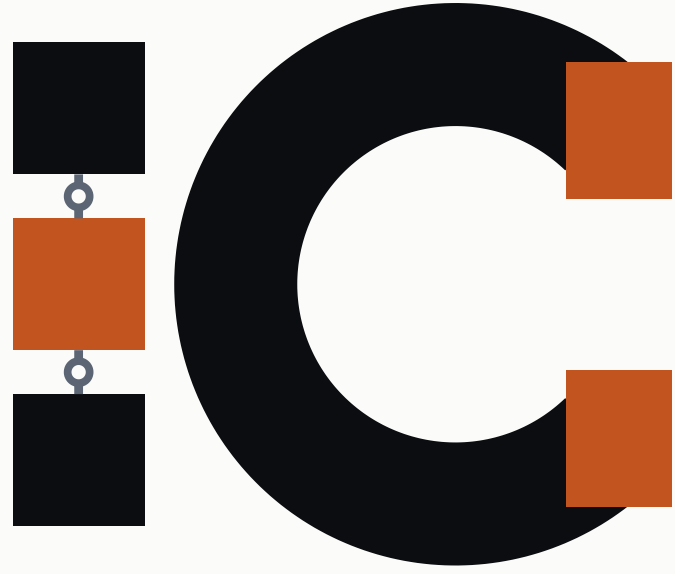


INFRA CONVICTION



The AI Infrastructure Atlas

How the AI stack works from the power plant to the agent, who wins at each layer, and where an early-stage investor should look next.

24

posters

9

layers + outcomes

43

micro-themes

150

startups mapped

114

early-stage startups

The stack on one page

Read bottom to top: each layer exists to make the one above cheaper, faster or safer.

★ Outcomes	What the customer pays for LLM assistants AI software engineering Agentic apps Vertical AI Physical AI	4 spotlight 5 outcomes
L8 Trust	Is the answer right, is the action safe and legal Evals & observability AI security & guardrails Data privacy & security Governance & compliance	8 spotlight 12 early-stage 4 micro-themes
L7 Agent runtime	Hands, memory and ID cards for agents Orchestration Tools & MCP Web access Memory & context Identity & auth Payments Voice & realtime	4 spotlight 12 early-stage 7 micro-themes
L6 Builder tools	The developer's workbench Frameworks & SDKs Sandboxes & dev boxes AI code review & testing	3 spotlight 6 early-stage 3 micro-themes
L5 Serving & efficiency	Run the brain fast and cheap Inference platforms Inference optimization Gateways & routing Local & on-device	3 spotlight 12 early-stage 4 micro-themes
L4 Model building	Make the brain Frontier labs Open-weight models Specialized models World models & sim Training & ML workflow Fine-tuning & continual learning RL environments	4 spotlight 12 early-stage 7 micro-themes
L3 Data	The fuel: what models learn from and look up Data rights Expert data & traces Synthetic data Warehouses & pipelines Storage & file systems Retrieval & vector search Robot & sensor data	6 spotlight 16 early-stage 7 micro-themes
L2 Compute cloud	Rent the engines Hyperscalers Neoclouds GPU markets & finance Orchestration & utilization	1 spotlight 8 early-stage 4 micro-themes
L1 Silicon	The engines and their wiring Accelerators & memory Networking & optics Edge chips Chip design & verification	2 spotlight 7 early-stage 4 micro-themes
L0 Power & buildings	Electricity, cooling and campuses Power & grid Cooling & power gear DC build & operations	1 spotlight 7 early-stage 3 micro-themes

PART 1 History	1–3	PART 2 Value chains	4–5	PART 3 Layer deep dives	6–18
How every layer of today's AI stack existed before LLMs, and what changed		One request traced through the SaaS stack, then the AI stack		Power to trust: micro-themes, leaders, startups, spotlight startups, exits	
PART 4 Outcomes	19–21	PART 5 Investor lens	22–24	APPENDIX Reference	A–C
How the stack moves LLMs, agents, vertical AI and physical AI		Where value accrues, spotlight startups ranked, and where to look next		Startup index, spotlight index, inference close-up, glossary, method and sources	

The stack before LLMs

1995 to 2019. Every layer of today's AI stack already existed, doing the same job for websites and SaaS apps.

HOW TO READ

Rows are layers. The bottom row is closest to the power plant.
Columns are eras. Each card is a company or event, with its date.
The job in each row never changes. The company doing it does.

LAYER
ERA →

TURNING POINTS

1995–2005
Dot-com & on-prem
Buy the box
BOTTLENECK
Hardware cost
UNIT OF WORK
Page view

2006–2011
Cloud & SaaS
Rent the box
BOTTLENECK
Ops & deploy time
UNIT OF WORK
API request

2012–2019
Big data & deep learning
Wire the data
BOTTLENECK
Data plumbing & ML talent
UNIT OF WORK
Query / event

2006 AWS opens2007 CUDA2012 AlexNet2017 Transformer

★ Outcomes

Amazon.com1995
Online store for everyone

Google1998
Search for the whole web

Netflix streaming2007
Video over the internet

App Store2008
Apps in every pocket

Dropbox2008
Files live in the cloud

UberX2012
Tap a button, get a ride

Slack2014
Chat replaces internal email

Alexa2014
Voice assistant at home

L8 Trust layer

Splunk2003
Search engine for machine logs

Check Pointest. 1993
The firewall goes commercial

New Relic2008
App monitoring as SaaS

CrowdStrike2011
Cloud-native endpoint security

Snyk2015
Security scanning for developers

GDPR / OneTrust2018
Privacy becomes law and software

Datadog IPO2019
Cloud observability goes public

L7 Agent runtime

TIBCO1997
Message bus wiring enterprise apps

PayPal1998
Payments by email

Twilio2008
Phone and SMS through an API

Stripe2010
Payments in a few lines of code

Zapier2011
Glue between SaaS apps

Plaid2013
Apps connect to bank accounts

Temporal2019
Workflows that survive crashes

L6 Builder tools

Eclipse2001
Open-source code editor

Selenium2004
Automated browser testing

GitHub2008
Social coding on Git

Jenkins2011
Default build-and-test server

Jupyter2014
Notebooks for data science

VS Code2015
Free editor everyone adopts

L5 Serving & efficiency

F51996
Load balancers keep sites up

Akamai1998
Content served from the edge

Cloudflare2010
CDN and DDoS shield for all

Fastly2011
Programmable edge cloud

TensorFlow Serving2016
Running ML models in production

TensorRT2017
Faster inference on NVIDIA GPUs

L4 Model building

Deep Blue1997
Computer beats the world chess champion

scikit-learn2010
Classic ML toolkit for Python

IBM Watson2011
Wins Jeopardy!

Google × DeepMind2014
Big Tech buys an AI lab

TensorFlow2015
Deep learning goes open source

Transformer2017
The architecture behind every LLM

L3 Data

Informatica IPO1999
Moving data between systems

MapReduce2004
Crunching data on cheap clusters

Hadoop2006
Big data goes open source

Cloudera2008
Selling Hadoop to enterprises

MongoDB2009
Database for fast-moving developers

Snowflake2012
The warehouse moves to the cloud

Databricks2013
Spark's creators start a company

Scale AI2016
Human labeling at scale

L2 Compute cloud

Rackspace1998
Managed hosting for the web

VMware1998
Virtual machines on ordinary servers

AWS2006
Servers rented by the hour

Heroku2007
Deploy with one command

Azure2010
Microsoft's public cloud

Docker2013
Apps in portable containers

Kubernetes2014
Orchestrating containers at scale

CoreWeave2019
Crypto miner pivots to GPUs

L1 Silicon

Intel Xeon1998
Server chip for data centers

GeForce 2561999
NVIDIA's first GPU

CUDA2007
GPUs opened to general computing

AlexNet2012
Two gaming GPUs spark deep learning

Google TPU2016
Custom chip for neural nets

Cerebras WSE2019
A chip the size of a wafer

L0 Power & buildings

Exodus IPO1998
Colocation goes mainstream

Equinix1998
Where networks interconnect

Google The Dalles2006
Hyperscale campus on cheap hydro

Open Compute2011
Facebook open-sources DC designs

DeepMind cooling2016
AI cuts cooling energy 40%

Switch IPO2017
Hyperscale colocation lists

What the pre-LLM era built

1. **Renting replaced buying.** AWS turned servers into a monthly bill. AWS revenue went from \$7.9B in 2015 to \$35.0B in 2019.
2. **Every job became an API.** Payments, messaging, identity and monitoring each became a company a developer could plug in.
3. **GPUs were waiting in the wings.** CUDA (2007) and AlexNet (2012) set up the AI era years before anyone needed it.

Sources

Company releases and filings; Amazon 10-K (FY2015, FY2019); Wikipedia used only as a backstop for founding dates.

The AI decade, 2020 to 2026

Read each row left to right to see the same job move to new companies. 2025 and 2026 each get their own column because so much happened in them.

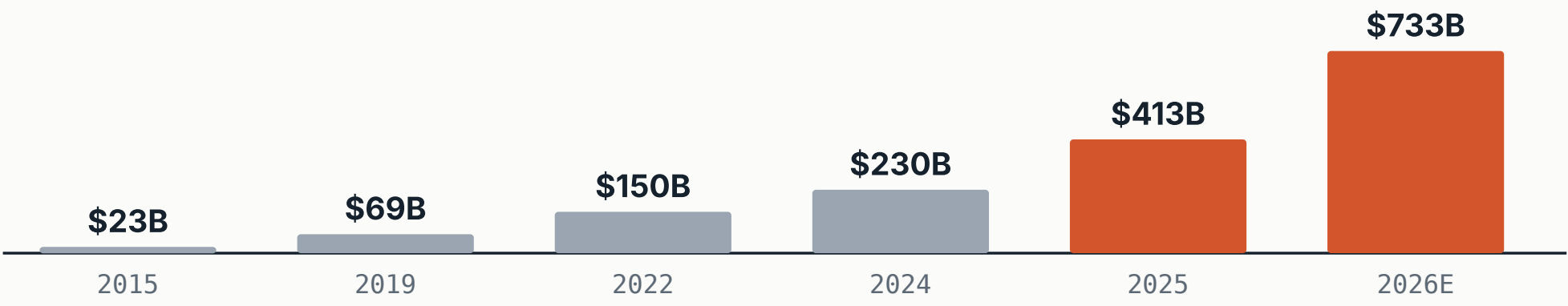
HOW TO READ

Rows are layers. The bottom row is closest to the power plant. Columns are eras. Each card is a company or event, with its date. **Amber** = spotlight startups, in the 2026 column.

	2020–2022	2023–2024	2025	2026 · Jan–Oct
LAYER	Foundation models Train the giant BOTTLENECK GPUs & training data UNIT OF WORK Training run	LLM apps Wrap the model BOTTLENECK GPUs & reliability UNIT OF WORK Token	Agents arrive Let it act BOTTLENECK Power & reliability UNIT OF WORK Task	Agents at scale Pay for outcomes BOTTLENECK Power, memory chips & trust UNIT OF WORK Outcome
ERA →				
TURNING POINTS	2020 GPT-3	Nov 2022 ChatGPT	Nov 2024 MCP	Jan 2025 DeepSeek-R1
				Mar 2026 Vera Rubin
				Jun 2026 Claude Fable 5
★ Outcomes	2020 AlphaFold 2 AI solves protein folding	Nov 2022 ChatGPT LLMs reach everyone	2023 Copilot Business AI pair programmer at work	2023 Harvey AI for law firms
			2023 Perplexity Answer engine for the web	2025 Claude Code Agents write and ship code
				Oct 2025 Figure 03 Humanoids enter pilots
				Sep 2026 ChatGPT 1.2B weekly users
				Sep 2026 Waymo ~500K paid rides a week, 15 cities
				Sep 2026 Harvey \$15.5B, used by 80% of Am Law 100
				SPOTLIGHT STARTUPS Infragrid Nexor AI Sterling
L8 Trust layer	2020 Arize Watching ML models in production	2021 Lakera Defending against prompt injection	2023 Langfuse Open-source LLM tracing	2023 Braintrust Evals for AI products
				2025 Palo Alto × Protect AI Security giants buy AI security
				2025 Check Point × Lakera Prompt defense gets absorbed
				Jan 2026 ClickHouse × Langfuse Database absorbs LLM tracing
				Mar 2026 OpenClaw crisis Open-source agent flaws spread malware
				Apr 2026 Cisco × Galileo Networking giant buys agent evals
				Jul 2026 EU AI Omnibus High-risk rules pushed to Dec 2027
				SPOTLIGHT STARTUPS AgentCat Harden Kairo Abliteration.ai Skyflow Aurva Trusys ZebraTruth
L7 Agent runtime	2021 UiPath IPO Software robots do office work	2021 Okta × Auth0 \$6.5B bet on developer identity	2023 Function calling LLMs learn to call APIs	2024 Browserbase Browsers built for agents
			Nov 2024 MCP One plug between AI and tools	Apr 2025 Google A2A Agents talk to agents
				Sep 2025 Stripe ACP Agents can check out
				Dec 2025 MCP → Linux Fdn Neutral home for the standard
				Feb 2026 OpenAI Frontier Platform to run company agents
				Feb 2026 Nebius × Tavily ~\$275M (reported) for agent search
				Apr 2026 Claude Managed Agents Anthropic hosts agents for firms
				Jul 2026 MCP 2026 spec ~500M SDK downloads a month
				SPOTLIGHT STARTUPS Swytchcode Cluster Rollup Switchfrog
L6 Builder tools	2021 GitHub Copilot First mainstream AI pair programmer	2022 LangChain Chaining LLM calls into apps	2023 Cursor AI-native code editor	2023 E2B Sandboxes where agents run code
				May 2025 Claude Code Agents code from the terminal
				May 2025 OpenAI Codex Cloud agent writes code in parallel
				Nov 2025 Cursor \$1B annual revenue run rate
				Feb 2026 Claude Code Over \$2.5B run rate
				Apr 2026 OpenAI Codex 4M active users
				Jun 2026 SpaceX × Cursor \$60B all-stock deal
				Aug 2026 CodeRabbit \$143M at \$1.5B for AI code review
				SPOTLIGHT STARTUPS Cube Computer TrueCourse PreMan
L5 Serving & efficiency	2022 FlashAttention Cheaper attention, longer prompts	2022 Fireworks AI Hosted inference for open models	2023 vLLM Memory paging boosts LLM throughput	2023 OpenRouter One API, hundreds of models
				2024 GroqCloud Record tokens per second
				Oct 2025 Fireworks AI \$250M at \$4B
				Dec 2025 NVIDIA × Groq ~\$20B license and team deal
				Jan 2026 vLLM → Inferact Open-source engine becomes a startup
				May 2026 Google 3.2 quadrillion tokens a month, 7× in a year
				Jun 2026 Baseten \$1.5B at \$13B
				Jul 2026 Fireworks AI \$1.5B at \$17.5B
				SPOTLIGHT STARTUPS Pipeshift Hold Minima
L4 Model building	2020 GPT-3 Scale unlocks new abilities	2022 Stable Diffusion Open image model for anyone	2023 Llama Meta releases open weights	2023 Claude Anthropic's first model
				2023 Mistral Europe's open-weight lab
				Jan 2025 DeepSeek-R1 Cheap reasoning model jolts markets
				Jan 2025 NVIDIA Cosmos World models for robots
				Feb 2025 Thinking Machines New frontier lab
				Mar 2026 OpenAI \$122B round at \$852B
				Apr 2026 DeepSeek V4 Near-frontier open weights, low price
				May 2026 Anthropic \$65B at \$965B, \$47B run rate
				Jun 2026 Claude Fable 5, Mythos 5 Paused Jun 12 for US export rules, back Jul 1
				SPOTLIGHT STARTUPS Oruk Zorbe Proper Motion Outerbounds
L3 Data	2020 Snowflake IPO Largest software IPO to date	2020 Surge AI Expert feedback for AI labs	2021 Pinecone Vector database launches	2023 Mercor Experts hired to train models
				2024 Reddit × Google ~\$60M a year for training data
				Jun 2025 Meta × Scale AI \$14.3B for a 49% stake
				Jan 2026 ClickHouse \$400M at \$15B
				Aug 2026 Databricks \$5B at \$190B, \$7B+ run rate
				Oct 2026 Mecka AI \$60M for robot training data
				SPOTLIGHT STARTUPS SF Data SenseLab Embrasure Space Munari Vound
L2 Compute cloud	2020 Microsoft × OpenAI 10,000-GPU supercomputer on Azure	2022 Together AI Open-source AI cloud	May 2024 CoreWeave \$8.6B raised in one month	2024 NVIDIA × Run:ai GPU scheduling bought by NVIDIA
				Mar 2025 CoreWeave IPO First AI neocloud goes public
				Sep 2025 Nebius × Microsoft \$17.4B GPU capacity deal
				Mar 2026 Meta × Nebius \$27B capacity deal
				Apr 2026 Meta × CoreWeave \$21B through 2032
				Jun 2026 Oracle \$638B contracted backlog
				Jul 2026 Nscale × Anyscale Neocloud buys the Ray company
				SPOTLIGHT STARTUPS Harbor AI
L1 Silicon	2020 NVIDIA A100 The foundation-model workhorse	2020 NVIDIA × Mellanox \$6.9B for chip networking	2023 NVIDIA H100 The scarce currency of AI	May 2023 NVIDIA \$1T Chipmaker joins trillion club
				Oct 2025 NVIDIA \$5T First company worth \$5 trillion
				Oct 2025 OpenAI × Broadcom 10 GW of custom AI chips
				Feb 2026 AMD × Meta Up to 6 GW of AMD GPUs
				Mar 2026 NVIDIA Vera Rubin Next platform enters production
				Apr 2026 Anthropic × Google TPU 3.5 GW of TPUs from 2027
				May 2026 Cerebras IPO \$5.55B raised, year's largest tech IPO
				SPOTLIGHT STARTUPS CIMTech VerifAIX
L0 Power & buildings	2020 Vertiv Power and cooling maker lists	Sep 2024 Microsoft × Constellation Restarting a nuclear reactor for AI	Oct 2024 Crusoe Abilene AI campus breaks ground in Texas	Jan 2025 Stargate Up to \$500B pledged for AI DCs
				Jan 2026 Meta nuclear deals 6.6 GW of nuclear by 2035
				Mar 2026 Ecolab × CoolIT \$4.75B for liquid cooling
				Jul 2026 PJM auction Grid power price hits the cap
				Sep 2026 Crusoe \$3.9B at \$30.9B
				SPOTLIGHT STARTUPS Cybernetic Physics

The bill for all this

Combined capital spending of Microsoft, Alphabet, Amazon and Meta, \$B per year. 2026 is guidance after Q2 2026 earnings: \$720–745B, midpoint shown.



What 2026 changed

- Labs became the biggest private companies on earth.** OpenAI raised \$122B and Anthropic \$65B in a single year.
- The bottleneck moved to power and memory.** Grid prices hit their cap and memory makers posted record margins.
- Tooling got bought.** Cursor, Langfuse, Galileo, Tavily and Anyscale were bought, or agreed to be bought, by larger platforms.

Sources

Company releases and SEC filings; Platformonomics and Q2 2026 earnings calls (capex); 10-K cash capex for 2015–24, approximate (Microsoft by fiscal year); TechCrunch, Bloomberg, CNBC, DCD for 2026 deals. Waymo, ChatGPT and revenue run rates are company-reported.

Same jobs, new names

Each layer does the same job it did in the SaaS era. What changed is how the job is done and how a developer judges if it is done well.

- 1

The unit of work

page view → API call → token → task
- 2

The user

a person → code a person wrote → an agent
- 3

The bottleneck

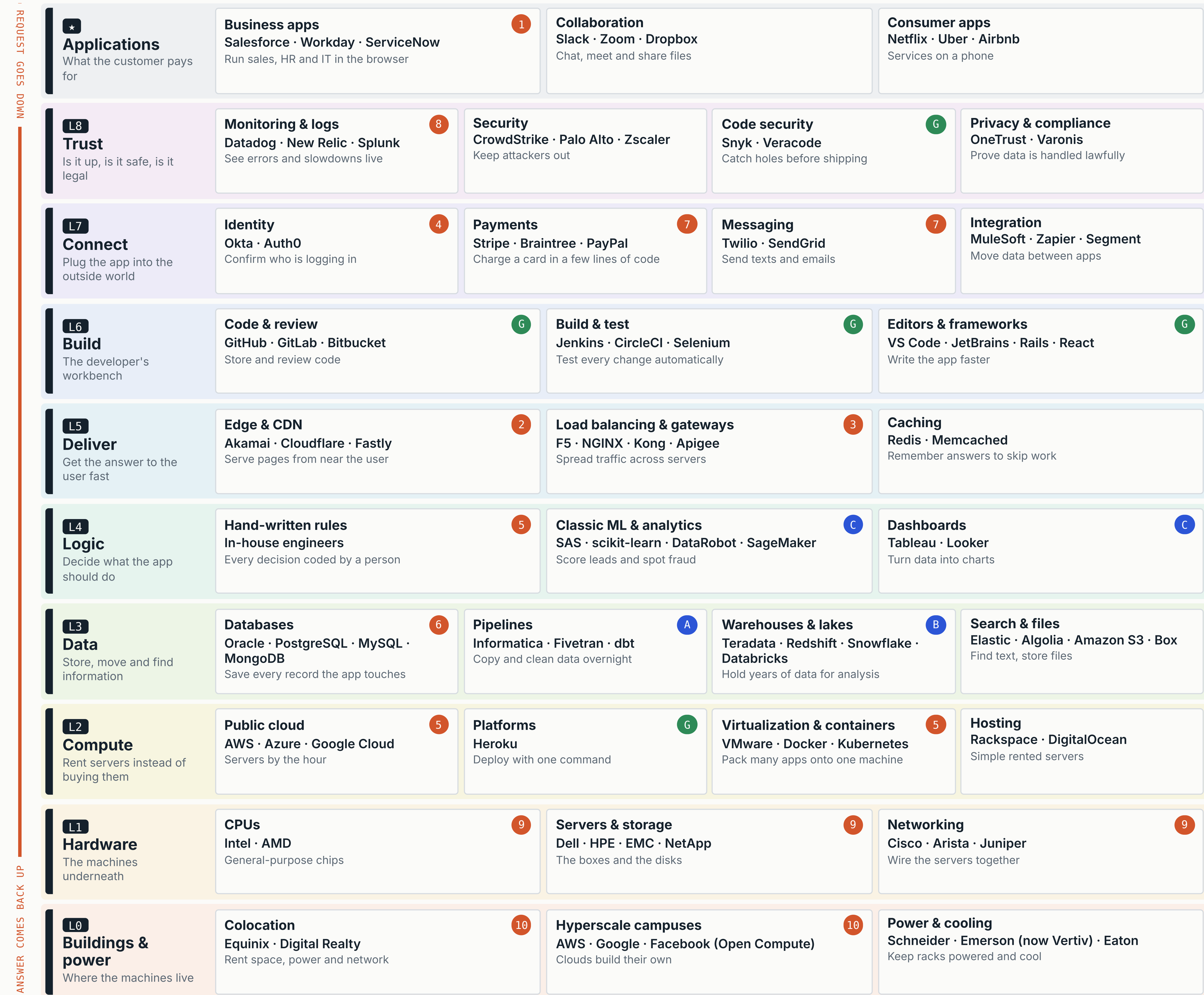
hardware → code → data → GPUs → power

SAAS ERA · 2006–2019			WHAT CHANGED	AI ERA · 2023–2026	
★ Outcomes Deliver software people pay for	Salesforce, Workday, Dropbox Humans click through apps to get work done	SCORED BY Seats sold	seat → outcome	Cursor, Harvey, Waymo Software does the task itself	SCORED BY Tasks completed
L8 Trust layer Know it works and keep it safe	Datadog, CrowdStrike, OneTrust Is the server up? Was the network breached?	SCORED BY Uptime %	uptime → correctness	Braintrust, Noma, Credo AI Is the answer right? Did the agent do something it should not?	SCORED BY Eval pass rate
L7 Agent runtime Connect software to the outside world	Twilio, Stripe, Okta, Zapier APIs called by code a human wrote in advance	SCORED BY API calls	person → agent	MCP, Composio, Stripe ACP, WorkOS The model picks and calls the tools itself	SCORED BY Tool-call success
L6 Builder tools Help developers ship	GitHub, Jenkins, Selenium Humans write the code, machines test it	SCORED BY Deploys per day	writing → reviewing	Claude Code, Cursor, E2B, CodeRabbit AI writes the code, other tools check it	SCORED BY AI code merged safely
L5 Serving & efficiency Answer fast and cheap	Akamai, Cloudflare, F5 Cache one page and send it to millions	SCORED BY Page load time	cached → generated	Fireworks, vLLM, Baseten Every answer is computed fresh, word by word	SCORED BY \$ per million tokens
L4 Model building Turn knowledge into working logic	Engineers + SAS, scikit-learn Logic is written by hand as rules	SCORED BY Rules shipped	coded → trained	OpenAI, Anthropic, Meta Llama Logic is learned from data	SCORED BY Benchmark score
L3 Data Store, clean and find information	Oracle, Snowflake, Elastic Rows and tables feed reports	SCORED BY Query speed	tables → meaning	Databricks, Pinecone, Scale AI Text, video and expert judgment feed models	SCORED BY Retrieval accuracy, data quality
L2 Compute cloud Rent computing on demand	AWS, VMware, Heroku Cheap, plentiful CPUs rented by the hour	SCORED BY \$ per server-hour	CPU → GPU	CoreWeave, Nebius, Oracle Scarce GPUs booked for years in advance	SCORED BY \$ per GPU-hour, utilization
L1 Silicon Do the math	Intel, Cisco General chips doing one thing at a time	SCORED BY Clock speed	serial → parallel	NVIDIA, Broadcom, SK hynix Specialized chips doing millions of sums in parallel	SCORED BY Compute per watt, memory bandwidth
L0 Power & buildings House and feed the machines	Equinix, Digital Realty Racks drew under 6 kW. Power was cheap and easy	SCORED BY Square feet	space → megawatts	Crusoe, Constellation, Vertiv One AI rack draws ~130 kW. Builders sign nuclear deals	SCORED BY Megawatts

The SaaS-era value chain

How one click in a business app traveled through every layer from about 2015 to 2019, and who got paid at each step. The AI-era version on the next page uses the same layers.

- 1 The request, top to bottom
- A The data, overnight
- G The developer, before launch



One click, end to end

- | | | |
|----|--|-------------------------|
| 1 | You click Save in a sales app | Salesforce |
| 2 | The edge network nearest you picks it up | Akamai, Cloudflare |
| 3 | A load balancer sends it to a free server | F5, NGINX |
| 4 | The app checks who you are | Okta |
| 5 | Code a person wrote runs on a rented server | AWS, Docker |
| 6 | The record is saved to the database | Oracle, PostgreSQL |
| 7 | An email goes out and a card is charged | SendGrid, Stripe |
| 8 | Everything is logged and watched | Datadog, Splunk |
| 9 | All on general-purpose chips | Intel, Dell |
| 10 | Inside a rented building on the grid | Equinix, AWS |
| A | Overnight, data is copied out | Fivetran |
| B | Stored for analysis | Snowflake |
| C | Analysts and classic ML score the leads | Tableau, SAS |
| G | Meanwhile a developer writes, tests and ships the code | GitHub, Jenkins, Heroku |

Who got paid

33 / 18 / 8%

Cloud share of AWS, Microsoft, Google, Q4 2019

\$35.0B

AWS revenue 2019, up from \$7.9B in 2015

\$17.1B

Salesforce revenue, fiscal 2020

\$35B

Stripe valuation, Sep 2019

\$5.6B

Equinix revenue 2019

\$7.5B

Microsoft buys GitHub, 28M+ developers, 2018

Where the value pooled

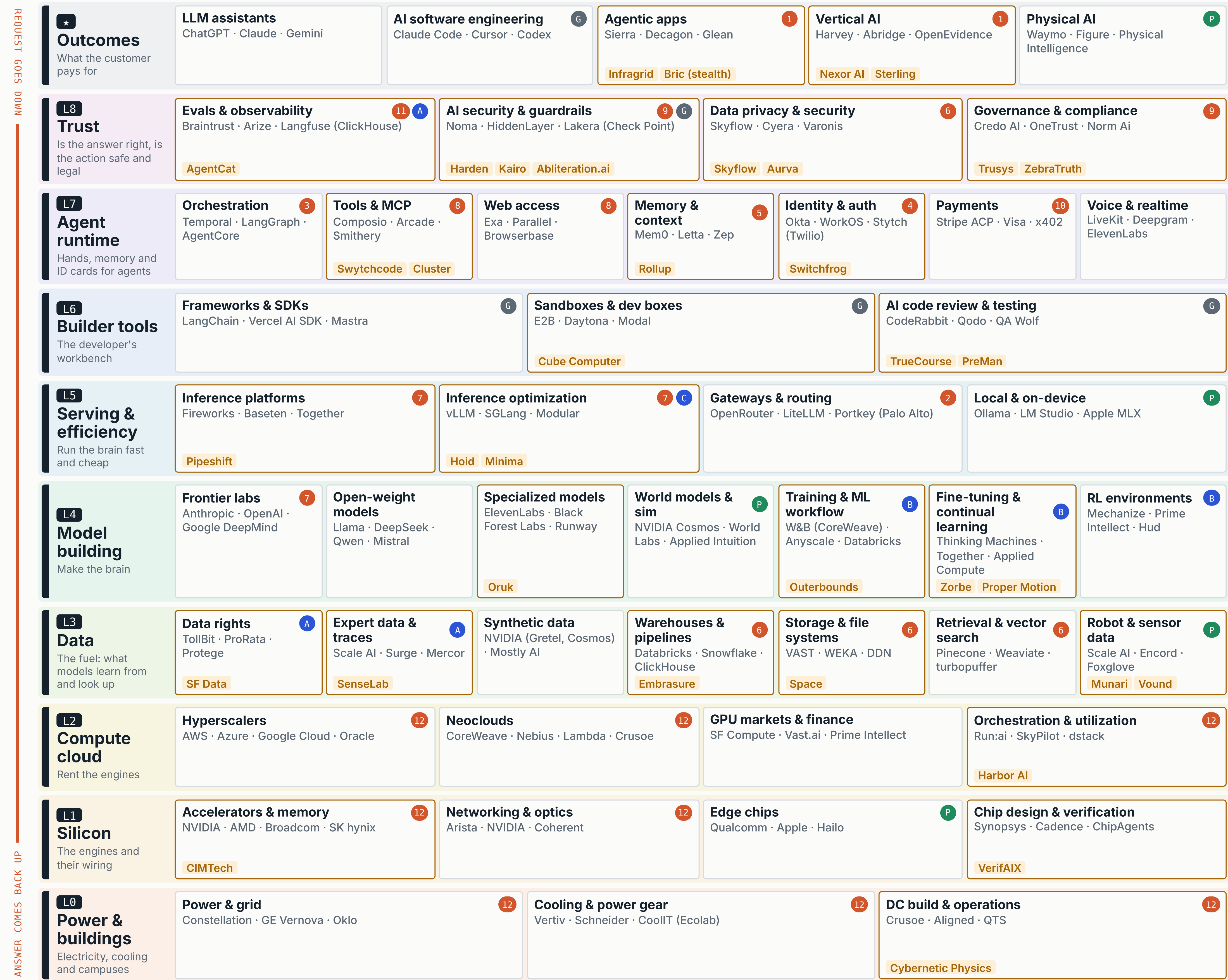
At the two ends. Apps kept about **73%** gross margins (2022 survey), and three clouds took most of the infrastructure. Everything in between became a rentable API, and **84%** of companies ran containers in production by 2019.

Sources: Amazon 10-K; Synergy Research via SDxCentral; Salesforce and Equinix 8-Ks; CNBC (Stripe); Microsoft (GitHub); KeyBanc 2022 SaaS survey; CNCF 2019 survey.

The AI-era value chain

The same ten layers as the SaaS map, now serving an agent instead of a person clicking. Follow one request down the stack, then the loop that makes the model better. Amber names are spotlight startups.

- 1 The request, top to bottom
- A The learning loop
- G The developer, before launch
- P The physical AI branch



One request, end to end

- | | | |
|----|---|--|
| 1 | You ask an agent: renew Acme's contract and email them | Sierra, Glean |
| 2 | A gateway picks the right model for the job | OpenRouter |
| 3 | The runtime plans the steps and keeps the job alive | Temporal, LangGraph |
| 4 | The agent proves who it is and what it may touch | Okta, Styx |
| 5 | It recalls what it knows about Acme | Mem0, Zep |
| 6 | It looks up the contract in company data | Databricks, Pinecone |
| 7 | A model thinks, served on fast inference | Claude via API; open models on Fireworks, vLLM |
| 8 | It calls the CRM, e-sign and email tools via MCP | Composio, Exa |
| 9 | Guardrails check the action before it runs | Noma, Norm AI |
| 10 | If money moves, an agent payment rail settles it | Stripe ACP |
| 11 | Every step is traced and scored | Braintrust, Langfuse |
| 12 | All on GPUs in a neocloud, on a gigawatt campus | CoreWeave, NVIDIA, Crusoe |
| A | Good and bad runs are saved as training data | Scale AI, Langfuse |
| B | They train a cheaper, sharper model | Thinking Machines, RL envs |
| C | The new model is shrunk and sped up, then goes back to step 7 | vLLM, Modular |
| G | Before launch, a developer builds and tests the agent with AI | Claude Code, E2B, CodeRabbit |
| P | Physical AI branch: robot data, world models, edge chips, robot | Encord, Cosmos, Qualcomm, Figure |

Who gets paid in 2026

\$720-745B

Big Tech capex guidance for 2026

\$96.2B
NVIDIA revenue in one quarter (ended Jul 26, 2026)

~\$65B
Anthropic revenue run rate, end of Jul
2026 (reported)

40 / 27 / 21%
Enterprise LLM API share: Anthropic,
OpenAI, Google (2025)

3.2 quadrillion	~500M
Tokens a month through Google, May 2026	MCP SDK downloads a month, Jul 2026

Where the value pools now

SaaS value pooled at the ends. AI value pools at the **bottom**: every answer burns compute, so chips, power and the labs collect first, and AI apps keep about **45%** gross margin vs **73%** for SaaS. The middle layers win by cutting that cost or that risk. That is where most spotlight startups sit.

Sources: company Q2 2026 earnings and releases; Anthropic Series H post; CNBC (Aug 2026); Menlo Ventures 2025; Google I/O 2026; MCP blog (Jul 2026); KeyBanc 2022; ICONIQ 2026. Example flow is illustrative; any vendor in a box can fill that step.

LO Power & buildings

AI is now limited less by chips than by electricity and buildings. Power is only about 7% of an AI campus's yearly cost, but it decides who can build, where and when.

4.4% → 6.7-12%

US electricity used by data centers, 2023 → 2028
LBNL, Dec 2024

4+ years

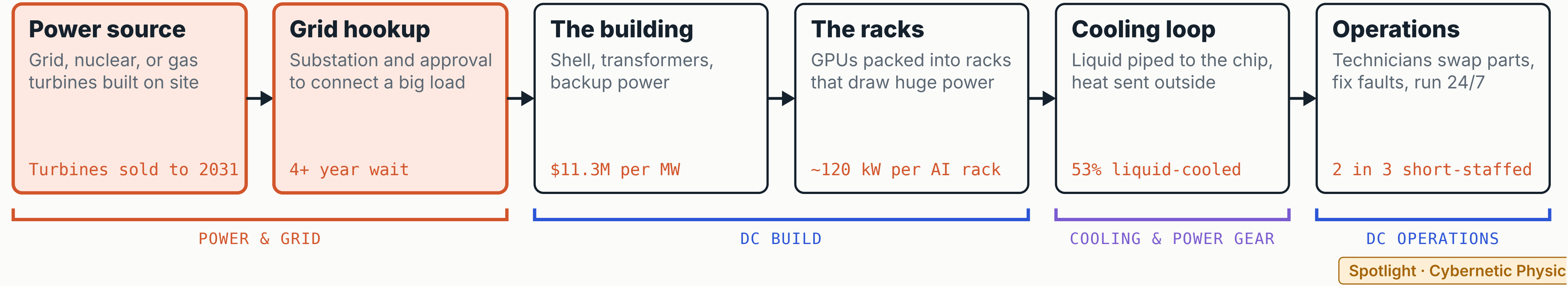
Wait for a grid connection in top markets
JLL, Jan 2026

<8 → ~120 kW

Power per rack: average 2024 vs one NVIDIA GB200 rack
Uptime; SemiAnalysis

How it works

Follow the electricity from the power plant to the chip. Red boxes are today's bottleneck.

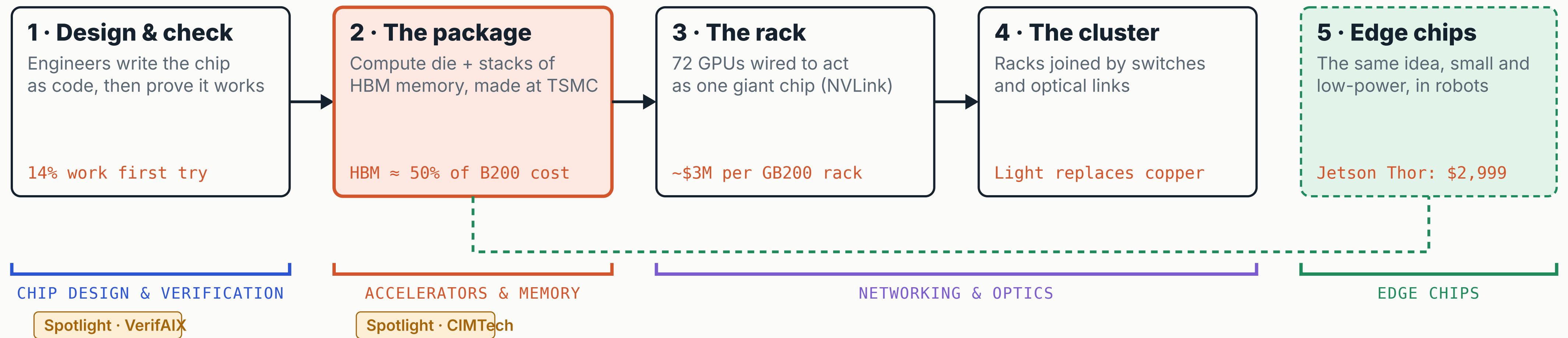


L1 Silicon

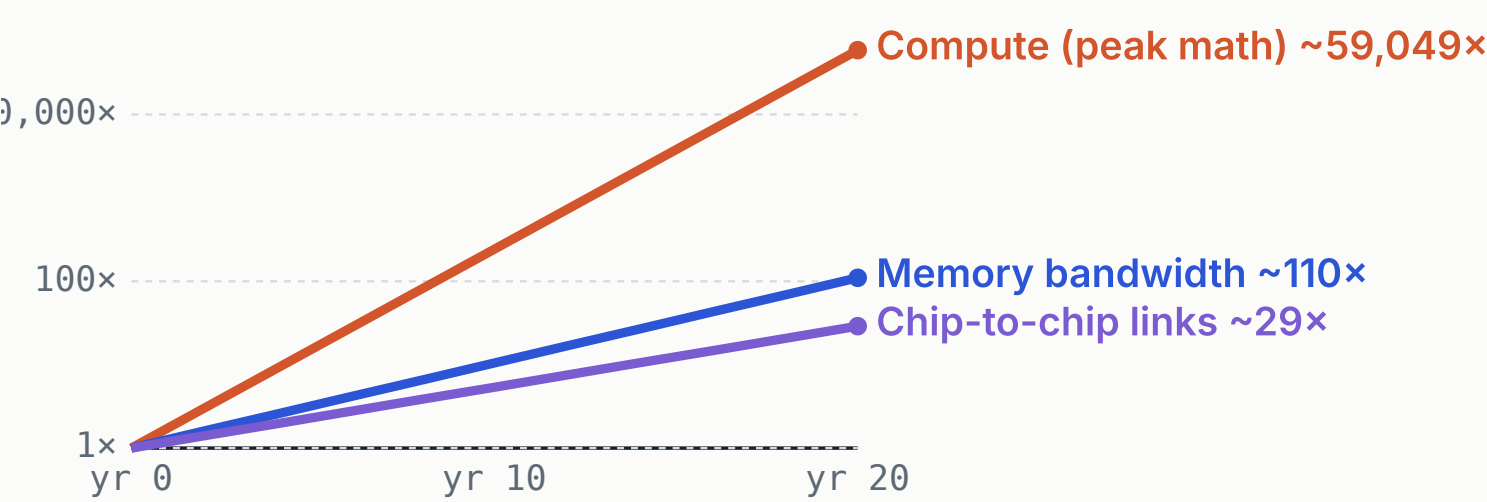
Chips do the math. In 2026 the limit is feeding them: memory and the links between chips. Whoever moves data faster per watt sets much of the price of every AI answer.

How it works

From a design file to a cluster of thousands of chips. Red boxes are today's bottleneck.

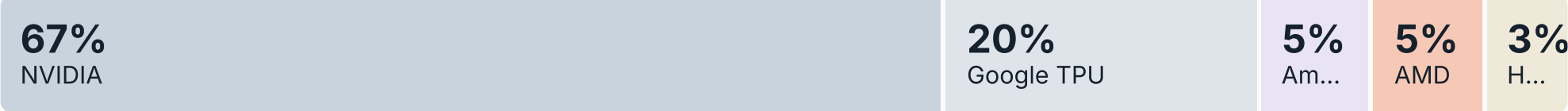


The memory wall: chips calculate far faster than they can be fed data



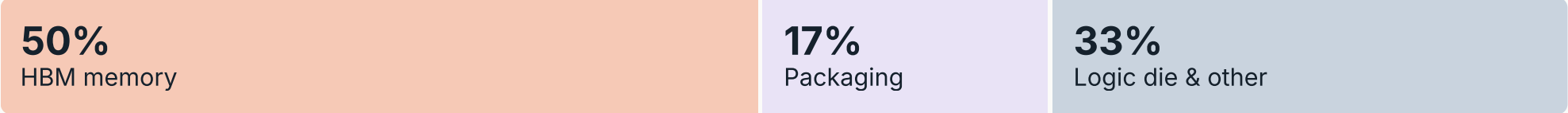
Growth over 20 years at 3.0x, 1.6x and 1.4x every two years. Gholami et al., 2024. When an AI writes an answer, each word means re-reading memory, so memory speed sets the pace.

Who shipped the world's AI compute, Q4 2025



Share of compute shipped, measured in H100-equivalents. Epoch AI, Oct 2026.

What it costs to build an NVIDIA B200

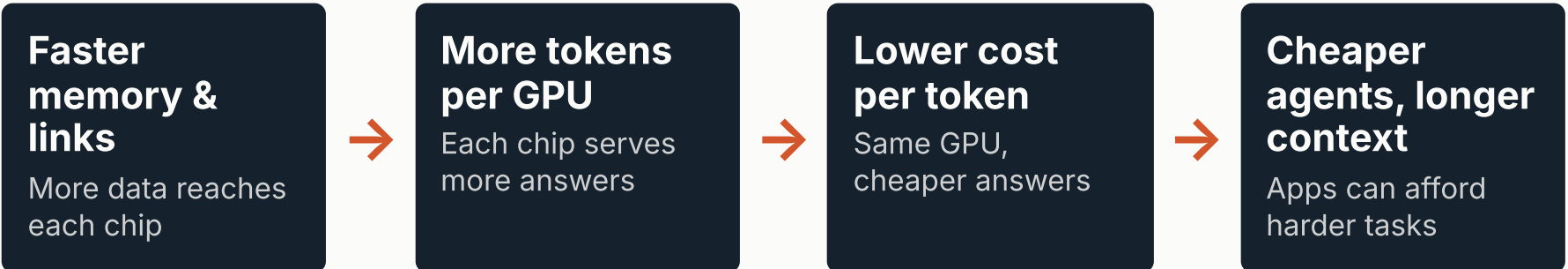


Approximate share of ~\$6,400 build cost. Epoch AI, Dec 2025. HBM prices are forecast to rise 121% in 2027 (TrendForce).

Same job, new problem: the micro-themes

Accelerators & memory Do the math and hold the model	SAAS ERA General Intel CPUs. Memory was cheap and plentiful.	AI ERA, 2026 GPUs and custom chips. HBM memory is about half of a B200's build cost and in short supply.	LEADERS NVIDIA · AMD · Google TPU · Broadcom · SK hynix 2026 MOVES • NVIDIA Vera Rubin claims up to 10x lower cost per token vs Blackwell • AMD signs Meta (6 GW) and Anthropic (2 GW) • Cerebras IPO raises \$5.55B; SK hynix posts 76% operating margin	MOVES THIS NUMBER Cost per token, tokens per second per watt EARLY-STAGE STARTUPS CIMTech Great Sky
Networking & optics Make thousands of chips act as one	SAAS ERA Cisco switches between ordinary servers.	AI ERA, 2026 GPUs wired inside a rack (scale-up), across racks (scale-out) and across sites. Light replaces copper.	LEADERS NVIDIA · Arista · Broadcom · Coherent · Ayar Labs 2026 MOVES • Marvell closes \$3.25B Celestial AI deal (Feb) • Ayar Labs raises \$650M in 2026 • Arista targets \$3.6B AI networking revenue	MOVES THIS NUMBER GPU utilization, time to train EARLY-STAGE STARTUPS Netris NcodiN
Edge chips Run AI inside robots, cars and PCs	SAAS ERA Phone chips ran apps, not models.	AI ERA, 2026 Robot-grade chips: NVIDIA Jetson Thor gives 7.5x its predecessor's AI compute for \$2,999 (module; dev kit \$3,499).	LEADERS NVIDIA Jetson · Qualcomm · Apple · Axelera 2026 MOVES • Qualcomm robot chip at 350 TOPS (Jan) • Microchip agrees to buy Hailo (Jul); Axelera raises \$250M+	MOVES THIS NUMBER AI per watt on the device, response time EARLY-STAGE STARTUPS HYFIX HRDWYR Mosaic SoC
Chip design & verification Design chips and prove they work	SAAS ERA Synopsys and Cadence tools for human designers.	AI ERA, 2026 AI writes chip designs, so checking them matters more. Engineers already spend 49% of their time verifying.	LEADERS Synopsys · Cadence · ChipAgents · Cognichip 2026 MOVES • Ricursive raises \$300M Series A (Jan) • ChipAgents expands Series A to \$134M (Jul) • VerifAIX \$5M seed (Sep)	MOVES THIS NUMBER First-pass success, months to tape-out EARLY-STAGE STARTUPS VerifAIX Architect Labs

How this layer moves the outcome



If this layer fails

Everything above stalls. A memory shortage already squeezes NVIDIA, which guides next quarter's gross margin to about 74% as memory costs rise. A chip bug found after manufacturing means a costly respin and months lost.

White space

No spotlight startup yet in networking, optics or edge chips. Silicon is capital-heavy, so look for software that makes chips go further: memory pooling, interconnect software, chip verification and compilers.

Spotlight startups here

CIMTech	ACCELERATORS & MEMORY · ALSO L5 SERVING
THE BET	An SRAM card placed beside GPUs speeds up the memory-bound part of long answers, so operators serve more users per watt.
PROOF SO FAR	SkyDeck Batch 22, Georgia Tech spinout, co-founder is an IEEE Fellow. Claims 10x faster long-context inference (company claim, untested).
MUST BE TRUE	Operators accept an add-in card, it plugs into common serving software like vLLM, and it beats NVIDIA's own roadmap per watt.
RISKS	NVIDIA closes the gap (Rubin's HBM4, Groq-based racks); very well-funded peers: d-Matrix (~\$2B value), Positron (\$5B), Etched (\$10.3B).
ASK THE FOUNDER	Tokens per second per watt vs Rubin at 128K context? Tape-out status and foundry? Who pilots it first?

VerifAIX	CHIP DESIGN & VERIFICATION
THE BET	As AI writes more chip code, automated formal checking becomes the bottleneck, and buyers pay to avoid respins.
PROOF SO FAR	\$5M seed (Sep 2026, Endiya, Bluehill VC); says it is deployed at several unnamed chip companies; founders from Intel, Apple and Cadence formal teams.
MUST BE TRUE	Chipmakers trust a startup's proofs at sign-off, and it fits inside Synopsys and Cadence workflows.
RISKS	Incumbents bundle AI into their suites; better-funded AI design peers (ChipAgents \$134M, Ricursive \$300M).
ASK THE FOUNDER	Bugs found that incumbent tools missed? Time saved to full coverage? Price per seat or per design?

What changes by 2027–2030

- Inference is two-thirds of AI compute in 2026; inference chips pass \$50B. Deloitte, Nov 2025
- Custom AI chip (ASIC) shipments triple by 2027. Counterpoint, Jan 2026
- HBM memory prices rise 121% in 2027, keeping memory the bottleneck. TrendForce, Oct 2026
- Optical links move onto the chip; Marvell targets a \$1B photonics run rate by FY29. Marvell

Sources

NVIDIA Q2 FY27 release and call (Aug 2026); AMD 8-Ks and newsroom (2025–26); Broadcom FQ3'26 release; Epoch AI (compute share Oct 2026; B200 cost Dec 2025); Gholami et al. arXiv 2403.14123 (2024); Counterpoint HBM share (Sep 2026); SK hynix Q2 2026; TrendForce (HBM, Oct 2026); Siemens/Wilson Research verification studies (2022, 2024); Synopsys and Cadence filings; Marvell and Ayar Labs releases; SemiEngineering startup funding Q1 2026; company sites for CIMTech and VerifAIX.

L2 Compute cloud

Cloud used to mean cheap computers rented by the hour. AI cloud means scarce GPUs, bought with debt and booked years ahead. The winners keep every GPU busy and every contract filled.

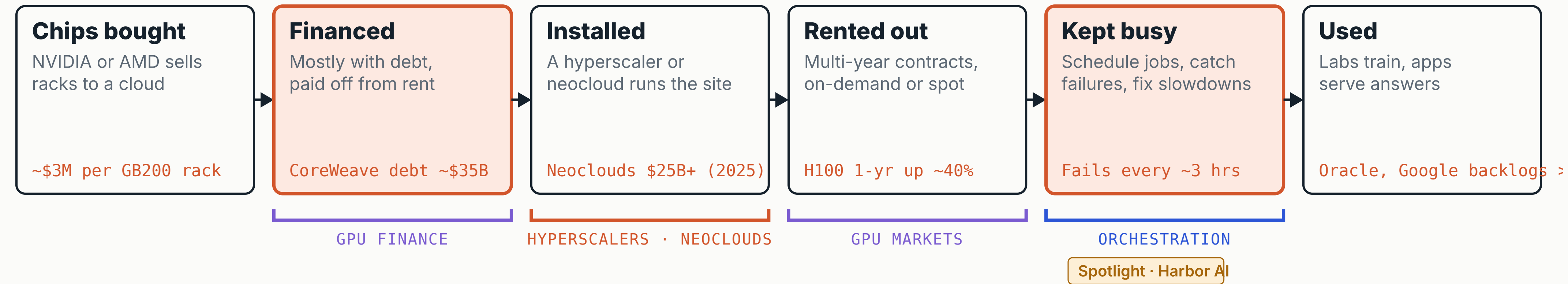
\$143.4B Cloud infrastructure spend in one quarter, +43%
Synergy, Q2 2026

\$6.62 → \$2.80 H100 rental per hour, 2H 2023 → Mar 2026
SemiAnalysis index

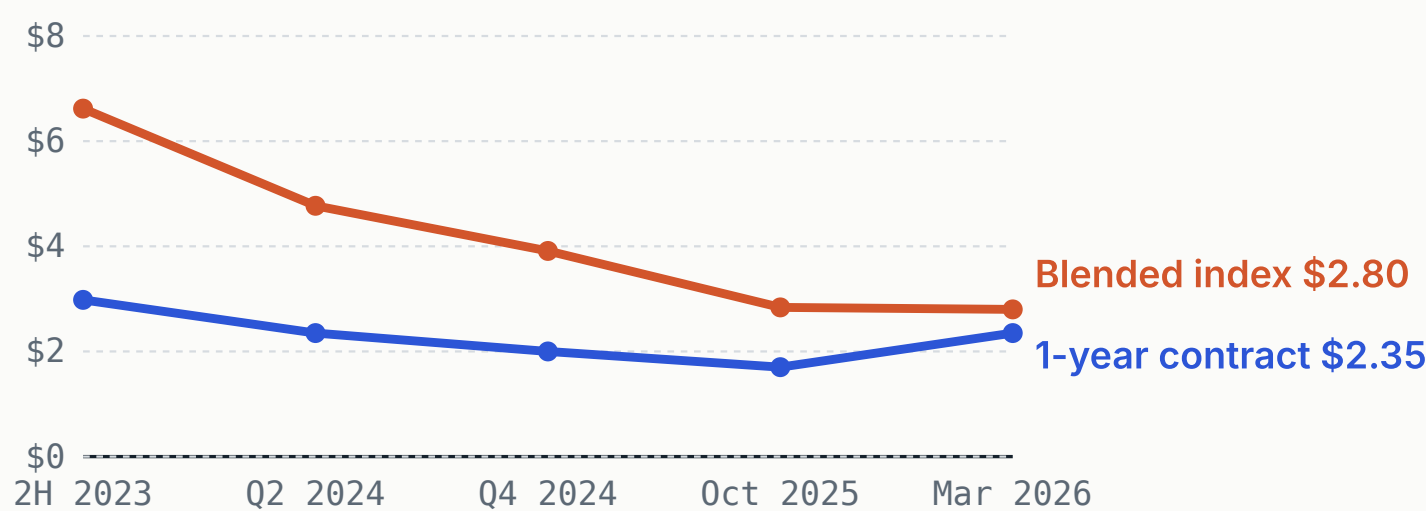
~\$35B CoreWeave's debt, with a \$104B contract backlog
CoreWeave 10-Q, Jun 2026

How it works

How a GPU becomes rented compute. Red boxes are today's bottleneck.

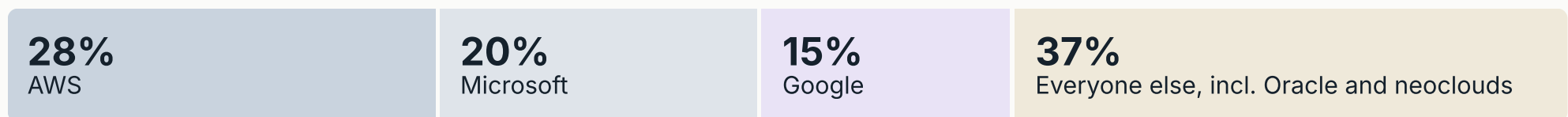


Renting one H100 GPU for an hour: the crash, then the squeeze



SemiAnalysis GPU index. Prices fell about 58% as supply caught up, then contract rates rose ~40% from Oct 2025 to Mar 2026 and on-demand H100s sold out. 1-year values are range midpoints.

Who rents the world's cloud, Q2 2026



\$143.4B spent in the quarter, +43% a year. Synergy Research, Jul 2026.

Why keeping GPUs busy is hard

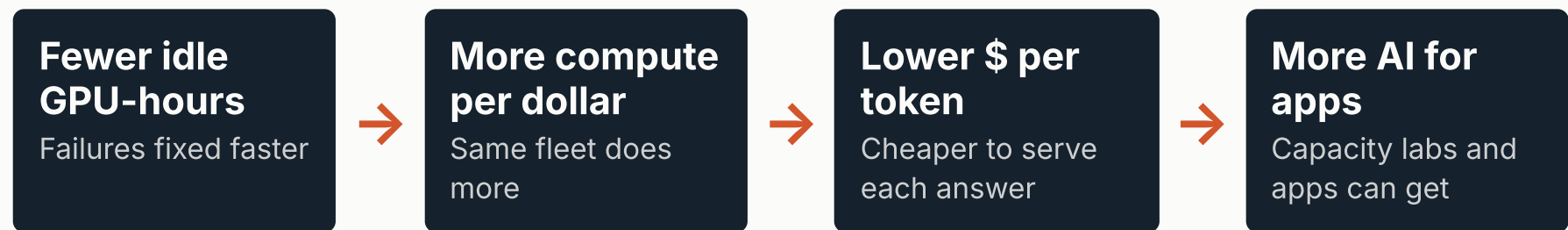
Meta's Llama 3 run on 16,384 H100s hit 419 unexpected interruptions in 54 days, about 1 every 3 hours. Almost half came from faulty GPUs or their memory. Even so it used only 38–43% of the chips' peak math.

Meta, Llama 3 paper, Jul 2024

Same job, new problem: the micro-themes

Hyperscalers Rent computing to everyone	SAAS ERA Cheap, plentiful CPUs rented by the hour.	AI ERA, 2026 GPU capacity sold years ahead. Google Cloud's backlog is \$514B and Oracle's is \$638B.	LEADERS AWS · Azure · Google Cloud · Oracle 2026 MOVES - Cloud market grows 43% in a year (Q2) - Google Cloud revenue +82% (Q2) - Oracle backlog jumps to \$638B (Jun)	MOVES THIS NUMBER Capacity delivered, \$ per GPU-hour EARLY-STAGE STARTUPS Hosted.ai (sells to clouds)
Neoclouds GPU-only clouds for AI	SAAS ERA Small hosts like Rackspace and DigitalOcean.	AI ERA, 2026 Specialists with giant contracts and giant debt. Revenue passed \$25B in 2025, growing over 200% a year.	LEADERS CoreWeave · Nebius · Lambda · Crusoe · Nscale 2026 MOVES - Meta signs \$27B with Nebius and \$21B with CoreWeave - Lambda raising \$4B ahead of a 2027 IPO - Crusoe raises \$3.9B at \$30.9B	MOVES THIS NUMBER Contract backlog, cost of borrowing EARLY-STAGE STARTUPS Nava (GPU cloud) General Compute Hyperbolic NexGen Cloud
GPU markets & finance Price, trade and fund GPU time	SAAS ERA AWS spot instances and equipment leasing.	AI ERA, 2026 GPUs are financed with debt and traded like a commodity. Futures contracts on GPU-hours are launching.	LEADERS SF Compute · Vast.ai · Silicon Data · Ornn 2026 MOVES - CME and ICE announce GPU compute futures (pending approval) 1-year H100 rates rise ~40% (Oct 2025 → Mar 2026)	MOVES THIS NUMBER Price spread, idle capacity sold EARLY-STAGE STARTUPS Liquid Compute
Orchestration & utilization Keep every GPU working	SAAS ERA VMware packed many apps onto one server.	AI ERA, 2026 Huge clusters fail every few hours and use well under half their peak power. Software that schedules and repairs them saves real money.	LEADERS NVIDIA Run:ai · Anyscale · SkyPilot · Clockwork 2026 MOVES - Nscale agrees to buy Anyscale, reported \$1.65B (Jul) - Clockwork raises \$31M; SkyPilot \$20M seed	MOVES THIS NUMBER GPU utilization, hours lost to failures EARLY-STAGE STARTUPS Harbor AI Zymtrace Chamber (YC)

How this layer moves the outcome



If this layer fails

Labs and apps can't get capacity, as when on-demand H100s sold out in early 2026. On the money side, a neocloud carrying billions in debt that loses a customer is left with GPUs it writes off over 4–6 years while NVIDIA ships a new generation every 12–18 months.

White space

Only Harbor AI here. Gaps: GPU pricing data and marketplaces, compute financing and residual-value insurance for GPUs, and scheduling across clouds.

Spotlight startups here

Harbor AI	ORCHESTRATION & UTILIZATION
THE BET	Owned GPU fleets waste money on failures and slowdowns. Software that finds the cause, proposes a fix and checks the recovery pays for itself.
PROOF SO FAR	Founded 2026, two people. CEO ran GPU deployments for frontier labs at Microsoft AI Supercomputing.
MUST BE TRUE	Operators let a third party act on production fleets, and the fixes work across NVIDIA, AMD and different clouds.
RISKS	Clouds and chipmakers build this themselves (CoreWeave and NVIDIA Mission Control, Run:ai). Funded peers: Clockwork (\$73M total), Zymtrace (\$12.2M).
ASK THE FOUNDER	GPU-hours recovered per 1,000 GPUs per month? Which fixes run automatically and which need sign-off? Size of the first paying fleet?

What changes by 2027–2030

1. **Neocloud revenue approaches \$400B by 2031, growing 58% a year.** Synergy, Apr 2026
2. **Neoclouds take 20% of a \$267B AI cloud market by 2030.** Gartner, Jun 2026
3. **Lambda plans a 2027 IPO; Crusoe has met banks.** WSJ via TechCrunch, 2026
4. **GPU-hours trade on futures exchanges once approved.** CME, ICE releases 2026

Sources

Synergy Research (Q2 2020; neocloud forecasts Oct 2025, Apr 2026); company filings and releases (CoreWeave, Nebius, Oracle, Crusoe); Motley Fool on Q2 cloud results; SemiAnalysis GPU index; Meta Llama 3 paper (Jul 2024); CME and ICE releases; TechCrunch, Bloomberg, DCD, Forbes, Fortune, SiliconANGLE (deals and raises); Gartner (Jun 2026); a16z Speedrun (Harbor AI).

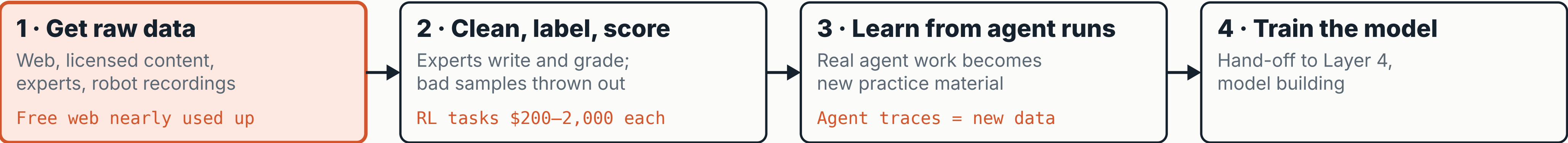
L3 Data · part 1

The free internet is nearly used up. Data now has to be bought, written by experts, recorded in the real world or captured from agents at work, then stored and found fast.

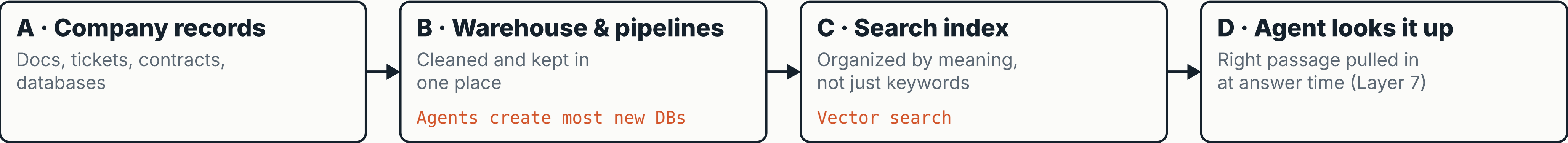
How it works

Two journeys for data: into the model when it is trained, and into the answer when an agent looks something up.

TRAINING · teach the model



ANSWERING · find the right facts



Storage & file systems under both: feeds GPUs fast and saves progress after crashes

Expert data is now a billion-dollar business

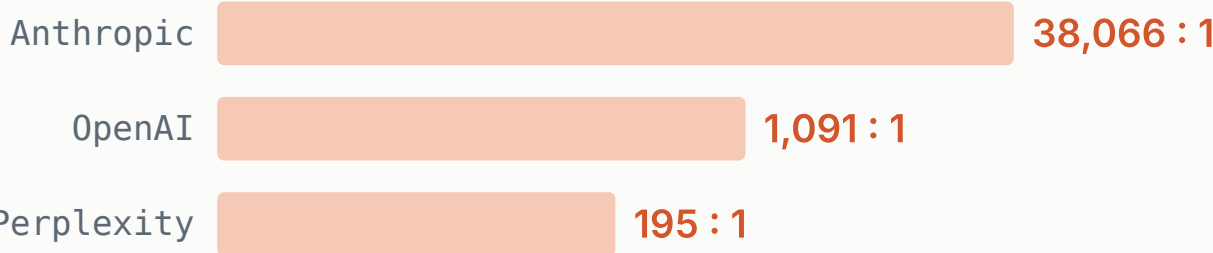
Mercor >\$2B run rate · Jul 2026

Surge ~\$1.2B revenue · 2024

Handshake AI ~\$1B run rate · Apr 2026

Annual revenue or run rate in \$B, latest reported. Bloomberg, Reuters, The Information.

AI crawlers take far more than they give back



Pages crawled per visitor sent back to the site, July 2025, log scale. Cloudflare. This is why publishers now charge for access.

Same job, new problem: the micro-themes

Spotlight startupOther early-stage startup* \$100M+ seed

Data rights & licensing Pay for the right to use data	SAAS ERA Search engines scraped the web and sent visitors back.	AI ERA, 2026 AI crawlers take far more than they return, so data gets licensed or sued over.	LEADERS Reddit · News Corp · TollBit · Cloudflare · Protege 2026 MOVES • Meta pays News Corp up to \$50M a year (Mar) • Anthropic's \$1.5B authors settlement approved (Jul) • Cloudflare buys data marketplace Human Native (Jan)	MOVES THIS NUMBER Cost per usable token, legal risk EARLY-STAGE STARTUPS SF Data Origin Lab Redpine
Expert data, labels & traces Teach models with human judgment	SAAS ERA Cheap click-work labeling (Mechanical Turk, Appen).	AI ERA, 2026 Doctors, lawyers and coders write and grade answers. Practice tasks for agents cost \$200–2,000 each.	LEADERS Scale AI · Surge · Mercor · Handshake AI 2026 MOVES • Mercor passes \$2B run rate, talks at \$20B (Jul) • Prime Intellect raises \$130M for RL environments (Jul) • Scale AI names a new CEO (Jul)	MOVES THIS NUMBER Model quality per \$ of data EARLY-STAGE STARTUPS SenseLab Datacurve
Synthetic data Make training data with machines	SAAS ERA Fake test data for software QA.	AI ERA, 2026 Models write data for other models. It scales, but errors compound if nobody checks it.	LEADERS NVIDIA (Gretel, Cosmos) · Mostly AI 2026 MOVES • Gartner forecast (2025): 60% of data leaders face synthetic-data failures by 2027	MOVES THIS NUMBER Quality without new human data EARLY-STAGE STARTUPS Rockfish Data Datology AI Bifrost
Warehouses & pipelines Keep company data clean and in one place	SAAS ERA Snowflake and Teradata fed human dashboards.	AI ERA, 2026 Agents query company data, and even create databases. Over 80% of new Neon Postgres databases are made by agents.	LEADERS Databricks · Snowflake · ClickHouse · Fivetran + dbt 2026 MOVES • Databricks raises \$5B at \$190B (Aug) • ClickHouse at \$15B (Jan); Snowflake buys Observe (~\$1B)	MOVES THIS NUMBER Right answers from company data EARLY-STAGE STARTUPS Embrasure Eventual (Daft) Spiral Llamaindex Extend
Storage & file systems Feed GPUs and save their progress	SAAS ERA NetApp and Dropbox stored files for people.	AI ERA, 2026 Storage must keep thousands of GPUs fed and save checkpoints after frequent crashes.	LEADERS VAST Data · WEKA · DDN · Pure Storage 2026 MOVES • VAST raises \$1B at \$30B (Apr) • DDN seeks strategic investors (Jun)	MOVES THIS NUMBER GPU time lost waiting on data EARLY-STAGE STARTUPS Space Tigris Data Akave Peak:AI
Retrieval & vector search Find the right facts by meaning	SAAS ERA Elasticsearch matched keywords.	AI ERA, 2026 Agents search by meaning to pull the right passage. It is turning into a database feature.	LEADERS Pinecone · turbopuffer · Qdrant · Postgres pgvector 2026 MOVES • turbopuffer reports \$100M run rate with little outside capital (Mar) • Qdrant raises \$50M Series B (Mar)	MOVES THIS NUMBER Retrieval accuracy, cost per query EARLY-STAGE STARTUPS LanceDB Chroma Infino
Robot & sensor data Record how the physical world works	SAAS ERA Mapping fleets filmed roads (HERE, Mobileye).	AI ERA, 2026 There is no internet of robot actions. People wear cameras or steer robots by hand to make it.	LEADERS Scale AI · Encord · Foxglove · Mecka 2026 MOVES • Mecka raises \$60M led by Sequoia (Oct) • Encord raises \$60M Series C (Feb) • Scale delivers 1,000+ hours of robot data a day	MOVES THIS NUMBER Useful demo hours, robot success rate EARLY-STAGE STARTUPS Munari Vound Midcentury Human Archive Vision Lab

How this layer moves the outcome



If this layer fails

Models stop improving or make things up. Without licensed data there is legal risk, as the \$1.5B authors settlement shows. Without clean company data, agents give confident wrong answers.

Spotlight startups on this layer

SF Data DATA RIGHTS Brokers coding-agent traces to labs	SenseLab AGENT TRACES Turns agent runs into training data	Embrasure WAREHOUSES A warehouse agents build and run
Space STORAGE Cloud drive that acts like a local disk	Munari ROBOT DATA Scores robot demonstration quality	Vound ROBOT DATA First-person worker video for robot labs

Full investor cards and how the six connect: next page.

Sources

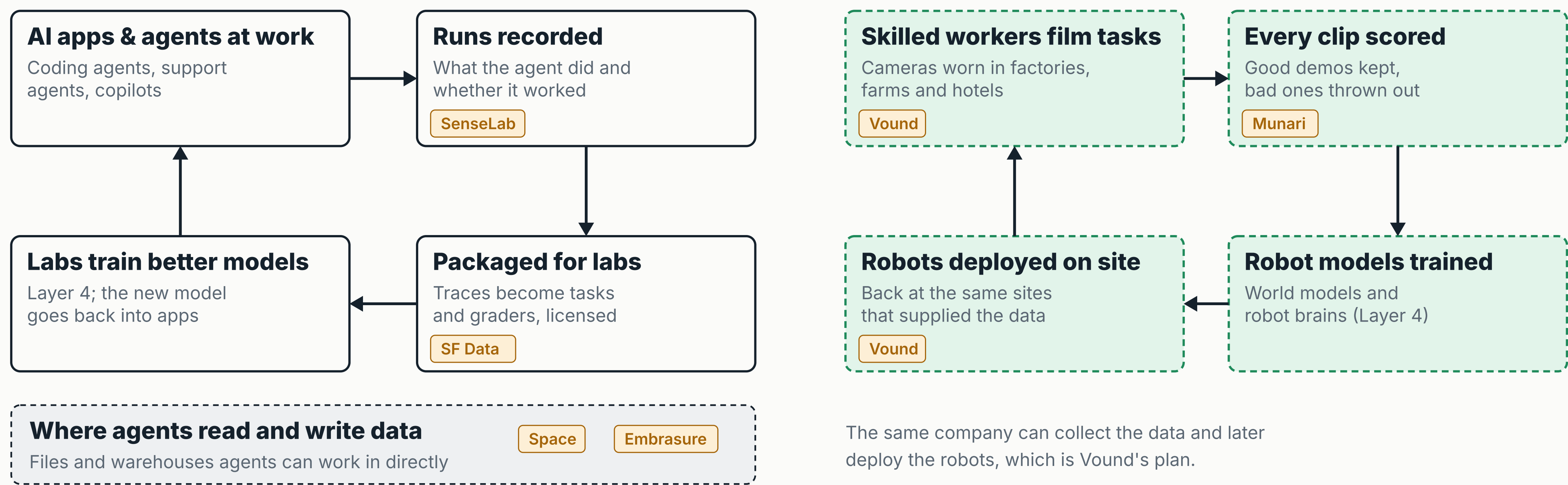
Epoch AI (data stock 2024; RL environments Jan 2026); Cloudflare (Aug 2025); TollBit via The Register (Feb 2026); Bloomberg, Reuters, The Information (Mercor, Surge, Handshake AI); Databricks, Snowflake, ClickHouse, VAST, Qdrant releases; Bloomberg Law (Anthropic settlement); Editor & Publisher (Meta–News Corp); Scale AI blog; Dealroom and PR Newswire (Mecka, Encord); Gartner (Jun 2025). Figures mix years; each label says which.

L3 Data · part 2

Six spotlight startups sit in the data layer, the most of any layer except trust. They cluster in two loops: one that teaches software agents, and one that teaches robots.

How the six connect

How the six spotlight data startups connect. Left: the digital learning loop. Right: the physical AI loop.



Investor cards

SF Data DATA RIGHTS & LICENSING <table><tr><td>THE BET</td><td>Real records of AI apps at work are the scarcest training data. It buys them from app companies, turns them into tasks and graders, and licenses them to labs.</td></tr><tr><td>PROOF SO FAR</td><td>Not yet verified. Its a16z Speedrun cohort membership could not be confirmed.</td></tr><tr><td>MUST BE TRUE</td><td>App companies have the right to sell user traces, labs pay again and again, and privacy holds up.</td></tr><tr><td>RISKS</td><td>Labs buy direct or build it. Mercor (>\$2B run rate), AfterQuery and Protege compete. Customer code ownership and consent.</td></tr><tr><td>ASK THE FOUNDER</td><td>Which app partners have signed? How are consent and code ownership handled? Do labs come back to buy more?</td></tr></table>	THE BET	Real records of AI apps at work are the scarcest training data. It buys them from app companies, turns them into tasks and graders, and licenses them to labs.	PROOF SO FAR	Not yet verified. Its a16z Speedrun cohort membership could not be confirmed.	MUST BE TRUE	App companies have the right to sell user traces, labs pay again and again, and privacy holds up.	RISKS	Labs buy direct or build it. Mercor (>\$2B run rate), AfterQuery and Protege compete. Customer code ownership and consent.	ASK THE FOUNDER	Which app partners have signed? How are consent and code ownership handled? Do labs come back to buy more?	SenseLab AGENT TRACES · ALSO L7 MEMORY <table><tr><td>THE BET</td><td>Agents don't learn from production. It records decisions and outcomes, shares lessons across agents, and exports good runs as training data.</td></tr><tr><td>PROOF SO FAR</td><td>A free tier and paid self-serve plans; pricing on its site is inconsistent. Its site claims 10,000+ teams (unverified).</td></tr><tr><td>MUST BE TRUE</td><td>Teams want to train smaller models on their own runs, and memory and training data can be one product.</td></tr><tr><td>RISKS</td><td>Crowded memory layer (Mem0, Zep). Eval platforms like Braintrust (\$800M value) add exports. Model makers build memory in.</td></tr><tr><td>ASK THE FOUNDER</td><td>Paying teams vs free users? Share who actually export to train? Retention after 3 months?</td></tr></table>	THE BET	Agents don't learn from production. It records decisions and outcomes, shares lessons across agents, and exports good runs as training data.	PROOF SO FAR	A free tier and paid self-serve plans; pricing on its site is inconsistent. Its site claims 10,000+ teams (unverified).	MUST BE TRUE	Teams want to train smaller models on their own runs, and memory and training data can be one product.	RISKS	Crowded memory layer (Mem0, Zep). Eval platforms like Braintrust (\$800M value) add exports. Model makers build memory in.	ASK THE FOUNDER	Paying teams vs free users? Share who actually export to train? Retention after 3 months?
THE BET	Real records of AI apps at work are the scarcest training data. It buys them from app companies, turns them into tasks and graders, and licenses them to labs.																				
PROOF SO FAR	Not yet verified. Its a16z Speedrun cohort membership could not be confirmed.																				
MUST BE TRUE	App companies have the right to sell user traces, labs pay again and again, and privacy holds up.																				
RISKS	Labs buy direct or build it. Mercor (>\$2B run rate), AfterQuery and Protege compete. Customer code ownership and consent.																				
ASK THE FOUNDER	Which app partners have signed? How are consent and code ownership handled? Do labs come back to buy more?																				
THE BET	Agents don't learn from production. It records decisions and outcomes, shares lessons across agents, and exports good runs as training data.																				
PROOF SO FAR	A free tier and paid self-serve plans; pricing on its site is inconsistent. Its site claims 10,000+ teams (unverified).																				
MUST BE TRUE	Teams want to train smaller models on their own runs, and memory and training data can be one product.																				
RISKS	Crowded memory layer (Mem0, Zep). Eval platforms like Braintrust (\$800M value) add exports. Model makers build memory in.																				
ASK THE FOUNDER	Paying teams vs free users? Share who actually export to train? Retention after 3 months?																				
Embrasure WAREHOUSES & PIPELINES <table><tr><td>THE BET</td><td>A data warehouse that agents build and maintain, on storage the customer owns, cheaper than Snowflake.</td></tr><tr><td>PROOF SO FAR</td><td>Waitlist only. Co-CEOs from Palantir, Uber and OpenAI's data and compute team.</td></tr><tr><td>MUST BE TRUE</td><td>Companies let agents own their data pipelines, and accuracy is good enough for business numbers.</td></tr><tr><td>RISKS</td><td>Databricks (\$190B) and Snowflake add agents to their own platforms. Its speed and cost claims are unbenchmarked.</td></tr><tr><td>ASK THE FOUNDER</td><td>Accuracy vs a human data team on real questions? First design partners? How do customers move off Snowflake?</td></tr></table>	THE BET	A data warehouse that agents build and maintain, on storage the customer owns, cheaper than Snowflake.	PROOF SO FAR	Waitlist only. Co-CEOs from Palantir, Uber and OpenAI's data and compute team.	MUST BE TRUE	Companies let agents own their data pipelines, and accuracy is good enough for business numbers.	RISKS	Databricks (\$190B) and Snowflake add agents to their own platforms. Its speed and cost claims are unbenchmarked.	ASK THE FOUNDER	Accuracy vs a human data team on real questions? First design partners? How do customers move off Snowflake?	Space STORAGE & FILE SYSTEMS <table><tr><td>THE BET</td><td>A cloud drive that behaves like a local disk lets people and agents work on terabytes without downloading them.</td></tr><tr><td>PROOF SO FAR</td><td>\$2.4M pre-seed led by a16z Speedrun (Aug 2026). Paying users in video and 3D work.</td></tr><tr><td>MUST BE TRUE</td><td>Agents need shared file workspaces, and it can grow beyond creative teams.</td></tr><tr><td>RISKS</td><td>LucidLink (\$75M Series C), Archil (\$11M Series A, focused on agents), Dropbox and Box, cloud storage giants.</td></tr><tr><td>ASK THE FOUNDER</td><td>Share of revenue from AI and agent use? Cost per terabyte vs LucidLink? Enterprise security story?</td></tr></table>	THE BET	A cloud drive that behaves like a local disk lets people and agents work on terabytes without downloading them.	PROOF SO FAR	\$2.4M pre-seed led by a16z Speedrun (Aug 2026). Paying users in video and 3D work.	MUST BE TRUE	Agents need shared file workspaces, and it can grow beyond creative teams.	RISKS	LucidLink (\$75M Series C), Archil (\$11M Series A, focused on agents), Dropbox and Box, cloud storage giants.	ASK THE FOUNDER	Share of revenue from AI and agent use? Cost per terabyte vs LucidLink? Enterprise security story?
THE BET	A data warehouse that agents build and maintain, on storage the customer owns, cheaper than Snowflake.																				
PROOF SO FAR	Waitlist only. Co-CEOs from Palantir, Uber and OpenAI's data and compute team.																				
MUST BE TRUE	Companies let agents own their data pipelines, and accuracy is good enough for business numbers.																				
RISKS	Databricks (\$190B) and Snowflake add agents to their own platforms. Its speed and cost claims are unbenchmarked.																				
ASK THE FOUNDER	Accuracy vs a human data team on real questions? First design partners? How do customers move off Snowflake?																				
THE BET	A cloud drive that behaves like a local disk lets people and agents work on terabytes without downloading them.																				
PROOF SO FAR	\$2.4M pre-seed led by a16z Speedrun (Aug 2026). Paying users in video and 3D work.																				
MUST BE TRUE	Agents need shared file workspaces, and it can grow beyond creative teams.																				
RISKS	LucidLink (\$75M Series C), Archil (\$11M Series A, focused on agents), Dropbox and Box, cloud storage giants.																				
ASK THE FOUNDER	Share of revenue from AI and agent use? Cost per terabyte vs LucidLink? Enterprise security story?																				
Munari ROBOT & SENSOR DATA <table><tr><td>THE BET</td><td>Robot labs waste training on bad demonstrations. Scoring every clip, task and operator saves money and improves robots.</td></tr><tr><td>PROOF SO FAR</td><td>Founders from Hedra and StockX (former CTO). Reads the common robot data formats.</td></tr><tr><td>MUST BE TRUE</td><td>Labs buy quality checks separately from collection, and the scores predict how well robots learn.</td></tr><tr><td>RISKS</td><td>Encord (\$60M Series C), Foxglove (\$40M Series B), Scale AI building it in.</td></tr><tr><td>ASK THE FOUNDER</td><td>Proof that filtering by score raises robot success? Who pays: labs or data collectors?</td></tr></table>	THE BET	Robot labs waste training on bad demonstrations. Scoring every clip, task and operator saves money and improves robots.	PROOF SO FAR	Founders from Hedra and StockX (former CTO). Reads the common robot data formats.	MUST BE TRUE	Labs buy quality checks separately from collection, and the scores predict how well robots learn.	RISKS	Encord (\$60M Series C), Foxglove (\$40M Series B), Scale AI building it in.	ASK THE FOUNDER	Proof that filtering by score raises robot success? Who pays: labs or data collectors?	Vound ROBOT & SENSOR DATA · ALSO ★ PHYSICAL AI <table><tr><td>THE BET</td><td>Pays skilled workers to wear cameras, sells labeled first-person video to robot labs, then deploys robots at the same sites.</td></tr><tr><td>PROOF SO FAR</td><td>\$2M run rate in 8 weeks, 15 staff (self-reported to a16z Speedrun).</td></tr><tr><td>MUST BE TRUE</td><td>Human first-person video teaches robots real skills, and access to work sites becomes a moat.</td></tr><tr><td>RISKS</td><td>Mecka (\$60M, Sequoia), Build AI (100K+ hours), Scale AI (1,000+ hours a day).</td></tr><tr><td>ASK THE FOUNDER</td><td>Hours captured per month and cost per hour? Exclusive deals? How many labs buy again?</td></tr></table>	THE BET	Pays skilled workers to wear cameras, sells labeled first-person video to robot labs, then deploys robots at the same sites.	PROOF SO FAR	\$2M run rate in 8 weeks, 15 staff (self-reported to a16z Speedrun).	MUST BE TRUE	Human first-person video teaches robots real skills, and access to work sites becomes a moat.	RISKS	Mecka (\$60M, Sequoia), Build AI (100K+ hours), Scale AI (1,000+ hours a day).	ASK THE FOUNDER	Hours captured per month and cost per hour? Exclusive deals? How many labs buy again?
THE BET	Robot labs waste training on bad demonstrations. Scoring every clip, task and operator saves money and improves robots.																				
PROOF SO FAR	Founders from Hedra and StockX (former CTO). Reads the common robot data formats.																				
MUST BE TRUE	Labs buy quality checks separately from collection, and the scores predict how well robots learn.																				
RISKS	Encord (\$60M Series C), Foxglove (\$40M Series B), Scale AI building it in.																				
ASK THE FOUNDER	Proof that filtering by score raises robot success? Who pays: labs or data collectors?																				
THE BET	Pays skilled workers to wear cameras, sells labeled first-person video to robot labs, then deploys robots at the same sites.																				
PROOF SO FAR	\$2M run rate in 8 weeks, 15 staff (self-reported to a16z Speedrun).																				
MUST BE TRUE	Human first-person video teaches robots real skills, and access to work sites becomes a moat.																				
RISKS	Mecka (\$60M, Sequoia), Build AI (100K+ hours), Scale AI (1,000+ hours a day).																				
ASK THE FOUNDER	Hours captured per month and cost per hour? Exclusive deals? How many labs buy again?																				

What changes by 2027–2030

1. **Labs use up the stock of public human text between 2026 and 2032 (median ~2028).** Epoch AI
2. **Old data vendors become sellers of practice worlds for agents (RL environments).** Epoch AI, Jan 2026
3. **By 2027, 50% of business decisions are augmented or automated by AI agents.** Gartner, Jun 2025
4. **By 2027, 60% of data leaders hit critical failures managing synthetic data.** Gartner, Jun 2025

White space

Synthetic data quality checks and RL environments are growing fast with no spotlight startup yet. Retrieval is becoming a database feature, so standalone bets need a sharp cost or speed edge, as turbopuffer showed.

6 spotlight startups in the data layer
Infra Conviction

2 loops
Digital learning loop and physical AI loop
see diagram

2 gaps
No spotlight startup in synthetic data or retrieval
Infra Conviction

Sources

a16z Speedrun company pages (SR007); company sites; Speedrun traction is self-reported. Competitor rounds: Bloomberg (Mercor), Axios (Braintrust), Archil, PR Newswire (Encord, LucidLink), BusinessWire (Foxglove), Dealroom (Mecka), Humanoids Daily (Build AI), Databricks release. Forecasts: Epoch AI, Gartner.

L4 Models · part 1

This is where data and compute become a brain. A few labs spend hundreds of millions per model, open models trail by about four months, and the new race is after the first training run.

~4–5× a year

Growth in compute for the biggest training runs
Epoch AI

\$852B · \$965B

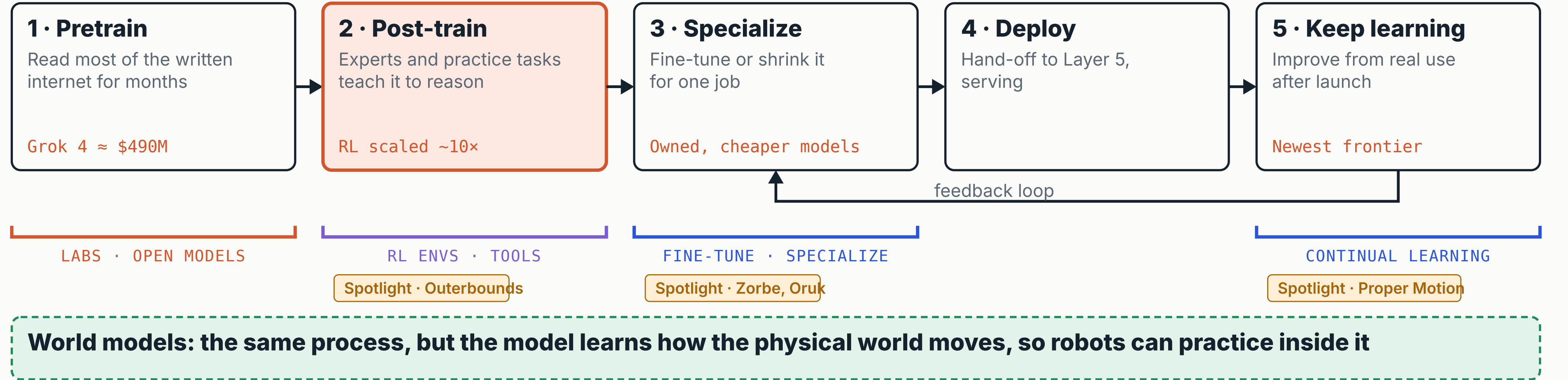
OpenAI and Anthropic valuations after 2026 rounds
Company releases

~3 months

Lag of open-weight models behind closed
Epoch AI, Oct 2025

How it works

The life of a model. Red box is where the new race is: what happens after the first training run.

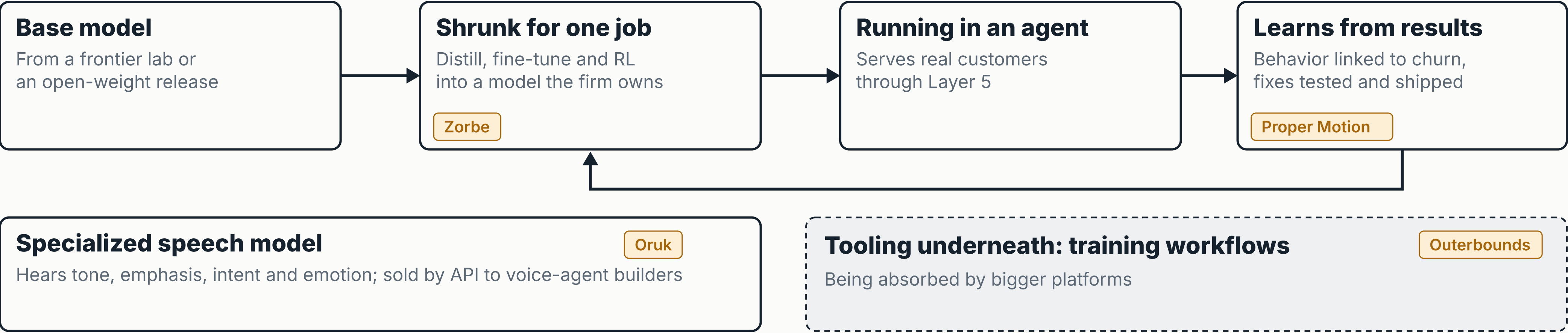


L4 Models · part 2

The four spotlight model-layer startups avoid the frontier race. They shape models for one job, keep them improving after launch, or sell one sense really well.

Where the four sit

Where the four spotlight startups sit around a model. Labs build the base; these startups shape it for one job, keep it improving, or run the tooling.



Investor cards

Oruk

SPECIALIZED MODELS · ALSO L5 API

THE BET	Transcripts lose how something was said. Models that capture tone, emphasis, intent and emotion make voice agents and call analytics smarter.
PROOF SO FAR	Its open 0.6B speech model beats NVIDIA Parakeet on 61 of 74 tests (own benchmark). API from \$9 to \$199 a month. Founders from Stanford NLP and a 14-patent engineer.
MUST BE TRUE	Buyers pay more for emotion and intent than for plain transcription, and it stays ahead of big-lab voice models.
RISKS	Labs absorb the capability: Google licensed Hume's tech and hired its CEO (Jan 2026). ElevenLabs (\$22B in a Sep tender) and Deepgram (\$1.3B) are well funded.
ASK THE FOUNDER	Accuracy on emotion and intent vs Hume-style models? Who pays today? Any licensing interest from labs?

Zorbe

FINE-TUNING & POST-TRAINING

THE BET	A company's repetitive AI jobs run cheaper and better on a small model it owns, built by distillation, fine-tuning and RL.
PROOF SO FAR	a16z Speedrun SR007 and Pareto Holdings. Three people. No public customers yet.
MUST BE TRUE	Frontier prices stay high enough to justify custom models, and customers want to own the weights.
RISKS	Applied Compute (~\$3B talks, ~\$50M run rate), Thinking Machines' Tinker and Together AI. Frontier prices fall 9–900× a year, which shrinks the savings.
ASK THE FOUNDER	Cost and quality vs a frontier model on a live workload? Who owns the weights? How does the model keep up when the frontier improves?

Proper Motion

CONTINUAL LEARNING · ALSO L8 EVALS

THE BET	Deployed agents slowly get worse. Linking agent behavior to churn and testing fixes lets them improve from real use.
PROOF SO FAR	\$270K+ ARR in 3 weeks of beta and ~\$2M pipeline (self-reported). A co-founder started OpenAI's user reward model team.
MUST BE TRUE	Customers share outcome data like churn and revenue, and the gains are measurable and repeatable.
RISKS	Part service (engineers placed at customers), which lowers margins. Trajectory (\$15M), Adaption (\$50M), NeoCognition (\$40M); labs adding on-the-job learning.
ASK THE FOUNDER	How much is software vs people? Measured lift in a customer's own metric? Time to first result?

Outerbounds

TRAINING & ML WORKFLOW · EXITED

THE BET	A managed version of Netflix's open-source Metaflow for teams moving ML from notebooks to production.
PROOF SO FAR	About \$24M raised. Acquired by Anaconda on April 29, 2026; terms not disclosed.
STATUS	Exited. Keep it out of the active pipeline and use it as a comparable.
RISKS	The pattern it shows: standalone ML workflow tools get absorbed (W&B by CoreWeave, Anyscale by Nscale).
USE IT FOR	Benchmarking exits for workflow tools: what Anaconda paid vs capital raised, if it becomes public.

Who buys model-layer startups: the exit map

CoreWeave bought Weights & Biases (~\$1.7B, 2025) and OpenPipe (2025) NEOCLOUD BUYING TOOLS	Nscale agreed to buy Anyscale, the Ray company (reported ~\$1.65B, Jul 2026) NEOCLOUD BUYING TOOLS	Rubrik bought Predibase, fine-tuning (\$100–500M reported, 2025) ENTERPRISE SOFTWARE BUYER
Anaconda bought Outerbounds (Apr 2026, terms not disclosed) DEVELOPER PLATFORM BUYER	Google licensed Hume AI's emotion-voice tech and hired its CEO (Jan 2026) LAB BUYING A CAPABILITY	NVIDIA ~\$20B license and team deal with Groq (Dec 2025) CHIPMAKER BUYING SERVING

The pattern: clouds and platforms buy model tooling to sell more compute; labs buy a sense or skill they lack. Plan each investment around which of these buyers it would be worth most to.

4 spotlight startups in model building, one already acquired
Infra Conviction

~\$3B Applied Compute's talks valuation: a sign that owned models sell
The Information, Aug 2026

4 gaps No spotlight startup in frontier, open-weight, world models or RL environments
Infra Conviction

What changes by 2027–2030

1. **Training runs of ~2e29 FLOP become feasible by 2030, about 10,000× GPT-4.** Epoch AI
2. **Training clusters cost over \$100B and draw gigawatts by 2030.** Epoch AI, Sep 2025
3. **Companies use small task models 3× more than general LLMs by 2027.** Gartner, Apr 2025
4. **Robot foundation models reach a \$150B addressable market by 2036.** ABI Research, 2026

White space

RL environments and world-model tooling are seed-friendly and growing fast. Frontier labs are out of reach for a seed fund's size; specialized models only work with a clear exit to a lab or platform.

Sources

a16z Speedrun pages (SR007; traction self-reported); Oruk Hugging Face model card; Anaconda release (Apr 2026); TechCrunch via Yahoo (Google–Hume); ElevenLabs, Deepgram releases; The Information via Dealroom (Applied Compute); Dealroom (Trajectory); Fortune (Adaption); TechCrunch (NeoCognition); Epoch AI; Gartner; ABI Research.

L5

Serving & efficiency

Every AI answer is computed fresh, word by word. Serving is where AI's cost is decided. A fixed level of ability gets 9–900× cheaper each year, but usage grows even faster, so shaving cost keeps getting more valuable.

~2/3

Share of all AI compute spent answering, not training, in 2026
Deloitte, Dec 2025

~200×

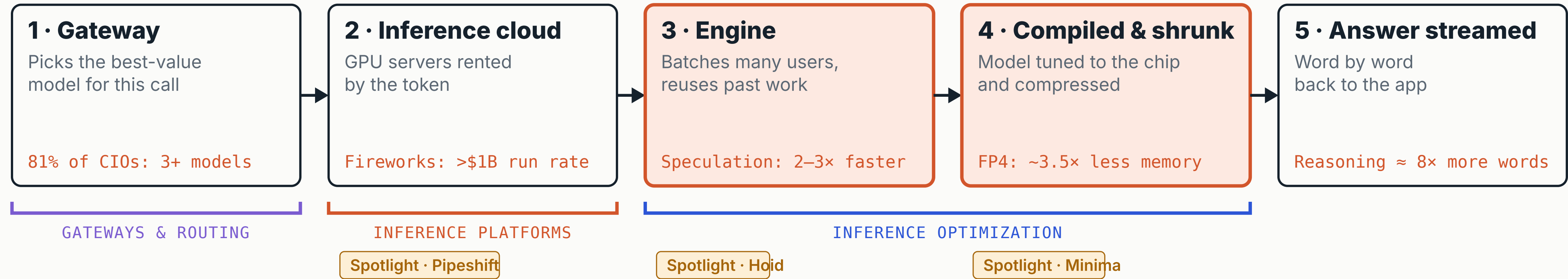
Cheaper GPT-4-level answers, Mar 2023 → Feb 2025
Epoch AI

5–30×

More tokens an agent uses per task than a chatbot
Gartner, Mar 2026

How it works

What happens to one request. Red boxes are where most of the cost is won or lost.



L5

The inference stack, up close

Between a model and the power plant sit six sub-layers that decide how fast and how cheap every answer is. Every token travels down the stack; every dollar travels up.

\$300M

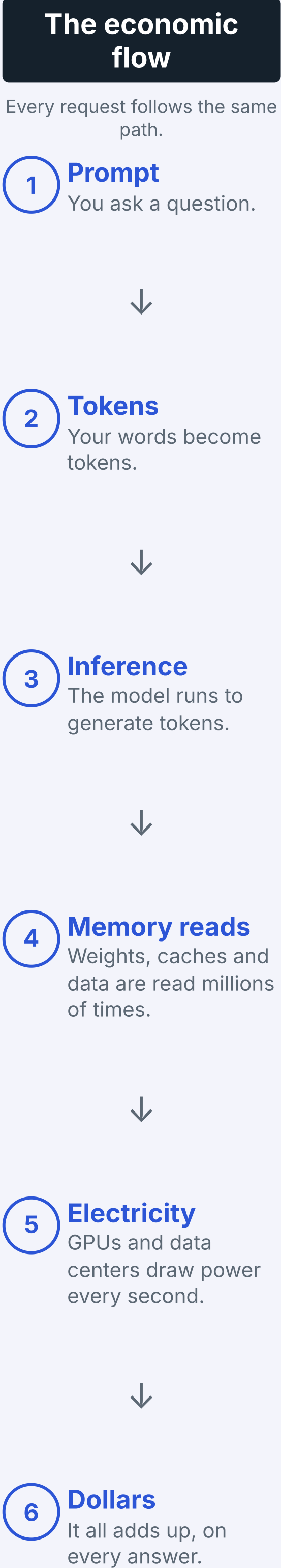
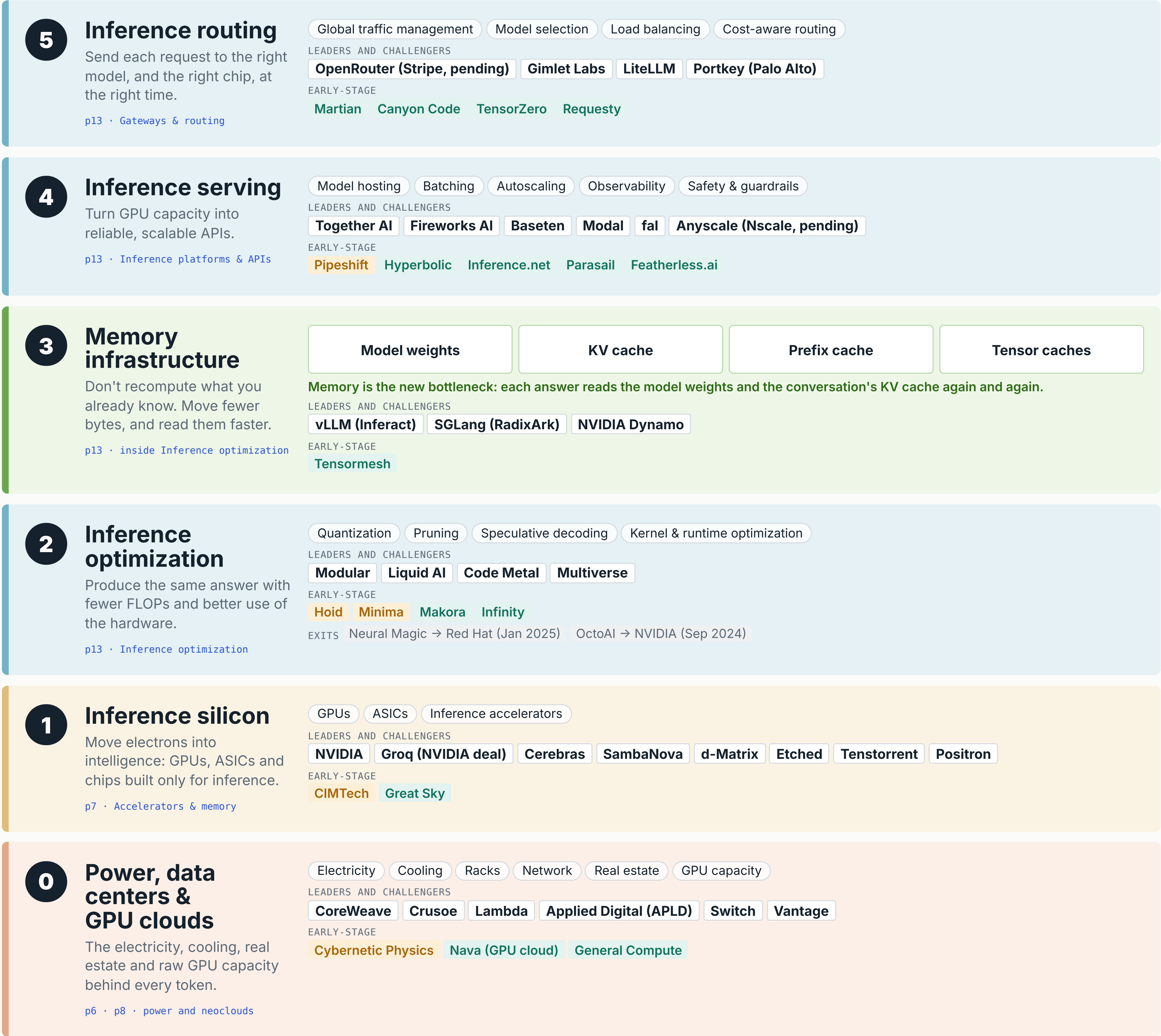
Gimlet Labs Series B, an inference cloud that splits work across GPUs, CPUs and accelerators
a16z Lead, Sep 2026

\$11B

SambaNova's valuation after a \$1B round for inference chips
General Atlantic lead, Jul 2026

2 exits

Optimization startups bought: Neural Magic by Red Hat, OctoAI by NVIDIA
Red Hat, Jan 2025; GeekWire, Sep 2024



Who's who: names added from the inference-stack map

Gimlet Labs ROUTING Inference cloud splitting work across GPUs, CPUs and accelerators	Canyon Code ROUTING Controls and optimizes multi-agent apps above model serving	fal SERVING Fast inference for image, video and audio models	Liquid AI OPTIMIZATION Small, efficient non-transformer models for devices
Code Metal OPTIMIZATION Verified translation of code onto edge and defense hardware	Makora OPTIMIZATION Automatically tunes GPU kernels to cut inference cost	SambaNova SILICON Inference chips and systems, plus a cloud service	Applied Digital POWER & DC Builds and hosts AI data centers (Nasdaq: APLD)
Vantage POWER & DC Builds and runs hyperscale data-center campuses	Switch POWER & DC Large data-center campuses, mainly in Las Vegas	Neural Magic EXIT Model compression behind vLLM work; bought by Red Hat	OctoAI EXIT Model serving platform; bought by NVIDIA, then shut down

Recent rounds in the inference stack (valuation, \$B)



Rounds: SambaNova \$1B Series F; fal \$140M Series D; Gimlet Labs \$300M Series B. Also: Code Metal \$125M Series B (Feb 2026, Salesforce Ventures); Liquid AI \$250M (Dec 2024, AMD); Canyon Code \$5M pre-seed (May 2026, Cota Capital).

What this adds to the layer pages

- Memory is its own fight.** Caching weights, KV and prefixes is where serving cost is now won; the atlas folds it into inference optimization.
- Routing is moving down to chips.** Gimlet routes work across different hardware, not just different models.
- Optimization tools get bought.** Neural Magic and OctoAI went to a platform and a chipmaker; Makora and Infinity are the next wave.

The one-line takeaway

Smarter routing, better serving, faster memory, leaner models, faster silicon and cheaper power: each layer makes intelligence cheaper per answer.

Sources

Company releases and press: Pulse 2.0 (Gimlet Labs, Code Metal), ABVEN (SambaNova), Tech.eu (fal), ODSC (Liquid AI), VCA (Canyon Code), Caplight (Makora), Red Hat (Neural Magic), GeekWire (OctoAI), Yahoo Finance (Applied Digital). Layer details on pages 6, 7, 8 and 13.

L6

Builder tools

AI now writes most new code at leading companies. The bottleneck moved from writing code to checking it and giving agents a safe place to run it.

75%

Of Google's new code is written by AI
Semafor, Apr 2026

~1.7×

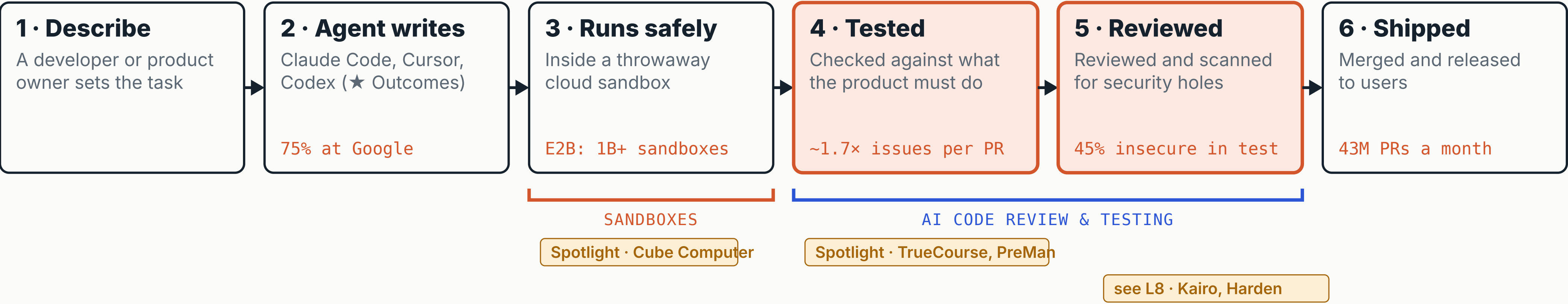
More issues per pull request in AI-written code
CodeRabbit study, Dec 2025

1B+

Sandboxes started on E2B, a cloud for agents
E2B, Jun 2026

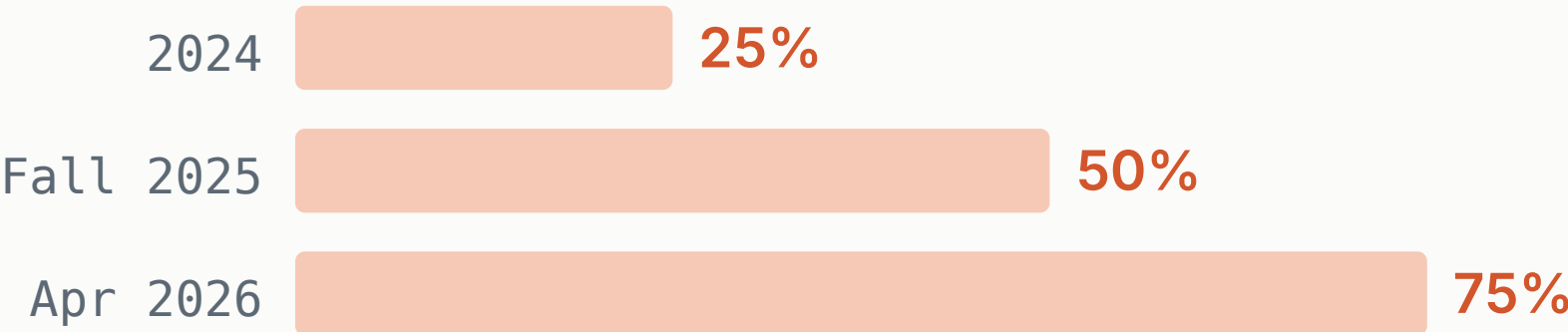
How it works

How code gets made in 2026. Red boxes are the new bottleneck: checking what the AI wrote.



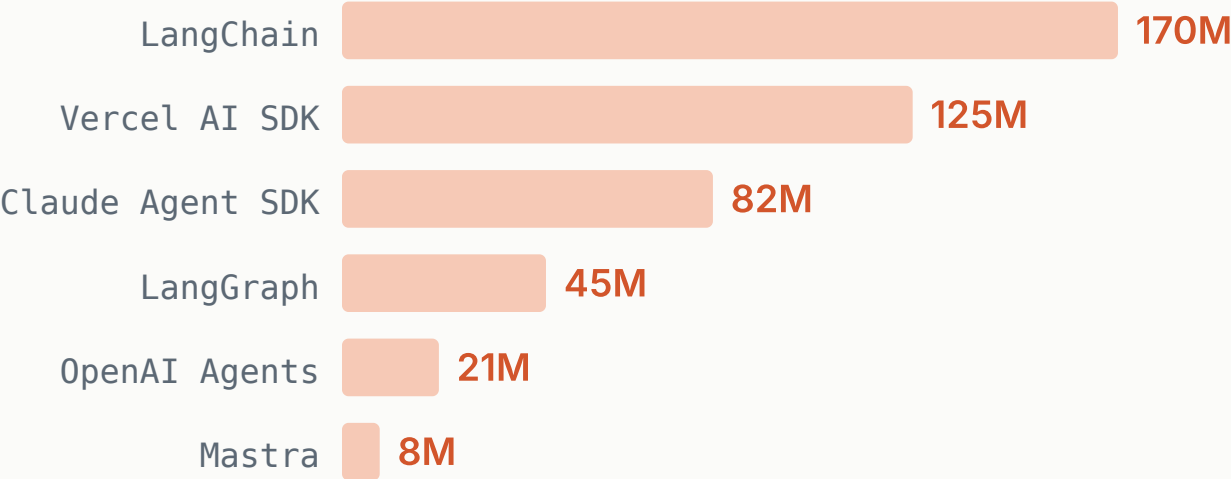
Frameworks & SDKs underneath: the code kits agents are built with (LangChain, Vercel AI SDK, Claude Agent SDK)

Share of Google's new code written by AI



Sundar Pichai, via Semafor (Apr 2026). Anthropic says over 80% of its merged code is written by Claude.

Agent kits downloaded in the last 30 days

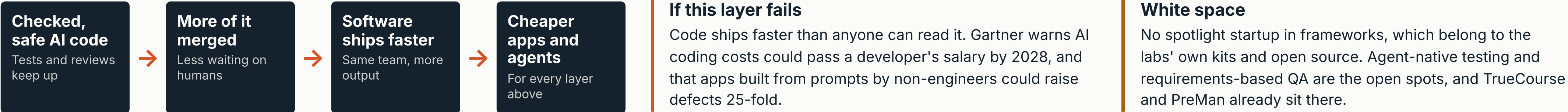


npm and PyPI, to Oct 10, 2026. Includes automated installs, so compare kits rather than count users.

Same job, new problem: the micro-themes

Frameworks & SDKs Code kits for building agents	SAAS ERA Rails and React for web apps.	AI ERA, 2026 Kits from labs and startups, downloaded tens of millions of times a month.	LEADERS LangChain · Vercel AI SDK · Claude Agent SDK · OpenAI Agents SDK · Mastra 2026 MOVES • Mastra raises \$22M Series A (Apr) • Claude Agent SDK passes ~80M downloads a month	MOVES THIS NUMBER Time to build a working agent EARLY-STAGE STARTUPS CopilotKit CrewAI Pydantic AI
Sandboxes & dev boxes Safe computers where agents run code	SAAS ERA Docker and GitHub Codespaces for people.	AI ERA, 2026 Agents start thousands of throwaway machines a day, and each must boot in under a second.	LEADERS E2B · Daytona · Modal · Namespace · Coder 2026 MOVES • OpenAI buys Ona, formerly Gitpod (Jun) • Modal raises \$355M at \$4.65B (May) • Namespace raises \$42M Series B (Oct)	MOVES THIS NUMBER Boot time, cost per agent-hour EARLY-STAGE STARTUPS Cube Computer Blaxel Boxd
AI code review & testing Prove AI code does what was asked	SAAS ERA SonarQube and Selenium for human-written code.	AI ERA, 2026 More code than people can read. 45% of AI code samples had security flaws in one test.	LEADERS CodeRabbit · Qodo · Greptile · Momentic · QA Wolf 2026 MOVES • CodeRabbit raises \$143M at \$1.5B (Aug) • Qodo raises \$70M Series B (Mar) • Cursor buys Graphite (Dec 2025)	MOVES THIS NUMBER Bugs that reach users EARLY-STAGE STARTUPS TrueCourse PreMan Gitar

How this layer moves the outcome



Spotlight startups here

Cube Computer

SANDBOXES & DEV BOXES

THE BET

Developers want a personal always-on cloud machine where several coding agents run side by side, off their laptop.

PROOF SO FAR

400+ users and \$12K ARR five weeks after launch (self-reported). Plans \$24.50–184 a month. CPO founded Kosmik (\$10M raised).

MUST BE TRUE

Developers pay for their own box rather than use labs' cloud agents, and it grows into teams.

RISKS

OpenAI bought Ona (Jun 2026). Coder (\$90M Series C), Daytona (\$24M Series A), and labs bundling cloud agents for free.

ASK THE FOUNDER

Paid conversion and churn? Revenue per user? How does it survive labs bundling cloud agents?

TrueCourse

AI CODE REVIEW & TESTING

THE BET

Turn the specs product owners already write into tests that run against the live product and block code that breaks them.

PROOF SO FAR

Open source (MIT), about 535 GitHub stars. Berkeley SkyDeck Batch 22.

MUST BE TRUE

Product owners adopt it, not only engineers, and requirements are clear enough to test automatically.

RISKS

Qodo (\$70M Series B), Momentic (\$15M), Meticulous (\$15M); code-review tools adding spec checks.

ASK THE FOUNDER

Paying teams? Share of blocked changes that were real bugs? Open-source to paid conversion?

PreMan

AI CODE REVIEW & TESTING

THE BET

Simulates AI agents using a product, finds where workflows break, and turns failures into regression tests.

PROOF SO FAR

\$170K ARR (self-reported). CEO was founding engineer at Surt AI (\$0 to \$1M ARR).

MUST BE TRUE

Tests stay reliable without constant fixing, and buyers switch from QA services.

RISKS

QA Wolf (\$36M Series B), Momentic (\$15M), Spur (\$4.5M). Its 'millions of agents' pitch is not yet the product.

ASK THE FOUNDER

Flaky-test rate? Time to cover a new app? Share of revenue from software vs people?

What changes by 2027–2030

- 75% of enterprise software engineers use AI code assistants by 2028. Gartner, Apr 2024
- Over 65% of teams using coding agents treat the code editor as optional by 2027. Gartner, May 2026
- AI coding costs pass the average developer's salary by 2028. Gartner, Jun 2026
- Apps built from prompts by non-engineers raise defects 2,500% by 2028. Gartner Predicts 2026, via ArmorCode

Sources

Semafor (Google, Apr 2026); VentureBeat (Anthropic, Jun 2026); GitHub Octoverse 2025; Stack Overflow 2025 survey; Veracode (Jul 2025); CodeRabbit report (Dec 2025, vendor study); METR (Jul 2025); npm and pypistats (Oct 2026); E2B, Modal, Daytona, Namespace, Mastra, Qodo releases; Axios (CodeRabbit); SiliconANGLE (Graphite); CIO (Ona); Gartner press releases; a16z Speedrun and company sites (Cube, TrueCourse, PreMan).

L7 Agent runtime · part 1

Bots now make more web requests than people. Agents need what people needed online: a way to stay on task, plug into apps, browse, remember, prove who they are and pay.

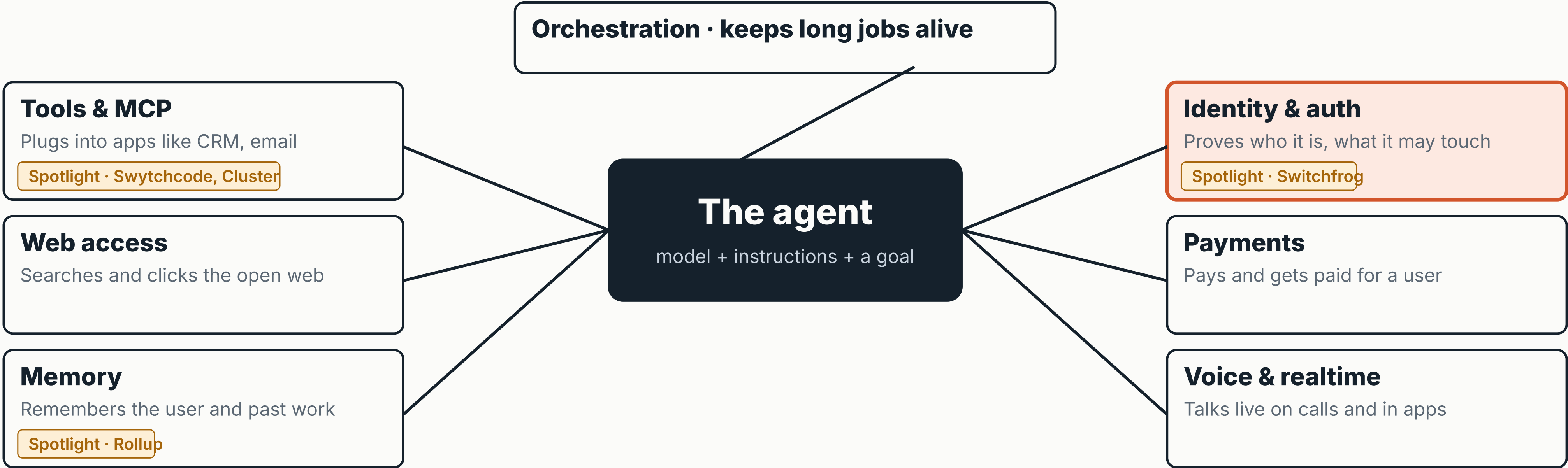
~60% Of web requests now come from bots
Cloudflare via Fortune, Sep 2026

~500M MCP software downloads a month
MCP blog, Jul 2026

\$12.55B Temporal's value: keeping long jobs alive is big business
Sep 2026

How it works

Everything an agent needs to work in the real world. spotlight startups shown where they plug in.



Bots passed people online

53% of web traffic was automated in 2025
Thales/Imperva, Apr 2026

+7,851% growth in AI agent traffic in a year
HUMAN Security, Apr 2026

~60% of requests to Cloudflare came from bots by Sep 2026
Fortune, Sep 2026

Feb 6 last day people used more tokens than agents on
OpenRouter
OpenRouter, 2026

Money into agent plumbing in 2026



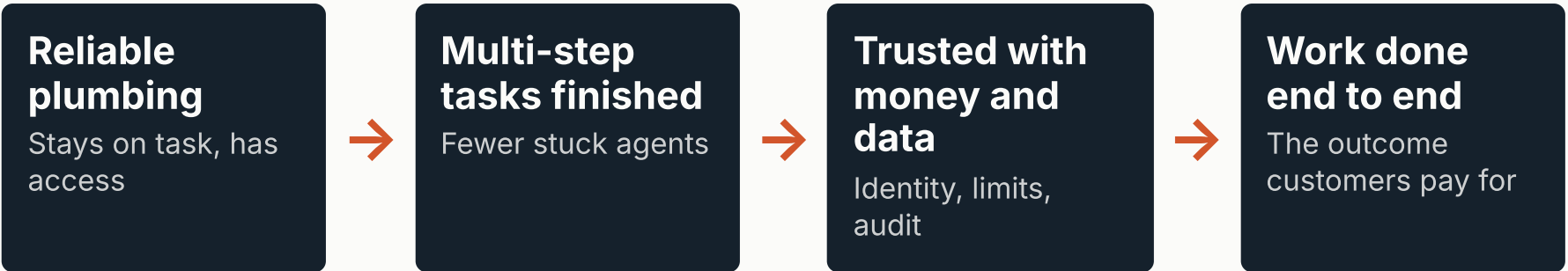
Company releases, TechCrunch, SiliconANGLE.

Same job, new problem: the micro-themes

Spotlight startup Other early-stage startup * \$100M+ seed

Orchestration & durable runtimes Keep long jobs alive	SAAS ERA Airflow and scheduled jobs.	AI ERA, 2026 Agents run for hours or days, crash, wait for people, and must pick up where they stopped.	LEADERS Temporal · LangGraph · Inngest · Restate · AWS AgentCore 2026 MOVES • Temporal raises \$550M at \$12.55B, \$250M+ ARR (Sep) • Claude Managed Agents opens in beta (Apr)	MOVES THIS NUMBER Tasks finished without restarting EARLY-STAGE STARTUPS Xpander Kestra Trigger.dev DBOS BAND
Tools, connectors & MCP Plug agents into business apps	SAAS ERA Zapier and MuleSoft moved data between apps.	AI ERA, 2026 MCP became the standard plug between AI and software, and went stateless in July.	LEADERS Composio · Arcade · Zapier MCP · Pipedream (Workday) 2026 MOVES • Arcade raises \$60M, then buys Smithery (Jun, Aug) • MCP spec goes stateless (Jul)	MOVES THIS NUMBER Tool calls that succeed EARLY-STAGE STARTUPS Swytchcode Cluster Runlayer Obot AI Orthogonal
Web access Search and use the open web	SAAS ERA Bing's search API and Selenium scripts.	AI ERA, 2026 Search engines and browsers built for agents, not people.	LEADERS Exa · Parallel · Browserbase · Firecrawl 2026 MOVES • Exa raises \$250M at \$2.2B (May); Parallel \$2B (Apr) • Nebius buys Tavily (Feb); Firecrawl \$75M (Sep)	MOVES THIS NUMBER Right page found, cost per lookup EARLY-STAGE STARTUPS Keenable Browser Use Kernel Linkup Anchor Browser
Memory & context Remember users and past work	SAAS ERA Redis caches and Segment customer profiles.	AI ERA, 2026 No standard yet. Re-feeding context into every call is slow and costly.	LEADERS Mem0 · Engram · Letta · Supermemory 2026 MOVES • Engram launches with \$98M at \$600M (Jun)	MOVES THIS NUMBER Fewer tokens, more personal answers EARLY-STAGE STARTUPS Rollup SenseLab (also L3) Cognee
Identity & auth Prove who the agent is and what it may do	SAAS ERA Okta single sign-on for people.	AI ERA, 2026 Agents need their own IDs and limited access. 88% of firms in one survey had agent security incidents.	LEADERS Okta · Styth (Twilio) · WorkOS · Keycard · Cloudflare 2026 MOVES • Okta for AI Agents launches (Apr); WorkOS \$2B (Mar) • Okta buys Permiso, ~\$200M reported (Jul)	MOVES THIS NUMBER Agents with scoped access EARLY-STAGE STARTUPS Switchfrog Oak NewCore Hush Security
Payments & commerce Let agents pay and get paid	SAAS ERA Stripe and PayPal for people at checkout.	AI ERA, 2026 New protocols let agents buy things. Volume is still tiny: x402 has 205M payments but ~\$53M in total.	LEADERS Stripe · Visa · Mastercard · Google · Coinbase 2026 MOVES • Google commerce protocol (Jan); Stripe machine payments (Mar) • Visa payments come to ChatGPT (Jun)	MOVES THIS NUMBER Purchases completed by agents EARLY-STAGE STARTUPS Natural Nekuda SolvaPay
Voice & realtime Talk live on calls and in apps	SAAS ERA Twilio phone and SMS APIs.	AI ERA, 2026 Voice agents answer calls at scale. Vapi alone has handled over 1B calls.	LEADERS LiveKit · Vapi · Deepgram · ElevenLabs · Retell 2026 MOVES • LiveKit raises \$100M at \$1B (Jan) • Vapi raises \$50M Series B (May)	MOVES THIS NUMBER Calls solved without a person EARLY-STAGE STARTUPS Rime Smallest.ai VoiceRun

How this layer moves the outcome



If this layer fails

Agents get stuck, forget, act with too much access or can't pay. Gartner expects over 40% of agentic AI projects to be canceled by the end of 2027.

Spotlight startups on this layer

Swytchcode TOOLS & MCP Safe, governed API calls for enterprise agents	Cluster TOOLS & MCP · GTM Sales data and outreach tools for agents
Rollup MEMORY & CONTEXT Personal context from a user's photos, on device	Switchfrog IDENTITY Tells a SaaS app if a session is a human or an agent

Full investor cards on the next page.

Sources

Fortune (Cloudflare, Sep 2026); Thales/Imperva Bad Bot Report 2026; HUMAN Security (Apr 2026); OpenRouter "Inferred" (Sep 2026); Claude blog on MCP (Jul 2026); MCP blog (Dec 2025); Temporal, Restate, Arcade, Exa, WorkOS, Okta, Mem0, LiveKit, Vapi releases; TechCrunch (Parallel, Okta-Permiso); Bloomberg via Times of Israel (Tavily); PYMNTS (Google UCP); Fortune (Stripe MPP); Visa (Jun 2026); CryptoRank (x402); Gartner (Jun 2025). Agent platform launch dates from third-party write-ups.

L7 Agent runtime · part 2

Three spotlight startups help agents reach out: call software safely, sell, and know the user. One helps the apps they reach: telling a person from an agent.

4 spotlight startups in the agent runtime
Infra Conviction

~\$200M Okta's reported price for an agent-security startup
TechCrunch, Jul 2026

4 gaps No spotlight startup in orchestration, web access, payments or voice
Infra Conviction

Where the four sit

Two sides of the same exchange: the agent reaching out, and the app it reaches.

THE AGENT SIDE · built by developers

Calls software safely
Auth, retries, policy and audit before any API call leaves the building
Swytchcode

Sales data and outreach
50+ data providers, buying signals, email and LinkedIn through MCP
Cluster

Knows the user on day one
Photo library turned into context, on the device, with consent
Rollup

The agent at work
logs in, reads, acts

THE APP SIDE · the SaaS it logs into

Is this a person or an agent?
Each session labeled human, agent or unknown, then governed or billed
Switchfrog

Charge for agent seats
Turns agent use into revenue instead of lost seats

Investor cards

SwytchcodeTOOLS, CONNECTORS & MCP

THE BET

Agents break on undocumented API call sequences. A governed layer handles auth, retries, policy and audit before any SaaS or legacy API call leaves the company.

PROOF SO FAR

Pre-seed with angels; open CLI. SkyDeck Batch 22 listing not confirmed on SkyDeck's own pages.

MUST BE TRUE

Enterprises want a separate control point for agent API calls, beyond the MCP servers vendors ship themselves.

RISKS

Arcade (\$60M Series A, bought Smithery), Runlayer (\$30M Series A), Composio (\$25M). Workday bought Pipedream.

ASK THE FOUNDER

Which APIs are covered today? Paying enterprise pilots? How is it different from an MCP gateway?

ClusterTOOLS & MCP · GO-TO-MARKET

THE BET

Give Claude, Codex and Cursor sales abilities through MCP and a CLI: 50+ data providers, buying signals and outreach.

PROOF SO FAR

a16z Speedrun SR007. Publishes its own GTM MCP benchmark. Founders ex-Remitly ML and PeakMetrics.

MUST BE TRUE

Sales teams run outbound through AI clients instead of sales apps, and data costs stay below what customers pay.

RISKS

Clay (\$7.1B, already ships MCP and a CLI), ZoomInfo's MCP data layer (Jun 2026), Apollo.

ASK THE FOUNDER

Revenue per seat vs data cost? Reply rates vs Clay? Are buyers developers or sales leaders?

RollupMEMORY & CONTEXT

THE BET

New apps know nothing about new users. An on-device SDK turns a user's photo library, with consent, into structured context from day one.

PROOF SO FAR

a16z Speedrun SR007. CTO ex-Databricks and Notion. Demo apps only so far.

MUST BE TRUE

Users opt in to share photo-based context, apps pay per user, and keeping it on the device protects privacy.

RISKS

Apple or Google build personal context into the phone. Memory players Mem0 (\$24M) and Engram (\$98M). App store rules.

ASK THE FOUNDER

Opt-in rate? First paying apps? What happens if Apple ships this?

SwitchfrogIDENTITY & AUTH · AGENT DETECTION

THE BET

When users hand their logins to agents, SaaS loses seat revenue and control. Labeling each session lets the app govern or bill it.

PROOF SO FAR

\$500K pre-seed (Jul 2026, per Wellfound, unverified). CEO ex-Microsoft security and Faire data infrastructure.

MUST BE TRUE

SaaS firms want to charge for agent seats rather than block them, and detection stays accurate as agents act more human.

RISKS

Cloudflare's signed-agent checks (free), Keycard (\$38M), Vouched Agent Shield. Exit comps: Twilio-Stytch, Okta-Permiso (~\$200M).

ASK THE FOUNDER

Detection accuracy and false alarms? Who buys: security or pricing teams? First customer billing agent seats?

Who buys agent plumbing: the exit map

Twilio bought Stytch, identity for apps and agents (announced Oct 2025) COMMS GIANT BUYING IDENTITY	Okta bought Permiso, identity threat detection (~\$200M reported, Jul 2026) IDENTITY LEADER BUYING AGENT SECURITY	Workday agreed to buy Pipedream, 3,000+ app connectors (Nov 2025) APP SUITE BUYING CONNECTORS
Nebius bought Tavily, web search for agents (~\$275M reported, Feb 2026) NEOCLOUD BUYING AGENT TOOLS	Arcade bought Smithery, an MCP registry (Aug 2026) STARTUP ROLL-UP	OpenAI bought Ona, formerly Gitpod, agent workspaces (Jun 2026) LAB BUYING RUNTIME

What changes by 2027–2030

- Over 40% of agentic AI projects are canceled by end-2027. Gartner, Jun 2025
- By 2028, 33% of enterprise software includes agentic AI. Gartner, Jun 2025
- AI agents handle over \$15T of business purchasing by 2028. Gartner, Nov 2025
- Agentic commerce reaches \$3–5T globally by 2030. McKinsey, Oct 2025

White space

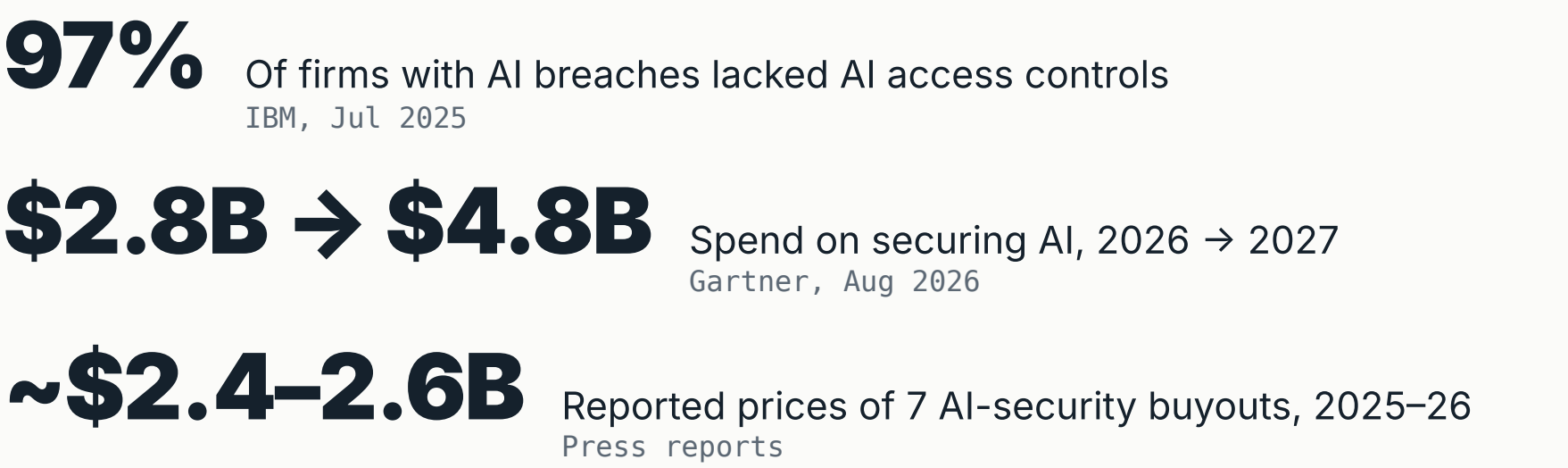
Agent payments and voice are early and growing. Orchestration and web access now have well-funded leaders. Memory still has no winner or standard.

Sources

a16z Speedrun pages (SR007); company sites; Wellfound (Switchfrog). Peers and exits: Arcade, Clay, ZoomInfo, Mem0 releases; TechCrunch (Okta-Permiso, Runlayer via SecurityWeek); Twilio; SiliconANGLE (Workday-Pipedream); Bloomberg via Times of Israel (Tavily); CIO (Ona). Forecasts: Gartner, McKinsey.

L8 Trust · part 1

Agents now hold real keys to email, code, money and data. One poisoned web page, plugin or package can make them delete, leak or spend. The trust layer tests, watches and limits them.



How it works

Four checkpoints on every agent action. Trust runs across all other layers, from chip design to the agent itself.



L8 Trust • part 2

Eight spotlight startups sit here, the most of any layer. The question is concentration, not gaps: which ones have a clear buyer, and which overlap.

8 spotlight startups in the trust layer
Infra Conviction

10+ AI-security or eval startups bought by big firms since 2025
Press reports

3 overlaps Trusys with AgentCat and ZebraTruth; Harden with Kairo
Infra Conviction

Where the eight sit on the four checkpoints

Before launch TrusysAbliteration.aiKairo	At action time HardenZebraTruthTrusys	On the data SkyflowAurva	After the fact AgentCatTrusys
---	--	---	--

Trusys appears three times because it sells a platform across checkpoints. That breadth is both its pitch and its risk.

Investor cards

AgentCat EVALS & OBSERVABILITY THE BET Product teams can't see what AI agents try to do inside their product. Analytics and session replay for agents using your MCP server, Claude connector or ChatGPT app. PROOF SO FAR a16z Speedrun SR007. Renamed from MCPcat (Jul 2026). Open-source SDK; Webflow and Rootly shown as customers on its site. MUST BE TRUE 'Agent experience' becomes a product discipline like UX, and teams pay for it apart from LLM tracing. RISKS PostHog and Mixpanel add agent traffic; Raindrop (\$35M), Braintrust (\$800M). ASK THE FOUNDER How many MCP servers instrumented? Paid conversion? What decisions do customers change from the data?	Harden AI SECURITY · CODING AGENTS THE BET A guard on the developer's machine checks every action a coding agent takes and blocks secret leaks, data theft and destructive commands. PROOF SO FAR \$2M pre-seed reported by NewsBytes (Sep 2026), not on its own site. Free for individuals. CEO ex-Verkada AI. MUST BE TRUE Developers keep it switched on (few false alarms, no lag), and security teams buy a paid tier. RISKS Runlayer (\$30M), Palo Alto bought Koi (~\$400M), Snyk, and coding-agent makers adding their own guards. ASK THE FOUNDER Block rate vs false alarms? Paid team seats? Which agents are covered?
Kairo AI SECURITY · APPSEC THE BET AI-speed shipping outruns security review. It reviews every code change and keeps attack-testing the live app. PROOF SO FAR a16z Speedrun SR007 and Superteam USA. GitHub app. CEO a competitive hacker (ex-BCG), CTO ex-AWS. MUST BE TRUE Findings are accurate enough to trust without a human tester, and it beats scanners bundled into big platforms. RISKS XBOW (\$1B+), Aikido (\$1B), Snyk, DryRun (\$8.7M seed). ASK THE FOUNDER True-positive rate? Bugs found that Snyk missed? Price per repository?	Abliteration.ai AI SECURITY · RED TEAMING THE BET Security teams need models that won't refuse attack simulations: hosted open models with refusals removed, plus a policy gateway. PROOF SO FAR \$7M ARR (self-reported, unverified). Revenue-funded before Speedrun. MUST BE TRUE Buyers are vetted so the tool isn't misused, and enterprises accept the reputational risk. RISKS Misuse and press risk; free self-hosted versions; Dreadnode, Mindgard (\$30M). ASK THE FOUNDER How are customers vetted? Share of revenue from vetted enterprises? Any misuse incidents?
Skyflow DATA PRIVACY & SECURITY THE BET A vault keeps sensitive data out of apps and AI models; agents work with tokens, not raw values. PROOF SO FAR ~\$97–100M raised (last: \$30M Series B extension, Khosla, Mar 2024). India data-law vault (Dec 2025); Skyflow for Glean (Aug 2026). MUST BE TRUE Firms route AI data through a vault rather than mask it inside each platform, and India's law drives demand. RISKS No new round since 2024. Cyera (\$12B) and platforms bundling privacy. A pre-AI company entering a new market. ASK THE FOUNDER Share of new revenue from AI and agent use? India pipeline? Next financing?	Aurva DATA PRIVACY & SECURITY THE BET Watch every data access in real time and tie it to a person, a service or an AI agent. PROOF SO FAR \$2.2M seed (Nexus, DeVC, Sep 2025). Razorpay and Meesho. Founders ex-Meta data security. MUST BE TRUE India's rules (1-year access logs, 72-hour breach reports) make it a must-buy, and it can expand to the US. RISKS Cyera (\$12B), Varonis (bought Cyral for this), Sentra (\$50M). ASK THE FOUNDER Production deployments? Average contract size? US pipeline?
Trusys GOVERNANCE · ALSO EVALS THE BET One platform to test, attack, monitor and enforce policy on enterprise AI agents. PROOF SO FAR LTIMindtree partner listing. CEO a serial founder (Pokkt, Koovs). Funding not found. MUST BE TRUE Buyers want one platform over best-of-breed tools, and the IT-services partnership turns into deals. RISKS Spans four crowded markets; Virtue AI, Credo AI, Lasso, Galileo (now Cisco). ASK THE FOUNDER Paying customers vs pilots? Revenue through LTIMindtree? Which module wins deals?	ZebraTruth GOVERNANCE & COMPLIANCE THE BET Check an agent's proposed real-world action against law, regulation and company policy before it happens. PROOF SO FAR CEO co-founded Scientific Beta (sold to SGX). Site shows BlackRock and Roche logos, unverified as customers. Funding not found. MUST BE TRUE Regulated firms trust an outside rules engine at the moment of action, and the rules stay current across countries. RISKS Norm Ai (\$1.2B, Khosla), smaller policy-layer startups; liability if a check is wrong. ASK THE FOUNDER Which logos are paying? How are rules kept current? Who is liable for a wrong answer?

Who buys trust companies: the exit map

Palo Alto Networks Protect AI (~\$500–700M, reports vary), Koi (~\$400M), Console (~\$500M) MOST ACTIVE BUYER	SentinelOne · CrowdStrike Prompt Security (~\$250M) · Pangea (~\$260M) ENDPOINT LEADERS	Check Point · F5 Lakera (~\$300M) · CalypsoAI (\$180M) NETWORK SECURITY
Cisco · ClickHouse Galileo (Apr 2026) · Langfuse (Jan 2026) OBSERVABILITY BUYERS	OpenAI · Harvey Promptfoo (Mar 2026) · Guardrails AI (2026) AI COMPANIES BUYING SAFETY	Okta · Varonis Permiso (~\$200M) · Cyral (2025) IDENTITY AND DATA SECURITY

Most natural buyer for each (our view, based on 2025–26 deals)

AgentCat Observability and analytics platforms: Datadog, Cisco, PostHog	Harden Endpoint and code security: Palo Alto, CrowdStrike, Snyk	Kairo AppSec platforms: Snyk, Palo Alto, Check Point	Abliteration.ai Unclear: reputational risk narrows buyers to offensive-security firms
Skyflow Identity and data security: Okta, Varonis, a cloud provider	Aurva Data security: Varonis, Cyera, Zscaler	Trusys IT services or governance suites: LTIMindtree, ServiceNow, OneTrust	ZebraTruth Legal and regulatory data: Thomson Reuters, Wolters Kluwer, Norm Ai

What changes by 2027–2030

- Spend on securing AI reaches \$4.8B in 2027, up about 70% in a year.** Gartner, Aug 2026
- Guardian agents that watch other agents take 10–15% of the agentic AI market by 2030.** Gartner, Jun 2025
- AI governance platforms pass \$1B in spend by 2030.** Gartner, Feb 2026
- By 2030, insurers rather than regulators drive AI governance.** Gartner, Sep 2026

How to use this page

Rank the eight by who would buy them. Palo Alto, CrowdStrike, SentinelOne, Cisco, Okta and Varonis have all paid for similar products since 2025. Companies without an obvious buyer need a much bigger standalone business.

Sources

a16z Speedrun pages (SR007); company sites; NewsBytes (Harden); FinTech Global (Aurva); Clay and Skyflow press (Skyflow); LTIMindtree partner page (Trusys). Peers and exits: SecurityWeek, SiliconANGLE, Globes, Dark Reading, BusinessWire, BankInfoSecurity, CyberScoop, Security Boulevard, TechCrunch, Nasdaq, FinSMEs. Forecasts: Gartner. Self-reported traction flagged on each card.

★ Assistants & agents

This is what end customers pay for: assistants for everyone and agents that finish work. Every layer below exists to lower the cost of a finished task and raise the chance it succeeds.

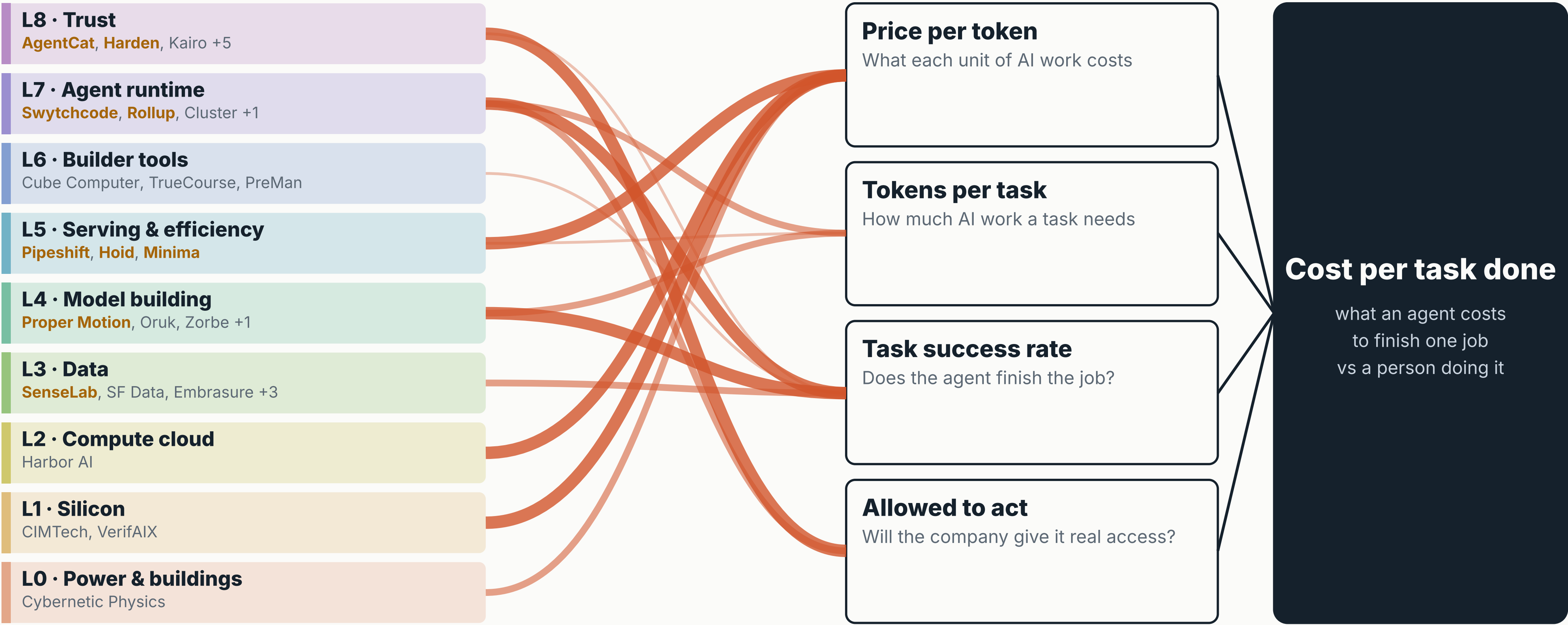
1.2B ChatGPT weekly users
OpenAI DevDay, Sep 2026

76% Average share of chats Intercom's Fin resolves for customers, \$0.99 each
Intercom, Jun 2026

~17 hrs Task length top agents finish half the time, doubling about every 3 months
METR, 2026

How the stack moves this outcome

Thicker line = bigger effect on the outcome (our judgment, from the layer research). Bold amber = spotlight startups that matter most here.



Who earns from the outcome (latest run rates)

~\$50B OpenAI annualized, Sep 2026 Reuters	>\$65B Anthropic run rate, Jul 2026 Reuters
>\$4B Cursor annualized, Jun 2026 Dealroom	\$1.5B Salesforce Agentforce ARR, Aug 2026 Earnings call
\$300M Glean ARR, May 2026 Glean	\$150M Sierra ARR at \$15.8B value, May 2026 CNBC

Pricing moves from seats to outcomes

\$0.99 Intercom Fin, per resolved conversation Intercom	76% Fin's average resolution rate, Jun 2026 Intercom
\$0.10 Agentforce per action (Flex Credits) Salesforce, May 2025	45% → 52% AI app gross margin, 2025 → 2026 expected ICONIQ
~20-23% Share of AI product cost spent on inference ICONIQ, SaaStr	~17 hrs Task length a frontier agent finishes half the time METR, Mar 2026; noisy above 16 hrs

Spotlight startups' levers on this outcome

Lower price per token Hoid tunes models to each chip Minima shrinks models to fewer GPUs Pipeshift fast serving of open models CIMTech memory card for long answers Harbor AI fewer idle GPU-hours	Fewer tokens per task Rollup user context without re-asking SenseLab agents reuse what worked Zorbe small owned models for repeat jobs	Higher success rate Swytchcode API calls that don't break Proper Motion agents improve from real use Embrasure clean data agents can query TrueCourse · PreMan AI code that does what was asked	Allowed to act Harden · Kairo safe actions and code ZebraTruth · Trusys legal and policy checks Skyflow · Aurva sensitive data protected Switchfrog · AgentCat agents identified and watched
---	---	---	---

How good agents are now

95.0% Top score on SWE-bench Verified, fixing real code issues (Claude Fable 5, self-reported) llm-stats, Oct 2026	86.1% Top score on OSWorld-Verified, using a computer like a person; people scored ~72% on the original test llm-stats, Oct 2026	~89 days How often the length of task agents can finish doubles, since 2024 METR
--	--	--

Spotlight startups at this outcome

Infragrid AGENTIC APP · DESKTOP AUTOMATION THE BET A desktop assistant watches how you work, records a task once, turns it into an editable runbook and runs agents to repeat it. PROOF SO FAR a16z Speedrun SR007. Founders ex-Palantir and ex-Mercor, AWS. No public traction. WHY IT MATTERS Not infrastructure. It is a buyer of memory, computer-use and model layers, and a test of whether 'record once, automate forever' works. RISKS Simular (\$21.5M A, same idea), ChatGPT agent, Claude Cowork; Meta's Manus deal was blocked by China (Apr 2026). ASK THE FOUNDER Tasks automated per user per week? Retention? Which model and computer-use stack does it run on?	Bric STEALTH · ENTERPRISE AGENTS THE BET Unknown. Listed as an enterprise-agent company in a16z Speedrun SR007; site says only 'access by request'. PROOF SO FAR Founders Sohom Paul, Ashwin Ayyasamy and Austin Vaday (listed by Frontrun). Prior roles at Slack AI and Snap not independently confirmed. WHY IT MATTERS Placement waits for launch. A founder call is the only way to place it on the map. RISKS Stealth companies in crowded agent markets often pivot before launch. ASK THE FOUNDERS What job does the agent do, for whom? Infra or app? What does it need from the layers below?
---	---

What changes by 2027–2035

- Over 40% of agentic AI projects are canceled by end-2027. Gartner, Jun 2025
- Agents resolve 80% of common customer-service issues by 2029. Gartner, Mar 2025
- Agentic AI spending reaches \$1.3T in 2029. IDC, Aug 2025
- Up to \$234B of seat-based software spending is at risk from agents by 2030. Gartner via CIO Dive, Jul 2026

Sources

OpenAI DevDay (Sep 2026); Google (Gemini, Aug 2026); Reuters via secondary outlets (OpenAI, Anthropic revenue); Dealroom (Cursor); Salesforce Q2 FY27 call; Glean release; CNBC via Yahoo (Sierra); Bloomberg (Decagon); Intercom blog; ICONIQ via SaaSletter; Vista Equity (inference share); METR and AGI Tracker; TechCrunch (Simular, Manus); a16z Speedrun and Frontrun (Infragrid, Bric).

★

Vertical AI

Vertical AI sells finished expert work: a legal brief, a medical note, a credit memo. It wins on domain accuracy, fit with how professionals work, and proof that it is compliant.

~400%

Yearly growth of AI-native vertical companies
Bessemer, Jan 2026

\$15.5B

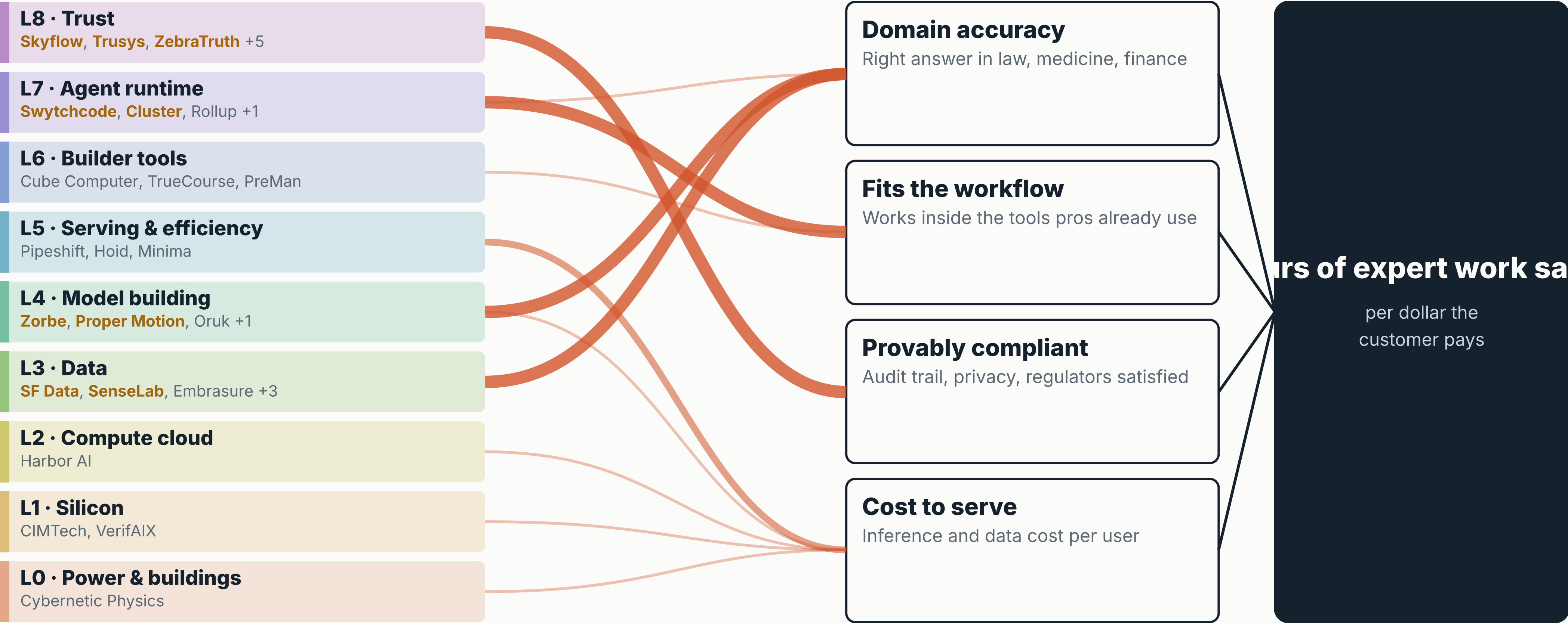
Harvey, used by 80% of the largest US law firms
Sep 2026

~65%

Vertical AI gross margin, well above AI apps overall
Bessemer

How the stack moves this outcome

Thicker line = bigger effect on the outcome (our judgment, from the layer research). Bold amber = spotlight startups that matter most here.



Physical AI

Physical AI puts models into machines that move: cars, humanoids, warehouse robots. It depends on the same stack, plus three things software AI does not need: robot data, simulation and chips on the robot itself.

~500K

Waymo paid robotaxi rides a week, 15 US cities
TechCrunch, Sep 2026

+432%

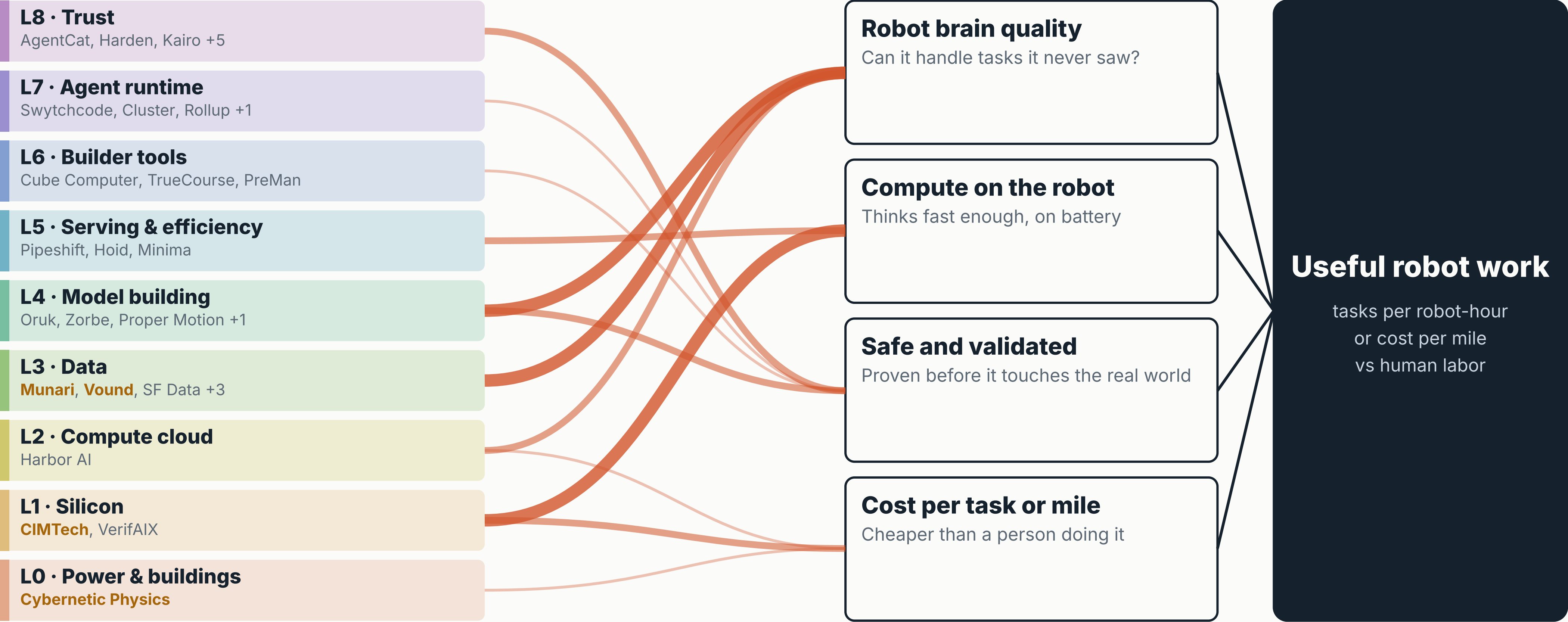
Humanoid shipments, H1 2026 vs a year earlier
IDC via Forbes

~1M

Demonstrations in the largest open robot dataset
Google DeepMind

How the stack moves this outcome

Thicker line = bigger effect on the outcome (our judgment, from the layer research). Bold amber = spotlight startups that matter most here.



Where physical AI stands, 2026

~500K

Waymo paid rides a week, 15 cities; \$126B valuation
TechCrunch; Bloomberg

~25,000

Humanoids shipped in H1 2026, +432%; >95% Chinese
IDC via Forbes

\$39B

Figure's valuation; 350+ robots delivered by Apr
Figure

~\$66B

Unitree's value after its first day of trading, Aug
Fortune

\$14B+

Skild AI, robot brains (Jan)
Bloomberg

\$3B

Generalist, trained on 270K hours of robot data
TechCrunch

Costs are falling fast

Waymo today

~\$150K per city

Waymo on Zeekr

~\$70K

Cybercab target

~\$30K

BNP via Benzinger, Sep 2026. Humanoid average price fell from ~\$37K to ~\$29.6K (IDC); BofA sees Chinese parts cost under \$17K by 2030.

The data gap

The largest open robot dataset has about 1M demonstrations. Language models learn from trillions of words. That gap is why Vound and Munari exist.

Spotlight startups' levers on this outcome

Robot brain quality Vound first-person video of skilled work Munari keeps only good demonstrations	Compute on the robot CIMTech memory-near-compute ideas could reach the edge (not its focus today)	Safe and validated None yet simulation and robot evaluation are open	Cost per task Cybernetic Physics robots doing data-center work Vound plans to deploy robots at its data sites
---	---	--	--

The race for robot brains

5% → 95%

Physical Intelligence's π0.7 on a new task after ~30 minutes of better instructions
TechCrunch, Apr 2026

Open

NVIDIA released GR00T N1.6, a free humanoid model, to sell more chips
NVIDIA, Jan 2026

270K hrs

Robot-hand data behind Generalist's GEN-0, growing 10K hours a week
Generalist, Nov 2025

Spotlight exposure and the gaps

Physical-AI spotlight startups

VOUND · MUNARI · CYBERNETIC PHYSICS

THE BET

All three are picks-and-shovels for robots: Vound collects first-person work video, Munari scores robot training data, Cybernetic Physics runs robots inside data centers.

PROOF SO FAR

Vound \$2M run rate in 8 weeks; Cybernetic Physics ~\$828K live ARR in 8 weeks (both self-reported). Munari: no public traction yet.

THE READ

Vound and Munari are two early bets on robot data, the most-cited constraint in robotics, without betting on which robot body wins.

RISKS

Scale AI (1,000+ hrs a day), Mecka (\$60M), Encord (\$60M); labs building their own data operations (Figure budgets ~\$1B a year).

NEXT STEP

Source in the gaps: robot simulation and validation, edge inference software, fleet operations and safety.

Open gaps

NO SPOTLIGHT STARTUP YET

THE BET

World models and simulation, robot foundation models, edge-chip software, robot safety and validation, fleet operations.

PROOF SO FAR

Money is flowing: World Labs \$1B, AMI Labs \$1.03B, Skild \$14B+, Generalist \$3B, Apptronik \$520M.

WHERE SEED CHECKS FIT

Tools for robot developers (simulation, evaluation, data pipelines) need less capital than robot makers or model labs.

RISKS

NVIDIA gives away much of the tooling (Isaac, Cosmos, GR00T) to sell chips.

SCREEN FOR

Products that work across many robot makers, and a buyer beyond NVIDIA's free stack.

What changes by 2027–2035

- Robotaxis become a ~\$415B market by 2035, with ~6M vehicles. Goldman Sachs, Apr 2026
- Humanoid shipments pass 750K a year by 2030. IDC via Forbes, Sep 2026
- Humanoids become a ~\$38B market by 2035. Goldman Sachs, Feb 2024
- Industrial robot installs pass 700K a year by 2028. IFR, Sep 2025

Sources

TechCrunch, InsideEVs, Bloomberg (Waymo); Tesla Q2 2026 call; Baidu and Pony.ai Q2 2026 results; IDC via Forbes (humanoid shipments, Sep 2026); Figure blog; 1X release; Fortune (Unitree); CNBC (Apptronik); Bloomberg (Skild); TechCrunch (Generalist, Physical Intelligence); NVIDIA (GR00T, Cosmos); Google DeepMind (Open X-Embodiment); BNP via Benzinger; BofA via AI Prediction Index; Goldman Sachs; IFR.

I\$ Where value accrues

Across the stack, money flows down to chips, memory and power, and up to the labs and vertical apps. The middle earns by cutting cost or risk, and usually exits by being bought.

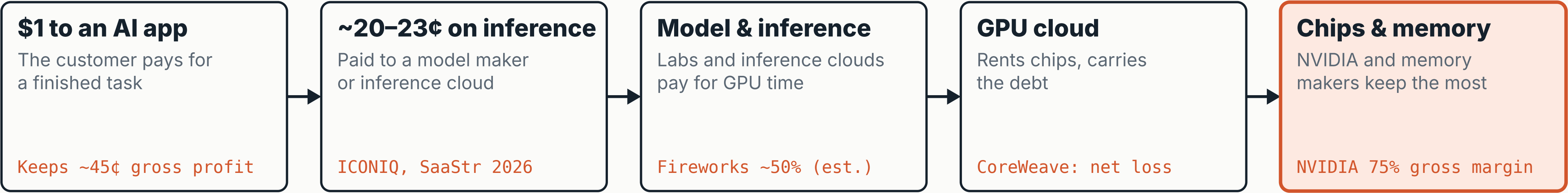
75% NVIDIA gross margin, quarter ended Jul 2026
NVIDIA

~45% Gross margin of a typical AI app in 2025
ICONIQ

-24% CoreWeave net margin in Q2 2026 (\$626M loss on \$2.58B)
CoreWeave 10-Q

Where \$1 of AI app revenue goes

Follow \$1 that a customer pays an AI app. Inference is the app's biggest cost, and that money flows down the stack.



#

Spotlight startups, ranked

A first-pass screen of every atlas company on four questions, scored from public evidence only. It decides where to spend diligence time, not what to invest in.

9

Tier A: diligence first
score 9–10 of 12

14

Tier B (5) and B– (9): diligenc
score 7–8

13

Tier C, or exited or stealth: watch or meet
score ≤6 or not scorable

EVIDENCE				POSITION		EXIT PATH		ROOM		
3 = verified funding and named customers or revenue · 2 = self-reported traction · 1 = little public				How hard to copy or bundle: data, workflow lock-in, technical depth		3 = a big firm has recently paid for something close · 1 = unclear buyer		3 = few funded rivals · 1 = rivals with \$100M+ or a giant's free product		
#	COMPANY	LAYER	MICRO-THEME	EVIDENCE	POSITION	EXIT PATH	ROOM	SCORE	TIER	WATCH-OUT
1	Aurva	L8	Data security	●●●	●●●	●●●	●●●	10	A	
2	VerifAIX	L1	Chip verification	●●●	●●●	●●●	●●●	10	A	
3	Cybernetic Physics	L0	DC operations	●●●	●●●	●●●	●●●	9	A	self-reported ARR
4	Oruk	L4	Speech models	●●●	●●●	●●●	●●●	9	A	
5	Hoid	L5	AI compiler	●●●	●●●	●●●	●●●	9	A	speed claim unaudited
6	Switchfrog	L7	Agent identity	●●●	●●●	●●●	●●●	9	A	round unverified
7	AgentCat	L8	Agent observability	●●●	●●●	●●●	●●●	9	A	
8	Harden	L8	Coding-agent security	●●●	●●●	●●●	●●●	9	A	round in one report
9	Skyflow	L8	Privacy vault	●●●	●●●	●●●	●●●	9	A	later stage (Series B)
10	Harbor AI	L2	GPU fleet ops	●●●	●●●	●●●	●●●	8	B	
11	Space	L3	AI file system	●●●	●●●	●●●	●●●	8	B	
12	Vound	L3	Robot data	●●●	●●●	●●●	●●●	8	B	self-reported run rate
13	Proper Motion	L4	Continual learning	●●●	●●●	●●●	●●●	8	B	part service
14	TrueCourse	L6	Spec-based testing	●●●	●●●	●●●	●●●	8	B	
15	CIMTech	L1	Memory accelerator	●●●	●●●	●●●	●●●	7	B–	deep tech, capital heavy
16	Munari	L3	Robot data QA	●●●	●●●	●●●	●●●	7	B–	
17	Minima	L5	Compression	●●●	●●●	●●●	●●●	7	B–	4× claim internal
18	Cube Computer	L6	Agent dev boxes	●●●	●●●	●●●	●●●	7	B–	
19	PreMan	L6	Self-healing QA	●●●	●●●	●●●	●●●	7	B–	
20	Swytchcode	L7	Governed API calls	●●●	●●●	●●●	●●●	7	B–	
21	Rollup	L7	User context SDK	●●●	●●●	●●●	●●●	7	B–	platform risk (Apple)
22	Abliteration.ai	L8	Red-team models	●●●	●●●	●●●	●●●	7	B–	misuse and reputation risk
23	ZebraTruth	L8	Pre-action compliance	●●●	●●●	●●●	●●●	7	B–	co-founder unverified
24	SF Data	L3	Trace brokerage	●●●	●●●	●●●	●●●	6	C	cohort unconfirmed
25	SenseLab	L3	Agent traces	●●●	●●●	●●●	●●●	6	C	
26	Embrasure	L3	Agent warehouse	●●●	●●●	●●●	●●●	6	C	
27	Pipeshift	L5	Inference cloud	●●●	●●●	●●●	●●●	6	C	giants well funded
28	Kairo	L8	AppSec	●●●	●●●	●●●	●●●	6	C	
29	Nexor AI	★	Sales agents	●●●	●●●	●●●	●●●	6	C	not infra
30	Sterling	★	Revenue agents	●●●	●●●	●●●	●●●	6	C	not infra
31	Zorbe	L4	Owned models	●●●	●●●	●●●	●●●	5	C	
32	Cluster	L7	GTM tools for agents	●●●	●●●	●●●	●●●	5	C	Clay overlap
33	Trusys	L8	AI governance suite	●●●	●●●	●●●	●●●	5	C	no funding found
34	Infragrid	★	Desktop agents	●●●	●●●	●●●	●●●	5	C	not infra
—	Outerbounds	L4	ML workflow	Exited: acquired by Anaconda, Apr 2026. Use as a comparable.					Exit	
—	Bric	?	Stealth	Unplaceable until launch. Founder backgrounds claimed, not yet confirmed; take the meeting.					Meet	

How to read it

Scores are a judgment from the research on pages 6–21, made to rank diligence time. Tier A means the public evidence, exit path and room line up; each still needs a founder call on its watch-out. Not investment advice.

Sources

All facts come from the company cards on pages 6–21 and their cited sources. Self-reported figures are flagged there.

→ Where to look next

Twenty micro-themes have no spotlight startup yet, though early-stage startups are active in every one. Six are good fits for a seed fund now. Here is where to look and what to ask.

20 / 43 Micro-themes with no spotlight startup
Infra Conviction

6 High-priority gaps for a seed fund
our view

8 Spotlight startups in trust, the most of any layer
Infra Conviction

Coverage map: 43 micro-themes, by layer



The empty micro-themes, prioritized

Priority	Empty Micro-theme	Layer	Why Now	Early-stage startups to meet
High	RL environments	L4	Labs pay \$200–2,000 per practice task; Bespoke Labs \$40M, Halluminate \$30M (2026)	Proximal, Bespoke Labs, Halluminate
High	World models & robot simulation tools	L4	World Labs \$1B, AMI Labs \$1.03B; tools for robot developers need less capital	Moonlake AI, SpAltial, Aether AI
High	Agent payments	L7	Stripe, Visa, Google launched rails in 2025–26; volume still tiny	Natural, Nekuda, SolvaPay
High	Voice & realtime infra	L7	LiveKit \$1B, Vapi 1B+ calls; vertical voice niches still open	Rime, Smallest.ai, VoiceRun
High	GPU markets & compute finance	L2	CME and ICE GPU futures; GPU-backed debt everywhere	Liquid Compute
High	Local & on-device inference	L5	55% of 2026 PCs forecast to be AI PCs; robots need on-board inference	Mirai, Argmax, Cactus (YC)
Medium	Synthetic data quality	L3	Gartner forecast: 60% of data leaders face synthetic-data failures by 2027	Rockfish Data, Datology AI, Bifrost
Medium	Agent orchestration niches	L7	Temporal leads at \$12.55B; vertical or edge cases remain	Xpander, Kestra, Trigger.dev
Medium	Web access niches	L7	Exa \$2.2B, Parallel \$2B; specialist data sources remain	Keenable, Browser Use, Kernel
Medium	Grid hookup & power software	L0	4+ year grid waits; software that speeds approvals	Niv-AI, Critical Loop, Soma Energy
Medium	Cooling parts & service	L0	53% of AI chips liquid-cooled; Ecolab paid \$4.75B for CoolIT	Omen AI, Claros, Corintis
Low	Edge chips	L1	Capital heavy; Microchip is buying Hailo	HYFIX, HRDWYR, Mosaic SoC
Low	Networking & optics	L1	Capital heavy; Ayar Labs raised \$650M	Netris, NcodiN
Low	Retrieval & vector search	L3	Turning into a database feature	LanceDB, Chroma, Infino
Low	Gateways & routing	L5	Thin margins; Stripe is buying OpenRouter	TensorZero, Requesty, Martian
Low	Frameworks & SDKs	L6	Labs' own SDKs dominate	CopilotKit, CrewAI, Pydantic AI
Low	Hyperscalers · neoclouds	L2	Billions in capex and debt	Nava (GPU cloud), General Compute, Hyperbolic
Low	Frontier · open-weight labs	L4	Rounds of \$1B+; out of seed range	Deep Cogito

Priority is our judgment for a seed-stage fund: capital needed, room to win, and a visible buyer. The 20 gaps are grouped into 18 rows.

Where to look

DEMO DAYS	YC, a16z Speedrun, Berkeley SkyDeck, Bessemer India, Neo and South Park Commons
OPEN SOURCE	GitHub trending in agent, inference and robotics repos; Hugging Face model and dataset uploads
USAGE DATA	OpenRouter rankings, the MCP registry and npm and PyPI downloads show what developers actually adopt
RESEARCH	arXiv authors leaving labs; Epoch AI and SemiAnalysis notes on bottlenecks
ACQUIRER MOVES	Every acquisition on pages 12, 16 and 18 marks a theme where the next startup has a buyer

Five questions for every infra founder

- Which layer are you in, and which outcome number do you move?
- Who would buy you, and what have they paid for something similar?
- What gets cheaper every year that could make you unnecessary?
- What proof exists beyond what you report yourselves?
- What happens when a lab, NVIDIA or a cloud ships this for free?

Signals to watch

H2 2026	NVIDIA Vera Rubin ramps; CME and ICE GPU futures begin trading after approval
2027	HBM prices forecast to more than double; Lambda plans its IPO
DEC 2027	EU AI Act high-risk rules apply (Annex III)
ONGOING	MCP adoption, agent payment volume, Waymo toward 1M weekly rides

Close the loop

Re-run this atlas each quarter: add new demo-day companies to the taxonomy on page 0, update the 2026 column on page 2, and re-score page 23 as evidence comes in.

Sources

Gap list from the taxonomy (page 0) and pages 6–21. Market facts: Epoch AI, SiliconANGLE (Prime Intellect), Bloomberg and TechCrunch (World Labs, AMI Labs, Exa, Parallel), Fortune, PYMNTS and Visa (payments), SiliconANGLE and GlobeNewswire (LiveKit, Vapi), CME and ICE, Gartner, JLL, TrendForce, SEC (Ecolab), Calcalist (Hailo), Ayar Labs, NVIDIA, TrendForce (HBM), TechCrunch (Lambda), EU sources. Startup names: company funding releases, TechCrunch, SiliconANGLE, YC batches, SemiEngineering funding tables, r1-list.com, Axomap; checked Oct 2026.

A

Early-stage startup index

Every green name in the atlas, with what it does and its latest disclosed round. Pre-seed to Series A, founded 2020 or later, not acquired, as of October 2026.

114

Startups across 41 micro-themes

Infra Conviction

3

Dropped in verification: raised too much or unconfirmed source check, Oct 2026

*

Mega-seeds of \$100M+ in frontier and world-model labs listed for completeness

L0 Power & buildings		
POWER & GRID		
Niv-AI	Frees stranded power inside AI data centers	Seed \$12M · Mar 2026
Critical Loop	Modular microgrids that bring power faster	Series A \$26M · Apr 2026
Soma Energy	AI control layer for data-center power	Seed \$7M · Apr 2026
Piq Energy	AI agents for grid hookup studies	Seed \$5M · Jul 2026
COOLING & POWER GEAR		
Omen AI	Sensors that monitor data-center coolant	Series A \$31M · Aug 2026
Claros	Chip-to-grid power management	Seed \$30M · Mar 2026
Corintis	Microfluidic cooling built into chips	Series A \$24M + \$25M · 2025

L1 Silicon		
ACCELERATORS & MEMORY		
Great Sky	Optoelectronic brain-inspired AI chips	Seed \$14M · Mar 2026
NETWORKING & OPTICS		
Netris	Network automation for GPU clouds	Series A \$15M · Jun 2026
NcodiN	Optical links replacing copper between chiplets	Seed €16M · 2025
EDGE CHIPS		
HYFIX	One chip for drones and robots	Seed \$15M · Apr 2026
HRDWYR	Adaptive AI chips for embedded devices	Series A \$13M · May 2026
Mosaic SoC	Ultra-low-power vision chips for wearables	Pre-seed \$3.8M · Apr 2026
CHIP DESIGN & VERIFICATION		
Architect Labs	AI that designs and verifies custom chips	Seed \$24M · Q2 2026

L2 Compute cloud		
HYPERSCALERS		
Hosted.ai (sells to clouds)	GPU pooling software for clouds	Seed \$19M · Mar 2026
NEOCLOUDS		
Nava (GPU cloud)	GPU cloud for India and Southeast Asia	Series A \$22M · Apr 2026
General Compute	Inference cloud on custom chips	Seed \$15M · May 2026
Hyperbolic	Open AI cloud and GPU marketplace	Series A \$12M · Dec 2024
NexGen Cloud	Sovereign GPU cloud (Hyperstack)	Series A \$45M · Apr 2025
GPU MARKETS & FINANCE		
Liquid Compute	Marketplace and planned exchange for compute	Seed \$15M · Sep 2026
ORCHESTRATION & UTILIZATION		
Zymtrace	GPU profiling to cut waste	\$12.2M raised · 2026
Chamber (YC)	Agents that reclaim idle GPUs	YC W26 · undisclosed

L3 Data		
SYNTHETIC DATA		
Rockfish Data	Synthetic operational data for enterprises	Seed \$4M · Jan 2025
Datology AI	Training-data curation and rephrasing	Series A \$46M · May 2024
Bifrost	Synthetic 3D data for vision and robots	Series A \$8M · Oct 2024
WAREHOUSES & PIPELINES		
Eventual (Daft)	Multimodal data engine	Seed + A \$30M · Jun 2025
Spiral	Database for multimodal data on object storage	Seed + A \$22M · Sep 2025
LlamaIndex	Document parsing and knowledge agents	Series A \$19M · Mar 2025
Extend	Document processing cloud for AI	Seed + A \$17M · Jun 2025
STORAGE & FILE SYSTEMS		
Tigris Data	Global object storage for AI	Series A \$25M · Oct 2025
Akave	S3-compatible storage, no egress fees	Seed \$6.65M · Mar 2026
Peak:AIO	Software-defined storage for AI	Seed \$6.8M · Oct 2025
RETRIEVAL & VECTOR SEARCH		
LanceDB	Multimodal lakehouse and vector database	Series A \$30M · 2025
Chroma	Open-source embedding database	Seed \$18M · Apr 2023
Infino	Text and vector search on object storage	Seed \$7.5M
ROBOT & SENSOR DATA		
Midcentury	First-person human data for physical AI	Seed \$15M · Sep 2026
Human Archive	First-person and touch data from workers	\$8.2M · May 2026
Vision Lab	Real factory video for robot models	Seed \$6M · Jun 2026

L4 Model building		
FRONTIER LABS		
Ineffable Intelligence*	RL-first superintelligence lab	Seed \$1.1B · Apr 2026
humans&*	Human-centric frontier lab	Seed \$480M · Jan 2026
Flapping Airplanes*	Data-efficient frontier research lab	Seed \$180M · Jan 2026
Inception	Diffusion language models for text and code	Seed \$50M · Nov 2025
OPEN-WEIGHT MODELS		
Deep Cogito	Open-weight hybrid reasoning models	Series A \$43M · Aug 2026
TRAINING & ML WORKFLOW		
Callosum	Runs AI jobs across mixed chip types	Seed \$10.25M · Feb 2026
FINE-TUNING & CONTINUAL LEARNING		
Refiant AI	Compresses and retrains models for edge cost	Seed \$5M · Apr 2026
RL ENVIRONMENTS		
Proximal	Long-horizon coding environments for labs	\$15M · Sep 2026
Bespoke Labs	RL environments and post-training data	Series A \$31.75M · Jul 2026
Halluminate	Finance and consulting RL environments	Series A \$30M · Oct 2026
Taste Labs	Preference data and RL environments	Seed \$18.5M · Jun 2026
Veris AI	Simulation sandboxes to train agents	Seed \$8.5M · Jun 2025

L5

Serving & efficiency

B

Spotlight startup index

The spotlight startups in stack order: where each sits, what it does in one line, its first-pass score from page 23, and the page with its full card.

32

Infrastructure companies across L0 to L8
Infra Conviction

4

Apps or stealth, shown under outcomes
Infragrid, Nexor AI, Sterling, Bric

9 · 14 · 13

Tier A · Tier B and B- · Tier C, exited or stealth
page 23

Layer	Company	Micro-theme	What it does	Score	Tier	Card
L0	Cybernetic Physics	DC build & operations	Puts other makers' humanoid and mobile robots inside data centers to rack servers, run cables and do maintenance around the clock.	9	A	p6
L1	CIMTech	Accelerators & memory	An add-on memory accelerator card that sits beside inference GPUs and speeds up long-context answers.	7	B-	p7
L1	VerifAIX	Chip design & verification	Checks that chip designs, including AI-written ones, actually do what the spec says.	10	A	p7
L2	Harbor AI	Orchestration & utilization	Finds why GPU jobs fail or slow down, proposes fixes for approval, then confirms performance recovered.	8	B	p8
L3	Space	AI storage & file systems	A cloud drive that behaves like a local disk, so a laptop or an agent can open terabytes of files without downloading them.	8	B	p10
L3	SF Data	Data rights & licensing	Buys real records of AI coding agents at work from app companies, turns them into training tasks and graders, and licenses them to frontier labs.	6	C	p10
L3	SenseLab	Expert data, labels & agent traces	Records what agents decided and whether it worked, shares lessons across agents, and exports good runs as training data for smaller models.	6	C	p10
L3	Munari	Robot & sensor data	Scores the quality of the human demonstrations used to train robots, per episode, task and operator.	7	B-	p10
L3	Vound	Robot & sensor data	Pays skilled workers in factories, farms and hotels to wear cameras, labels the first-person video and sells it to robotics labs.	8	B	p10
L3	Embrasure	Warehouses & pipelines	A data warehouse that AI agents build and maintain themselves, on the customer's own open-format storage.	6	C	p10
L4	Proper Motion	Fine-tuning & continual learning	Makes deployed agents improve from real use by linking their behavior to customer churn and testing fixes.	8	B	p12
L4	Zorbe	Fine-tuning & continual learning	Turns a company's repetitive AI jobs into a smaller model the company owns, trained with distillation, fine-tuning and RL.	5	C	p12
L4	Oruk	Specialized models	Speech models that capture emotion, tone and intent as well as the words.	9	A	p12
L4	Outerbounds	Training & ML workflow	Managed enterprise version of Metaflow, Netflix's open-source tool for building and shipping ML and AI workflows.	—	Exit	p12
L5	Hoid	Inference optimization	Uses AI agents to tune a model so it runs fast on a given chip, work engineers normally do by hand for weeks.	9	A	p13
L5	Minima	Inference optimization	Shrinks large open models so they need fewer GPUs and answer faster, sold as a self-hosted engine or API.	7	B-	p13
L5	Pipeshift	Inference platforms & APIs	An ultra-low-latency cloud for running open-source models in real time.	6	C	p13
L6	PreMan	AI code review & testing	Simulates AI agents using a product, finds where workflows break, and turns the failures into regression tests.	7	B-	p14
L6	TrueCourse	AI code review & testing	Turns product requirements into tests that run against the live product and block code that breaks them.	8	B	p14
L6	Cube Computer	Sandboxes & dev boxes	A personal always-on cloud machine for running coding agents side by side.	7	B-	p14
L7	Switchfrog	Agent identity & auth	Tells a SaaS company whether a logged-in session is a human or an AI agent, so it can govern or charge for agent use.	9	A	p16
L7	Rollup	Memory & context	An SDK that lets a new app understand a user from day one by analyzing their photo library on the device, with consent.	7	B-	p16
L7	Cluster	Tools, connectors & MCP	Gives agents like Claude or Cursor go-to-market abilities through MCP: 50+ data providers, buying signals and outreach.	5	C	p16
L7	Swytchcode	Tools, connectors & MCP	Lets agents call business APIs like Stripe or HubSpot correctly and safely, with auth, retries, policy and audit built in.	7	B-	p16
L8	Abliteration.ai	AI security & guardrails	Hosted open models with their refusal behavior removed, so security teams can simulate real attackers.	7	B-	p18
L8	Harden	AI security & guardrails	Runs on the developer's machine and checks every action a coding agent tries, blocking secret leaks and destructive commands.	9	A	p18
L8	Kairo	AI security & guardrails	Reviews each code change for security holes before it ships and keeps attack-testing the live app.	6	C	p18
L8	Aurva	Data privacy & security	Shows in real time who or what, including AI agents, is touching a company's sensitive data.	10	A	p18
L8	Skyflow	Data privacy & security	A vault for sensitive data, offered as an API, so apps and agents can use data without holding the raw values.	9	A	p18
L8	AgentCat	Evals & observability	Analytics and session replay showing how AI agents use your product through MCP, Claude connectors or ChatGPT apps.	9	A	p18
L8	Trusys	Governance & compliance	Tests, attacks, monitors and enforces rules on a company's AI agents from one platform.	5	C	p18
L8	ZebraTruth	Governance & compliance	Checks an agent's proposed real-world action against laws, regulations and company policy before it happens.	7	B-	p18
★	Infragrid	Agentic apps	A desktop assistant that watches how you work, records a task once, and runs agents to repeat it.	5	C	p19
★	Nexor AI	Vertical AI	AI agents that run revenue follow-ups across calls, iMessage, email and Instagram, tied into the customer's CRM and payments.	6	C	p20
★	Sterling	Vertical AI	AI operators that take over revenue work such as renewals and ad spend, built on each customer's own data definitions.	6	C	p20
★	Bric	Watchlist (stealth)	Stealth.	—	Meet	p19

EVIDENCE
Verified funding and named customers score highest; self-reported traction scores 2

POSITION
How hard it is to copy or bundle: data, workflow lock-in, technical depth

EXIT PATH
Whether a large company has recently paid for something close

ROOM
How few funded rivals there are, and whether a giant gives it away

Reading the score

Each factor is scored 1 to 3 from public evidence only, for a maximum of 12. The score ranks where to spend diligence time. It is not a view on whether to invest.

Sources

Company sites, a16z Speedrun and Frontrun pages, Berkeley SkyDeck, funding releases; detail and caveats on each company card.

IC Glossary, method, sources

The words used across the atlas in plain English, how companies and facts were chosen and checked, and where the numbers come from.

35 Terms defined
glossary

~100 Facts corrected in the October 2026 check
independent review

Oct 12 2026: date every figure was last checked
Infra Conviction

Glossary

Accelerator A chip built for AI math, such as a GPU or Google's TPU.
Inference Running a trained model to answer a request. Paid for every time it is used.
Neocloud A cloud built only to rent GPUs, such as CoreWeave or Nebius.
Optics Using light instead of copper to move data between chips and racks.
Open-weight model A model whose trained weights anyone can download and run.
World model A model that learns how the physical world behaves, used to train robots in simulation.
Gateway / router One entry point that sends each request to the best-value model.
Sandbox A throwaway computer where an agent can run code safely.
Durable execution Keeping a long job alive through crashes so it resumes where it stopped.
Evals Tests that score whether AI output is right, safe and on-task.
EU AI Act The EU law grading AI by risk; most high-risk rules apply from Dec 2027.
Mega-seed A first round of \$100M or more, typical for new frontier labs.

HBM High-bandwidth memory stacked next to an AI chip. Scarce, and about half a top chip's build cost.
Token A chunk of text, about three-quarters of a word. AI is priced per token.
Colocation Renting space, power and cooling in someone else's data center.
Edge chip A low-power AI chip inside a phone, camera, car or robot.
Fine-tuning Further training a model on narrow data so it does one job better.
Synthetic data Training data made by models rather than collected from people.
KV cache A model's short-term memory of a conversation. Reusing it saves compute.
Agent An AI system that plans steps and uses tools to finish a task, not just answer.
Agent identity Proving which agent is acting, for whom, and what it may touch.
Observability Tracing what an AI system did at each step, to debug and improve it.
Run rate Latest month's revenue times twelve. Common for fast-growing AI firms.
Self-reported A figure the company states that no third party has confirmed.

Training Teaching a model from data. Done once per model, at huge cost.
Hyperscaler AWS, Microsoft Azure, Google Cloud or Oracle: the giant general clouds.
Liquid cooling Pumping coolant to the chips because air can no longer carry the heat away.
Frontier lab A company building the most capable general models, such as OpenAI or Anthropic.
RL environment A practice world with tasks and graders where a model learns by trial and error.
Vector search Finding the most similar stored text or images, used to give models the right facts.
Compiler / kernel Software that turns a model into fast instructions for a specific chip.
MCP Model Context Protocol: the standard plug between AI models and software tools.
x402 An open protocol that lets software pay for an API call with stablecoins.
Prompt injection Hidden text that tricks a model into following an attacker's instructions.
Gross margin Revenue left after the direct cost of serving customers.

Method

WHAT IS A SPOTLIGHT STARTUP Early-stage companies profiled in depth, each re-checked against its own site and public funding records, and placed by what it sells today.	WHAT IS EARLY-STAGE Pre-seed to Series A (or a small Series B), founded 2020 or later, under about \$75M raised, not acquired. Mega-seed labs are marked *.	HOW FACTS WERE CHECKED Every page was reviewed by an independent agent that had not written it, checking numbers, dates, deals and logic against primary sources. About 100 items were corrected in October 2026.	HOW TO READ NUMBERS “Reported” means press, not the company. “Self-reported” means the company only. Forecasts carry the forecaster and year. Run rates are not annual revenue.	WHAT IS STILL OPEN About 60 smaller 2026 figures could not be re-confirmed in the last pass and should be checked before external use. Placement calls noted on cards: SenseLab spans L3 and L7; Nexor AI and Sterling sell function-specific agents.
--	---	---	---	---

What changed in the October 2026 check

DEALS Acquired leaders tagged with their buyer; pending deals reworded as “agrees to buy” (Nscale–Anyscale, Microchip–Hailo, Stripe–OpenRouter).	NUMBERS CoreWeave debt ~\$35B; Ollama \$88M; AI-security spend +70%; Goldman humanoids \$38B by 2035; H100 price fall ~58%; capex \$735–750B.	LOGIC Page 22 quadrant counts now sum to 31 with exclusions named; EU AI Act dates split Annex III (Dec 2027) and products (Aug 2028).	WORDING Self-reported claims flagged; overclaims softened (robot data, founder backgrounds); Fable 5 and Mythos 5 pause dated Jun 12 to Jul 1.	STARTUPS 114 early-stage startups added and checked against source pages; Nous Research, Gadium and one unconfirmed company dropped.
--	---	--	--	--

Sources by type

Filings and company results NVIDIA, CoreWeave 10-Q, SK hynix, Microsoft, Alphabet, Amazon, Meta, Oracle, Vertiv, Cerebras	Research Epoch AI, SemiAnalysis, METR, IEA, LBNL and DOE, EPRI, Uptime Institute, JLL, TrendForce, Synergy, Google DeepMind	Surveys and forecasts Menlo Ventures, ICONIQ, Bessemer, KeyBanc, Gartner, IDC, McKinsey, Goldman Sachs, IBM, OWASP
News Reuters, Bloomberg, The Information, TechCrunch, Fortune, CNBC, SiliconANGLE, DCD, Axios, Calcalist	Startups Company sites and releases, a16z Speedrun, Frontrun, Berkeley SkyDeck, Y Combinator, SemiEngineering, rl-list.com, Axomap, Rye	Protocols and data MCP blog, OpenRouter rankings, npm and PyPI download counts, Hugging Face, Cloudflare Radar

Disclaimer

For research and discussion. Built from public information; company figures may be self-reported, and private-market data changes quickly. Nothing here is investment advice.

Re-running it

Refresh each quarter: add new startups from demo days and funding news, update the 2026 columns, re-check any figure marked reported or self-reported, and re-score page 23.