

The Model Serving & Compute Landscape 2026

134

COMPANIES MAPPED

10

SUB-SECTORS

\$114bn

TOTAL CAPITAL RAISED*

~\$22bn

OF IT. VENTURE FUNDING

69%

US-HEADQUARTERED

\$92bn of the \$114bn raised in this category is debt and public-market financing, concentrated in a handful of compute providers. Strip it out and genuine venture equity is just **\$22.5bn**, with over half the field still at Seed or Series A. Compute is a far earlier and more contestable market than its headline capital implies.

LLM & Model Inference Serving 33


baseten


beam

ACQUIRED


BentoML


Berget


cerebrum

ACQUIRED


Cortex
GPU Orchestration & Cluster Management


Cumulus
Model Gateways & Routing


deepinfra


Doubleword


Featherless


Fireworks AI


Flex


Gelu AI

ACQUIRED


Groq


Impala AI


inference
Model Gateways & Routing


Inferless

ACQUIRED


Lepton AI
GPU Cloud & Compute Rental


Modal

ACQUIRED

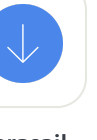

Modular

ACQUIRED


OctoML
Inference Optimization & Compiles

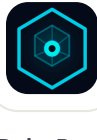

Ollama


OpenGradient


Parasail


Pipeshift


PolarGrid


PolarPeer

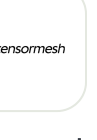
ACQUIRED


Replicate
GPU Cloud & Compute Rental


SambaNova Systems
AI Compute Hardware: Chips & Networking


Simplismart


Synthetic


tensormesh

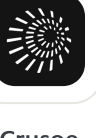

Together
GPU Cloud & Compute Rental

GPU Cloud & Compute Rental 31


Andromeda


Coreweave


Corvex
AI Data Privacy & Confidential Computing


Crusoe


DigitalOcean

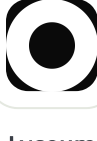

Firmus Technologies


Gensyn

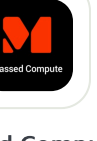

Hydra Host


io.net


Lambda


Lyceum


Manifold Labs
AI Data Privacy & Confidential Computing


Massed Compute


Nava
AI Data Centre & Power Infrastructure


Nebius
MLOps & ML Lifecycle Platforms


neysa
GPU Orchestration & Cluster Management


Novita
Agent Execution & Sandboxing


Nscale


PaleBlueDot
GPU Orchestration & Cluster Management


Qumulus


runpod


Scaleway


Sharon AI

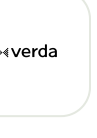

TensorWave


The San Francisco Compute Company


Thunder Compute


Trainy


Vast


Verda


VESSEL AI
MLOps & ML Lifecycle Platforms


Voltage Park

AI Compute Hardware: Chips & Networking 21


Arago


Axelera


Ayar Labs

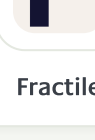

Blumind
Edge & On-Device AI


Cerebras Systems

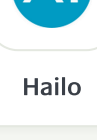

d-Matrix


DriveNets

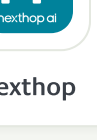

Extropic


Fractile


FuriosaAI

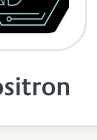

Hailo

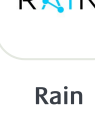

Mythic


Nexthop


OLIX


Opticore

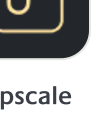

Positron


Rain


Rebussions


Syntiant
Edge & On-Device AI


Unconventional AI


Upscale

Model Gateways & Routing 12


Alsa
Agent Tool Integration & MCP Gateways


Barndoor AI
AI Runtime Security & Guardrails


Clod


Coworker
Enterprise AI Platforms & Operating Systems

ACQUIRED


Helicone
LLM & Agent Observability


LiteLLM


Maitai
LLM & Agent Observability


Not Diamond


OpenRouter

ACQUIRED


Portkey
LLM & Agent Observability


TensorZero
LLM & Agent Observability


TrueFoundry
MLOps & ML Lifecycle Platforms

Inference Optimization & Compilers 12

ACQUIRED


CentML
Model Training & Fine-Tuning

ACQUIRED


Clarifai

ACQUIRED


Deci


General Compute


Literal Labs


Luminal


Mako

ACQUIRED


Neural Magic


OpenInfer


Pruna Ai


Refiant


TheStage AI

GPU Orchestration & Cluster Management 11


Anyscale


Clear ML
MLOps & ML Lifecycle Platforms


Coiled


Hammerhead


https://dstack.ai/

ACQUIRED


Mirantis


Mithril


Rafay

ACQUIRED


Run.ai


Saturn


ScaleOps

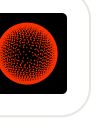
AI Data Centre & Power Infrastructure 8


Applied Digital
GPU Cloud & Compute Rental


DayOne


Fluidstack


Nexus Data Centers


Radiant


WEKA


WhiteFiber


Zettabyte

Other Sub-Sectors 2


Decompute
On-Device / Edge AI Inference


vCluster
GPU Orchestration & Control Plane

Hyperscaler / SaaS Giant 4


Amazon Web Services


Google


Microsoft


NVIDIA

*Funding from Crunchbase, 112 of 134 companies matched. "Total capital raised" includes all funding types (venture, debt, post-IPO, grants); "venture funding" excludes debt, private equity, post-IPO and grant rounds. Companies may appear in more than one sub-sector. Logos via logo.dev.