAZARD

AI

LAZARD VGB AI INFRA 40

VGB INSIGHTS

JULY 2024

Lazard Venture & Growth Banking

Christina Frankopan
christina.frankopan@lazard.com

Winifred Awolope
winifred.awolope@lazard.com

Christopher Britton
christopher.britton@lazard.com

Amy Cozamanis
amy.cozamanis@lazard.com

Nick James
nick.james@lazard.com

Artificial Intelligence | Enterprise Software | Consumer | InfraTech | Healthcare | DeepTech | FinTech

Table of Contents

I. EXECUTIVE SUMMARY 3

II. LAZARD VGB AI INFRA 40 LIST 4

III. REPORT SCOPE AND METHODOLOGY 6

IV. KEY MARKET TRENDS 6

V. SEGMENTATION OF AI INFRASTRUCTURE 14

VI. APPENDIX I: COMPANY PROFILES 21

VII. APPENDIX II: REFERENCES 42

I. Executive Summary

We are delighted to introduce the Lazard VGB AI Infra 40 list which showcases a selection of some of the most exciting growth-stage AI infrastructure companies we have identified through hundreds of discussions with CEOs and investors.

4 key market trends emerged through our research and company interviews:

Rethinking the Stack for the AI Age
The optimal architectures of hardware, software and associated tooling appear to be shifting significantly for the AI Era and technical efficiency. Architectures will likely become increasingly driven by AI use-case (e.g., specialized for edge compute) and flexible (agnostic to both underlying chips and foundational models).

Focus on AI Sustainability
Rapidly increasing energy requirements of AI training and inference is leading to a focus on architectural energy efficiency and is increasing the need for optimized energy and AI infrastructure integration.

Looming Data Constraints
Limited supply of high-quality language data is leading to opportunities in alternative and synthetic data sources.

The Dawn of Autonomous AI Agents
Co-pilot agents are just the start of harnessing the power of LLMs for agents. We see emerging opportunities in multi-agent and autonomous agent models as a possible key theme over the next 12 months.

This report looks at AI infrastructure as a space for potential investments at the growth stage. McKinsey[1] estimates that while up to US$25.6 trillion p.a. in economic value could be added to the global economy by Generative Artificial Intelligence (GenAI) and its applications, the build-out of essential AI infrastructure is fundamental to realizing this potential: "*infrastructure is destiny*", as one OpenAI executive recently put it.[2]

The current phase of AI infrastructure growth is however being hampered by a series of stark shortages: of hardware (GPUs and other AI chips), of sustainable energy sources, of high-quality training data and of talent. We believe that technologists from many companies including our VGB AI Infra 40 list are aiming to respond with innovative rethinking of hardware, software, data and tooling stacks, increasingly led by AI use cases, including the growing need for edge compute.

II. Lazard VGB AI Infra 40 List

We are delighted to introduce the VGB AI Infra 40 list of selected companies arranged alphabetically by segment. Detailed profiles of the VGB AI Infra 40 companies are available in the Appendix.

#	Companies	Segment	Sub-segment	Raised to Date ($m)	Last Valuation ($m)	HQ Location
1	AXELERA ARTIFICIAL INTELLIGENCE	Hardware / Silicon	Edge/Embedded AI	$130	-	Netherlands
2	AyarLabs	Hardware / Silicon	I/O/Networking	$219	$500	US
3	celestial AI	Hardware / Silicon	I/O/Networking	$339	$1,175	US
4	CORNAMI Intelligent Computing	Hardware / Silicon	Specialized workloads	$124	$244	US
5	CORNELIS NETWORKS	Hardware / Silicon	I/O/Networking	$127	$228	US
6	d-Matrix	Hardware / Silicon	Inference Accelerator	$161	$385	US
7	enfabrica	Hardware / Silicon	I/O/Networking	$175	$295	US
8	Lightelligence	Hardware / Silicon	Photonic Compute	$232	-	US
9	LIGHTMATTER	Hardware / Silicon	Photonic Compute	$421	$1,200	US
10	NEXTSILICON	Hardware / Silicon	HPC Supercompute	$300	-	US
11	SiMa.ai	Hardware / Silicon	Edge/Embedded AI	$270	$770	US
12	SIPEARL	Hardware / Silicon	HPC Supercompute	$131	-	France
13	UNTETHER AI	Hardware / Silicon	Inference Accelerator	$154	-	Canada
14	Lambda	Hyperscalers & Compute	Compute-as-a-Service	$892	$1,520	US
15	NEXGEN CLOUD	Hyperscalers & Compute	Compute-as-a-Service	$13	$185	UK
16	together.ai	Hyperscalers & Compute	Compute-as-a-Service	$233	$1,250	US
17	anyscale	Model Serving & Inference	Optimization / Acceleration & Hosting	$260	$1,014	US
18	Modular	Model Serving & Inference	Optimization / Acceleration & Hosting	$130	$600	US
19	OctoAI	Model Serving & Inference	Optimization / Acceleration & Hosting	$133	$850	US

Source: PitchBook Data, Inc.; Public Sources; Lazard VGB Insights

Lazard VGB AI Infra 40 List

#	Companies	Segment	Sub-segment	Raised to Date ($m)	Last Valuatio n ($m)	HQ Location
20	ABACUS.AI	MLOps	AI/ML Platforms	$90	$374	US
21	acceldata	MLOps	Monitoring and Observability	$106	$360	US
22	ANACONDA.	MLOps	Model Development Tools	$77	-	US
23	DOMINO DATA LAB	MLOps	AI/ML Platforms	$224	$800	US
24	EDGE IMPULSE	MLOps	Model Development Tools	$54	$229	US
25	FeatureBase	MLOps	Feature Engineering	$30	$51	US
26	fiddler	MLOps	Monitoring and Observability	$45	-	US
27	gretel	MLOps	Synthetic Data	$68	$335	US
28	HIDDENLAYER	MLOps	Model Security	$56	$200	US
29	Human Signal	MLOps	Data Labeling	$30	$125	US
30	LatentAI	MLOps	Model Development Tools	$31	$52	US
31	Lightning AI	MLOps	AI/ML Platforms	$62	$290	US
32	MAD MAD STREET DEN	MLOps	Data Preparation	-	-	US
33	MOSTLY·AI	MLOps	Synthetic Data	$31	-	Austria
34	PROTECT AI	MLOps	Model Security	$49	$110	US
35	RelationalAI	MLOps	Data Preparation	$122	$595	US
36	rescale	MLOps	Model Development Tools	$157	-	US
37	Snorkel	MLOps	Data Labeling	$138	$1,000	US
38	unravel	MLOps	Monitoring and Observability	$128	-	US
39	Unstructured AI	MLOps	Data Preparation	$68	$223	US
40	Weights & Biases	MLOps	Model Development Tool	$265	$1,250	US

Source: PitchBook Data, Inc.; Public Sources; Lazard VGB Insights

III. Report Scope & Methodology

Since publishing our work on AI commercialization and go-to-market strategies in mid-2023 (see report here), the Lazard VGB Insights team has spoken to 100+ CEOs of private AI companies, across the stack.

This report focuses on our selected AI infrastructure growth-stage companies and a complementary report on enterprise-ready **Horizontal and Vertical AI Applications** will follow at a later date.

Our screening methodology can be summarized as follows:

1. We utilized extensive public information, industry reports and databases such as PitchBook to screen 2,000+ companies in our core markets of North America and Europe which we then identified for detailed research and interview.
2. We only included companies categorized as "AI-centric", meaning AI functionality was assessed to be core to the company's native architecture or product offering. We excluded companies employing single-use features or enhancements.
3. Companies considered by us to be incumbent or dominant players were not included, as well as foundational model builders, such as Anthropic or OpenAI, which we see as a maturing market segment.
4. We limited our selection to companies with known valuations under US$1.5 billion.

IV. Key Market Trends

Introduction

The immense potential of the AI Era has been catalyzed at an astonishing pace in the 18 months since the launch of ChatGPT. Yet it can be challenging for observers to read signal versus noise amidst the use of both superlatives and skepticism. Indeed, a recent survey[3] indicated that experts are split on whether AI-related public stocks are in a bubble: 40% said "Yes", 45% said "No".

It could be that both are correct. The investor Bill Janeway (who recently delivered a keynote at our VGB T500 Conference in London) argues that not all bubbles are alike, and that in "*productive bubbles*", speculation attaches itself to transformational general-purpose technology but with the real potential to create new economies.[4]

Boom or Bubble?

"Artificial intelligence and generative AI may be the most important technology of any lifetime."

Marc Benioff, CEO Salesforce

"A bubble within a bubble...is totally unprecedented. The best guess is that this [AI bubble]will at least temporarily deflate."

Jeremy Grantham, GMO

Textbox Sources[7]

However, the potential opportunity is too large to ignore. Generative AI has catalyzed a step-change and represents a reported[5] US$1.3 trillion market opportunity by 2032, potentially quadrupling to as much as 12% of total global technology spend. Including the impact of new use cases and productivity gains, McKinsey estimates GenAI could add upwards of US$25.6 trillion of economic value per annum to the global economy.[6]

- While there are still notable challenges to adoption, we are seeing significant enterprise-pull for use cases which is "*mostly top down, coming from CEO/C-suite or a steering committee*", according to 71% of recently surveyed executives.[8]
- More than 80% of enterprises are expected to have used GenAI by 2026, up from less than 5% in 2023.[9]
- Almost all the companies in our AI Infra 40 list have paying marquee enterprise customers, typically on multi-year recurring contracts and deploying in production. We have, however, included a few pre-revenue companies by exception.

As outlined in our previous report on AI commercialization, we note a continued trend for VC-backed AI companies to aim to reach large audiences through monetized multi-year strategic partnerships, often backed by equity investment.

- Recent examples of such partnerships since our previous report include those between Amazon/ Anthropic, Snowflake/Mistral, and Google/HuggingFace.[10]
- We note, however, that the Big Tech partnership approach may have limited runway due to antitrust concerns. The US Justice Department and the FTC have reportedly agreed an approach to potential investigations into Nvidia, OpenAI and Microsoft (including the recent Inflection AI transaction)·[11]

While the US received the vast majority of VC investment with 70% of the total in 2023[12], and 3x more than Europe[13], we note emerging vibrant ecosystems in Canada (Toronto and Waterloo), UK and France.

- Although outside the detailed scope of this report, we note that moves towards "*Sovereign AI*"[14] by some countries in the interests of cultural and linguistic preservation, economic growth, talent development, and cybersecurity may materially impact the funding environment in coming years.
- For more on the Geopolitics of AI, see this October 2023 report from Lazard Geopolitical Advisory.[15]

> *"US$1 trillion worth of current equipment in data centers would have to be replaced with AI chips."*
>
> JENSEN HUANG, NVIDIA

The energy and AI infrastructures to satisfy escalating potential demand is being built out today for the AI Era of tomorrow. While Nvidia has captured a dominant market position in the making of specialized chips for running generative AI models in the first phase of the AI Era, we believe that the broader AI infrastructure sector, or "picks and shovels", could represent a significant investment opportunity, estimated to be valued at US$309 billion by 2031.[16] The second phase of AI evolution will likely involve a broad range of companies building specialized AI-related infrastructure, across software and hardware, AIOps (automating IT systems using ML and big data), MLOps (standardizing the process of deploying ML systems), data infrastructure and more.

1. Rethinking the Stack for the AI Age

The AI infrastructure stack will be significantly different from historic datacenters, cloud, and software infrastructures due to the many unique characteristics and demands of AI workloads. Training currently accounts for the majority of computational requirements, but as market demand scales, efficient real-time inference and low latency may become equally critical.

While the transition from CPUs to AI chips including GPUs, ASICs and TPUs has significantly improved the efficiency of AI workloads, there is room for further optimization and innovation as well as rethinking hardware/software integration needs.

- The majority of GPUs are underutilized during peak times. Improving effective GPU deployment efficiency is set to become a key issue in 2024 through 2025.[17]

- A *"staggering 74% of companies are dissatisfied with their current job scheduling tools and face resource allocation constraints regularly, while limited on-demand and self-serve access to GPU compute inhibits productivity"*.[18]

Architecture innovations such as interconnect technologies and High Bandwidth Memory (HBM) have become key to the AI stack in order to optimize GPU usage:

- **Ayar Labs'** optical I/O solution seeks to address data movement bottlenecks in AI systems, to result in higher bandwidth and lower latency with greater power efficiency. "*We're on the cusp of a new era in high performance computing as optical I/O becomes a 'must have' building block for meeting the exponentially growing, data-intensive demands of emerging technologies like generative AI*".[19]

- Meanwhile **Celestial AI's** photonic fabric interconnect addresses the "*Memory Wall*"[20] by aiming to enable bandwidth delivery directly to the point of compute within the chip.

- **Cornelis Networks** delivers end-to-end high-performance interconnect solutions with a proprietary scale-out architecture, incorporating telemetry-based adaptive routing, congestion control with low latency and enhanced support for messaging, memory models and AI optimization in large-scale hyperscaler, cloud AI and on-prem AI/HPC environments.

> *"The bottleneck to many companies' growth quickly became not customer demand but access to the latest GPUs fron Nvidia."*
>
> SEQUOIA

At the hardware/silicon level, there could also be significant opportunity for new and specialized players given market demand potential, the opportunity to diversify supply bases given the dominance of Nvidia's estimated 80% market share[21] as well as potential future supply chain constraints.[22]

- **SiPearl** seeks to provide a high-performance, low power microprocessor for high performance computing (HPC) and AI workloads which integrates Samsung's HBM solution, to improve processing speed with reduced thermal resistance, rather than by simply adding more GPUs.[23]

Edge acceleration will require a range of AI-ready and energy efficient solution sets, whether for automotive, defense or enterprise for all of whom we are already seeing noticeable adoption.

- Edge computing market size was estimated at US$16 billion in 2023 and is expected to grow at 37% CAGR to US$156 billion by 2030.[24]

- Meanwhile, Gartner predicts that by the end of 2026, 100% of enterprise PC purchases will be an AI PC, with an integrated Neural Processing Unit (NPU).[25]

- **SiMa's** embedded edge machine learning system-on-chip (MLSoC) aims to allow customers to run entire applications on a chip, while **Axelera's** AI acceleration platform seeks to enable inference processing with YOLO (You Only Look Once, convolutional neural networks for real-time object detection) for edge AI computer vision applications.

> *"We are in the midst of a massive technological shift—innovation within this emerging AI infrastructure stack is progressing at an unprecedented pace."*
>
> BESSEMER VENTURE PARTNERS

Most of the emergent semiconductor/hardware category companies we spoke to are working on innovative technical approaches to workload management, recognizing that the computational requirements of data preparation, training and inference vary significantly. Integrated hardware/software solutions and use of computationally appropriate models will also vary significantly by use case.

- The use of industry or function specific GenAI models used by enterprise is expected to increase from approximately 1% in 2023 to 50% by 2027.[26]

2. Focus on AI Sustainability

The rapid growth in generative AI has sparked concerns about intensity of energy use. The International Energy Agency estimates that by 2026, datacenters could globally consume more than 1,000 terawatt-hours of electricity, more than double that of 2022 and roughly equal to Japan's total energy usage.[27] As energy consumption shifts from training towards inference as modern models are deployed at scale, energy efficiency throughout the model life cycle is increasingly under scrutiny.[28] We identify two potential emerging trends resulting: Energy and AI infrastructure integration, and innovation in energy-efficient hardware and models.

Interdependence of energy supply chains and datacenter infrastructure is seeing "*The Magnificent 7*" (Mag7)[29] integrating with energy infrastructure, increasingly co-locating datacenters with sustainable energy sources. Yet it is not clear that sufficient clean energy resources can meet demand.

- Microsoft and OpenAI are reportedly planning a massive US$100 billion, 5GW "*Stargate*" AI datacenter, potentially powered by alternative energy sources including nuclear, at an unspecified location.[30] While Microsoft signed an agreement with nuclear power producer Constellation Energy in 2023,[31] the analysis reports that this would not provide the scale of power required, and few existing global nuclear facilities could.[32]

- Microsoft also recently announced it is backing an estimated US$10 billion in renewable energy projects in a partnership with Brookfield AM.[33]

- Meanwhile in March 2024, Amazon acquired Talen Energy's datacenter campus at a nuclear power plant in Pennsylvania for US$650 million.[34]

The capital intensity of the AI Era is reflected in the Mag7's burgeoning Capex and R&D spend of US$374 billion in 2023. Amazon, Meta, Alphabet and Microsoft alone have already pledged to spend a combined US$200 billion on Capex in 2024, mostly on AI infrastructure, up 35% on 2023 figures. This represents as much as 21% of the total capex of the entire S&P500, up from 4% a decade ago.[35]

- Some analysts are beginning to question however how and when this "*AI Arms Race*" will generate commensurate returns.[36] "*Given the billions of dollars that Big Tech companies have been pouring into the AI boom, investors are cautious that this may ultimately result in infrastructure overbuild minus the promised future profits*," according to one.[37]

- There is a question about the sustainability of such capital and energy intensive AI datacenters and potentially a complementary need for innovation solutions to address the issue.

While many hyperscalers have long claimed to be 100% renewable or carbon neutral, there are broader issues surrounding the displacement of renewable energy supplies from alternative uses such as EVs, which could lead to increasing scrutiny of the need for intensive AI energy use.[38] We believe that innovations in chip technology, model architecture, orchestration, scheduling and energy efficient acceleration will likely play an important role and may present investment opportunities in the context of sustainability.

- Sustainability of energy resourcing is expected by analysts to result in 30% of GenAI implementations using energy-conserving computational methods by 2028.[39]

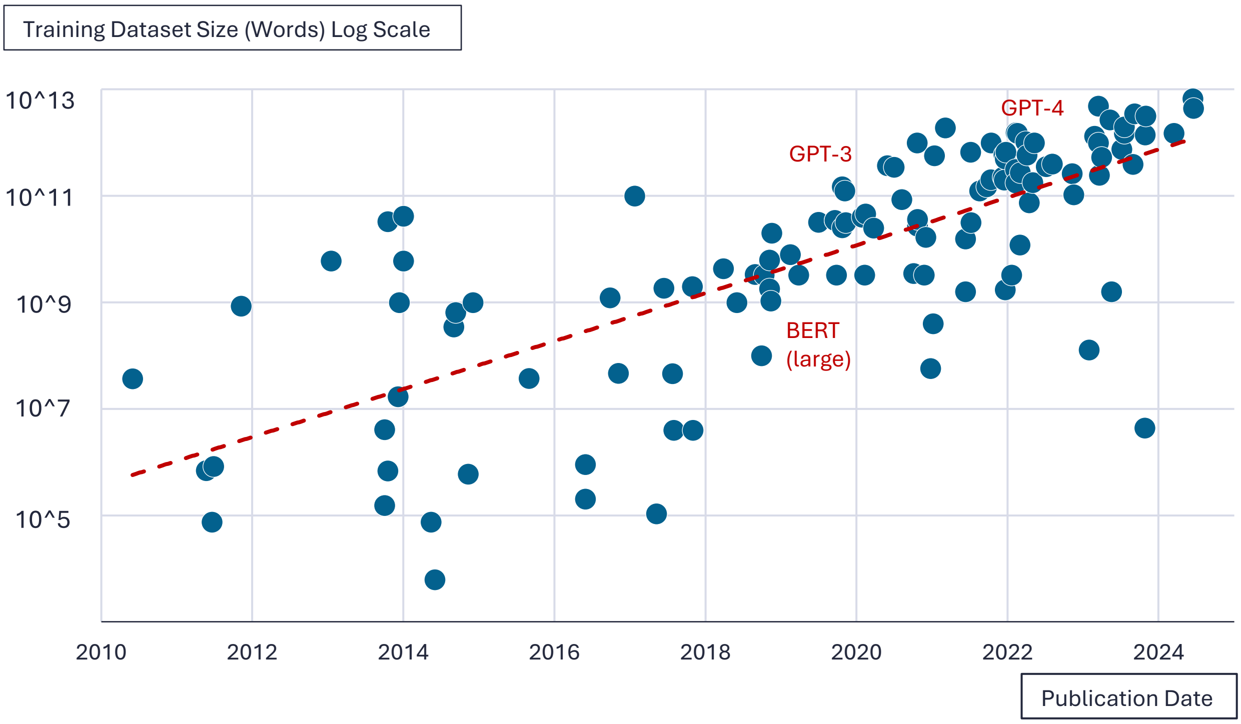
- We see the increasing use of more energy efficient Small Language Models such as Microsoft Phi3 as a trend for 2024, a trend which may present positive opportunities for AI developer platforms such as **Anaconda**, which supports the creation and flexible deployment of efficient models and a high-performance version of Python that maximizes efficiency of AI workloads, as well as **Weights & Biases**.[40]

- Nvidia's acquisition of Run.ai could be indicative of the need for dynamic resource allocation and workload orchestration to minimize idle resource and optimize energy efficiency. Similarly, **Anyscale** aims to enable dynamic scaling, advanced scheduling and resource management tools as well as reduced data movement to improve energy efficiency across AI workloads. **Octo AI**'s strategy is to optimize models for energy efficient deployment on edge devices through advanced model compression and optimization of specific target edge devices.

- Cerebras' wafer-scale chips and software stack seek to enable faster, more efficient processing of large AI models, reducing energy consumption. It is also designed to be air-cooled, addressing certain sustainability issues surrounding water-cooling.[41] **Lightmatter's** photonic AI accelerators and **Untether AI's** approach to at-memory computation are also designed to be air-cooled.

3. Looming Data Constraints, Synthetic and Alternative Data

According to research from EpochAI, models are using increasing quantities of training datasets, growing at a rate of 2.8x per annum. There are limited resources of public human-generated text, with researchers estimating that high-quality data stocks for model training may become fully exploited by 2026-2032.[42] This timing is uncertain as on the one hand, there may be advances in data efficiency, yet overtraining (using more data over reduced parameters with a view to optimizing compute for inference) could also exploit available public language data stocks even sooner.



Source: EpochAI, Lazard VGB Insights

There are a number of potential mitigating strategies and technologies to this bottleneck, including: speech recognition systems (e.g., OpenAI Whisper) enabling use of abundant audio data for training, Optical Character Recognition (OCR) enabling visual transformer models for paper academic documents (e.g., Meta's Nougat), as well as the generation and use of synthetic data.

- By 2025, it is estimated that 75% of businesses will use generative AI to create synthetic customer data, up from less than 5% in 2023.[43]

We believe there could be a possibility of medium-term constraints on ML models transitioning from compute to data and synthetic data providers including those such as **Gretel** and **Mostly.ai**, as well as multi-modal data ingestion like **Unstructured.io** may present possible opportunities.

Our view is that there might also be a continuation of the trend for alternative data strategic partnerships such as those recently announced, including between both Google and OpenAI and Reddit and Stack Overflow.[44]

The scramble for training and fine-tuning data has begun to extend to proprietary data sources. OpenAI announced a data licensing deal with News Corp (Wall Street Journal, New York Post, Barron's, The Times UK) in May 2024, "*worth US$250 million in the next five years*", alongside existing agreements with The Associated Press, Financial Times, and Axel Springer (Politico).[45]

While there are further unused private data sources that might yet be exploited for training or fine-tuning, these are currently being hampered by privacy concerns and threat vectors. Accordingly, there may be opportunities for model and data security solutions such as **Hidden Layer** and **Protect AI**.[46]

4. The Dawn of Autonomous AI Agents

While GenAI has delivered significant opportunity over the last 18 months, we believe that the trifecta of transformers, internet-scale data and human feedback could be just the start of what is possible. While autonomous agents in the form of coding companions such as co-pilots have already started to emerge, multi-agent systems (MAS) for cooperative GenAI applications may also potentially transform the opportunity space.

- AI MAS are complex systems of multiple interacting intelligent agents, which learn from and collaborate with each other. Modern AI (GPT, BERT)[47] may enable new opportunities for MAS, which have previously been constrained by complexity and coordination limitations.

- The market for autonomous agents is forecast to grow from US$5 billion in 2023 to US$29 billion by 2028 at a CAGR of 43%.[48]

- Established solutions such as Microsoft 365 Copilot and Github's Copilot are already widely used by developers, albeit with some accuracy and security risks.[49] Microsoft has further developed AutoGen as a framework to enable generation of multi-agent systems based on high-level specifications.

While still a relatively immature space for growth-stage investors, we believe that the use of multi-agent systems in AI infrastructure and the embedding of autonomous agents in applications could develop and become one of the important trends to watch over the next 12 months.

- Cognition Labs, the creator of autonomous coding agent Devin emerged from stealth only 6 months after formation raising US$175 million in April 2024 led by Founders Fund at a reported US$2 billion valuation.[50]

- The Open-Source project LangGraph seeks to facilitate efficient and complex information flows between multi-agent systems enabling them to collaborate and make decisions based on shared knowledge. The developer of LangGraph, LangChain recently (February 2024) announced a US$25 million Series A from Sequoia, Benchmark, and others.[51]

- Platforms for building and deploying AI applications using complex multi-agent approaches such as **Abacus.ai** could enable the creation of agents by chaining together user code, data transforms, ML models and LLM prompts which can access both LLMs and abacus APIs in one place.

- Agent monitoring-as-a-service will likely be necessary to develop multi-agent "artificial immune systems" with observability tooling including **Fiddler** and **AccelData** potentially seeing positive opportunities from the agent and trend towards MAS.

V. Segmentation of AI Infrastructure

There are many ways to segment the AI infrastructure market, and our approach is designed to provide an overall market view while recognizing the distinct features of each sub-segment.

- This market map provides an overview of some companies we have identified as incumbent players and some selected emergent players, with our selected **VGB AI Infra 40** companies highlighted in dotted boxes.

- Companies have been categorized by us based on what we have identified as their specific AI use case, using category abstractions. Certain companies, however, do not fit neatly into any single box but may have capabilities across multiple categories.

Segmentation of AI Infrastructure (cont'd)

MLOPS

Orchestration
ClearML, dataiku, Google Cloud Vertex AI, Apache Airflow, orkes, PREFECT

Explainability, Monitoring, Observability & Governance
acceldata, arize, fiddler, Grafana, unravel, Weights & Biases

LLM Ops
LangChain, PyTorch, TensorFlow

Synthetic Data
gretel, MOSTLY·AI, Synthesis.ai, TONIC

Model Development Tools
ANACONDA, comet, kumo, Determined AI, EDGE IMPULSE, LatentAI, Predibase, rescale, Weights & Biases

Feature Engineering
FeatureBase, tecton

Model Security
CALYPSOAI, HIDDENLAYER, PROTECT AI, ROBUST INTELLIGENCE

Data Prep
DataRobot, MAD, RelationalAI, UNSTRUCTURED

Data Labeling
Human Signal, Labelbox, Kili, scale, Snorkel

AI/ML Platforms
ABACUS.AI, DataRobot, DOMINO, Lightning AI

Model Serving and Inference

Optimization / Acceleration & Hosting
ANACONDA, anyscale, deci, Modular, NEURAL MAGIC, OctoAI

Model Repository & Hosting
ANACONDA, BENTOML, Hugging Face, Modal, SELDON

Foundational Models
ANTHROPIC, cohere, Gemini, imbue, MISTRAL AI_, LLaMA by Meta, OpenAI

Data Infrastructure

Data Lake / Warehouse
databricks, dremio, snowflake

Vector DB
DATASTAX, KX, Pinecone, Weaviate

Data Streaming
Apache Kafka, Cloud DataFlow, CONFLUENT

Data Management
bodo.ai, DATASTAX, WEKA, zilliz

Hyperscalers and Compute

Hyperscalers
aws, Azure, Google Cloud Platform, ORACLE

Compute-as-a-Service
anyscale, baseten, CoreWeave, groq, Lambda, NEXGEN CLOUD, ori, together.ai

Hardware / Silicon

I/O/Networking
AyarLabs, celestial AI, CORNELIS NETWORKS, enfabrica

Specialized Workloads
Achronix Data Acceleration, CORNAMI Intelligent Computing

Edge/Embedded AI
AXELERA, SiFive, SiMa.ai, SYNTIANT

Photonic Compute
LIGHTELLIGENCE, LIGHTMATTER, Salience Labs.

Inference Accelerator
cerebras, d-Matrix, groq, SambaNova SYSTEMS, tenstorrent, UNTETHER AI

HPC Supercompute
NEXTSILICON, SIPEARL

LAZARD

Source: Lazard VGB Insights

Segmentation of AI Infrastructure (cont'd)

The following section provides our outline of what we perceive to be the features of each AI Infrastructure subsector as well as identifying some incumbents and some selected growth stage companies identified during our research.

Hardware / Silicon

Sub-Segment	Description	Primary End Market	Selected Players Identified
Specialized Workloads	Designers of custom compute chips tailored to accelerate specific applications. Unlike general-purpose processors which handle a wide range of tasks, these chips are optimized for specialized workloads	healthcare, banking, fintech, pharma	Achronix Data Acceleration RAIN CORNAMI Intelligent Computing
HPC Supercompute	Designers of datacenter processors capable of handling high performance and/or high parallelism vs. traditional server CPUs	climate, security, energy, academia, healthcare, industrial	NEXTSILICON SIPEARL
Inference Accelerators	Chip developers focused on accelerating and optimizing efficiency for inference workloads for AI / ML	hyperscalers, LLM developers	cerebras d-Matrix groq SambaNova SYSTEMS tenstorrent UNTETHER AI
IO/Networking	Networking chip designers focused on reducing networking costs, latency, and power consumption often for AI and HPC workloads	data center ecosystem	AyarLabs CORNELIS NETWORKS celestial AI enfabrica
Edge/ Embedded AI	Developers of chips for compute at the edge and/or in applications with requirements for low power, high efficiency, and small footprint	surveillance, aerospace, manufacturing, automobile, wearable devices	AXELERA ARTIFICIAL INTELLIGENCE SiFive SiMa.ai SYNTIANT
Photonic Compute	Designers of chips that leverage light waves instead of electricity to handle computing, data storage, or communication which can be more powerful than traditional circuits as photons have higher bandwidth and are not affected by electromagnetic interference	cloud service providers, semiconductor companies, enterprises	LIGHTELLIGENCE LIGHTMATTER Salience Labs

Source: PitchBook Data, Inc.; Public Sources; Lazard VGB Insights

Segmentation of AI Infrastructure (cont'd)

Hyperscalers and Compute

Sub-Segment	Description	Primary End Market	Selected Players Identified
Hyperscalers	Large tech companies that provide extensive and scalable cloud computing services	Enterprise	aws Azure Google Cloud Platform ORACLE
Compute-as-a-Service	Third-party cloud service providers for users who require high-performance computing power. They run a cloud computing model that provides customers with a platform to develop, run, and manage applications without the complexity of building and maintaining the infrastructure	Enterprise, MLOps developers	baseten CoreWeave Lambda NEXGEN CLOUD together.ai

Data Infrastructure

Sub-Segment	Description	Primary End Market	Selected Players Identified
Data Lake / Warehouse	Large-scale, centralized storage repositories that hold structured data or raw data in its native format until it is needed for analysis	Enterprise, MLOps developers	databricks dremio snowflake
Vector DB	Specialized type of database designed to efficiently store, manage, and query high-dimensional vector data	Enterprise, MLOps developers	DATASTAX KX Pinecone Weaviate
Data Management	Centralized management systems for data and metadata leveraged by ML models that streamline the deployment, scaling, and monitoring of these models in production environment	Enterprise, MLOps developers	bodo.ai WEKA zilliz
Data Streaming	Platforms / frameworks that enable AI and ML models to operate on live data, facilitating real-time analytics, decision-making, and automated responses	Enterprise, MLOps developers	Apache Kafka Cloud DataFlow CONFLUENT

Source: PitchBook Data, Inc.; Public Sources; Lazard VGB Insights

Segmentation of AI Infrastructure (cont'd)

Model Serving and Inference

Sub-Segment	Description	Primary End Market	Selected Players Identified
Optimization	Platforms focused on enhancing the performance, scalability, and efficiency of applications, particularly those involving complex computations and large-scale data processing	Enterprise, MLOps developers	ANACONDA anyscale Modular NEURAL MAGIC OctoAI
Compute-as-a-Service	Centralized storage and management systems for ML models that streamline the deployment, scaling, and monitoring of these models in productions environment	Enterprise, MLOps developers	ANACONDA BENTOML Hugging Face Modal SELDON

Foundational Models

Sub-Segment	Description	Primary End Market	Selected Players Identified
Foundational Models	Foundational models seeking to compete with Chat GPT, Meta's Llama, and Microsoft's Copilot	Enterprise, MLOps developers	cohere imbue LLaMA by Meta OpenAI

MLOps

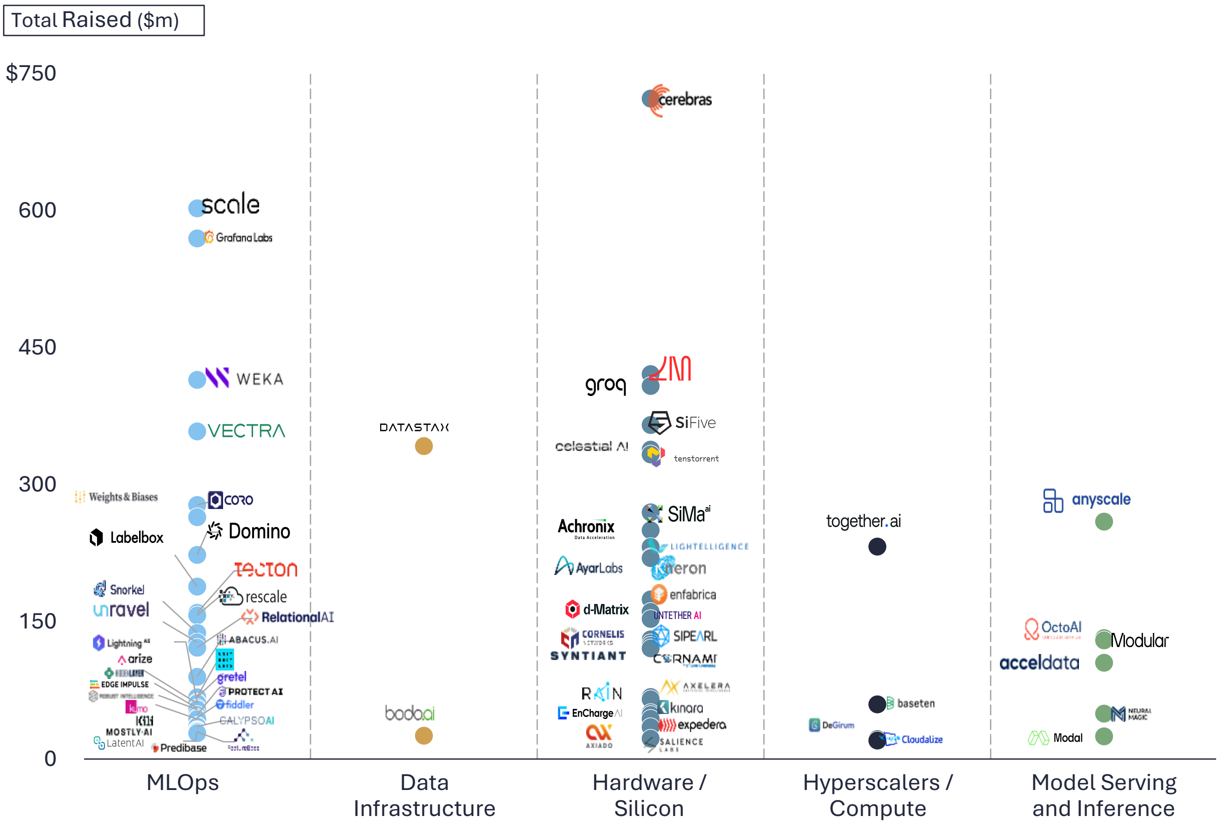
Sub-Segment	Description	Primary End Market	Selected Players Identified
MLOps	Includes technologies focused on the ingestion, cleansing, and transformation of data, to the training and optimization of models, and ultimately to the deployment and monitoring of the completed models	Enterprise, MLOps developers	ABACUS.AI ANACONDA DataRobot DOMINO EDGE IMPULSE FeatureBase fiddler gretel HIDDENLAYER LangChain LatentAI Lightning AI PyTorch RelationalAI TensorFlow rescale UNSTRUCTURED Weights & Biases

Source: PitchBook Data, Inc.; Public Sources; Lazard VGB Insights

Market Map – Company Maturity

We looked at the maturity stages of companies across our AI infrastructure market map and from our analysis have observed the following:

1. Significant capital seems to be being deployed to certain disruptors in the Hardware/Silicon and MLOps sector, possibly indicating that these segments are beginning to mature.
2. The market landscape appears highly fragmented across all segments, in our view with no clearcut winners in each category, likely encouraging enterprises to adopt one of three strategies: embrace end-to-end solutions, construct their own bespoke systems, or select the "best of breed" companies.
   - We might see category leaders continuing to attract through-the-cycle funding, and some mid-tier players may potentially be:
     i. forced into defensive mergers,
     ii. acquired by industry leaders, or
     iii. acquired by strategic investors
3. Companies including Cerebras, Scale and Grafana that have raised >US$500 million, may possibly be looking towards the public markets within the next 12-18 months depending on IPO market sentiment.



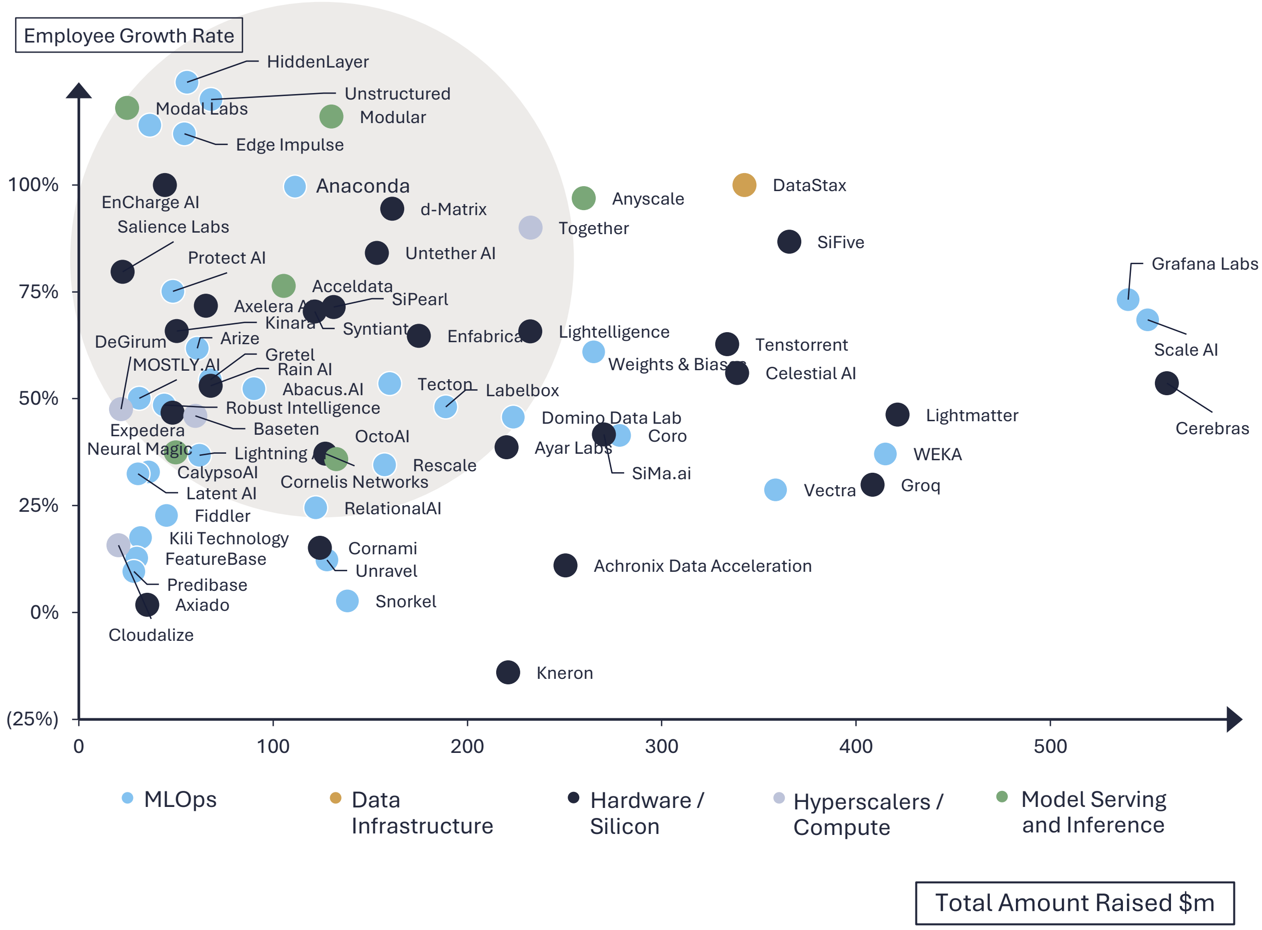
Source: PitchBook Data, Inc.; Lazard VGB Insights

Growth Trajectories across AI Infrastructure Segments

Relying on the aforementioned market segmentation map and companies which we have selected, we used employee growth figures since 2021 relative to total funds raised as a hypothetical proxy for overall growth. Based entirely on that measure of growth and applied to our selected landscape of companies, our analysis revealed that:

1. Many larger companies are continuing to experience rapid expansion (>50%+) even as they continue to scale.

2. A concentrated group of companies seem to be experiencing rapid growth, particularly at the relatively early stages of funding (highlighted in the grey section of the chart below).



Source: PitchBook Data, Inc.; Lazard VGB Insights

VI. Lazard VGB AI Infra 40 Profiles

AXELERA ARTIFICIAL INTELLIGENCE | *edge/embedded AI*

Hardware / Silicon

Eindhoven, Netherlands | www.axelera.ai

Founded 2021	Selected Investor(s) • Platinum Capital • Samsung Catalyst Fund
Total Raised $130m	
Employees 180	
Last financing Raised $68m in a Series B round	

AI Use Case

- Axelera AI is a developer of a hardware and software platform for AI, designed to deliver exceptional performance within a power envelope of just a few watts while maintaining the flexibility to support multiple networks
- The platform combines a custom dataflow architecture with multicore in-memory computing. This enables clients to minimize power consumption and deliver edge applications for a sustainable future, promoting both efficiency and environmental responsibility

AyarLabs | *I/O/networking*

Hardware / Silicon

San Jose, California | www.ayarlabs.com

Founded 2015	Selected Investor(s) • Alumni Ventures • HP • IAG Capital Partners • NVIDIA
Total Raised $219m	
Employees 154	
Last financing Raised $25m in a Series C1 round	

AI Use Case

- The company develops electronic-photonic chipsets designed for applications demanding high bandwidth, low latency, and power-efficient short-reach interconnects
- Utilizing industry-standard, cost-effective silicon processing techniques, the company creates optical-based interconnect chipsets and lasers to replace traditional electrical-based input-output systems. This technology allows companies to manage large volumes of data more effectively by miniaturizing fiber optic transceivers

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

celestial AI | *I/O/networking*

Hardware / Silicon

Santa Clara, California | www.celestial.ai

Founded
2020

Total Raised
$339m

Employees
99

Selected Investor(s)

- AMD Ventures
- IAG Partners
- Koch Disruptive Technologies
- Temasek

Last financing
Raised $175m in a Series C round

AI Use Case

- Celestial AI is a developer of an innovative data center and AI computing platform that aims to cater to deep learning and machine learning applications
- The company's technology combines the benefits of photonics, mixed-signal ASICs, and packaging to provide a substantial enhancement in computing performance. This enables clients to offer AI acceleration hardware and software alternatives, fostering advanced solutions for diverse applications.

CORNAMI Intelligent Computing | *specialized workloads*

Hardware / Silicon

Dallas, Texas | www.cornami.com

Founded
2011

Total Raised
$124m

Employees
49

Selected Investor(s)

- Raptor Group
- Softbank

Last financing
Raised $13m in SAFE notes

AI Use Case

- Cornami is a developer of reconfigurable computational fabric technology that aims to treat processor cores as scalable resources in the same way as memory, storage, and transistors
- The company's technology delivers scalability from thousands of cores on a single chip to millions across a system for real-time computing at the lowest cost, power, and latency available. It is focused on delivering real-time fully homomorphic encryption at market prices. This enables developers, large enterprises, and edge computing to deliver high performance anywhere and on any device with the lowest power consumption and latency, enhancing efficiency and productivity

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

CORNELIS NETWORKS | *I/O/networking*

Hardware / Silicon

Wayne, Pennsylvania | www.cornelisnetworks.com

Founded
2020

Total Raised
$127m

Employees
177

Selected Investor(s)

- Alumni Ventures
- IAG Capital
- Intel Capita

Last financing
Raised $25m in a Series B4 round

AI Use Case

- Cornelis Networks is a developer of purpose-built high-performance fabrics, designed for leading scientific, commercial, and government organizations
- The company offers high-performance fabrics that expedite computing, data analytics, and artificial intelligence workloads. This enables customers to effectively concentrate the computational power of multiple processing devices on a single problem, thereby enhancing both the result and accuracy simultaneously

d-Matrix | *inference accelerator*

Hardware / Silicon

Santa Clara, California | www.d-matrix.ai

Founded
2019

Total Raised
$161m

Employees
177

Selected Investor(s)

- A&E Investments
- M12
- Microsoft
- Temasek

Last financing
Raised $110m in a Series B round

AI Use Case

- d-Matrix is a developer of a computing platform tailored for GenAI and LLMS by offering an innovative in-memory computing technique for data centers
- The platform concentrates on addressing the physics of memory-compute integration using mixed-signal and digital signal processing techniques. This enables clients to benefit from enhanced computing efficiency, fostering improved performance and productivity

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

enfabrica | I/O/networking

Hardware / Silicon

Mountain View, California | www.enfabrica.com

Founded
2019

Total Raised
$175m

Employees
111

Selected Investor(s)

- Alumni Ventures
- IAG Partners
- NVIDIA
- Valor Equity Partners

Last financing
Raised $125m in a Series B round

AI Use Case

- Enfabrica, a developer of high-performance, ultra-scalable infrastructure silicon and software, aims to address the critical interconnect bottlenecks in next-generation computing workloads
- The company's foundational fabric technologies and products are designed to enable groundbreaking system efficiencies, topologies, and performance across various sectors, including hyper-scale cloud, edge, enterprise, fifth-generation/sixth-generation, and automotive infrastructure.

LIGHTELLIGENCE | phonic compute

Hardware / Silicon

Boston, Massachusetts | www.lightelligence.ai

Founded
2017

Total Raised
$232m

Employees
200

Selected Investor(s)

- CICC Capital
- Prosperity7

Last financing
Raised $20m in a Series C1 round

AI Use Case

- Lightelligence is an operator of an optical computing platform that aims to accelerate information processing
- The platform employs artificial intelligence and cutting-edge technology to transmit data via photons, enabling users to convey information with considerably lower latency and higher throughput compared to conventional electronic circuits

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

LIGHTMATTER | *photonic compute*

Hardware / Silicon

Mountain View, California | www.lightmatter.com

Founded
2017

Total Raised
$421m

Employees
154

Selected Investor(s)

- Fidelity
- GV
- HP
- Sequoia
- Viking Global

Last financing
Raised $155m in a Series C2 round

AI Use Case

- Lightmatter is leading the revolution in networking for AI.
- The company invented a leading 3D-stacked photonics engine, Passage™, capable of connecting thousands to millions of processors at the speed of light in extreme-scale data centers for the most advanced AI and HPC workloads

NEXTSILICON | *HPC Supercompute*

Hardware / Silicon

Austin, Texas | www.nextsilicon.com

Founded
2017

Total Raised
$300m

Employees
243

Selected Investor(s)

- Playground Global
- Standard Investments
- Corner Ventures
- Third Point Ventures

Last financing
Raised $128m in Series C2

AI Use Case

- NextSilicon develops advanced computing architecture technology focused on enhancing future computer processing methods
- The company specializes in chip design and software development, using innovative software algorithms to speed up compute-intensive applications. This technology provides high-performance architecture for supercomputers, offering a new approach to chip technology that supports organizations in achieving greater computational efficiency

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

SiMa.ai | *edge/embedded AI*

Hardware / Silicon

San Jose, California | www.sima.ai

Founded
2018

Total Raised
$270m

Employees
177

Selected Investor(s)

- Fidelity
- MSD Partners
- Maverick Capital
- Point72 Ventures

Last financing
Raised $70m in a Series B1 round

AI Use Case

- SiMa.ai facilitates the widespread adoption of high-performance machine learning inference at extremely low power in embedded edge applications
- It offers push-button performance, effortless deployment, and scaling at the embedded edge. This enables businesses to support traditional computing with high-performance, energy-efficient, safe, and secure machine learning inference

SIPEARL | *HPC Supercompute*

Hardware / Silicon

Maisons Laffitte, France | www.sipearl.com

Founded
2019

Total Raised
$131m

Employees
200

Selected Investor(s)

- ARM
- Bpifrance

Last financing
Raised $105m in a Series A round

AI Use Case

- SiPearl is a manufacturer of microprocessors for the European exascale supercomputing industry
- The company designs high-performance, energy-efficient microprocessors for various applications, including computing, artificial intelligence, medical research, climate change mitigation, and energy management. This provides scientific researchers, supercomputing centers, and leading entities from the IT, electronics, and automotive sectors with cutting-edge microprocessors, thereby fostering innovation and progress

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

UNTETHER AI | *inference accelerator*

Hardware / Silicon

Toronto, Canada | www.untether.ai

Founded
2018

Total Raised
$154m

Employees
127

Selected Investor(s)

- CPP Investments
- Intel Capital

Last financing
Raised $125m in a Series B1 round

AI Use Case

- Untether AI is a developer of AI chips designed to pioneer new frontiers in artificial intelligence applications
- The company's chips integrate near-memory design with digital processing to facilitate neural net inference that minimizes the distance data must travel. This enables clients to enhance inference efficiency while consuming fewer resources and energy, and requiring less supporting infrastructure, thereby promoting sustainable and efficient operations

Lambda | *compute-as-a-service*

Hyperscalers and Compute

San Jose, California | www.lambdalabs.com

Founded
2012

Total Raised
$892m

Employees
143

Selected Investor(s)

- Alumni Ventures
- B Capital
- IAG Capital
- Macquarie Group

Last financing
Raised $500m in debt funding

AI Use Case

- Lambda is a developer of a cloud computing platform tailored for large-scale artificial intelligence training and inference.
- The company's product portfolio spans from on-prem GPU hardware to hosted GPUs in the cloud and includes Lambda's proprietary software which enables deep learning / AI teams to access the tools they need via a singular interface regardless of the location of the compute resources (on-prem or cloud
- The solution is ideal for tasks such as natural language processing and drug discovery, enabling organizations to accelerate their AI initiatives with ease and precision

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

NEXGEN CLOUD | *compute-as-a-service*

Hyperscalers and Compute

London, UK | www.nexgencloud.com

Founded
2020

Total Raised
$13m

Employees
29

Selected Investor(s)

- DARMA Capital

Last financing
Raised $13m in a Seed round

AI Use Case

- NexGen Cloud operates as an Infrastructure-as-a-Service (IaaS) provider with a mission to bridge the gap between Web 2.0 and Web 3.0. Specializing in decentralized blockchain storage services, the company caters to the AI/ML, Meta, and Omniverse industries
- By offering robust and scalable solutions, it empowers large-scale projects to overcome challenges related to cost, transparency, and centralization, facilitating a seamless transition to the next generation of web technologies

together.ai | *compute-as-a-service*

Hyperscalers and Compute

Menlo Park, California | www.together.ai

Founded
2022

Total Raised
$233m

Employees
76

Selected Investor(s)

- Coatue
- Kleiner Perkins
- Lux Capital
- NVIDIA
- Prosperity7

Last financing
Raised $106m in a Series B round

AI Use Case

- Together AI is an operator of a technology services platform dedicated to offering a decentralized cloud for AI
- The platform focuses on constructing extensive, open models that are user-friendly and open-source. This enables researchers, developers, and companies to harness and enhance artificial intelligence through a seamless integration of data, models, and computation platforms

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

anyscale | *optimization / acceleration & hosting*

Model Serving and Inference

San Francisco, California | www.anyscale.com

Founded
2019

Total Raised
$260m

Employees
355

Selected Investor(s)

- a16z
- Addition
- Ant Group
- NEA

Last financing
Raised $199m in a Series C round

AI Use Case

- Anyscale aims to streamline distributed computing. The company's software specializes in an open-source framework that simplifies the creation of complex, compute-intensive applications by easing the underlying hardware decisions
- This enables software developers of all skill levels to build applications capable of running at any scale, from a laptop to a data center, promoting versatility and efficiency

Modular | *optimization / acceleration & hosting*

Model Serving and Inference

Palto Alto, California | www.modular.com

Founded
2022

Total Raised
$130m

Employees
184

Selected Investor(s)

- General Catalyst
- GV
- Greylock

Last financing
Raised $100m in a Series B round

AI Use Case

- Modular's is an integrated, composable suite of products that simplifies customers' AI infrastructure so they can develop, deploy, and innovate faster
- Modular provides an engine that tries to improve the inferencing performance of AI models on CPUs and GPUs while delivering on cost savings
- Modular's other flagship product, Mojo, is a programming language that aims to combine the usability of Python with features like caching, adaptive compilation techniques, and metaprogramming

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

OctoAI | *optimization / acceleration & hosting*

Model Serving and Inference

Seattle, Washington | www.octo.ai

Founded
2019

Total Raised
$133m

Employees
109

Selected Investor(s)

- Addition
- Amplify Partners
- Tiger Global

Last financing
Raised $85m in a Series C round

AI Use Case

- OctoAI is a developer of a technology is designed to assist engineering teams in swiftly deploying machine learning models on any hardware, cloud provider, or edge device
- The company's technology offers a managed service that utilizes machine learning to automate machine language code generation and optimization in multi-cloud environments. This ensures that the models continue to operate at peak efficiency, providing businesses with secure deployments of deep learning models, thereby enhancing productivity and security

ABACUS.AI | *AI/ML platforms*

MLOPs

San Francisco, California | www.abacus.ai

Founded
2019

Total Raised
$90m

Employees
122

Selected Investor(s)

- General Catalyst
- GV
- Greylock

Last financing
Raised $50m in a Series C round

AI Use Case

- Abacus.AI is an autonomous AI platform that aims to assist organizations in creating large-scale, real-time customizable deep learning systems
- The platform offers an end-to-end autonomous AI service that trains machine and deep learning models for common enterprise AI use cases such as churn prediction, time-series forecasting, and deep-learning-based personalization. It also allows for the creation of custom, specific models with a state-of-the-art toolset. This enables clients to integrate cutting-edge deep learning models into their business processes or customer experiences, fostering innovation and improved efficiency

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

acceldata | *monitoring and observability*

MLOPs

Campbell, California | www.acceldata.io

Founded
2018

Total Raised
$106m

Employees
259

Selected Investor(s)

- Aramco Ventures
- Lightspeed Venture Partners
- Prosperity7

Last financing
Raised $60m in a Series C round

AI Use Case

- Acceldata is a data and analytics platform designed to simplify data operations
- The platform provides information integration and data streaming services, enabling clients to stream, collect, and process data, construct data clusters, and gain actionable insights from the data. It also allows for optimization of workflow operations and capitalization on opportunities identified through predictive analytics. This empowers enterprises to proactively manage performance, security, data quality, and workflow, fostering improved efficiency and decision-making

ANACONDA. | *model development tools*

MLOPs

Austin, Texas | www.anaconda.com

Founded
2012

Total Raised
$77m

Employees
366

Selected Investor(s)

- Blackrock
- In-Q-Tel
- Morningside
- Snowflake Ventures

Last financing
Raised $31m in a Series B round

AI Use Case

- Anaconda provides an enterprise grade platform for open-source software development with a focus on AI and data science capabilities. Anaconda's platform enables AI developers, data scientists and IT teams with secure access to curated open-source software artifacts, with security, policy and governance control, capabilities to develop and deploy on-prem LLMs, and a high-performance version of Python that optimizes workload performance for both numerical and general code.
- The platform helps AI developers and data scientists ensure complete reproducibility and reliability of their projects and in-production workloads, and supports IT teams with reobust governance and control over how developers leverage open-source software artifacts.

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

DOMINO DATA LAB | *AI/ML platforms*

MLOPs

San Francisco, California | www.domino.ai

Founded
2013

Total Raised
$224m

Employees
323

Selected Investor(s)

- Coatue
- Amgen Capital
- NVIDIA
- Sequoia
- Snowflake Ventures

Last financing
Raised $100m in a Series F round

AI Use Case

- Domino Data Lab is a developer of an enterprise data science management platform designed to assist companies in building and deploying ideas through collaborative, reusable, and reproducible analysis
- The platform expedites research, accelerates model deployment, and fosters collaboration for code-first data science teams at scale. This enables data scientists to contribute to various fields, such as developing medicines, increasing crop productivity, adapting risk models to major economic shifts, constructing cars, and enhancing customer support

EDGE IMPULSE | *model development tools*

MLOPs

San Jose, California | www.edgeimpulse.com

Founded
2019

Total Raised
$54m

Employees
111

Selected Investor(s)

- Canaan Partners
- Coatue
- In-Q-Tel

Last financing
Raised $34m in a Series B round

AI Use Case

- Edge Impulse is an ML development platform designed to bring about positive societal change through machine learning.
- The platform streamlines the process of building, deploying, and scaling embedded ML applications. This enables developers to create intelligent devices by simplifying the collection of real sensor data, live signal processing from raw data to neural networks, testing, and deployment to any target device, fostering innovation and efficiency in creating smart solutions

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

FeatureBase | *feature engineering*

MLOPs

Austin, Texas | www.featurebase.com

Founded
2017

Total Raised
$30m

Employees
30

Selected Investor(s)

- Drive Capital
- Oracle

Last financing
Raised $24m in a Series A round

AI Use Case

- Featurebase is a data virtualization software is designed to secure access to large, fragmented, and geographically dispersed datasets
- The company's software aids in the mass parallelization of large, high cardinality, ad hoc queries for managing massive, unbounded datasets and takes advantage of the efficiency, performance, and simplicity of bitmaps as a foundation for powering AI with real-time information. This enables clients to perform time-based sharing to capture streaming data and carry out segmentation based on historical data or time ranges, ultimately enhancing data accessibility and analysis capabilities

fiddler | *monitoring and observability*

MLOPs

Palo Alto, California | www.fiddler.ai

Founded
2018

Total Raised
$45m

Employees
77

Selected Investor(s)

- Lightspeed Venture Partners
- Insight Partners
- Lux Capital

Last financing
Raised $32m in a Series B round

AI Use Case

- Fiddler is an enterprise AI platform designed to create AI services that are transparent, explainable, and understandable
- The platform utilizes the AI engine to provide statistical metrics, performance monitoring, and security services through a common language, centralized controls, and actionable insights. This enables businesses to analyze, manage, and deploy their machine learning models at scale, enhancing efficiency and decision-making capabilities

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

gretel™ | *synthetic data*

MLOPs

San Diego, California | www.gretel.ai

Founded
2017

Total Raised
$68m

Employees
77

Selected Investor(s)

- Drive Capital
- Oracle

Last financing
Raised $52m in a Series B round

AI Use Case

- Gretel is a data categorization and identification platform, is designed to automatically generate an anonymized version of a dataset
- The platform leverages machine learning to categorize data across various customer identifiers such as names and addresses. It features automatic data labeling, power testing, and synthetics. This enables developers to safely and swiftly experiment, collaborate, and build with customer data, promoting innovation and data privacy

HIDDENLAYER | *model security*

MLOPs

Austin, Texas | www.hiddenlayer.com

Founded
2022

Total Raised
$56m

Employees
50

Selected Investor(s)

- Capital One Ventures
- IBM Ventures
- M12
- Moore Strategic Ventures

Last financing
Raised $50m in a Series A round

AI Use Case

- HiddenLayer's AISec Platform is an AI/ML Protection Suite that ensures the integrity of customers' models throughout the MLOps pipeline
- By ensuring the security of pretrained models, detecting malicious injections, and monitoring algorithm inputs and outputs for potential threats, The AISec Platform delivers an automated and scalable defense tailored for ML
- This enables proactive responses to attacks without necessitating access to private data or models

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

Human Signal | *data labelling*

MLOPs

San Francisco, California | www.humansignal.com

Founded
2019

Total Raised
$30m

Employees
53

Selected Investor(s)

- Bow Capital
- 500 Global
- Unusual Ventures

Last financing
Raised $25m in a Series A round

AI Use Case

- Human Signal specializes in the development of advanced data labeling software engineered to centralize and streamline training data management
- The platform enhances operational efficiency by integrating robust management and annotator functionalities, which facilitate and optimize collaborative data labeling processes, quality assurance, and analytical evaluations. This enables enterprises to expedite dataset annotation and achieve precise, high-performance machine learning and artificial intelligence models at scale, thereby maintaining a competitive edge in the industry

LatentAI | *model development tools*

MLOPs

Menlo Park, California | www.latentai.com

Founded
2018

Total Raised
$31m

Employees
45

Selected Investor(s)

- Lockheed Martin Ventures

Last financing
Raised $27m in a Series A1 round

AI Use Case

- Latent AI is an inference platform designed to support edge computing workloads
- The company offers a quantization optimizer for edge AI devices to automate the exploration of low-bit-precision training deploying efficient neural networks for on-device intelligence and inference. This enables software developers to feasibly access, deploy, and manage AI for the edge, promoting innovation and efficiency in edge computing applications

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

Lightning AI | *AI/ML platforms*

MLOPs

New York, New York | www.lightning.ai

Founded
2019

Total Raised
$62m

Employees
64

Selected Investor(s)

- Bain Capital Ventures
- Coatue
- Index Ventures

Last financing
Raised $40m in a Series B round

AI Use Case

- Lightning AI is a multi-cloud machine learning systems designed to aid in building simple research demos
- The platform offers monitoring, training management, single command cloud training, experiment analysis, engineering automation, and automated artifact backups. This enables engineers, data scientists, and AI researchers to save time and train machine learning models on the cloud directly from their laptops, enhancing efficiency and convenience

MAD Mad Street Den | *data preparation*

MLOPs

Redwood City, California | www.madstreetden.com

Founded
2013

Total Raised
Undisclosed

Employees
313

Selected Investor(s)

- Alpha Wave Global
- Chimera Capital
- Sequoia Capital

Last financing
Undisclosed

AI Use Case

- Mad Street Den is a cloud-based AI platform is designed to build models of generalizable intelligence and create actionable ways to contextualize AI on a large scale
- The company's platform offers artificial intelligence and computer vision modules to facilitate various features, including object recognition, gaze tracking, emotion-expression detection, head and facial gestures, as well as 3D facial reconstruction. This enables clients to build models of generalizable intelligence on a grand scale, which can be deployed through meaningful applications across various industries, enhancing efficiency and innovation

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

MOSTLY·AI | *synthetic data*

MLOPs

Vienna, Austria | www.mostly.ai

Founded
2017

Total Raised
$31m

Employees
61

Selected Investor(s)

- Citi Ventures
- 42CAP
- Molten Ventures

Last financing
Raised $25m in a Series B round

AI Use Case

- Mostly AI is a pioneer in GPU-powered technology designed to simulate synthetic customer data at scale. This cutting-edge technology enables the generation of an unlimited number of realistic and representative synthetic customers, closely mirroring the patterns and behaviors of actual customers with unprecedented accuracy
- By leveraging this advanced simulation capability, businesses can unlock a wealth of opportunities from previously inaccessible data, driving faster innovation while mitigating risks and reducing costs. This transformative approach empowers organizations to harness the full potential of their data assets, opening new avenues for growth and efficiency

PROTECT AI | *model security*

MLOPs

Seattle, Washington | www.protectai.com

Founded
2022

Total Raised
$49m

Employees
46

Selected Investor(s)

- Evolution Equity Partners
- Salesforce Ventures

Last financing
Raised $35m in a Series A round

AI Use Case

- Protect AI is a cybersecurity platform is designed to concentrate on machine learning workflows and pipelines
- The company's platform offers innovative security products and performs security scans using machine learning models and artificial intelligence systems to access curated resources, learn best practices in machine learning security, listen to podcasts with thought leaders, and connect with a thriving community. This enables enterprises to build a safer, AI-powered world, fostering enhanced security and innovation

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

rescale | *model development tools*

MLOPs

San Francisco, California | www.rescale.com

Founded
2011

Total Raised
$157m

Employees
234

Selected Investor(s)

- A&E Investments
- a16z
- DCVC
- DST Global

Last financing
Raised $105m in a Series C round

AI Use Case

- Rescale, a developer of a cloud-based infrastructure platform, is designed to streamline scientific and engineering simulations
- The platform offers infinite scalability, customization tools, and the ability to make adjustments optimized for specific workloads, ultimately reducing turnaround times. This enables businesses to transform their information technology into unified, agile environments, enhancing overall outcomes and efficiency

RelationalAI | *data preparation*

MLOPs

Berkeley, California | www.relational.ai

Founded
2017

Total Raised
$122m

Employees
169

Selected Investor(s)

- Addition
- Madrona Venture Group
- Menlo Ventures
- Tiger Global Management

Last financing
Raised $75m in a Series B round

AI Use Case

- RelationalAI is a relational knowledge graph system designed to address complex business challenges
- The company concentrates on the rich interdependencies and structures inherent in every business, complementing the modern data stack to expedite the development of intelligent data applications. This enables clients to implement intelligent applications with semantic layers on a data-centric foundation, lowering the barrier to codifying and utilizing knowledge, and ultimately enhancing business efficiency and decision-making

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

Snorkel | *data labelling*

MLOPs

Redwood City, California | www.snorkel.ai

Founded
2019

Total Raised
$138m

Employees
157

Selected Investor(s)
- A&E
- Accel
- Addition
- Blackrock
- Greylock
- GV

Last financing
Raised $85m in a Series C round

AI Use Case

- Snorkel is an AI-powered programmatic data labeling tool is designed for extracting information from text documents such as scientific articles and electronic health records
- The company's tool leverages theoretically grounded techniques to perform data augmentation and slicing data into different critical subsets, and then identifies subsets of the data. This enables users to quickly leverage structured data resources available in domains such as bioinformatics, enhancing efficiency and productivity in data processing and analysis

unravel | *monitoring and observability*

MLOPs

Palo Alto, California | www.unraveldata.com

Founded
2012

Total Raised
$128m

Employees
140

Selected Investor(s)
- Bridge Bank
- GGV Capital
- M12
- Menlo Ventures

Last financing
Raised $70m in a Series D round

AI Use Case

- Unravel leverages AI, ML and analytics to offer actionable recommendations and automation, enabling businesses to understand and optimize their data-driven applications
- Unravel's purpose-built AI data observability and FinOps for Databricks, Snowflake, BigQuery and other modern data stacks provides granular visibility for cost allocation, metadata correlation for data reliability, and AI-powered insights for data performance management

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

UNSTRUCTURED | *data preparation*

MLOPs

Rocklin, California | www.unstructured.io

Founded
2022

Total Raised
$68m

Employees
50

Selected Investor(s)

- Alumni Ventures
- Bain Capital Ventures
- Menlo Ventures
- NVIDIA

Last financing
Raised $43m in a Series B round

AI Use Case

- Unstructured.io is an open-source data transformation platform designed to simplify the preprocessing of natural language data for downstream machine learning services
- The platform utilizes open-source libraries and application programming interfaces to construct custom preprocessing pipelines for labeling, training, or production machine learning pipelines. This enables clients to convert simple data into language data, fostering innovation and improved efficiency in processing natural language information

Weights & Biases | *model development tools*

MLOPs

San Francisco, California | www.wandb.ai

Founded
2017

Total Raised
$265m

Employees
262

Selected Investor(s)

- BOND Capital
- Coatue
- Felicis
- NVIDIA
- Sapphire Ventures

Last financing
Raised $65m in an Undisclosed round

AI Use Case

- Weights & Biases is a dataset optimization tool, is dedicated to creating high-quality software tools for deep learning practitioners
- The company's tool offers performance visualization for machine learning and assists teams in tracking their models, visualizing model performance, and effortlessly automating the training and enhancement of models. This enables companies to transform deep learning research projects into deployed software, fostering innovation and efficiency

Source: Company Websites; Funding Press Releases; PitchBook Data, Inc.; Lazard VGB Insights

Disclaimer

This document has been prepared by Lazard Frères & Co. LLC ("Lazard") solely for general information purposes and is based on publicly available information which has not been independently verified by Lazard. The information contained herein is preliminary and should not be relied upon for any purpose. No liability whatsoever is accepted, and neither Lazard nor any member of the Lazard Group (being Lazard Ltd and its direct and indirect subsidiary and associated undertakings) nor any of their respective directors, partners, officers, employees, representatives or other agents is, or will be, making any warranty, representation or undertaking (expressed or implied) concerning the accuracy or truthfulness of any of the information, ideas, forecasts, projections or of any of the views or opinions contained in this document or any other written or oral statement provided in connection herewith or for any errors, omissions or misstatements contained herein or for any reliance that any party may seek to place upon any such information. Lazard undertakes no obligation to provide the recipient with access to any additional information or to update or correct any information contained herein. Interested parties should conduct their own investigation and analysis of the companies and information referenced herein. Lazard only acts for those entities whom it has identified as its client in a signed engagement letter and no-one else and will not be responsible to anyone other than such client for providing the protections afforded to clients of Lazard nor for providing advice. Nothing contained in this document constitutes, or should be relied upon as, (i) the giving of financial, investment or other advice or recommendations by, or the issuance of research by, Lazard, or (ii) a promise or representation as to any matter whether as to the past or the future. Recipients are recommended to seek their own financial and other advice and should entirely rely solely on their own judgment, review and analysis of this document. Lazard or other members of the Lazard Group (A) may have acted in the past, or act currently or in the future as adviser to some of the companies referenced herein, (B) may receive fees in connection with any such advisory engagements, (C) may at any time be in contact with such companies in order to solicit them to enter into advisory engagements and (D) may from time to time have made, and may in the future make, investments in such companies. Nothing contained in this document constitutes, or should be deemed to constitute, an offer or solicitation for the purchase or sale of any security. You undertake to keep this document confidential and to not distribute it to any third party, or excerpt from or reproduce this document (in whole or in part), without the prior written consent of Lazard.

VII. References

[1] McKinsey Digital, *"The Economic Potential of Generative AI: The Next Productivity Frontier",* June 2023

[2] Axios, *"OpenAI's Chris Lehane says AI is Critical Infrastructure"*, April 2024, as quoted in Axios, *"Behind the Curtain: AI's Ominous Scarcity Crisis",* May 2024

[3] Bank of *America "Global Fund Manager Survey"*, March 2024, as quoted in Morningstar*, "Is the World in an AI Bubble? Money Managers are Split?",* March 2024

[4] Bill Janeway, *"Productive Bubbles",* in Noema Magazine, July 2021

[5] Bloomberg Intelligence, *"2023 Generative AI Growth Report"*, June 2023, as quoted in Bloomberg Press Release, *"Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds"*, June 2023

[6] McKinsey Digital, *"The Economic Potential of Generative AI: The Next Productivity Frontier"*, June 2023

[7] Additional quotations in text boxes from: Mark Benioff, *"Salesforce+ Salesforce AI Day"* video, cited in Forbes, *"Shifting the AI Narrative: From Doomsday Fears to Pragmatic Solutions"*, March 2024; GMO*, "The Great Paradox of the U.S. Market"*, March 2024; Jensen Huang, as cited in Reuters, "*Chip Giant Nvidia Nears Trillion-Dollar Status on AI Bet*", May 2023; Bessemer Venture Partners, *"Roadmap: AI Infrastructure"*, June 2024

[8] Gartner CEO Survey November 2023, quoted in Harvard Business Review, *"5 Forces That Will Drive the Adoption of GenAI"*, December 2023

[9] Gartner Press Release, "*More than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026*", October 2023

[10] See Press Releases: "*Amazon and Anthropic Deepen Their Shared Commitment to Advancing Generative AI",* March 2024; *"Snowflake Partners with Mistral AI to Bring Industry-Leading Language Models to Enterprises Through Snowflake Cortex",* March 2024; *"Hugging Face and Google Partner for Open AI Collaboration",* January 2024

[11] Reuters*, "US sets Stage for Antitrust Probes into Microsoft, OpenAI and Nvidia"*, June 2024; Fortune, *"Why Microsoft's Surprise Deal with $4 Billion Startup Inflection is the Most Important Non-Acquisition in AI"*, March 2024; Reuters, *"Microsoft Pays Inflection $650 Million in Licensing Deal While Poaching Top Talents"*, March 2024

[12] Stanford Institute for Human-Centered Artificial Intelligence, *"Artificial Intelligence Index 2024",* Chapter 4: Economy, Figure 4.3.8, p.35

[13] Dealroom.co, *"State of AI Investing"*, May 2024

References (cont'd)

[14] See Anupam Chander, Haochen Sun (eds), *"Data Sovereignty: From the Digital Silk Road to the Return of the State",* Oxford University Press 2023, Chapter 5: Andrew Keane *Woods "Digital Sovereignty + Artificial Intelligence"*; World Economic Forum, *"Sovereign AI: What it is, and 6 Strategic Pillars for achieving it"*, April 2024

[15] Lazard Geopolitical Advisory, *"The Geopolitics of Artificial Intelligence",* October 2023

[16] Allied Market Research, *"AI Infrastructure Market 2023"*, September 2023

[17] AI Infrastructure Alliance, *"The State of Infrastructure at Scale 2024"*, March 2024

[18] AI Infrastructure Alliance, March 2024, *ibid.*

[19] Ayar Labs CEO Charlie Wuischpard, quoted in Press Release, *"Ayar Labs Showcases 4 Tbps Optically-Enabled Intel FPGA at Supercomputing 2023"*, November 2023

[20] The *"Memory Wall"* refers to a mismatch between the slow growth of on-chip memory capabilities and the dramatic expansion of data requirements for advanced AI. See Optics and Photonics News, *"Celestial AI Cultivates a Photonic Fabric Ecosystem"*, April 2024

[21] The Wall Street Journal, *"Nvidia's Business is Booming. Here's What Could Slow It Down"*, May 2024

[22] Quote from Sequoia, *"Generative AI's Act Two"*, Sonya Huang and Pat Grady, September 2023

[23] eeNews Embedded*, "SiPearl Partners with Samsung for built-in HBM in Rhea"*, May 2024

[24] Grand View Research, *"Edge Computing Market Size and Trends",* March 2024

[25] Gartner*, "Gartner Forecasts Worldwide AI Chips Revenue to Grow 33% in 2024"*, May 2024

[26] Gartner, *"3 Bold and Actionable Predictions for the Future of GenAI",* April 2024

[27] International Energy Agency (IEA), *"Electricity 2024"*, January 2024: IEA, Paris https://www.iea.org/reports/electricity-2024, Licence: CC BY 4.0

[28] Luccioni, Sasha, Yacine Jernite, and Emma Strubell. *"Power Hungry Processing: Watts Driving the Cost of AI Deployment?"* In The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 85-99. 2024, arXiv:2311.16863 [cs.LG]

References (cont'd)

[29] The *"Magnificent Seven"* includes Meta, Apple, Tesla, Nvidia, Amazon, Microsoft, and Alphabet

[30] Data Center Dynamics*, "Microsoft and OpenAI Consider $100 Billion, 5GW 'Stargate' AI Data Center"*, March 2024

[31] Data Center Dynamics, *"Microsoft Signs 24/7 Nuclear Power Deal with Constellation for Boydton Data Center"*, June 2023

[32] Data Center Dynamics, *"Is Microsoft and OpenAI's 5GW Stargate Supercomputer Feasible?"*, April 2024

[33] Financial Times, "*Microsoft to Power Data Centers with Big Brookfield Renewables Deal"*, May 2024

[34] Data Center Dynamics, *"AWS Acquires Talen's Nuclear Data Center Campus in Pennsylvania"*, March 2024. See also Talen Energy, *"Mar-24 Business Update Presentation"*, available at https://talenenergy.investorroom.com/financials-presentations

[35] Capital Group, *"Tech Giants Ratchet Up Spending in AI Race"*, May 2024

[36] See, for example, The Economist, *"Big Tech's Capex Splurge May be Irrationally Exuberant"*, May 2024

[37] Nicole Tanenbaum, Chequers Financial Management, quoted in *"Big Tech's AI Spending Spree Comes with a Catch"*, May 2024

[38] Google, *"100% Renewable is Just the Beginning",* December 2016

[39] Gartner, April 2024, *ibid*

[40] Venture Beat, *"Why Small Language Models are the Next Big Thing in AI"*, April 2024

[41] Each ChatGPT search uses an estimated gallon of water. Microsoft, Meta, and Alphabet have all set targets to become water positive by 2030. See Liontrust, *"The New Investment Landscape for the Water Industry"*, June 2024

[42] Epoch AI, *"Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data"*, June 2024. See also full paper: Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. *"Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data"*. ArXiv arXiv:2211.04325v2 [cs.LG], 2024. https://doi.org/10.48550/arXiv.2211.04325 Graph generated from selected model data in EpochAI's Notable AI Models set, under Creative Commons Attribution license. Epoch AI, *"Data on Notable AI Models"*. Published online at epochai.org. Retrieved from 'https://epochai.org/data/notable-ai-models' [online resource]. Accessed 1 Jul 2024.

References (cont'd)

[43.] Gartner, April 2024, *ibid*

[44.] Company Press Releases*, "Stack Overflow and Google Cloud Announce Strategic Partnership to Bring Generative AI to Millions of Developers"*, February 2024; *"Stack Overflow and OpenAI Partner to Strengthen the World's Most Popular Large Language Models",* May 2024; *"OpenAI and Reddit Partnership", May 2024; Reddit, "Expanding our Partnership with Google"*, February 2024

[45.] The Verge, *"OpenAI's News Corp Deal Licenses Content from WSJ, New York Post and More",* May 2024; OpenAI Announcement, *"A Landmark Multi-Year Global Partnership with News Corp"*, May 2024

[46.] For more on Security for AI, see Menlo Ventures, *"Security for AI: The New Wave of Startups Racing to Secure the AI Stack"*, February 2024

[47.] Bidirectional Encoder Representations from Transformers, LLM introduced by Google in October 2018

[48.] See for example, Markets and Markets, *"Autonomous AI and Autonomous Agents Market",* June 2023

[49.] Axios, *"When AI-Produced Code Goes Bad"*, June 2024

[50.] Quartz, *"An AI Startup That's Not Even Six Months Old Says It's Worth $2 Billion",* April 2024

[51.] Next Unicorn, *"LangChain Secures $25 Million in Funding",* February 2024. See also LangChain Press Release, "*Announcing the General Availability of LangSmith and Our Series A Led by Sequoia Capital*", February 2024