

WOODSIDE
CAPITAL
PARTNERS



A Guide to AI in 2026

Understanding the Moving Parts of the AI Ecosystem

Mark Bagley
Managing Director

Woodside Capital Partners | June 2026

What We'll Cover

1 Frontier Models & Benchmarks

Who's building what — 12 frontier models compared

2 The Release Race

10 releases in H1 2026 alone — what's driving it

3 WCP Model Value Score

DeepSWE 1.1 + cost efficiency: bang for the buck

4 Tokens: The New Currency of Work

Why AI bills are going up, not down

5 AI Coding & The SaaSocalypse

90% adoption, vibe coding, and what it means for SaaS

6 Local Models and Agents:

Apple is All-In on Local AI and Privacy

7 Architecture: Dense vs MoE

How models are built and self-improvement loops

8 Funding & Economics

\$373B raised, two IPOs imminent

9 Infrastructure & Bottlenecks

\$725B capex and five constraints that matter

10 The Chip Race

NVIDIA at 77%, custom silicon rising

11 The AI Stack

Seven layers between model and value

12 Agents, Context & Memory

From chatbots to persistent AI personal assistants

13 What Comes Next

Six trends shaping 2027 and beyond

The Frontier Models: Who's Building What

Model	Developer	Access	Released
Llama 4	Meta	Open-weight	Apr 2025
Gemini 3.1 Pro	Google	Closed	Feb 2026
Mistral Large 3	Mistral AI	Open-weight	Feb 2026
GPT-5.5	OpenAI	Closed	Apr 2026
Qwen 3.7 Max	Alibaba	Closed (API)	Apr 2026
Claude Opus 4.8	Anthropic	Closed	May 2026
DeepSeek V4	DeepSeek	Open (MIT)	May 2026
Kimi K2.7	Moonshot AI	Open-weight	Jun 2026
GLM-5.2	Zhipu AI	Open (MIT)	Jun 2026
Claude Fable 5	Anthropic	Closed	Jun 2026
MAI-Thinking-1	Microsoft	Closed	Jun 2026
Claude Sonnet 5	Anthropic	Closed	Jun 2026
Apple AFM 3	Apple	Closed (on-device)	Jun 2026
Grok 4.5	xAI	Closed	Jul 2026

Key Terms



Closed

API access only. You can't see or modify the model weights.



Open-weight

Weights downloadable. Run on your own hardware, fine-tune for your use case, modify freely. No cloud dependency, no API costs, no vendor lock-in. In practice: download once, run forever.



Open vs Closed: The Strategic Choice

Open-weight models now within 5-10% of frontier on most tasks. Closed = max capability but vendor lock-in and higher cost. Open = control, privacy, 5-25x cheaper, but you run the infra.



Dense vs MoE

Dense: every parameter fires every query. Simple but costly. MoE: trillions of params but only a fraction activate per query. Only Claude remains fully dense.

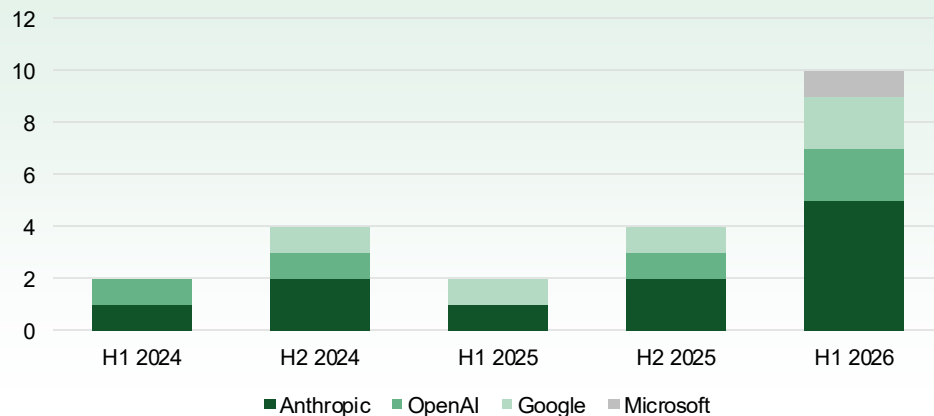


Claude Fable 5 / Mythos

Anthropic's most powerful public model. Mythos-class (above Opus). 80.3% SWE-Bench-Pro (vs GPT-5.5: 58.6%, Gemini 3.1: 54.2%). \$10 input / \$50 output per million tokens (2x Opus pricing). 1M context window, 128K max output.

The Release Race: Models Shipping Faster, Getting Bigger

Major frontier model releases per half-year (stacked by lab)



Milestones

2024

~5 major releases across all labs

Claude 3, GPT-4o, o1, Gemini 1.5

2025

~10 releases, pace doubles

Claude 4, GPT-5, Gemini 3, Opus 4.5

H1 2026

10 releases in 6 months

5 Anthropic models in 5 months alone

What's driving this?

Self-improvement loops

80%+ of Anthropic code is Claude-authored.
Engineers ship 8x more code per quarter.

Competitive pressure

Labs must respond within weeks of a rival's release. The window for advantage shrank from years to days. Ship or be irrelevant.

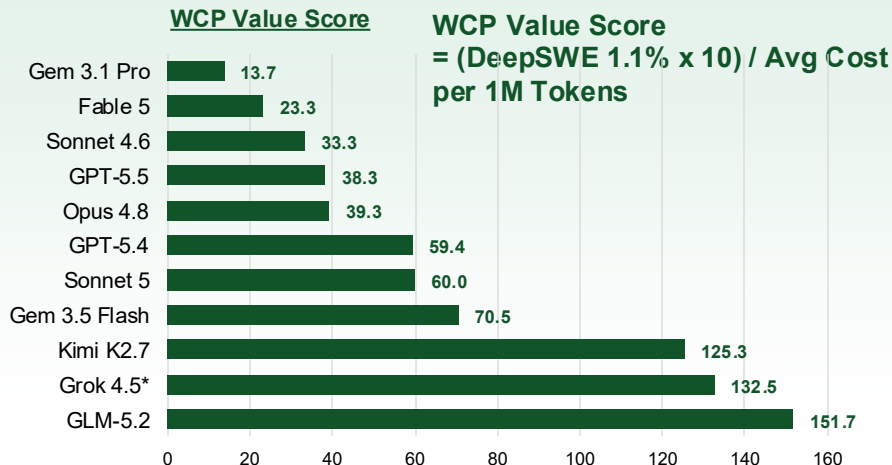
Training costs commoditized by AI

AI writes 80%+ of code at labs, cutting engineering costs. More output per dollar spent on development.

Net effect: 2024: ~1 major model per lab per year → 2026: every 1-2 months per lab

Cheap Models Win on Value: Frontier Models Pay 20x More for Only 2-3x the Capability

Major frontier models



Most Coding Capability per Dollar

- **GLM-5.2 leads on value (152)**
Open-weights: 44% DeepSWE at ~\$2.90/M blended.
- **Kimi K2.7 close behind (125)**
31% coding at ~\$2.48/M avg.
- **Grok 4.5 would rank high (133*)**
On xAI's self-reported 53% at \$4/M — unverified.
- **Frontier models = worst value**
Fable 5 (23), GPT-5.5 (38), Opus 4.8 (39): top raw coding, bottom of the value chart.

Key takeaway

Frontier costs 6–12x more per token for ~2x the coding capability.

Source: Datacurve DeepSWE v1.1 (Jul 1) + provider list pricing; Grok self-reported

What is DeepSWE 1.1?

A contamination-free benchmark: real bugs in real codebases the model has never seen. Unlike SWE-bench (widely memorized by 2025), DeepSWE uses fresh, unpublished tasks. Scores from Datacurve DeepSWE v1.1 (Jul 1, 2026). Models absent from v1.1 are excluded.

* Grok 4.5: 53% is xAI self-reported (not yet listed on Datacurve). Treat as indicative.



Model Routing Is Key

Route to a model that is exactly matched to the task
Simple → local (\$0). Medium → mid-tier. Hard → frontier API.

Examples:

Apple orchestrator: AFM 3 Core → Cloud → Cloud Pro
OpenRouter.ai: 300+ models, auto-route by cost

Tokens: The New Currency of Work — Why AI Bills Are Going Up

Model Pricing (per million tokens)

Model	Input	Output
Claude Fable 5	\$10.00	\$50.00
GPT-5.5	\$5.00	\$30.00
Claude Opus 4.8	\$5.00	\$25.00
Claude Sonnet 4.6	\$3.00	\$15.00
Gemini 3.1 Pro	\$2.50	\$15.00
Llama 4 Scout	\$0.18	\$0.59
DeepSeek V3	\$0.27	\$1.10

What Tasks Cost in Tokens

Task	Tokens	Open-src	Frontier
Simple chatbot reply	200-2K	\$0.001	\$0.06
Email draft (500 words)	~1,500	\$0.004	\$0.08
Summarize 10-page doc	10K-15K	\$0.05	\$0.75
RAG question + context	2K-12K	\$0.02	\$0.36
Code review (single file)	5K-20K	\$0.06	\$1.00
Agentic coding task	15K-1 M+	\$0.05	\$50.00
Full app build (vibe)	500K-5M	\$1.50	\$250

Costs per task. Open-src = Llama 4 hosted. Frontier = Fable 5. Mid-2026 API pricing, USD.

Why AI Bills Keep Rising



Frontier is 10-55x more expensive than open-source.

The range: Fable 5 (\$30/M avg) vs DeepSeek V3 (\$0.69/M avg) = 43x.
Gemini 3.1 Pro (\$8.75/M) vs Llama 4 Scout (\$0.39/M) = 22x.



Consumption is exploding

Weekly token volume: +3,800% YoY (OpenRouter).
Goldman Sachs projects 24x increase by 2030.



Reasoning models use 8x more tokens

Chain-of-thought, tool use, and verification loops multiply token counts per request dramatically.



Agentic coding: up to 1,000x more per task

An agentic coding session can cost \$50+ on a frontier model. Full app builds hit \$250. Even "cheap" models add up when agents loop thousands of times.



Price per token falls, total spend rises

Unit costs drop 10x/year but usage grows faster.
Jevons paradox: cheaper AI means more AI consumption, not lower bills. Enterprise AI budgets are rising 40-60% YoY even as per-token prices fall.



The fix: model routing

Route each task to just enough model. Simple → local (\$0).
Medium → mid-tier. Hard → frontier.
Most tasks don't need the most expensive model.

AI Coding & The SaaSpocalypse: Fear, Reality, and What Actually Changes

90%

of devs use AI tools

46%

of code is
AI-generated

63%

of vibe coders
aren't developers

\$22B

coding market
by 2030

Why Coding Is AI's Sweet Spot



Coding is the highest-ROI use of AI today. It's structured, testable, and improvable.

Developer tools:

GitHub Copilot (29% adoption, 26M+ users, \$2B+ ARR), Cursor (\$2B+ ARR), Claude Code (18% adoption, 46% most-loved), Codex, Windsurf, Devin, Augment

Vibe coding (non-developers building apps):

Lovable (8M users, \$50M ARR), Replit (\$9B val, 40M users), Bolt.new, v0 by Vercel (\$9.3B val)

Productivity Impact



DeepSWE: Fable 5 scores 80.3% on real-world software engineering tasks.
GPT-5.5: 70% on same benchmark. Models are genuinely capable coders.

18% faster (2026 study) — up from 19% slower in 2025.

Developers learned when and how to use the tools effectively.

Anthropic: 80%+ of code written by Claude. 8x more code shipped per quarter.

Goldman Sachs: 30% developer velocity improvement across enterprise.

Gartner: 40% of enterprise apps use vibe coding by 2028.

The SaaSpocalypse: AI Coding Triggers a SaaS Valuation Crisis

The Fear



Vibe coding lets anyone build software in minutes. If AI can replace SaaS products, why pay premium valuations?

SaaS traded at an average of 6-8x revenue through 2025. The SaaSpocalypse drove private multiples down to 3.1x by Mar 2026.

Software ETFs worst quarter since 2008. SaaS stocks fell while Nasdaq +18%.

The Reality



Large enterprise systems are systems of record. Payroll engines, compliance workflows, ERP logic — these can't be vibe-coded in a weekend.

What AI actually replaces: the UI layer. Agents build bespoke, highly customized interfaces — but they still need access to the data in these systems of record.

What Actually Changes



1. SaaS companies must open their systems to agents via APIs. Those that don't get bypassed.

2. New software products can be built far more quickly and cheaply with minimal staff. Startups launch in weeks, not years.

May 2026: best month for software stocks since 2001. Fear subsiding as AI seen as additive, not replacement.



The real disruption isn't replacing enterprise software — it's making the traditional UI irrelevant.

Local agents build custom interfaces that access data via APIs. SaaS companies that don't make their systems agent-accessible risk being bypassed entirely.

Sources: JetBrains AI Pulse Jan 2026, Stack Overflow Dev Survey 2025, METR/Turing Study 2026, CNBC May 2026, Gartner 2026

Local Models and Agents: Apple is All-In on Local AI and Privacy

Advantages of Local Models

\$0 per token

Eliminate API costs entirely. As token consumption explodes, local is the only way to control spend.

Zero latency, full privacy

Data never leaves the device. No network round-trip. Works offline.

Advantages of Local Agents

Agents run on your machine

OpenClaw, Claude Code, Cursor orchestrate tasks, route to models, read local files.

Smart routing = smart spending

Local model: STT, TTS, summarization. Cloud only for reasoning + world knowledge.

Use of Intelligent Routing

Route each task to just enough model. Simple → local (\$0). Medium → private cloud. Hard → frontier API.

Apple Is All-In on Local AI (WWDC 2026)

2.5B active devices. 1.5B iPhones.

What Apple does with AI matters.

Trusted privacy brand + massive install base = AI at global scale.

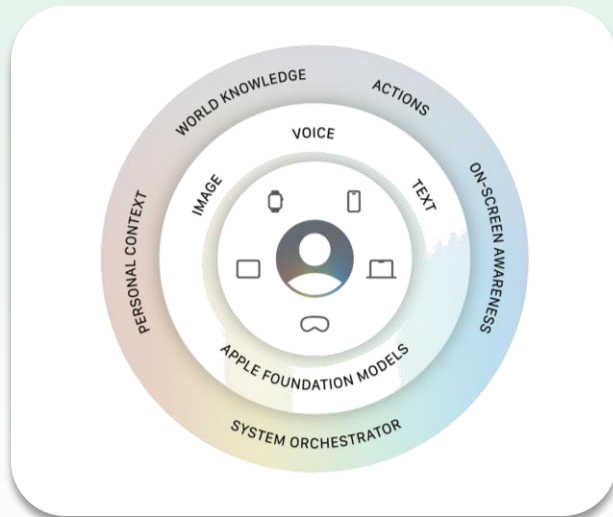
5 Foundation Models:

On-Device:

AFM 3 Core (3B dense)
AFM 3 Core Advanced (20B sparse, 1-4B active, multimodal)

Private Cloud Compute:

AFM 3 Cloud (Apple silicon)
ADM 3 Cloud Image (image gen)
AFM 3 Cloud Pro (NVIDIA GPUs in Google Cloud, agentic/reasoning)



Results: Most Requests Never Leave the Device

Evaluated OpenAI and Anthropic. Chose Google.

Key innovation: AFM 3 Core Advanced stores 20B params in flash, loads only 1-4B into RAM per prompt

Private Cloud Compute: data never stored on servers, cryptographic attestation, no access even by Apple.

How Models Are Built: Dense vs MoE Architecture

MoE (Mixture-of-Experts) is the dominant architecture for frontier models.

A 3T-param MoE model activates only 2-5% of params per query, delivering trillion-param intelligence at a fraction of the cost.

Only Claude remains fully dense: every parameter fires every query. Simpler, but far more expensive to run.

Mixture-of-Experts: Total vs Active Params

Model	Total	Active	Architecture
Grok 4.20	~3T	Undiscl.	Sparse MoE
Llama 4 Behemoth	2T	288B	128+ experts
DeepSeek V4 Pro	1.6T	49B	MoE hybrid
MAI-Thinking-1	~1T	35B	Sparse MoE
Kimi K2	1T	32B	384 experts
GLM-5	745B	~42B	256 experts
Mistral Large 3	675B	41B	Granular MoE
Llama 4 Maverick	400B	17B	128 experts
Qwen3-235B	235B	22B	MoE
Claude Opus 4.8	Undiscl.	All (dense)	NO MoE
Claude Fable 5	Undiscl.	All (dense)	NO MoE

Grok 4: 3T total but ~100B active = 30x cheaper inference.

Claude bets dense quality justifies the cost premium.

Why This Matters for Development Speed

Cheap Inference = Faster AI Development

MoE cuts inference costs 10-30x. Cheaper inference means labs can run their models 24/7 on their own code, training data generation, and testing. That is the link to faster release cadence.

The Self-Improvement Loop

80%+ of Anthropic code written by Claude.
Engineers ship 8x more code per quarter.

Task Horizon (doubles every 4 months):

2024: 4 min → 2025: 1.5 hr → 2026: 12 hr
→ **2027: ~48 hr (projected)**

Mythos Preview: 52x code speedup
(vs 3x for Opus 4)

AlphaEvolve: AI optimized its own
training kernel

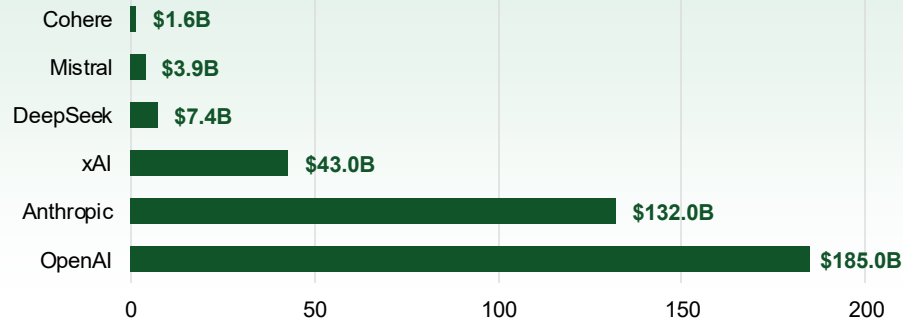
Loop not fully closed; humans still control
goals and architecture decisions.

Costs Still Outpacing Revenue — But Revenue Is Building Fast

Revenue vs Spending: The Gap

Player	Revenue (ARR)	Profitability	Note
OpenAI	\$24B (run rate)	Losses: ~\$14-17B in 2026	Profitable target: 2029-30
Anthropic	\$47B	Q2 2026: First profitable qtr	Winning efficiency race
xAI	~\$2B (est.)	Deeply unprofitable	50M Grok users
AI Sector Total	\$50-80B direct rev	\$725B capex, >8:1 ratio	Largest bet in tech history

\$373B Raised Since 2019



Valuations & IPO Pipeline

Anthropic: \$965B (S-1 filed Jun 1)

\$65B Series H (May 28), \$47B ARR
Confidential filing, Fall 2026 target

OpenAI: \$852B (S-1 filing imminent)

Goldman + Morgan Stanley advising
Sep-Nov 2026 window, \$1T target

SpaceX: \$1.75T (IPO priced \$135/sh)

Raising \$75B, largest IPO ever
All-primary shares (no insider sales)

xAI: \$250B, DeepSeek: \$45-59B

Mistral: \$14B, Cohere: ~\$6B

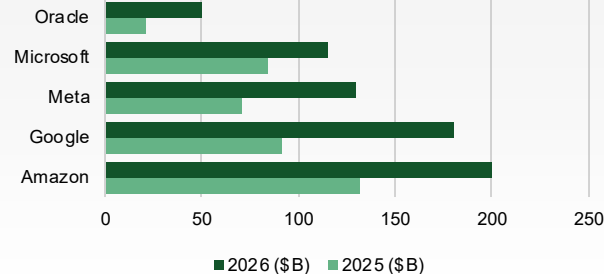
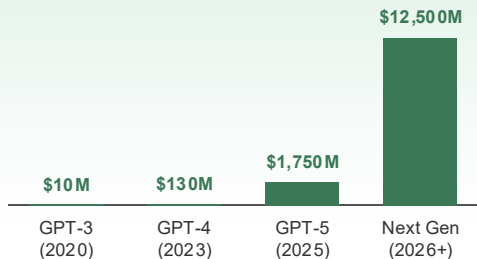


\$373B raised, \$725B capex, but only \$50-80B in AI revenue. \$8 in capex for every \$1 earned.

Revenue is accelerating: Anthropic went from \$1B to \$30B ARR in 15 months (30x). But only companies with massive capital access survive. Winner-takes-most.

The Physical Cost of Intelligence: \$725B Capex and Five Bottlenecks

Training Cost Explosion



Hyperscaler Capex: \$725B in 2026

The Key Numbers

- **\$725B hyperscaler capex**
- in 2026 alone
- **Training cost: \$10M to \$12.5B**
- 1,250x increase in 6 years
- Only 5-7 orgs can afford frontier training runs.
- Inference costs dropping 10x per year
- Deployment is getting radically cheaper.

The Five Bottlenecks



Power

41 GW now; nuclear deals are standard



Talent

~5,000 frontier researchers worldwide



Water

5M gal/day per DC; permits being denied



Data Quality

Running out of clean text; synthetic helps



Chip Supply

TSMC 3nm fully booked; 12-18 month lead times

Frontier Lab Infrastructure Commitments

OpenAI / Stargate

\$600B thru 2030. Renting + building hybrid.

Anthropic

\$200B / 5yr GCP + \$1.25B/mo to xAI Colossus.

SpaceX / xAI

\$21B capex in 2025. \$13B on AI alone.

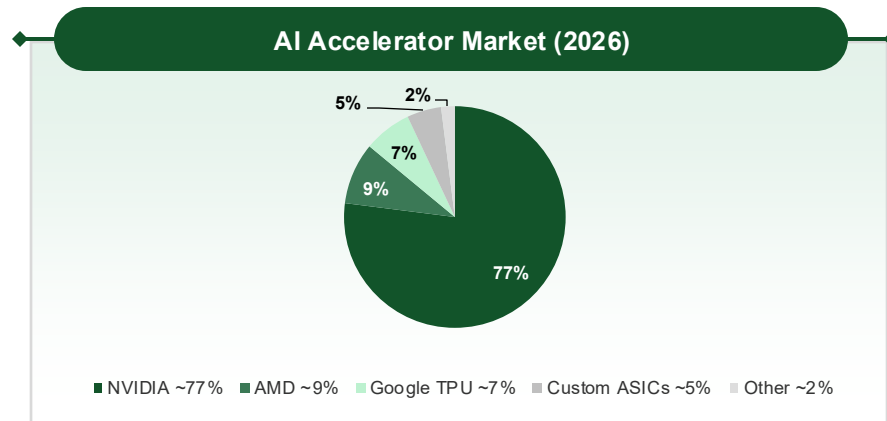
Google / Amazon

\$180B + \$200B capex. Both expanding custom silicon.

Infrastructure Bottlenecks: Permitting adds 2-4 years. NIMBYism + grid upgrades slow data center builds in the US and EU.

The Chip Race: NVIDIA Leads at 77%, But Custom Silicon Is Rising

Chip	Performance	Memory	Price
NVIDIA B200	4,500 TFLOPS	192 GB HBM3e	\$40-50K
AMD MI350X	4,600 TFLOPS	288 GB HBM3e	\$20-30K
Google TPU v7	4,614 TFLOPS	HBM (GCP)	GCP only
Amazon Trainium 3	2,517 TFLOPS	144 GB HBM3e	AWS only
Microsoft Maia 200	>5 PFLOPS	216 GB HBM3e	Internal
Huawei Ascend 910C	~800 TFLOPS	128 GB HBM2e	China only



Cloud GPU Pricing (\$/GPU-hour, June 2026)

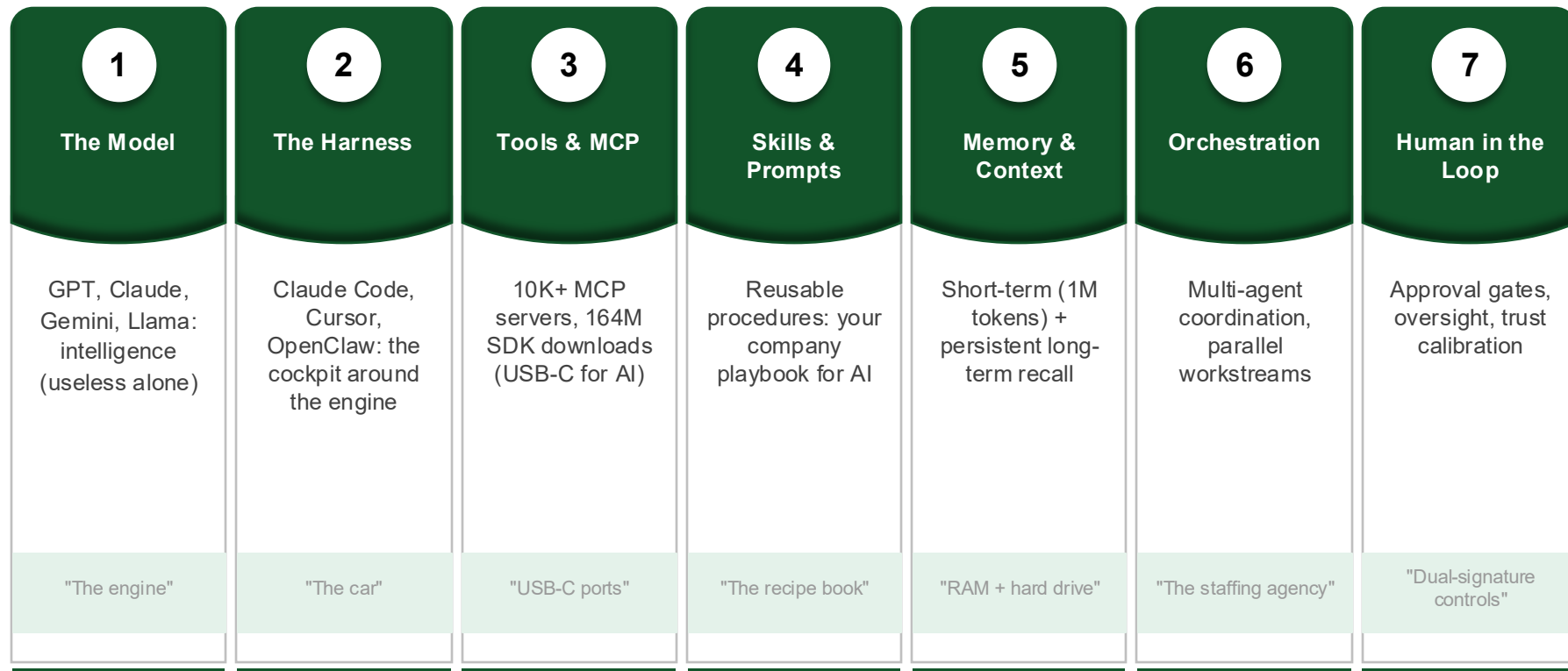
Chip	Spot/Low	On-Demand	Notes
NVIDIA H100	\$1.03	\$3.11	Legacy; prices crashed 60% as Blackwell ramps
NVIDIA B200	\$2.12	\$14.24	Current workhorse for training + inference
NVIDIA B300	\$6.10	\$18.00	Latest gen; limited availability
AMD MI300X	\$1.99	\$7.86	Best value for inference workloads

MAI-Thinking-1: Azure AI Foundry, pricing TBD. Meta MTIA: internal only; Llama 4 hosted via third parties at \$0.15-0.50/M input.

NVIDIA: \$75.2B DC revenue in Q1 FY27 (+92% YoY), ~\$3.5T market cap. But share fell from 87% to 77%. Every hyperscaler now builds custom silicon. AMD validated by Meta \$100B deal.

The AI Stack: Seven Layers Between Model and Value

The model is just the engine. Prompt quality, tool access, memory design, and orchestration determine the actual experience.



From Chatbots to Personal Assistants: How Memory Changes Everything

The shift: An agent with persistent memory turns AI from a tool into a personal assistant that knows your work.

Run multiple projects simultaneously — no re-briefing, no lost context.

The Problem: Traditional AI

One conversation at a time

Every session starts from zero. You re-explain your role, your project, your preferences.

Context evaporates on close

Close the tab, lose the thread. No recall of past decisions, documents, or reasoning.

Prompt engineering, not project management

You spend time crafting the right question instead of getting the right answer.

Before:

15 min re-explaining context each session.

Every new chat = starting over.

The Solution: Agents with Memory

Persistent context across sessions

Walk away for a week. Pick up exactly where you left off. Every decision is retained.

Multiple workstreams in parallel

Each project gets its own memory space: documents, history, tasks, timelines.

No re-briefing. No lost context.

The agent manages the context, recall, and execution. You manage the work.

After:

0 min. Agent already knows.

Like a personal assistant with perfect recall.

Key Agent Capabilities That Make This Possible

Short-term Memory

1M+ tokens per session.
Entire codebases in context.

Long-term Memory

Files, decisions, preferences
persist across sessions.

Tool Use

Email, CRM, code, web, file
systems — all accessible.

Multi-agent

Parallel sub-agents for
research, writing, analysis.

What Comes Next: Six Trends Through 2027+

1. Agents Become the Interface



AI agents replace SaaS dashboards. Your CRM, analytics, and project tools become API endpoints consumed by agents, not screens you click through. 40% of enterprise apps will feature agents by end of 2026.

2. Data Is the New Moat



As model capabilities converge, proprietary data becomes the durable advantage. Training data is running out. Synthetic data pipelines, licensing deals, and 'data flywheels' separate winners from followers. Your company's data strategy is your AI strategy.

3. Training Costs Hit \$10B+



Next-gen frontier training runs cost \$5-20B. Only 5-7 organizations on earth can afford it. But inference costs drop 10x per year; deployment is getting radically cheaper.

4. Power Is the Bottleneck



Not GPUs, not data: electricity is the constraint. Nuclear power deals are now standard. Stargate targets 15 GW. AI data centers consume more power than many countries.

5. Self-Improvement Accelerates



AI task horizons double every 4 months. 80%+ of code at top labs is AI-written. Models are optimizing their own training. The loop hasn't closed, but it's tightening fast.

6. Regulation Arrives



EU AI Act enforcement begins Aug 2026. California SB-833 takes effect Jul 2026. Audit trails, human oversight gates, and risk classification are now required, not optional.