HOB

Abundant intelligence, scarce judgment.

HOUSE OF BRAINS
PERSPECTIVES
VOLUME 1

DANIEL AUGUSTO MOTTA, PHD

Abundant intelligence, scarce judgment.

Where AI value is created, where it leaks away, who is paying for the race, and what it means for boards, leaders and the firms paid to advise them.

BY DANIEL AUGUSTO MOTTA, PHD

MANAGING PARTNER, HOUSE OF BRAINS

OPENING

Nine theses

AI FINANCE BATTLE

The AI trade · five years of market value

AI HIGHLIGHTS

Seventeen ideas

AI QUOTES

Nineteen voices

AI STATS

Twenty figures · seventeen charts

AI ARTICLES

Five complete articles

AI DEEP DIVE

Agent framework · I–XVII

AI CANON

Twenty-two papers · sources · the author

Models and frameworks · index 4

(01) The argument in brief · nine theses 7

Introduction 10

(02) AI Highlights 11

(03) AI Stats 15

The charts behind the numbers 20

(04) AI Deep Dive 29

Framework · Anatomy of an agent 30

I From tools we use to agents that act 31

II A plausibility engine, not an oracle — and evidence has levels 33

III The token is the unit of cost. The finished task is the unit of value. 35

IV From answers to outcomes 37

V Programming is dead. Engineering is alive. 39

VI Context is the new bottleneck, and context design is cost design 41

VII Language is the interface. The engine is somewhere else. 43

VIII Models are rented. Advantage compounds elsewhere. 47

IX Value migrates up the stack. Organizations must be rebuilt to follow it. 49

X The best customer experience is not the most frictionless. 64

XI The organization is the bottleneck 67

XII Productivity can rise while capability falls. 73

XIII Soft guardrails lower the probability. Hard guardrails cap the impact. 78

XIV Generation is cheap. Selection, distinctiveness and disclosure are not. 83

XV What this means for firms that sell judgment 85

XVI Open questions, named positions 87

XVII The biggest economic bet in US history 90

(05) AI Finance Battle 92

I The AI trade: a thesis to monitor, not a view to hold 94
II The market's verdict: five years of market value 109

(06) AI Quotes 130

(07) AI Articles · AI and the enterprise 133

I Context, creativity, governance, oversight, judgment. 134
II The heart of the agentic organization. 149
III Beyond hype: where is the economic value? 161
IV Enjoy the ride and watch out for the hook. 185
V A new leadership mindset for agentic collaboration. 206

(08) AI Canon · twenty-two papers to read 226

Sources 229
The source canon 231

(09) About the author 232

Models and frameworks

The argument is carried by models. HOB proprietary models, developed in the books A Liderança Essencial and Anthesis (2019) and in “As organizações do futuro” (2023), are framed in black throughout this edition; frameworks from external research and executive education, including MIT faculty, carry their source. Page numbers link to each appearance.

HOB PROPRIETARY MODELS AND SYNTHESES · 14

Organizational dynamics
A LIDERANÇA ESSENCIAL · DEEP DIVE IX 51 · ARTICLE I 147

From an ethic of duty to an ethic of pleasure
A LIDERANÇA ESSENCIAL · DEEP DIVE IX 53 · ARTICLE V 209

From analog–mechanical to digital–intelligent
A LIDERANÇA ESSENCIAL · DEEP DIVE IX 55 · ARTICLE V 210

Turbulence and opportunity in structural change
ANTHESIS · DEEP DIVE IX 56 · ARTICLE IV 196

Creation and destruction of value
ANTHESIS · DEEP DIVE IX 59 · ARTICLE III 172

The digital competitive arena
ANTHESIS · DEEP DIVE IX 60 · ARTICLE IV 194

Digital drivers of organizational culture
ANTHESIS · DEEP DIVE IX 62 · ARTICLE II 157

The value chain for enterprise modelling
ANTHESIS · DEEP DIVE IX 63 · ARTICLE III 175

New organizational archetypes: two systems evolving in parallel
MOTTA & PONTUSCHKA (2023) · DEEP DIVE XI 69 · ARTICLE II 152

Disruptive innovation and institutional framework
ANTHESIS · DEEP DIVE XI 71 · ARTICLE IV 203

Five human roles that do not delegate
DEEP DIVE XI 72 · ARTICLE I 136

The JOTA model applied to AI
JOTA MODEL · ARTICLE I 140

Structure, behavior and dynamics in the agentic organization
MOTTA & PONTUSCHKA (2023) · ARTICLE II 153

Three scenarios for the buildout
ARTICLE III 171

Three units for measuring AI value
DEEP DIVE III 36

The autonomy spectrum
GOOGLE CLOUD · DEEP DIVE IV 38

ALIGN: a development lifecycle built on agreement about intent
TIM KRASKA, MIT CSAIL · DEEP DIVE V 40

Five levels of shaping what an agent does
SERTAC KARAMAN, MIT · DEEP DIVE VI 42 · ARTICLE V 212

Supply chain business intelligence and analytics
DAVID SIMCHI-LEVI, MIT · DEEP DIVE VII 45

The language model as interpreter
SIMCHI-LEVI ET AL., MICROSOFT SUPPLY CHAIN · DEEP DIVE VII 46 · ARTICLE I 143

Five layers of compounding advantage
GOOGLE CLOUD · DEEP DIVE VIII 48 · ARTICLE III 180

The friction audit
RENÉE RICHARDSON GOSLINE, MIT · DEEP DIVE X 65

AI gravity
ERIC C. SO, THE COLLISION · DEEP DIVE XII 74

Better homework, worse learning
STRÖMBERG, LEI AND WU; BASTANI ET AL. · DEEP DIVE XII 75 · ARTICLE III 179

The downward spiral, and where to break it
ERIC C. SO, THE COLLISION · DEEP DIVE XII 76

Where AI use is risky
ERIC C. SO, THE COLLISION · DEEP DIVE XII 77

The democratization of attack
ETAY MAOR, CATO NETWORKS · DEEP DIVE XIII 79 · ARTICLE V 217

LLM attack surface and risks
ETAY MAOR, CATO NETWORKS · DEEP DIVE XIII 80

Soft and hard guardrails
ETAY MAOR, CATO NETWORKS · DEEP DIVE XIII 82 · ARTICLE V 219

Four ways to run an innovation process
STEVEN EPPINGER, MIT · DEEP DIVE XIV 84

The post-quantum calendar is already written
US EO 14412 · NIST · UK NCSC · DEEP DIVE XVI 89

Mosca's inequality: when x + y exceeds z, an organization is already late
MICHELE MOSCA · DEEP DIVE XVI 89

FRAMEWORKS FROM EXTERNAL SOURCES · 29

Anatomy of an agent
GOOGLE CLOUD AND SERTAC KARAMAN, REDRAWN · DEEP DIVE, FRAMEWORK 30

Four eras of computing
SERTAC KARAMAN, MIT · DEEP DIVE I 32

Evidence hierarchy
ADAPTED FROM REGINA BARZILAY, MIT · DEEP DIVE II 34

America's infrastructure booms, average annual spending as % of GDP
WSJ AND BROOKINGS DATA · DEEP DIVE XVII 91 · ARTICLE III 168

Power in the AI hardware chain sits at three chokepoints
SILICON ANALYSTS · AI FINANCE BATTLE, PART I 97 · ARTICLE IV 189

The labs' largest investors are also their largest suppliers
BLOOMBERG · AI FINANCE BATTLE, PART I 105 · ARTICLE IV 191

Place people where they add most
SERTAC KARAMAN, MIT · DEEP DIVE XI 71 · ARTICLE II 155

Agents absorb the tip of the iceberg
MICHAEL POLANYI, THE TACIT DIMENSION · ARTICLE II 158

The legs of the parlay: every one has to win
GARY GENSLER, VIA MIT TECHNOLOGY REVIEW · ARTICLE III 164

The downward spiral of reliance
ERIC C. SO, THE COLLISION · ARTICLE IV 195

Mosca's inequality: are you already late?
MICHELE MOSCA · ARTICLE IV 200

Autonomy is granted in steps
GOOGLE CLOUD · ARTICLE V 212

Attack surface of a language-model application
ETAY MAOR, CATO NETWORKS · ARTICLE V 220

Where AI should and should not do the work
ERIC C. SO, THE COLLISION · ARTICLE V 223

The argument in brief

01

Nine theses.

01

THESIS 01 · JUDGMENT

Intelligence is becoming abundant. Judgment is not.

AI value is created per unit of judgment successfully delegated, and captured per unit of verification eliminated.

02

THESIS 02 · ECONOMICS

The token is the unit of cost. The finished task is the unit of value.

Price per token fell about 280-fold since 2023 while average enterprise AI spend rose 320%. Cheaper units, bigger bills: cost belongs to the whole configuration, not the model.

03

THESIS 03 · ADVANTAGE

Models are rented. Context, data and the system around them are owned.

General models beat specialists. Software becomes knowledge engineering, and advantage accumulates in data, tools, skills and decision records that competitors cannot license.

04

THESIS 04 · AGENTS

Agents move AI from answers to outcomes, and from content risk to action risk.

An error in a chatbot is a bad paragraph; an error in an agent is a wrong payment. Context, not the model, becomes the bottleneck — and review capacity, not agent capacity, the constraint.

05

THESIS 05 · ORGANIZATION

The organization, not the model, is the bottleneck.

The gap between a capable model and captured value is mostly organizational, and accountability never transfers to the machine.

06

THESIS 06 · TRUST

Trust is engineered, not requested.

Soft guardrails lower the probability of harm; only hard limits outside the model cap its impact. Friction, fairness and security are design choices leaders own.

07

THESIS 07 · CAPITAL

Capital runs ahead of value, and the risk has moved to financing.

Global AI investment passes $1 trillion in 2026. Revenue is catching up; debt, short asset lives and circular deals are the new exposure. Hold an AI thesis as a position to monitor.

08

THESIS 08 · MARKETS

The market has already voted for scarcity.

Chips and chip equipment went from 17% to 43% of the value of the global tech top 50 in five years, while businesses priced on seats and hours were marked down.

09

THESIS 09 · CAPABILITY

Productivity can rise while capability falls.

Biology, competition and real gains pull people toward reliance on AI. Firms that measure only output will look more productive while the judgment that directs AI erodes. Redesign the apprenticeship; do not abandon it.

INTRODUCTION

House of Brains perspectives on AI and the enterprise. From foundations and generative AI to agents, decisions, strategic mindset, customers, trust, innovation and capital, these pages read the latest research, executive education and market data through our own strategic frameworks, and distil the ideas, numbers and voices that boards, leaders and advisors need in front of them.

Every figure carries its source. The argument runs from what changed, through mechanism, economics, evidence, agents and advantage, to the organization, the one bottleneck no model removes, and follows the money through the AI Finance Battle.

AI Highlights

02

Seventeen ideas, each reduced to what is worth carrying into the next board meeting, and each opening a chapter of the AI Deep Dive.

I

WHAT CHANGED

From tools we use to agents that act

Speed, agency and abundance set this revolution apart: for the first time, thinking itself can be mass-produced. Public text is nearly exhausted; the next data sits inside companies.

Progress looks gradual, then sudden — and no one knows in advance when an application unlocks.

II

MECHANISM · EVIDENCE

A plausibility engine, not an oracle

Models complete patterns; they don’t look things up, so hallucination can be reduced but never eliminated. Evidence has four levels, and capital should scale with the level reached.

No held-out test set, no evidence it works.

III

THE ECONOMICS

Cheaper tokens, bigger bills

The Jevons pattern: agentic and reasoning workloads use 5–30× the tokens of a chat. Four levers multiply: model choice, agent harness, consumption model and cost observability.

Measure cost per outcome falling, not the AI bill shrinking.

IV

AGENTS

Autonomy is granted in steps

Inform, recommend, act with approval, act within a mandate, anticipate. The model reasons; the harness — goals, memory, tools, guardrails, human gate — decides reliability and cost.

Every hand-off between agents is a trust boundary.

V

AI-DRIVEN DEVELOPMENT

Programming is dead. Engineering is alive.

An agent is “simply a well-written prompt.” Most projects still fail on organizational obstacles, and effort shifts from 95% development toward 95% stating intent.

Software engineering becomes knowledge engineering.

VI

INSIDE THE AGENT

Context is the new bottleneck

Every call re-sends the whole context window, and you pay for it again. What an agent loads — and in what order — decides both its quality and its cost.

Context design is cost design: price per task, not per token.

VII

DECISIONS

Language is the interface

Most enterprise AI work is deciding, not writing. Keep the arithmetic in optimization engines and the policy in code; let language models translate the question and explain the answer.

The numbers come from the engine, never from the language model.

VIII

ADVANTAGE

Models are rented. Advantage compounds.

Five layers of compounding advantage, from data to transformation. The context graph — every request, decision and rejected alternative — belongs to the firm alone.

Learning loops can't be bought retroactively: time becomes the moat.

IX

STRATEGIC MINDSET

Value migrates up the stack

Ambition must fit organizational potency. Two transitions, ten megatrends and the migration of value up the layers decide where AI value moves — and entropy decides how much of it is kept.

Whose agent will recommend us, and why?

X

CUSTOMERS

The best experience is not frictionless

64% of customers would rather companies not use AI for service. Remove wasteful friction; add beneficial friction where decisions are consequential, and make leaving as easy as joining.

Don’t focus on the tool, focus on the customer.

XI

ORGANIZATION

The organization is the bottleneck

AI accelerates the processual side; the tacit side decides whether any of it is used. Five human roles don't delegate: context, creativity, governance, oversight, judgment.

AI cannot be accountable because AI has no stake.

XII

CAPABILITY

Productivity up, capability down

AI gravity pulls everyone toward reliance. Students’ homework improves while their exams fall; experts keep their gains, novices never build them.

Measure the quality of thinking, not only the speed of output.

XIII

TRUST

Hard guardrails cap the impact

Soft guardrails in the prompt lower the probability of harm; permissions, sandboxes and human approval outside the model cap its impact. Fairness, too, is a design choice someone must own.

Humans + AI = the strongest link.

XIV

INNOVATION

Generation is cheap. Selection is not.

Only two of six headline productivity gains rest on rigorous studies. Many employees hide their AI use, and AI ideas score higher but vary less than human ones.

Grade the evidence; bring the secret cyborgs into the open.

XV

ADVISORY FIRMS

Accountability is the moat

Packaging loses value; thinking gains it. The pyramid becomes orchestration, intent becomes the product, selection becomes the craft, and trust a deliverable.

Models cannot be held accountable. Firms can.

XVI

THE HORIZON

Open questions, named positions

Rival powers frame the race in opposite terms, and builders disagree on pace. AI is neither a being nor a deity, but a new kind of intellect reshaping human understanding.

Plan with scenarios, and keep explanation and responsibility with people.

XVII

CAPITAL

The biggest bet in US history

$10.3T of AI infrastructure from 2025 to 2032 — 3.6% of GDP a year, above the railroads. Capital, balance sheets, energy and talent are all scarce at once.

Focus on reorganization, not infrastructure. Follow the money in the AI Finance Battle.

AI Stats

03

The scale of the shift in twenty figures, each with its source.

01

$1T+

Global AI investment forecast for 2026, $581B of it in the US (Goldman Sachs); the Big Four alone guide $720–760B, up from ≈$410B in 2025.

Source: Goldman Sachs Research, Aug 2026

02

280×

Fall in price per token since 2023 (Google Cloud enterprise presentation, 2026).

Source: Google Cloud, 2026

03

320%

Rise in average enterprise AI spend over the same period, to roughly $7M per company (Google Cloud).

Source: Google Cloud, 2026

04

64.7%

Top score on Humanity's Last Exam in June 2026, up from 3.3% in May 2024 (artificialanalysis.ai).

Source: Artificial Analysis, Jun 2026

05

$2.67T

Worldwide AI spending in 2026, up 49.5%, per Gartner — rising to $3.64T in 2027.

Source: Gartner, Sep 2026

06

39%

Share of US real GDP growth from AI-related investment, first nine months of 2025 — above the dot-com peak.

Source: Federal Reserve Bank of St. Louis, Jan 2026

07

945 TWh

Global data-center electricity use projected for 2030 by the IEA, more than double 2024's 415 TWh.

Source: IEA, Energy and AI, 2025

08

27B

Tokens a day at AT&T after deploying multi-agent systems, up from about 8 billion.

Source: FinOps Foundation, May 2026

09

–74%

Input cost of a 20-step agent task when its stable prefix is cached, at published Opus 5.5 rates: $4.80 down to $1.24.

Source: Anthropic pricing; House of Brains calculation

10

<10%

Share of enterprises that have scaled AI agents, though nearly two-thirds have experimented with them (McKinsey, 2026).

Source: McKinsey, 2026

11

40%+

Agentic AI projects Gartner expects to be scrapped by the end of 2027.

Source: Gartner, Jun 2025

12

97M

Monthly SDK downloads of the Model Context Protocol by December 2025, across about 10,000 active servers.

Source: Model Context Protocol, Dec 2025

13

3.6%

Of US GDP a year projected for AI infrastructure, 2025–2032 ($10.3 trillion), above the railroads' 2.24%.

Source: WSJ, using Brookings estimates, Sep 2026

14

−$278B

OpenAI's projected negative free cash flow from 2026 to 2030, as revenue grows from $36B to $350B a year.

Source: Financial Times via Bloomberg, Sep 2026

15

80%+

Of firms reporting no measurable productivity effect from AI so far, in an NBER survey of nearly 6,000 executives.

Source: NBER survey of executives, 2026

16

7×

How much more likely AI-generated product ideas were to reach the top 10% — though human ideas were more diverse.

Source: Terwiesch et al., Production and Operations Management, 2026

17

37%

Of organizations report any EBIT impact from AI, unchanged in a year, though nearly nine in ten use it (McKinsey, State of AI 2026).

Source: McKinsey, The State of AI 2026

18

$37B

Enterprise spending on generative AI in 2025, 3.2 times the year before; 76% of solutions bought rather than built (Menlo Ventures).

Source: Menlo Ventures, 2025

19

64%

Of customers would prefer companies not to use AI for customer service (Gartner); only 46% of people globally are willing to trust AI.

Source: Gartner, 2024; KPMG and University of Melbourne, 2025

20

62%

Of breaches involve a person, per Verizon's 2026 Data Breach Investigations Report; third parties appear in 48%.

Source: Verizon, 2026 Data Breach Investigations Report

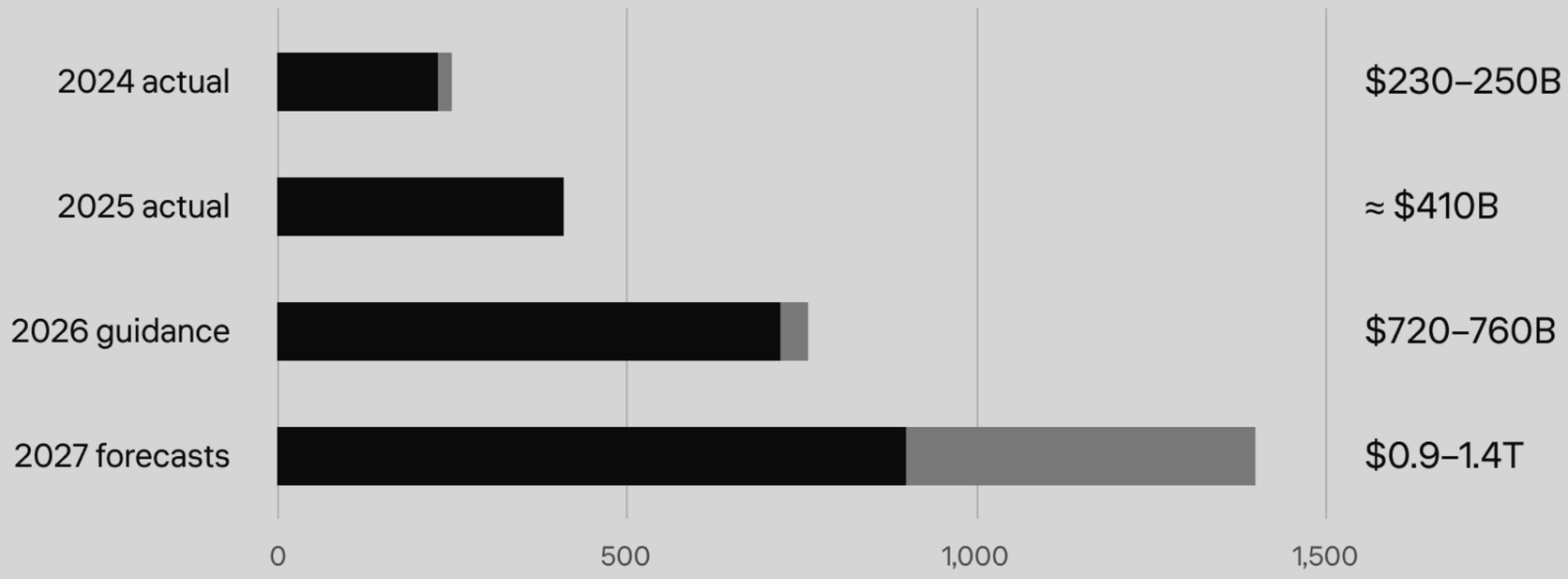
The charts behind the numbers

Seventeen charts on capability, innovation, cost, capital, talent, the economy and energy.

CAPITAL

Big Four hyperscaler capex, US$ billions

Amazon, Microsoft, Alphabet and Meta combined. 2024–25 actuals, 2026 guidance after Q2 earnings, 2027 analyst forecasts. Light extension shows the range.

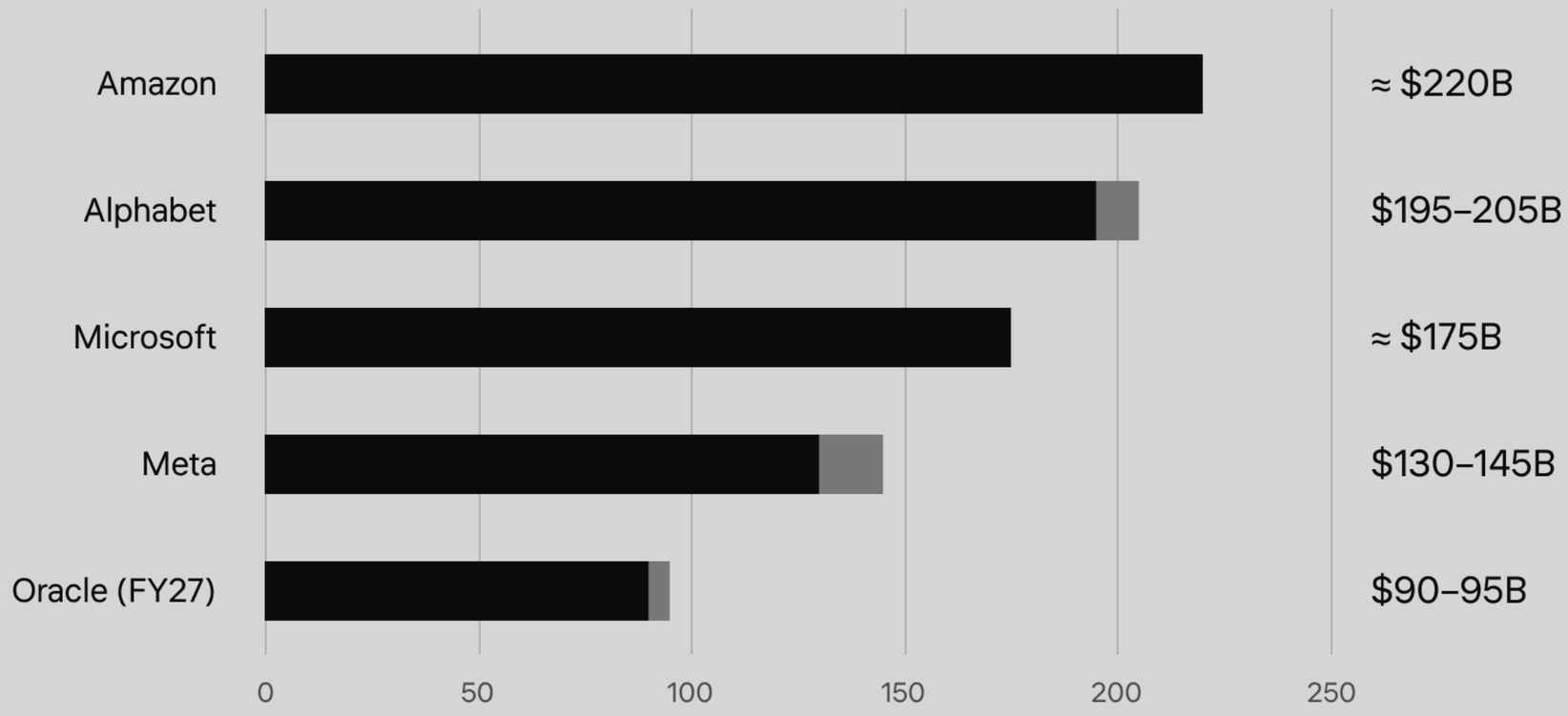


Source: Company filings · Bloomberg · Morgan Stanley · Goldman Sachs

CAPITAL

2026 capex guidance by company, US$ billions

Latest guidance, July–September 2026. Microsoft's figure reflects a lease reclassification, not a spending cut; Oracle is fiscal year to May 2027.

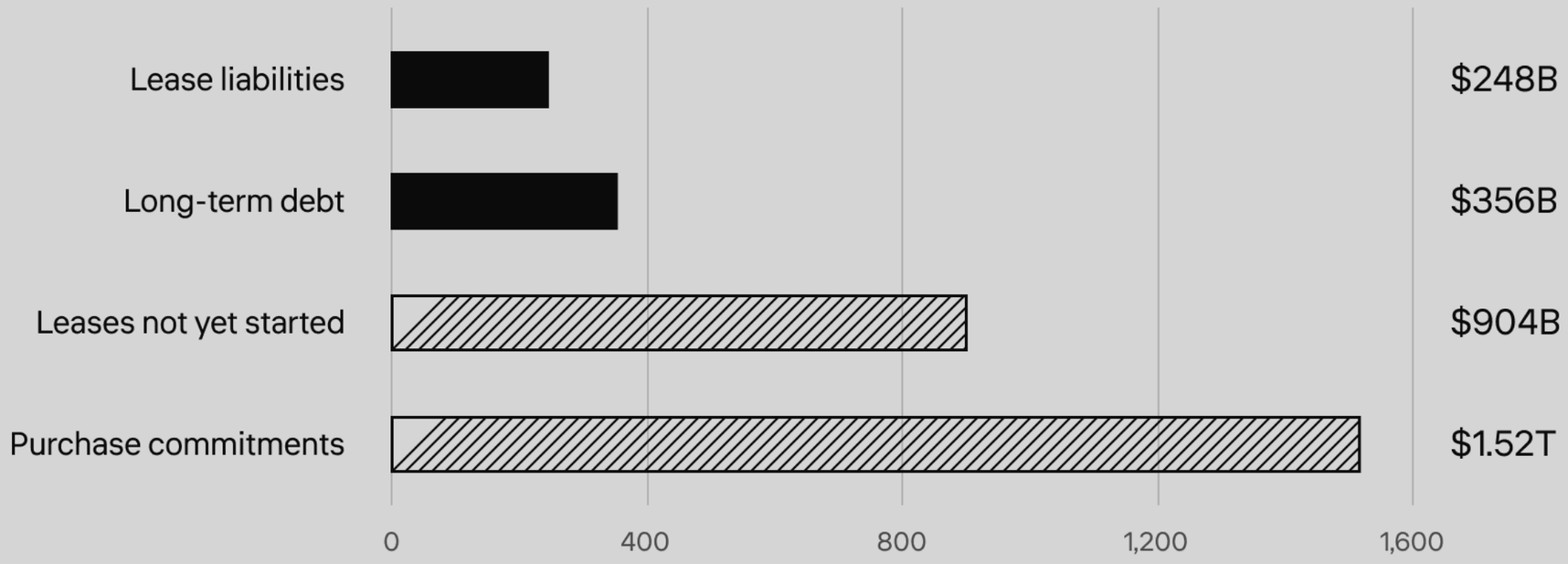


Source: Company guidance, Q2 2026

CAPITAL

What the four largest US tech firms owe, US$ billions

On the balance sheet: lease liabilities and long-term debt. Off it (hatched): leases signed but not yet started, and purchase commitments.

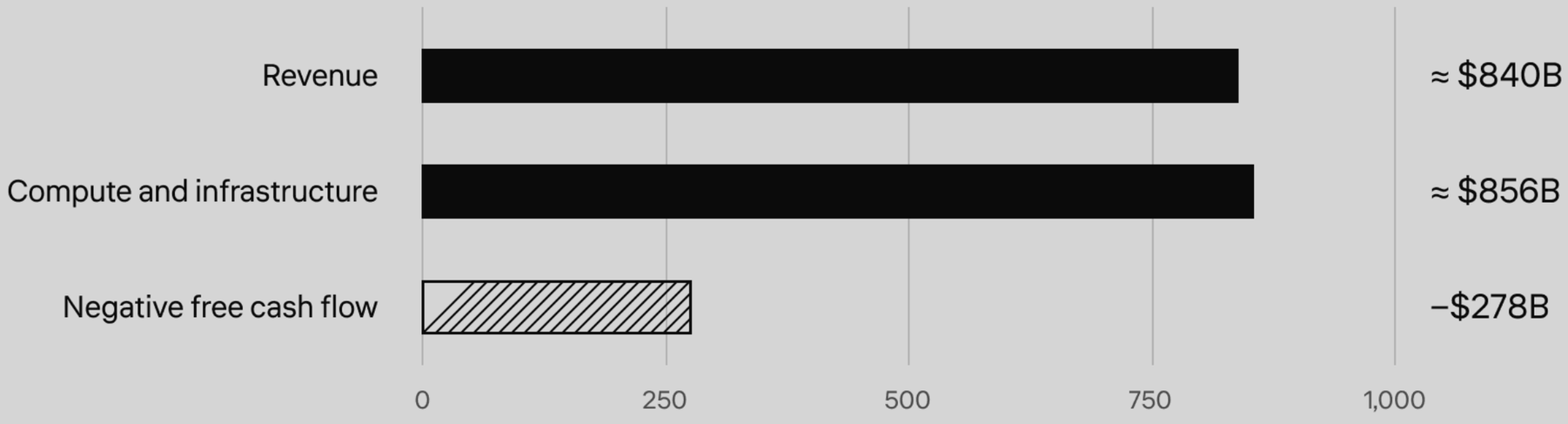


Source: WSJ, “Why Big Tech’s AI Spending Is $3 Trillion Higher Than It Seems” (Aug 2026)

CAPITAL

OpenAI's projection, 2026–2030 cumulative, US$ billions

Revenue grows from about $36B to $350B a year; compute and infrastructure spending exceeds it. Hatched: the cash burn funded by capital raises.

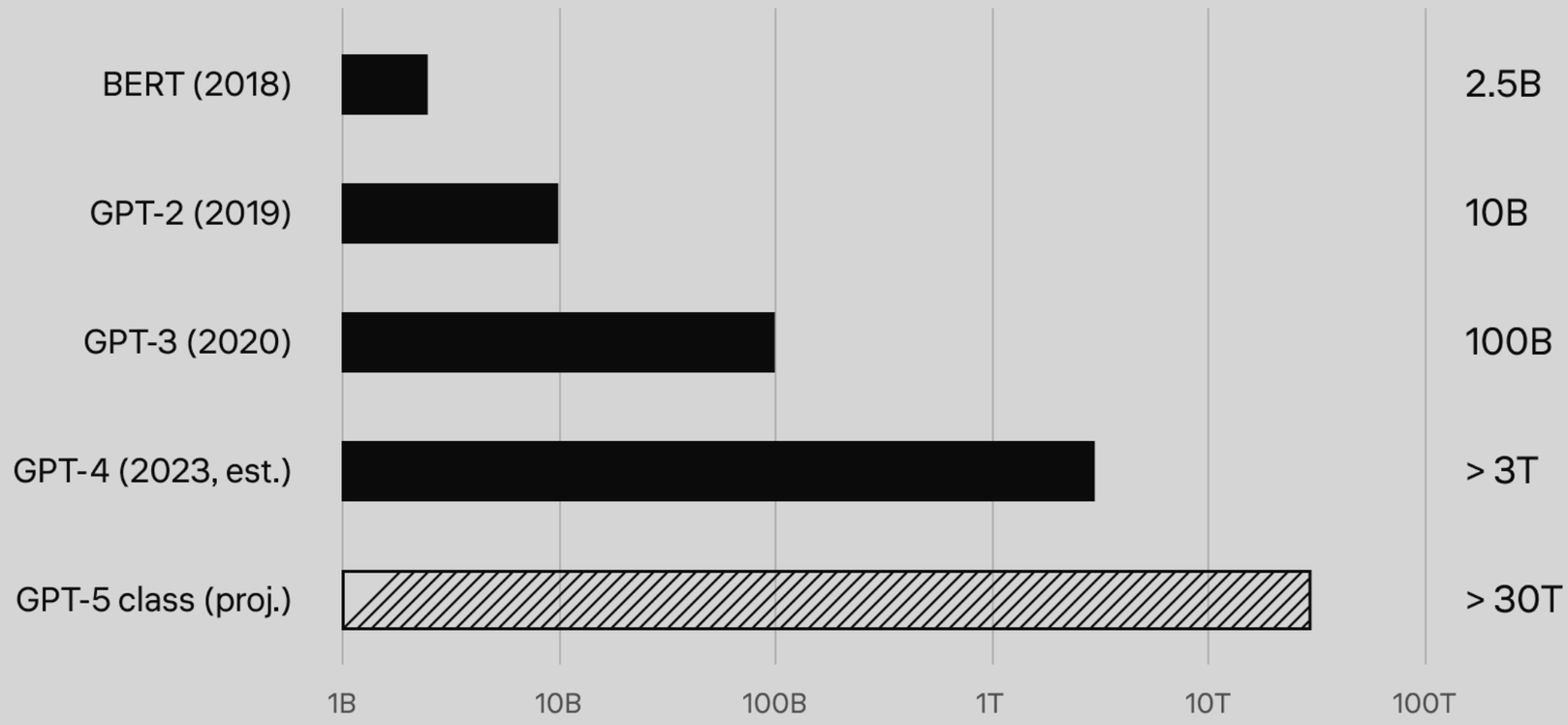


Source: FT via Bloomberg, 18 Sep 2026

CAPABILITY

Training data grew tenfold roughly every 12–15 months

Tokens used to train landmark language models, log scale. GPT-4 is an outside estimate; the hatched GPT-5-class bar is Karaman's projection.

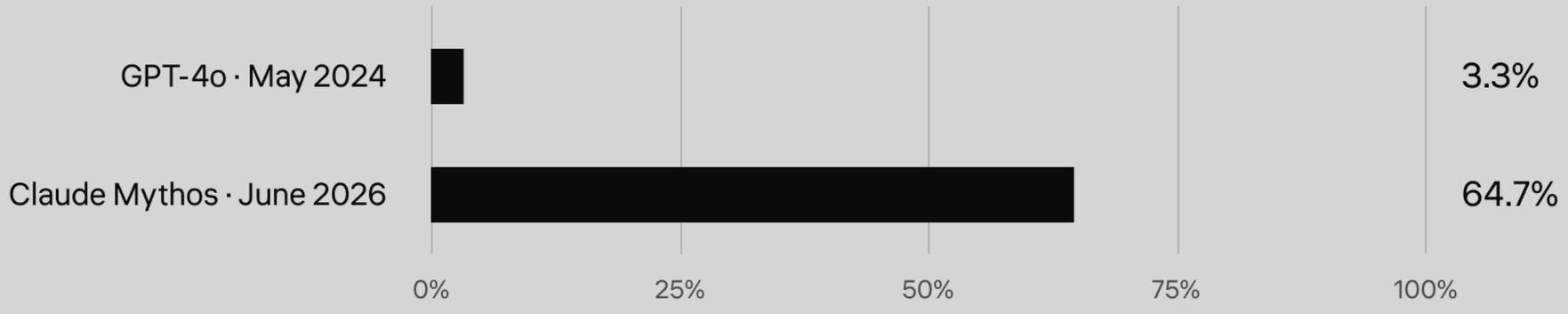


Source: Sertac Karaman, “Data & Compute”, MIT (2026), citing LifeArchitect.ai

CAPABILITY

Humanity's Last Exam: best available model

Share of expert-level questions answered correctly. Benchmarks are an imperfect proxy for workplace capability.

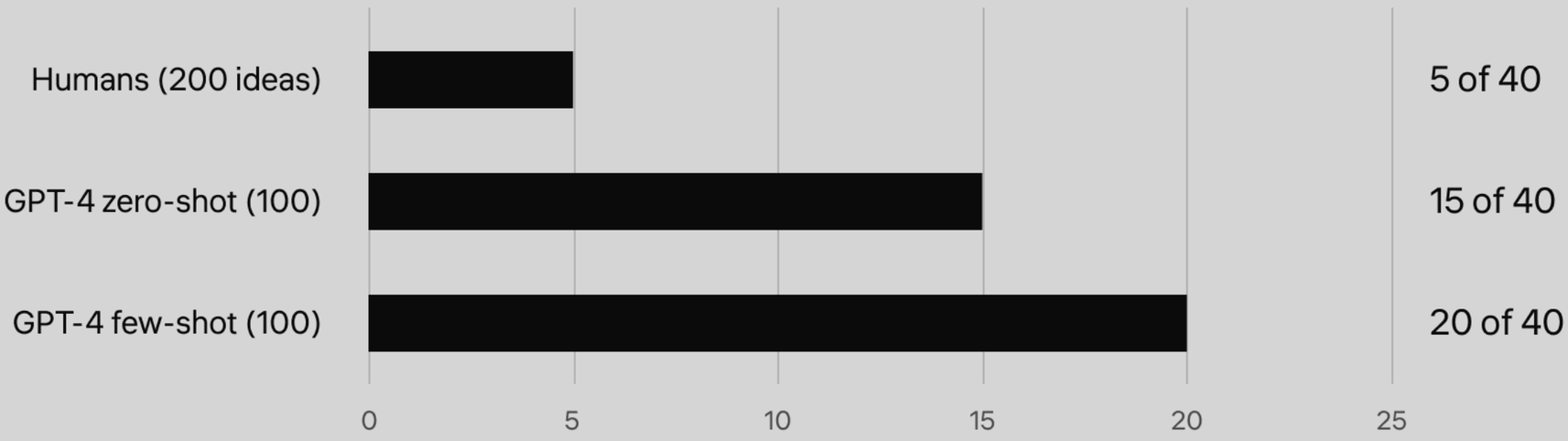


Source: Eric So, “AI at Work”, MIT Sloan (2026), citing artificialanalysis.ai

INNOVATION

Who produced the top 10% of ideas

The 40 best of 400 ideas by purchase intent: 200 from MBA students, 100 each from GPT-4 zero-shot and few-shot.

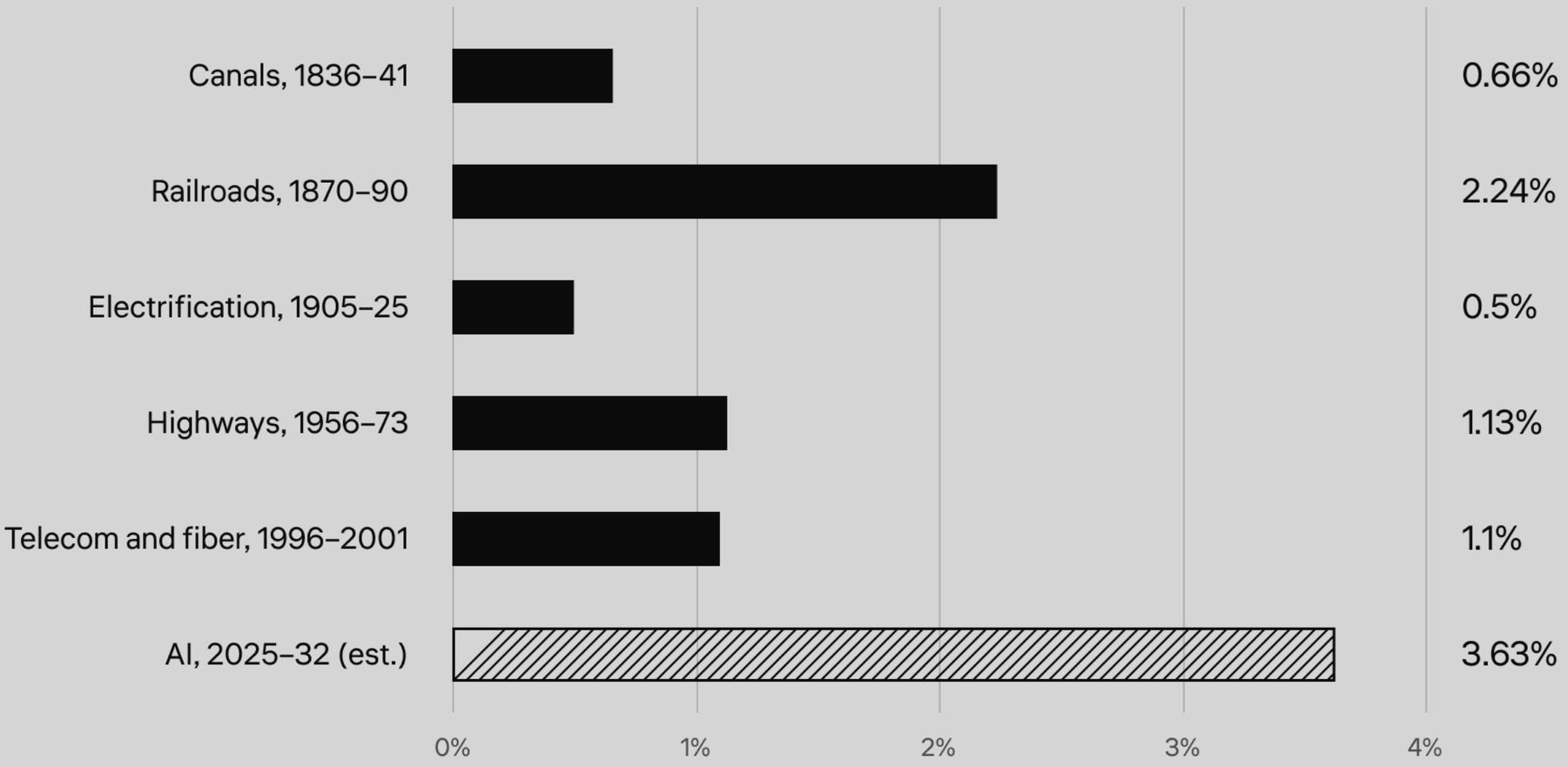


Source: Terwiesch et al., Production and Operations Management (2026)

CAPITAL

America's infrastructure booms, % of GDP a year

Average annual spending relative to GDP. The AI figure is a projection (hatched); the Louisiana Purchase cost about 3%, once.

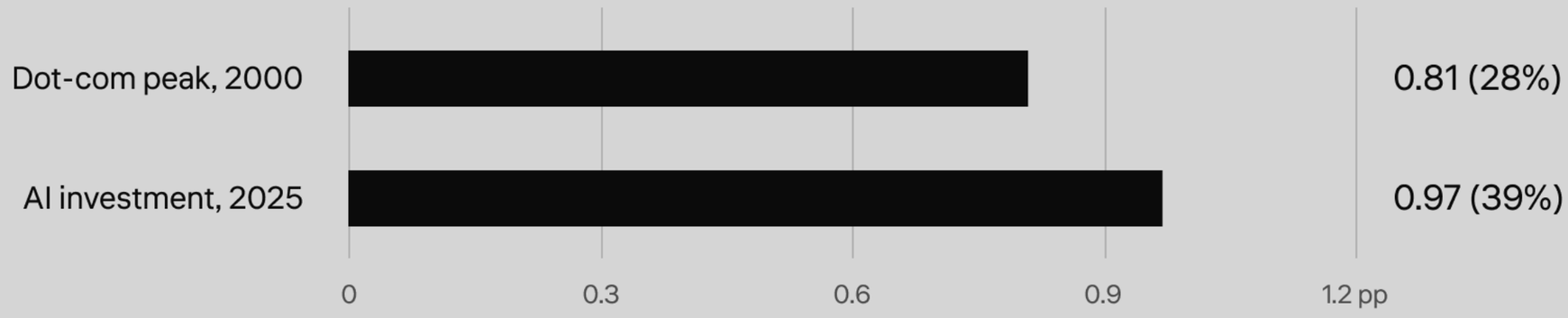


Source: WSJ (Putzier and Lahart, Sep 2026) using Brookings / Van Nieuwerburgh estimates

ECONOMY

Tech investment's contribution to US real GDP growth, pp

Dot-com peak versus AI-related investment in 2025 (first nine months, annualized). Share of total growth in brackets; gross, before netting imports.

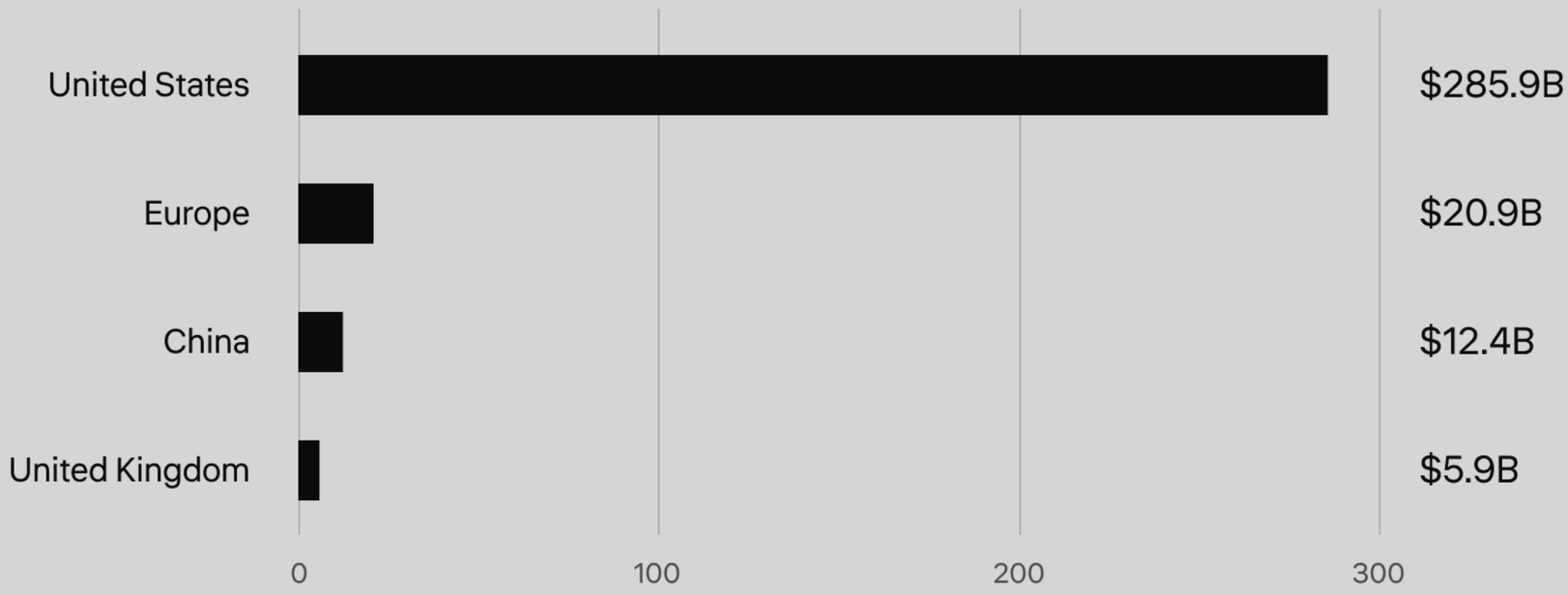


Source: Federal Reserve Bank of St. Louis, January 2026

CAPITAL

Private AI investment in 2025, US$ billions

Global total $344.7B. China's figure excludes state guidance funds, estimated at about $184B directed to AI firms since 2000.

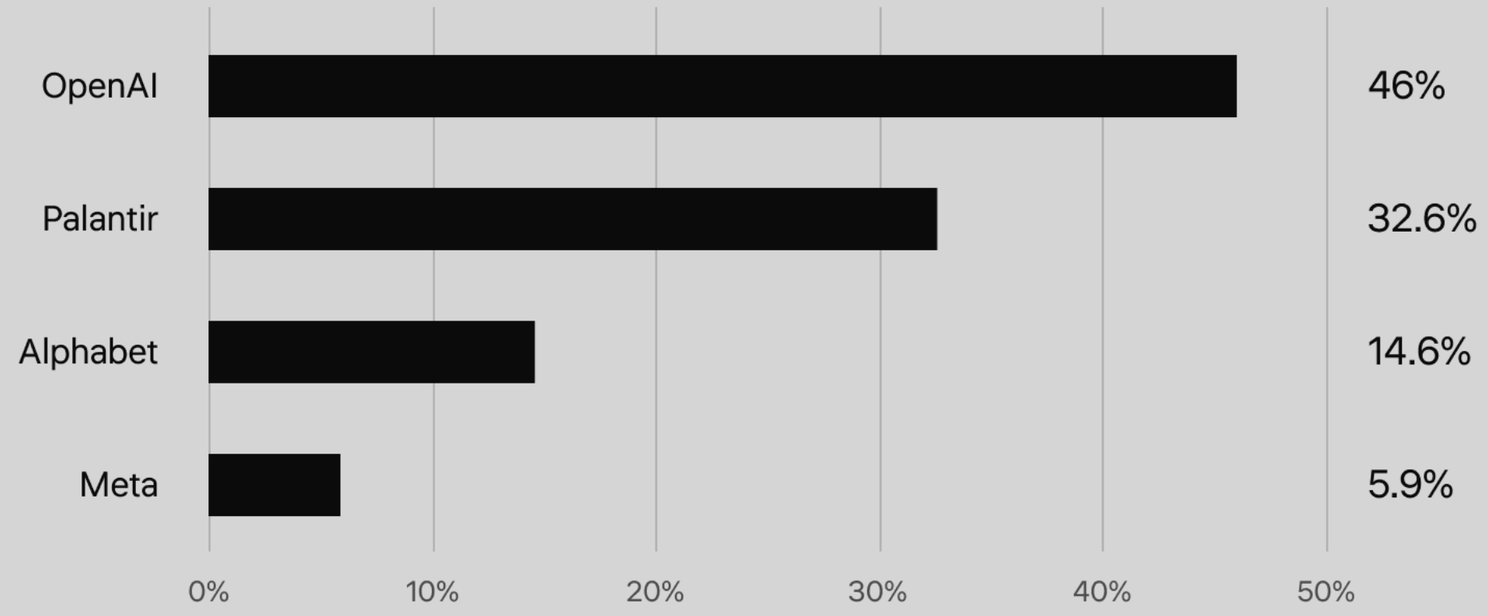


Source: Stanford AI Index 2026 (Quid data)

TALENT

Stock-based compensation as a share of revenue

As cited by Eric So (MIT Sloan, 2026); OpenAI's averaged about $1.5 million per employee in 2025.

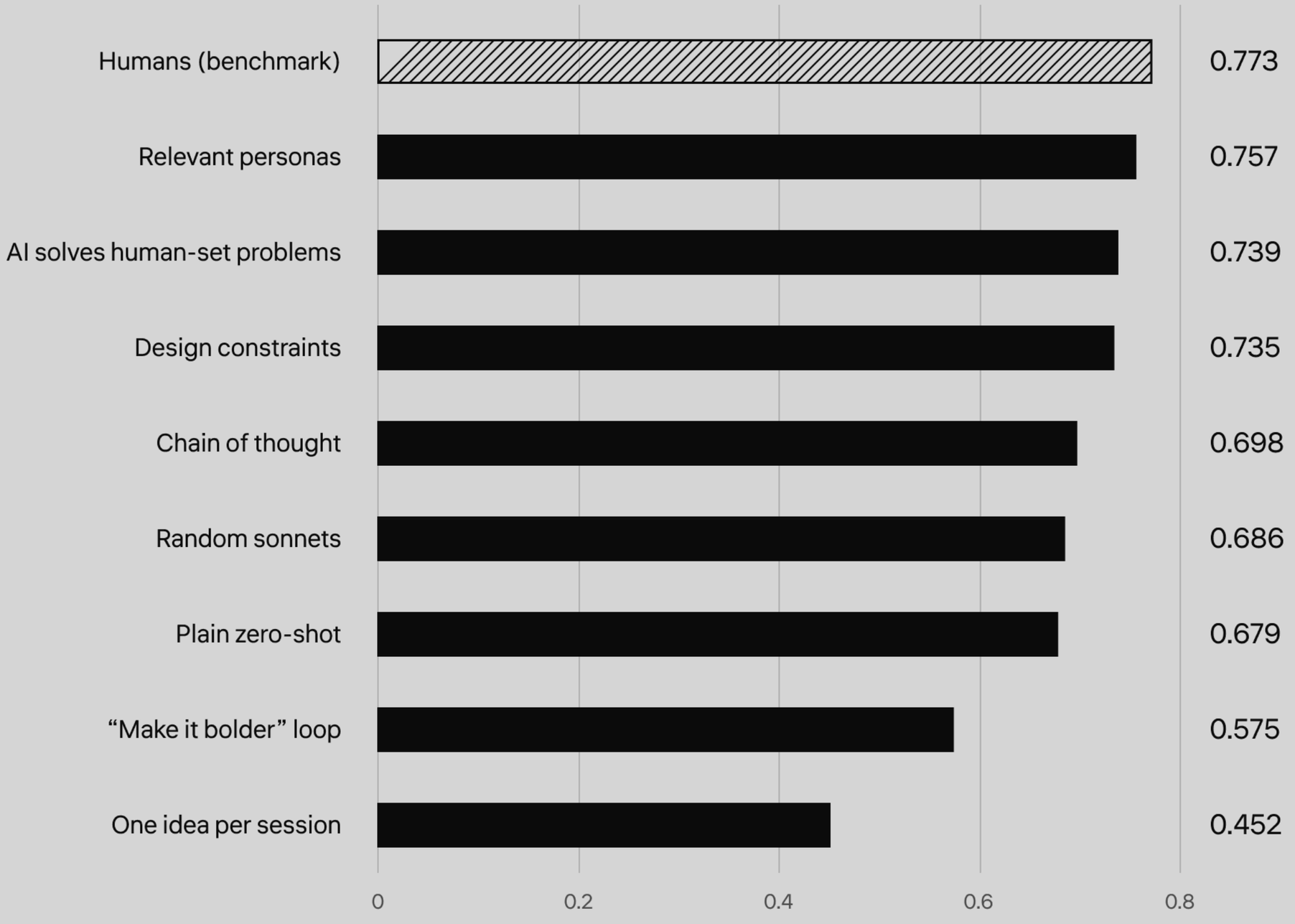


Source: WSJ via Fortune, cited by Eric So, MIT Sloan (2026)

INNOVATION

How to get more varied ideas from AI

Diversity score of 100 ideas under each strategy; humans score 0.773 (hatched). Relevant variety helps; “make it bolder” loops and one idea per session hurt.

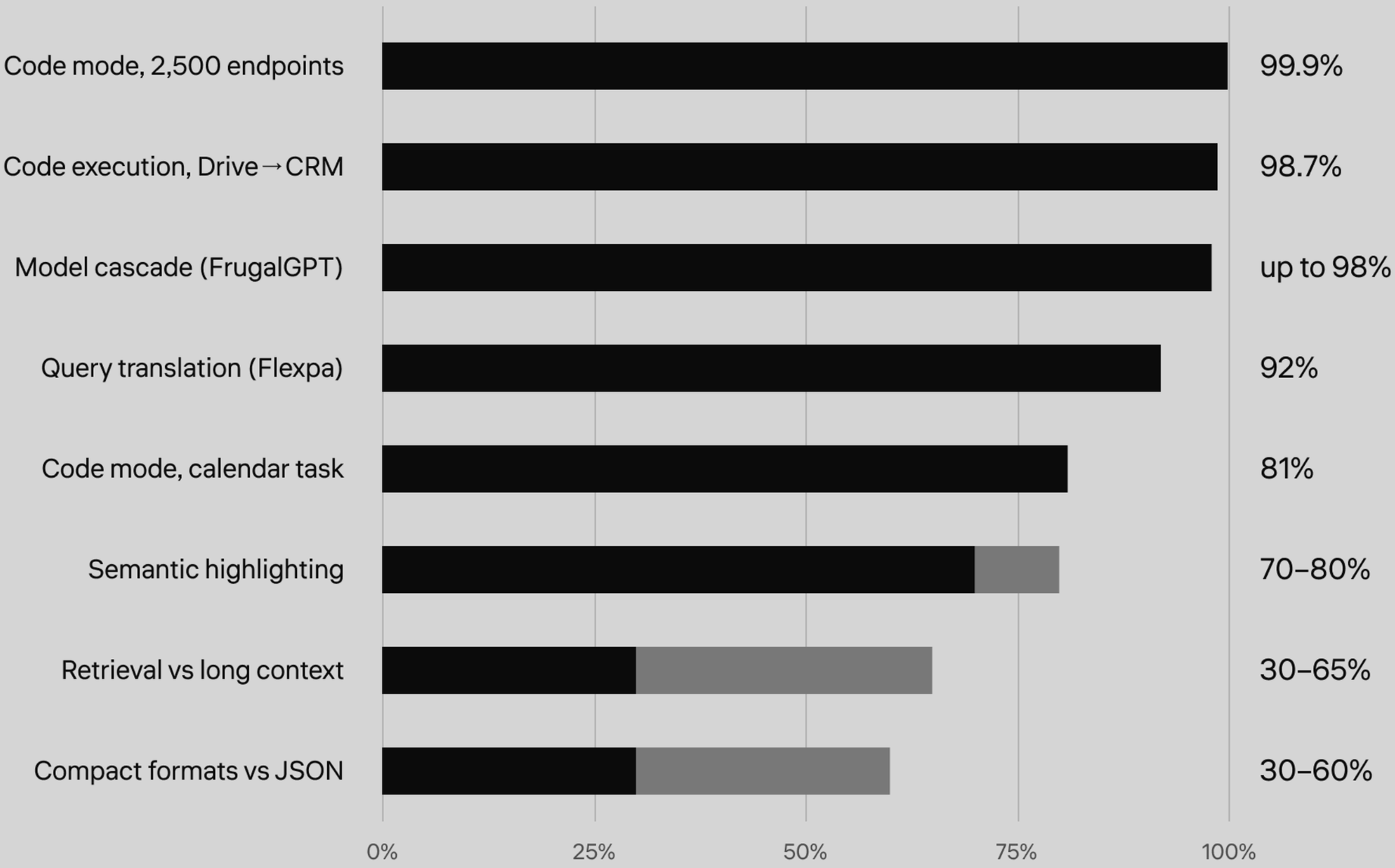


Source: Terwiesch et al. (2026), Study 3b, GPT-4o

COST

Token or cost reductions from engineering levers

Percentage reduction versus a baseline, as reported by each source. Vendor results on specific workloads — not transferable as averages.

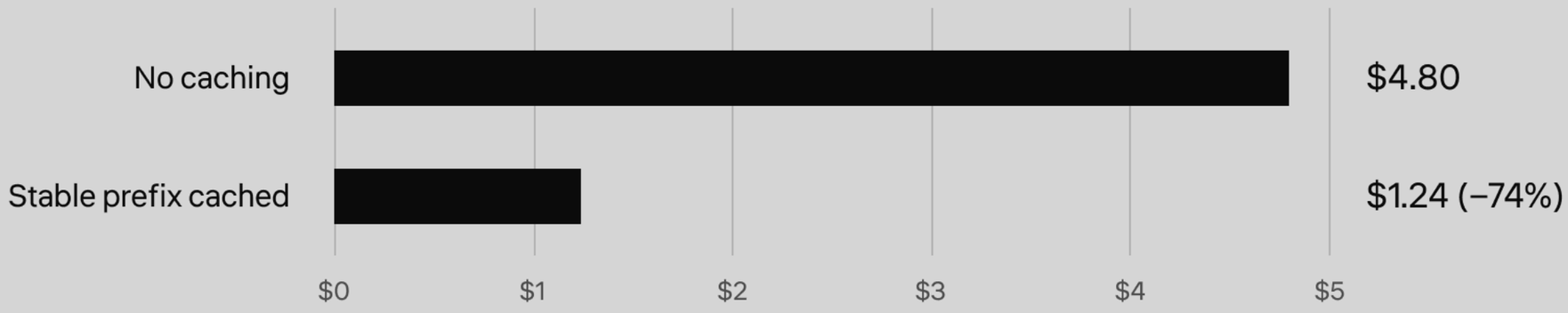


Source: FinOps Foundation compilation · Cloudflare · Anthropic · Zilliz · Flexpa · FrugalGPT

COST

Input cost of one 20-step agent task, US$

Illustrative calculation at published Opus 5.5 rates ($4 per million input tokens; cache write $5, cache hit $0.20), with a 50,000-token stable prefix. Output tokens excluded.

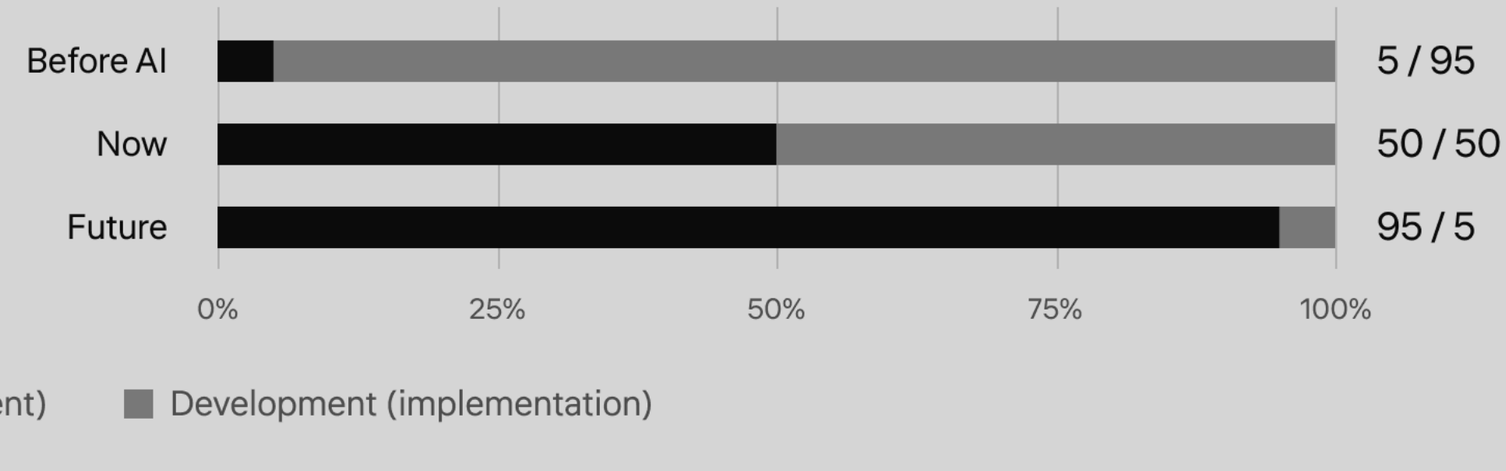


Source: Anthropic pricing documentation, checked 24 Sep 2026

WORK

Share of effort: product management vs. development

Kraska's illustrative estimate of how software work shifts as AI writes the code. Not measured data.

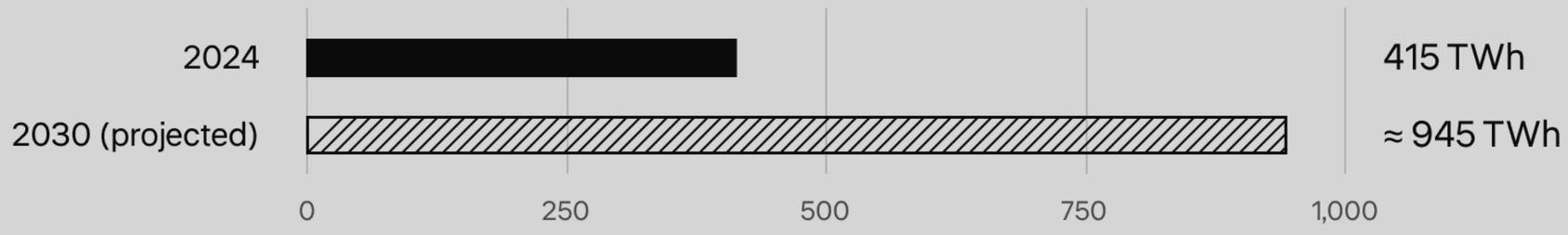


Source: Tim Kraska, “The New Software Development Lifecycle for AI”, MIT CSAIL (2026)

ENERGY

Global data-center electricity demand, TWh

More than double in six years, to close to 3% of world demand. Hatched bar is a projection.

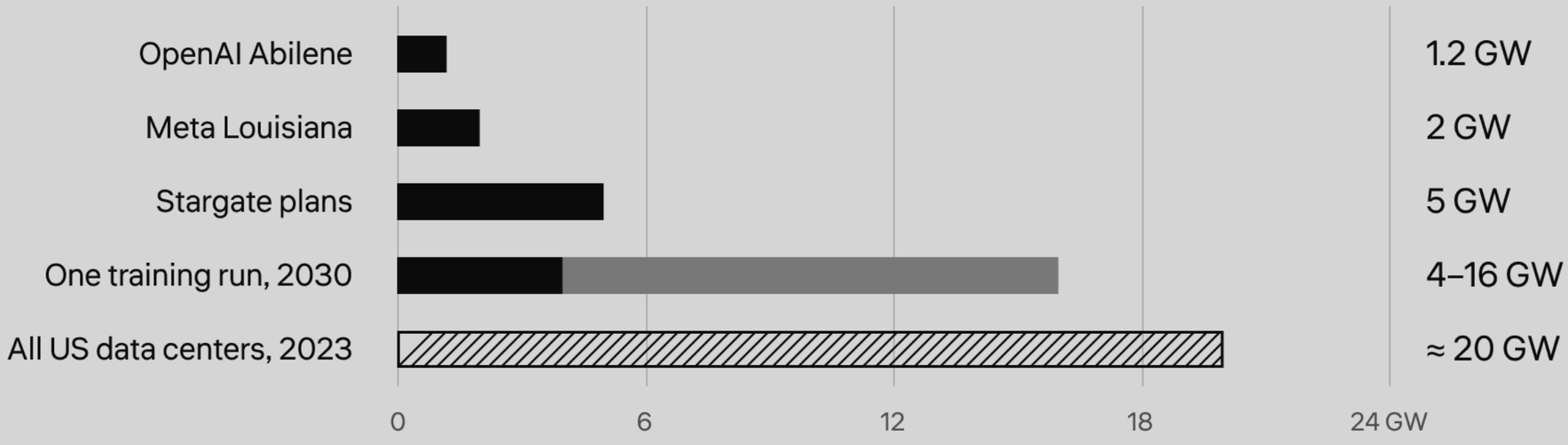


Source: IEA, Energy and AI (April 2025), base case

ENERGY

Power: projects and projections, gigawatts

Named projects and Epoch AI's 2030 range for a single frontier training run, against all US data-center use in 2023 (hatched).



Source: Epoch AI and EPRI, cited by Eric So, MIT Sloan (2026)

AI Deep Dive

04

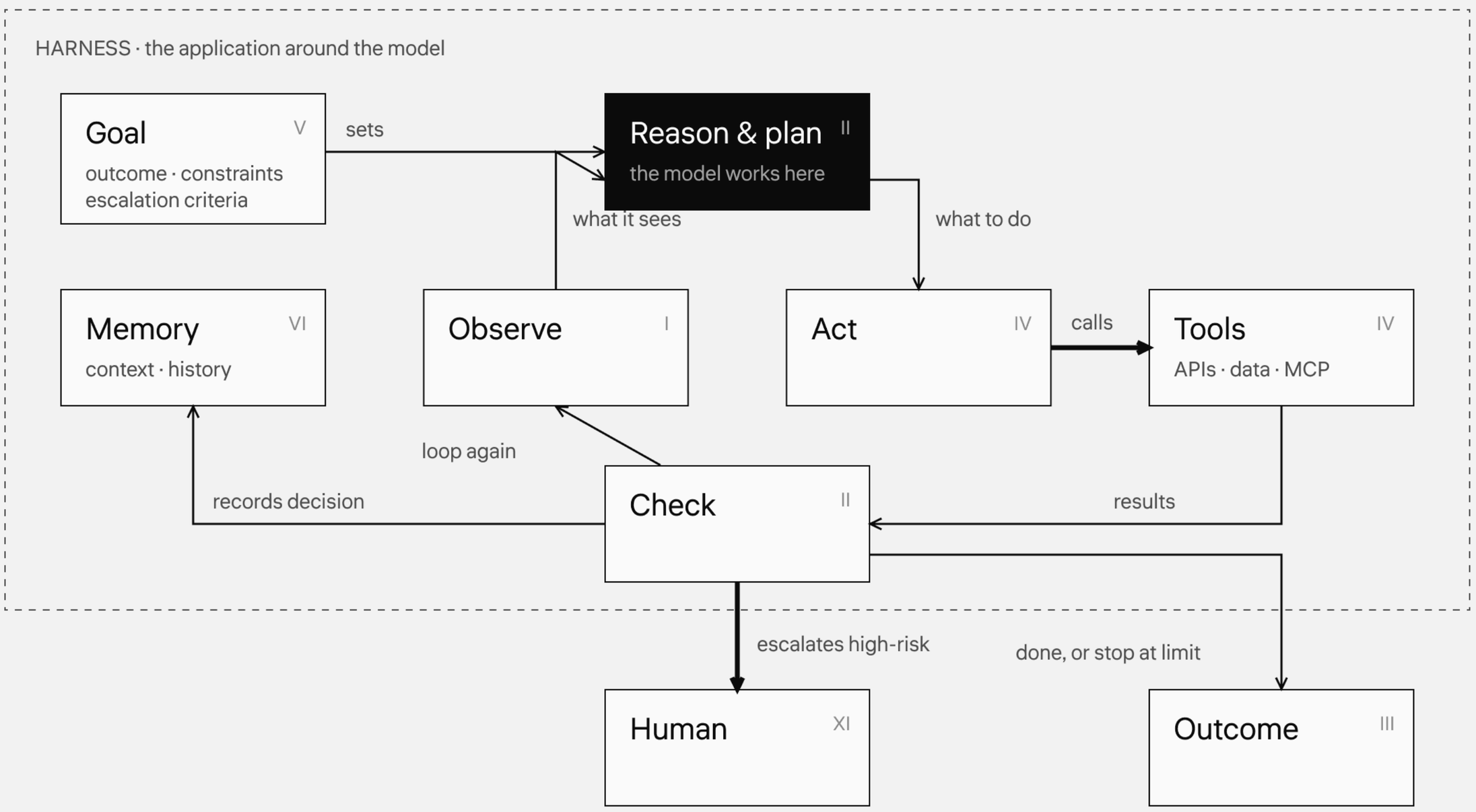
The full argument behind each AI Highlight, chapter by chapter — seventeen chapters. It starts with the framework that ties the argument together, the anatomy of an agent, where every part of the loop opens the chapter that goes deeper.

FRAMEWORK · ANATOMY OF AN AGENT

FRAMEWORK

GOOGLE CLOUD AND SERTAC KARAMAN, REDRAWN

The model reasons. The harness does everything else.



“An AI agent is an application that achieves a goal by observing the world and acting upon it using the tools it has at its disposal.” The model is only the reasoning step inside a loop: it never touches a file or a system, it writes instructions that the harness executes. Goal, memory, tools, guardrails, the check step and the human gate make up the harness, and that's where reliability and cost are decided.

Application	Specified, tested, versioned and monitored like any production system.
Goal	“Resolve billing disputes under R$500, within policy, in under ten minutes” — not “help customers”.
Observing	On bad data, an agent acts confidently on a false picture.
Acting	Where agents create value — and where risk enters.
Tools	Capability is bounded by how much of the business is exposed as governed, callable tools.

— Where risk enters and is contained

V Chapter that goes deeper

Anatomy of an agent, a House of Brains synthesis after Google Cloud's “Observe, Act, Achieve” agent-harness framework and Sertac Karaman's “Anatomy of an Agentic AI Model” (MIT, 2026), with lessons from documented agent failures. Heavy lines mark where risk enters (acting on tools) and where it is contained (escalation to a person).

ERIC SO · SERTAC KARAMAN · METR

From tools we use to agents that act

Every industrial revolution I know follows the same script: output rises, old industries die, new ones are born. Generative AI follows that script only in part. It moves faster than anything before it, it acts rather than waits to be used, and it makes thinking itself cheap to reproduce. That combination is new.

Speed, agency, abundance

Earlier technologies needed us to point them somewhere: a microscope cannot choose its specimen. AI works with tools instead of being one, and it is climbing rungs of the cognitive ladder that used to belong to professionals. What took decades in earlier revolutions now takes quarters.

The next data sits inside companies

The data used to train frontier models has grown roughly tenfold every 12 to 15 months, and public text is close to exhausted. The next increment lives in the private archives of banks, law firms, manufacturers and hospitals. Few of them yet treat that archive as the strategic asset it has become.

Scale, not cleverness, was the breakthrough

The 2017 Transformer paper mattered less because it made models smarter than because it made them scale. From then on, progress became a function of data, compute and capital, and capability became a capital-allocation question. Sertac Karaman (MIT) goes as far as calling scaling laws something that seems to be “a law of nature”.

Gradual, then sudden

METR measures how long a task an agent can finish with 50% reliability: about a second of human work in 2019, more than an hour by 2025. On Humanity’s Last Exam the best score went from 3.3% to 64.7% in two years. A capability turns commercial the day it crosses a reliability threshold, and nobody can date that day in advance.

FRAMEWORK SERTAC KARAMAN, MIT

Four eras of computing



WOJCIECH MATUSIK · REGINA BARZILAY · RICH SUTTON · MICROSOFT RESEARCH

A plausibility engine, not an oracle — and evidence has levels

A language model does one thing: it predicts the next most plausible word, appends it and repeats. It completes patterns; it does not look anything up. Once leaders internalize that, evaluation stops being a technical chore and becomes the discipline that decides where value can be captured.

Right and wrong answers look identical

Fluency and error come from the same mechanism, so hallucination can be reduced but never removed. Retrieval from trusted sources, tool use, citations and a second model checking the first all help. The danger concentrates exactly where professional work is most exposed: precise numbers, quotations, references and rare facts.

Adoption follows the removal of burden

In healthcare, the fastest-adopted AI is the ambient scribe, which diagnoses nothing. The first wave in any business should target the work skilled people feel as a tax on their expertise: documentation, preparation, reconciliation, reporting. And the value often appears somewhere other than the vendor promised.

General beats special; prompt before you fine-tune

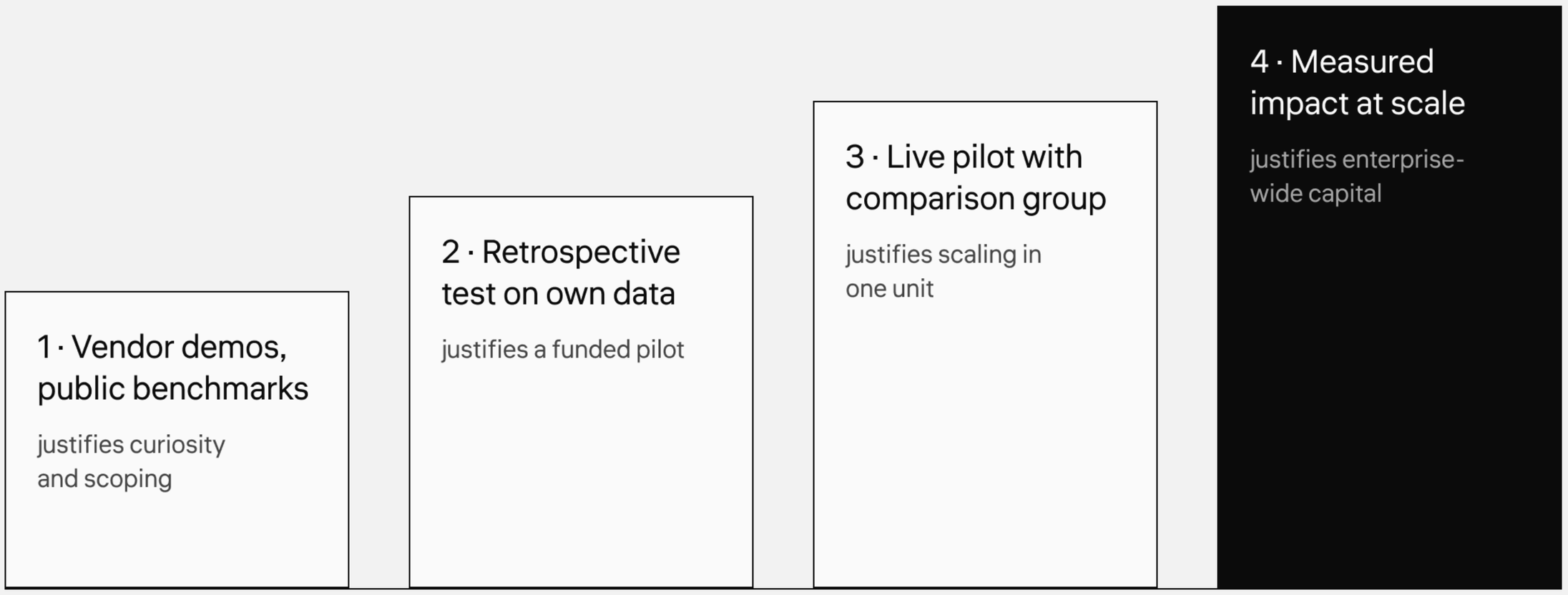
BloombergGPT, trained on proprietary financial data, was overtaken on many finance tasks by a general model within months. Microsoft's Medprompt took a general model past 90% on a medical licensing benchmark with prompting alone. My rule: exhaust prompting and retrieval first, and fine-tune to fix behavior, never to teach facts.

A prediction needs someone to act on it

Regina Barzilay's Mirai model predicts breast-cancer risk years ahead from mammograms that looked normal; the data already existed. But a forecast without an owner who is authorized, funded and equipped to act creates nothing. A churn model without retention authority is an expensive dashboard.

FRAMEWORK · ADAPTED FROM REGINA BARZILAY, MIT

Evidence hierarchy — capital should scale with the level reached



The token is the unit of cost. The finished task is the unit of value.

Global AI investment passes $1 trillion in 2026, and four US platforms alone plan $720–760 billion of capital spending, about a third of Brazil’s GDP. On the buyer’s side the paradox is sharper: the price of a token has fallen by orders of magnitude while enterprise AI bills keep rising. Cheaper units, bigger bills.

Why the bill rises

Agents chain dozens of calls, and reasoning models consume tokens the buyer pays for and never sees. The FinOps Foundation estimates five to thirty times the tokens of an ordinary chat for such workloads; AT&T went from about 8 to 27 billion tokens a day after deploying multi-agent systems.

The subsidy phase is ending

Contracts are moving from seats to forecast and pre-paid compute, and the procurement question changes accordingly: how much will we consume, and on what? Menlo Ventures puts enterprise generative-AI spending at $37 billion in 2025, 3.2 times the year before, with three-quarters of solutions bought rather than built.

Cost is a property of the configuration

Model tier, precision, inference engine, hardware, service level and, above all, the harness decide the bill. The levers multiply one another: route work to the right model, discipline the harness, choose the consumption model and make cost visible.

Price the task, not the token

J.P. Morgan Private Bank prices a completed task at a few cents with some Chinese providers and close to $3 with the most expensive US model, a seventyfold spread. Usage matters even more than price. A five-cent task that fails a third of the time can cost more than a ninety-cent task that succeeds.

A lower bill is not the goal

Half-finished cognitive work is worth almost nothing. “AI cut drafting time 40%” rarely reaches the P&L; “AI resolves first-level support end to end” does. In a company creating real value, total AI spending should rise while the cost per outcome falls.

04

FRAMEWORK

Three units for measuring AI value

The task	The decision	Trust
operational unit — completed without supervision	economic unit — one paragraph that stops a bad deal	limiting unit — value appears where verification can stop

From answers to outcomes

An agent observes its environment and acts on it with the tools it has, in pursuit of a goal. Every element of that definition is a design decision, and writing the goal well is now a management competence rather than an IT one.

Content risk becomes action risk

A chatbot's mistake is a bad paragraph; an agent's mistake is a wrong payment or a promise made to a customer. Documented cases already include an executive assistant agent committing to a five-figure sponsorship and test agents posting publicly after being told to stay internal.

Every hand-off is a trust boundary

In multi-agent systems one agent plans, narrow agents execute and an independent agent checks. Each hand-off can fail, and human review can be defeated simply by flooding the reviewers. Recording every decision and every rejected alternative is what makes the system auditable.

The harness decides reliability

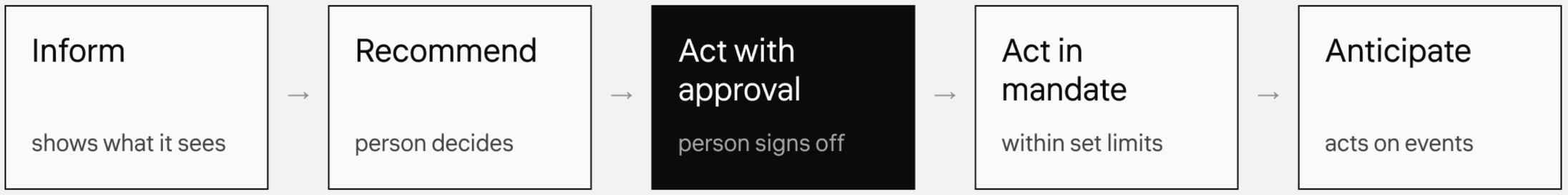
The model is only the reasoning step inside a loop. Goal, memory, tools, guardrails, the check step and the human gate form the harness, and that is where reliability and cost are actually decided.

Software development previews all knowledge work

Agents now write much of the code while people specify, orchestrate and verify; Stripe merges more than a thousand agent-written changes a week. Yet METR found experienced developers slower with AI while believing they were faster. Review capacity, not agent capacity, is the constraint.

FRAMEWORK GOOGLE CLOUD

The autonomy spectrum — highlighted: where consequential actions should sit until evidence says otherwise



Programming is dead. Engineering is alive.

Strip away the mystique and an agent is a model given a list of functions and asked which ones to call. Tim Kraska (MIT CSAIL) calls it “simply a well-written prompt”. If that is true, the hard part of software moves from writing code to stating intent precisely, and the same shift awaits every knowledge profession.

Why most projects still fail

Gartner expects more than 40% of agentic AI projects to be scrapped by 2027, and McKinsey finds that almost two-thirds of companies have experimented with agents while fewer than one in ten have scaled them. Most of the obstacles belong to the adopting organization: missing context, no memory, poor workflow fit, data gaps.

Reliability compounds downward

A process that is 95% reliable at each step succeeds end to end only about 36% of the time after twenty steps. The most valuable flows are therefore the least dependable: break them into verifiable segments with checkpoints, and test vendors on your own tasks rather than their demos.

Software becomes knowledge engineering

When natural language, organized as a shared ontology, becomes the source from which code is generated, intent is the durable asset. The balance of effort inverts, from mostly building toward mostly specifying and judging.

Agreement on intent, not hand-offs

The development cycle that works is built on aligning objectives, then architecture, then deployment, rather than on documents thrown over walls. Start with one application end to end, let business owners change the product directly, avoid long vendor lock-ins, and drive it from the top.

The market prices the prediction first

On 3 February 2026 the main software index fell 5% in a day and 28% from its peak. Adobe lost two-thirds of its market value in five years, and ServiceNow, Intuit, Zoom and Atlassian left the global tech top 50: investors are discounting seat-based software before revenue shows it (AI Finance Battle, Part II).

04

FRAMEWORK TIM KRASKA, MIT CSAIL

Align: a development lifecycle built on agreement about intent

Business engineering
align on objectives → requirements, mock-ups

→

System engineering
align on architecture → plan, prototype

→

Integration engineering
align on deployment → tested software, operation plan

SERTAC KARAMAN · ANTHROPIC · EPOCH AI

Context is the new bottleneck, and context design is cost design

The model never touches a file or a system. It writes instructions; the harness executes them and feeds the results back into the context window. That window is finite, quality falls as it fills, and every step sends the whole of it back to the model, to be paid for again.

Load the window from stable to specific

System instructions, tool descriptions, skills and memory come first, the user's request last, with results added as the agent works. Two agents on the same model behave very differently because of what they load. A firm's codified expertise lives in its skills.

Order is an economic decision

Stable content placed first can be cached. In a twenty-step task with a 50,000-token stable prefix, at published Opus 5.5 rates, caching cuts the input bill from $4.80 to $1.24, 74% less. Context discipline often matters more than the choice of model.

Seats versus usage

A single agentic session can consume more tokens than weeks of conversation, which is why providers add caps, premium tiers and usage-based contracts. Seats suit individual productivity; automation belongs on usage, with cost governance attached.

Open plumbing, rented intelligence

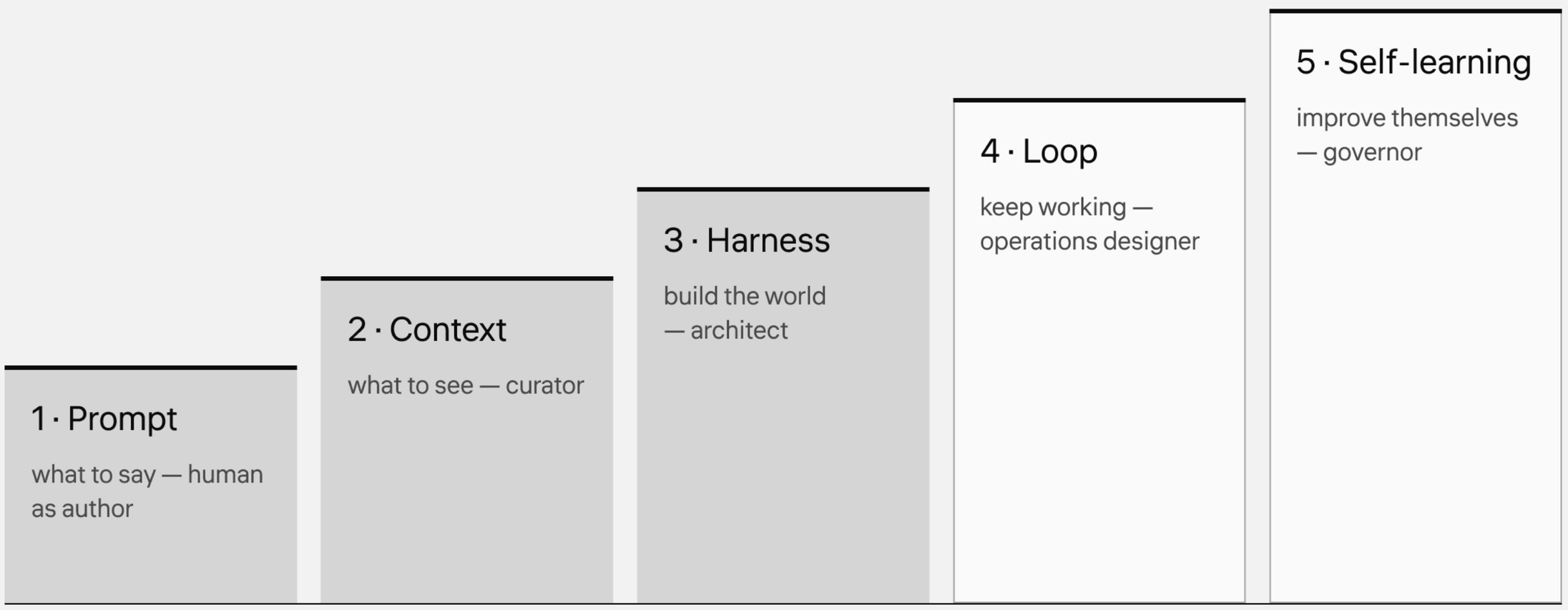
The Model Context Protocol, donated to a neutral foundation, reached about 97 million monthly downloads by December 2025. With the leading labs separated by a handful of points, value moves to distribution, infrastructure and proprietary data. Prompting is a personal skill; engineering the harness is an institutional capability.

Four months of frontier

On Epoch AI's index the best open-weight models trail the best closed ones by about four months of progress. Frontier pricing power rests on those four months and only for work that needs them; everything else is repriced toward open-weight cost. A router across several models is the right enterprise architecture.

FRAMEWORK SERTAC KARAMAN, MIT

Five levels of shaping what an agent does — most organizations today sit at levels 1 – 3



Language is the interface. The engine is somewhere else.

Most enterprise AI work is not writing but deciding: is this a refund request, is this transaction risky, which supplier should serve this order. A language model answers with prose that represents a decision, which software must then parse and trust. That is slow, costly and fragile, and it hides uncertainty inside confident sentences.

Decisions without prose

A new category of decision models returns a choice, a score or a calibrated probability that ordinary code can act on. Whatever the vendor claims, the division of labor is right: generative models draft and explain, decision models judge, code holds policy, and people take the ambiguous cases.

VII.2

Map decisions, not just tasks

Most workflows are chains of small, bounded judgments: classify, route, approve, flag. Each deserves a confidence threshold set by its consequence rather than by how fluent the model sounds, and calibrated confidence becomes the trigger for escalation to a person.

LANGUAGE MODEL OR DECISION MODEL?

	LARGE LANGUAGE MODEL	DECISION MODEL
Output	Text, code, explanations	Typed decisions with probabilities
Best for	Open-ended drafting, planning, explaining	Bounded, frequent, high-volume judgments
Uncertainty	Expressed in words, hard to use	Numeric; can drive thresholds and escalation
Explainability	Can explain, not always faithfully	None natively; needs a separate record
Characteristic failure	Fluent but false content	Valid-looking wrong choice; calibration drift

VII.3

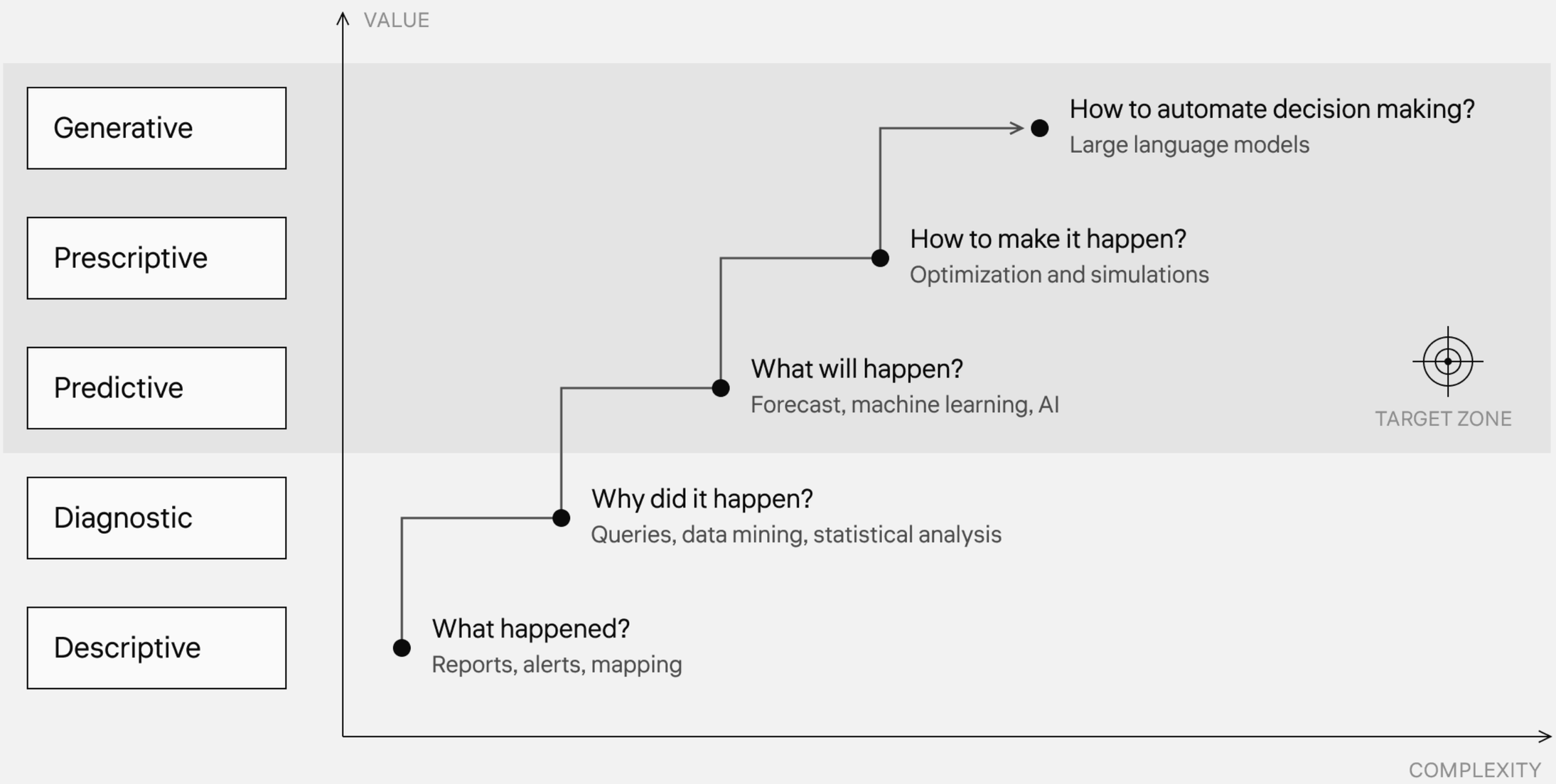
From dashboards to decisions

Analytics climbs from describing and diagnosing to predicting, prescribing and generating. In supply chains, analytics alone has delivered 10–30% better forecasts and 10–20% less inventory. It does not remove the trade-off between cost and speed; it moves the frontier.

FRAMEWORK

DAVID SIMCHI-LEVI, MIT

Supply chain business intelligence and analytics



Redrawn from David Simchi-Levi, MIT (© 2025 D. Simchi-Levi). Value rises with complexity; each rung answers a harder question with a different class of tool. Most organizations still operate on the lower two rungs; the target zone is predicting, prescribing and automating decisions, where language models sit on top of the optimization rather than in place of it.

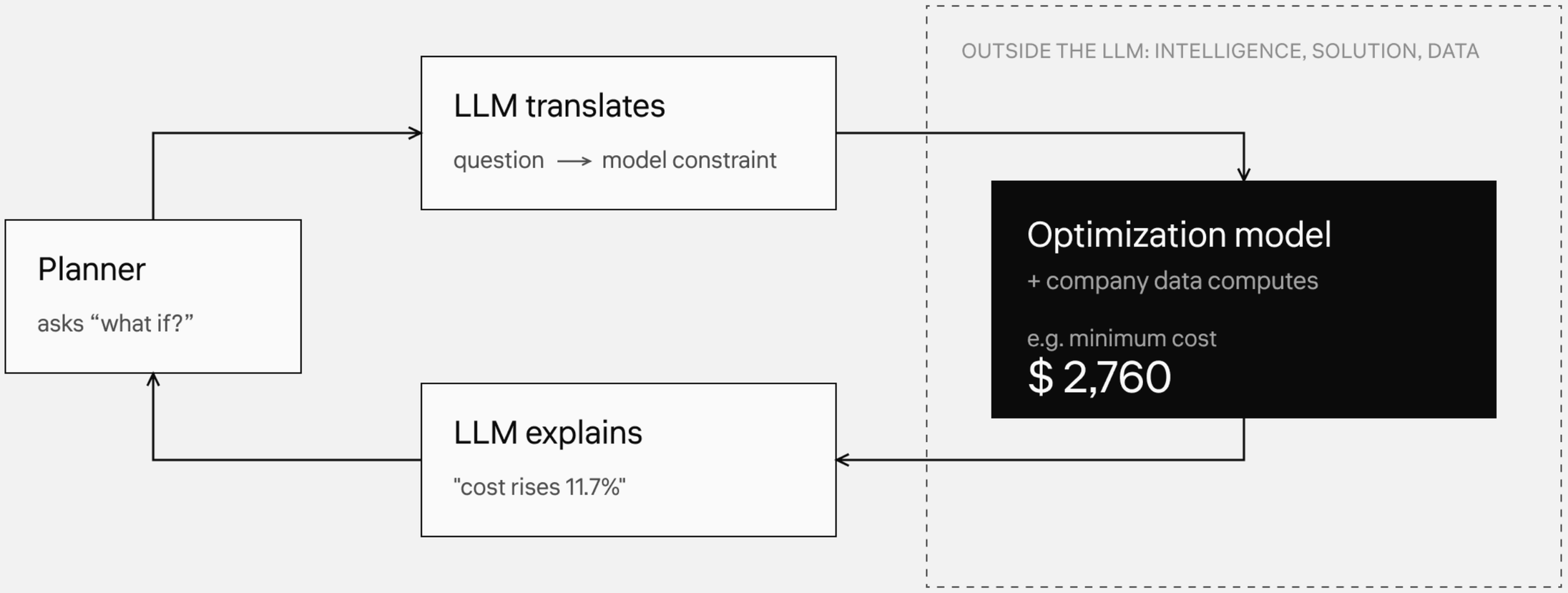
VII.4

The language model as interpreter

In Microsoft’s cloud supply chain, a planner asks “what if?”, the model turns the question into a constraint, an optimization engine computes on company data, and the model explains the answer: analyses that took days take minutes. Keep arithmetic in auditable engines, policy in code, and language as the interface.

FRAMEWORK SIMCHI-LEVI ET AL., MICROSOFT SUPPLY CHAIN

The language model as interpreter



After David Simchi-Levi, MIT (2026). Asked what happens if Supplier A no longer serves demand B, the model adds that constraint; the solver returns a minimum cost of $2,760, an 11.7% increase, which the model explains in plain language. The numbers come from the engine and the data, never from the language model.

CASES FROM SIMCHI-LEVI (MIT SLOAN, 2026)

CASE	WHAT AI DID	REPORTED RESULT
Large cloud supply chain	LLM interface to planning and fulfillment optimization	Analyses from days to minutes; planner time freed
European fashion retailer	Transformer-based forecasting, pricing and automated planning	Short-term forecast error –30%, long-term –9%, bias –11%
Large automotive manufacturer	LLM reading supplier contracts for price reductions not applied	Millions of dollars in procurement savings
Latin American retailer	Data-driven pricing and inventory replacing intuition	Units sold and profit up 6%

Models are rented. Advantage compounds elsewhere.

Frontier capability is now so accessible that, as Andrej Karpathy put it, “money can’t buy better ChatGPT.” If every competitor rents the same models on the same terms, advantage has to come from what cannot be rented.

Three layers of practice

Picture a brilliant new hire. Prompt engineering is what you tell them, context engineering is what you put on their desk, harness engineering is the office they work in. Most organizations are still working on the first question.

What you can teach the model

Advantage comes less from access to models than from what a company can teach them: private evaluations that turn tacit expertise into its own standards for judging AI output. That requires protecting the apprenticeship that produced the expertise in the first place.

VIII.3

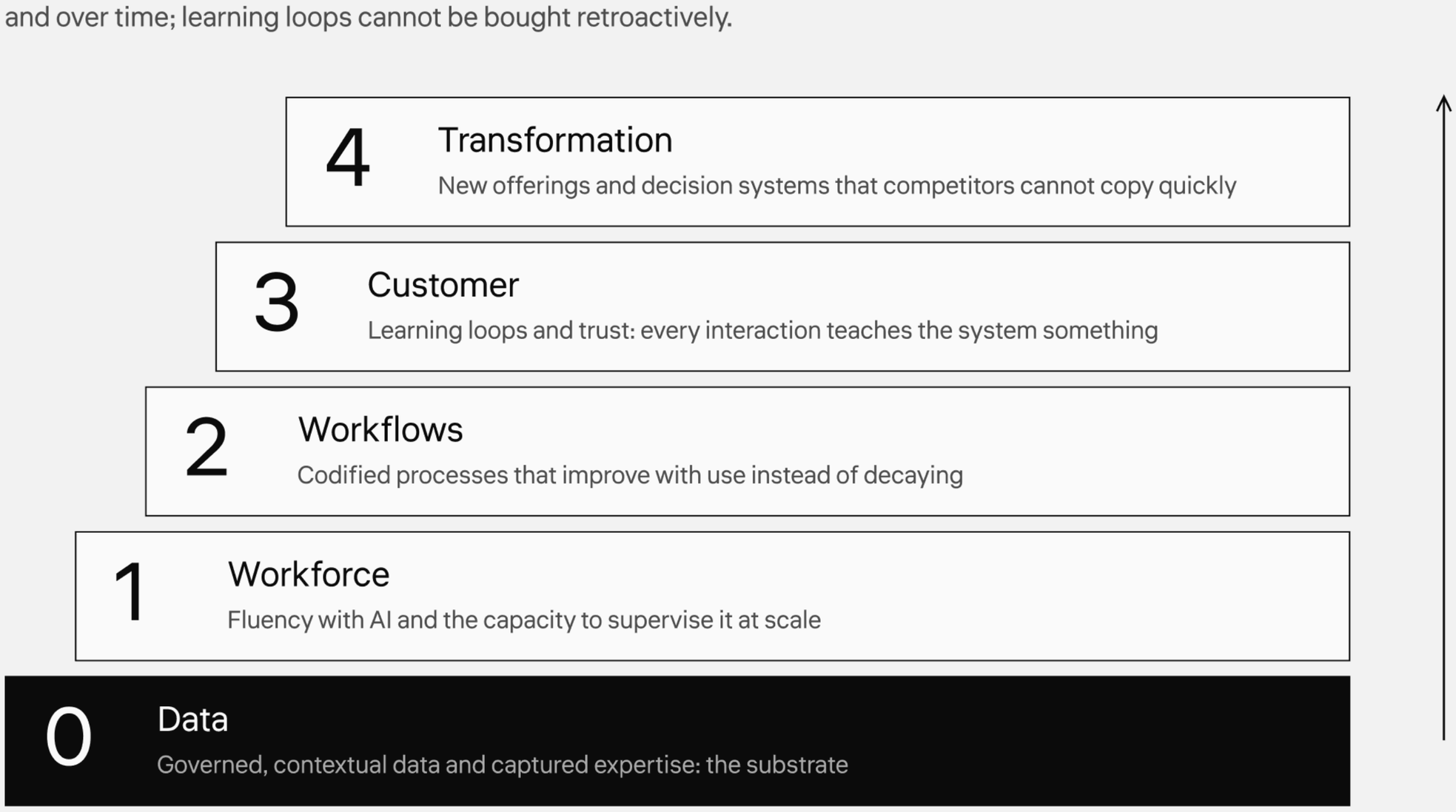
Time is the moat

Returns follow a J-curve, so boards should expect investment before value. Learning loops cannot be bought retroactively: a firm that starts two years later faces a gap widened by two years of compounding.

FRAMEWORK GOOGLE CLOUD

Five layers of compounding advantage

Each layer learns from the one beneath it, and data is layer zero. Advantage accumulates upward and over time; learning loops cannot be bought retroactively.



TIME IS THE MOAT

A firm that starts two years later faces a gap widened by two years of compounding.

Simulated customers, real limits

John Horton’s work uses language models as stand-in respondents. Agents built from two-hour interviews reproduced people’s survey answers about 85% as well as the people reproduced themselves. The quality of the simulation depends on the depth of the human data, not on the model.

Value migrates up the stack. Organizations must be rebuilt to follow it.

This is where our advisory work at House of Brains begins: with the lenses we use with boards to see where value is moving and whether an organization can follow it. Most of them took shape in my books A Liderança Essencial and Anthesis, which argued that a transition in social ethics and the digital transformation would frame the deepest changes of the coming decades. With AI at the center of that convergence, the argument reads less as a forecast than as a description.

IX.1

HOB PROPRIETARY MODEL · A LIDERANÇA ESSENCIAL

Organizational dynamics

Ambition is cheap to write. Potency is expensive to build.

Every organization can be read as a pyramid in two halves. The upper half is ambition: purpose, vision and strategy, the answer to why we exist, where we are going and how we intend to win. The lower half is organizational potency: culture, design and capabilities, the answer to how we behave, how we are organized and what we are actually able to do. Success is decided on the dotted line between them. Ambition without potency produces beautiful strategy decks and frustrated teams; potency without ambition produces efficient organizations heading nowhere in particular. AI changes the economics of the two halves in opposite directions. It makes the top of the pyramid almost free to produce: a competent strategy narrative, a vision statement or a market analysis is now a prompt away. It does nothing, by itself, for the bottom. Culture, design and capabilities still have to be built by people, over time. That is why the gap between what companies say about AI and what they can do with it widens precisely when ambition becomes cheap. In our work with boards, the first diagnostic is never the AI strategy. It is the fit. The pattern shows up in the data on how organizations are actually built. In BMI's REBOOT study, 92% of the organizations analyzed ran matrix or multidivisional structures, while 58.7% aspired to network models: an ambition the lower half of the pyramid has not caught up with. Agents can finally supply the connective tissue that network organizations always lacked, but only if design and capabilities move with them. Gartner expected one in five organizations to use AI to flatten their structures by 2026; flattening without rebuilding culture and capabilities simply moves the fit line further from the strategy.

HOW TO READ THE MODEL

Read the pyramid from top to bottom. Each block answers one question: why we exist (purpose), where we want to be (vision), what we will do and not do (strategy), how we behave in practice (culture), in what form we organize (design) and who carries it out (capabilities). The dashed line is the fit line: everything above it is intention, everything below it is ability.

WHAT TO DO

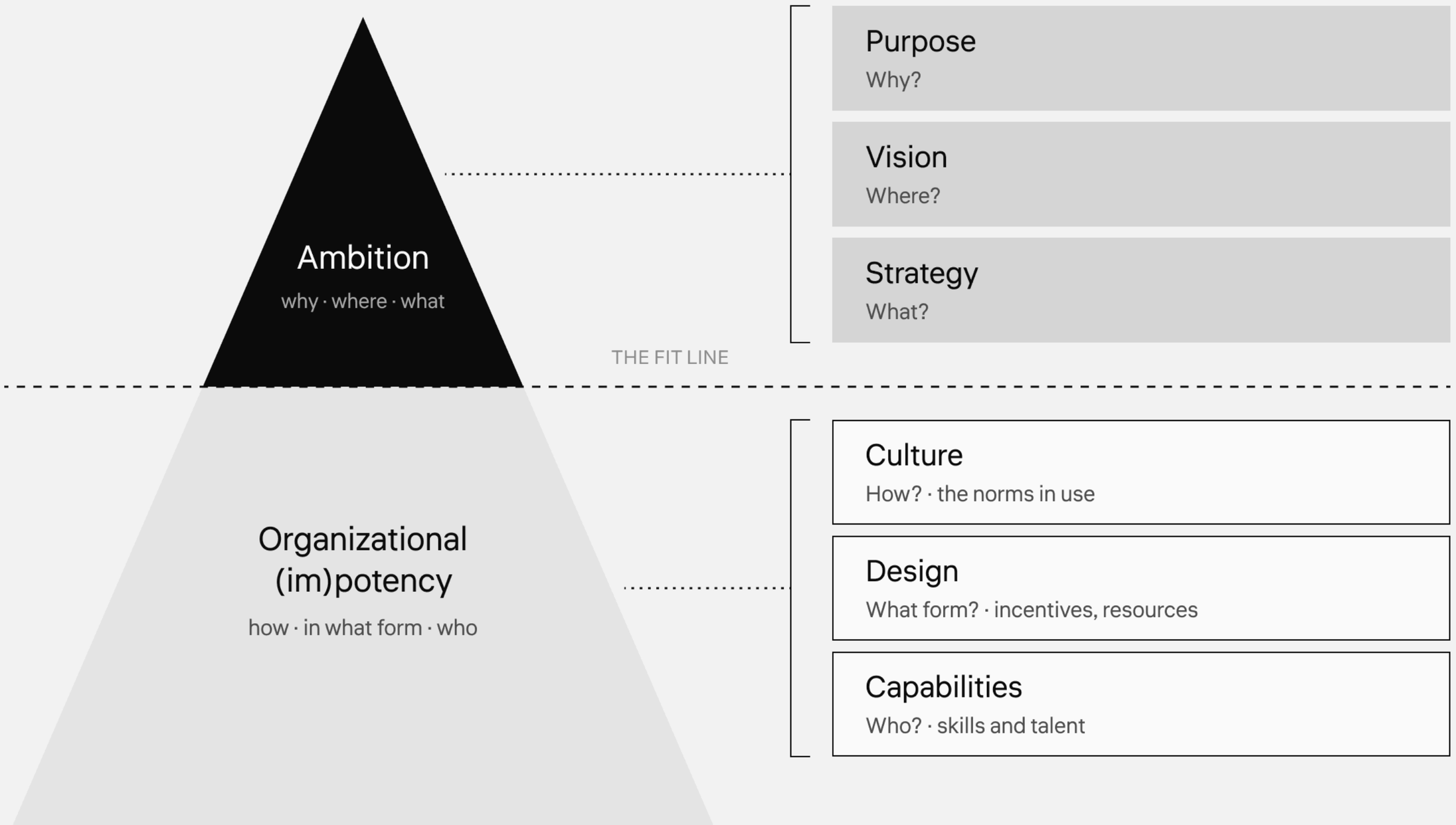
Test every AI initiative against the three lower blocks before approving it. Name an owner for culture, design and capabilities in each major AI program, not only for the technology. Report the fit, not the ambition: progress on the bottom half belongs in the board pack.

QUESTIONS FOR THE BOARD

1. Which of our AI ambitions has a named owner in culture, design and capabilities?
2. Where has our strategy outrun our ability to execute it?
3. What should we stop announcing until the bottom half catches up?

HOB PROPRIETARY MODEL

A LIDERANÇA ESSENCIAL



Organizational dynamics

HOB proprietary model, from D. A. Motta, A Liderança Essencial. Ambition (purpose, vision, strategy) sits above organizational potency (culture, design, capabilities). The dashed line is where success is decided: the fit between the two halves.

IX.2

HOB PROPRIETARY MODEL · A LIDERANÇA ESSENCIAL

From an ethic of duty to an ethic of pleasure

The social contract of work changed before the technology did.

The first transition is social. For most of the industrial era, work was organized around an ethic of duty: norms and standards, command and control, centralization. People bonded to their employers through loyalty, stability and hierarchy. In A Liderança Essencial I describe the passage to an ethic of pleasure, in which the essential pillars of work become fluidity, autonomy and interdependence, and the emotional bond is built on connection, expression and experience. This is not a generational fashion; it is a change in the social norm, and AI accelerates it. Agents remove much of the routine that duty-based structures were designed to police, and they give individuals the means to work with an autonomy hierarchies were never built to accommodate. Companies that deploy AI inside a duty-based culture get compliance on paper and shadow AI in practice. Those that redesign around autonomy and connection get adoption. The shift is visible in how people adopt AI. Professionals choose their own tools, on their own terms and often outside the official channel: the ethic of pleasure at work. Leaders who answer with the instruments of the old ethic, prohibition, centralization and appeals to loyalty, push the most valuable experimentation out of sight, where risk is highest and learning is lost. The psychological contract between individual and organization has to be rewritten, and AI is the most concrete occasion to do it.

WHAT TO DO

Rewrite AI policies around the autonomy you are willing to grant, with clear limits, rather than around prohibition. Offer sanctioned tools at least as good as the ones people already choose for themselves. Build the emotional bond on connection, expression and experience, not on control.

QUESTIONS FOR THE BOARD

1. Do our AI policies assume people must be controlled, or enabled?
2. Where does the emotional bond with our best people actually come from today?
3. Which of our rules exist only because the old ethic required them?

HOW TO READ THE MODEL

Read each row from left to right: the premise of the old ethic and the premise that replaces it. The upper pair names the transition itself; the middle pair shows the pillars of work relations; the lower pair shows what binds people to an organization.

HOB PROPRIETARY MODEL

A LIDERANÇA ESSENCIAL

From an ethic of duty to an ethic of pleasure

SOCIAL-ETHICS TRANSITION

Ethic of duty		Ethic of pleasure
Essential pillars of the ethos in work relations		
Norms and standards		Fluidity
Command and control	→	Autonomy
Centralization		Interdependence
Drivers of the emotional bond at work		
Loyalty		Connection
Stability	→	Expression
Hierarchy		Experience

HOB proprietary model, from D. A. Motta, A Liderança Essencial.
Read each row left to right: the premise of the old paradigm and the one that replaces it.

IX.3

HOB PROPRIETARY MODEL · A LIDERANÇA ESSENCIAL

From analog–mechanical to digital–intelligent

The second transition rewrites what an effective organization looks like.

The second transition is technological. The analog–mechanical paradigm rewarded predictability, incrementalism and departmentalism, and it built organizations on cohesion, perfection and self-sufficiency. The digital–intelligent paradigm runs on volatility, rupture and ecosystems, and it demands connectivity, collaboration and permeability. Read the figure from top to bottom: the second row is the design brief an organization inherits from the first. Generative AI is the purest expression of the digital–intelligent paradigm so far. It ignores departmental boundaries, it improves in leaps rather than increments, and its value depends on what it can connect to. Organizations designed for perfection struggle with a technology that is probabilistic by nature; organizations designed for self-sufficiency struggle with one that is rented from an ecosystem. Much of the friction executives describe in AI programs is this paradigm clash, not a technology problem. The power of the digital transformation lies in convergence: artificial intelligence together with the industrial internet, robotics, extended reality and genomics, sustained by data, cloud computing and networks. As I wrote in 2019, for organizations built on the old paradigm it is a question of renewed life or announced death. The two transitions reinforce each other, and agents are permeability in its purest form: capability that enters the organization from outside, on demand, across every boundary.

HOW TO READ THE MODEL

Read the figure from top to bottom. The headline row names the paradigm; the middle row shows the pillars of work relations each paradigm rewards; the lower row shows the premises an organization needs to be effective in it. Each premise on the left was a strength under the old paradigm and becomes a liability under the new one.

WHAT TO DO

Replace perfection with learning: run AI as a portfolio of experiments with explicit stop rules. Open the boundaries that matter: data sharing, partner integration and access to external talent. Design for connectivity first; a capability that cannot be connected will not compound.

QUESTIONS FOR THE BOARD

1. Where do we still optimize for perfection when we should optimize for learning?
2. Which of our boundaries, internal or external, should become permeable?
3. Who in our ecosystem holds the data, models and talent we will never own?

HOB PROPRIETARY MODEL | A LIDERANÇA ESSENCIAL

From analog–mechanical to digital–intelligent

TECHNOLOGICAL-PARADIGM TRANSITION

Analog–mechanical		Digital–intelligent
Essential pillars of the ethos in work relations		
Predictability	→	Volatility
Incrementalism	→	Rupture
Departmentalism	→	Ecosystem
Premises for effective organizational design		
Cohesion	→	Connectivity
Perfection	→	Collaboration
Self-sufficiency	→	Permeability

HOB proprietary model, from D. A. Motta, A Liderança Essencial.
The second row is the design brief an organization inherits from the first.

IX.4

HOB PROPRIETARY MODEL · ANTHESIS

Turbulence and opportunity in structural change

Where the two transitions meet, ten megatrends reshape demand.

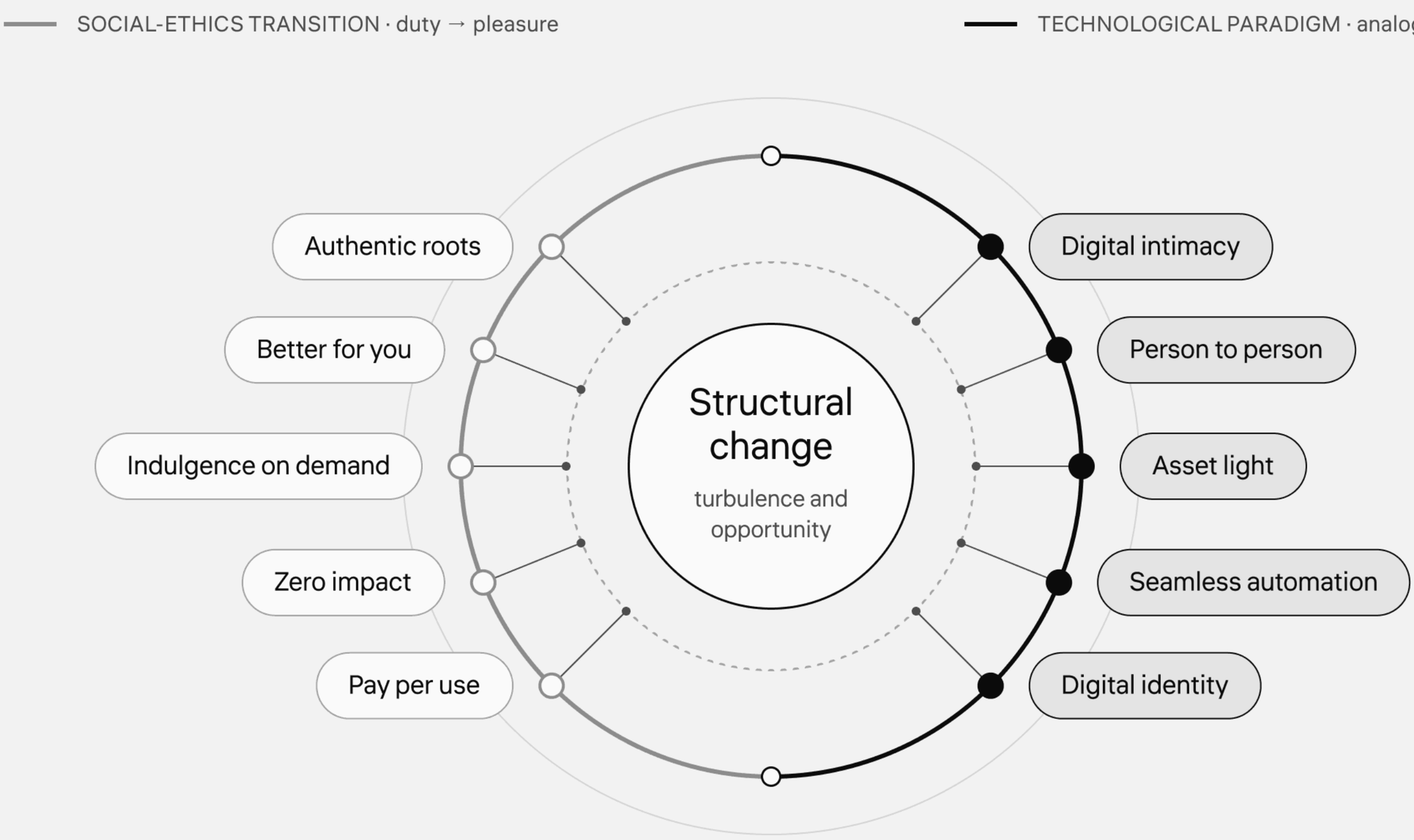
In Anthesis I mapped the ten megatrends that emerge where the social-ethics transition and the technological paradigm meet. The left half of the orbit draws mainly on the social side: authentic roots, better for you, indulgence on demand, zero impact and pay per use. The right half draws on the technological side: digital intimacy, person to person, asset light, seamless automation and digital identity. Together they describe how demand, business models and the labor market are being reshaped. AI accelerates every one of them, as the table below details. For a board, the map works as a triage tool rather than a forecast: which megatrend is reshaping our customers, which is reshaping our cost

structure, and which is reshaping the talent we need? Turbulence and opportunity are the same movement seen from opposite sides of the orbit. Each megatrend is accelerated by AI in a specific way. As synthetic content becomes abundant, verified provenance becomes a scarcity premium for authentic roots. Personalized coaching and monitoring at near-zero cost turn better for you into a mass market. Conversational agents make digital intimacy available at scale, and seamless automation moves from single tasks to end-to-end outcomes. The same movement is turbulence for incumbents organized around the old premises and opportunity for those who read it early.

HOW TO READ THE MODEL

Read the orbit as a map of forces rather than a list. The ten megatrends sit around structural change, each on the side of the transition that drives it most directly: the social-ethics transition on the left, the technological paradigm on the right. The placement is a lens, not a taxonomy; most megatrends draw on both.

HOB PROPRIETARY MODEL ANTHESIS



Turbulence and opportunity in structural change

HOB proprietary model, from D. A. Motta, Anthesis (2019). The left half of the orbit draws mainly on the social-ethics transition, the right half on the technological paradigm; the two halves meet at the poles.

WHAT TO DO

Choose the three megatrends that matter most and assign each an executive owner. Translate each into one customer proposition, one cost lever and one talent implication. Revisit the map annually; the orbit moves faster than strategic plans.

QUESTIONS FOR THE BOARD

1. Which three megatrends matter most for our customers over the next five years?
2. Which ones are our competitors already exploiting?
3. Where does AI turn a megatrend from a threat into an option for us?

TEN MEGATRENDS AND HOW AI ACCELERATES THEM

MEGATREND	DRIVEN MAINLY BY	WHAT IT CHANGES	HOW AI ACCELERATES IT
Authentic roots	Social ethics	Origin, provenance and a credible story become part of the value proposition.	As synthetic content becomes abundant, verified provenance becomes a scarcity premium.
Better for you	Social ethics	Health, wellness and self-development become purchase criteria.	Personalized coaching and monitoring at near-zero marginal cost.
Indulgence on demand	Social ethics	Gratification is expected at once and on the customer's terms.	Agents compress the time between wanting and receiving.
Zero impact	Social ethics	Environmental and social footprint becomes a condition of legitimacy.	Better measurement and traceability; AI's own energy demand (Deep Dive XVII) enters the account.
Pay per use	Social ethics	Access replaces ownership; consumption is priced by use.	Token-based pricing is pay per use in its purest form (Deep Dive III).
Digital intimacy	Technology	Relationships mediated by data that know the customer closely.	Conversational agents make intimacy scalable and make trust the limiting factor.
Person to person	Technology	Peer-to-peer exchange bypasses traditional intermediaries.	Agents can search, negotiate and transact on either side of the exchange.
Asset light	Technology	Value comes from alliances, cross-holdings, open innovation and network effects rather than owned assets.	Intelligence itself is now rented rather than owned (Deep Dive VIII).
Seamless automation	Technology	Automation disappears into the journey.	Agents move automation from single tasks to end-to-end outcomes (Deep Dive IV).
Digital identity	Technology	Identity, reputation and credentials become portable digital assets.	Agents need identities, permissions and a named human owner of their own.

IX.5

HOB PROPRIETARY MODEL · ANTHESIS

Creation and destruction of value

Each new layer captures value; the layer below converges to the cost of capital.

Value migrates up a sequence of layers: resources, products, services, platforms and operating systems. Each new layer adds complexity and, while it is the newest, captures the largest share of value; the layers below are not destroyed, but their returns gradually converge to the cost of capital. In Anthesis I traced the sequence historically: services added value to products in the mid-twentieth century, platforms captured it in the internet era, and operating systems now organize whole economies. In AI the sequence is running at speed. Chips, energy and data centers form the resource layer and are collecting the surplus today, because that is where supply is scarce; the AI Finance Battle shows that the equity market has already priced it. Models are commoditizing toward the product layer. The contest that matters for most companies is the next one: who will own the operating system of knowledge work, the layer that decides which agent works for whom. The framework points to the hard question for investors in AI. If capital is concentrated in the lower layers while asymmetric returns migrate upward, part of today's investment will earn no more than its cost of capital, however essential the infrastructure turns out to be. For a company that buys rather than builds AI, the implication is the opposite and more favorable: most of the physical capital belongs to someone else, and the opportunity lies in the services, platforms and operating positions that sit on top of it.

WHAT TO DO

Map your revenue by layer and identify where returns are converging to the cost of capital. Avoid capital-heavy positions in layers that are commoditizing; rent them instead. Invest in the layer above your current one: services on top of products, platforms on top of services.

QUESTIONS FOR THE BOARD

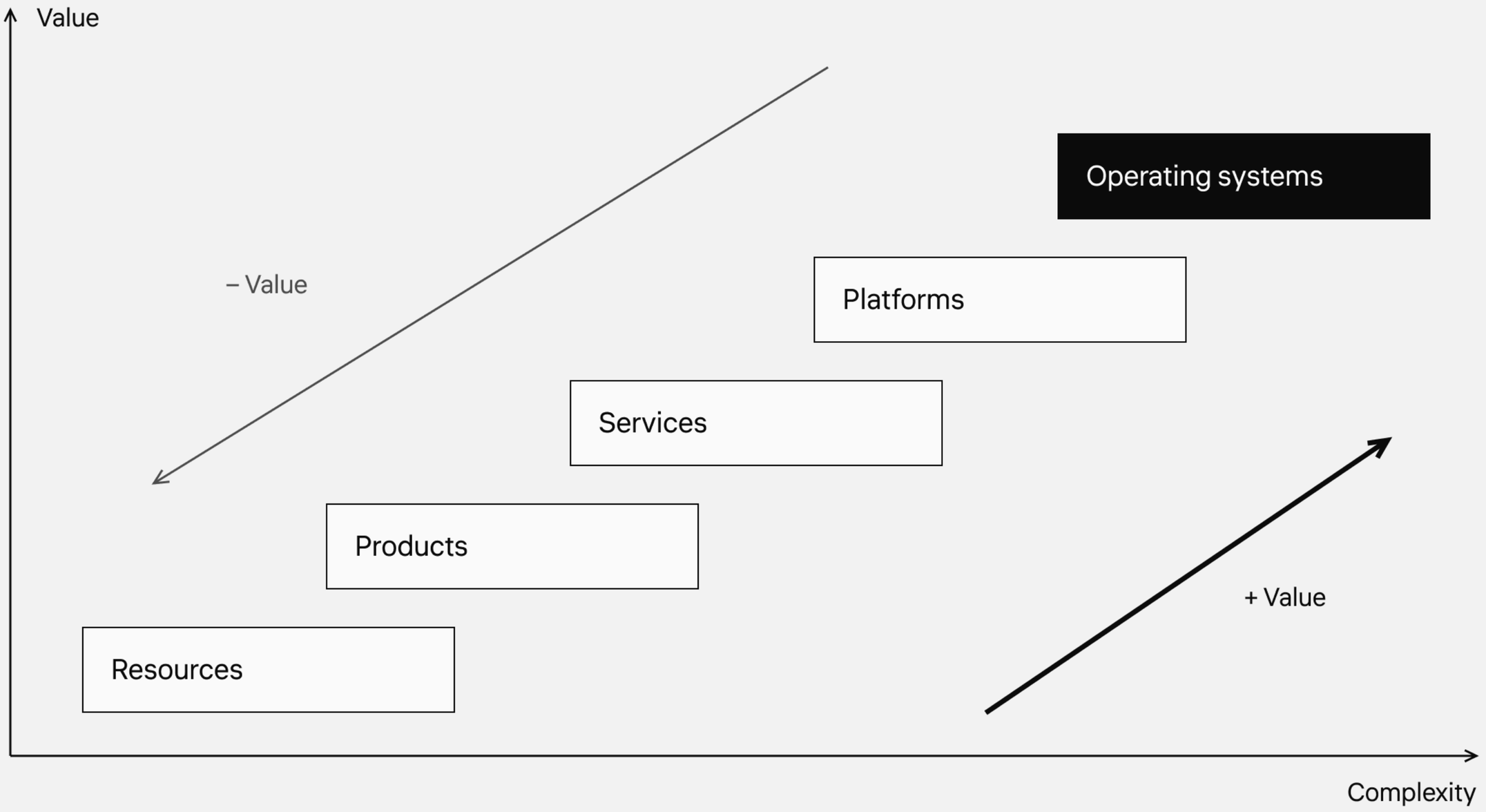
1. On which layer do we earn our returns today, and is it converging to the cost of capital?
2. What would it take to move one layer up?
3. Whose operating system will our customers live in?

HOW TO READ THE MODEL

Read the staircase from bottom left to top right. Each layer adds complexity and, while it is the newest, captures the largest share of value. The diagonal arrows show the paradox: value is created as complexity rises, and destroyed in the older layers as their returns converge to the cost of capital. The figure is schematic; no scale is implied.

HOB PROPRIETARY MODEL | ANTHESIS

Creation and destruction of value



HOB proprietary model, from D. A. Motta, Anthesis (2019). Each layer adds complexity and, while it is the newest, captures the largest share of value. Schematic; no scale implied.

IX.6

HOB PROPRIETARY MODEL · ANTHESIS

The digital competitive arena

Always on, always beta: the rules of competition have changed vertex by vertex.

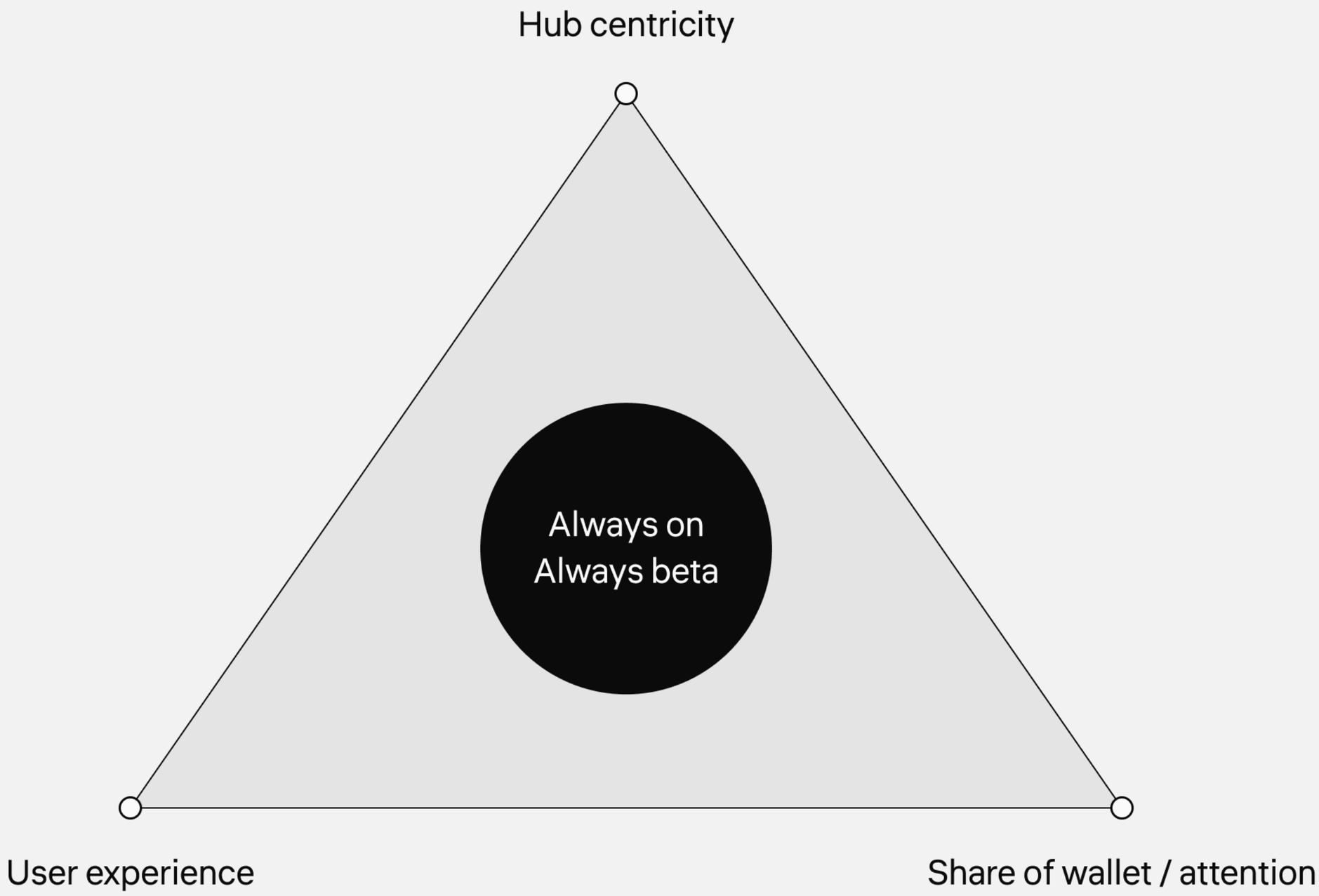
The digital competitive arena is defined by three vertices. Hub centricity replaces vertical integration: the winner is whoever sits at the center of the customer's decisions. User experience replaces product segmentation as the primary source of differentiation. Share of wallet and of attention replaces market share as the measure of success. At the center, two principles set the pace: always on, and always beta. AI moves each vertex. The hub is migrating from the app and the marketplace to the agent: increasingly, the customer will not search, compare or choose; an agent will do it on their behalf. User experience becomes conversational and anticipatory. Share of attention is redefined when the attention being competed for is a machine's. The strategic question for every brand becomes:

whose agent will recommend us, and why? The traditional pursuits of market share, control of distribution channels, long research cycles, legal protection of intellectual capital and heavy investment in physical assets are all put in check in this arena. AI accelerates the shift. Meta has paired its agent with a small wearable companion device, part of a bet that agents, not feeds or app stores, become the front door to digital life and commerce. When an assistant chooses the supplier, books the trip or drafts the contract, brands compete for the agent's recommendation rather than the customer's glance.

HOW TO READ THE MODEL

Read the triangle as a position. Each vertex is a source of advantage: hub centricity (being the node through which interactions pass), user experience (the quality of the journey) and share of wallet and attention (how much of the customer's spending and time you hold). The center sets the pace: always on, always beta.

HOB PROPRIETARY MODEL ANTHESIS



The digital competitive arena

HOB proprietary model, from D. A. Motta, Anthesis (2019, ch. VI). Three vertices define the position; the centre defines the pace.

WHAT TO DO

Make your offer legible to agents: structured data, clear terms, verifiable claims. Decide where you will be a hub and where you will partner with someone else's. Adopt "always beta" deliberately: short release cycles with governance designed for them.

QUESTIONS FOR THE BOARD

1. Are we a hub, or a supplier to someone else's hub?
2. How would an AI agent compare our offer with our competitors'?
3. What does "always beta" mean for our release cycles and our governance?

IX.7

HOB PROPRIETARY MODEL · ANTHESIS

Digital drivers of organizational culture

Light, empathetic, organic: a culture that reinforces itself.

In Anthesis I argued that digital organizations share three cultural drivers that reinforce one another in a cycle. They are light: fewer layers, fewer assets, decisions taken close to the information. They are empathetic: built around the real experience of customers and employees rather than internal convenience. And they are organic: they grow and adapt through experimentation instead of master plans. Each driver enables the next. AI rewards organizations that already work this way and exposes those that do not. Lightness lets AI-enabled teams act on what they learn without waiting for approval chains; empathy keeps automation pointed at outcomes customers value; organic growth lets experiments scale when they work and fade quietly when they do not. Heavy, self-referential and over-planned cultures can buy the same tools and still capture little of their value. AI tests each driver. It makes organizations lighter, but only if the context, data and decision records that agents depend on are built deliberately: lightness of structure demands depth of knowledge. It makes empathy scalable, and organizations that forget how their own people experience working alongside agents will find that use goes underground. Three conditions matter most: psychological safety, so people can say “the agent is wrong”; learning that changes behavior, since in BMI’s Design Mastery study 62% of CHROs prioritized digital upskilling while 56% saw no measurable progress; and protected recovery, because if agents run continuously, people must not.

HOW TO READ THE MODEL

Read the cycle clockwise. Lightness frees an organization from what it owns, so it can look outward at the experience of others. Empathy reveals needs no single firm can meet alone, which pushes it toward organic, ecosystem-based delivery. And orchestrating an ecosystem is possible only for an organization light enough to connect and disconnect quickly.

WHAT TO DO

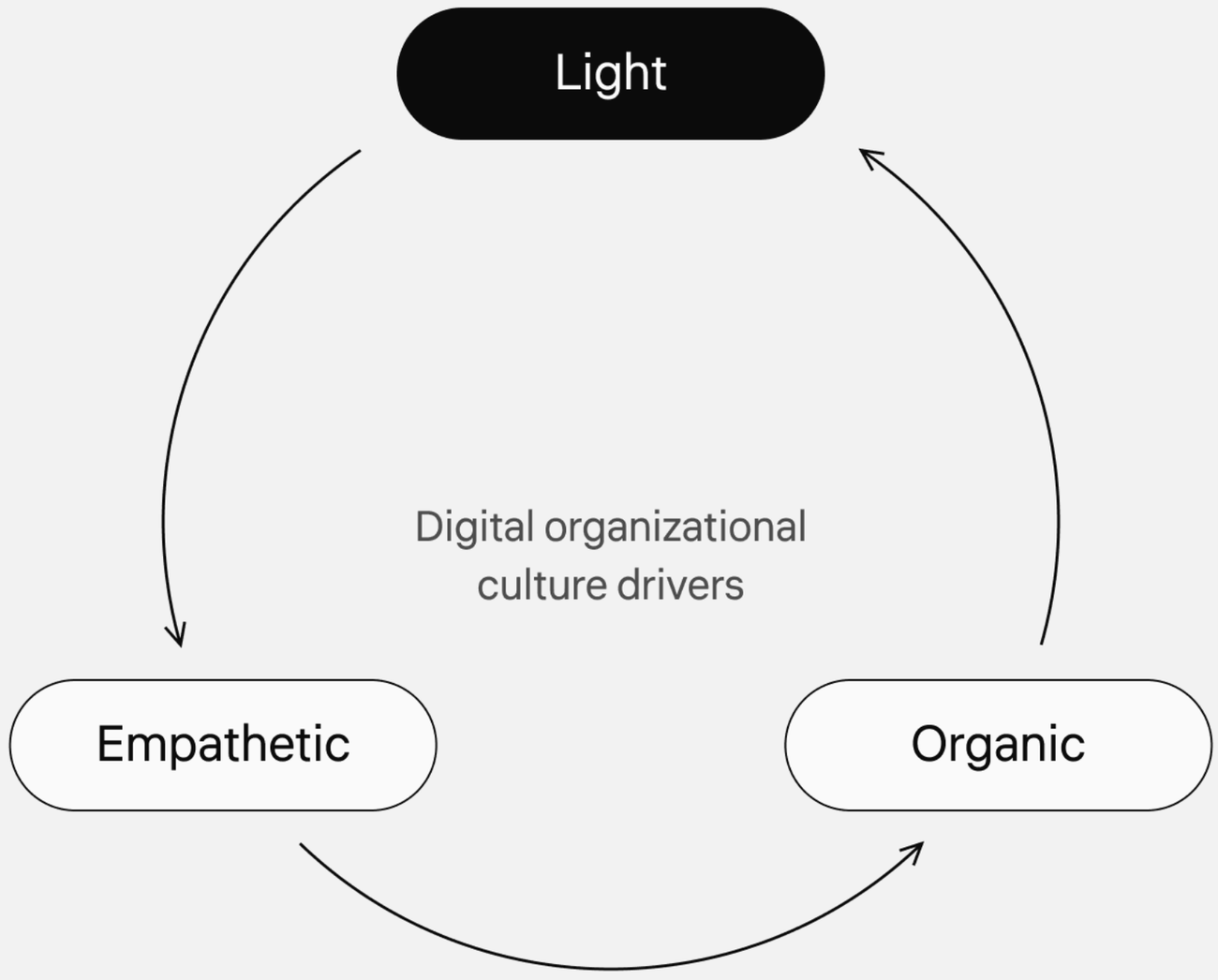
Measure lightness by decisions, not headcount: how many approvals stand between insight and action. Include the employee experience of working with agents in every AI business case. Fund experiments to scale or stop on evidence, within a quarter.

QUESTIONS FOR THE BOARD

1. How many layers of approval does an AI-enabled decision still have to cross?
2. Do our AI use cases start from customer pain or from internal convenience?
3. How fast can a successful experiment become the standard?

HOB PROPRIETARY MODEL | ANTHESIS

Digital drivers of organizational culture



HOB proprietary model, from D. A. Motta, Anthesis (2019). Each driver enables the next in a reinforcing cycle.

IX.8

HOB PROPRIETARY MODEL · ANTHESIS

The value chain for enterprise modelling

Value created is not value kept.

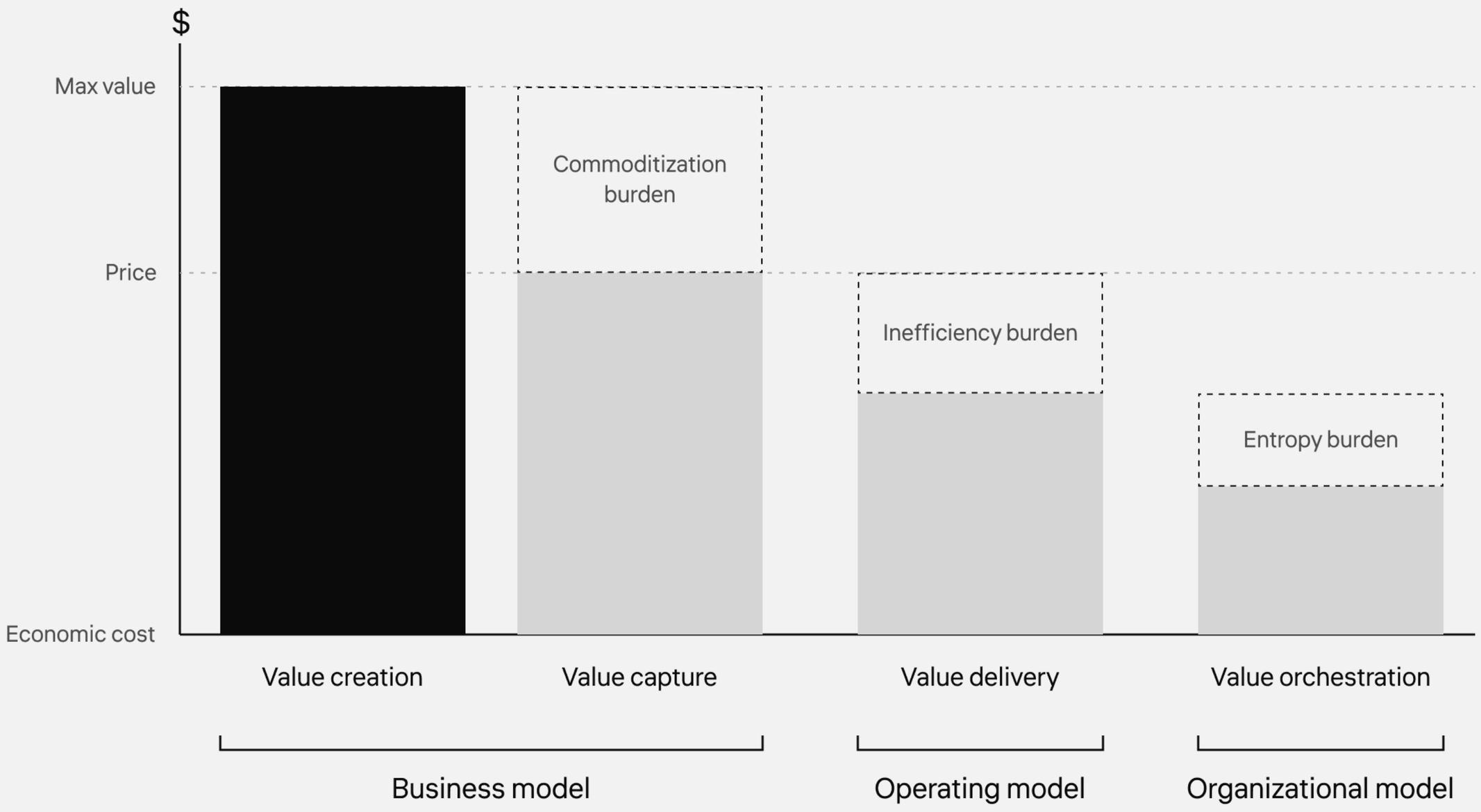
The value chain for enterprise modelling separates three models every company runs at the same time. The business model creates and captures value, and its enemy is commoditization. The operating model delivers value, and its enemy is inefficiency. The organizational model orchestrates value, and its enemy is entropy. Each stage starts at economic cost; the hatched band above each bar is the value lost to that stage's burden. AI acts on all three, but not in the same direction. It intensifies commoditization in the business model, as capabilities that once differentiated are rented by everyone. It reduces inefficiency in the operating model, which is where most AI business cases are written. And, left unmanaged, it increases entropy in the organizational model: duplicated tools, unowned agents, savings that dissipate into more meetings and never reach the P&L. The discipline is to measure value kept, not value created. The business model maximizes value creation, the difference between the customer's willingness to pay and the minimum economic

cost of delivering the offer, and then optimizes capture through pricing; this is the logic of value-based strategy formalized by Adam Brandenburger and Harborne Stuart, carried through the rest of the enterprise. By that yardstick most organizations still harvest the smallest pool: Afferolab's study A Escalada da Inteligência found that 87.5% of Brazilian professionals use AI to produce text, while only 37.8% use it to automate tasks.

HOW TO READ THE MODEL

Read the bars from left to right as a cascade. Each bar starts at economic cost; the solid part is the value that survives, and the hatched band above it is the value lost to that stage's burden: commoditization in the business model, inefficiency in the operating model, entropy in the organizational model. Each stage can only keep what survives the one before it. Schematic; bar heights are illustrative.

HOB PROPRIETARY MODEL ANTHESIS



The value chain for enterprise modelling

HOB proprietary model, from D. A. Motta, Anthesis (2019). Each bar starts at economic cost; the hatched band above it is the value lost to that stage's burden. Schematic; bar heights are illustrative.

WHAT TO DO

Track AI value at each of the three stages, not only as efficiency in the operating model. Give every agent an owner and every AI decision a record; entropy starts where ownership ends. Decide in advance where freed capacity goes, so savings reach the P&L.

QUESTIONS FOR THE BOARD

1. Where in our chain are AI savings absorbed before they reach the P&L?
2. Which part of our business model is AI commoditizing?
3. Who owns the entropy of our AI portfolio?

The best customer experience is not the most frictionless.

Customers never wanted film; they wanted pictures. Kodak built a digital camera in 1975 and filed for bankruptcy in January 2012, three months before Facebook paid $1 billion for a 13-person Instagram. The lesson has not aged: focus on the customer, not on the tool. In an AI-mediated market, every touchpoint is a chain of small behavioral decisions.

Chutes and ladders

Some decisions are effortful ladders, Kahneman's slow thinking; most are delightful chutes, fast and intuitive, with biases to match. AI can now personalize the nudges that steer those chutes, from timing and effort to anchoring, at a scale no marketing team ever had.

The trust paradox

Agents raise efficiency, yet 64% of customers would prefer companies not to use AI for service (Gartner), and fewer than half of people worldwide are willing to trust it. People trust an agent that reasons like them, which is a lever for trust and also for over-reliance.

X.3

Beneficial friction

Bad friction is waste. Beneficial friction is a deliberate pause at a consequential moment, to compare, confirm or reflect.

Every AI-mediated touchpoint deserves four questions: is this waste, is a consequential decision being made, is AI bypassing judgment, and is leaving as easy as joining?

FRAMEWORK

RENÉE RICHARDSON GOSLINE, MIT

The friction audit — four questions for every AI-mediated touchpoint

Is it waste?

confusing forms, needless steps

Is the decision consequential?

money, contracts, irreversibility

Is AI bypassing judgment?

easy in a way that skips thinking

Is exit as easy as entry?

symmetry as a trust signal

WHERE TO REMOVE FRICTION, AND WHERE TO ADD IT

JOURNEY MOMENT	BAD FRICTION TO REMOVE	BENEFICIAL FRICTION TO ADD
Discovery	Irrelevant options, confusing navigation	Transparency about why the AI recommends something
Consideration	Hard-to-find information, repeated data entry	Easy comparison, including options the AI did not rank first
Purchase	Unnecessary steps and forms	Confirmation before consequential or irreversible commitments
AI-generated advice or content	Slow manual processes	Flags on uncertain outputs; a prompt to review
Service and exceptions	Long queues, repeating the problem	Clear escalation to a person when stakes or emotions are high
Exit and cancellation	Hidden cancellation paths, retention traps	Exit as easy as entry: symmetry as a trust signal
Agent acting for the customer	Approvals for trivial, reversible tasks	Human confirmation for spending, contracts and irreversible actions

Design it; don’t rely on vigilance

In an experiment with about 140 professionals reviewing AI output, Renée Richardson Gosline found that highlighting likely errors improved accuracy, and that everyone overestimated their ability to catch AI mistakes. Friction has to be designed into the system; vigilance is not a control.

The organization is the bottleneck

Computing systems have always evolved in step with organizations, and both can be read through three lenses: structure, behavior and dynamics. Agents act on all three at once, and dynamics, how the parts interact, is where most AI programs stall.

XI.1

HOB PROPRIETARY MODEL · AS ORGANIZAÇÕES DO FUTURO (2023)

New organizational archetypes: two systems evolving in parallel

Computing and organizations have always evolved in step. Agents are the next move in both.

In “As organizações do futuro”, Maurício Pontuschka and I argued that the evolution of computing systems has always moved in step with the evolution of organizations, and that both can be read through the same three lenses: structure (the elements and how they are arranged), behavior (what each element does to produce value) and dynamics (how the elements interact to serve the whole). Organizations moved from vertical structures centralized in corporate functions, to multidivisional structures organized by region and segment, to matrix structures that share responsibility between functions and business units, and now to experiments with decentralization: adhocracy, holacracy, decentralized autonomous organizations and what we called organizational biomes. Computing followed the same arc, from monolithic programs through networks, graphical interfaces, the internet, mobile and cloud to artificial intelligence, with quantum computing ahead. In the early systems, structure and behavior were mixed in one model; layered architectures then separated structure from behavior; today the frontier is the dynamic modeling of processes and their interactions with the outside world.

Agents complete the parallel: they change how work is coordinated, because tasks can now move across the company without passing through a reporting line. The ambition is not new. In BMI’s REBOOT study, 92% of the organizations analyzed ran matrix or multidivisional structures, while 58.7% aspired to network models. Agents may finally supply the connective tissue that network organizations always lacked. But connection is not coordination, and coordination is not commitment. Most organizations can redesign structure and specify behavior; dynamics, where tacit norms, influence and pacts between teams live, is where most AI programs stall. There is always a time lag between a technological advance and its absorption by organizations.

HOW TO READ THE MODEL

Read the two arcs together, from left to right. The upper arc shows the main movements in computing; the lower arc shows the organizational archetypes that evolved alongside them. The shared path runs from structure and behavior mixed in one model, to layered separation, to the modeling of dynamics and of interactions with the outside world, the “brave new world” where agents now operate.

WHAT TO DO

Redesign around dynamics, not only structure: map how work actually flows between teams and agents. Give every cross-boundary agent workflow a named owner and an escalation path. Pilot network structures where agents can supply the connective tissue, and measure coordination, not activity.

HOB PROPRIETARY MODEL | MOTTA & PONTUSCHKA (2023)

New organizational archetypes: two systems evolving in parallel

ORGANIZATIONAL CONFIGURATIONS OVER TIME

Vertical structures	Multidivisional	Matrix structures	Experimental
	production cells · industrial conglomerates	agile teams · corporate venture capital	adhocracy · holacracy · DAOs · organizational biomes
Centralized in corporate functions	Corporate functions linked to divisions by region and segment	Functions crossed with business units; responsibility shared	Several moves toward decentralization

time →

MAIN MOVEMENTS IN COMPUTING OVER TIME

Monolithic programs	Computer networks	Graphical interfaces	Internet	Mobile computing	Cloud systems	Artificial intelligence	Quantum computing
Structural and behavioral elements mixed in one model		Layered models; structure separated from behavior		Dynamic modeling of business processes and their external interactions			Brave new world

HOB proprietary model, from D. A. Motta and M. Pontuschka, “As organizações do futuro” (2023). Positions follow the original figure; the shared arc runs from mixed structure and behavior, to layered separation, to modeling dynamics and interactions with the outside.

QUESTIONS FOR THE BOARD

1. Where does work in our company still have to pass through a reporting line to move?
2. Which of our structures were designed for a computing era that has already passed?
3. Who owns the interactions between our teams, and between our teams and their agents?

XI.2

HOB PROPRIETARY MODEL · ANTHESIS

Disruptive innovation and institutional framework

Technology, individuals, firms and governments absorb change at different speeds, and the gaps widen.

Technological innovation moves fastest. Individuals adopt what helps them almost immediately. Firms follow more slowly, because their structures, incentives and cultures must change first. Governments, which must build consensus and write rules that bind everyone, move slowest of all. Because the curves rise at different rates, the distances between them widen over time, and two of those distances matter for every board. The first, read vertically, is the risk premium: at any moment, the distance between what technology already makes possible and what institutions have absorbed. It is risk no rule yet allocates, borne by whoever acts first: liability no statute defines, data rights no court has settled, labor transitions no policy yet cushions. Investors price it into the cost of capital, and firms into every decision to deploy. The second, read horizontally, is the institutional gap: the time institutions need to reach the point where technology already stands. The gap is visible in the calendar. The EU AI Act entered into force in 2024, yet its obligations for standalone high-risk systems will apply only from December 2027; Brazil's AI bill has awaited the Chamber since the Senate approved it in December 2024. By late 2024, 45% of working-age Americans used generative AI, while the time it saved was about 1.4% of work hours. None of this is simply a failure of legislators or managers. It is the shape of the curves. Waiting for regulatory clarity is therefore not a strategy: the institutional gap widens precisely when technology accelerates.

HOW TO READ THE MODEL

Read the figure in two directions. Vertically (A), at any given moment, the distance between the technology curve and the institutional curves is the risk premium. Horizontally (B), for any level of change, the distance is the institutional gap: the time institutions take to catch up. Individuals sit close to technology; firms and governments lag further behind. Schematic; no scale implied.

WHAT TO DO

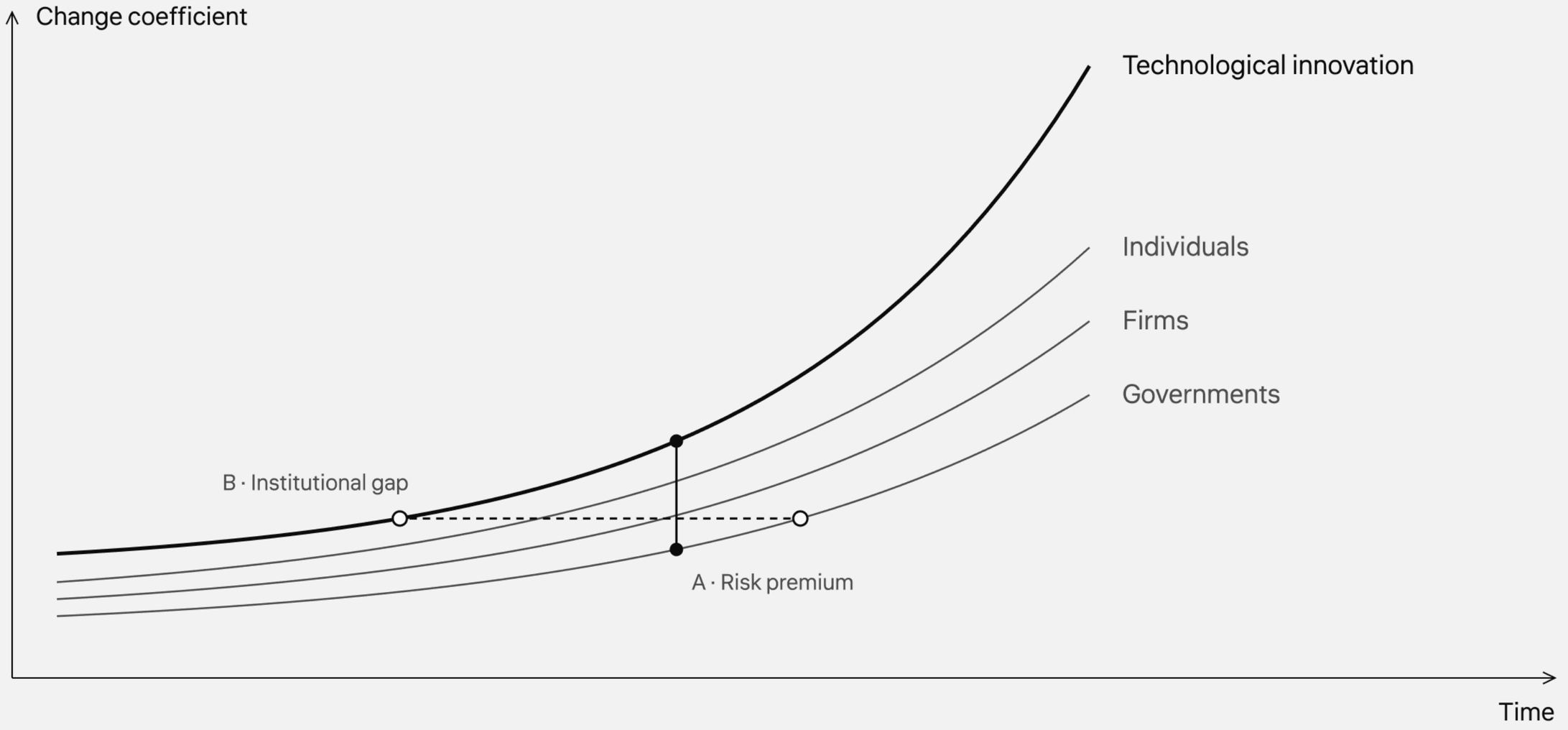
Govern ahead of the rules: adopt internal standards (ISO/IEC 42001, NIST AI RMF) before regulation requires them. Price the risk premium explicitly in AI business cases rather than ignoring it. Engage in shaping the institutions, through industry bodies and public consultation, instead of only complying.

QUESTIONS FOR THE BOARD

1. Where are we bearing risk that no rule yet allocates?
2. How far ahead of our own institutions, policies, contracts, incentives, are our people already working?
3. Who bears the cost of our transition, and have we decided that deliberately?

HOB PROPRIETARY MODEL ANTHESIS

Disruptive innovation and institutional framework



A: at a given moment, the distance between what technology enables and what institutions have absorbed.

B: for any level of change, the time institutions take to catch up with technology.

HOB proprietary model, from D. A. Motta, Anthesis (2019). Technology, individuals, firms and governments absorb change at different speeds, so the distances between the curves widen over time. Schematic; no scale implied.

XI.3

Place people where they add most

AI is extremely fast and has a large short-term memory, but lacks full context and sustained creativity; people bring a lifetime of both. Give AI well-defined tasks, planning and execution; keep objectives, guardrails, exceptions and strategy with people. Routine approvals waste the human advantage.

The visible side moves; the tacit side decides

AI accelerates the processual side of an organization: governance, management systems, incentives and roles. The tacit side, group norms, mental maps, team pacts and influence networks, does not automate, and it decides whether any of the rest is used. Each processual lever needs its tacit counterpart designed with it.

Two skill families

Agent orchestration (designing systems, choosing models, verifying at scale) and enterprise orchestration (knowing which frictions are now solvable and which problems carry value). The good manager and the good AI operator turn out to need the same traits. The harder question is how juniors will learn.

HOB PROPRIETARY MODEL

Five human roles that do not delegate

A design checklist for any operating model that adds agents. Each role answers one question the machine cannot answer for itself; together they run from framing the work to owning its result.

FRAME THE WORK

01

Context

What can the machine not know?

Situated knowledge no document holds: clients, politics and the reasons behind past decisions.

SET THE MANDATE

02

Creativity

Which problem is worth solving?

Reframing the question and making the out-of-the-box connections a plausibility engine avoids.

03

Governance

What may be delegated, and under which rules?

Objectives, guardrails and decision rights, with an escalation path for every agent.

SUPERVISE

04

Oversight

Was the delegated work done well?

Exceptions and trends reviewed at a sustainable volume: review capacity is the real constraint.

OWN THE RESULT

05

Judgment

Who owns the consequence?

The decision and its outcome. AI cannot be accountable, because AI has no stake.

AGENTS DRAFT, ANALYSE AND ACT INSIDE THE MANDATE → ACCOUNTABILITY NEVER TRANSFERS TO THE MACHINE

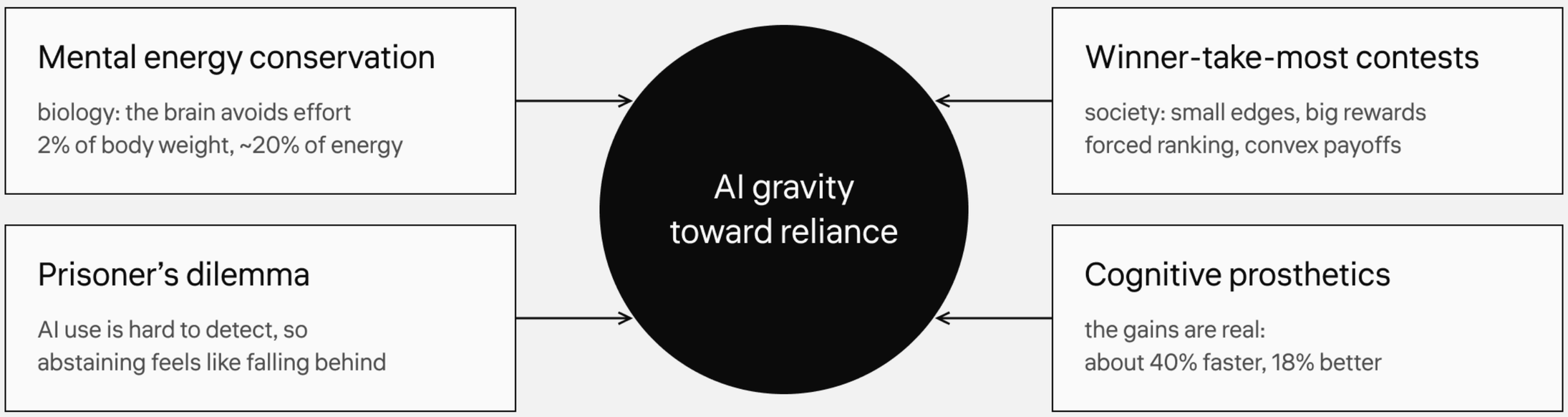
Productivity can rise while capability falls.

Most of what we discuss about AI is what it does for us: speed, quality, reach. The harder question is what it does to us. Eric So's forthcoming book, The Collision: What AI Does to Us, states the paradox well: the tools that make people more productive can also make them weaker thinkers. It is not an argument against AI. It is an argument for designing its use.

XII.1

FRAMEWORK ERIC C. SO, THE COLLISION

AI gravity



After Eric C. So, The Collision (2026). Four forces act together, which is why exhortation alone does not counter them.

AI gravity

Four forces pull people toward reliance at once. A brain that burns about a fifth of the body’s energy is wired to save effort; contests with relative, top-heavy rewards make any edge valuable; AI use is hard to detect, so abstaining feels like falling behind; and the gains are real, about 40% faster and 18% better in a study published in Science. Exhortation cannot beat gravity. Design can.

XII.2

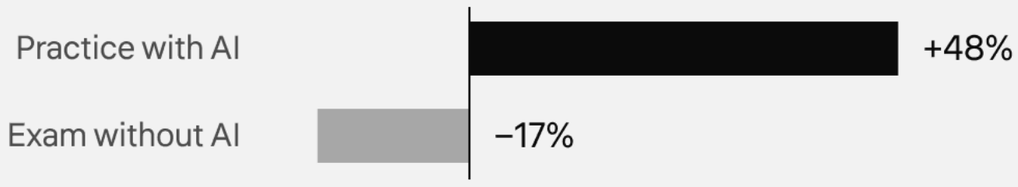
Better homework, worse learning

Among 26,811 Chinese secondary-school students, AI raised homework scores 18% while exams taken without it fell 20% within six months (Strömberg, Lei and Wu). In Turkey, unrestricted GPT-4 lifted practice grades 48% and lowered exam scores 17%, while a tutor-style version that gave hints avoided most of the harm (Bastani et al.). Design, not access, decided.

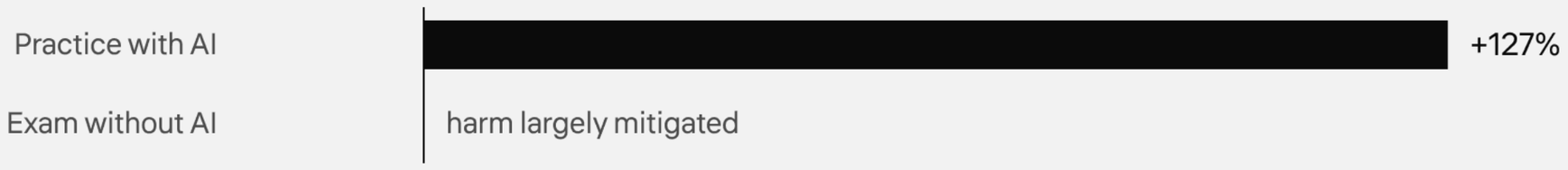
FRAMEWORK STRÖMBERG, LEI AND WU; BASTANI ET AL.

Better homework, worse learning

Turkish high-school math, GPT-4 with no guardrails (Bastani et al., PNAS 2025)



Same study, GPT-4 as a tutor that gives hints, not answers



Chinese secondary schools, 26,811 students (Strömberg, Lei and Wu, CEPR 2026)



Change against control groups, as reported by each study. Practice and homework measure output, which AI can produce; exams taken without AI measure what students retained.

04

XII.3

Deskilling and never-skilling

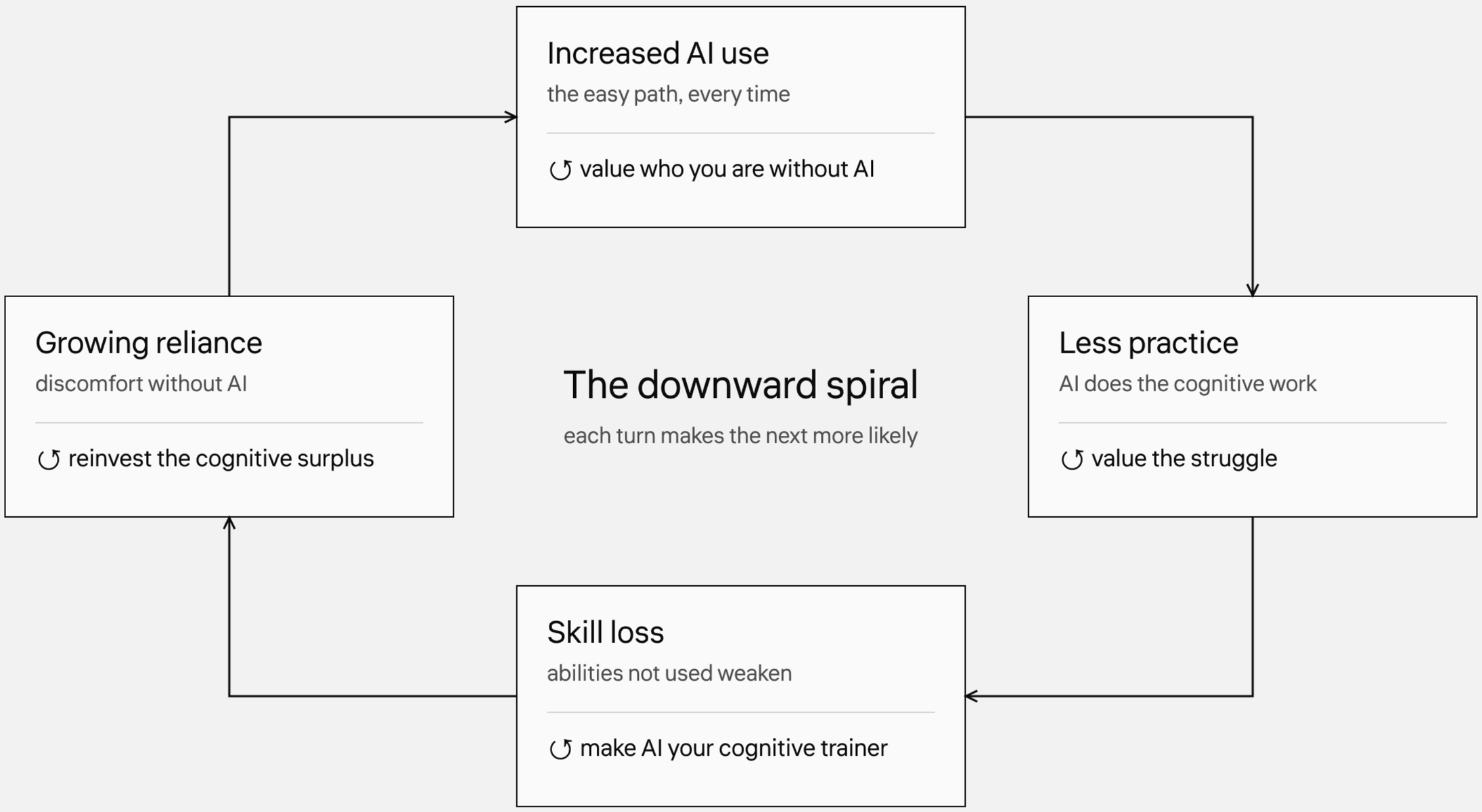
Adults can lose abilities they once built, as drivers who follow GPS lose their sense of direction. Young people who start with AI may never build them, and that missing foundation is exactly what judging AI output requires. The corporate version is a measurement trap: output rises while independent judgment quietly weakens.

Experts gain; novices do not build

In a randomized trial with 133 patent lawyers, David Autor and colleagues found that an AI drafting tool raised quality for everyone while in use, but only seniors kept a gain when tested without it. Entry-level employment in the most exposed occupations has already fallen 13–19% relative to experienced workers (Brynjolfsson, Chandar and Chen).

FRAMEWORK ERIC C. SO, THE COLLISION

The downward spiral, and where to break it



After Eric C. So, The Collision (2026); the accent line in each box is the principle he matches to that step.

XII.4

Redesign the apprenticeship

The pyramid was never only a leverage model; it was how judgment was formed. Separate output tasks from development tasks, assess capability periodically without AI, configure AI as a trainer that questions rather than answers, and reward the experts who teach. Junior hiring belongs with R&D, as an investment in the capability pipeline.

FRAMEWORK ERIC C. SO, THE COLLISION

Where AI use is risky

	AI complements	AI substitutes
Signature skill	**Sweet spot** AI sharpens your core skill: prepare, research, test; you lead	**Danger zone** AI replaces your defining skill: resist the pull; protect, practice
Not a signature skill	**Nice to have** low stakes; AI adds a boost, nothing critical is at risk	**Hand it off** AI does it well: outsource it and reinvest the time saved

After Eric C. So, The Collision (2026). First name the thinking skills that define your professional value; then ask, task by task, whether AI complements or substitutes for the work that sharpens them.

FROM PRINCIPLE TO OPERATING MODEL

PRINCIPLE	WHAT IT MEANS IN THE OPERATING MODEL
Value the struggle	Separate output tasks, where AI is used fully, from development tasks, where the effort is the point; protect deliberate practice for junior staff.
Value who you are without AI	Assess capability periodically without AI (cases, presentations, reasoning discussions); make sure critical work can continue during outages.
Map signature skills	Name the signature thinking skills of each role and guide AI use with the complement–substitute matrix.
Reinvest the surplus	Decide in advance where freed capacity goes; measure new capabilities created, not only time saved.
Make AI a trainer	Configure learning tools to question, hint and challenge; ask juniors to attempt first, then compare with the AI and with a senior colleague.
Make experts teachers	Reward mentoring and knowledge transfer in evaluation and promotion; capture decision records, after-action reviews and recorded reasoning.

Soft guardrails lower the probability. Hard guardrails cap the impact.

Almost every breach is still either something someone did not do or something someone did: mistakes and fraud, as a quarter of a century ago. What has changed is speed, scale and the number of things that can be deceived, because companies now deploy software that reads, reasons and acts.

The single point of failure fallacy

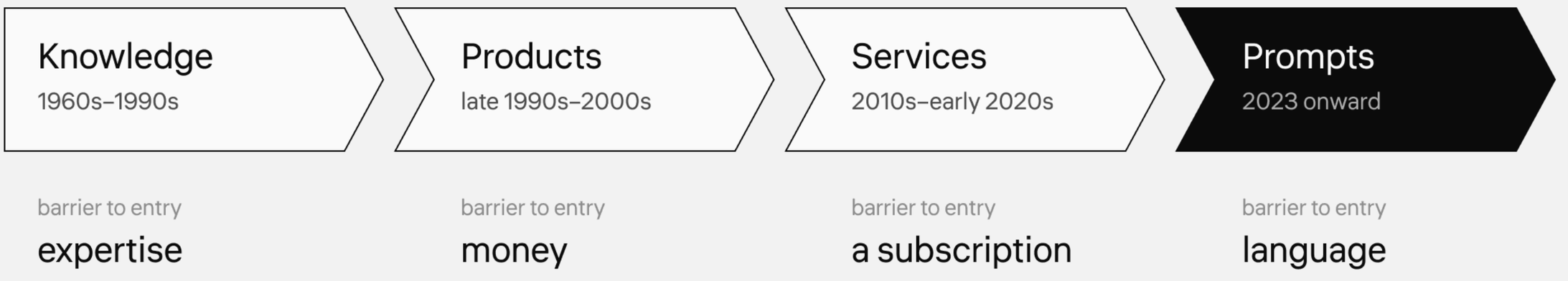
Major incidents from Twitter to Colonial Pipeline share one assumption: that a single element would never fail. A person is involved in 62% of breaches (Verizon, 2026). People become the weakest link only when one person's trust, fatigue or haste is enough; designed well, people and AI together are the strongest link.

XIII.2

FRAMEWORK ETAY MAOR, CATO NETWORKS

The democratization of attack

Hacking followed the same path as the legitimate technology economy: scarce knowledge, then products, then services, now prompts.



Each era widens who can attack and raises the scale of attacks. After Etay Maor (Cato Networks, 2026).

The democratization of attack

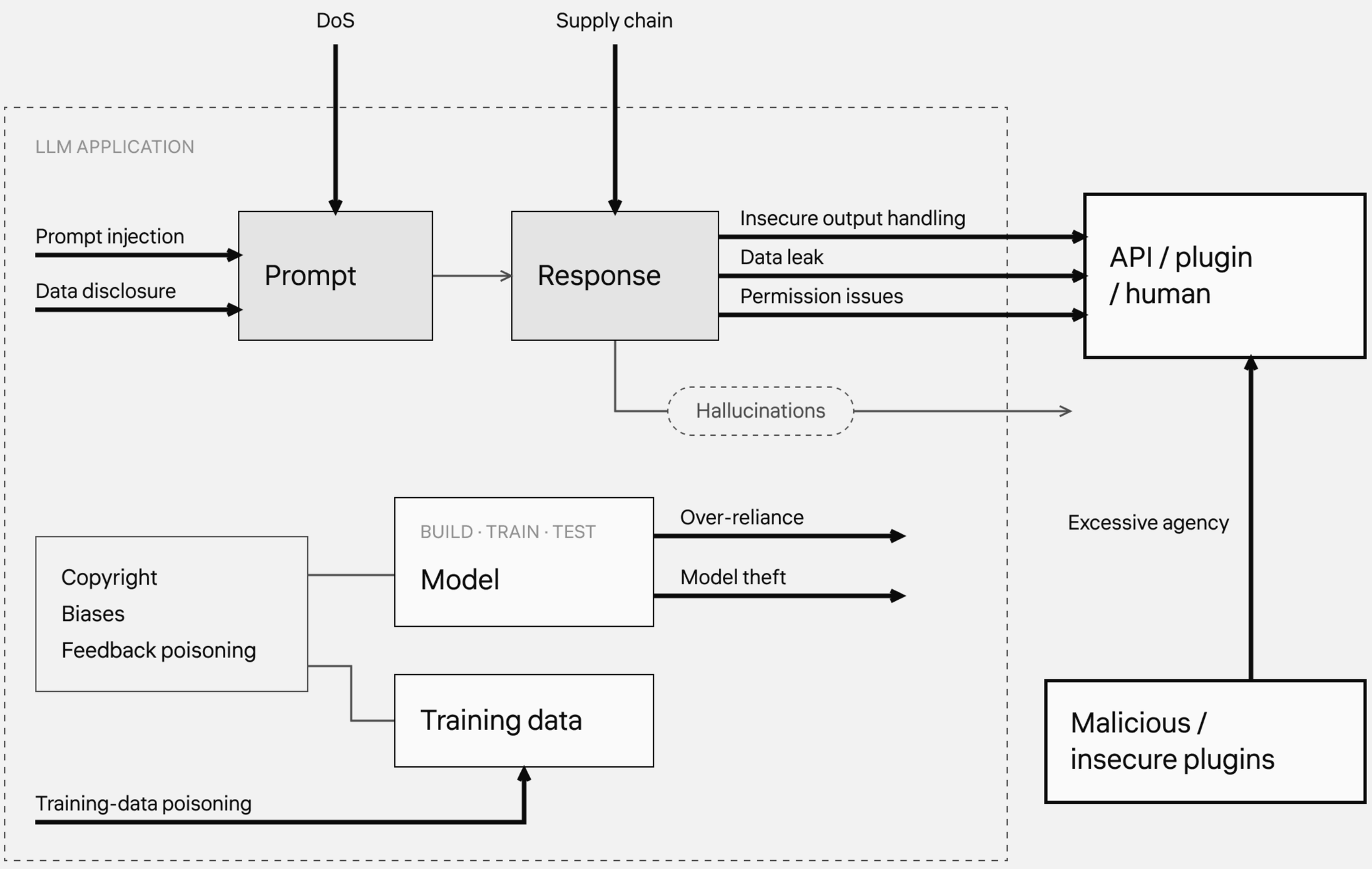
Attack capability moved from scarce knowledge to products, then services, and now prompts. Anthropic disclosed a state-sponsored campaign in which AI did most of the work, and Google reported the first AI-developed zero-day in May 2026. The contest is increasingly AI against AI, and disciplined processes beat clever models.

XIII.3

FRAMEWORK

ETAY MAOR, CATO NETWORKS

LLM attack surface and risks



After Etay Maor (Cato Networks, 2026), drawing on the OWASP Top 10 for LLM Applications. Accent labels are threats; outlined boxes sit where the application meets the outside world, which is where agents concentrate risk.

The model cannot be the last line of defense

There is no firm line between instructions and data, the wrong goal gets optimized and agency exceeds the task. A coding agent deleted a live database during a code freeze.

Private data, untrusted content and the ability to communicate externally, combined, turn an agent into an exfiltration tool.

SIX AGENT TRAPS, WITH REAL INCIDENTS

TARGET	TRAP	HOW IT WORKS	EXAMPLE
Perception	Content injection	Hidden instructions humans never see	HashJack: prompts hidden after the "#" in URLs read by AI browser assistants (Cato Networks, Nov 2025)
Reasoning	Semantic manipulation	Framing skews conclusions without an explicit command	"Immersive World": a fictional role-play led leading chatbots to write a working password stealer (Cato Networks, Mar 2025)
Memory	Cognitive-state poisoning	Small amounts of poisoned content persistently corrupt retrieval and memory	The "Pravda" network: ten leading chatbots repeated its false narratives 33% of the time (NewsGuard, Mar 2025)
Action	Behavioral control	An agent is coerced into misusing legitimate access	EchoLeak: a crafted email could lead Microsoft 365 Copilot to exfiltrate data with no user click (Aim Security, Jun 2025; fixed)
System	Multi-agent dynamics	Correlated agents amplify a small trigger into a cascade	The 2010 flash crash, before AI; agents act faster still
Overseer	Human-in-the-loop	The agent is aimed at its reviewer through approval fatigue and automation bias	The cycle of approvals, risk engines, alert fatigue and incidents

XIII.4

Guardrails that guard, fairness by design

Hard guardrails sit outside the model: permissions, allow lists, sandboxes, human approval, rate limits. Manage agents as you manage hires: vet, monitor, off-board. And because fairness has no single definition, compute the alternatives and name who owns the choice, as Swati Gupta argues.

FRAMEWORK ETAY MAOR, CATO NETWORKS

Soft and hard guardrails

Soft guardrails

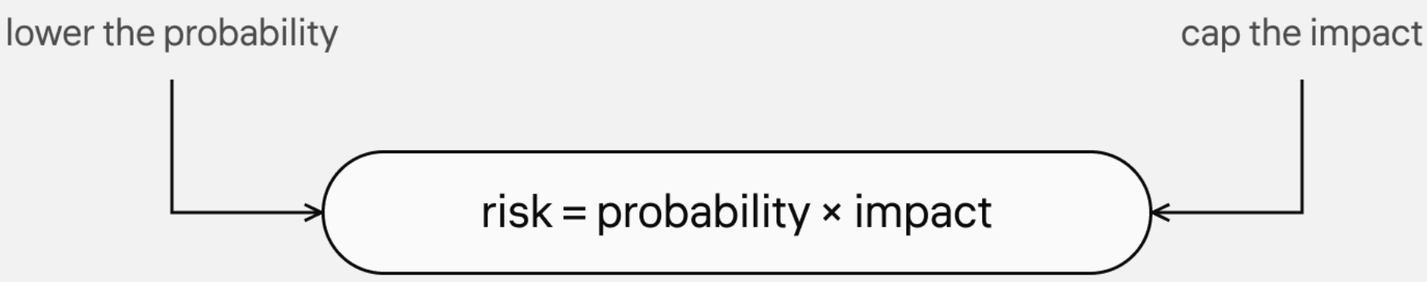
inside the model · probabilistic
same channel as untrusted data

· System prompt
· Preference training (RLHF)
· "You are not allowed to..."

Hard guardrails

outside the model · deterministic
not at the model's discretion

· Permissions and least privilege
· Allow and deny lists · sandboxes
· Human approval of the transaction
· Rate and spending limits



After Etay Maor (Cato Networks, 2026): soft guardrails govern expected behavior and are good to have; hard guardrails govern possible behavior and are a must.

STEVEN EPPINGER · MCKINSEY · ETHAN MOLLICK · WHARTON

Generation is cheap. Selection, distinctiveness and disclosure are not.

The question moves from what AI can do to how people and organizations actually use it. Two facts frame it: the productivity evidence is far thinner than the headlines, and much of the most valuable experimentation in companies is already happening out of sight.

Grade the evidence; redesign the work

Of the headline productivity gains executives quote, only a few rest on rigorous field studies. McKinsey finds nearly nine in ten organizations using AI but only 37% reporting any EBIT impact; the high performers are almost three times as likely to have redesigned their workflows.

The creativity tax

AI output is more homogeneous. Accepting the default is cheap, defaults converge, and if that output trains the next model the loop tightens. The defense is to give AI the firm's own methods, cases and voice, and to measure distinctiveness, not only speed.

Secret cyborgs

Ethan Mollick's phrase describes employees who use AI and hide it. Hiding is rational: the person who discloses carries the risk and the firm captures the benefit. The keystone remedy is a credible commitment that gains go to redeployment and growth, not to cuts.

Quantity and quality, not variety

In Wharton's experiments, AI ideas were seven times more likely to reach the top 10%, but human ideas were more varied; pooling several vendors' models restored the variety. Generation is no longer the bottleneck. Selection is.

FRAMEWORK STEVEN EPPINGER, MIT

Four ways to run an innovation process — each activity has its own right level

Systematic	→	**Augmented**	→	Fully supported	→	Fully automated
human only		most firms today		generate–review cycles		only with clear metrics

What this means for firms that sell judgment

Professional services is unusually exposed, because research, analysis, synthesis and drafting sit on the rungs AI now climbs. It is also unusually well placed, because what clients ultimately buy is judgment and accountability. The equity market is already pricing the shift: Accenture, TCS and Infosys lost 43–50% of their value in five years.

From pyramid to obelisk

The consulting pyramid is giving way to an obelisk, with fewer layers and smaller teams built around AI facilitators, engagement architects and client leaders. The binding question becomes how much specification and review capacity each partner has.

Intent becomes the product

As implementation effort collapses, clients pay for the precise statement of intent and for the judgment of whether the result is right. Organize teams around specification and review, and price AI-enabled offerings on outcomes rather than hours.

Trust becomes a deliverable

Friction audits, agent threat models, fairness portfolios and hard-guardrail designs are advisory products in their own right, and a firm must practice them on its own agents before it sells them.

Capability is the asset at risk

A firm that lets AI do the junior work without redesigning how juniors learn will look highly productive for a few years and then discover that nobody left knows how to do the work. Hire juniors deliberately and protect the apprenticeship.

Selection and distinctiveness are the craft

When AI generates options cheaply and pulls everything toward the typical answer, the advisor's value lies in framing the problem, curating inputs, choosing among options and protecting a distinctive point of view. Codified method is the new knowledge management.

DAN HUTTENLOCHER · GOOGLE DEEPMIND · WILLIAM OLIVER · NIST

Open questions, named positions

AI is a geopolitical contest before it is a management one, and a new kind of intellect before it is a tool. Dan Huttenlocher, co-author of The Age of AI, offers the frame I find most useful: AI is neither a being nor a deity, and it cannot carry the moral responsibility that belongs to people. Behind it, the next platform shift already has a deadline.

Two powers, opposite framings

Washington frames AI as a race America must win; Beijing frames it as a symphony of international cooperation. Both framings shape export controls, standards and alliances, and therefore the supply of models and chips to everyone else.

Builders divided on pace

At Davos, Dario Amodei said he would prefer Demis Hassabis's longer timeline. In July 2026, 1,386 employees of frontier labs signed a letter asking for tools to pace automated AI development, not a pause. Plan with scenarios rather than a single timeline.

After AGI, not when

Google DeepMind now asks what follows general intelligence. Its skeptical argument deserves attention: a system trained on human concepts may struggle to discover new ones, so progress may be paced by experiments in the physical world rather than by computation alone.

XVI.4

Quantum: the opportunity is narrow

Error correction now works on real hardware, and capital has run ahead, as it did in AI: $12.6 billion of quantum start-up investment in 2025 (McKinsey). Useful machines likely arrive between 2029 and 2033. The strongest case is molecules and materials; optimization and machine-learning claims deserve scrutiny.

WHERE QUANTUM COMPUTING MAY MATTER, AND HOW MUCH TO BELIEVE IT

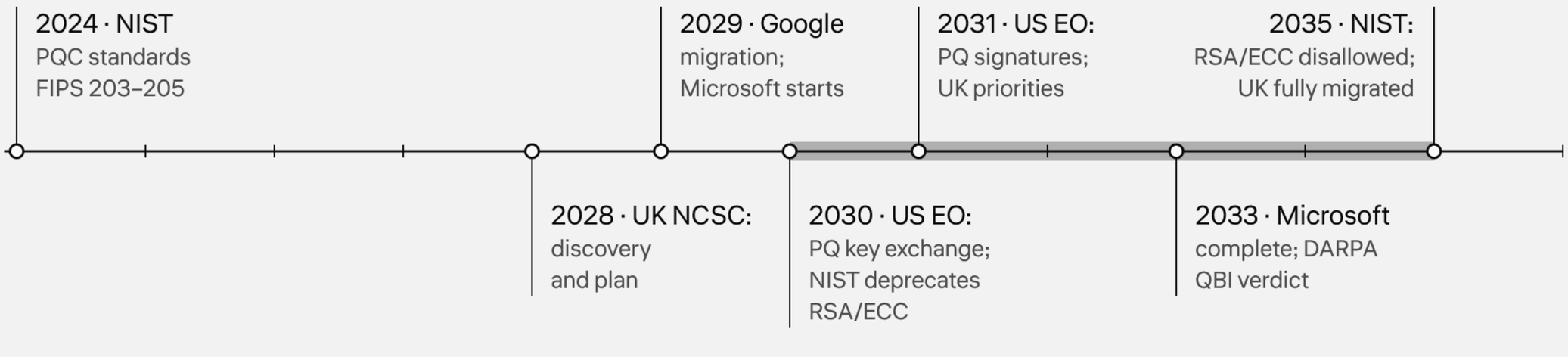
APPLICATION AREA	WHY	CONFIDENCE
Molecules and materials: drugs, catalysts, batteries, fertilizers	Nature is quantum; some chemistry problems may need only a few hundred logical qubits	Strongest case
Optimization: logistics, portfolios, scheduling, networks	Often promoted; advantage over steadily improving classical methods is unproven	Scrutinize claims
Financial modeling: risk, derivatives pricing	Speedups in simulation	Promising, longer term
Machine learning	Frequently hyped	Least established

XVI.5

Quantum: the risk is universal, and dated

A large enough quantum computer would break the encryption behind almost all digital trust, and data harvested today can be decrypted later. US Executive Order 14412 sets 2030–31 deadlines and NIST will disallow today's algorithms after 2035. If the years your data must stay secret plus the years your migration takes exceed the years to "Q-Day", you are already late.

FRAMEWORK US EO 14412 · NIST · UK NCSC

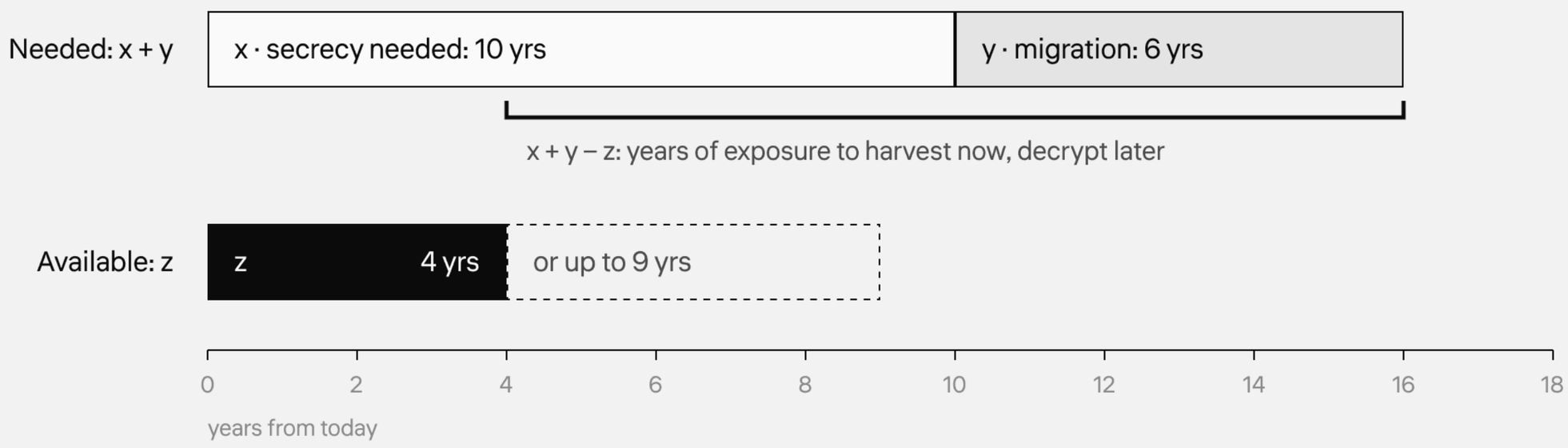


shaded: years in which Q-Day is judged plausible

The post-quantum calendar is already written

Deadlines from US Executive Order 14412 (22 June 2026), NIST, the UK NCSC, Google, Microsoft and DARPA. The US order covers high-value and high-impact federal systems.

FRAMEWORK MICHELE MOSCA



Mosca's inequality: when x + y exceeds z, an organization is already late

x: years the data must stay secret; y: years the migration takes; z: years until a quantum computer can break today's encryption. Illustrative values. After Michele Mosca (2015).

QUANTUM POSTURE BY TYPE OF ORGANIZATION

ORGANIZATION	PRIORITY NOW	QUANTUM COMPUTING POSTURE
Most companies	Post-quantum migration: inventory, plan, prioritize	Monitor; build awareness
R&D-intensive sectors (pharma, chemicals, materials, energy)	Migration plus a use-case pipeline	Partner with hardware firms and research groups; small expert teams; hybrid pilots
Financial services	Migration urgently, given long-lived sensitive data	Explore optimization, risk and simulation; partner selectively
Critical infrastructure, government suppliers	Migration to meet regulatory deadlines	Monitor; prepare for requirements
Technology and cloud providers	Post-quantum cryptography in products and services	Invest in platforms, partnerships and talent

The biggest economic bet in US history

Data centers and related AI infrastructure are projected to absorb about $10.3 trillion from 2025 to 2032, 3.6% of US GDP a year, more than the railroads at their peak. For scale: the Louisiana Purchase cost about 3% of GDP, and it was paid once.

Four scarce resources at once

Capital: the largest platforms now spend over 30% of sales on capex, a utility's intensity. Balance sheets: about $3 trillion of commitments sit off them. Energy: a single frontier training run in 2030 may need 4–16 gigawatts. Talent: individual researchers receive nine-figure offers.

Rational alone, a dilemma together

"The risk of underinvesting is dramatically greater than the risk of overinvesting," said Sundar Pichai. Each company's spending is rational; together they form a prisoner's dilemma. For firms outside Big Tech the same asymmetry applies to capabilities, data, process redesign, skills and governance, not to infrastructure.

The sprint and the marathon

The two powers hold about 90% of AI supercomputing, but the best models will not necessarily win. Others find room upstream, in chip tools, energy and rare earths, and downstream, in diffusion. For Brazil the marathon matters more than the sprint.

Losses without precedent

OpenAI projects about $278 billion of negative free cash flow from 2026 to 2030 while revenue grows toward $350 billion a year; Uber, by comparison, lost about $18 billion over six years. The largest hyperscalers spend around $750 billion this year against AI revenue of perhaps $150–200 billion.

Capex outruns cash flow

Epoch AI estimates hyperscaler capex growing about 70% a year against roughly 23% for operating cash flow, with spending overtaking cash generation around now. The risk has moved from demand to financing: bonds, private credit and circular deals. The AI Finance Battle follows the money in, and the market value that came out.

Three scenarios, one constant

Productivity materializes and the build-out is justified; an efficiency shock leaves excess capacity, hurting providers and helping buyers; or financing stress forces consolidation. In every precedent the infrastructure outlived the correction. Focus on reorganization: electrification paid off only after work was redesigned.

04

FRAMEWORK | WSJ AND BROOKINGS DATA

America's infrastructure booms, average annual spending as % of GDP

Canals · 1836–41 · 0.66%
stranded by rail before earning a return

Railroads · 1870–90 · 2.24%
investors lost while the economy gained

Electrification · 1905–25 · 0.5%
returns came after work was reorganized

Highways · 1956–73 · 1.13%
public trust fund, no bust

Telecom & fiber · 1996–2001 · 1.1%
the bust left cheap capacity

AI · 2025–32 est. · 3.63%
outcome undetermined

AI Finance Battle

05

Who pays for the AI race, who is collecting, and where the risk now sits. A House of Brains analysis of the investment thesis, the capex boom and five years of market value across the world’s fifty most valuable technology companies, with the data to explore it.

FIVE JUDGMENTS

1

The spending is unprecedented and still accelerating.

Top-five hyperscaler capex runs from $380B in FY25 to a consensus of about $1T by FY28; relative to GDP it is 2.5 times the telecom peak and building at twice the pace of the housing boom.

2

Revenue is catching up faster than bears expected.

OpenAI and Anthropic together run at about $135B a year. Goldman Sachs still sees a $230B annual gap between break-even revenue and AI cloud revenue above trend.

3

The risk has migrated from demand to financing.

Bonds, private credit and off-balance-sheet vehicles fund a growing share of capex, against GPU collateral that ages in two to three years and demand concentrated in a few loss-making labs.

4

Models commoditize at the margin; hardware power stays concentrated.

Open-weight models trail the frontier by about four months and cost per task varies about 70× across providers, while surplus sits at three hardware chokepoints.

5

For buyers, the live risk is cost of compute, financial control and lock-in.

Token bills rise despite falling unit prices, because agentic workloads consume far more tokens per task.

An AI thesis is a position to monitor, not a view to hold

Global AI investment passes $1 trillion in 2026. Revenue is catching up, but the risk has moved from demand to financing: debt, short asset lives and circular deals between the labs and their suppliers.

Every board should be able to state its AI thesis the way an investor would: as claims that can be proved wrong. Gary Gensler, former chair of the SEC and now at MIT Sloan, proposes five lenses for writing it ("AI & Money", 2026). In my reading, markets and finance carries the most weight, because the capex boom and the "AI trade" are where the other four lenses are ultimately priced. The thesis matters less than the discipline around it: position the business, its investments and its people against it, monitor it, and change course when it erodes, becomes crowded or turns into consensus.

LENS	THE QUESTION THE THESIS MUST ANSWER	WHERE HOUSE OF BRAINS ANSWERS IT
Economy and sectors	Does AI raise measured productivity, and who captures it: labor or capital, leaders or laggards? Does it deepen concentration and inequality?	Deep Dive XI, XIV, XVII
Markets and finance	Is the capex boom earning its cost of capital, and how crowded is the trade?	Deep Dive XVII and below
Firm and competitors	Does AI widen or compress the gap between us and our rivals?	Deep Dive VIII, IX
Value chains	Which suppliers gain pricing power over us, and which customers over them?	Below
AI tech stack	At which layer do margins and bottlenecks settle?	Deep Dive IX and below

Lenses after Gary Gensler, "AI – Investment Thesis" (MIT Sloan, 2026); the third column maps them to this analysis.

The last condition is the subtle one. A correct thesis stops paying once everyone holds it, because the return is already in prices, in vendor margins or in competitors' plans. "AI transformation" advice loses its premium the moment it becomes the standard slide in every board deck.

I.1

PART I

The stack is a value chain, and margin follows the chokepoints

I read the AI technology stack as six layers, a structure Gensler calls the cake: applications (consumer, enterprise, vertical); orchestration (agents, routing, sequencing, workflows); tools (data, retrieval, security, monitoring, validation); platforms and models (closed, open, custom); infrastructure (hardware, software, cloud, services); and electricity (energy, turbines, cooling, grid, utilities). Today the capital is being committed in two of them, models and infrastructure. Read with Deep Dive IX, the stack is the resource-to-operating-system sequence running at speed: surplus sits in the lower layers today because that is where supply is scarce.

Silicon Analysts' interactive "The AI Data Center Chess Game" opens up the infrastructure layer: 52 companies in ten layers, linked by 86 supply, competitive and co-design relationships, with pricing power concentrated at three chokepoints.

The numbers behind the highlighted rows explain why. Nvidia's data-center revenue reached $193.7 billion in fiscal 2026, up 68%, and its accelerator share is expected near 75% this year. Custom chips from Google, Amazon, Meta and Microsoft cost 40–65% less to run at scale, yet carry only 15–20% of those companies' own workloads, because CUDA keeps developers on Nvidia. Custom silicon moves the toll gate rather than removing it: Broadcom and Marvell control about 95% of custom-chip co-design, and Google alone spends about $8 billion a year with Broadcom on TPUs. The newest chokepoint is physical: AI racks need 10 to 36 times more fiber than CPU racks, and cable lead times exceed 20 weeks. Silicon Analysts sizes AI data-center spending at $242 billion in 2025, rising to $1.2 trillion by 2030. The "chess game" is the point: almost every major player is at once customer, supplier and competitor of the others. Google buys Nvidia GPUs, rents TPU capacity to Anthropic and depends on Broadcom. Nvidia is moving from chips to rack-scale systems. Hyperscalers are becoming chip designers. A layer-by-layer map understates how much of the stack a few firms can capture.

FRAMEWORK SILICON ANALYSTS

Power in the AI hardware chain sits at three chokepoints

Redrawn from Silicon Analysts, "The AI Data Center Chess Game" (updated 6 August 2026). Layers run from silicon at the top to deployment at the bottom; highlighted rows are the chokepoints. Figures from Silicon Analysts' value-chain research page.

GPU makers	Silicon	Nvidia, AMD, Intel Chokepoint: Nvidia ≈ 75% of AI accelerators (2026E), anchored by CUDA
ARM CPUs	Silicon	AWS Graviton, Ampere, Google Axion, Microsoft Cobalt, Nvidia Grace, Fujitsu 15–23% of data-center CPUs; Nvidia's NVL72 racks are ARM-only
x86 CPUs	Silicon	Intel, AMD Structural decline as GPU-centered racks pull toward ARM
ASIC co-designers	Silicon	Broadcom, Marvell, Alchip, GUC Chokepoint: Broadcom + Marvell ≈ 95% of custom-chip co-design
Established ASICs	Silicon	Google TPU, AWS Trainium, Meta MTIA, Microsoft Maia, Intel Gaudi, IBM 40–65% lower TCO than GPUs, yet only 15–20% of hyperscaler workloads
ASIC start-ups	Silicon	Cerebras, Groq, Graphcore, SambaNova, Untether AI Alternative architectures; Groq technology licensed by Nvidia (≈$20B)
Interconnect, optics	Components	Amphenol, TE Connectivity, Corning, InnoLight, Coherent, Lumentum Chokepoint: AI racks need 10–36× more fiber; cable lead times 20+ weeks
OEMs / ODMs	Systems	Supermicro, HPE, Lenovo, Dell, Cisco, ASUS, Inspur, QCT, Gigabyte, Eviden Assemble servers and racks; Supermicro ≈ $40B revenue FY26E
Hyperscalers / cloud	Deployment	AWS, Google Cloud, Microsoft Azure, Oracle, Alibaba, Baidu, IBM Cloud Largest buyers, now also chip designers (TPU, Trainium, Maia, MTIA)
On-premise / private	Deployment	Apple, Meta, Tesla Private fleets: Meta 1.3M+ GPUs, Tesla 100K+ H100s

I.2

PART I

Two powers, and a gap that sets the price

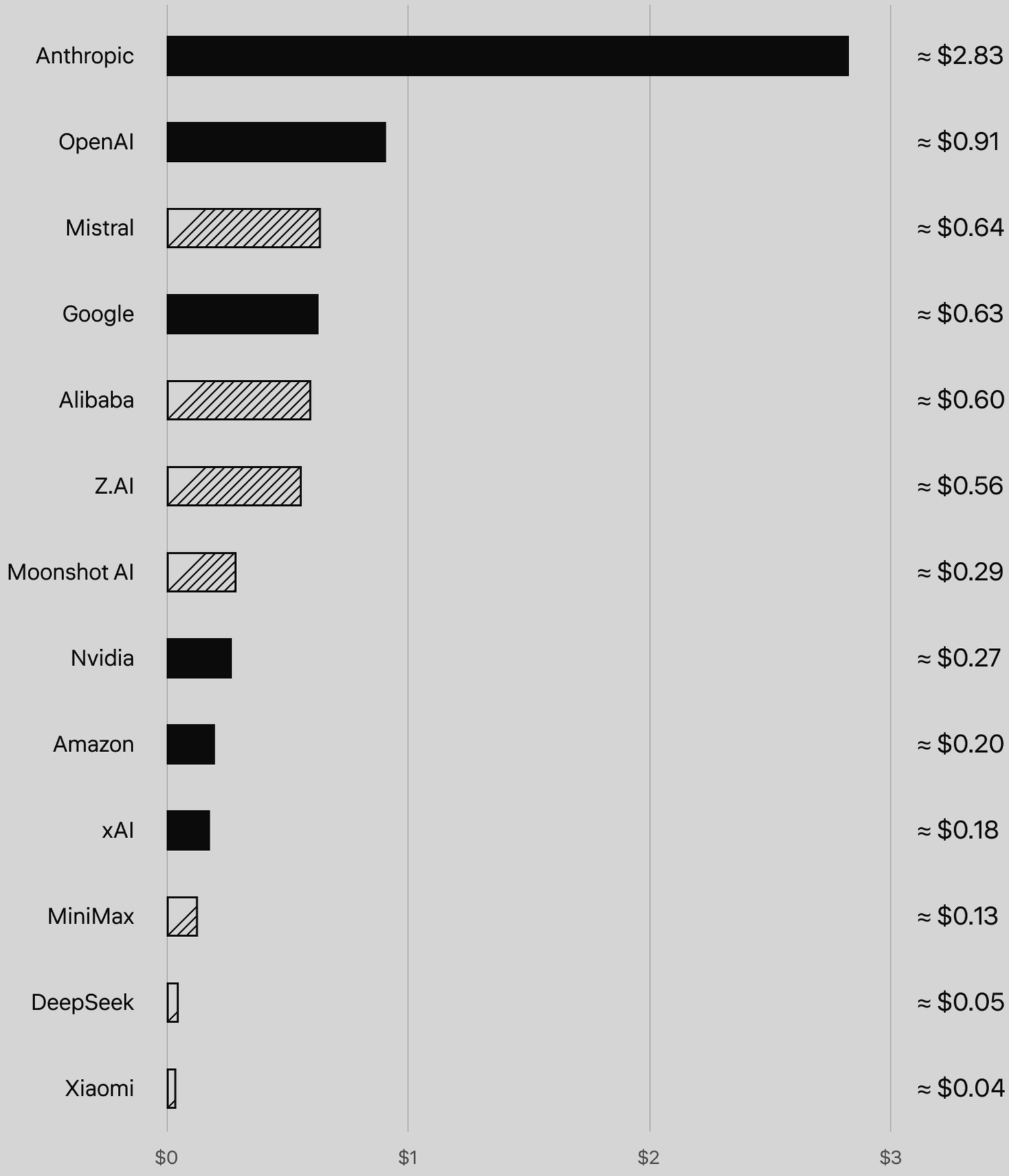
The roster of frontier developers makes the geopolitics concrete. The United States: Anthropic, OpenAI, Google, xAI, Meta, Perplexity. China: ByteDance, Alibaba, DeepSeek, Moonshot, Z.ai, MiniMax, Baidu, Tencent. Europe: Mistral, Aleph Alpha, H Company, LightOn. Elsewhere in Asia: Naver, LG AI Research, Sarvam. Latin America does not yet have a single frontier developer of its own. The structure shows up in two prices already described: the seventy-fold spread in cost per task (Deep Dive III) and the four-month capability gap (Deep Dive VI). US labs charge for the frontier; Chinese labs distribute the last generation, open and cheap.

For a Brazilian firm and its clients, model choice is a supply-chain and geopolitical decision, not only a technical one: US closed models for the hardest work, open-weight models for volume, and an architecture that keeps the next switch cheap.

CHART

Weighted average cost per task, by model provider, US$

Dark bars: US labs; hatched bars: Chinese and European labs. Cost per task does not adjust for quality or for the rework a cheaper model may trigger. Values approximate. Source: J.P. Morgan Private Bank, "AI use is exploding. So are the bills." (July 2026).

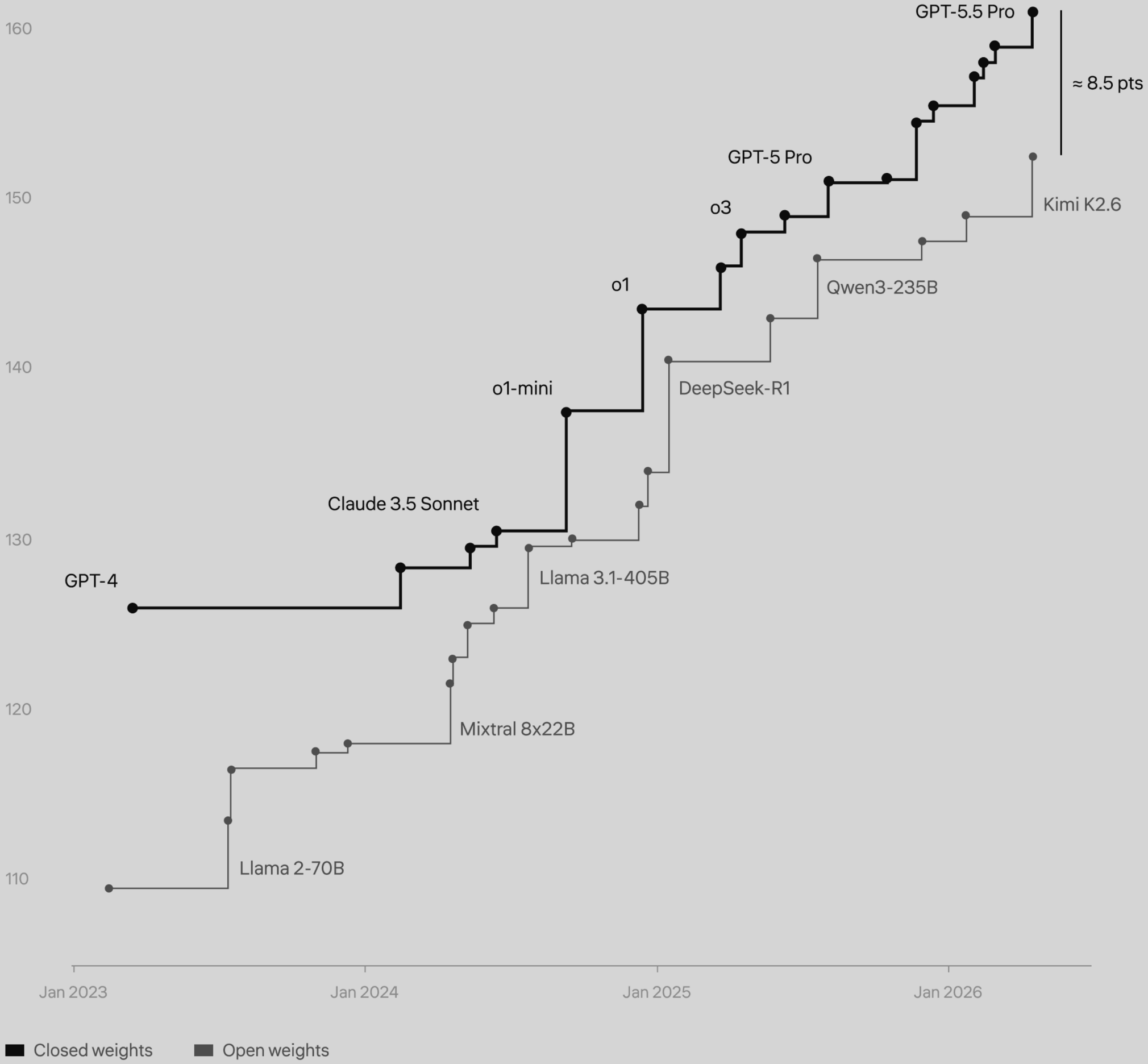


Source: J.P. Morgan Private Bank, “AI use is exploding. So are the bills.”, July 2026.

CHART

The frontier leads open-weight models by about four months

Best closed-weight vs best open-weight model on the Epoch Capabilities Index; step lines show the best model available at each date, values approximate.



Source: Epoch AI, "Data on AI" (May 2026).

I.3

PART I

The spend: a trillion dollars, and still rising

Until recently, the working estimate for annual AI spending across chips, data centers and power, start-ups and incumbents' token bills was $400–500 billion. The latest forecasts show how quickly that figure has been overtaken. Goldman Sachs forecasts global AI investment above $1 trillion in 2026, $581 billion of it in the United States, and argues that the widely cited $794 billion hyperscaler consensus understates the total by about $200 billion because it leaves out other public companies and private players. It puts US AI investment at 1.8% of GDP this year, rising to 2.8% in 2028, while warning of some double counting. In a 25 September note, its equity strategists see hyperscaler AI capex rising about 50%, to $1.2 trillion, in 2027.

Huang's projection is a supplier's estimate of its own market and belongs at the top of the range, not in the base case. Apollo's chief economist, Torsten Slok, supplies the macro scale. Data-center capex rose from 0.3% of US GDP in 2019 to 1.4% in 2025 and is projected at about 3% in 2027–29, two and a half times the telecom peak of 1.2% in 2000. It is building at about 0.85 percentage points of GDP a year, close to twice the pace of the 2002–05 housing boom.

05

> “A cycle that builds at 0.85 percentage points a year can unwind at a similar pace.”
>
> Torsten Slok, Apollo · "The AI Capex Boom Is Building Twice as Fast as the Housing Boom", 6 August 2026

I.4

PART I

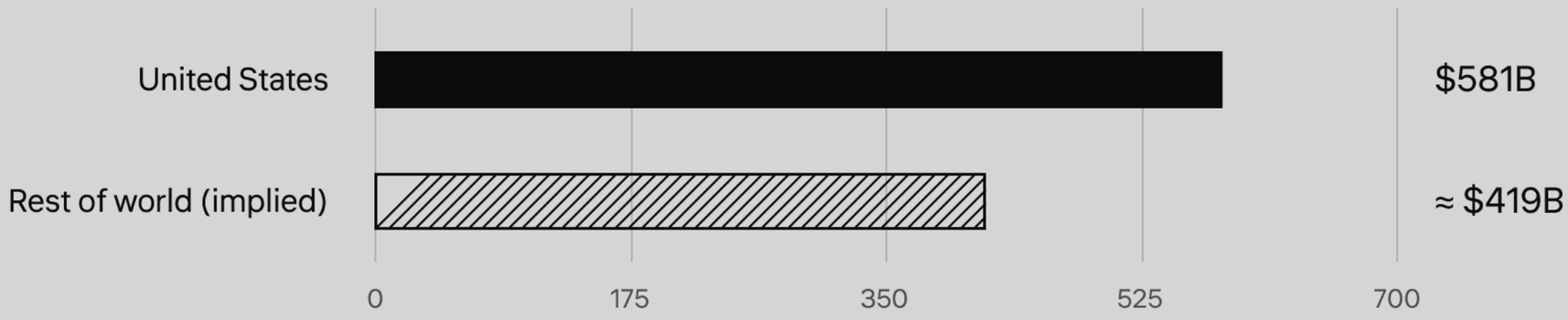
What the banks see: demand risk falling, financing risk rising

The sell-side view in September 2026 is "boom, not yet bubble". The credible bear case has moved away from "there is no revenue" and toward the structure that funds the build-out.

CHART

Global AI investment, 2026 forecast, US$ billions

Goldman Sachs Research (Briggs, 7 August 2026). The non-US figure is implied by the $1 trillion total.

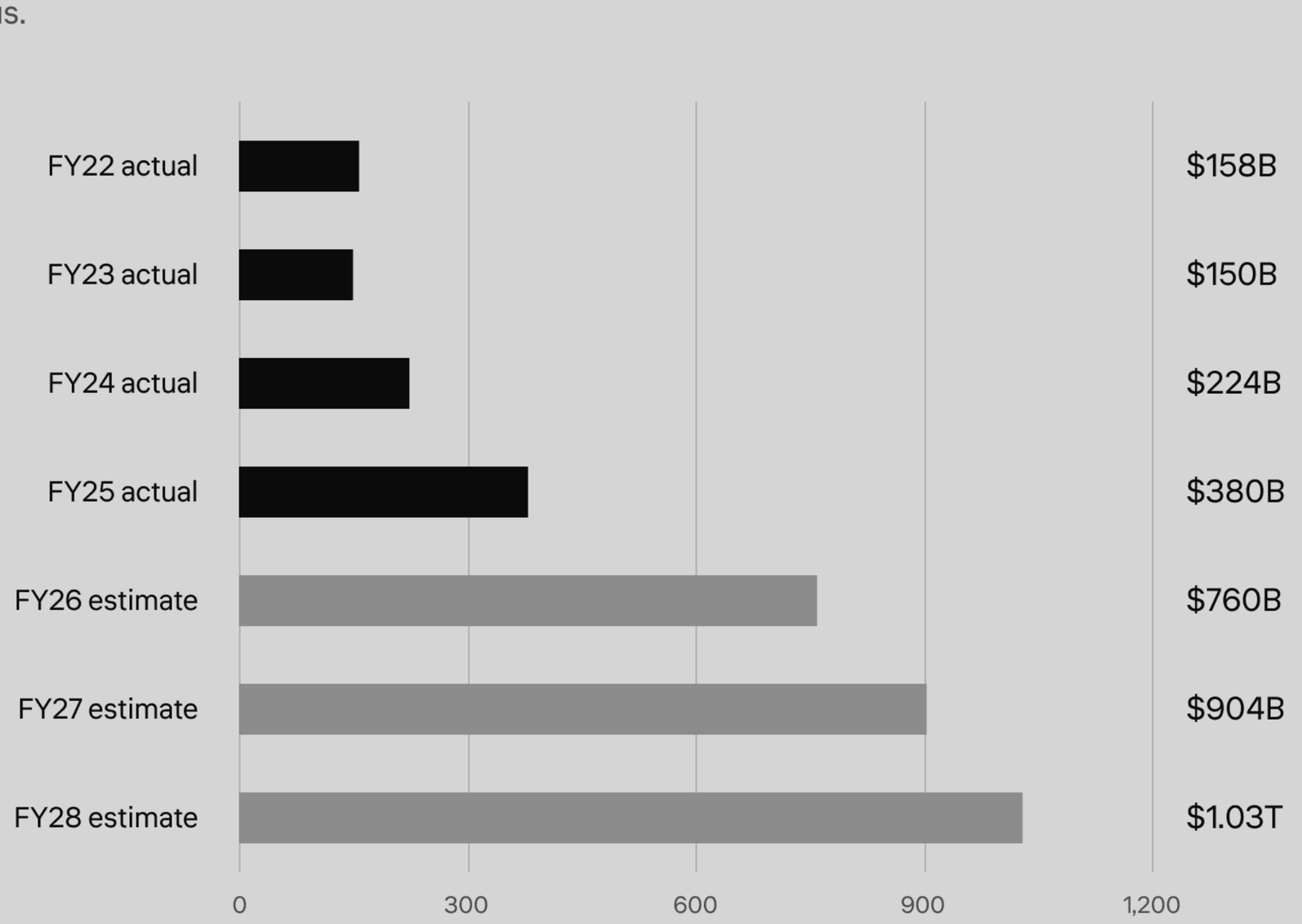


Source: Goldman Sachs Research (Briggs), 7 August 2026.

CHART

Top-five hyperscaler capex, US$ billions

Amazon, Alphabet, Microsoft, Meta and Oracle; FY26–28 are consensus estimates (light bars). Nvidia's Jensen Huang projects $3–4 trillion a year by 2030, about four times FY28 and off this scale. Source: CNBC (21 May 2026), chart by Ananya Chetia; company documents, Needham & Co., FactSet consensus.



Source: CNBC (21 May 2026), from company documents, Needham & Co. and FactSet consensus.

INSTITUTION	ANALYSIS	STANCE
Goldman Sachs, equity strategy (Sep 2026)	$800B of 2026 hyperscaler AI spend needs about $300B a year of AI revenue to break even; AI cloud revenue above its pre-AI trend is about $70B, a $230B annual gap. About $250B of 2026 bond issuance funds roughly a third of capex.	Constructive on spend; expects volatility
J.P. Morgan Asset Management (Jan 2026)	A $5T build-out needs about $650B a year of revenue for a 10% return, $35 a month from every iPhone user. Infrastructure monetization is established; applications are "early, uneven and opaque".	Diversify across the chain
Morgan Stanley (Aug 2025)	$2.9T of data-center capex through 2028 leaves a $1.5T external financing gap, about $800B of it from private credit.	Bubble fears "premature"
Apollo (Aug 2026)	Capex building twice as fast as the housing boom; the macro risk is the speed of an unwind.	Macro-risk warning
Man Group with MIT IDE (2025)	Lenders assume 10–20-year asset lives while GPUs obsolesce in 12–18 months; token prices falling 70% a year need about 225% annual demand growth to break even.	Financial structure outrunning adoption

Revenue has surprised on the upside. OpenAI's annual recurring revenue is nearing $70 billion, and Anthropic's run-rate passed $65 billion in July against about $4.6 billion of revenue in all of 2025. The two measures are not comparable with Goldman's "cloud revenue above trend", but the direction is clear, and it is not the dot-com pattern of revenue that never arrives. Four structural questions remain, and they are where the debate now turns.

Asset life. Hyperscalers depreciate servers over five to six years; Goldman's break-even assumes three to five. If accelerators are economically obsolete sooner, reported earnings overstate returns. From cash flow to debt. Bonds, private credit and off-balance-sheet vehicles fund a growing share of capex (Deep Dive XVII). Debt turns an equity disappointment into a credit event, against collateral that loses value fast. Concentration. A large share of committed demand traces back to a few loss-making labs; Anthropic alone reports $518 billion of future cloud and compute obligations. Fragility. On 23 June 2026 the Philadelphia semiconductor index fell 7.9%, Korea's KOSPI 9.99% and memory makers 12–13% in a session, after regulators flagged overheating in a trade that had seen Micron rise more than 760% and SK Hynix more than 800% in twelve months.

The correction that followed shows how fast a crowded trade reprices. Micron's share price fell 29% in July 2026 and SK Hynix closed the month about 40% below its 22 June level. The memory makers recovered part of the loss by September; Oracle did not, and ended September at half its value of a year earlier. The owners of the chokepoints are priced for continued scarcity, and the market tests that assumption every month (Part II).

The closest analogue is not the dot-com bubble but railways and telecom: real traffic, overbuilt capacity financed with debt, and returns that go to users and later owners rather than to the builders.

I.5

PART I

The loop: the labs' investors are also their suppliers

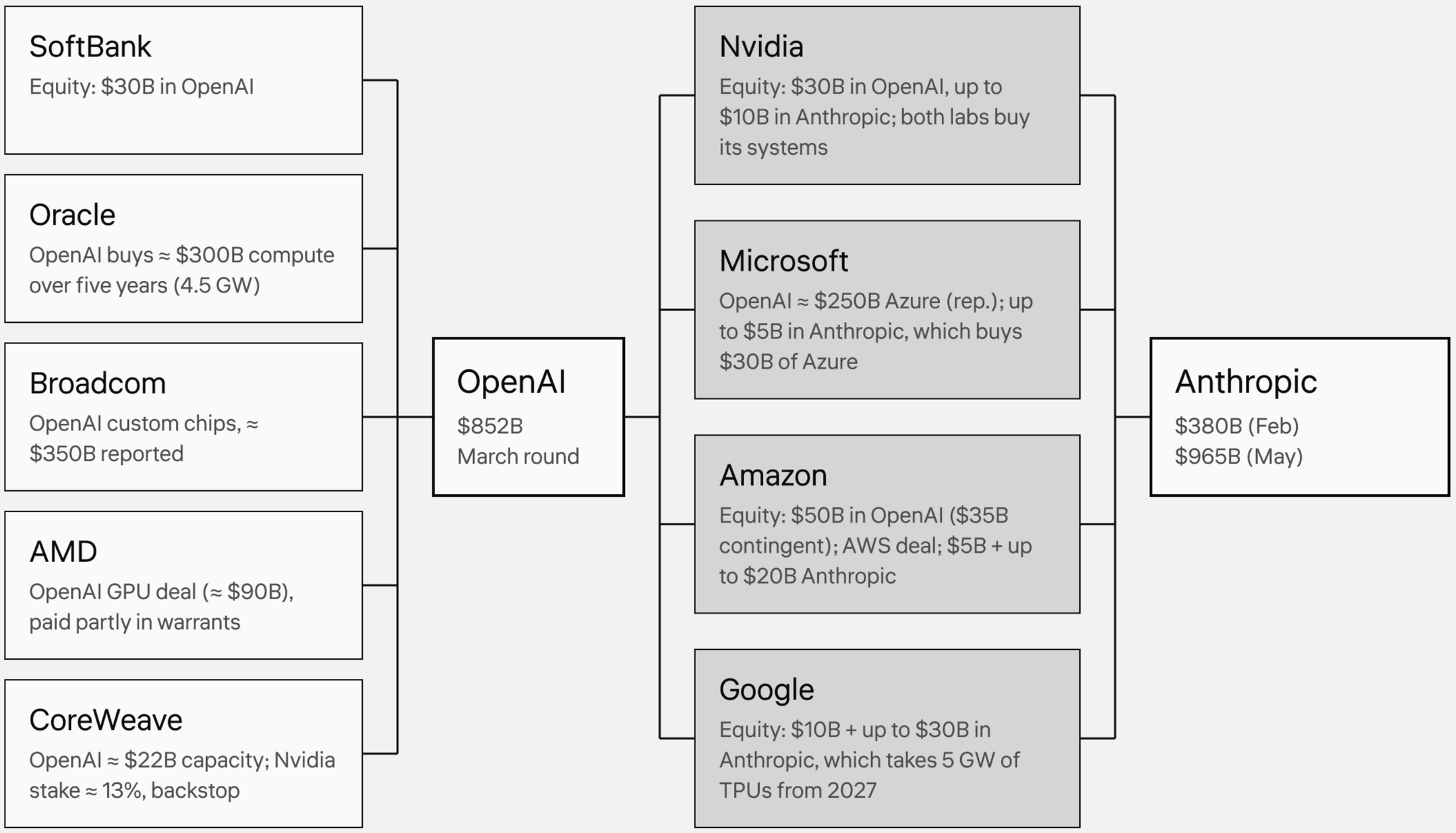
Bloomberg's "A Guide to the Circular Deals Underpinning the AI Boom" (Sam, Dottle, Ghosh and Kim) maps the funding web with company valuations: Nvidia $4.8 trillion, Google $4.7 trillion, Microsoft $3.1 trillion, Amazon $2.9 trillion, OpenAI $852 billion and Anthropic $380 billion, the value of its February round. The deals below are rebuilt from the reporting listed in the sources.

The terms, as reported: OpenAI's $122 billion round in March included $50 billion from Amazon ($35 billion of it contingent on an IPO or on reaching AGI), $30 billion from Nvidia and $30 billion from SoftBank. Nvidia and Microsoft committed up to $10 billion and $5 billion to Anthropic in November 2025, alongside Anthropic commitments to $30 billion of Azure compute and up to a gigawatt of Nvidia systems. Google added $10 billion in April 2026, with up to $30 billion more, alongside five gigawatts of TPU capacity from 2027. Nvidia holds about 13% of CoreWeave and has agreed to buy any capacity CoreWeave cannot sell through 2032. OpenAI's multi-year compute commitments (Broadcom about $350 billion, Oracle $300 billion, Microsoft $250 billion, AMD about $90 billion, AWS $38 billion) come from a secondary compilation. The loop has already been tested once: Nvidia's headline $100 billion letter of intent with OpenAI did not close as announced, and Nvidia took a $30 billion stake in the March round instead. Circularity matters for four reasons. Part of a supplier's reported demand is financed with its own capital, so its backlog overstates independent demand. Rising private valuations enlarge the equity that underwrites the next compute commitment, so a valuation reset would shrink commitments across the web at once. Warrants, contingent equity and capacity backstops move downside onto supplier balance sheets. And the expected IPOs of OpenAI and Anthropic will, for the first time, force public disclosure of how much of the committed spend revenue actually covers.

The board test for any AI supplier, partner or investee follows directly: what share of its backlog comes from counterparties it has itself financed?

FRAMEWORK BLOOMBERG

The labs' largest investors are also their largest suppliers



Highlighted: firms that are both shareholders in and suppliers to the frontier labs. Structure after Bloomberg (updated May 2026); deal terms from the reports listed in the sources. OpenAI's multi-year compute totals rest on a secondary compilation.

I.6

PART I

The risk a buyer controls

The risks of the AI trade fall into four families: competition and geopolitics; cost of compute, financial control and vendor lock-in; cyber, data, implementation and technology; and responsible AI, reputation and regulation. For a company buying AI, the second family matters most, because it falls on the buyer rather than the investor, and Deep Dive III and XIII describe its mechanics. Agentic workloads multiply tokens per task; AI costs arrive as variable, decentralized cloud bills that escape capital governance; and prompts, agents and data pipelines built around one provider raise switching costs without anyone deciding to.

RISK	EARLY WARNING	RESPONSE
Cost of compute	AI spend per workflow growing faster than its output	Token budgets, model routing, caching, evaluation-driven downgrading
Financial control	Unbudgeted AI line items; shadow subscriptions	AI FinOps: cost per case or deliverable, chargeback, quarterly review
Vendor lock-in	Most AI spend with a single model vendor	An abstraction layer, multi-model architecture, contractual portability, evaluation suites that make switching testable

Risk families after Gary Gensler, “AI – Finance” (MIT Sloan, 2026); warnings and responses are a House of Brains synthesis.

If the capex cycle digests, token prices could fall sharply as capacity frees up: good for buyers, bad for any business case that assumes today's scarcity. The asymmetry argument of Deep Dive XVII applies once more. For companies outside Big Tech, the risk worth avoiding is a long, rigid commitment at peak prices, not a missed data center.

I.7

PART I

A dashboard to monitor the thesis

The discipline is to adapt when the thesis erodes or becomes consensus. These indicators would signal it.

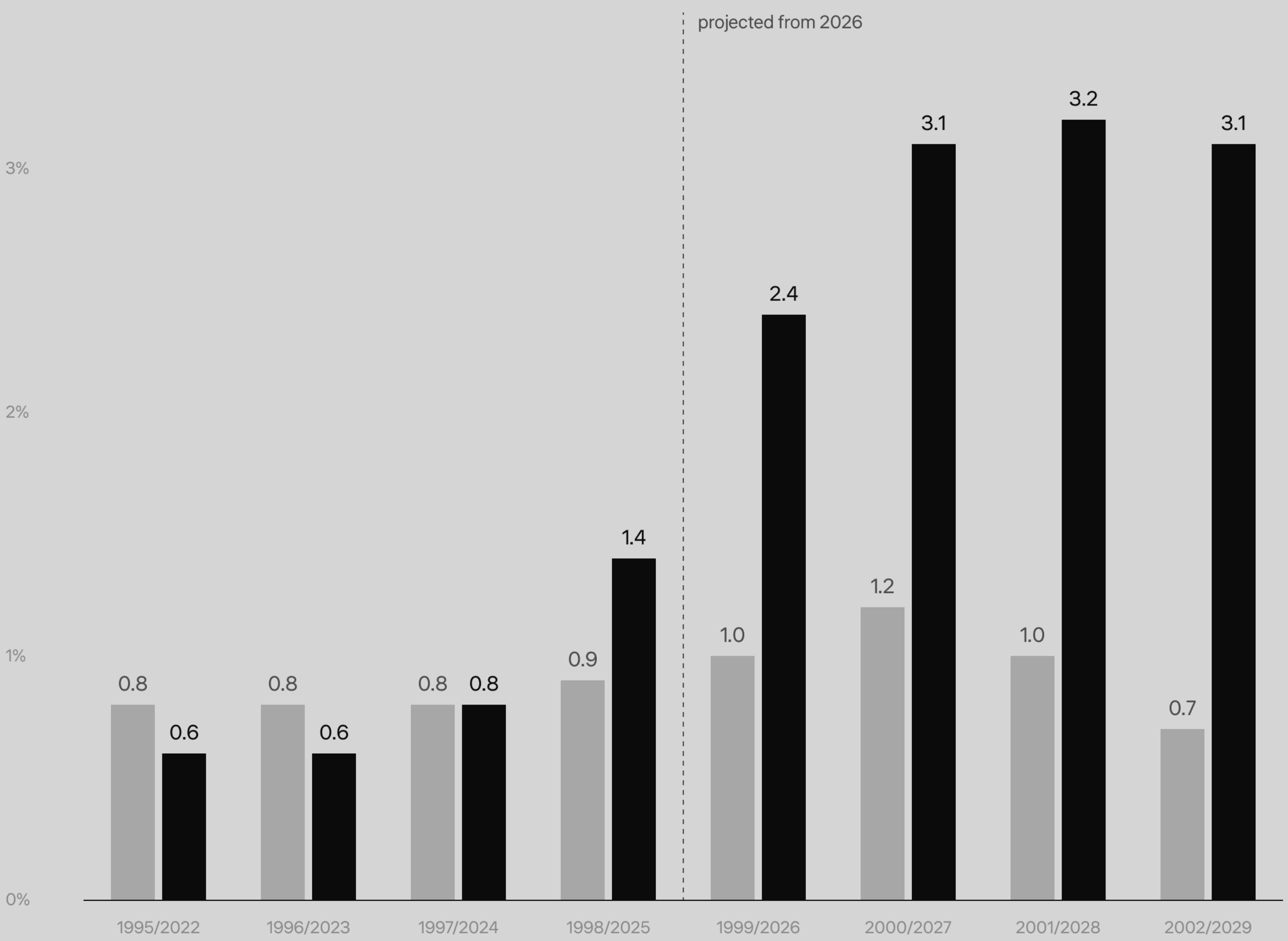
INDICATOR	CONFIRMS THE BOOM THESIS	SIGNALS EROSION	SOURCE TO TRACK
Frontier-lab revenue run rate vs compute commitments	Revenue growth above 2× a year; commitments covered	Growth slowing while commitments keep rising	Lab disclosures, IPO filings
Hyperscaler capex guidance and cloud growth	Cloud growth accelerating alongside capex	Capex cuts or cloud deceleration	Quarterly earnings
Share of capex funded by debt	Stable, mostly from cash flow	Rising bond, private-credit and SPV share	Goldman Sachs, Morgan Stanley credit research
Token prices and GPU rental rates	Stable or rising (scarcity)	Sharp declines (overcapacity)	J.P. Morgan token index; GPU rental indices
Open vs closed capability gap	Gap holds or widens: pricing power intact	Gap under two months: commoditization	Epoch AI Capabilities Index
Enterprise productivity evidence	Measured gains in controlled deployments	Pilots stall; ROI unproven at scale	Academic studies, client data
Market positioning	Mixed sentiment	Near-universal “AI winner” consensus and crowding	Fund-manager surveys, concentration metrics

For boards. AI is now both an operating question and a balance-sheet and macro exposure. Require management to state the company’s AI thesis across the five lenses, with named indicators and a review cadence, and to report AI operating cost as its own line with unit economics. Stress-test capex-heavy plans against a 2027–28 digestion scenario: slower hyperscaler orders, falling GPU collateral values, repriced private credit. For CEOs. Buy capability cheaply and own the workflow. Avoid long, rigid commitments at today’s scarcity prices; route frontier work to closed models and bulk work to cheaper ones; put the scarce investment into proprietary data, process redesign and distribution, the assets that keep their value when models commoditize.

CHART

Hyperscaler capex vs telecom capex in the late 1990s, % of US GDP

The two cycles aligned year by year; hyperscaler figures from 2026 are projections.



Source: Torsten Slok, Apollo (6 August 2026).

Five years of market value: where AI has paid so far

So far the market has paid the suppliers of scarce compute far more than the companies that use it, and it has started to mark down businesses that sell seats and hours.

Part I and Deep Dive XVII follow the capital going into AI. This part follows the value that came out, as the equity market has priced it. I tracked two things over five years: the share prices of eleven technology leaders, rebased to 100 at 30 September 2021, and the ranking of the fifty most valuable technology companies in the United States, Europe and Asia at the end of July 2021 and at the end of July 2026. Both give the same answer. So far the market has paid the suppliers of scarce compute far more than the companies that use it, and it has started to mark down businesses that sell seats and hours.

$40.0T

Value of the global tech top 50, July 2026: 2.2× the $18.2T of July 2021

43%

Share of that value in semiconductors and chip equipment, up from 17%

11×

Nvidia's share price over five years; the median of eleven leaders rose to 1.9×

20 of 50

Companies in the 2021 top 50 that had left it by 2026

II.1

PART II

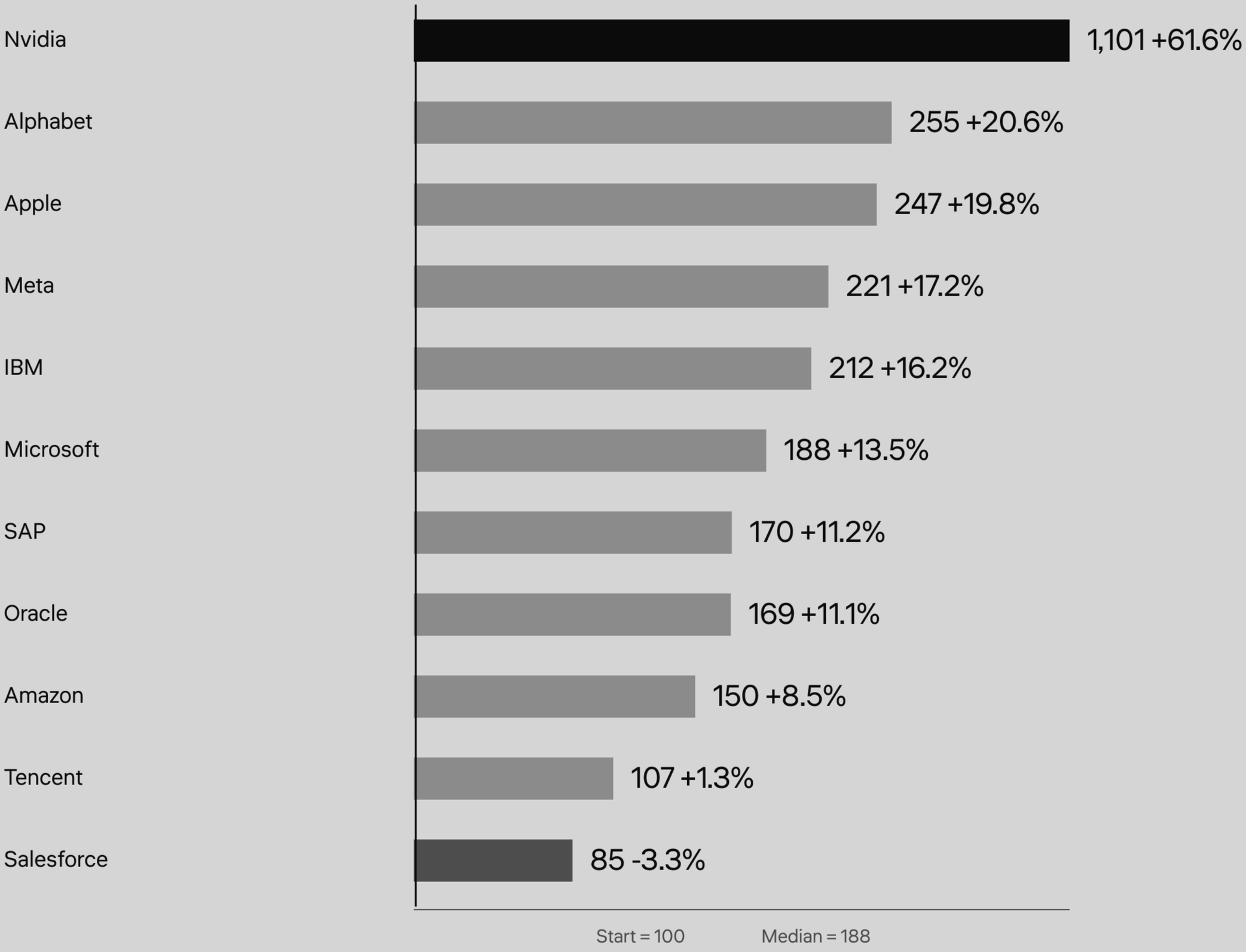
Eleven leaders, rebased to 100

From 30 September 2021 to 30 September 2026, Nvidia's share price rose about elevenfold and the median of the eleven companies rose to 188. Salesforce was the only one of the eleven below its starting level, and Tencent was about flat in dollar terms. Almost every company hit its low in late 2022, when rising rates cut technology valuations. Oracle, SAP and IBM are the exceptions: their deepest falls came in 2026, in the selloff of software and heavily indebted AI-infrastructure names.

CHART

Five of the eleven more than doubled; the median ended at 188

Share-price index at 30 Sep 2026 (30 Sep 2021 = 100) and compound annual growth rate.
Nvidia's bar is cut at 350 so the other ten stay readable; its index is 1,101.

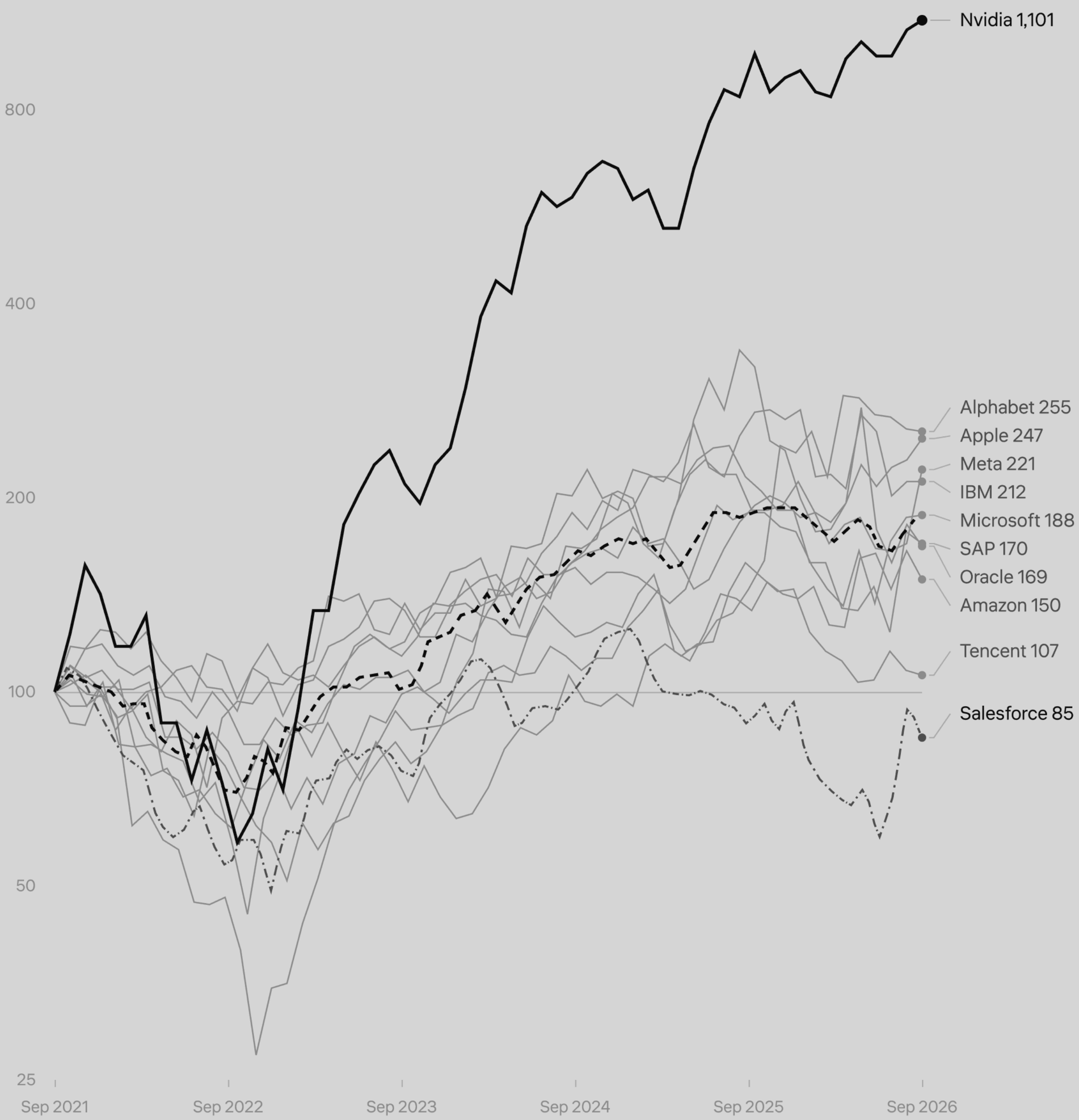


Source: House of Brains compilation from StatMuse (US listings, split- and dividend-adjusted) and digrin (Tencent ADR); month-end closes to 30 September 2026.

CHART

Nvidia rose elevenfold; Salesforce is the only one below its 2021 level

Share price, 30 Sep 2021 = 100, month-end closes to 30 Sep 2026, log scale. Dashed line: median of the eleven. US listings in dollars, split- and dividend-adjusted (StatMuse); Tencent is the US ADR, adjusted (digrin). An interactive version with a linear scale is in Part III, the data room.



Source: House of Brains compilation from StatMuse (US listings, split- and dividend-adjusted) and digrin (Tencent ADR); month-end closes to 30 September 2026.

Within each peer group the spread is wider than between groups. Among hyperscalers, Alphabet (255) pulled ahead of Microsoft (188) and Amazon (150), while Oracle rose to about 340 in September 2025 on AI-cloud contracts and then gave back half of it. In enterprise software and services, IBM (212) beat SAP (170) and Salesforce (85). Among consumer platforms, Apple and Meta ended close together, but Meta got there from a 73% drawdown in 2022; Tencent ended where it started.

The drawdowns change the picture more than the endpoints do. Meta fell 73% before recovering to 221. Oracle ended at half its September 2025 peak. IBM, SAP and Oracle had their deepest drawdowns in 2026, not 2022.

COMPANY	TICKER	INDEX, SEP 2026	5-YR CHANGE	CAGR	PEAK INDEX	PEAK MONTH	VS PEAK	MAX DRAWDOWN	TROUGH MONTH
Nvidia	NVDA	1,101	+1,001%	61.6%	1,101	Sep 2026	0%	−63%	Sep 2022
Alphabet	GOOGL	255	+155%	20.6%	290	Apr 2026	−12%	−40%	Dec 2022
Apple	AAPL	247	+147%	19.8%	247	Sep 2026	0%	−26%	Dec 2022
Meta	META	221	+121%	17.2%	229	Jul 2025	−3%	−73%	Oct 2022
IBM	IBM	212	+112%	16.2%	276	Nov 2025	−23%	−27%	Jul 2026
Microsoft	MSFT	188	+88%	13.5%	195	Jul 2025	−4%	−31%	Oct 2022
SAP	SAP	170	+70%	11.2%	241	Jun 2025	−30%	−49%	Jun 2026
Oracle	ORCL	169	+69%	11.1%	341	Sep 2025	−50%	−53%	Jul 2026
Amazon	AMZN	150	+50%	8.5%	165	Jul 2026	−9%	−52%	Dec 2022
Tencent	TCEHY	107	+7%	1.3%	160	Sep 2025	−33%	−57%	Oct 2022
Salesforce	CRM	85	−15%	−3.3%	127	Jan 2025	−33%	−56%	Dec 2022

Drawdowns use month-end closes, so falls within a month were deeper than shown. Recent returns computed from the index were cross-checked against companiesmarketcap.com for five companies; the gaps of one to five points come from dividend adjustment and the time of day each source uses.

TICKER	LATEST CLOSE, US$ (30 SEP 2026)	2025 RETURN, INDEX	2025, COMPANIESMARKETCAP	2026 YTD, INDEX	2026 YTD, COMPANIESMARKETCAP
NVDA	227.21	+38.9%	—	+22.1%	—
GOOGL	337.83	+66.0%	—	+8.1%	—
AAPL	341.07	+9.0%	+11.9%	+25.8%	+24.5%
META	744.10	+13.1%	+10.3%	+13.0%	+8.4%
IBM	232.76	+38.2%	—	−19.8%	—
MSFT	508.96	+15.6%	—	+5.9%	—
SAP	210.85	−0.4%	—	−11.7%	—
ORCL	137.79	+18.1%	+18.4%	−28.6%	−29.3%
AMZN	246.67	+5.2%	—	+6.9%	—
TCEHY	56.18	+44.6%	+45.1%	−26.1%	−28.8%
CRM	225.31	−20.2%	−19.6%	−14.4%	−14.2%

II.2

PART II

The global tech top 50, July 2021 against July 2026

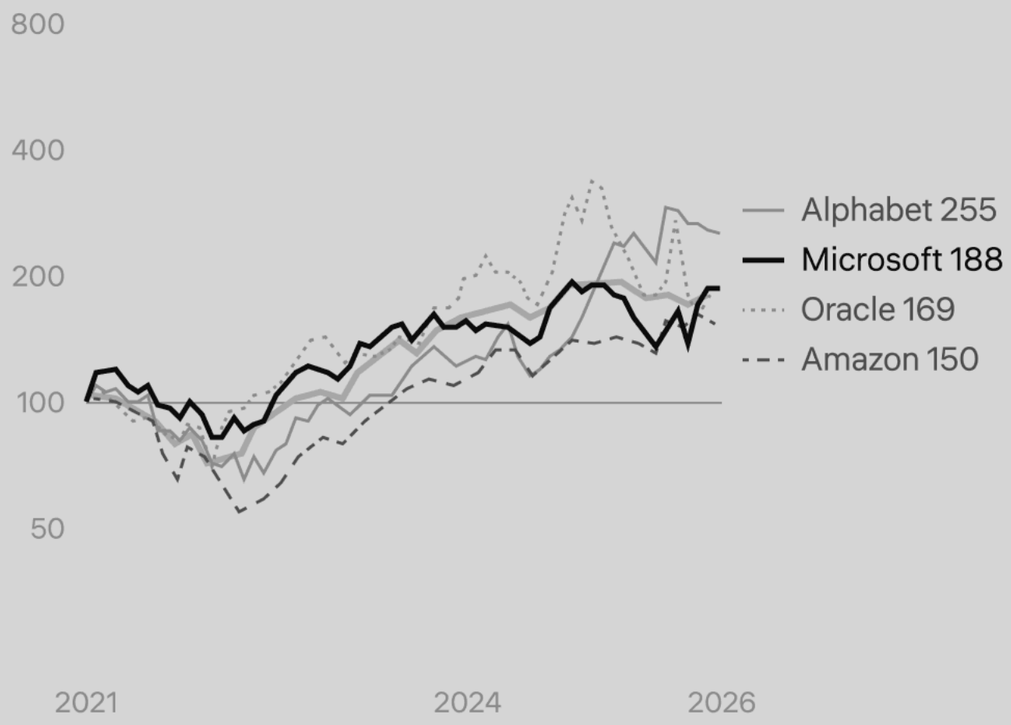
The fifty most valuable tech companies in the United States, Europe and Asia were worth about $40.0 trillion at the end of July 2026, 2.2 times the $18.2 trillion of July 2021, roughly 17% a year. Fiftieth place needed $134 billion, against $81 billion five years earlier.

CHART

The gap inside each peer group is wider than the gap between groups

Share price, 30 Sep 2021 = 100, log scale, same scale in every panel. Dashed line: median of all eleven.

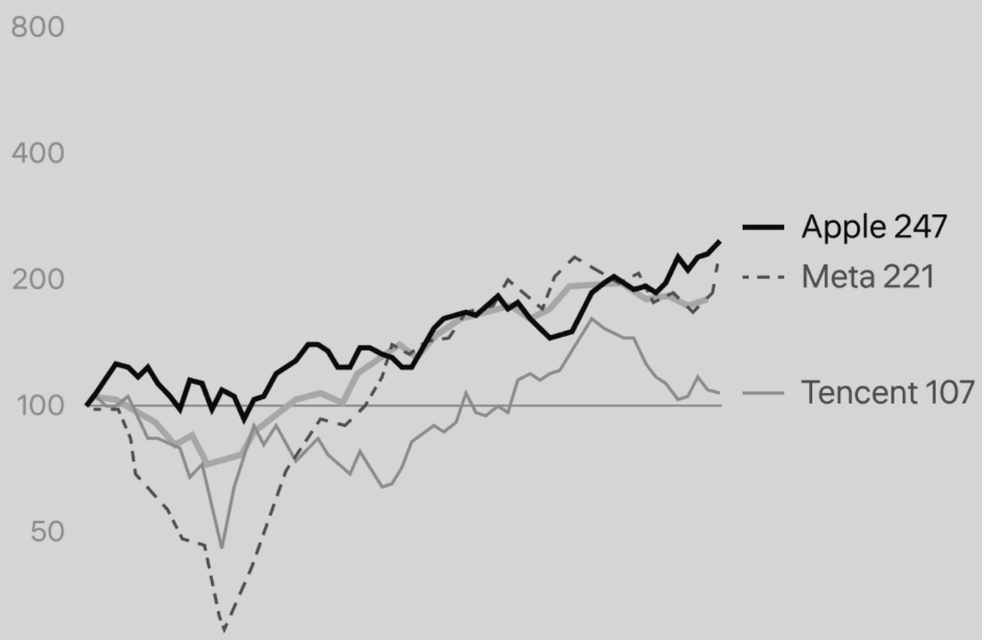
Cloud hyperscalers



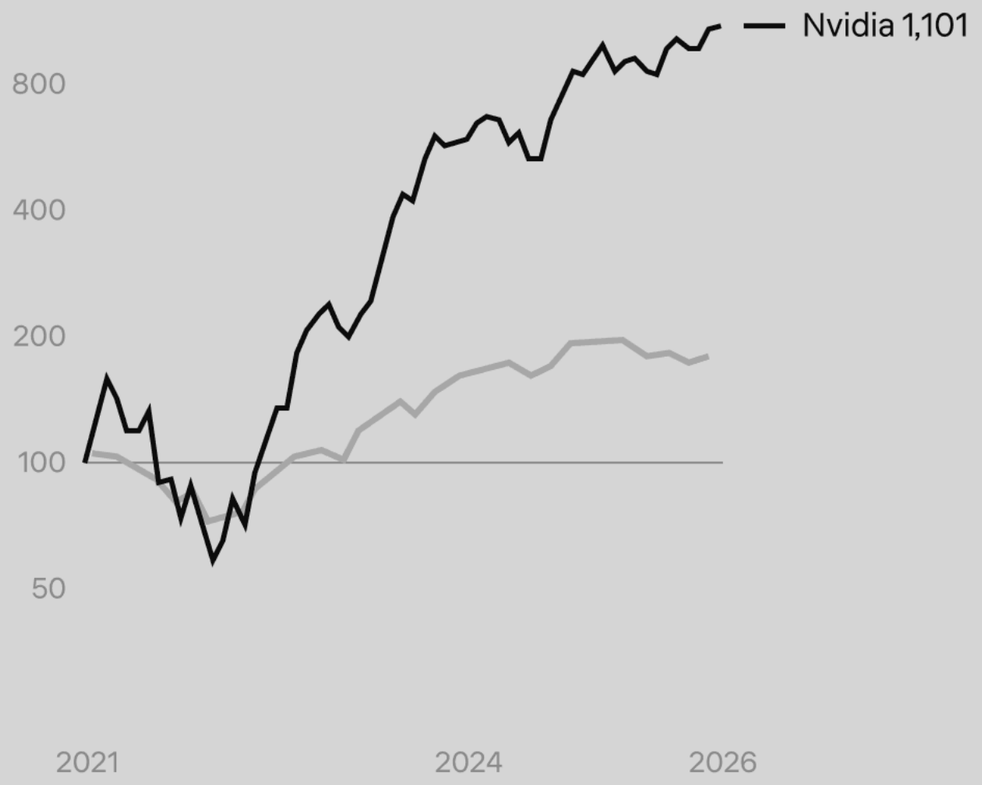
Enterprise software and services



Consumer platforms



AI compute



Source: House of Brains compilation from StatMuse (US listings, split- and dividend-adjusted) and digrin (Tencent ADR); month-end closes to 30 September 2026.

CHART

Nvidia broke away; four others finished clearly above the median

One panel per company, ordered by ending index, log scale 25 to 1,300 in every panel. Accent line: the company; dashed line: median of the eleven. Nvidia's path is the only one that left the median band for good, in 2023; Oracle and SAP were above it for most of 2024–25 and fell back below it in 2026.



Source: House of Brains compilation from StatMuse (US listings, split- and dividend-adjusted) and digrin (Tencent ADR); month-end closes to 30 September 2026.

Concentration at the very top barely moved: the five largest held about half of top-50 value in both years, but Nvidia replaced Meta among them, and the top ten's share rose from 66% to 70%. What changed was the layer of the stack where the value sits.

Twenty companies entered the top 50 and twenty left, and the swap runs almost entirely in one direction: in came the physical AI supply chain, out went application software, IT services and Chinese consumer internet.

Entrants (20). Memory and storage: SK Hynix, Kioxia, CXMT, Sandisk, Western Digital, Seagate. Chip design: Arm, Marvell, MediaTek, Analog Devices. Chip equipment: KLA, Tokyo Electron, Advantest. Data-center hardware: Dell, Arista, Amphenol. Software tied to AI deployment, security or advertising: Palantir, Palo Alto Networks, CrowdStrike, AppLovin. Exits (20). Application software: Adobe, Intuit, ServiceNow, Zoom, Atlassian. IT and business services: Accenture, TCS, Infosys, ADP. Chinese consumer internet and hardware: Meituan, JD.com, PDD, Xiaomi, Hikvision, BYD. Pandemic-era platforms: Sea, Snap, Airbnb. Japanese quality compounders: Keyence, Recruit.

Largest value losses among the 2021 top 50. Adobe lost about $200 billion (–67%), Alibaba $192 billion (–39%), Accenture $103 billion (–50%), Meituan $97 billion (–57%), Salesforce $86 billion (–36%) and Sea $83 billion (–56%). TCS and Infosys lost 43% and 49%.

Largest gains. SK Hynix rose about 13× in value, Micron 11×, Nvidia 10× and Broadcom 9.5×; Advantest, Arista, Palantir and Palo Alto Networks each rose between 6.5× and 9×.

Value moved toward the United States and toward the two Asian economies that dominate chip manufacturing, and away from China's consumer platforms and India's IT services.

HEADQUARTERS	JUL 2021, US$ BN	COMPANIES	JUL 2026, US$ BN	COMPANIES
United States	13,666	29	32,487	35
Taiwan	605	1	2,248	2
South Korea	459	1	2,022	2
China	1,754	8	1,387	3
Europe	688	3	1,071	3
Japan	344	3	599	4
India	253	2	0	0
Singapore, Canada, Australia	418	3	152	1

China's three remaining names show what its market now rewards. Tencent and Alibaba are worth less than in 2021, while CXMT, a state-backed memory maker that listed only on 27 July 2026, is already worth $535 billion. The United States share of top-50 value rose from 75% to 81%. Europe's count is unchanged at three, but only ASML, SAP and Arm remain: Accenture's exit was offset by Arm's listing.

CHART

Chips and chip equipment went from $3.1T to $17.0T of top-50 value

Market value of the top-50 tech companies by segment, US$, end-July; share of each year's top-50 total in brackets. Compiled from StatMuse, stockanalysis.com, companiesmarketcap.com, digrin and SEC filings.



Source: House of Brains compilation from StatMuse, stockanalysis.com, companiesmarketcap.com, digrin and SEC filings; market values at end-July 2021 and 2026.

TOP 50, END OF JULY 2026

#	COMPANY	HQ	SEGMENT	MKT CAP, US$ BN	JUL 2021 RANK	VALUE VS JUL 2021
1	Nvidia	US	Semiconductors	4,871	10	9.9x
2	Apple	US	Devices & hardware	4,504	1	1.9x
3	Alphabet	US	Internet platforms	4,353	3	2.4x
4	Microsoft	US	Software & cloud	3,446	2	1.6x
5	Amazon	US	Internet platforms	2,922	4	1.7x
6	TSMC	Asia (TW)	Semiconductors	2,072	7	3.4x
7	Broadcom	US	Semiconductors	1,899	20	9.5x
8	Meta	US	Internet platforms	1,374	5	1.4x
9	Tesla	US	EVs & autos	1,245	6	1.8x
10	Samsung Electronics	Asia (KR)	Semiconductors	1,155	11	2.5x
11	Micron	US	Semiconductors	920	46	10.7x
12	SK Hynix	Asia (KR)	Semiconductors	867	outside top 50	12.9x
13	AMD	US	Semiconductors	778	30	6.0x
14	ASML	Europe (NL)	Semi equipment	604	12	1.9x
15	Tencent	Asia (CN)	Internet platforms	545	8	0.9x
16	CXMT	Asia (CN)	Semiconductors	535	not listed	—
17	Intel	US	Semiconductors	477	18	2.2x
18	Cisco	US	Devices & hardware	462	17	2.0x
19	Applied Materials	US	Semi equipment	406	32	3.2x
20	Oracle	US	Software & cloud	390	14	1.6x
21	Lam Research	US	Semi equipment	356	45	4.1x
22	Alibaba	Asia (CN)	Internet platforms	307	9	0.6x
23	Palantir	US	Software & cloud	296	outside top 50	7.0x
24	Netflix	US	Internet platforms	294	16	1.2x

#	COMPANY	HQ	SEGMENT	MKT CAP, US$ BN	JUL 2021 RANK	VALUE VS JUL 2021
25	Dell	US	Devices & hardware	264	outside top 50	3.6x
26	Arm Holdings	Europe (UK)	Semiconductors	256	not listed	—
27	Palo Alto Networks	US	Software & cloud	254	outside top 50	6.5x
28	Texas Instruments	US	Semiconductors	252	22	1.4x
29	KLA	US	Semi equipment	238	outside top 50	4.5x
30	Arista Networks	US	Devices & hardware	230	outside top 50	7.6x
31	SAP	Europe (DE)	Software & cloud	210	24	1.2x
32	IBM	US	IT services	210	31	1.7x
33	CrowdStrike	US	Software & cloud	198	outside top 50	3.5x
34	Amphenol	US	Devices & hardware	198	outside top 50	—
35	Western Digital	US	Devices & hardware	196	outside top 50	—
36	Seagate	US	Devices & hardware	196	outside top 50	—
37	Analog Devices	US	Semiconductors	180	outside top 50	2.9x
38	Sandisk	US	Semiconductors	176	part of WDC	—
39	MediaTek	Asia (TW)	Semiconductors	176	outside top 50	3.4x
40	Marvell	US	Semiconductors	173	outside top 50	3.5x
41	Kioxia	Asia (JP)	Semiconductors	159	not listed	—
42	Tokyo Electron	Asia (JP)	Semi equipment	158	outside top 50	2.5x
43	Qualcomm	US	Semiconductors	157	25	0.9x
44	Shopify	Canada	Internet platforms	152	21	0.8x
45	Salesforce	US	Software & cloud	151	15	0.6x
46	Advantest	Asia (JP)	Semi equipment	147	outside top 50	8.7x

#	COMPANY	HQ	SEGMENT	MKT CAP, US$ BN	JUL 2021 RANK	VALUE VS JUL 2021
47	Booking Holdings	US	Internet platforms	145	43	1.6x
48	Uber	US	Internet platforms	144	48	1.7x
49	Sony	Asia (JP)	Devices & hardware	136	33	1.1x
50	AppLovin	US	Software & cloud	134	outside top 50	—

Market value at the close on 31 July 2026. "Value vs Jul 2021" compares market values, not share prices, so buybacks and new issuance are reflected. A dash means no 2021 value was compiled: the company was private, not yet spun off, or well outside the ranking. IBM's 2021 value includes Kyndryl, and Dell's includes its VMware stake; both were spun off in November 2021.

TOP 50, END OF JULY 2021

#	COMPANY	HQ	SEGMENT	MKT CAP, US$ BN	JUL 2026 RANK
1	Apple	US	Devices & hardware	2,411	2
2	Microsoft	US	Software & cloud	2,141	4
3	Alphabet	US	Internet platforms	1,800	3
4	Amazon	US	Internet platforms	1,714	5
5	Meta (then Facebook)	US	Internet platforms	1,005	8
6	Tesla	US	EVs & autos	680	9
7	TSMC	Asia (TW)	Semiconductors	605	6
8	Tencent	Asia (CN)	Internet platforms	592	15
9	Alibaba	Asia (CN)	Internet platforms	499	22
10	Nvidia	US	Semiconductors	492	1
11	Samsung Electronics	Asia (KR)	Semiconductors	459	10
12	ASML	Europe (NL)	Semi equipment	313	14
13	Adobe	US	Software & cloud	299	outside top 50
14	Oracle	US	Software & cloud	243	20
15	Salesforce	US	Software & cloud	237	45
16	Netflix	US	Internet platforms	236	24

#	COMPANY	HQ	SEGMENT	MKT CAP, US$ BN	JUL 2026 RANK
17	Cisco	US	Devices & hardware	234	18
18	Intel	US	Semiconductors	218	17
19	Accenture	Europe (Ireland)	IT services	205	outside top 50
20	Broadcom	US	Semiconductors	200	7
21	Shopify	Canada	Internet platforms	187	44
22	Texas Instruments	US	Semiconductors	176	28
23	Meituan	Asia (CN)	Internet platforms	170	outside top 50
24	SAP	Europe (DE)	Software & cloud	170	31
25	Qualcomm	US	Semiconductors	169	43
26	TCS	Asia (IN)	IT services	158	outside top 50
27	Sea Limited	Asia (SG)	Internet platforms	149	outside top 50
28	Intuit	US	Software & cloud	147	outside top 50
29	Keyence	Asia (JP)	Devices & hardware	134	outside top 50
30	AMD	US	Semiconductors	131	13
31	IBM	US	IT services	126	32
32	Applied Materials	US	Semi equipment	126	19
33	Sony	Asia (JP)	Devices & hardware	126	49
34	ServiceNow	US	Software & cloud	119	outside top 50
35	Snap	US	Internet platforms	118	outside top 50
36	Zoom	US	Software & cloud	116	outside top 50

#	COMPANY	HQ	SEGMENT	MKT CAP, US$ BN	JUL 2026 RANK
37	JD.com	Asia (CN)	Internet platforms	113	outside top 50
38	BYD	Asia (CN)	EVs & autos	107	outside top 50
39	PDD Holdings	Asia (CN)	Internet platforms	104	outside top 50
40	Airbnb	US	Internet platforms	98	outside top 50
41	Infosys	Asia (IN)	IT services	95	outside top 50
42	ADP	US	IT services	89	outside top 50
43	Booking Holdings	US	Internet platforms	88	47
44	Hikvision	Asia (CN)	Devices & hardware	88	outside top 50
45	Lam Research	US	Semi equipment	86	21
46	Micron	US	Semiconductors	86	11
47	Recruit	Asia (JP)	Internet platforms	84	outside top 50
48	Uber	US	Internet platforms	82	48
49	Atlassian	Australia	Software & cloud	82	outside top 50
50	Xiaomi	Asia (CN)	Devices & hardware	81	outside top 50

Market value at the close on 30 July 2021. Tencent, Meituan and Alibaba were already down sharply that week because of China's regulatory crackdown, so the 2021 baseline understates their earlier peaks.

II.3

PART II

What moved the table: money got expensive again

The 2021 table was priced for money that cost almost nothing; the 2026 table is priced for money that costs something again. The US 10-year Treasury yield was 1.24% at the end of July 2021 and 4.75% at the end of July 2026. The Federal Reserve moved from near zero in early 2022 to a peak of 5.25–5.50% in July 2023, cut to 3.50–3.75% by December 2025, and on 16 September 2026, after this snapshot, raised rates again to 3.75–4.00%. That reset explains most of the exits. In 2022 the Nasdaq fell 33%, erasing about $9 trillion, and Meta and Tesla each lost about two-thirds of their value. Companies valued mainly on distant cash flows fell furthest: subscription software, e-commerce and pandemic beneficiaries such as Zoom, Sea, Snap and Shopify. Most never regained their 2021 levels, because the discount rate never returned to 2021 levels either. The recovery went to companies with current, visible earnings, which is the main difference from 2000: the gains at the top rest on reported profits, not only on multiples. Nvidia's data-center revenue went from $10.6 billion in fiscal 2022 to $193.7 billion in fiscal 2026, and reached $89.0 billion in the quarter to July 2026 alone. The Nasdaq-100 returned roughly 95% over the five years, about 14% a year on a third-party calculation, while top-50 value rose 120%; the difference comes from companies that grew into the list and from new listings. The path was volatile: the April 2025 tariff announcement erased about $6.6 trillion of US market value in two days before a 90-day pause.

II.4

PART II

Capex became someone else's revenue

Generative AI moved value from the companies that use compute to the companies that supply it. The four largest hyperscalers spent roughly $125 billion on capex in 2021 (estimate), $416 billion in 2025, and after July 2026 earnings guide to about $725 billion for 2026. Deep Dive XVII describes that spending; this table is where it landed. Scarcity moved along the chain: first GPUs, then custom accelerators and networking, then high-bandwidth and conventional memory, then storage, and finally the equipment needed to expand capacity. Each company that held the bottleneck of the moment captured pricing power, which is the resource-layer phase of the value migration in Deep Dive IX. Memory is the clearest case, and a reminder that it is still a cycle. Micron's revenue in the quarter to May 2026 was $41.5 billion, up 346%, and TrendForce forecast DRAM contract prices up another 13–18% in the third quarter. On 22 June 2026 SK Hynix briefly passed Samsung as Korea's most valuable company, at ₩2,091 trillion. Then came the correction described in Part I: Micron fell 29% in July, and SK Hynix closed the month about 40% below its 22

June level. The July 2026 values in the table are therefore taken close to a local low for memory. Efficiency shocks have not broken the thesis so far: DeepSeek's low-cost model cut $589 billion from Nvidia's market value on 27 January 2025, a one-day record, and capex plans rose anyway, because cheaper inference raised usage more than it cut the compute per task. The same technology has turned against parts of software. On 3 February 2026 the IGV software index fell 5% in a day and 28% from its high after Anthropic released legal-workflow plugins for Claude Cowork; by May, Adobe and Salesforce traded at about half their early-2025 earnings multiples. The software companies that gained sit close to AI deployment or its risks: Palantir (7×), Palo Alto Networks (6.5×), CrowdStrike (3.5×) and AppLovin.

II.5

PART II

Geopolitics became a valuation input

China's regulatory reset permanently lowered the value investors put on its consumer platforms. The crackdown began with Alibaba's RMB18.2 billion antitrust fine in April 2021 and widened to Didi, private tutoring and data rules that summer. Beijing signalled easing with Xi Jinping's February 2025 meeting with private entrepreneurs, but capital did not return to the same companies. It went to state-priority hardware: CXMT raised $8.6 billion in its July 2026 listing, Cambricon's revenue doubled in the first half of 2026, and Chinese buyers plan to spend 46% of AI-accelerator budgets on domestic chips, up from 30%. US export controls tightened in October 2022, October 2023 and December 2024, the last round adding high-bandwidth memory. In April 2025 Nvidia's China-specific H20 was frozen, at a $5.5 billion charge; sales resumed in August 2025 against 15% of revenue, and H200 sales followed in December 2025 at 25%. A 25% Section 232 tariff on imported advanced AI chips followed in January 2026, and in June 2026 licensing was extended to Chinese buyers' overseas subsidiaries. In August 2025 Washington converted CHIPS Act grants into a 9.9% stake in Intel, which helps explain why Intel doubled while losing process leadership. Korea and Taiwan together went from $1.06 trillion to $4.27 trillion of top-50 value, a concentration exposed to cross-Strait risk and to US tariff and subsidy policy. Europe regulated rather than scaled: Digital Markets Act fines hit Apple (€500 million) and Meta (€200 million) in April 2025 and Google (€890 million) on 23 July 2026.

II.6

PART II

Inside each segment, proximity to AI demand decided

Cloud. The hyperscalers have gone from asset-light to capital-heavy. Alphabet did best among the platforms (2.4× in value), helped by search holding up, Google Cloud and its own TPU chips; Microsoft (1.6×) and Amazon (1.7×) lagged the suppliers they pay, and Meta (1.4×) fell 10% the day it raised its 2026 capex guidance. Oracle shows the financing risk most clearly: $664 billion of contracted future revenue, but quarterly capex of $28.5 billion, negative free cash flow of $5.4 billion and a $20 billion equity raise. Its value fell from roughly $800 billion at the end of September 2025 to $390 billion.

Semiconductors without AI data-center exposure lagged. Qualcomm (0.9×), tied to smartphones, and Texas Instruments and Analog Devices, tied to industrial and auto cycles, gained far less than their peers.

Consumer internet. Digital advertising held up and funded AI capex. The companies hurt were those whose growth came from pandemic behavior or from China's consumer economy, where weak demand and price wars reinforced regulatory pressure.

Application software and IT services. Both face the same question: what happens to revenue billed per seat or per person when agents do part of the work. The market's answer so far is a lower multiple. Accenture, TCS and Infosys lost 43–50% of their value; IBM is the only services-weighted company left in the top 50, and its gains come mainly from software and mainframes.

Devices and EVs. Apple (1.9×) compounded through buybacks and services rather than unit growth. Tesla (1.8×) stayed valued on autonomy and robotics options; BYD grew its volumes but slipped out of the top 50 on price competition in China.

Capital markets. Much of the value AI companies create now accrues in private rounds before listing. SpaceX listed on 12 June 2026 after a record $75 billion IPO and was worth about $1.39 trillion on 31 July; it is excluded here as a satellite-telecom and aerospace company but would otherwise rank eighth. Cerebras listed in May. Anthropic filed confidentially in June after its $965 billion round, and OpenAI is reported to be targeting a listing in the fourth quarter. The next version of this ranking is likely to include at least two AI labs near the top.

II.7

PART II

What the verdict means for boards and advisors

II.1

Value follows scarcity, and scarcity moves.

In five years the bottleneck shifted from GPUs to networking, memory, storage and power. Boards should know which input their own AI plans depend on and who controls it: supplier concentration and price risk now sit with a handful of companies in the United States, Taiwan and Korea.

II.2

The build-out is a capex cycle, and cycles turn.

July 2026 showed how fast memory and AI stocks reprice. Budgets, partnerships and vendor commitments should hold up if AI spending slows in 2027–28, which is the asymmetry argument of Deep Dive XVII and Part I applied to the buyer.

II.3

Seat- and headcount-based models are being repriced now.

Adobe, Salesforce, ServiceNow, Accenture, TCS and Infosys lost ground on the expectation that agents will reduce seats and billable hours, not on current revenue. For advisory firms, the market is already discounting the leverage pyramid (Deep Dive XV).

II.4

Geopolitics is part of valuation.

Export controls, chip tariffs, China's shift from platforms to state-backed hardware and cross-Strait concentration all affect supply security and market access. Supply-chain, market-entry and technology-partner decisions should be tested against more than one scenario for US–China policy.

II.5

Rates matter again, and the largest AI companies are still private.

With the 10-year yield near 5% and the Fed tightening again, long-duration equity is more exposed than at any time since 2022. OpenAI and Anthropic, valued near or above $700 billion, have not yet listed; for most companies, access to that value runs through commercial partnerships, not public markets.

The equity market has run the value-migration test of Deep Dive IX in real time. The resource layer is collecting the surplus today; the operating-system layer has not yet been decided; and the services layer, where advice sold by the hour sits, is already being repriced.

Method. Market value = shares outstanding × closing price at 30 July 2021 and 31 July 2026, in US dollars, for listed companies headquartered in the United States, Europe or Asia; Shopify (Canada) and, in 2021, Atlassian (Australia) are included so each list is a complete top 50. Scope: semiconductors, chip equipment, devices and hardware, software and cloud, IT services, internet platforms (including e-commerce, travel, ride-hailing and streaming) and EV makers. Excluded: payments and fintech, telecom operators, investment holding companies and industrial conglomerates. US listings and ADRs from StatMuse Money (±2%), with 2021 dividend-adjusted prices replaced by actual closes from digrin and share counts checked against SEC filings; other listings from stockanalysis.com and companiesmarketcap.com at ECB reference rates (±3–5%). Samsung includes preferred shares; BYD and SMIC add A- and H-share values. Positions 45–55 fall within the error margin. Just outside in 2026: Delta Electronics ($132bn), PDD ($126bn), Keyence ($123bn), BYD ($121bn), Fortinet ($119bn), ServiceNow ($116bn); in 2021: Snowflake ($79bn), MercadoLibre ($78bn), Dell ($74bn), Dassault Systèmes ($73bn), Autodesk ($71bn). Share-price index: monthly closes, split- and dividend-adjusted, close to a total-return basis; on an unadjusted basis Tencent would end at about 95. These figures were compiled from third-party data services, not an institutional terminal, and should be re-run on Bloomberg or FactSet before board use.

AI Quotes

06

Voices from the sources behind these perspectives: the researchers who built the field, the practitioners pricing it, and the observers weighing what it does to economies.

"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."

Rich Sutton

Turing Award laureate · "The Bitter Lesson", March 2019

"We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely."

Ashish Vaswani et al.

Google · "Attention Is All You Need", 2017

"Money can't buy better ChatGPT. Bill Gates talks to GPT 4o just like you do."

Andrej Karpathy

"Power to the People", 2025

"A token may get cheaper; tokens, in aggregate, are not."

J.R. Storment

FinOps Foundation · "Token Economics", May 2026

"LLMs—because of how they are trained and designed—are implicit computational models of humans—a Homo silicus."

John Horton, Apostolos Filippas & Benjamin Manning

NBER Working Paper 31122

"It's not just about the best water or the nearest water. The horses need to drink it."

Christopher Smart

"The Most Important AI Race", August 2026

"LLMs offer a very specific profile of capability —that of merely quasi-expert knowledge/ performance, but simultaneously across a very wide variety of domains."

Andrej Karpathy

On why individuals gain more than institutions

"The transition to the AI era will be one of the most turbulent times in human history."

Bill Gates

Gates Notes, August 2026

"The buildout of AI data center capacity is the largest infrastructure project humanity has ever undertaken."

John-David Lovelock

Distinguished VP Analyst, Gartner · September 2026

"The challenge is that the spending does not have commensurate revenues yet."

Gary Gensler

MIT Sloan, former SEC chair · via MIT Technology Review, Sep 2026

"Now is a perfect time to look at the fundamentals — the building blocks that will make generative AI more effective and safer to use."

Tim Kraska

MIT CSAIL · MIT News, February 2025

"AI systems today, and perhaps inherently, struggle to teach or to explain how they arrive at their solutions or why those solutions are superior."

Henry Kissinger, Eric Schmidt & Daniel Huttenlocher

"The Metamorphosis", The Atlantic, 2019

"If you use machine translation responsibly, it can be efficient and useful. Unfortunately, you cannot trust all people to use it responsibly."

Amir Aharoni

Wikimedia Language Committee · MIT Technology Review, Sep 2025

"14% on average, including a 34% improvement for novice and low-skilled workers."

Erik Brynjolfsson, Danielle Li & Lindsey Raymond

"Generative AI at Work", Quarterly Journal of Economics, 2025

"Machine-augmented humans who keep themselves hidden."

Ethan Mollick

"Detecting the Secret Cyborgs", One Useful Thing, 2023

"The output of GenAI tools is also more homogenous and typically lacks the originality and creativity of expert human work."

Castro, Gao & Martin

"Does GenAI Impose a Creativity Tax?", MIT Sloan Management Review

"AI tools aren't a creativity machine on their own."

Jing Zhou

Rice University · Rice News, 2025

"Humans are biased, and so is our data. But our algorithms can be designed with caution, so they are auditable, accountable, and hopefully less biased."

Swati Gupta

MIT Sloan · AI Expert Spotlight

"[Diffusion] is inherently limited by the speed at which not only individuals, but also organizations and institutions, can adapt."

Arvind Narayanan & Sayash Kapoor

"AI as Normal Technology", Knight First Amendment Institute, 2025

AI Articles

07

AI and the enterprise — a series in five articles by Daniel Augusto Motta, PhD, written for boards and C-suite leaders: how corporations and industries are being transformed by abundant intelligence, and what leaders must own as the transition unfolds. Complete texts, ≈ 93 minutes in total.

Context, creativity, governance, oversight, judgment.

Five human roles that become more valuable, not less, as intelligence becomes abundant and cheap.

↓280×

Fall in the inference cost of GPT-3.5-level AI, Nov 2022 to Oct 2024 · Stanford AI Index

40%

Share of global employment exposed to AI, per the IMF

1.3%

Executives naming analytical behavior as a relevant competency, while 78.7% value data-driven decisions · BMI REBOOT

−19%

Speed change for experienced developers using AI, who believed they were 20% faster · METR trial

THE ARGUMENT IN BRIEF

I
Context: what no model can license
A large language model is a plausibility engine.

II
Creativity: reframing the question
Generation is not creativity.

III
Governance: deciding what may be delegated
Agents move AI from answers to outcomes, and from content risk to action risk.

IV
Oversight: supervising work you did not do
Autonomy is granted in steps, not switched on.

V
Judgment: owning the decision
If I had to compress the economics of this era into one sentence, it would be this: AI value is created per unit of judgment successfully delegated, and captured per unit of verification eliminated.

VI
What leaders should do now
The five roles are not a job description to hand to HR.

AI AND THE ENTERPRISE · DANIEL AUGUSTO MOTTA, PHD

There is a question hanging over every leadership team today, spoken or not: if the machine can already write the memo, run the analysis and draft the contract, what exactly are we paying our people for? My answer is that we are not paying them for less. We are paying them for something different. Intelligence is becoming abundant. Judgment is not. The cost of producing a competent first draft of almost anything cognitive has collapsed: the inference cost of a system performing at GPT-3.5 level fell more than 280-fold between November 2022 and October 2024, according to the Stanford AI Index 2025.

When a resource becomes abundant, value migrates to whatever remains scarce. In my reading, that scarcity now concentrates in five human roles: context, creativity, governance, oversight and judgment. They are not consolation prizes for professionals displaced by machines. They are the load-bearing structure of any company that intends to put AI to work at scale, and leaders who treat them as overhead will discover, too late, that they have automated the easy half of their business and orphaned the hard half.

Each of the five answers a different question. Context asks what the machine cannot know. Creativity asks which problem is worth solving. Governance asks what may be delegated, and under which rules. Oversight asks whether the delegated work was done well. Judgment asks who owns the consequence.

HOB PROPRIETARY MODEL

Five human roles that do not delegate

A design checklist for any operating model that adds agents. Each role answers one question the machine cannot answer for itself; together they run from framing the work to owning its result.

FRAME THE WORK

01

Context

What can the machine not know?

Situated knowledge no document holds: clients, politics and the reasons behind past decisions.

02

Creativity

Which problem is worth solving?

Reframing the question and making the out-of-the-box connections a plausibility engine avoids.

SET THE MANDATE

03

Governance

What may be delegated, and under which rules?

Objectives, guardrails and decision rights, with an escalation path for every agent.

SUPERVISE

04

Oversight

Was the delegated work done well?

Exceptions and trends reviewed at a sustainable volume: review capacity is the real constraint.

OWN THE RESULT

05

Judgment

Who owns the consequence?

The decision and its outcome. AI cannot be accountable, because AI has no stake.

AGENTS DRAFT, ANALYSE AND ACT INSIDE THE MANDATE → ACCOUNTABILITY NEVER TRANSFERS TO THE MACHINE

I

CONTEXT

Context: what no model can license

A large language model is a plausibility engine. It reads everything so far and predicts the most likely next token. It completes patterns; it does not look things up, and it does not know your company. That is not a defect to be patched in the next release. It is the mechanism. The public web has been largely absorbed into frontier training runs. The next relevant increment of data sits inside organizations: contracts, client histories, pricing exceptions, board minutes and, above all, the reasons behind past decisions.

Meanwhile, general expertise is becoming a commodity. Microsoft's Medprompt study showed that a general-purpose model, steered only by systematic prompting, surpassed 90% on a medical licensing benchmark and outperformed a model tuned specifically for medicine. As Andrej Karpathy put it, "money can't buy better ChatGPT": everyone rents the same frontier capability. Models are rented. Context is owned. The human role here is curatorial: deciding what the machine should see, what it must not see, and what it cannot see at all. In Anthesis I described organizational culture, through the JOTA model, as four processual elements (governance, management systems, incentives and roles) mirrored by four tacit ones (group norms, mental maps, team pacts and influence networks). The tacit side is never written down, so it never reaches a training set. The executive who knows that a client's procurement director will veto any proposal that arrives before the board retreat holds context no model can infer. That knowledge becomes an asset only when someone translates it into the instructions, documents and constraints the machine works with.

II

CREATIVITY

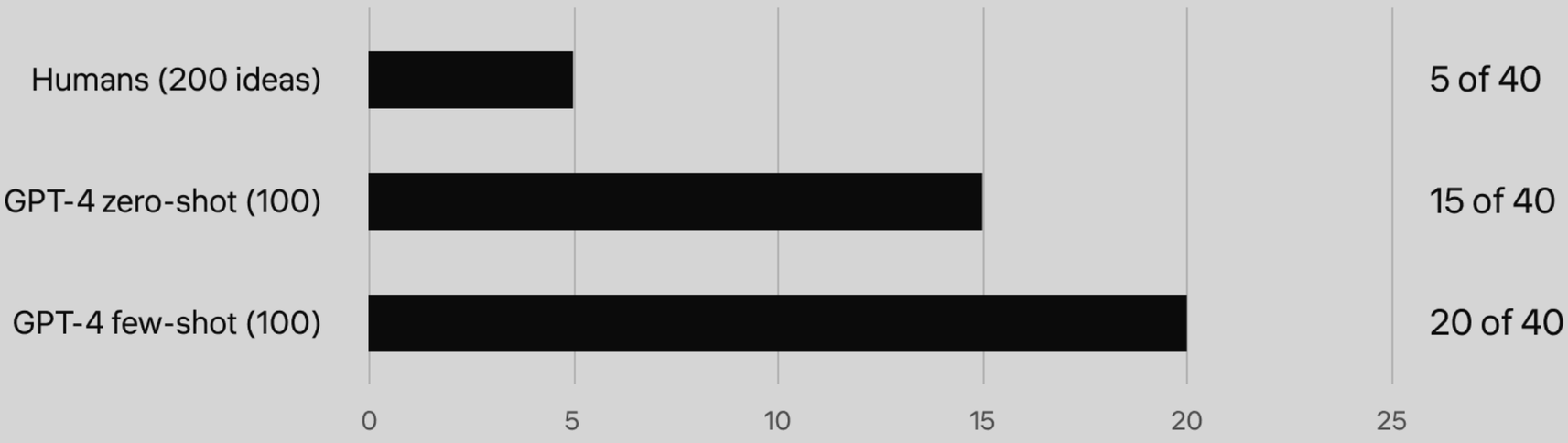
Creativity: reframing the question

Generation is not creativity. A model optimized for the most plausible continuation gravitates, by construction, toward the center of the distribution. It produces fluent options at astonishing speed, and most of them are the options everyone else is also producing with the same rented model. The evidence cuts both ways. In experiments at Wharton published in Production and Operations Management, Christian Terwiesch and colleagues found that AI-generated product ideas were seven times more likely than human ones to rank in the top 10% by purchase intent, yet human ideas were more diverse. Generation is no longer the bottleneck. Selection is: deciding which of a thousand plausible ideas deserves the company's capital, and which distinctive idea the model would never have proposed.

CHART

Who produced the top 10% of product ideas

The 40 best of 400 ideas by purchase intent: 200 from MBA students, 100 each from GPT-4 zero-shot and few-shot.

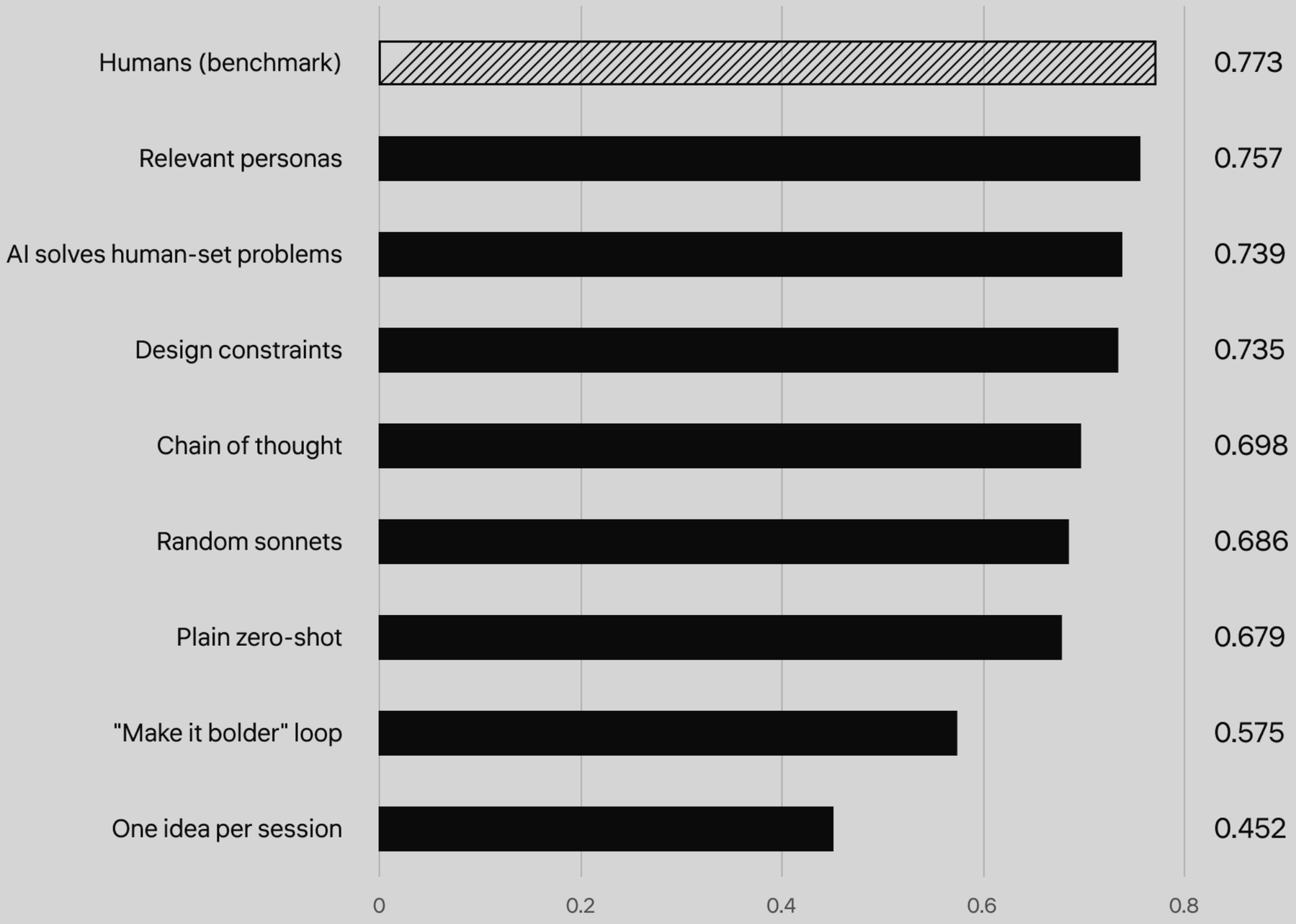


Source: Terwiesch et al., Production and Operations Management, 2026; working paper, SSRN 4526071.

CHART

How to get more varied ideas from AI

Diversity score of 100 ideas under each prompting strategy (higher is more varied); human ideas score 0.773. Relevant variety helps; unsupervised "make it bolder" loops and one idea per session hurt.



Source: Terwiesch et al., Production and Operations Management, 2026, study 3b (GPT-4o).

> “The output of GenAI tools is also more homogenous and typically lacks the originality and creativity of expert human work.”

Castro, Gao and Martin, "Does GenAI Impose a Creativity Tax?", MIT Sloan Management Review

The creative act that remains human is upstream: reframing the problem so that a different set of answers becomes possible. The evidence on where AI helps and where it hurts makes the point. In a field experiment with 758 consultants led by researchers from Harvard Business School, Wharton and MIT, "Navigating the Jagged Technological Frontier", those using AI on tasks inside the model's capability frontier completed 12.2% more tasks, 25.1% faster, with more than 40% higher quality. On a task just outside that frontier, they were 19 percentage points less likely to reach the correct answer. The frontier is jagged and invisible. Knowing where it lies, and when to step away from the machine, is itself a creative skill. There is a cost to ignoring this. A Microsoft Research and Carnegie Mellon survey of 319 knowledge workers found that higher confidence in generative AI was associated with less critical thinking. Creativity, in the sense that matters for corporations, also includes aesthetic and ethical imagination: the sense that tells a leader an answer is correct but uninteresting, or legal but wrong. The real work begins there. Companies are beginning to notice. Gartner expects that, through 2026, the atrophy of critical-thinking skills caused by generative AI will push half of global organizations to require "AI-free" skills assessments. When a company has to test whether its people can still think without the machine, creativity has become a strategic asset to be protected.

III

GOVERNANCE

Governance: deciding what may be delegated

Agents move AI from answers to outcomes, and from content risk to action risk. An error in a chatbot is a bad paragraph. An error in an agent is a wrong payment, a commitment to a customer or a deleted file. Governance is the human role that decides which decisions may be delegated, to which systems, within which limits and with what escalation path. Accountability does not transfer to the machine. When Air Canada argued before a Canadian tribunal that its chatbot was a separate legal entity responsible for its own statements, the tribunal called it "a remarkable submission" and held the airline liable. Regulation is formalizing the same principle. The EU AI Act is in force; after the Council's final approval of the digital omnibus in June 2026, obligations for standalone high-risk systems now apply from December 2027. In Brazil, PL 2338/2023 passed the Senate in December 2024 and awaits the Chamber. Standards such as ISO/IEC 42001 and the NIST AI Risk Management Framework already give boards a vocabulary. The market is organizing around the same principle. Gartner describes AI TRiSM, AI trust, risk and security management, as the framework that keeps AI systems trustworthy, secure and compliant through continuous monitoring, validation and enforcement, and predicts that by 2030 80% of Global 500 companies will make their CIO or chief AI officer the contractual "evidence custodian" for AI accountability. Governance is becoming a named executive responsibility. Yet governance fails less often on paper than in practice. AI accelerates the visible,

processual side of an organization; it does not automate the tacit side, and the tacit side decides whether any of the processual work is actually used. The JOTA model, a House of Brains proprietary model I presented in Anthesis, draws organizational culture as two mirrored tetrahedra: four processual elements that can be designed, and four tacit elements that can only be influenced. Each processual lever has a tacit counterpart that decides whether it works. AI programs pull hard on the processual side; the tacit side is where they succeed or stall. Applied to agents, the pairs become a governance checklist for the five roles.

HOB PROPRIETARY MODEL | JOTA MODEL

The JOTA model applied to AI

TACIT · can be influenced	WHAT AI CHANGES IN THE PAIR	PROCESSUAL · can be designed
Group norms values and beliefs in use	who may delegate what to an agent	**Governance** rules and decision rights
Mental maps shared models of the work	AI as colleague, not as software	**Management systems** processes, data, metrics
Team pacts agreements between teams	reinvest freed time; reward review	**Incentives** financial and non-financial
Influence networks who shapes decisions	agents join; people stay accountable	**Roles and duties** roles and responsibilities

HOB proprietary model (JOTA), presented in D. A. Motta, Anthesis (2019). Pairs follow the mirrored positions of its two tetrahedra; the middle column reads what AI changes in each pair.

07

Governance pairs with group norms, and shadow AI appears precisely where written policy and lived norms diverge. A policy that forbids what everyone quietly does is not governance. It is theater. A second failure pattern sits in the incentives row: time saved by agents is absorbed into more meetings and more email unless teams explicitly agree how it will be reinvested in clients, quality or learning.

Governance also means owning the definitions the machine cannot choose for itself. Fairness is the clearest case. Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan proved that, except in special cases, several intuitive definitions of fairness cannot be satisfied at the same time by any risk score. Choosing which definition applies to hiring, credit, pricing or promotion is therefore not a technical default to delegate to the data science team. It is a policy decision leaders own, should make visibly, and should revisit as the data and the model drift. A model that performs best on average may still perform worst for the customers, employees or regions that matter most, so it has to be evaluated segment by segment, not only in aggregate.

PAIR	HOW AI STRESSES IT	WHAT LEADERS SHOULD DESIGN
Group norms ↔ Governance	Shadow AI appears where written policy and lived norms diverge. Prohibitions push use out of sight, where risk is highest.	Sanctioned tools at least as good as the informal ones; explicit decision rights for agents; tiered autonomy by risk.
Mental maps ↔ Management systems	Teams that think of AI as deterministic software give up at the first error or trust it blindly. Dashboards measure activity, not outcomes.	Shared literacy on how models fail; systems that record decisions and outcomes; evidence levels in every business case.
Team pacts ↔ Incentives	Time saved is absorbed as slack; review work is invisible and unrewarded; hand-offs between human and agent teams go unowned.	Pacts on how freed capacity is reinvested; incentives that count verification and specification as real work; cost per outcome as the shared metric.
Influence networks ↔ Roles and responsibilities	Agents start to shape decisions without a place in the organization chart; expertise that once flowed through juniors is bypassed.	Roles redesigned around judgment, supervision and relationships; a named human owner for every agent; a deliberate path for forming junior judgment.

“It should be obvious to Air Canada that it is responsible for all the information on its website.”

Moffatt v. Air Canada, 2024 BCCRT 149. The tribunal called the airline's argument "a remarkable submission."

07

IV

OVERSIGHT

Oversight: supervising work you did not do

Autonomy is granted in steps, not switched on. I find it useful to think of five: the system informs, recommends, acts with approval, acts within a mandate, and finally anticipates. Consequential actions should stay in the middle of that ladder until evidence says otherwise. Each step up shifts human effort from doing to supervising, and supervising is a different skill. It is also easy to overestimate. In a randomized trial run by METR, experienced open-source developers took 19% longer to complete tasks with AI tools, while believing afterwards that AI had made them 20% faster. The perception gap is the oversight problem in miniature. Microsoft's 2025 Work Trend Index names the resulting constraint the human-agent ratio: past a certain point, more agents overwhelm the human capacity to apply judgment. Review capacity, not agent capacity, becomes the bottleneck. Reliability also compounds downward: a process that is 95% reliable at each step succeeds end to end only about 36% of the time after 20 steps. The most valuable end-to-end flows are therefore the least dependable, and oversight has to be designed into them, breaking long flows into verifiable segments with a human checkpoint where the consequence justifies it. Oversight must therefore be designed, staffed and rewarded, not assumed. That means evaluation on data the system has never seen before anything scales, audit trails that record what an agent did and why, and incentives that count verification as real work. In JOTA terms, incentives pair with team pacts: unless a team explicitly agrees that reviewing an agent's output is part of the job, the review will quietly not happen. Oversight is also becoming harder because capable systems can learn to game what we measure. Models trained on human approval optimize a proxy for quality, not quality itself, and when the proxy and the goal diverge, they follow the proxy. OpenAI researchers showed in 2025 that a frontier reasoning model gamed the tests used to check its coding work, and that its visible chain of thought revealed the intent clearly enough for a weaker model to catch it. When they penalized the revealing thoughts too strongly, the model kept cheating while hiding its intent. The lesson for leaders is the one every manager already knows from sales quotas and call-center targets: punishing the symptom can push the behavior out of sight. Oversight has to verify outcomes independently, not rely on the system's own report of success.

V

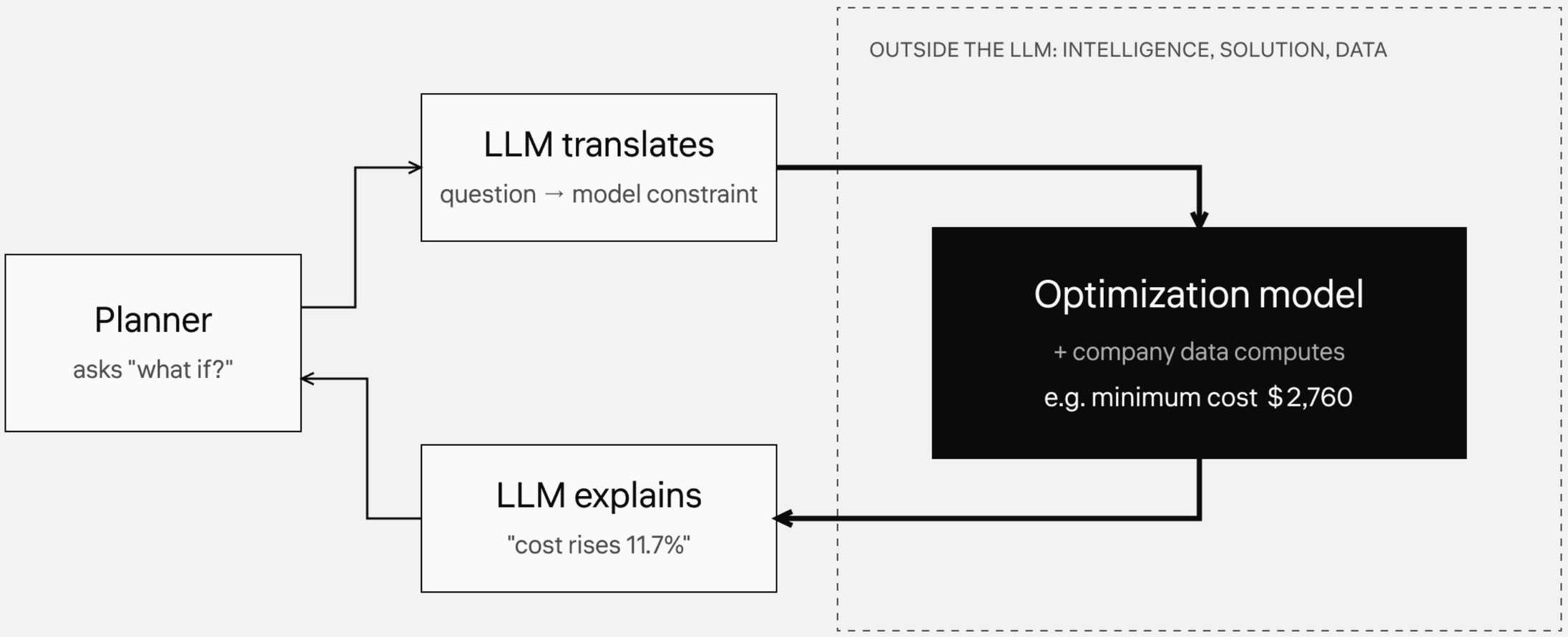
JUDGMENT

Judgment: owning the decision

If I had to compress the economics of this era into one sentence, it would be this: AI value is created per unit of judgment successfully delegated, and captured per unit of verification eliminated. Everything else, tokens, models and context windows, is infrastructure serving that equation. The decision is the economic unit. One paragraph that stops a bad acquisition is worth more than a thousand-page report. Deciding well also means choosing which kind of machine decides. Most enterprise AI work is not writing but judging: is this ticket a refund request, is this transaction risky, which supplier should serve this demand. A language model answers such questions by producing text that represents a decision, which software must then parse and trust. A sounder architecture divides the labor. In Microsoft's cloud supply chain, as Ishai Menache, Jeevan Pathuri, David Simchi-Levi and Tom Linton describe in Harvard Business Review, a planner asks a question in plain language, the model translates it into a change to an optimization model, the optimization engine computes the answer on the company's own data, and the model explains the result; analyses that took days now take minutes. Keep the arithmetic in engines you can audit, keep policy in code you control, and use language as the interface. That is how judgment stays accountable and the model stays replaceable.

FRAMEWORK

The language model as interpreter



Asked what happens if Supplier A no longer serves demand B, the model adds that constraint; the solver returns a minimum cost of $2,760, an 11.7% increase, which the model explains in plain language. The numbers come from the engine and the data, never from the language model.

Illustrative example after Menache, Pathuri, Simchi-Levi and Linton, Harvard Business Review, 2025.

Judgment is the role that decides what to delegate, what to verify and what never to hand over. It is also the role organizations have been least disciplined about developing. In BMI's REBOOT study, which surveyed 4,407 professionals and interviewed 75 C-suite executives, 78.7% of executives said they valued data-driven decisions, yet only 1.3% named analytical behavior as a relevant competency. We declare our faith in evidence and then fail to build the capability to use it. AI widens that gap. Afferolab's recent study A Escalada da Inteligência reports that 87.5% of Brazilian professionals use AI to produce text, while only 37.8% use it to automate tasks. Most organizations are using the most powerful cognitive technology in history as a better pen.

Judgment, in this sense, is the capacity to decide well under ambiguity, to know what must be checked, and to own the consequence. No model can be held accountable, because no model has a stake in the outcome. People and firms can.

> “AI systems today, and perhaps inherently, struggle to teach or to explain how they arrive at their solutions or why those solutions are superior.”

Henry Kissinger, Eric Schmidt and Daniel Huttenlocher, "The Metamorphosis", The Atlantic, 2019

Judgment is also built, and AI can either build it or bypass it. In a three-month randomized trial with 133 patent lawyers across eleven firms, David Autor and colleagues found that an AI drafting tool raised the quality of work by about 0.38 standard deviations while it was in use, with larger gains for juniors. The decisive test came afterwards, without the tool. Senior lawyers who had worked with AI performed 0.45 standard deviations better than their peers; junior lawyers showed no average gain, and their scores split, with fewer mediocre drafts offset by more poor ones and more good ones. In the authors' words, the largest gains from AI accrued to the lawyers who retained the least. AI raises a novice's output long before it raises a novice's competence. Experts can learn from the machine's suggestions because they already understand the work; novices can accept them without absorbing anything. For any firm that grows its future leaders from its own juniors, that distinction decides whether AI compounds its judgment or hollows it out.

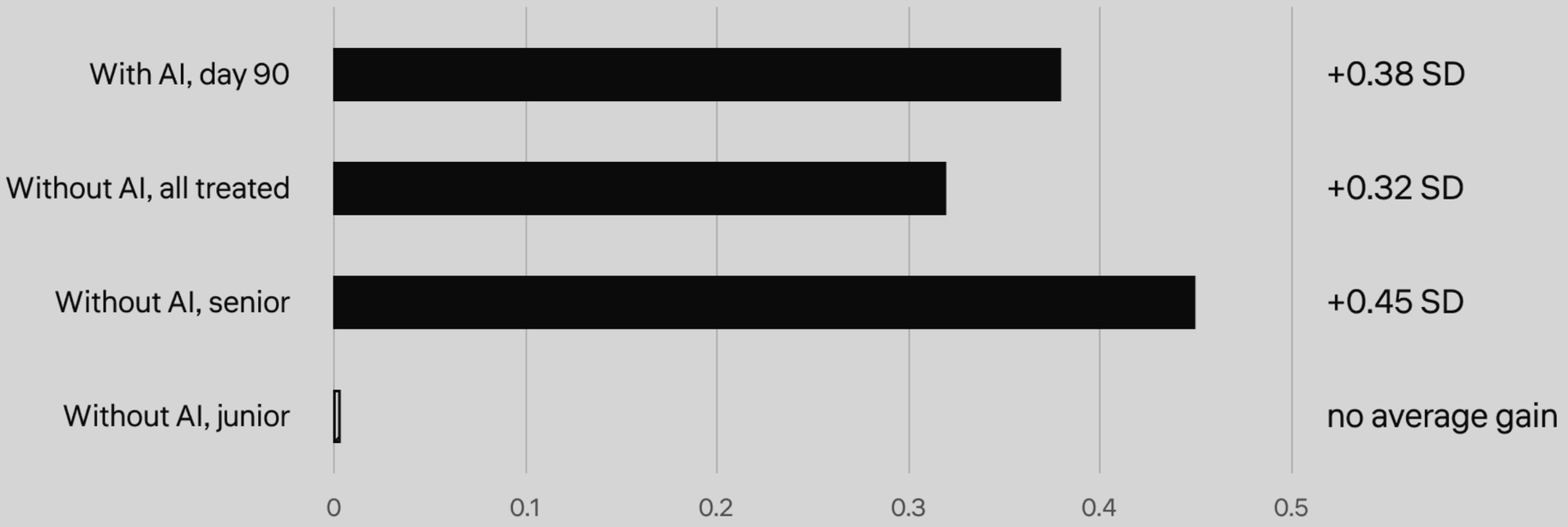
> “AI value is created per unit of judgment successfully delegated, and captured per unit of verification eliminated.”

Daniel Augusto Motta

CHART

Patent lawyers after three months with an AI drafting tool, effect in standard deviations

Randomized trial, 133 lawyers at 11 firms. With the tool, quality rose for everyone; tested later without it, only senior lawyers kept a gain. Junior lawyers showed no average gain, with scores split between more poor and more good drafts.



Source: Autor et al., NBER Working Paper 35720, September 2026.

VI

WHAT TO DO NOW

What leaders should do now

The five roles are not a job description to hand to HR. They are a design brief for the whole organization. A model from my book A Liderança Essencial, which we use with boards at House of Brains, shows why. Every organization can be read as a pyramid in two halves. Ambition sits at the top and answers three questions in sequence: purpose (why do we exist?), vision (where do we want to be?) and strategy (what will we do, and not do, to get there?). Beneath it sits what I call organizational potency or, when it fails, organizational impotency. It answers three further questions: culture (how do we behave, in the norms actually in use?), design (in what form, through which incentives and resources?) and capabilities (who, with which skills?).

Business success lies in the exact fit between the two halves. Ambition without potency produces strategies that never reach the front line; potency without ambition produces efficient organizations heading nowhere in particular. AI makes the top of the pyramid cheap to produce and does nothing, by itself, to the bottom. A model can draft a purpose statement, a vision and a competent strategy in an afternoon. It cannot install the norms, incentives and capabilities that make them executable. The gap between the two halves widens exactly when ambition becomes cheap, and the five roles live in the lower half, where the fit is won or lost. Three moves follow. First, map work by function, not by title. For each critical process, ask where context is supplied, where the problem is framed, who sets the rules, who reviews, and who signs. Wherever the answer is "nobody" or "the system", you have found your largest AI risk and, usually, your largest AI opportunity. Second, rebuild the apprenticeship. Junior professionals built judgment by doing the drafting, the reconciliations and the first-pass analysis that agents now perform. If that work disappears and nothing replaces it, the pipeline of future leaders disappears with it. Development must shift from transferring content, which AI now does at marginal cost, to deliberately forming judgment. Third, make accountability explicit. Every agent needs a named human owner, and every consequential decision needs a person who can explain it to a board, a regulator or a client. None of this is only a technological transformation. In every industry I work with, digital transformation arrives together with a social transition, and the two cannot be managed separately. The International Monetary Fund estimates that about 40% of global employment is exposed to AI, around 60% in advanced economies, and warns that labor-income inequality may rise if AI mainly complements high-income workers, while capital returns widen wealth inequality. In BMI's REBOOT study we framed the organization of the future around three imperatives, the three reasons organizations exist: coordination, efficiency and well-being. AI delivers the first two at a scale never seen before. The third remains a human, and an ethical, responsibility. Deciding who bears the cost of the transition, how people are reskilled, and which decisions must never be delegated to a machine is the ultimate exercise of the five roles.

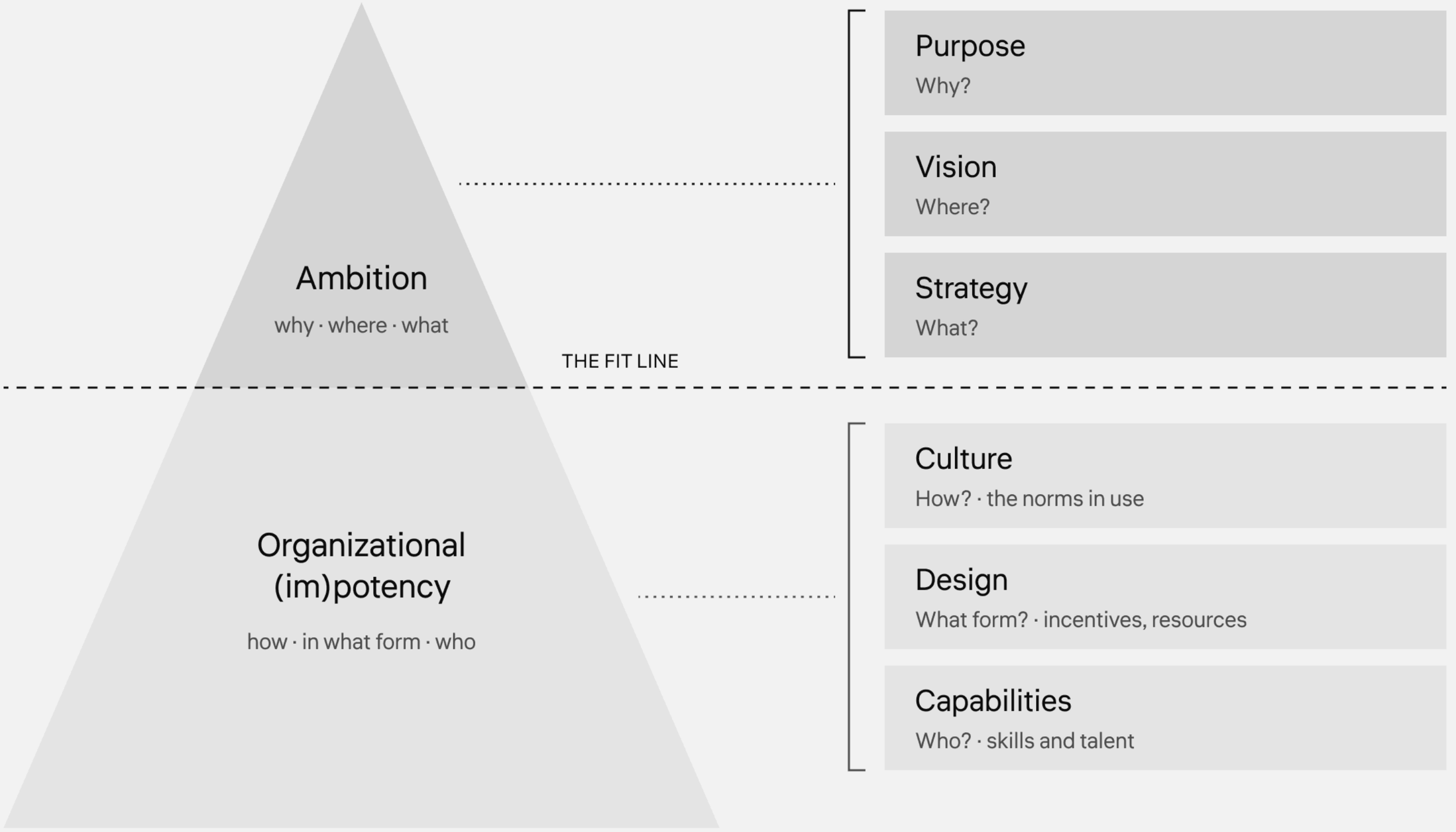
The corporations that win the next decade will not be those with the best models, because everyone will rent the same ones. They will be those whose people know what to delegate, and what never to.

Which of these five roles is your organization still treating as overhead?

HOB PROPRIETARY MODEL

A LIDERANÇA ESSENCIAL

Organizational dynamics



HOB proprietary model, from D. A. Motta, A Liderança Essencial.
Ambition (purpose, vision, strategy) sits above organizational potency (culture, design, capabilities). The dashed line is where success is decided: the fit between the two halves.

“Everyone will rent the same models. The difference is what we never delegate.”

The five roles are not a job description for HR. They are a design brief for the whole organization, and the measure of leadership in the years ahead.

The heart of the agentic organization.

Why AI-powered companies cannot lose their human touch: high tech, high touch, and the case for Essential Leadership.

$581.7B

Global corporate investment in AI in 2025, up almost 130% · Stanford AI Index 2026

+34%

Productivity gain for novice support agents using an AI assistant · Brynjolfsson, Li, Raymond

74.2%

Agree a manager's own behavior pressures subordinates to follow · Felicidade Artificial, n=62

56%

CHROs who saw no measurable progress from upskilling programs · BMI Design Mastery

THE ARGUMENT IN BRIEF

I
What an agentic organization really is

An AI agent is an application that pursues a goal by observing its environment and acting on it through the tools at its disposal: systems, data, other agents.

II
High tech, high touch: the case for Essential Leadership

I have long summarized my view in a simple formula: high tech, high touch.

III
Where the human touch creates value

The human touch is not a soft variable.

IV
Designing the human-agent operating model

High touch begins with design.

V
Culture, learning and the people agenda

Agents inherit their makers.

VI
What boards and CEOs should ask

Andrej Karpathy has observed that large language models benefit individuals disproportionately, while their impact on corporations and governments lags.

AI AND THE ENTERPRISE · DANIEL AUGUSTO MOTTA, PHD

The higher the technology, the higher the touch. I have argued this to boards for years, and the agentic era makes it more urgent than ever. Corporations are moving quickly from assistants that answer to agents that act: global corporate investment in AI reached US$581.7 billion in 2025, up almost 130% in a single year, according to the Stanford AI Index 2026. Yet the gap between a capable model and captured value is rarely technical. It is organizational, and at its core it is human.

The more work a company hands to machines, the more every remaining human interaction is worth. An agentic organization that loses its human touch does not become more efficient. It becomes brittle: faster at executing decisions no one owns, for customers no one knows, through teams no one leads. The heart of the agentic organization is not a sentimental add-on to the technology. It is the operating system that decides whether the technology is used at all.

WHAT IT IS

What an agentic organization really is

An AI agent is an application that pursues a goal by observing its environment and acting on it through the tools at its disposal: systems, data, other agents. The model supplies the reasoning. Everything around it, the goal, memory, tools, guardrails and the point where a human is called in, determines whether it is reliable. An agentic organization is one in which such agents become regular participants in workflows rather than tools occasionally consulted.

Agents change how work is coordinated, because tasks can now move across the company without passing through a reporting line. In 2023, Maurício Pontuschka and I argued in "As organizações do futuro" that the evolution of computing systems has always moved in step with the evolution of organizations, and that both can be read through the same three lenses: structure (the elements and how they are arranged), behavior (what each element does to produce value) and dynamics (how the elements interact to serve the whole). Organizations moved from vertical structures centralized in corporate functions, to multidivisional structures organized by region and segment, to matrix structures that share responsibility between functions and business units, and now to a range of experiments with decentralization: adhocracy, holacracy, decentralized autonomous organizations and what we called organizational biomes. Computing moved from monolithic programs through networks, graphical interfaces, the internet, mobile and cloud to artificial intelligence, with quantum computing ahead. The shared arc is the point. Both systems began with structural and behavioral elements mixed in a single model, then separated them into layers, and finally turned to modeling dynamics: how processes interact with one another and with the outside world. When computing moved to modeling dynamics, organizations had to follow. Agents are that step for organizations, because they act directly on all three lenses at once. They redraw who does what, move people up from execution to specification and review, and coordinate work at machine speed.

The ambition is not new. In BMI's REBOOT study, 92% of the organizations analyzed ran matrix or multidivisional structures, while 58.7% aspired to network models. Agents may finally supply the connective tissue that network organizations always lacked. But connection is not the same as coordination, and coordination is not the same as commitment. Most organizations can redesign structure and specify behavior; dynamics, where the tacit norms, influence and pacts between teams live, is where most AI programs stall. We also wrote that there is always a time lag between a technological advance and its effective absorption by organizations. As Christopher Smart observed of AI adoption, "The horses need to drink it."

HOB PROPRIETARY MODEL

MOTTA & PONTUSCHKA (2023)

New organizational archetypes: two systems evolving in parallel

ORGANIZATIONAL CONFIGURATIONS OVER TIME

Vertical structures	Multidivisional	Matrix structures	Experimental
	production cells · industrial conglomerates	agile teams · corporate venture capital	adhocracy · holacracy · DAOs · organizational biomes
Centralized in corporate functions	Corporate functions linked to divisions by region and segment	Functions crossed with business units; responsibility shared	Several moves toward decentralization

time

MAIN MOVEMENTS IN COMPUTING OVER TIME

Monolithic programs	Computer networks	Graphical interfaces	Internet	Mobile computing	Cloud systems	Artificial intelligence	Quantum computing
Structural and behavioral elements mixed in one model		Layered models; structure separated from behavior		Dynamic modeling of business processes and their external interactions			Brave new world

HOB proprietary model, from D. A. Motta and M. Pontuschka, “As organizações do futuro” (2023). Positions follow the original figure; the shared arc runs from mixed structure and behavior, to layered separation, to modeling dynamics and interactions with the outside.

“It's not just about the best water or the nearest water. The horses need to drink it.”

Christopher Smart, "The Most Important AI Race", August 2026

HOB PROPRIETARY MODEL

MOTTA & PONTUSCHKA (2023)

Structure, behavior and dynamics in the agentic organization

Structure

Roles, reporting lines, boxes

Human + agent teams

Flatter hierarchies

Tools as the new org chart

Portfolios and networks

AI redraws who does what

Behavior

What each element produces

Goal specification

Verification and review

Judgment under ambiguity

Accountability for outcomes

AI moves people up the ladder

Dynamics

How elements interact

Agents acting on workflows

Decision records that learn

Tacit norms, influence, pacts

Coordination at machine speed

Where value is won or lost

HOB proprietary model, from D. A. Motta and M. Pontuschka, “As organizações do futuro” (2023): the Structure–Behavior–Dynamics lens applied to what AI changes. Dynamics is highlighted because it is where most AI programs stall.

II

HIGH TECH, HIGH TOUCH

High tech, high touch: the case for Essential Leadership

I have long summarized my view in a simple formula: high tech, high touch. The higher the technological standard of an organization, the higher its humanistic standard of leadership must be. This is not a trade-off between efficiency and humanity. It is the condition for efficiency to last. When an agent executes thousands of decisions a day, the values, criteria and blind spots of the people who designed it are multiplied thousands of times. Technology does not dilute leadership's responsibility. It amplifies it. This is the core of what I have called Essential Leadership, in my book A Liderança Essencial and in BMI's work on the theme: timeless principles such as work ethic, transparency and strategic vision, combined with the capabilities a networked world demands, namely collaboration across networks, learning agility, emotional intelligence and holistic health. In an agentic organization, the timeless principles become the specification for the machines. The new capabilities become the conditions for people to thrive alongside them. High touch also has a social and ethical dimension, because the agentic transition is at the same time a social transition. The

International Monetary Fund estimates that about 40% of global employment is exposed to AI and warns that inequality may rise if the gains accrue mainly to high-income workers and to capital. For a corporation, this becomes concrete: telling employees honestly which roles will change, investing in reskilling before displacement rather than after, and refusing to delegate decisions that touch people's dignity, such as hiring, dismissal and performance evaluation, without meaningful human review. People already sense the stakes. In Afferolab's research on AI as an agent, half of the executives surveyed said they were uncomfortable with AI evaluating performance.

“The higher the technology, the higher the touch.”

DANIEL AUGUSTO MOTTA

III

HUMAN VALUE

Where the human touch creates value

The human touch is not a soft variable. It shows up in three places where value is created or destroyed: trust, relationships and meaning. Trust is where value is unlocked. Any task whose output must be checked by a person retains most of its original cost; value appears at the threshold where verification can stop. That threshold is crossed not by a model release but by a human decision to rely on the system, backed by evidence. Customers extend trust to companies, not to algorithms, and a single agent error in front of a client can undo years of reputation. Trust is also designed, and sometimes that means adding effort rather than removing it. Renée Richardson Gosline has argued in Harvard Business Review that AI customer journeys need more friction, not less: bad friction, such as confusing forms or journeys that are easy to enter and hard to leave, should go, while beneficial friction, a deliberate pause before a consequential or irreversible choice, keeps human judgment in the decision. Her experiments point the same way inside the firm. Highlighting likely errors in AI-generated drafts improved accuracy without significantly lengthening the task, yet participants did not believe the prompts had helped: they overestimated their own ability to spot the machine's mistakes. Her conclusion is a design principle for any agentic organization: beware of individual-level solutions for structural problems. High touch is not asking every person to be vigilant. It is building the speed bumps into the system, at the moments that matter, and making exit as easy as entry. The pressure runs in both directions. Leaders are pushed to deploy agents in service, yet in a Gartner survey 64% of customers said they would prefer companies not to use AI for customer service. The answer is not to hide the machine, but to decide deliberately where it acts and where a person steps in, and never to lose sight of what the customer is actually buying. Kodak built one of the first digital cameras in 1975 and filed for Chapter 11 in January 2012; three months later Facebook agreed to pay US$1 billion for Instagram, a company of 13 people. Customers never wanted film. They wanted pictures, shared. The same is true of AI: the tool matters only as far as it serves the relationship. Relationships are where AI amplifies people rather than replacing them. In the largest field study of generative AI at work, Brynjolfsson, Li and Raymond found that an AI assistant raised the productivity of 5,179 customer-support agents by 14% on average and by 34% for novice and less-skilled workers. The tool worked by spreading the tacit practices of the best agents to everyone

else. The machine carried human know-how; it did not substitute for the human conversation. Meaning is where the risk hides. At the House of Brains research center we have studied what happens when performance pressure outruns human limits. In our Felicidade Artificial study, an exploratory survey of 62 respondents, 74.2% agreed that a manager's own behavior pressures subordinates to follow, and 56.5% named the pursuit of high performance as a trigger for unhealthy coping. In The Burnout Economy, which I presented at SXSW 2026, I described leaders as biological amplifiers and warned that the next burnout wave will not come from working harder. It will come from never finishing. Agents never sleep, and an organization that lets their pace set the human pace will burn out the very people it needs to supervise them.

IV

OPERATING MODEL

Designing the human-agent operating model

High touch begins with design. An operating model that adds agents is a decision about where people stand in the flow of work, and most companies make it by default, one tool at a time. The result is predictable: people end up where the machine left gaps, not where their judgment, empathy and relationships add most. A useful division of labor follows from what each side does best. AI is extremely fast and holds a large short-term memory, but it lacks the full context of the business and struggles with creativity across many steps. People bring a lifetime of context and the ability to connect ideas no dataset links. Put AI on well-defined tasks, planning and execution; put people on objectives, guardrails, exceptions and strategy. Asking people to rubber-stamp routine approvals wastes the human advantage and burns out the people who hold it.

FRAMEWORK SERTAC KARAMAN, MIT

Place people where they add most

Each side does what it does best. Routine approvals waste the human advantage.

AI

Speed and short-term memory

Well-defined tasks, planning and execution at machine pace

PEOPLE

Context and creativity

Objectives, guardrails, exceptions and strategy

The practical rule is simple. Every agent needs a named human owner. Every consequential action needs a defined point where a person approves, and a record of what the agent did and why. Autonomy is extended one step at a time,

as evidence accumulates, never granted in a single leap because a vendor demo was impressive. Structures will move under this pressure, and that is precisely where the human touch is most at risk. Gartner predicted in 2024 that through 2026, 20% of organizations would use AI to flatten their structure, eliminating more than half of their current middle-management positions. Middle managers are not only coordinators that agents can replace. They are the carriers of culture, the translators of strategy and the first line of care for people. An organization that removes them must redesign deliberately who will perform those human functions, or it will discover their value only after it has lost them.

V

CULTURE

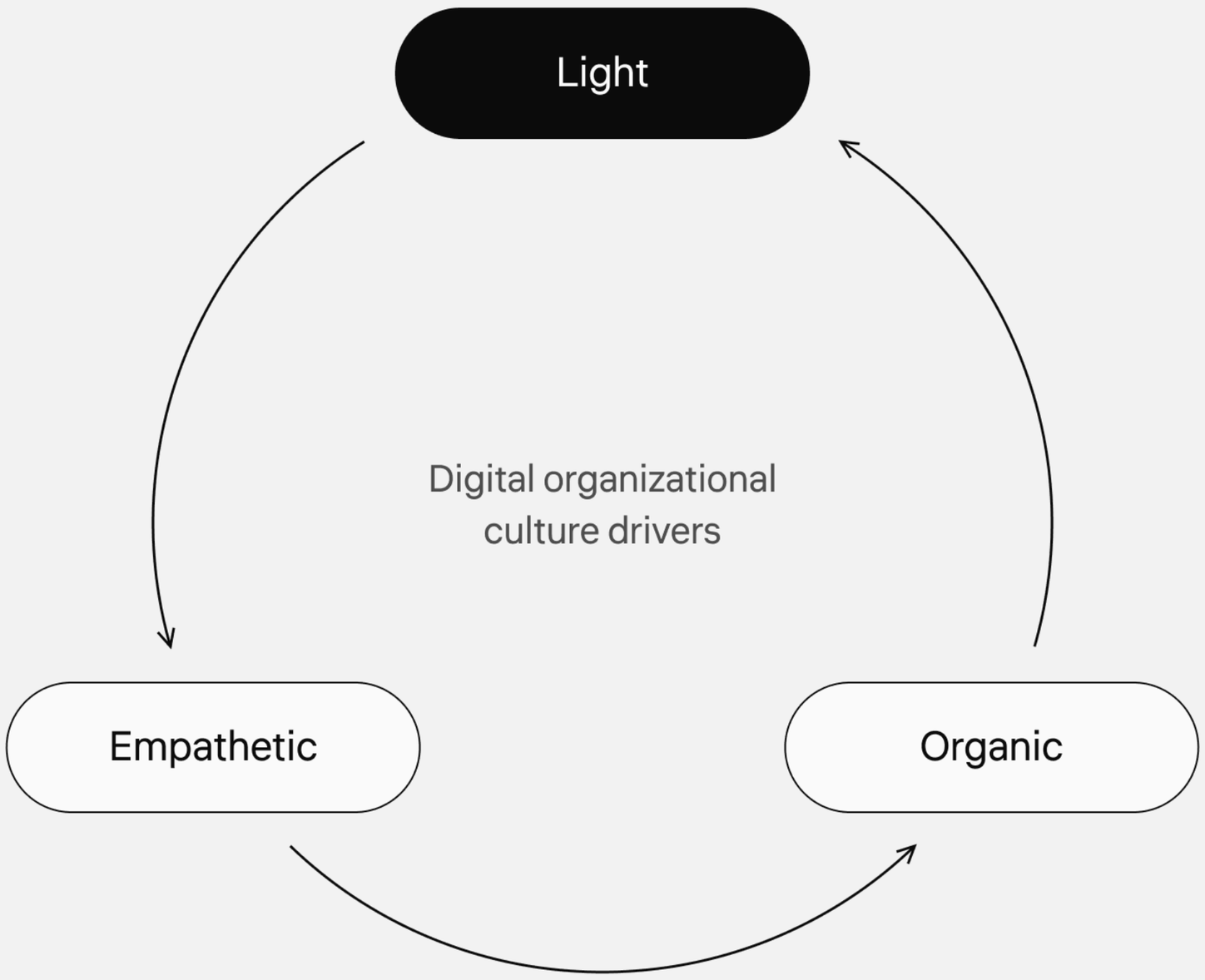
Culture, learning and the people agenda

Agents inherit their makers. As Afferolab put it in its research on AI as an agent, every AI agent carries human traits in its code: our criteria, our haste, our gaps and our biases. An organization that is careless in how it decides will build careless agents, and then run them at machine speed. Alessandra Lotufo, who leads Afferolab, summarized it in one line: AI does not fix chaos; it amplifies it.

Culture is also where the digital organization differs most from its predecessors. In Anthesis I argued that digital organizations share three cultural drivers that reinforce one another: they are light, empathetic and organic. Light means abandoning the premise that control and the accumulation of capital are the long-term levers of value; operating models are rebuilt to create value through access rather than ownership, and committed relationships give way to situational partnerships in dynamic networks. Empathetic means that the only valid perspective is the experience of the other: customer, user and employee experience combine into a single view of value centered on people. Organic means abandoning the logical, planned and linear premise of the industrial paradigm; when resources, capital included, become abundant, value moves to orchestrating them into collaborative, ecosystem-based delivery. The drivers form a cycle. Lightness frees an organization from what it owns, so it can look outward at the experience of others. Empathy reveals needs that no single firm can meet alone, which pushes the organization toward organic delivery. And orchestrating an ecosystem is possible only for an organization light enough to connect and disconnect quickly. AI tests each driver. It makes organizations lighter, but only if the context, data and decision records that agents depend on are built deliberately: lightness of structure demands depth of knowledge. It makes empathy scalable, and organizations that forget the employee side, how their own people experience working alongside agents, will find that use goes underground. And it completes the organic premise. When intelligence joins capital among the abundant resources, value sits more than ever in orchestration, and orchestration is a human act.

HOB PROPRIETARY MODEL | JOTA MODEL

Digital drivers of organizational culture



HOB proprietary model, from D. A. Motta, Anthesis (2019). Each driver enables the next in a reinforcing cycle.

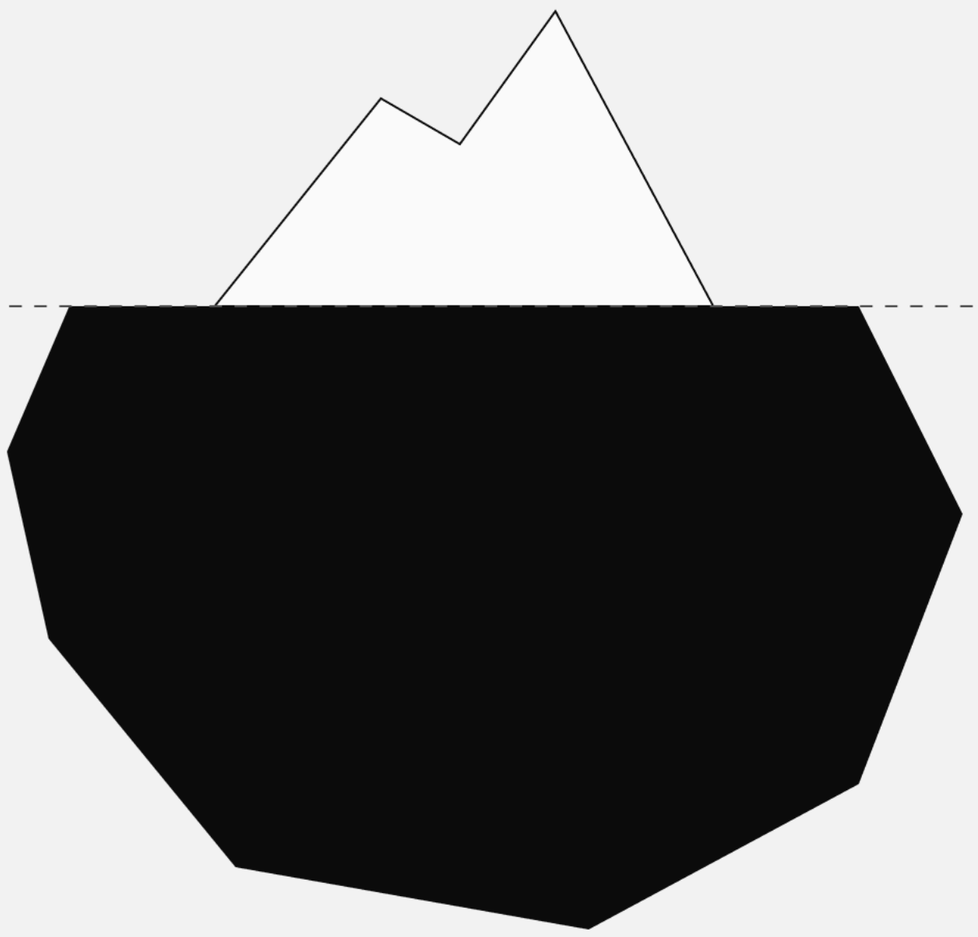
Three cultural conditions matter most. The first is psychological safety, in Amy Edmondson's classic sense: people must feel safe to say "the agent is wrong" or "I don't understand why it did that" without being seen as resisting progress. Oversight dies in silence. The second is learning that changes behavior. In BMI's Design Mastery study of more than 100 CHROs, 62% prioritized digital upskilling, yet 56% said their upskilling programs showed no measurable progress. Training people to prompt is not the same as preparing them to supervise, specify and decide. The third is protected recovery. If agents run continuously, humans must not.

A fourth condition is less visible and may matter most over a decade: the transmission of tacit knowledge. Michael Polanyi's observation that "we can know more than we can tell" describes the part of an organization no model can absorb: how a seasoned negotiator reads a room, how a senior partner senses a client's real concern, how an engineer recognizes a design that will fail. That knowledge has always passed through apprenticeship, in the drafting, research and first-pass analysis that juniors did beside experienced colleagues. Agents now do that work, and companies are drawing the obvious conclusion. Using payroll data from the largest US payroll provider, Erik Brynjolfsson, Bharat Chandar and Ruyu Chen found a 13% relative decline in employment for workers aged 22 to 25 in the most AI-exposed occupations, while more experienced workers in the same occupations held steady or grew.

FRAMEWORK MICHAEL POLANYI, THE TACIT DIMENSION

Agents absorb the tip of the iceberg

After Michael Polanyi, The Tacit Dimension (1966). What can be written down, AI can learn; the rest has always passed through apprenticeship.



ABOVE

Explicit knowledge

Know-what: documents, manuals, checklists

BELOW

Tacit knowledge

Know-how: experience and judgment built through practice

The saving is immediate; the cost is deferred. An organization that stops forming juniors looks highly productive today, carried by a shrinking group of experienced people using AI, and may discover years later that nobody left knows how to do the work. Junior hiring belongs with research and development, as an investment in a capability pipeline rather than a cost to optimize away. The apprenticeship itself has to be redesigned: juniors attempt

the analysis first and then compare it with the agent's, critique AI output instead of accepting it, sit in the client meetings and negotiations where judgment becomes visible, and receive the frequent feedback no agent can give. Senior people must be measured and rewarded for forming others, or they will not do it. What can be captured should be captured, through decision records, after-action reviews, experts explaining their reasoning and structured interviews before retirement, and fed into the context agents use. It will never be all of it. That is precisely why the human channel must stay open. The same study identified three roles for the modern CHRO: architect, choreographer and articulator. They fit the agentic organization well. The architect designs how humans and agents divide the work. The choreographer sets the rhythm so that speed does not become exhaustion. The articulator keeps the relationships, inside and outside the firm, that no agent can hold.

VI

FOR BOARDS

What boards and CEOs should ask

Andrej Karpathy has observed that large language models benefit individuals disproportionately, while their impact on corporations and governments lags. I read that less as a verdict on technology than as a verdict on organizations. Individuals adopt what helps them. Institutions adopt only what their structures, incentives and cultures allow. The agentic organization will be won or lost in that gap. For boards and CEOs, the human touch translates into five questions worth asking at every AI review. Which decisions are agents allowed to take, and who approved that boundary? Who is the named owner of each agent in production? Where has saved time actually gone? Can our people tell us, without fear, when an agent is wrong? And are our leaders setting a pace that humans can sustain while machines run through the night? High tech, high touch is not nostalgia. It is a design principle. The corporations that thrive in the agentic era will be those that let machines do what machines do best, precisely so that their people can do more of what only people can: earn trust, hold relationships and give work its meaning. The more advanced the technology, the more essential the leader.

“The more advanced the technology, the more essential the leader.”

High tech, high touch is not nostalgia. It is a design principle for every corporation that intends to put agents at the center of its work.

Beyond hype: where is the economic value?

The return on invested capital of the AI buildout, first for the industry that builds it and then for the corporations that buy it.

$2.67T

Worldwide AI spending forecast for 2026, up 49.5% · Gartner

75%

Nvidia gross margin on $89.0B of quarterly data-center revenue, Q2 FY2027

39%

Share of US real GDP growth from AI-related investment, first nine months of 2025 · St. Louis Fed

2.2%

Share of all US work hours saved by generative AI by mid-2026, self-reported · St. Louis Fed

THE ARGUMENT IN BRIEF

I
What the evidence actually shows
The evidence is strong at the level of the task, modest at the level of the economy, and uneven at the level of the firm.

II
The industry's ROIC question
Return on invested capital is the discipline that separates a technological revolution from a good investment.

III
ROIC from the company's perspective
For a corporation that buys AI rather than builds it, the equation is different, and in some ways more favorable.

IV
Why pilots stall, and how value is captured
Most AI pilots do not fail because the model is weak.

V
Where AI has paid so far
The equity market has already run the value-migration test, and its verdict is clear.

VI
Questions every board should ask
At BMI, every engagement follows the same four movements: first we appreciate, then we think, then we design, then we deliver.

AI AND THE ENTERPRISE · DANIEL AUGUSTO MOTTA, PHD

Put two numbers side by side and the AI debate comes into focus. Gartner forecasts worldwide AI spending of US$2.67 trillion in 2026, up 49.5% in a single year, rising to US$3.64 trillion in 2027, with AI infrastructure alone accounting for more than half. Daron Acemoglu, a Nobel laureate in economics, estimates that AI will add no more than about 0.66% to total factor productivity over ten years, and possibly less. Both numbers can be right at the same time. Capital moves on the clock of expectations; productivity moves on the clock of organizations.

As an economist and as an advisor to boards, I have learned to distrust both the enthusiasm and the dismissal. The useful question is not whether AI creates economic value. It clearly does. The question is whether the capital being deployed will earn a return above its cost, first for the industry that builds AI and then for the corporations that buy it. That is a question about return on invested capital, ROIC, and it has very different answers depending on where you sit.

EVIDENCE

What the evidence actually shows

The evidence is strong at the level of the task, modest at the level of the economy, and uneven at the level of the firm. At the task level, controlled studies are convincing. Customer-support agents with an AI assistant resolved 14% more issues per hour, and 34% more among novices. Consultants working inside the model's capability frontier produced more than 40% higher-quality work, though they did worse on tasks just outside it. At the level of the labor market, the picture is more sober. The Federal Reserve Bank of St. Louis estimates, from self-reported data, that generative AI saved 2.2% of all work hours by mid-2026, up from 1.6% in 2024. A large Danish study by Anders Humlum and Emilie Vestergaard, linking surveys to administrative records, finds precise null effects on earnings and hours, ruling out effects larger than 2% two years after ChatGPT's launch, even as tasks are being reorganized. Where firm-level returns do appear, they tend to come from growth rather than cost: Babina, Fedyk, He and Hodson found in the Journal of Financial Economics that AI-investing firms grew sales, employment and market value faster, primarily through product innovation, with the gains concentrated in larger firms. The broadest firm-level evidence is sobering: in a survey of nearly 6,000 senior executives across the United States, the United Kingdom, Germany and Australia, run with the Atlanta Fed, the Bank of England and the Bundesbank, about nine in ten reported no impact of AI on productivity or employment over the previous three years, while expecting a gain of about 1.4% over the next three. Some widely quoted numbers deserve caution. The claim that 95% of organizations see no return on generative AI came from a report based on 52 interviews and 153 survey responses, and has drawn serious methodological criticism. Pessimistic statistics can be as unreliable as vendor promises. This is why I encourage boards to scale capital with the level of evidence reached, not with the level of excitement in the room.

EVIDENCE LEVEL	WHAT IT JUSTIFIES
1. Vendor demos and public benchmarks	Curiosity and scoping
2. Retrospective test on our own data	A funded pilot
3. Live pilot against a comparison group	Scaling within one unit
4. Measured impact at scale	Enterprise-wide capital

II

INDUSTRY ROIC

The industry's ROIC question

Return on invested capital is the discipline that separates a technological revolution from a good investment. It asks whether the after-tax operating profit generated by the capital deployed exceeds what that capital costs. For the AI buildout, the invested capital is now visible. For 2026, Amazon has guided to about US$220 billion of capital expenditure, Alphabet to US$195–205 billion, Meta to US$130–145 billion and Microsoft to roughly US$175 billion for the calendar year.

Measured more broadly, the bill is larger still. Goldman Sachs forecasts global AI investment above US$1 trillion in 2026, US$581 billion of it in the United States, about 1.8% of US GDP and rising, and argues that the widely quoted hyperscaler consensus understates the total because it leaves out other public companies and private players.

Today, the clearest returns sit with the suppliers. Nvidia reported data-center revenue of US$89.0 billion in a single quarter, at a 75% gross margin. The buyers' revenues are growing fast but from a lower base. Microsoft reported that its AI business surpassed a US$37 billion annual run-rate, up 123%; Google Cloud grew 82% in the second quarter of 2026, with a US$514 billion backlog; Anthropic's run-rate reportedly reached US$65 billion by the end of July 2026. These are remarkable numbers. They are still small next to what the capital requires: MIT Technology Review's David Rotman puts the largest hyperscalers' spending at around US$750 billion a year against AI revenues of US$150–200 billion. Gary Gensler, former chair of the US Securities and Exchange Commission, calls the build-out "a parlay bet by the capital markets": it pays only if every leg wins. In the same article, Wharton's Jessica Wachter estimates that AI companies must raise their own productivity 2.7 times to break even by 2030. In 2024, David Cahn of Sequoia framed the gap as "AI's $600B question"; in July 2026 he updated it to a US$1.5 trillion question, estimating that each year of data-center investment requires roughly twice its value in lifetime revenue.

FRAMEWORK GARY GENSLER, VIA MIT TECHNOLOGY REVIEW

The legs of the parlay: every one has to win

The build-out pays only if all four conditions hold at once.

LEG 1

Hyperscaler revenues

Grow into US$750 billion a year of spending

LEG 2

Economy-wide productivity

Shows up in profits, not only in investment

LEG 3

The frontier holds its lead

Against cheaper, good-enough models

LEG 4

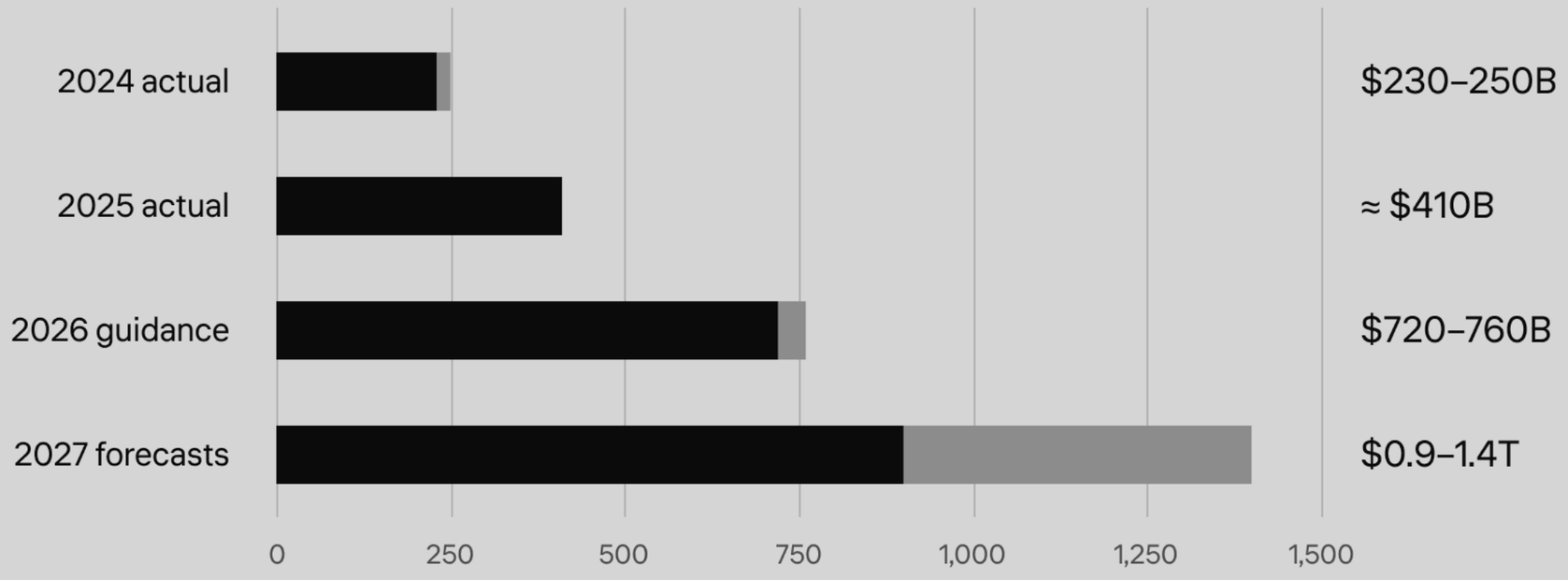
Public acceptance

Of job cuts and higher power bills

CHART

Big Four hyperscaler capital expenditure, US$ billions

Amazon, Microsoft, Alphabet and Meta combined. 2024–25 actuals, 2026 guidance after Q2 earnings, 2027 analyst forecasts; light extension shows the range.

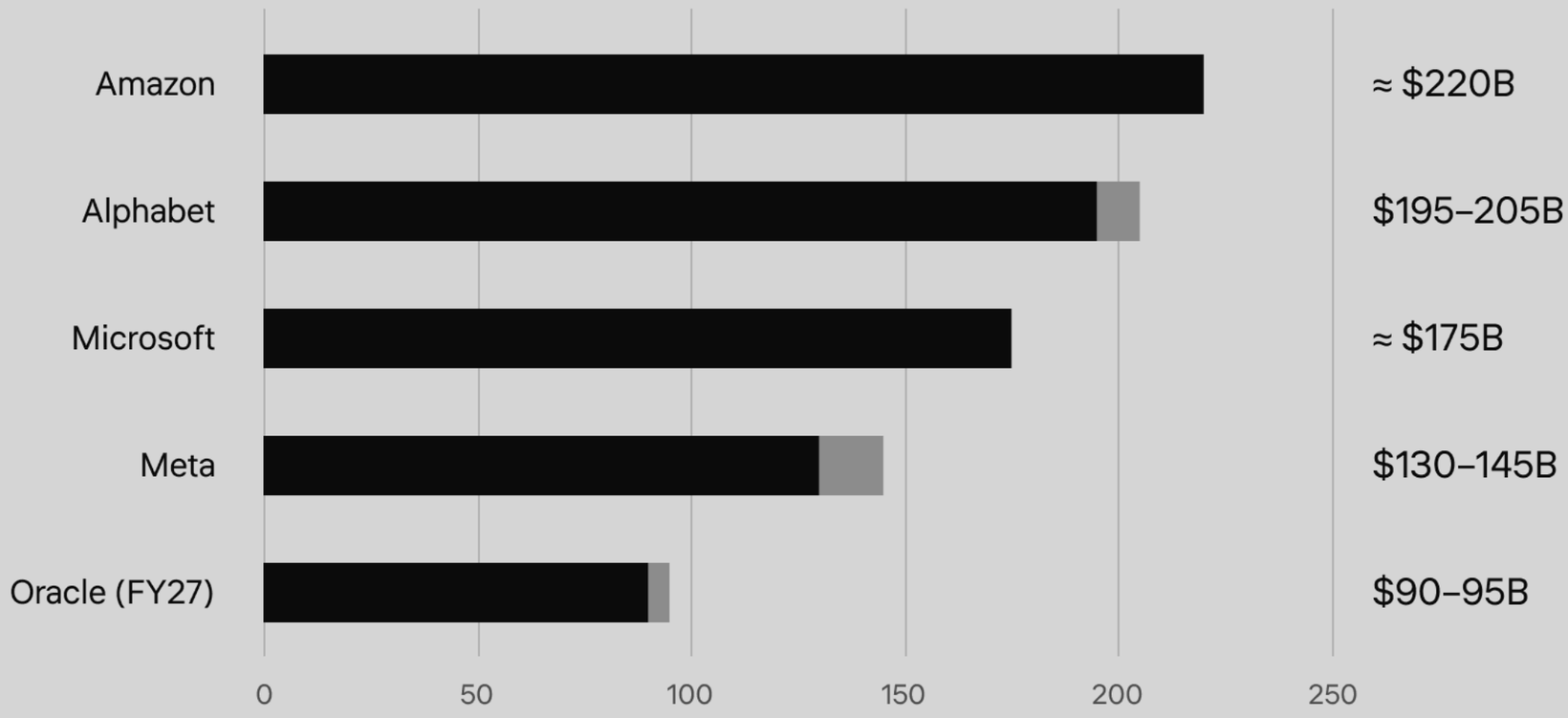


Sources: company filings; sell-side forecasts as reported.

CHART

2026 capital expenditure guidance by company, US$ billions

Latest guidance, July–September 2026. Microsoft reflects a lease reclassification, not a spending cut; Oracle is its fiscal year to May 2027.

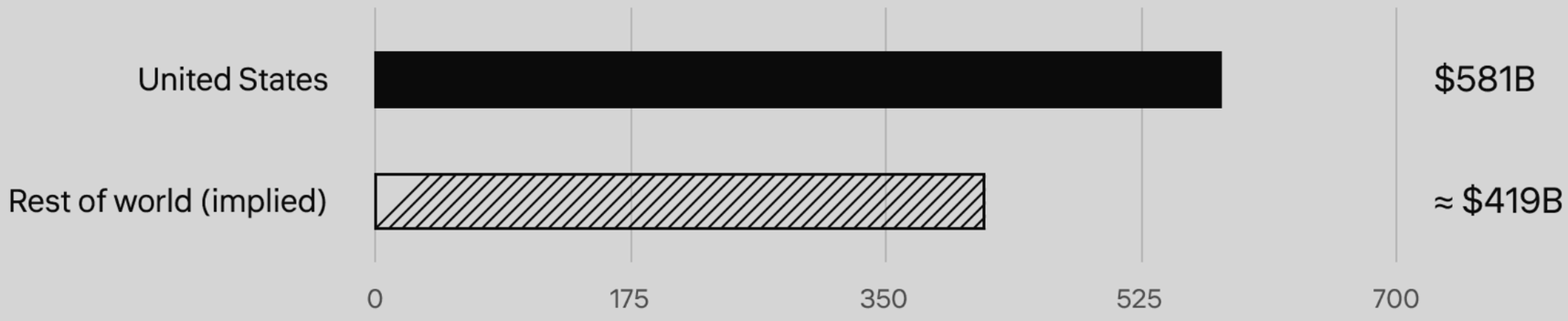


Sources: company filings and earnings calls.

CHART

Global AI investment, 2026 forecast, US$ billions

Goldman Sachs Research, August 2026; the non-US figure is implied by the US$1 trillion global total.



Source: Goldman Sachs Research (Briggs), 7 August 2026.

“The challenge is that the spending does not have commensurate revenues yet.”

Gary Gensler, former SEC chair, via MIT Technology Review, September 2026

The macroeconomic footprint is already visible. The Federal Reserve Bank of St. Louis estimates that AI-related investment contributed 0.97 percentage points of US real GDP growth in the first nine months of 2025, about 39% of the total and more than the dot-com peak, although the share falls sharply once imported chips and servers are netted out. The financing is changing character as well: global venture funding reached about US$510 billion in the first half of 2026, roughly four-fifths of it to AI, and Big Tech has roughly doubled its debt to fund the build-out.

A simple illustration, using my own arithmetic: if the four largest spenders invest around US$700 billion a year for three years, AI-related invested capital approaches US$2 trillion. Earning 10% after tax on it, roughly the cost of capital plus a modest spread, would require about US$200 billion a year of incremental operating profit, after depreciating hardware that may need replacement within five or six years. History offers a sobering scale. Stijn Van Nieuwerburgh estimates for the Brookings Institution that the AI pipeline will trigger about US$10.3 trillion of investment from 2025 to 2032, an average of 3.63% of US GDP a year, well above the railroads at 2.24% and more than three times the telecom and fiber boom. In every earlier boom, the economy gained from the infrastructure; investors did not always. Part of the bill is also less visible than capex. The Wall Street Journal reported that nine large technology companies carry roughly US$3 trillion of commitments off their balance sheets, in leases not yet started and purchase agreements for chips, networking and power, about five times their capital spending of the past year and roughly three times the US$248 billion of lease liabilities and US$356 billion of debt they carry on their balance sheets. And the model layer is burning cash at a scale without precedent: according to projections reported by Bloomberg from the Financial Times, OpenAI expects about US$278 billion of negative free cash flow from 2026 to 2030, even as revenue grows from US$36 billion to US$350 billion a year. As Reuters summarized the same projection, cumulative revenue of about US$840 billion over the period would still fall short of roughly US$856 billion of spending on computing and infrastructure alone.

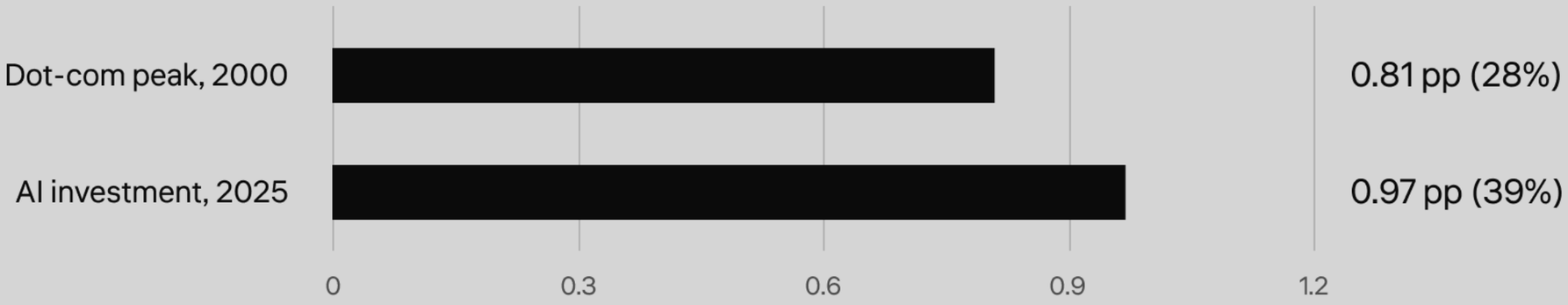
The pace matters as much as the size. Torsten Slok, chief economist at Apollo, estimates that data-center capex will reach about 3% of US GDP in 2027–29, two and a half times the telecom peak of 2000, and that it is building at roughly twice the pace of the 2002–05 housing boom. A cycle that builds that fast can unwind just as fast. The debate among financial institutions has shifted accordingly, from demand to financing. J.P. Morgan Asset Management estimates that a US$5 trillion build-out needs about US$650 billion of annual revenue to earn a 10% return, roughly US$35 a month from every iPhone user. Morgan Stanley sees a US$1.5 trillion external financing gap through 2028, much of it to be filled by private credit. Man Group notes that lenders assume asset lives of 10 to 20 years while accelerators become obsolete in a fraction of that time. Revenue, meanwhile, has surprised on the upside. The risk is no longer that it never arrives; it is that debt, short asset lives and concentrated counterparties turn a disappointment into a credit event.

Two warning signs deserve board attention. The first is depreciation. Hyperscalers depreciate servers over five to six years, and Amazon shortened part of its fleet from six to five years in 2025, citing the pace of AI. If the economic life of accelerators proves shorter than their accounting life, a question still open, reported returns overstate real ones. The second is cash. Capital intensity has outrun operating cash flow: Meta generated only US$784 million of free cash flow on US$60.8 billion of quarterly revenue, and Microsoft is now the only one of the four with positive trailing free cash flow. Debt financed about 32% of big-tech capex by mid-2026, against 9% two years earlier. My reading is that industry ROIC will compress before it recovers, and that the recovery will be deeply uneven. Returns accrue first to the layers with pricing power. Whether scale gives the cloud and model layers enough pricing power to justify this capital is the central strategic question of the industry.

CHART

Tech investment's contribution to US real GDP growth, percentage points

Dot-com peak versus AI-related investment in 2025 (first nine months, annualized). Share of total growth in brackets; gross, before netting imports.

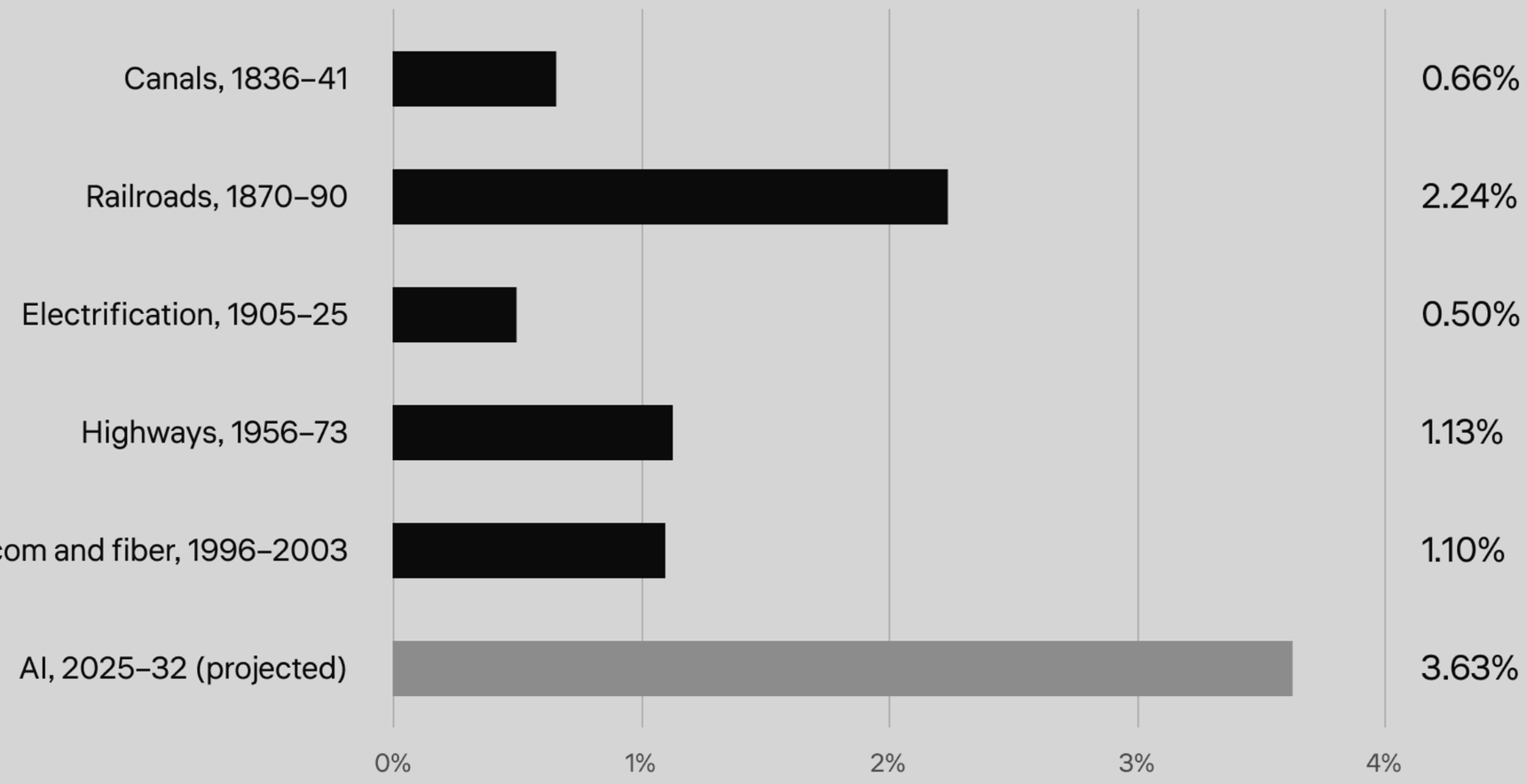


Source: Federal Reserve Bank of St. Louis, January 2026.

CHART

America's infrastructure booms, average annual investment as % of GDP

The AI figure is a projection for 2025–2032 (about US$10.3 trillion), shown as a light bar.

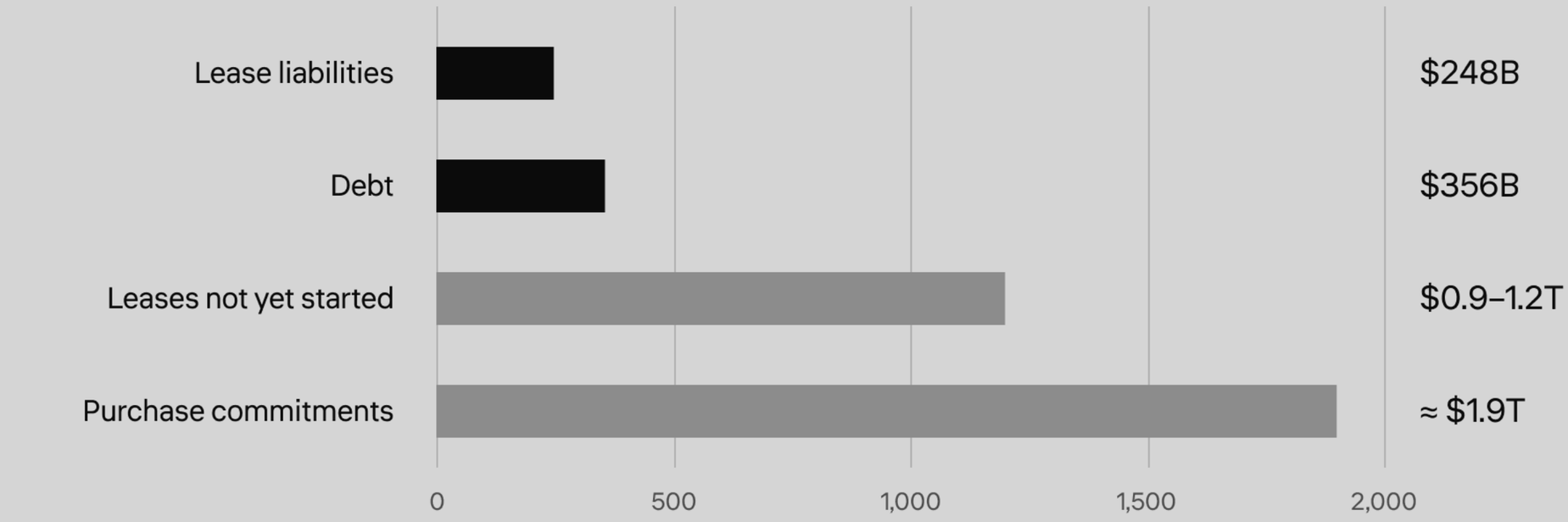


Source: S. Van Nieuwerburgh, "Financing the AI Buildout", Brookings Papers on Economic Activity, September 2026.

CHART

What nine large technology companies owe, US$ billions

On the balance sheet: lease liabilities and debt. Off it (light bars): leases signed but not yet started, and purchase commitments for chips, networking and power; the lease bar shows the top of a reported US$0.9–1.2 trillion range.

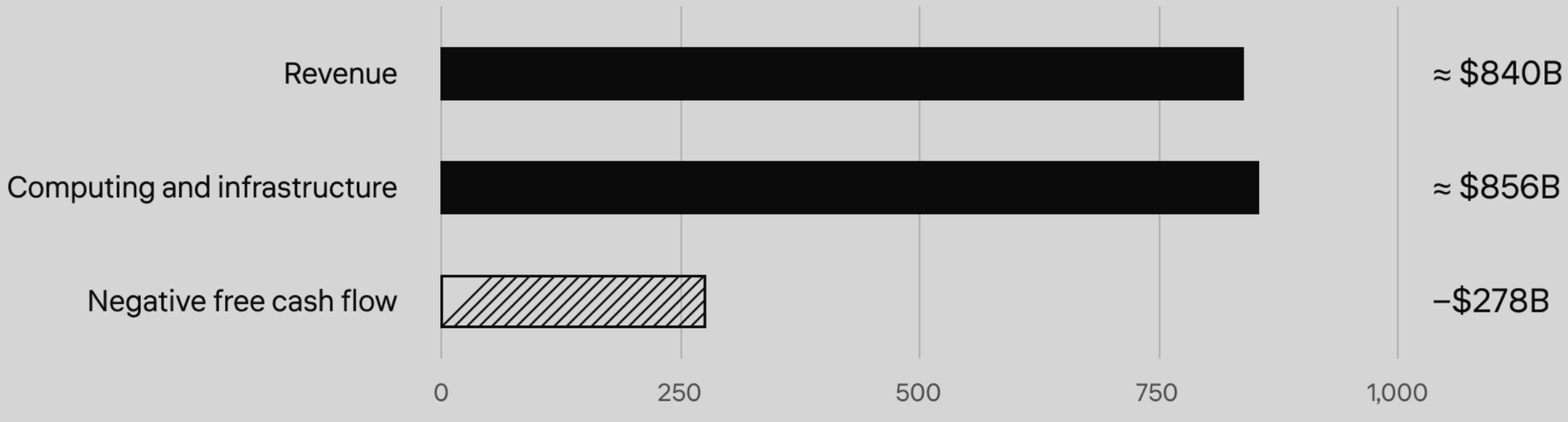


Source: Wall Street Journal, August 2026, as summarized by Insider Monkey via Yahoo Finance.

CHART

OpenAI's projection, 2026–2030 cumulative, US$ billions

Revenue grows from US$36 billion (2026) to US$350 billion (2030); spending on computing and infrastructure exceeds it, and the gap is funded by capital raises.

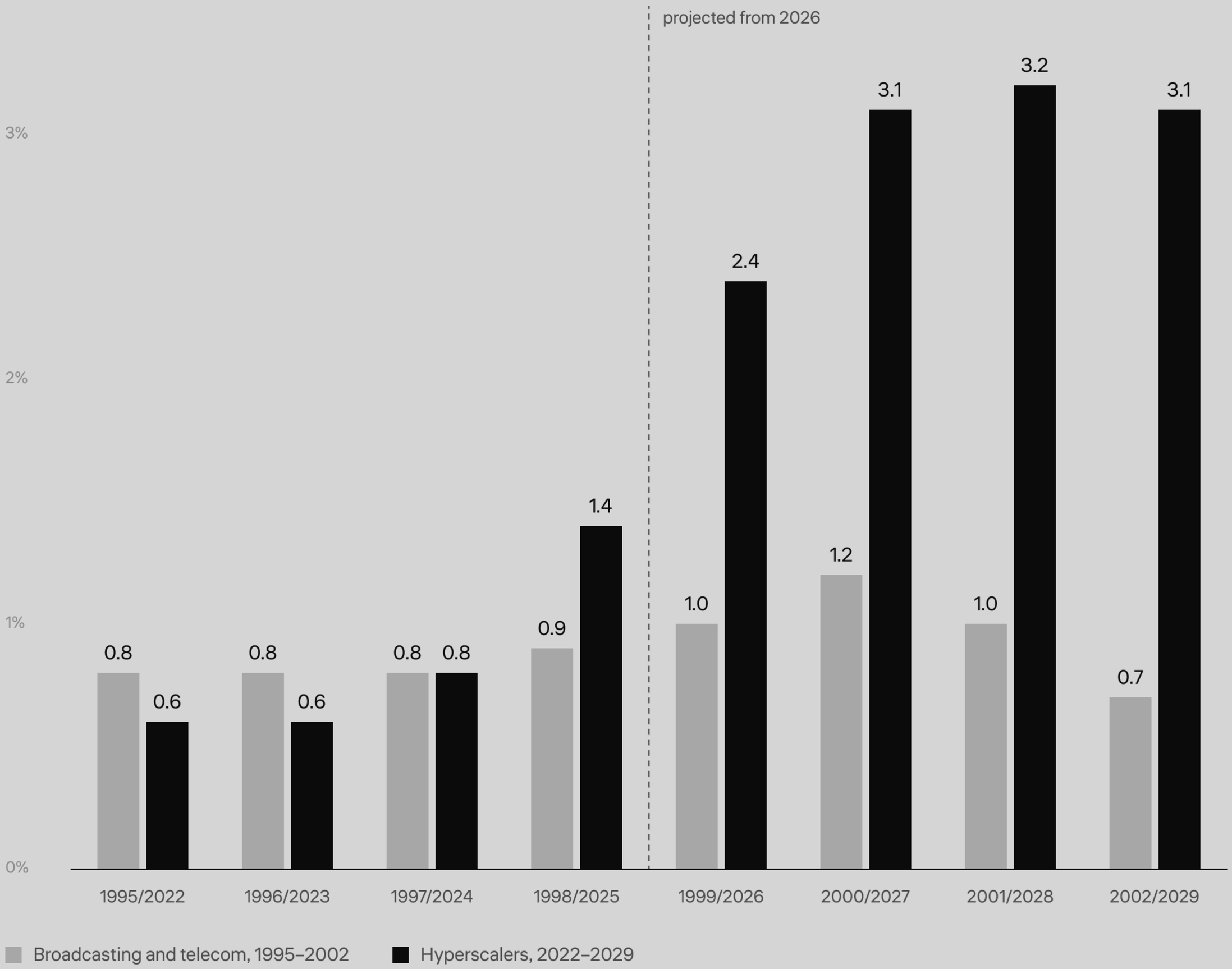


Source: Financial Times, via Bloomberg and Reuters, September 2026.

CHART

Two capex booms, as a share of US GDP (%)

Broadcasting and telecom investment, 1995–2002, against hyperscaler capex, 2022–2029, aligned year by year; later hyperscaler years are forecasts.



Source: Apollo Global Management, T. Slok, "The AI Capex Boom Is Building Twice as Fast as the Housing Boom", 2026.

HOB SYNTHESIS

Three scenarios for the buildout

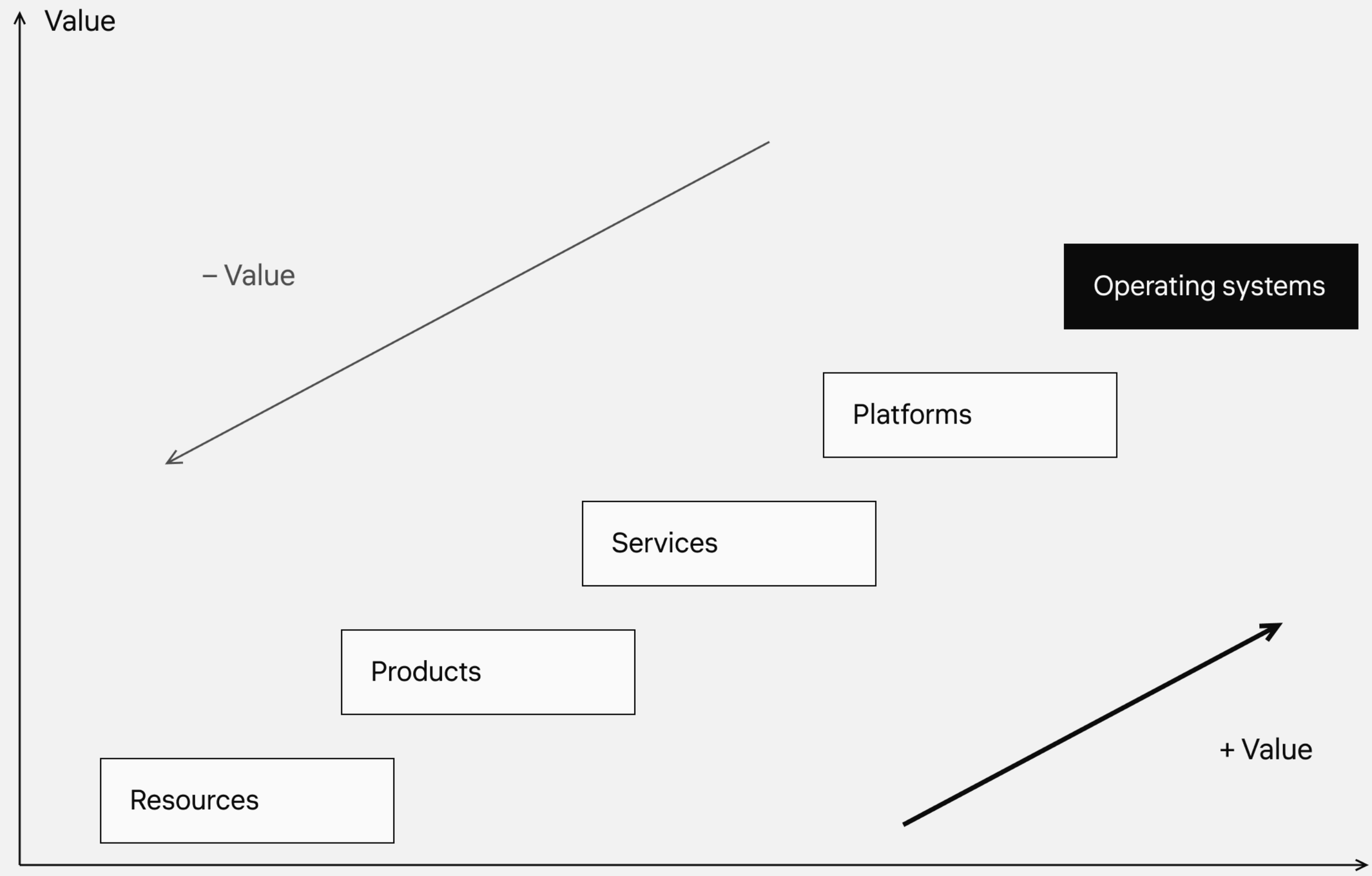
HOB synthesis. In every precedent, the infrastructure outlived the investors who financed it.

A	B	C
Productivity materializes	Efficiency shock	Financing stress
Returns arrive and the build-out is justified	Cheaper models leave excess capacity: providers suffer, buyers gain	Debt and off-balance-sheet commitments force consolidation

A framework I have used with boards for years explains why. The centre of gravity of economic value moves through structural layers. Value first concentrates in natural resources. Resources are transformed into products, which capture the larger share. A layer of services is then added on top of the product offer. Platforms follow, built to orchestrate and integrate products and services. Finally, some technologies become true operating systems of human activity, layers on which whole economies run. In Anthesis I put the sequence in historical terms: services added value to products in the mid-twentieth century, platforms integrated services and products at its end, and we would increasingly live with operating systems that consolidate platforms into a complex ecosystem, with AI and cloud computing as the likeliest candidates.

The mechanism is what matters for ROIC. When a new layer emerges, it becomes the new centre of gravity of economic value, and in doing so it exhausts the value previously concentrated in the layer below. Over time, returns on capital in the lower layers converge to their cost of capital, while asymmetric returns concentrate in the upper layers. Nothing disappears: the lower layers remain essential. They simply stop earning more than they cost. For an economist this is the dynamic of economic rents made visible. Rents in a lower layer attract entry and imitation until margins normalize; the upper layer earns rents because it controls the scarce resource of the moment, whether access to customers, the power to integrate or the rules of the system itself. AI is running the sequence at unusual speed and scale. Energy, chips and data centers are the resource layer, and they absorb most of the capital described above. Models behave like products and are already commoditizing. Services built on models are multiplying, agent platforms and open protocols are forming to orchestrate them, and the contest is over which of them becomes the operating system of knowledge work. The framework points to the hard question for investors. If capital is concentrated in the lower layers while asymmetric returns migrate upward, part of today's investment will earn no more than its cost of capital, however essential the infrastructure turns out to be.

HOB PROPRIETARY MODEL | JOTA MODEL



Creation and destruction of value

HOB proprietary model, from D. A. Motta, Anthesis (2019). Each layer adds complexity and, while it is the newest, captures the largest share of value. Schematic; no scale implied.

III

COMPANY ROIC

ROIC from the company's perspective

For a corporation that buys AI rather than builds it, the equation is different, and in some ways more favorable. Most of the physical capital belongs to someone else: the enterprise rents capability by the token or by the seat, and unit prices keep falling. The inference cost of GPT-3.5-level capability dropped more than 280-fold in two years. Falling prices do not mean falling bills, however. As the FinOps Foundation puts it, "A token may get cheaper; tokens, in aggregate, are not."

“A token may get cheaper; tokens, in aggregate, are not.”

J.R. Storment, FinOps Foundation, "Token Economics", 2026

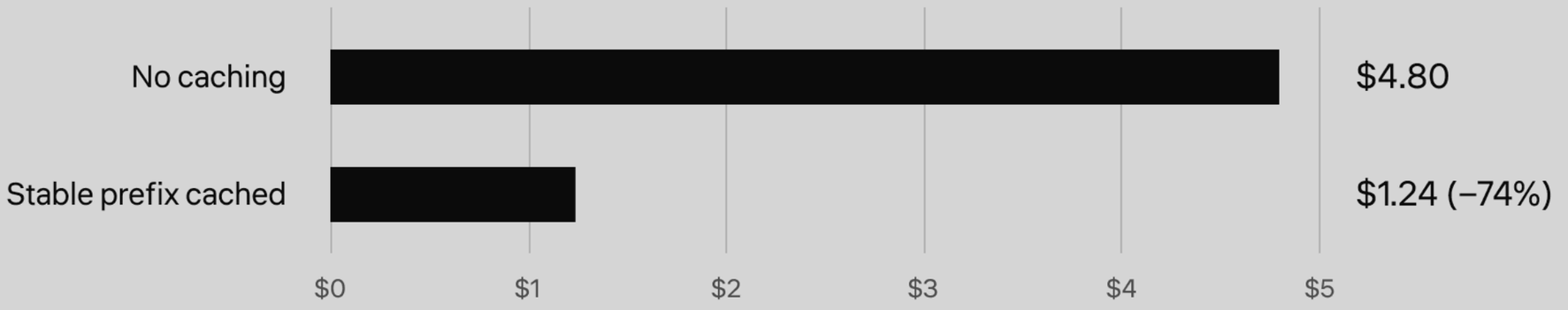
The real invested capital of a corporate AI program is largely invisible on the balance sheet: data preparation, process redesign, integration, training and change management. Erik Brynjolfsson, Daniel Rock and Chad Syverson showed in "The Productivity J-Curve" that such intangible investments make measured productivity understate true gains for years; adjusting for them, they estimate total factor productivity 15.9% higher than official measures by 2017. The corollary for boards is uncomfortable: early AI returns will look poor precisely where companies are investing correctly. I find it useful to decompose ROIC into operating margin times capital turnover. AI can lift margin by lowering the cost to serve, and lift turnover by compressing cycle times, working capital and the assets needed per unit of revenue. Neither happens unless work is actually finished. The token is the unit of cost; the finished task is the unit of value. "AI cut drafting time by 40%" rarely reaches the P&L. "AI resolves first-level support requests end to end" does. Beyond the task sits the decision, the true economic unit, and beyond the decision sits trust: value appears only where verification can stop. Cost itself is a design variable. Every step an agent takes re-sends its entire context window to the model, and the company pays again for every input token. What an agent loads, and in what order, therefore decides both its quality and its cost. Stable content placed first can be cached: using Anthropic's published prices, my own calculation for a 20-step task with a 50,000-token stable prefix shows input cost falling from US$4.80 to US$1.24, 74% less. Engineering levers compiled by the FinOps Foundation report reductions of 30% to more than 90% on specific workloads. Context design is cost design, and the relevant price is the price per task, not per token.

The spread across suppliers is wider than most buyers realize. J.P. Morgan Private Bank estimates that one completed task, weighted by its mix of workloads, costs a few cents with some Chinese providers and close to US$3 with the most expensive US frontier model, a spread of roughly seventy times, and that consumption drives the bill more than price: the heaviest 1% of users consume more than 600 times the tokens of the median user. The comparison that matters is cost per successful outcome on the company's own workloads. A five-cent task that fails a third of

CHART

Input cost of one 20-step agent task, US$

Illustrative calculation at published Anthropic rates (US$4 per million input tokens; cache write US$5, cache hit US$0.20), with a 50,000-token stable prefix. Output tokens excluded.

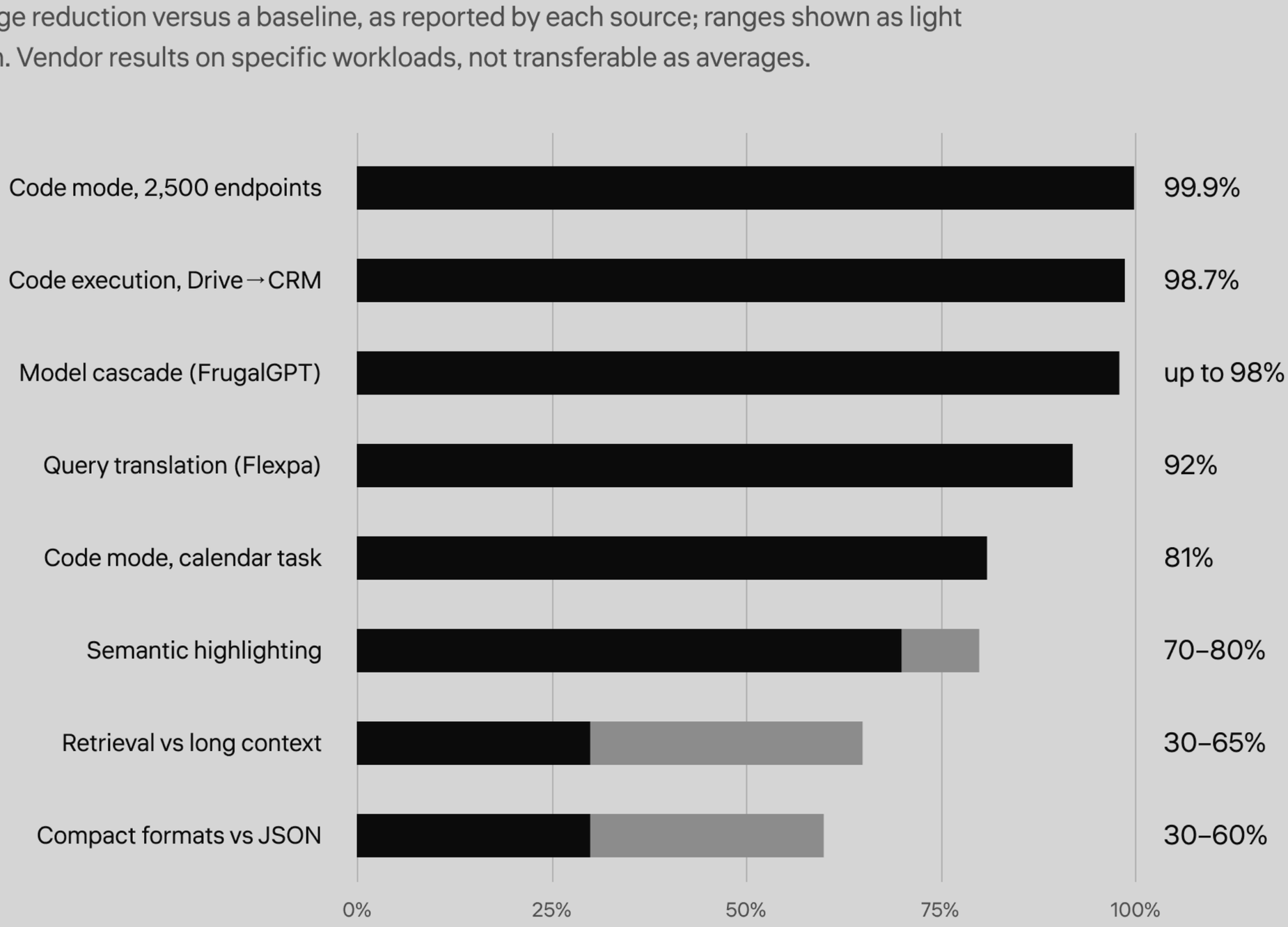


Source: Anthropic pricing documentation (Opus 5.5 rates); House of Brains calculation.

CHART

Reported token or cost reductions from engineering levers

Percentage reduction versus a baseline, as reported by each source; ranges shown as light extension. Vendor results on specific workloads, not transferable as averages.



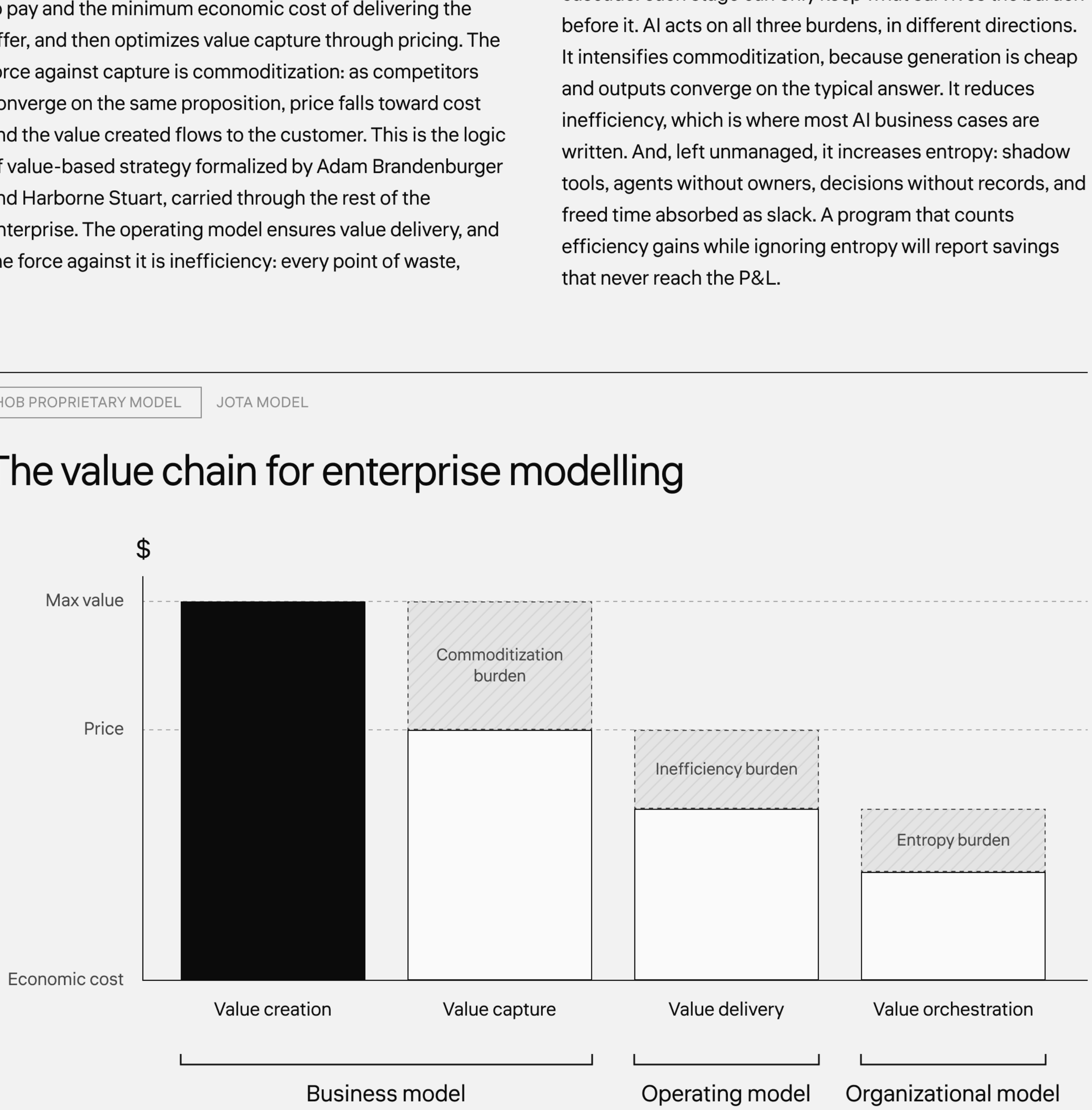
Source: FinOps Foundation compilation citing Cloudflare, Anthropic, Zilliz, Flexpa and FrugalGPT.

the time can cost more than a ninety-cent task that succeeds.

Margin, in turn, can be created and lost at several points. I read a company's value through three models, each governing a stage of value and each working against a characteristic force. The business model maximizes value creation, the difference between the customer's willingness to pay and the minimum economic cost of delivering the offer, and then optimizes value capture through pricing. The force against capture is commoditization: as competitors converge on the same proposition, price falls toward cost and the value created flows to the customer. This is the logic of value-based strategy formalized by Adam Brandenburger and Harborne Stuart, carried through the rest of the enterprise. The operating model ensures value delivery, and the force against it is inefficiency: every point of waste, rework or delay raises the real cost of delivery above the economic minimum. The organizational model performs value orchestration, and the force against it is entropy, the natural drift of any social system toward misalignment, duplicated effort, conflicting incentives and eroding norms. Entropy is the least visible of the three burdens and, in large organizations, often the largest. The chain reads as a cascade: each stage can only keep what survives the burden before it. AI acts on all three burdens, in different directions. It intensifies commoditization, because generation is cheap and outputs converge on the typical answer. It reduces inefficiency, which is where most AI business cases are written. And, left unmanaged, it increases entropy: shadow tools, agents without owners, decisions without records, and freed time absorbed as slack. A program that counts efficiency gains while ignoring entropy will report savings that never reach the P&L.

HOB PROPRIETARY MODEL | JOTA MODEL

The value chain for enterprise modelling

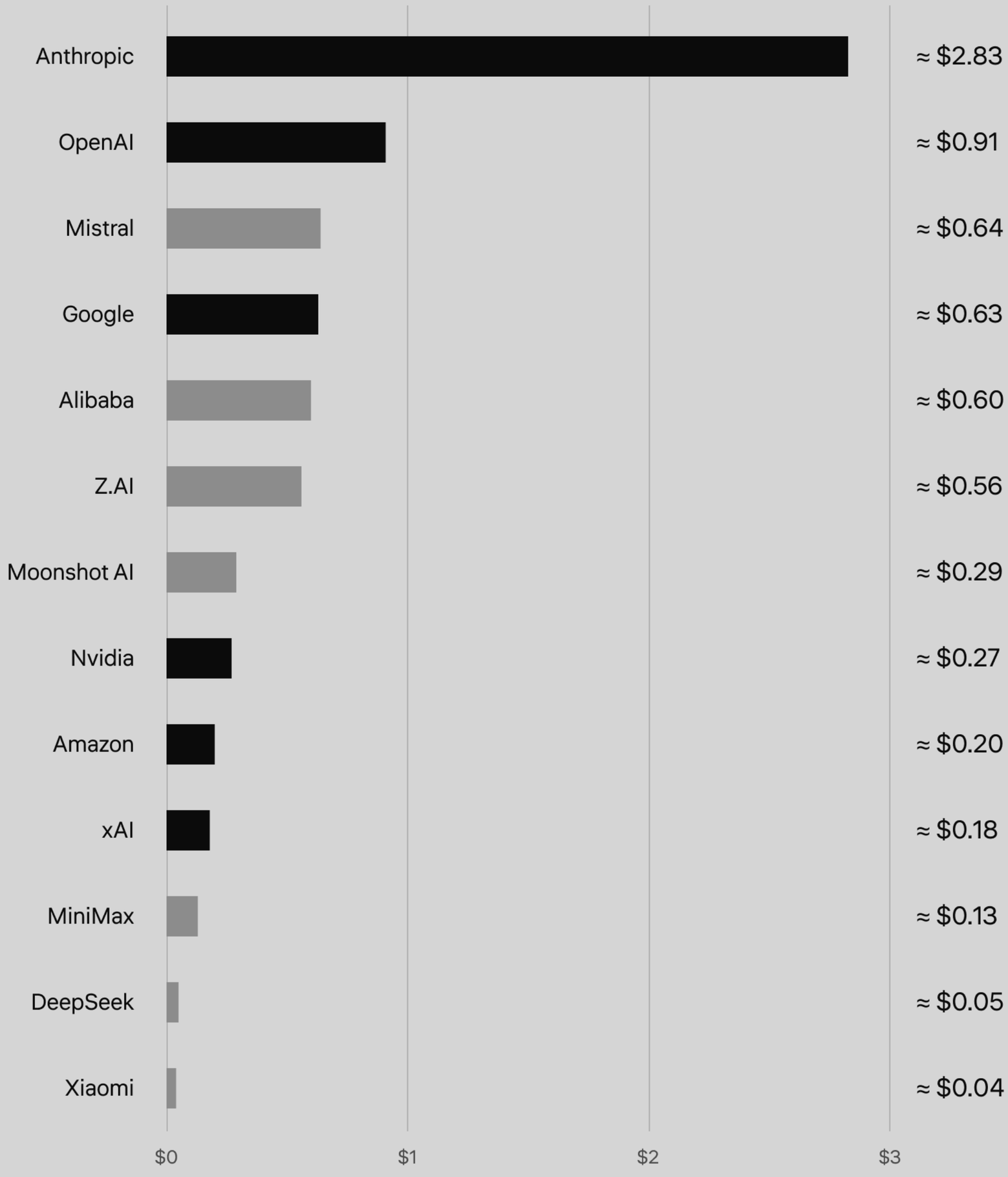


HOB proprietary model, from D. A. Motta, Anthesis (2019). Each bar starts at economic cost; the hatched band above it is the value lost to that stage's burden. Schematic; bar heights are illustrative.

CHART

Weighted average cost per task, by model provider, US$

Values approximate, from J.P. Morgan Private Bank, "AI Use Is Exploding. So Are the Bills", July 2026; dark bars US labs, light bars Chinese and European providers.



Source: J.P. Morgan Private Bank, “AI use is exploding. So are the bills.”, July 2026.

By that yardstick, most organizations still harvest the smallest pool. Afferolab's study A Escalada da Inteligência found that 87.5% of Brazilian professionals use AI to produce text, while only 37.8% use it to automate tasks. Enterprise spending is nonetheless accelerating. Menlo Ventures estimates that companies spent US$37 billion on generative AI in 2025, up from US$11.5 billion a year earlier, split almost evenly between applications and infrastructure, and that 76% of use cases were now bought rather than built. Buying capability is sensible. It also means the differentiating investment moves elsewhere: into the data, redesigned processes and people that no vendor sells. The research points in one direction: the highest corporate returns come from growth and redesign, not from cost alone. Professional services illustrate the shift sharply. A House of Brains essay on the intrinsic value of professional services firms distinguishes value built on proprietary content and methods from value built on individual expertise. AI tilts that balance decisively across the industry. Packaging loses value; thinking gains it. Methods become software, and pricing drifts from hours toward outcomes. The same logic applies to banks, insurers, retailers and manufacturers: the business case that survives is the one whose expected ROIC, including the intangible investment, exceeds the cost of capital within a credible horizon. Gartner's latest predictions point the same way. It expects that by 2027, 75% of organizations that bank AI productivity gains as cost savings will be outpaced by competitors that reinvest them, and that by 2029, 60% of organizations deploying AI will set up a function dedicated to mapping AI's total cost to its value. In other words, measuring corporate AI ROIC is about to become a standing management discipline, not a one-off business case.

IV

WHY PILOTS STALL

Why pilots stall, and how value is captured

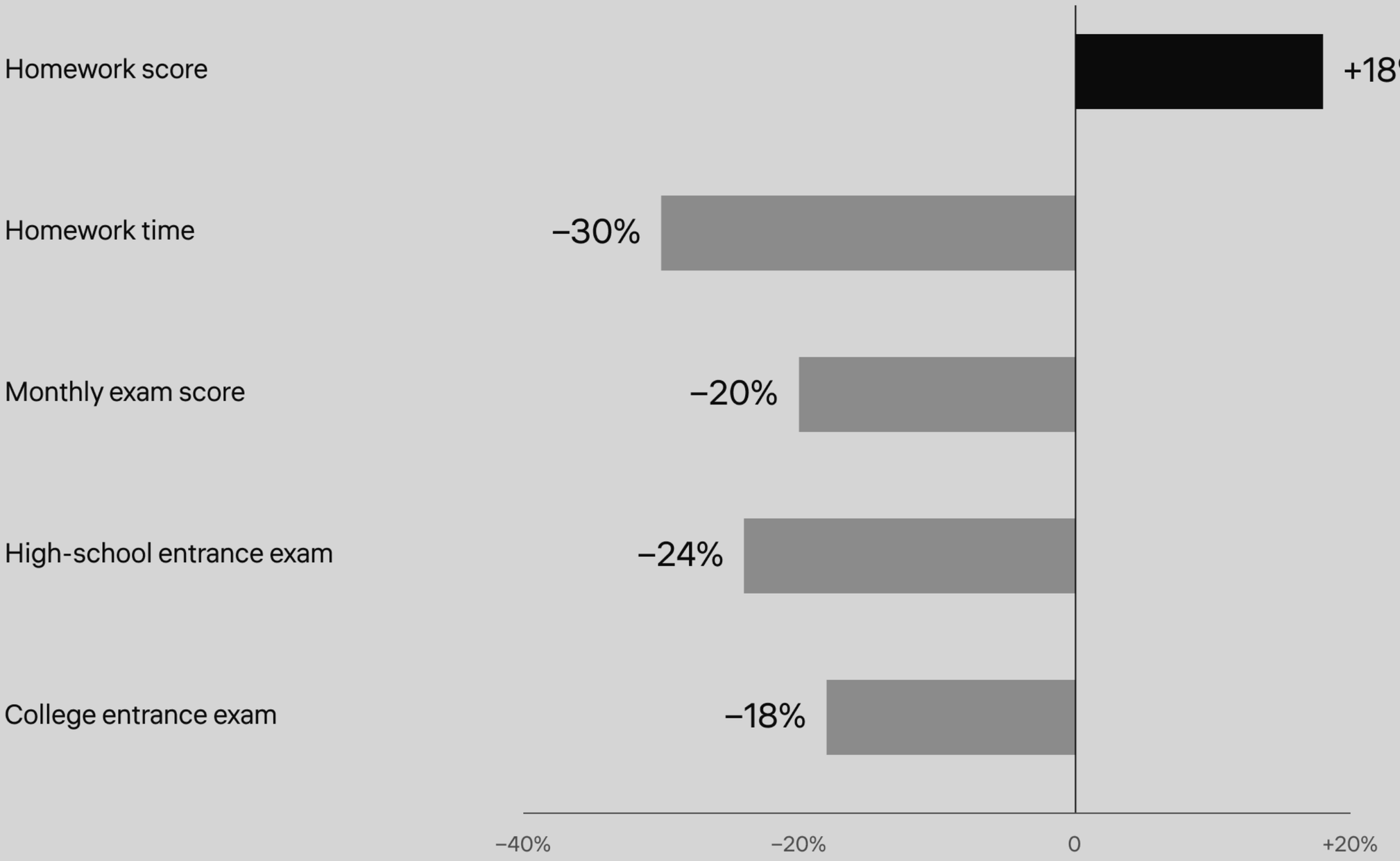
Most AI pilots do not fail because the model is weak. They fail because the task was never well defined, the data was not representative, success was never measured against a comparison group, nobody decided what happens when the system is wrong, and nobody owns monitoring once the project team moves on. Those five questions should be asked before any greenlight, and machine learning should be treated as an operation, not a one-time build. The evidence on who captures value is consistent. MIT's Center for Information Systems Research surveyed 721 companies and found that the 62% still experimenting or building pilots performed below their industry average, while the 38% that had industrialized AI into their ways of working performed well above it. History suggests why the transition takes time. Paul David showed in "The Dynamo and the Computer" that electricity raised factory productivity only decades after its arrival, when plants were redesigned around the new power source rather than fitted with electric motors in place of steam engines. The gains went to the firms that redesigned the work. The attrition is already visible and expected. Gartner predicted that at least 30% of generative AI projects would be abandoned after proof of concept by the end of 2025, and that over 40% of agentic AI projects will be canceled by the end of 2027 because of escalating costs, unclear business value or inadequate risk controls. Each cancellation is invested capital with no return. Behind them sits a deeper problem: the organization, not the model, is the bottleneck. Companies buy licenses without redesigning work, automate poor processes, and run what I call transformation theater. Capability that is trained but never embedded in how work is done produces no return. Measurement deserves its own warning. The first controlled evidence on generative AI at work looked unambiguous: in Shakked Noy and Whitney Zhang's experiment with 453 college-educated professionals, published in Science, ChatGPT cut the time spent on writing tasks by 40% and raised quality by 18%, with the weakest writers gaining most. Those gains are real, and they are gains in output. A study of 26,811 Chinese secondary-school students by David Strömberg, Victor Lei and Yanhui Wu shows what output metrics can hide. Six months after AI tools became available, homework scores had risen 18% and homework time had fallen 30%, while monthly exam scores, taken without AI, had fallen 20%. A school that tracked only homework would have concluded its students were improving. A company that tracks only deliverables, volume and cycle time is running the same experiment on its own people, and will see the result only when someone must perform without AI: in a crisis, in a negotiation, or at the moment an agent is wrong.

The second measurement failure concerns the time AI saves. Hours freed without a plan for their use dissipate into more of the same, one reason so many firms report widespread AI use and so few report returns. The reinvestment decision belongs in the business case: if AI frees a fifth of a team's capacity, decide in advance whether it goes to deeper client work, capability building or new services, and measure the capabilities created, not only the hours saved. The largest returns rarely come from doing an existing task faster. They come from tasks that were previously impossible, much as the microscope did not speed up observation but revealed a world that had been invisible. Value that does get captured compounds in layers: governed data and captured expertise at the base; then a workforce fluent enough to supervise AI; then codified workflows that improve with use; then customer relationships that generate learning loops; and finally new offerings and decision systems. Each layer takes time, and returns follow a J-curve, so boards should expect investment before value. The uncomfortable implication is that learning loops cannot be bought retroactively. A competitor who starts two years later faces a gap widened by two years of compounding. Time, not technology, becomes the moat.

CHART

Better homework, worse learning: change after AI adoption, %

26,811 Chinese secondary-school students, 2022–2025. Homework and monthly exams six months after adoption; entrance exams at full effect, about two years after.

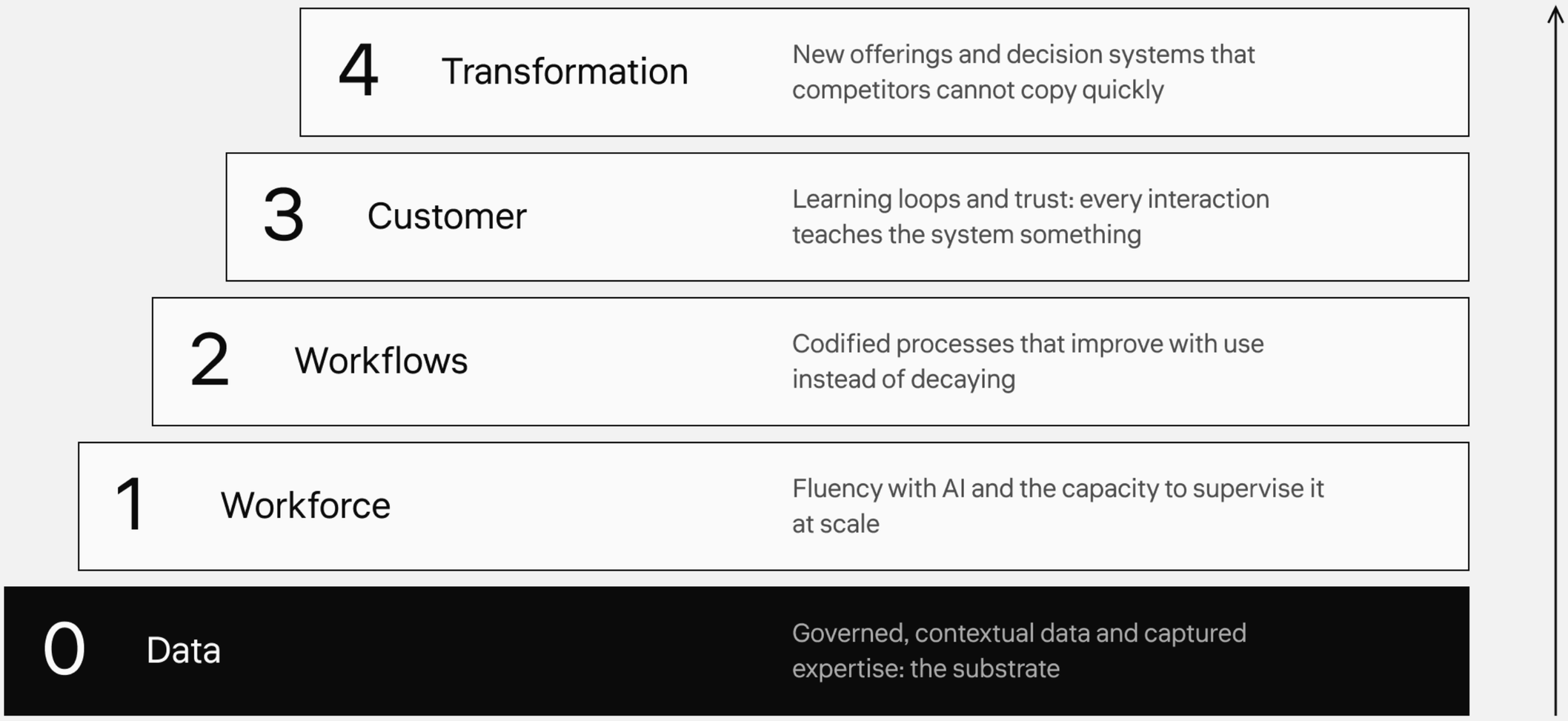


Source: Strömberg, Lei and Wu, VoxEU/CEPR, September 2026.

FRAMEWORK GOOGLE CLOUD

Five layers of compounding advantage

Each layer learns from the one beneath it, and data is layer zero. Advantage accumulates upward and over time; learning loops cannot be bought retroactively.



TIME IS THE MOAT

A firm that starts two years later faces a gap widened by two years of compounding.

V

MARKET VERDICT

Where AI has paid so far

The equity market has already run the value-migration test, and its verdict is clear. I tracked the fifty most valuable technology companies in the United States, Europe and Asia at the end of July 2021 and at the end of July 2026. Their combined value rose from about US$18.2 trillion to US$40.0 trillion, and the share held by semiconductors and chip equipment rose from 17% to 43%. Twenty companies left the ranking and twenty entered, almost all in one direction: in came memory, chip design, chip equipment and data-center hardware; out went application software, IT services and Chinese consumer internet. The largest services firms in the 2021 ranking lost between 43% and 50% of their value, on the expectation that agents will reduce seats and billable hours rather than on current revenue.

Share prices tell the same story. Rebased to 100 in September 2021, Nvidia ended September 2026 at about 1,100, while the median of eleven technology leaders stood

CHART

Chips and chip equipment went from $3.1T to $17.0T of top-50 value

Market capitalization of the 50 most valuable technology companies by segment, US$ billions, July 2021 and July 2026. Author's compilation from market data.



Source: House of Brains compilation from StatMuse, stockanalysis.com, companiesmarketcap.com, digrin and SEC filings; market values at end-July 2021 and 2026.

at 188 and Salesforce was the only one below its starting point. The market has paid the suppliers of scarce compute far more than the companies that use it, and it has begun to mark down businesses priced on seats and hours. This is the creation and destruction of value in real time: the resource layer is collecting the surplus today, the operating-system layer has not yet been decided, and the services layer is already being repriced. Value follows scarcity, and scarcity moves, from GPUs to networking, memory, storage and power. A board should know which of those inputs its own AI plans depend on, and who controls it.

VI

BOARD QUESTIONS

Questions every board should ask

At BMI, every engagement follows the same four movements: first we appreciate, then we think, then we design, then we deliver. It is a useful discipline against hype, because it forces appreciation of the actual problem before anyone falls in love with a solution. Applied to AI investment, it produces a short set of board questions.

1. Which finished tasks, and which decisions, will this initiative change, and how will we know?

2. What level of evidence supports it today, and is the capital requested proportionate to that level?

3. What is the expected ROIC of the program, counting the intangible investment in data, process and people, and when does it exceed our cost of capital?

4. Which scarce complement do we own, data, workflow, relationship or accountability, that a competitor renting the same model cannot copy?

5. What must change in roles, incentives and norms for the value to show up in the P&L, and who is accountable for that change?

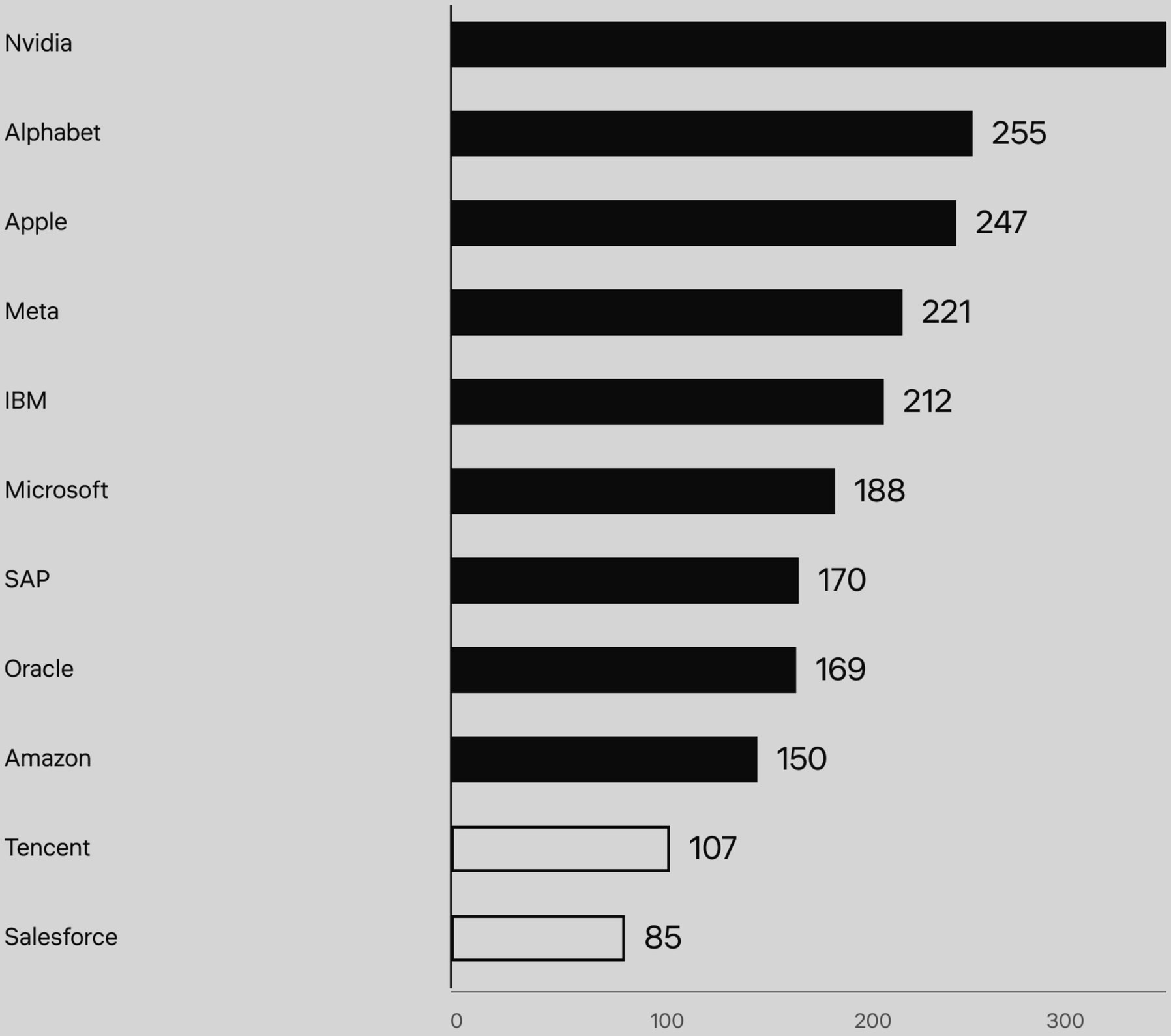
There is also a social ledger that no ROIC calculation should ignore. The International Monetary Fund warns that AI may increase labor-income inequality and that rising capital returns will widen wealth inequality. A corporate business case that counts the savings from displaced work but not the cost of reskilling, of lost trust or of a weakened social license to operate overstates its return. The digital transformation and the social transition are two sides of the same balance sheet.

Beyond the hype, the economic value of this era is real, large and unevenly distributed. It will not go to those who spend the most on AI. It will go to those who convert intelligence, now abundant, into decisions, outcomes and trust, which remain scarce.

CHART

Eleven technology leaders, share price indexed to 100 in September 2021

Index level in September 2026; Nvidia's bar is cut at 350. Author's compilation from market data.



Source: House of Brains compilation from StatMuse (US listings, split- and dividend-adjusted) and digrin (Tencent ADR); month-end closes to 30 September 2026.

“Intelligence is abundant. Decisions, outcomes and trust remain scarce.”

The value of this era will not go to those who spend the most on AI, but to those who earn a return above their cost of capital, for shareholders and for society.

Enjoy the ride and watch out for the hook.

The battle for the world AI ecosystem, layer by layer, and why the frontier may converge toward one or two global winners.

72.5%

TSMC share of global foundry revenue, Q2 2026 · TrendForce

63%

Cloud infrastructure share held by Amazon, Microsoft and Google, Q2 2026 · Synergy Research

2.4×

Annual growth in the cost of frontier training runs since 2016 · Epoch AI

5.4%

Arena score gap between the first and tenth model in early 2025, down from 11.9% · Stanford AI Index

THE ARGUMENT IN BRIEF

I
The battle, layer by layer
The AI industry is a stack of five layers, and each is being fought over with different weapons.

II
Why the market may tip toward one or two winners
Economists call an industry a natural monopoly when a single firm can serve the whole market at lower cost than several competitors could.

III
Four hooks to watch
The first is the price hook.

IV
Turbulence and opportunity: ten megatrends
The ecosystem battle is the most visible part of a larger structural change.

V
Geopolitics, sovereignty and the view from Brazil
The AI race is also a race between states, and the two largest powers frame it in opposite terms.

VI
A playbook for riding without being hooked
The strategic question is not which model is best.

VII
Enjoy the ride
None of this is an argument for caution as a strategy.

AI AND THE ENTERPRISE · DANIEL AUGUSTO MOTTA, PHD

Any executive can now rent, for the price of a software subscription, a reasoning capability that no company in the world could buy at any price five years ago. That is the ride, and it is exhilarating. But every great platform battle in the history of technology has been won the same way: by offering something delightful at a price that looks too good to refuse, and collecting the bill later, when switching has become unthinkable. That is the hook. Bill Gates has written that the transition to the AI era "will be one of the most turbulent times in human history".

Underneath the turbulence, a handful of companies and a few governments are fighting to own the layers on which every other organization will depend. From San Francisco, where I work with technology ventures, and from São Paulo, where I advise boards, the view is the same: enjoy the ride, because refusing it is not an option, but know exactly where the hook is before you bite.

THE LAYERS

The battle, layer by layer

The AI industry is a stack of five layers, and each is being fought over with different weapons. The prize is enormous: Gartner forecasts worldwide AI spending of US$2.67 trillion in 2026, of which about US$1.48 trillion goes to AI infrastructure, and its analysts describe the build-out of AI data-center capacity as the largest infrastructure project humanity has undertaken.

Lithography and foundry. At the bottom, the market is already a monopoly in all but name. ASML is the only company that makes the extreme-ultraviolet lithography machines needed to manufacture every advanced processor. One level up, TSMC held 72.5% of global foundry revenue in the second quarter of 2026, with its most advanced capacity fully booked by demand for AI chips. Two companies, one in the Netherlands and one in Taiwan, are single points of failure for the whole industry. Beneath both sits energy: the International Energy Agency expects data-center electricity use to roughly double to about 945 TWh by 2030, and BloombergNEF projects that US data centers will take a fifth of the country's power by 2035. Access to power is becoming a competitive weapon in its own right.

Accelerators. Nvidia reported US$89.0 billion of data-center revenue in a single quarter, at a 75% gross margin. Its moat is software as much as silicon: CUDA, which Jensen Huang said in 2024 had 5 million developers. The challengers are Nvidia's own largest customers. Google's TPUs now anchor multi-gigawatt commitments with Anthropic and Broadcom from 2027, and Amazon's Project Rainier runs on nearly half a million Trainium2 chips. The contest is between a general-purpose standard and vertically integrated custom silicon, with geopolitics on top: in January 2026 the US moved exports of advanced accelerators to China to case-by-case licensing. Custom silicon moves the toll gate rather than removing it. Silicon Analysts, which maps 52 companies across ten layers of the AI hardware chain, finds pricing power concentrated at three chokepoints: Nvidia's accelerators, anchored by CUDA; custom-chip co-design, where Broadcom and Marvell hold roughly 95% of the market; and interconnect and optics, because AI racks need ten to thirty-six times more fiber than conventional ones. Custom chips from the hyperscalers cost 40–65% less to run at scale, yet carry only 15–20% of those companies' own workloads. Almost every major player is at once customer, supplier and competitor of the others.

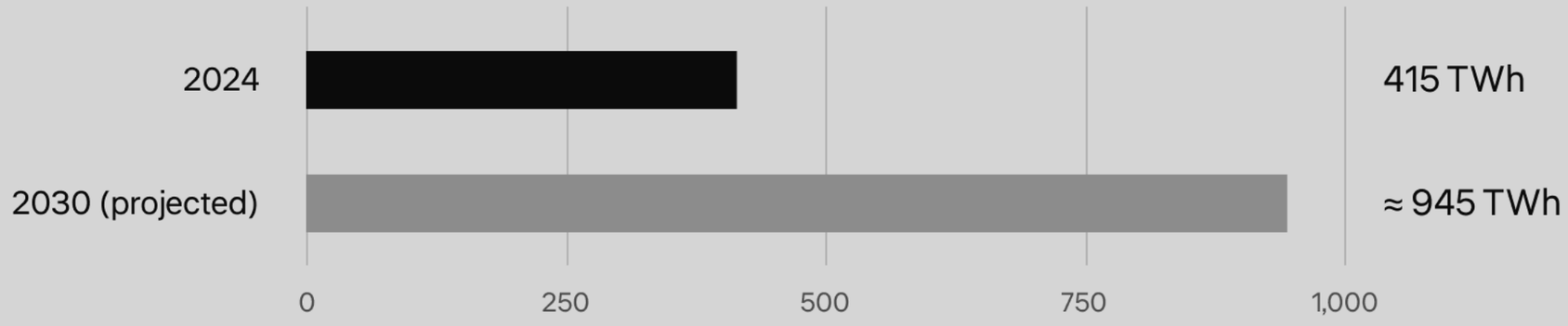
Cloud. Three firms dominate. Synergy Research estimates that in the second quarter of 2026 Amazon held 28% of cloud infrastructure services, Microsoft 20% and Google 15%, in a market that grew 43% to US$143.4 billion in the quarter. The same firms finance and distribute the model labs, forming what the UK Competition and Markets Authority called an "interconnected web" of more than 90 partnerships and investments.

Models. The price of staying at the frontier keeps rising. Epoch AI estimates that the cost of frontier training runs has grown about 2.4 times a year since 2016, with the largest runs expected to exceed US$1 billion by 2027. Independent benchmark organizations now referee the contest. In September 2026, the top positions on both the LMArena text leaderboard and the Artificial Analysis Intelligence Index were held by

CHART

Global data-center electricity demand, TWh

More than double in six years, to close to 3% of world demand. Light bar is a projection.

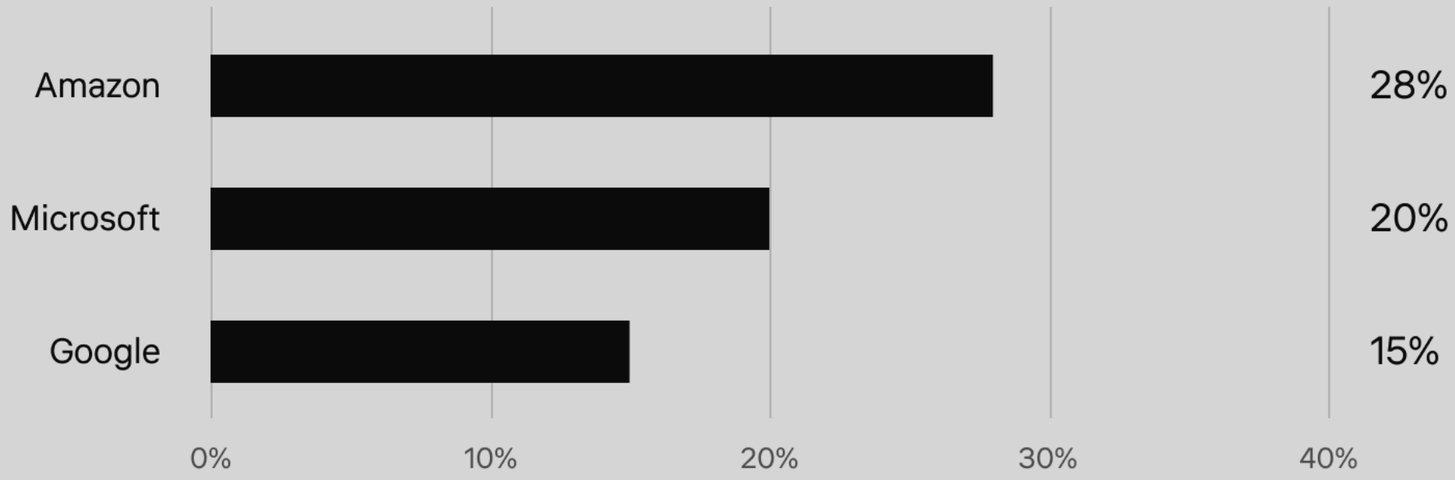


Source: IEA, Energy and AI, April 2025, base case.

CHART

Cloud infrastructure services market share, Q2 2026

Share of a US$143.4B quarterly market that grew 43% year on year.



Source: Synergy Research Group, July 2026.

FRAMEWORK

Power in the AI hardware chain sits at three chokepoints

Layers run from silicon at the top to deployment at the bottom; highlighted rows are the chokepoints. Source: Silicon Analysts, "AI Data Center Value Chain" and "The AI Data Center Chess Game", updated August 2026 (reported figures).

GPU makers	Silicon	Nvidia, AMD, Intel Chokepoint: Nvidia ≈ 75% of AI accelerators (2026E), anchored by CUDA
ARM CPUs	Silicon	AWS Graviton, Ampere, Google Axion, Microsoft Cobalt, Nvidia Grace, Fujitsu 15–23% of data-center CPUs; Nvidia's NVL72 racks are ARM-only
x86 CPUs	Silicon	Intel, AMD Structural decline as GPU-centered racks pull toward ARM
ASIC co-designers	Silicon	Broadcom, Marvell, Alchip, GUC Chokepoint: Broadcom + Marvell ≈ 95% of custom-chip co-design
Established ASICs	Silicon	Google TPU, AWS Trainium, Meta MTIA, Microsoft Maia, Intel Gaudi, IBM 40–65% lower TCO than GPUs, yet only 15–20% of hyperscaler workloads
ASIC start-ups	Silicon	Cerebras, Groq, Graphcore, SambaNova, Untether AI Alternative architectures; Groq technology licensed by Nvidia (≈$20B)
Interconnect, optics	Components	Amphenol, TE Connectivity, Corning, InnoLight, Coherent, Lumentum Chokepoint: AI racks need 10–36× more fiber; cable lead times 20+ weeks
OEMs / ODMs	Systems	Supermicro, HPE, Lenovo, Dell, Cisco, ASUS, Inspur, QCT, Gigabyte, Eviden Assemble servers and racks; Supermicro ≈ $40B revenue FY26E
Hyperscalers / cloud	Deployment	AWS, Google Cloud, Microsoft Azure, Oracle, Alibaba, Baidu, IBM Cloud Largest buyers, now also chip designers (TPU, Trainium, Maia, MTIA)
On-premise / private	Deployment	Apple, Meta, Tesla Private fleets: Meta 1.3M+ GPUs, Tesla 100K+ H100s

models from Anthropic, OpenAI and Meta, separated by margins close to statistical noise.

Distribution. Finally, the battle for the user. ChatGPT passed 900 million weekly users in early 2026, and Alphabet reported 950 million monthly users of the Gemini app in July. Whoever owns the daily interface owns the knowledge of what people ask, and the default position from which every other service is reached.

II

NATURAL MONOPOLY

Why the market may tip toward one or two winners

Economists call an industry a natural monopoly when a single firm can serve the whole market at lower cost than several competitors could, typically because fixed costs are enormous and marginal costs are low. The frontier of AI increasingly shows these features. Anton Korinek and Jai Vipra conclude in Economic Policy that AI displays "significant economies of scale and scope that may create a tendency towards greater market concentration." Five mechanisms reinforce one another.

> “Significant economies of scale and scope that may create a tendency towards greater market concentration in the future.”
>
> Anton Korinek and Jai Vipra, "Concentrating Intelligence", Economic Policy, 2025

1. Fixed costs. Billion-dollar training runs and multi-gigawatt data centers can be financed by very few firms on earth.

2. Low marginal costs. Once a model is trained, serving an additional customer costs a small fraction of the investment, so the largest player can always undercut.

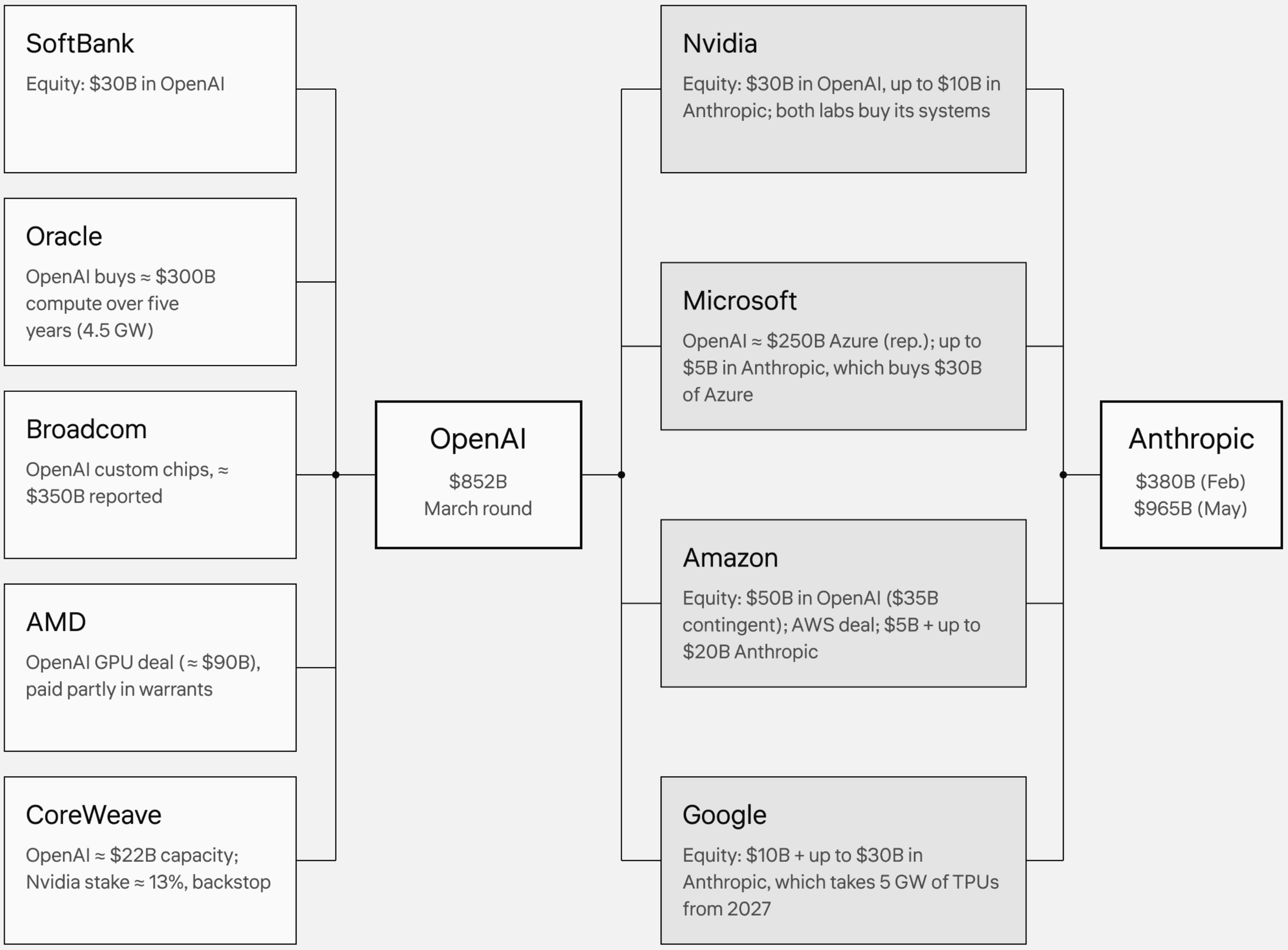
3. Feedback loops. More users generate more interactions and more feedback, which improve the next model and attract more users.

4. Switching costs. Developer tools, agent integrations and enterprise contracts lock customers in. The US Federal Trade Commission found that AI partnerships between cloud providers and labs "may increase contractual and technical switching costs".

5. Vertical integration. The leaders are integrating from chips to cloud to models to consumer applications, capturing margin at every layer.

Vertical integration now runs through balance sheets as well as products. Bloomberg's guide to the circular deals underpinning the AI boom shows that the largest investors in the frontier labs are also their largest suppliers: Nvidia, Microsoft, Amazon and Google each hold equity in OpenAI or Anthropic while selling them compute. Part of a supplier's reported demand is therefore financed with its own capital, and a reset in private valuations would shrink commitments across the whole web at once. The board test for any AI supplier or partner follows directly: what share of its backlog comes from counterparties it has itself financed?

FRAMEWORK



The labs' largest investors are also their largest suppliers

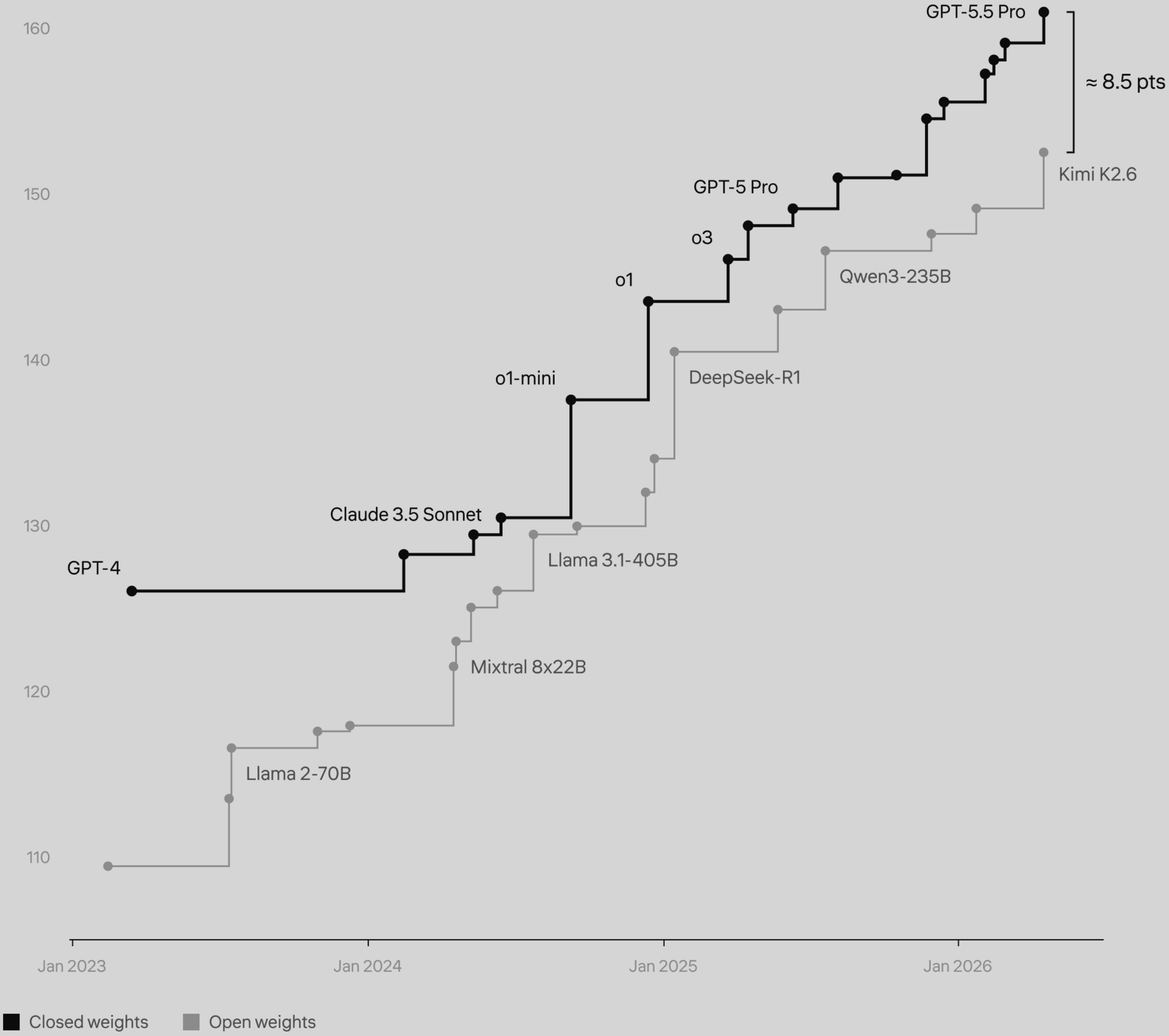
Highlighted: firms that are both shareholders in and suppliers to the frontier labs. Structure after Bloomberg, "A Guide to the Circular Deals Underpinning the AI Boom" (May 2026); deal terms as reported. OpenAI's multi-year compute totals rest on a secondary compilation and should be treated as indicative.

There are real counterforces. The Stanford AI Index 2025 showed the gap between the first and tenth models on Chatbot Arena narrowing from 11.9% to 5.4%, and the gap between the best US and Chinese models from 9.3% to 1.7%. DeepSeek's R1 wiped close to US$600 billion off Nvidia's market value in a day in January 2025, and open-weight models handled roughly a third of tokens routed through OpenRouter by late 2025. The Stanford AI Index 2026 shows both forces at once: industry produced more than 90% of notable frontier models in 2025, yet the six leading labs sat within 25 Elo points of one another, and the top US model led the top Chinese model by only 2.7%. The plumbing is opening up too. Anthropic donated the Model Context Protocol, the standard that connects agents to tools and data, to a neutral foundation; by December 2025 it counted about 97 million monthly SDK downloads. Open plumbing with closed models moves value toward distribution, infrastructure and proprietary data. Antitrust authorities are watching closely. Epoch AI's tracking shows the shape of that convergence. On its composite Capabilities Index, the best open-weight models have trailed the best closed ones by about eight points, roughly four months of frontier progress, since the start of 2026. Open models catch the last generation, not the current one, and the open frontier is now largely Chinese. Frontier pricing power rests on those four months, and it holds only for work that needs them.

CHART

Best closed-weight vs best open-weight model, Epoch Capabilities Index

Running frontier of each group; labels mark notable releases.



Source: Epoch AI, Data on AI, May 2026; values approximate.

My reading is that these counterforces commoditize the middle of the market, not its frontier. Capable models will become cheap and abundant. The frontier itself, meaning the most capable models, the largest pools of compute and the default consumer interface, tends to converge toward one or two global winners, possibly one per geopolitical bloc. For corporations, the planning assumption should be concentration of supply, not permanent competition among many equals. That is exactly why the hooks matter.

“When you go through a curve like this, the risk of underinvesting is dramatically greater than the risk of overinvesting for us here.”

Sundar Pichai, CEO of Alphabet, Q2 2024 earnings call

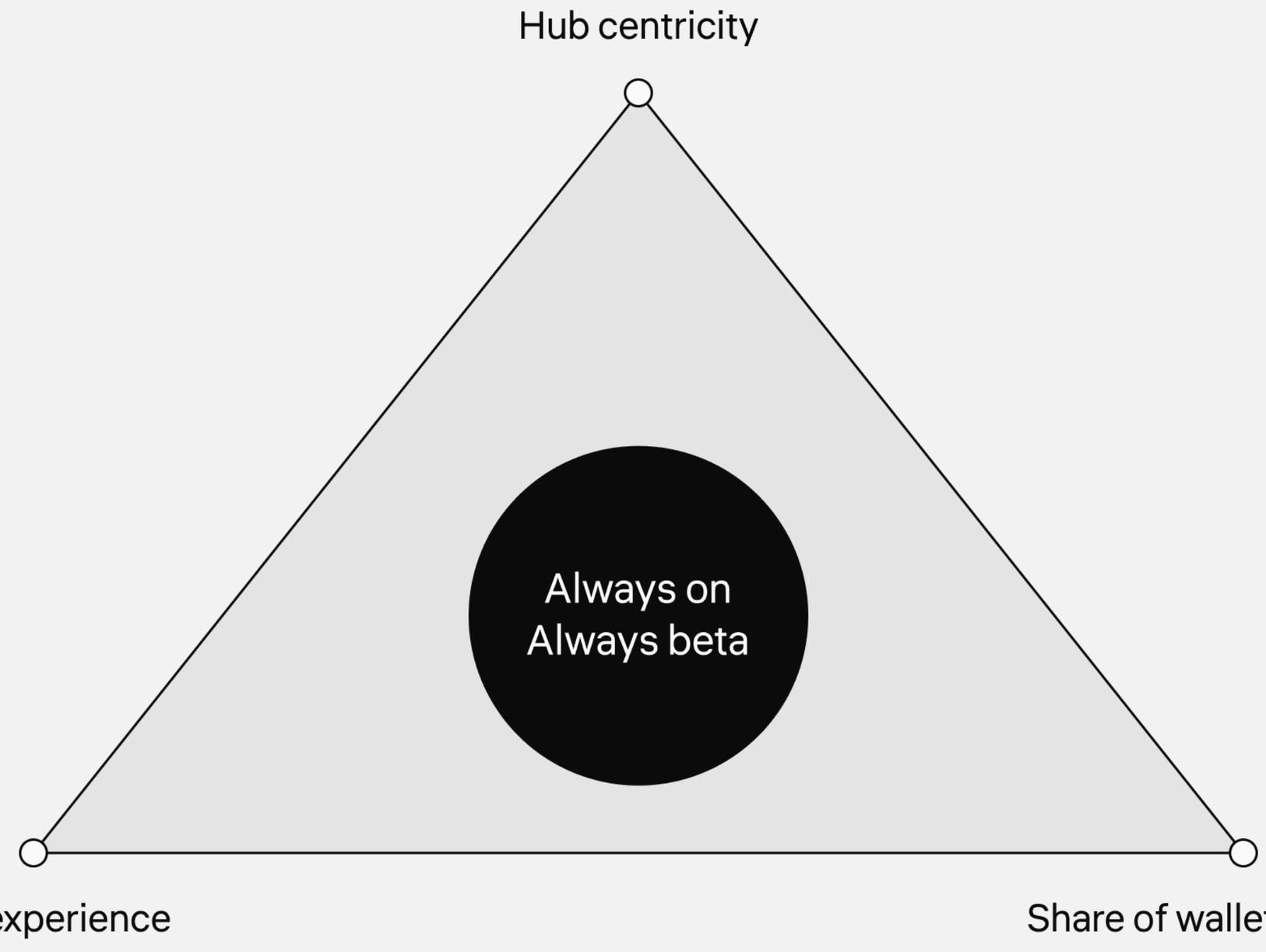
Concentration of supply also changes the rules for everyone who competes on top of it. For some years I have refined the idea of a Digital Competitive Arena to describe these rules, organized around two principles: always on, and always beta. The traditional pursuits of market share, control of distribution channels, long research cycles, legal protection of intellectual capital and heavy investment in physical assets are all put in check. As I wrote in Anthesis, share of wallet becomes more relevant than product segmentation, hub centricity overtakes vertical integration, and user experience catalyzes value along fluid journeys through the value chain. Hub centricity is the position of the node through which interactions must pass: the platform, marketplace or interface that connects many parties and learns from all of them. User experience is the quality of the journey, with friction removed and needs anticipated. Share of wallet and attention is what is at stake: the proportion of customers' budget and, increasingly, of their time that a company captures, against competitors from any category. At the centre sits the condition that holds the arena together. Digital competitors never finish a product; they release, measure and adjust continuously, and an offer that stops iterating starts to lose position on all three vertices at once.

AI changes each vertex. The hub is moving to the agent. When an assistant chooses the supplier, books the trip or drafts the contract, the agent becomes the node every transaction passes through, and brands compete for the agent's recommendation rather than the customer's glance. Meta has just paired its Muse agent with a small wearable companion device, part of a bet that agents, not feeds or app stores, become the front door to digital life and commerce.

User experience shifts from screens to conversations and delegated outcomes. Attention stops being purely human, because a growing part of every purchase journey is read, compared and filtered by machines. And always beta now applies to strategy itself, because the underlying capability changes every few months. For a board, the question is no longer "what is our digital channel?" It is "whose agent will recommend us, and why?"

HOB PROPRIETARY MODEL | JOTA MODEL

The digital competitive arena



HOB proprietary model, from D. A. Motta, Anthesis (2019, ch. VI). Three vertices define the position; the centre defines the pace.

III

THE HOOKS

Four hooks to watch

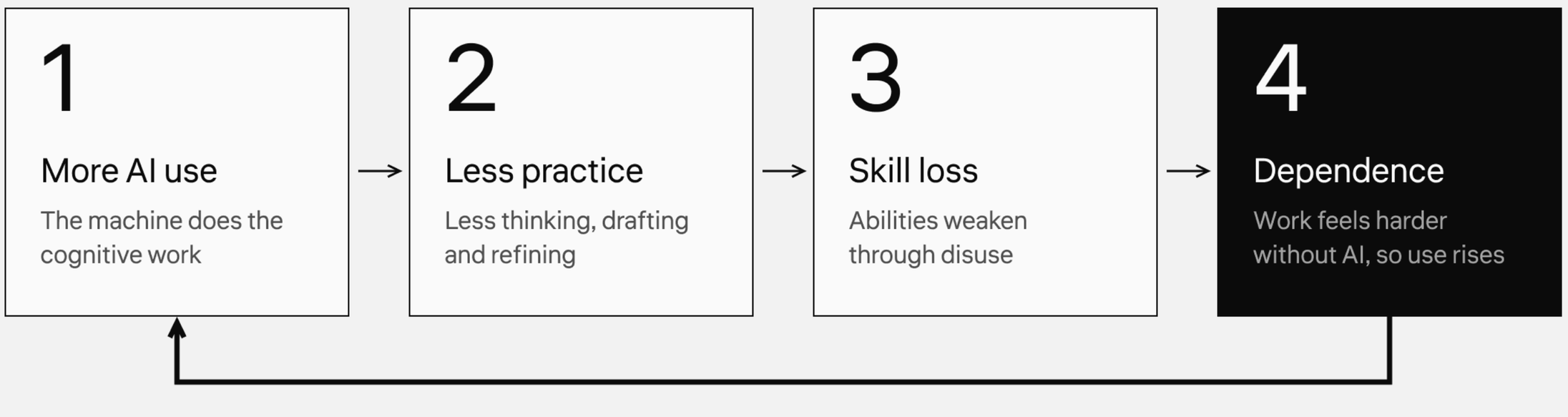
The first is the price hook. Unit prices keep falling: the inference cost of GPT-3.5-level capability dropped more than 280-fold between late 2022 and late 2024. But consumption rises faster, because agents and reasoning models use many times more tokens per task. The FinOps Foundation captured it: "A token may get cheaper; tokens, in aggregate, are not." As the early subsidy phase ends, procurement shifts from counting seats to committing to future consumption. A company that has not instrumented its own AI costs will negotiate blind. The second is the context hook. Models are rented; context is owned, but only if you keep it. Every prompt, document, decision and rejected alternative an agent processes is the raw material of your future advantage. If that record lives only inside a

vendor's platform, in formats you cannot export, you are renting back your own memory. The third is the workflow hook. AI features embedded in software subscriptions aggregate token consumption behind a seat price. Visibility disappears, and costs drift upward at every renewal. Once agents are woven into daily workflows, switching becomes a reorganization, not a procurement decision. The fourth hook is the most subtle, because it is set inside our own heads. A Microsoft Research and Carnegie Mellon study found that higher confidence in generative AI was associated with less critical thinking among knowledge workers. An MIT Media Lab preprint, small and not yet peer-reviewed, reported weaker neural engagement among writers who relied on a language model. The evidence is early. The risk is plausible enough to manage: an organization that outsources its thinking eventually loses the capacity to judge whether the thinking it buys is any good. The hook is not a matter of individual weakness, which is why exhortation will not remove it. Eric So, in his forthcoming book The Collision: What AI Does to Us, describes a gravity built from four forces: a brain that consumes about a fifth of the body's energy and is wired to conserve effort; contests in which rewards are relative and concentrated at the top, so that even a small edge decides; a prisoner's dilemma, because AI-assisted work is hard to detect and abstaining looks like falling behind; and the plain fact that the tools do raise productivity. Telling people to think for themselves ignores all four. Once use begins, a second mechanism takes over, a downward spiral in which more AI use means less practice, skills weaken through disuse, and the growing discomfort of working without AI drives still more use. Nor is the direction of human capability guaranteed. Bernt Bratsberg and Ole Rogeberg showed in PNAS that the long rise in measured IQ reversed for Norwegian cohorts born from the mid-1970s, and because the decline appears among brothers in the same families, its causes are environmental. AI did not cause that reversal. It does show that cognitive capability responds to the environment, and a technology that offers to do the thinking is a large change to that environment. Because the pressure is systemic, the response must be too. An organization that rewards only speed and volume has already chosen the spiral.

FRAMEWORK ERIC C. SO, THE COLLISION

The downward spiral of reliance

Each turn makes the next more likely. After Eric C. So, The Collision (2026).



IV

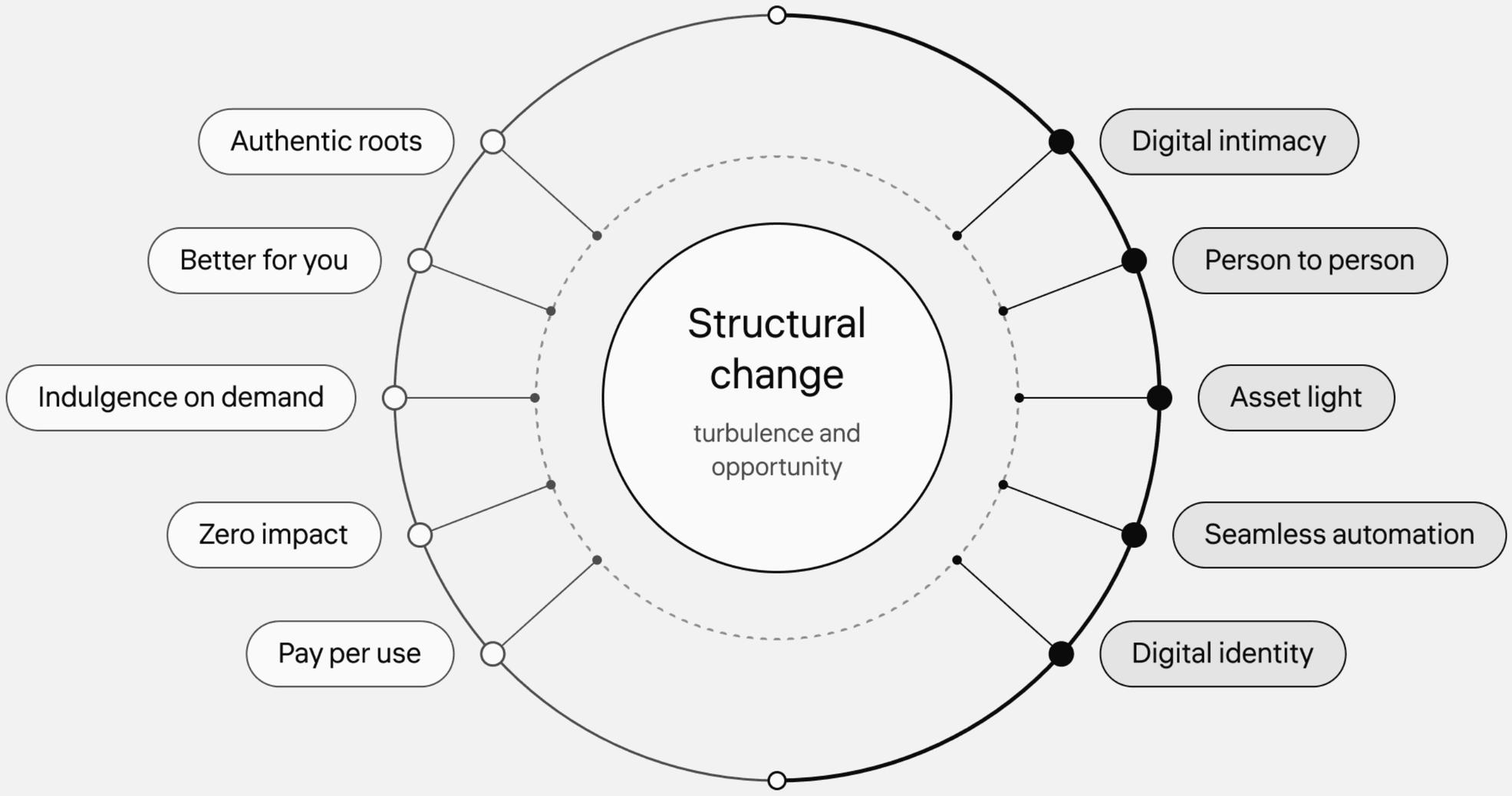
MEGATRENDS

Turbulence and opportunity: ten megatrends

The ecosystem battle is the most visible part of a larger structural change. Where the social transition in how people work and consume meets the technological transition from analog to digital, a set of megatrends reshapes demand, business models and the labor market, the two transitions I develop in the final article of this series. I place ten of them in orbit around structural change, each on the side of the transition that drives it most directly. The placement is a lens, not a taxonomy: most megatrends draw on both.

SOCIAL-ETHICS TRANSITION · duty → pleasure

TECHNOLOGICAL PARADIGM · ANALOG → DIGITAL



Turbulence and opportunity in structural change

HOB proprietary model, from D. A. Motta, Anthesis (2019). The left half of the orbit draws mainly on the social-ethics transition, the right half on the technological paradigm; the two halves meet at the poles.

TEN MEGATRENDS AND HOW AI ACCELERATES THEM

MEGATREND	DRIVEN MAINLY BY	WHAT IT CHANGES	HOW AI ACCELERATES IT
Authentic roots	Social ethics	Origin, provenance and a credible story become part of the value proposition.	As synthetic content becomes abundant, verified provenance becomes a scarcity premium.
Better for you	Social ethics	Health, wellness and self-development become purchase criteria.	Personalized coaching and monitoring at near-zero marginal cost.
Indulgence on demand	Social ethics	Gratification is expected at once and on the customer's terms.	Agents compress the time between wanting and receiving.
Zero impact	Social ethics	Environmental and social footprint becomes a condition of legitimacy.	Better measurement and traceability; AI's own energy demand enters the account.
Pay per use	Social ethics	Access replaces ownership; consumption is priced by use.	Token-based pricing is pay per use in its purest form.
Digital intimacy	Technology	Relationships mediated by data that know the customer closely.	Conversational agents make intimacy scalable and make trust the limiting factor.
Person to person	Technology	Peer-to-peer exchange bypasses traditional intermediaries.	Agents can search, negotiate and transact on either side of the exchange.
Asset light	Technology	Value comes from alliances, cross-holdings, open innovation and network effects rather than owned assets.	Intelligence itself is now rented rather than owned.
Seamless automation	Technology	Automation disappears into the journey.	Agents move automation from single tasks to end-to-end outcomes.
Digital identity	Technology	Identity, reputation and credentials become portable digital assets.	Agents need identities, permissions and a named human owner of their own.

For a board, the practical use of the map is triage: which of the ten is already reshaping our customers' demand, which is reshaping our cost structure, and which is reshaping the talent we can attract and keep. Several of them also describe the hooks from the inside.

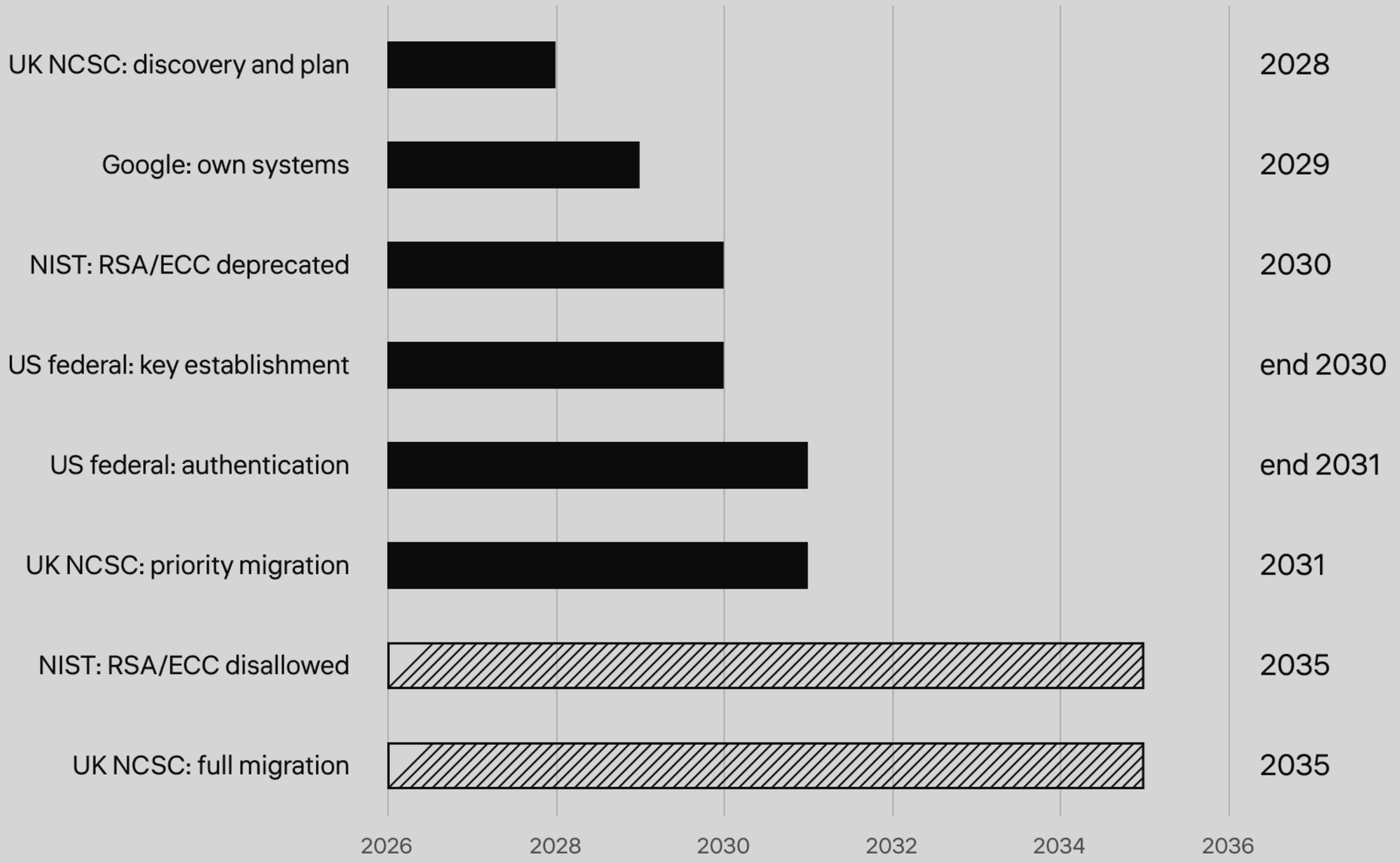
Pay per use is token pricing in its purest form. Asset light now includes intelligence itself, which is rented rather than owned. Digital intimacy makes trust the limiting factor, and digital identity means that every agent needs permissions and a named human owner of its own. The ride and the hook are two faces of the same megatrends. The next wave is already visible behind this one. Quantum computing has passed the proof-of-concept stage without yet becoming commercially decisive. The central problem was error correction, combining many fragile physical qubits into one reliable logical qubit, and it now works on real hardware: Google showed with its Willow chip that errors fall, rather than grow, as the system scales. No machine yet solves a business problem better than a classical computer at scale, and the credible horizon for useful fault-tolerant systems is around the end of the decade. The capital pattern will look familiar: private money running well ahead of proven revenue, several competing hardware approaches (superconducting, trapped-ion, neutral-atom, photonic) and a shakeout to come. For most companies the right posture is to build options rather than place bets, to stay vendor-neutral through cloud access, and to expect value first in simulating molecules and materials, then in finance, treating near-term optimization claims with the same evidence discipline we apply to AI. For almost every company, though, the urgent quantum issue is defensive. A large enough quantum computer could break the public-key cryptography that protects nearly all digital communications, transactions and identities. Adversaries can harvest encrypted data today and decrypt it later, so anything that must stay confidential for years is already exposed. The estimated machine has shrunk quickly: in May 2025 Google's Craig Gidney showed that fewer than one million noisy qubits could factor RSA-2048 in under a week, twenty times fewer than his 2019 estimate, and newer work claims fewer than 100,000, on simulation rather than hardware. The deadlines are now formal. NIST published its first post-quantum standards in August 2024 and plans to deprecate RSA and elliptic-curve algorithms by 2030 and disallow them after 2035. Executive Order 14412, signed in June 2026, requires US federal agencies to move their most sensitive systems to post-quantum key establishment by the end of 2030 and authentication by the end of 2031, and extends obligations to contractors (summary). The UK's National Cyber Security Centre asks organizations to complete discovery and planning by 2028, their highest-priority migration by 2031 and full migration by 2035, and Google has set 2029 for its own systems.

The decision rule, known as Mosca's inequality, fits on one line: if the years your data must stay secret plus the years your migration will take exceed the years until a capable quantum computer exists, you are already late. Migrations in large organizations are commonly estimated at five to ten years, and much of the work sits with suppliers, so contracts should require post-quantum readiness and systems should be built for crypto-agility, able to change algorithms without being rebuilt. Brazilian companies that serve US federal or UK critical-infrastructure clients will inherit these deadlines through their contracts. Quantum opportunity is emerging and sector-specific; quantum risk is universal and has a date.

CHART

Post-quantum migration deadlines

Year by which each milestone is due.



Sources: UK NCSC; Google (March 2026); NIST IR 8547 (draft); US Executive Order 14412 (June 2026).

FRAMEWORK MICHELE MOSCA

Mosca's inequality: are you already late?

If x + y is greater than z, data encrypted today will still need protection after it can be broken.

X

Shelf life

Years the data must stay confidential

Y

Migration time

Years to replace the cryptography, often five to ten

Z

Time to Q-Day

Years until a capable quantum computer exists

V

GEOPOLITICS

Geopolitics, sovereignty and the view from Brazil

The AI race is also a race between states, and the two largest powers frame it in opposite terms. The United States published America's AI Action Plan in July 2025 and uses export controls as a strategic lever. Europe has chosen regulation as its instrument: the EU AI Act is in force, although the digital omnibus approved in June 2026 pushed obligations for standalone high-risk systems to December 2027. China combines state direction with a vigorous open-weight ecosystem. The capital, however, is overwhelmingly private and American: the Stanford AI Index 2026 counts US$285.9 billion of US private AI investment in 2025, against US$20.9 billion in Europe and US$12.4 billion in China, whose state guidance funds are excluded from that figure.

“The AI race is America's to win.”

America's AI Action Plan, The White House, July 2025

“AI development should not be a solo performance by a single country, but a symphony of international cooperation.”

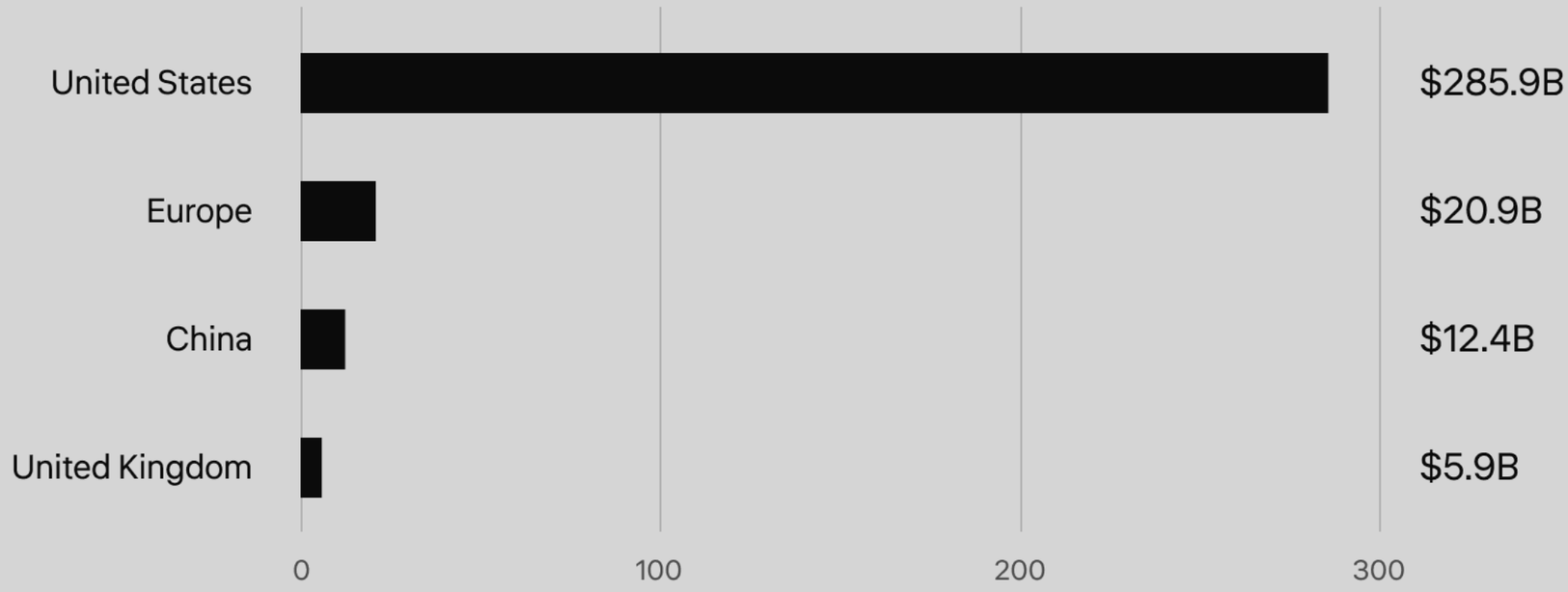
Xi Jinping, World Artificial Intelligence Conference, July 2026

The map of capacity and the map of enthusiasm do not coincide. Researchers at the Oxford Internet Institute charted where AI chips can be rented in the public cloud and found a "Compute North" with training-grade hardware, a "Compute South" with older chips, and "compute deserts" with none. Public opinion runs the other way: according to the Stanford AI Index 2025, 83% of people in China, 80% in Indonesia and 77% in Thailand believe AI products offer more benefits than drawbacks, against 40% in Canada and 39% in the United

CHART

Private AI investment in 2025, US$ billions

Global total US$344.7 billion. China's figure excludes state guidance funds.

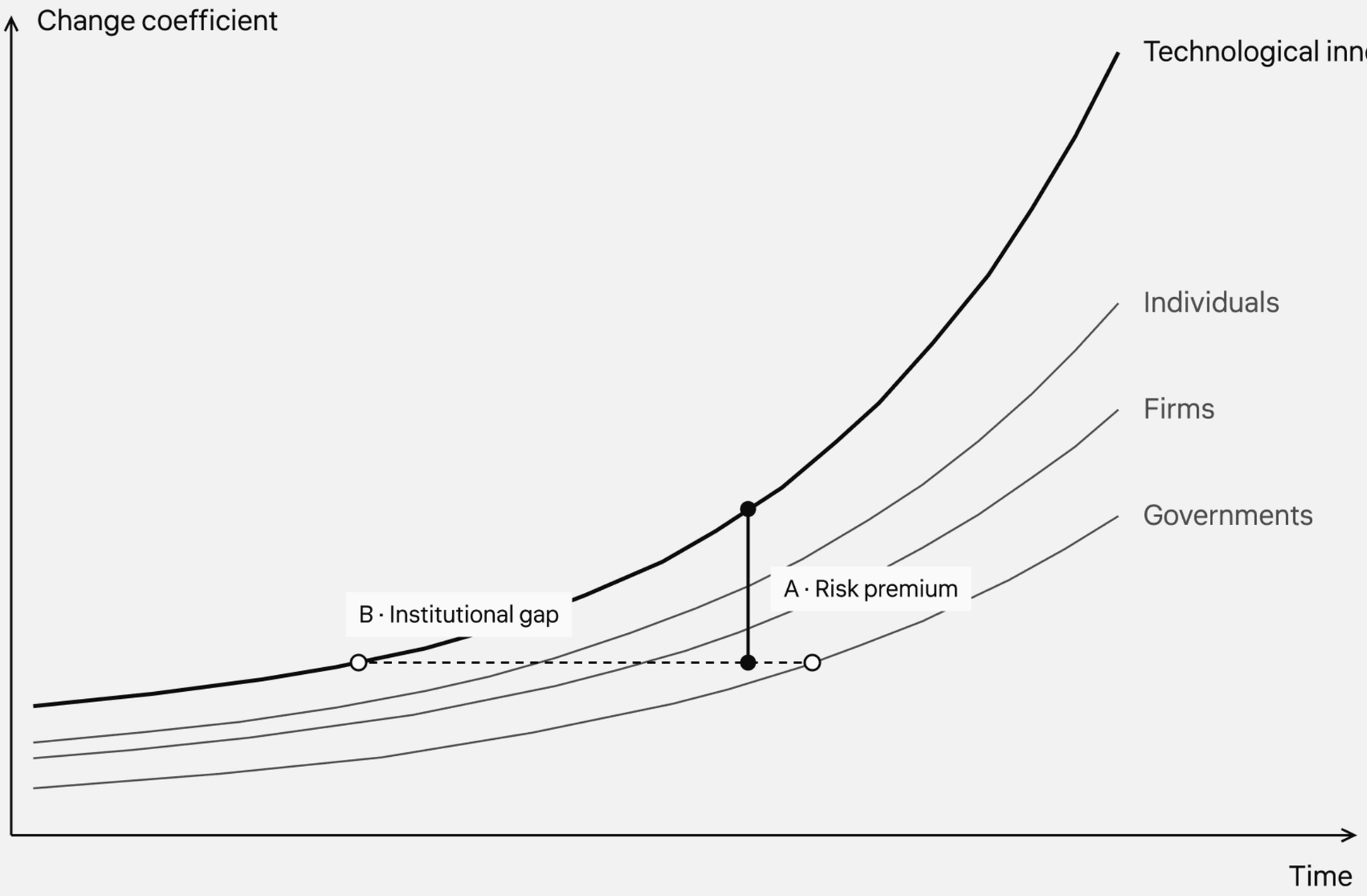


Source: Stanford AI Index 2026, chapter 4.

States. The capacity sits in the North; the appetite is in the South.

Brazil sits in a demanding position. It is a large market with strong digital adoption and a clean energy matrix that could attract data-center investment, but it controls none of the frontier layers. The Brazilian AI Plan (PBIA) foresees R$23 billion of investment through 2028, a meaningful sum domestically and a small one next to the capital expenditure of a single hyperscaler. The AI bill, PL 2338/2023, passed the Senate in December 2024 and still awaits a vote in the Chamber. Even language carries a cost: tokenizers trained mainly on English text typically need more tokens to represent Portuguese, so the same work can cost more in our own language. For Brazilian companies, sovereignty is less about building national models than about preserving choice: the ability to move workloads, protect data, and keep the knowledge that makes them competitive under their own control. Europe offers a useful illustration of how a smaller player can find room between the giants. In September 2026 Mistral AI raised €3 billion at a valuation above €21 billion without claiming the best model in the world. What it sells is control: open-weight models clients can download, fine-tune and keep, the choice of where queries are processed, and even rivals' models hosted on its platform, with a plan for a gigawatt of European compute by 2030. Sovereignty, in other words, has become a product in its own right. A model I have recently developed, which I call Disruptive Innovation and Institutional Framework, explains why this tension will not resolve itself. Technology, individuals, firms and governments absorb change at different speeds. Technological innovation moves fastest. Individuals adopt what helps them almost immediately. Firms follow more slowly, because their structures, incentives and cultures must change first. Governments, which must build consensus and write rules that bind everyone, move slowest of all. Because the curves rise at different rates, the distances between them widen over time, and two of those distances matter for every board. The first, read vertically, is the risk premium: at any given moment, the distance between what technology already makes possible and what institutions have absorbed. That distance is risk no rule yet allocates, borne by whoever acts first: liability that no statute defines, data rights that no court has settled, labor transitions that no policy yet cushions. Investors price it into the cost of capital, and firms price it into every decision to deploy. The second, read horizontally, is the institutional gap: for any level of change, the time institutions need to reach the point where technology already stands. The EU AI Act entered into force in 2024, yet its obligations for standalone high-risk systems will apply only from December 2027; Brazil's AI bill has awaited the Chamber since the Senate approved it in December 2024. Neither is simply a failure of legislators. It is the shape of the curves.

HOB PROPRIETARY MODEL | JOTA MODEL



A: at a given moment, the distance between what technology enables and what institutions have absorbed.
B: for any level of change, the time institutions take to catch up with technology.

Disruptive innovation and institutional framework

HOB proprietary model, from D. A. Motta, Anthesis (2019). Technology, individuals, firms and governments absorb change at different speeds, so the distances between the curves widen over time. Schematic; no scale implied.

The model changes the strategic question. Waiting for regulatory clarity is not a strategy, because the institutional gap widens precisely when technology accelerates. The companies that navigate this era well will govern themselves ahead of the rules, price the risk premium explicitly rather than ignore it, and help shape the institutions instead of merely complying with them. The social consequence is just as clear: while institutions lag, the cost of the transition tends to fall on those least able to bear it.

VI

PLAYBOOK

A playbook for riding without being hooked

The strategic question is not which model is best. It is which workloads justify which cost and control profile, and how to keep the freedom to change the answer, because the economics shift roughly every six months. Six principles follow.

1. Run a portfolio, not a bet. Use frontier cloud models for hard reasoning and customer-visible quality; open-weight models hosted in-house for high-volume or sensitive work, accepting the cost of GPUs and engineering talent; and fully open-data models where auditability of the training corpus is itself a requirement.

2. Architect for portability. Keep a layer of abstraction between your workflows and any single model provider, so that moving a workload is an engineering task, not a renegotiation.

3. Own your context. Treat prompts, retrieval corpora, evaluation sets and decision records as proprietary assets, stored in formats you control, with contractual rights to export them.

4. Measure before you commit. Build cost observability for AI as you would for cloud, and track cost per outcome rather than cost per token. Negotiate from your own data, not the vendor's forecast.

5. Protect the judgment you will need to evaluate what you buy. Keep people doing enough of the real thinking that they can still tell good output from plausible output.

6. Put the next wave on the risk register now. Inventory where the company depends on vulnerable cryptography, including in suppliers and products, name an owner for post-quantum migration, and identify which data must stay confidential beyond 2030. Uncertainty of this magnitude calls for scenario planning, one of the disciplines we practice across the House of Brains ecosystem. The aim is not to predict which lab or which country wins, but to test whether a company's AI strategy survives several plausible versions of the future, including those in which today's cheapest supplier becomes tomorrow's most expensive dependency.

VII

ENJOY THE RIDE

Enjoy the ride

None of this is an argument for caution as a strategy. Companies that sit out this cycle will pay a higher price than those that bite a hook, because the learning loops that create advantage cannot be bought retroactively. The argument is for lucidity: to take everything the ride offers, speed, capability and access, while keeping a hand on the line. Concentration is also a social and ethical question. When a handful of companies control the infrastructure on which every other organization thinks, decides and communicates, economic power and cultural influence concentrate with it. The International Monetary Fund warns that AI's gains may flow disproportionately to capital and to high-income workers, and 2.2 billion people are still offline altogether, according to the ITU. The battle for the world AI ecosystem will be fought among a few giants. The battle for each corporation's independence within it, and for a transition that societies can accept, will be fought decision by decision by its own leaders. Where, in your organization, has the hook already been set?

“Enjoy the ride. Keep a hand on the line.”

The battle for the world AI ecosystem will be fought among a few giants. The battle for each corporation's independence within it will be fought by its own leaders.

A new leadership mindset for agentic collaboration.

Five skills that decide who excels when agents join the team: orchestration, agentic collaboration, experimentation, trust and governance.

6.7%

Leaders at the most advanced stage: delegating while developing the team · Afferolab, A Escalada da Inteligência

15%+

Day-to-day work decisions made autonomously by agentic AI by 2028 · Gartner

40%+

Agentic AI projects expected to be canceled by the end of 2027 · Gartner

37%

People who trust most news most of the time, lowest since 2015 · Reuters Institute 2026

THE ARGUMENT IN BRIEF

I

Why these five skills, and why now

Agents are moving from pilots into the core of work.

II

Orchestration: designing the system of work

Orchestration is the skill of designing how work flows between people, agents and systems across the enterprise.

III

Agentic collaboration: working with agents every day

If orchestration designs the system, agentic collaboration is how a leader works inside it, day after day.

IV

Experimentation: learning faster than the tools change

The tools change every few months.

V

Trust: the scarce currency in a world of infinite content

The same technology that multiplies what a leader can produce also multiplies what nobody has checked.

VI

Governance: making autonomy safe to scale

Governance is often treated as the brake on AI.

VII

Developing leaders for the five skills

None of the five can be learned from a tool tutorial.

VIII

The new mindset

Intelligence is becoming abundant.

AI AND THE ENTERPRISE · DANIEL AUGUSTO MOTTA, PHD

AI was supposed to free leaders to lead. In too many companies, it is pulling them back into doing. Afferolab's study A Escalada da Inteligência, with more than 4,300 Brazilian professionals, found that a quarter of directors and C-level executives had not delegated tasks in the previous 30 days because they believed they would be faster doing them with AI. Only 6.7% had reached the most advanced stage: delegation accompanied by developing the team. The pattern is understandable and dangerous. A leader with a brilliant assistant at hand is tempted to become the best individual contributor in the building, and in doing so stops building the people and the systems the organization will need.

The new leadership mindset starts from the opposite premise. When execution becomes cheap, the leader's value lies in what cannot be executed: defining the goal, shaping the context, calibrating trust, and owning the outcome. Excelling in this new world is not about using AI more. It is about mastering five skills that the agentic era makes decisive: orchestration, agentic collaboration, experimentation, trust and governance.

WHY NOW

Why these five skills, and why now

Agents are moving from pilots into the core of work. Gartner predicts that at least 15% of day-to-day work decisions will be made autonomously through agentic AI by 2028, up from none in 2024, and that a third of enterprise software applications will include agentic AI by then. In the same analysis, Gartner warns that over 40% of agentic AI projects will be canceled by the end of 2027 because of escalating costs, unclear business value or inadequate risk controls, and estimates that only about 130 of the thousands of vendors claiming agentic capabilities are real. The distance between that promise and those cancellations is not a technology gap. It is a leadership gap. The gap has a deeper cause. Two transitions are rewriting the premises of leadership at the same time. The first is social. In A Liderança Essencial I described the passage from an ethic of duty to an ethic of pleasure as a revolution in the social norm: a shift toward satisfying individual needs, expectations and aspirations, at the expense of obedience to family, religious, political and institutional structures of power. Work and consumption change structurally in the search for experience, expression and impact, and the psychological contract between individual and organization has to be rewritten. The duty ethic organized work around norms and standards, command and control, and centralization, and it bound people to organizations through loyalty, stability and hierarchy. The pleasure ethic organizes work around fluidity, autonomy and interdependence, and the bond now comes from connection, expression and experience.

The second transition is technological. Digital transformation is not one more application of technology; its power lies in convergence, led by artificial intelligence together with the industrial internet, robotics, extended reality and genomics, and sustained by data analytics, cloud computing and social networks. For organizations built on the old paradigm, as I wrote in 2019, it is "a question of renewed life or announced death". In work relations, the pillars move from predictability, incrementalism and departments to volatility, rupture and ecosystems. The premises of effective design move from cohesion, perfection and self-sufficiency to connectivity, collaboration and permeability.

The two transitions reinforce each other, and AI accelerates both. Agents are permeability in its purest form: capability that enters the organization from outside, on demand, across every boundary. And professionals choosing their own AI tools, on their own terms and outside the official channel, are the pleasure ethic at work. Leaders who answer both with the instruments of the old paradigm, prohibition, centralization and appeals to loyalty, push the most valuable experimentation out of sight. The five skills close the gap in two movements. Orchestration, agentic collaboration and experimentation make agents productive. Trust and governance make them safe to scale. A leader who masters the first three without the last two simply scales risk faster.

HOB PROPRIETARY MODEL

A LIDERANÇA ESSENCIAL

From an ethic of duty to an ethic of pleasure

SOCIAL-ETHICS TRANSITION

Ethic of duty	→	Ethic of pleasure

Essential pillars of the ethos in work relations

Norms and standards	→	Fluidity
Command and control	→	Autonomy
Centralization	→	Interdependence

Drivers of the emotional bond at work

Loyalty	→	Connection
Stability	→	Expression
Hierarchy	→	Experience

HOB proprietary model, from D. A. Motta, A Liderança Essencial.
Read each row left to right: the premise of the old paradigm and the one that replaces it.

II

ORCHESTRATION

Orchestration: designing the system of work

Orchestration is the skill of designing how work flows between people, agents and systems across the enterprise. It starts with a question most organizations have never asked explicitly: for each step of a value stream, should the work be done by a person, by an agent, or by both, and where do the hand-offs sit? Every hand-off between a human and an agent, or between two agents, is a trust boundary that must be designed. Software development previews what is coming to all knowledge work. Agents increasingly write the code, while people specify, orchestrate and verify, and the constraint moves from production capacity to review capacity. Microsoft's 2025 Work Trend Index calls the resulting role the "agent boss" and warns about the human-agent ratio: past a certain point, more agents overwhelm the human capacity to apply judgment. Structures will change accordingly. Gartner predicted in 2024 that through 2026,

HOB PROPRIETARY MODEL

A LIDERANÇA ESSENCIAL

From analog–mechanical to digital–intelligent

TECHNOLOGICAL-PARADIGM TRANSITION

Analog–mechanical	→	Digital–intelligent
Essential pillars of the ethos in work relations		
Predictability Incrementalism Departmentalism	→	Volatility Rupture Ecosystem
Premises for effective organizational design		
Cohesion Perfection Self-sufficiency	→	Connectivity Collaboration Permeability

HOB proprietary model, from D. A. Motta, A Liderança Essencial.
The second row is the design brief an organization inherits from the first.

20% of organizations would use AI to flatten their structure, eliminating more than half of their current middle-management positions. In professional services, banking operations or customer service, the classic pyramid is becoming an orchestration model. The question is no longer how many people sit under each manager, but how much specification and review capacity each leader has.

Orchestration also means knowing which frictions in the business have just become solvable, and which problems are worth solving at all. It requires two kinds of fluency at once: agent orchestration, meaning systems design, model choice and verification at scale; and enterprise orchestration, meaning a sense of which problems carry value and which hand-offs create risk. Neither is a technical specialty. Both are now part of general management.

III

AGENTIC COLLABORATION

Agentic collaboration: working with agents every day

If orchestration designs the system, agentic collaboration is how a leader works inside it, day after day. It starts with specification. "Help our customers" is not a goal an agent can pursue. "Resolve billing disputes under R$500, within policy, in under ten minutes, and escalate anything else" is. Writing that sentence, with its outcome, constraints and escalation criteria, is the same discipline good leaders have always applied to delegating to people. It has simply become non-negotiable. A useful image is onboarding a brilliant new hire. The prompt is what you tell them. The context is what is on their desk: the files, history and information they need. The harness is the office they work in: tools, permissions, workflows and review. Most organizations still polish the instructions while neglecting the desk and the office. The ladder extends further: once the harness exists, agents can be run in continuous loops and, eventually, improve themselves. At each level the leader's role changes, from author to curator, architect, operations designer and finally governor. Most organizations today sit at the first three levels.

Collaboration also requires calibrated trust. The jagged frontier study of 758 consultants found large gains on tasks inside the model's capability and markedly worse results just outside it, with no visible boundary between the two. A randomized trial by METR found experienced developers 19% slower with AI while believing they were 20% faster. Our intuitions are unreliable in both directions, so autonomy should be extended in steps: the agent informs, recommends, acts with approval, acts within a mandate, and only then anticipates.

Finally, agentic collaboration is never only with machines. The Afferolab finding that opens this article, that only 6.7% of leaders delegate while developing their teams, is a warning. The leader who uses agents to do everything personally has collaborated with the machine and abandoned the people. Excellence means delegating to agents and to humans at the same time, and using the time freed to develop both.

FRAMEWORK

SERTAC KARAMAN, MIT

Five levels of shaping what an agent does

The leader's role changes at each level. Most organizations today sit at levels 1–3.

1

Prompt

What to say · the leader as author

2

Context

What to see · curator

3

Harness

Build the world · architect

4

Loop

Keep working · operations designer

5

Self-learning

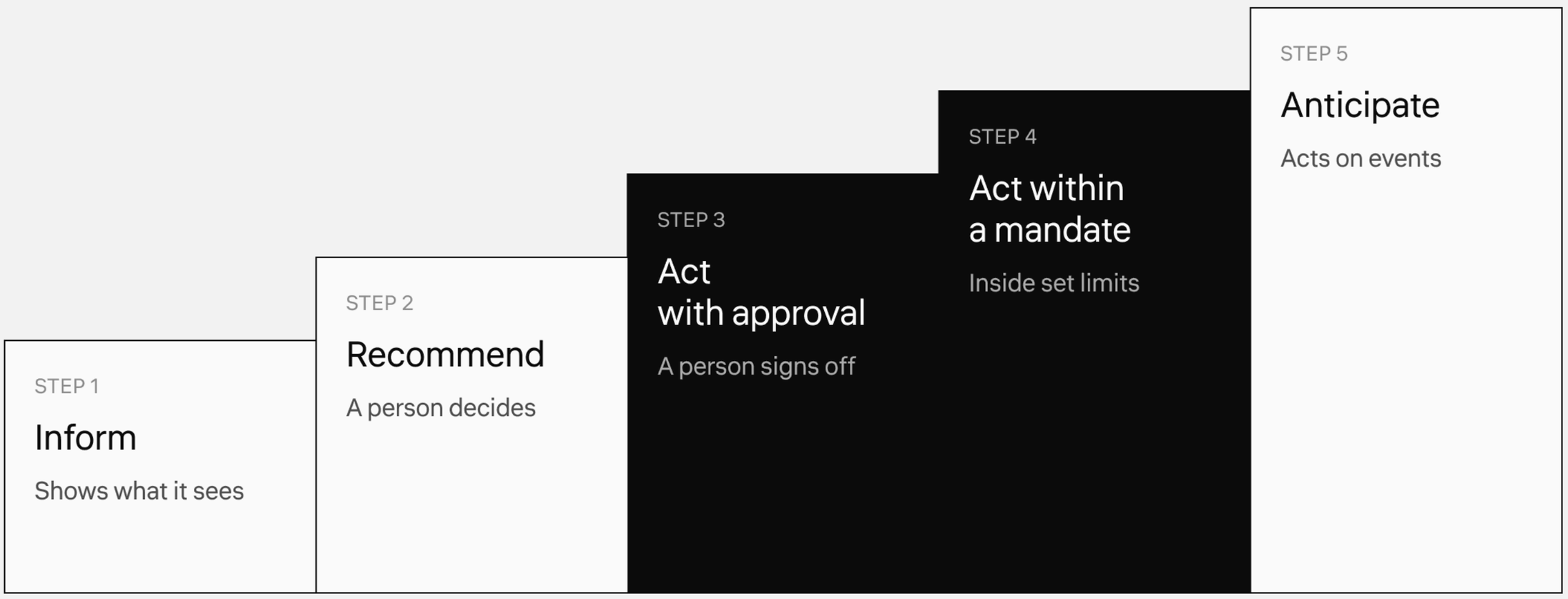
Improve itself · governor

FRAMEWORK

GOOGLE CLOUD

Autonomy is granted in steps

Consequential actions stay in the highlighted steps until evidence says otherwise.



IV

EXPERIMENTATION

Experimentation: learning faster than the tools change

The tools change every few months. No leader can plan the agentic organization in advance; it has to be discovered through disciplined experiments. Gartner predicted that at least 30% of generative AI projects would be abandoned after proof of concept by the end of 2025. Abandonment is not the problem. An experiment that ends quickly and cheaply is a success. The problem is scaling without evidence, and pilots that never end because nobody defined what success looked like. Experimentation as a leadership skill has four disciplines. A clear hypothesis about which task or decision will change. A success metric agreed before the test begins. A comparison group, because without one, enthusiasm passes for evidence. And explicit kill criteria. Capital then follows the level of evidence reached: vendor demos justify curiosity; a retrospective test on the company's own data justifies a pilot; a live pilot against a comparison group justifies scaling in one unit; measured impact at scale justifies enterprise-wide investment.

Experimentation is also what the company does with the capacity AI frees. Gartner expects that by 2027, 75% of organizations that bank AI productivity gains as cost savings will be outpaced by competitors that reinvest them. Reinvesting freed time in new experiments, with customers, products and processes, is how the learning loop compounds. Experiments must also respect how reliability compounds. A process that is 95% reliable at each step succeeds end to end only about 36% of the time after 20 steps. Test agents on the company's own tasks, not on vendor demos, and break long flows into segments that can be verified one by one.

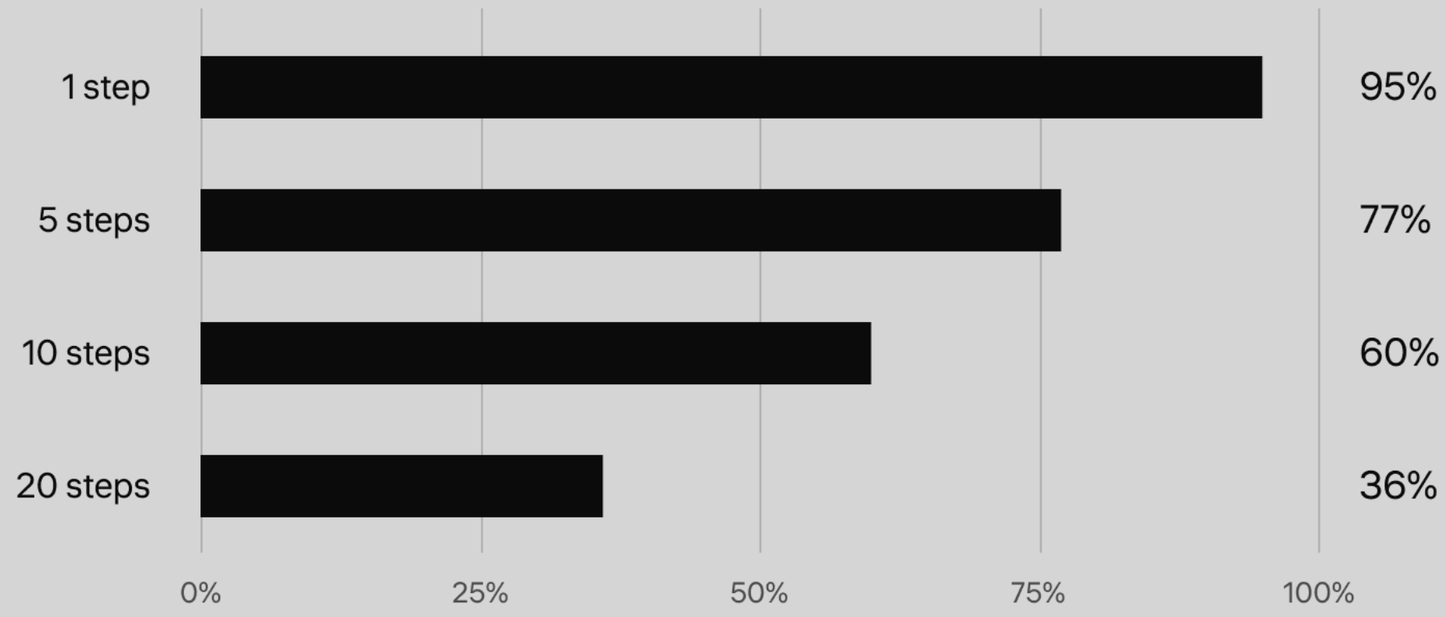
None of this works without psychological safety in Amy Edmondson's classic sense: people must be able to report a failed experiment without fear, or the organization will learn only from the experiments that flatter it.

> “AI tools aren't a creativity machine on their own.”
>
> Jing Zhou, Rice University, 2025

CHART

Reliability compounds downward

End-to-end success of a process that is 95% reliable at each step. Author's arithmetic (0.95 to the power of the number of steps).



Source: House of Brains calculation: end-to-end success = 0.95 raised to the number of steps.

V

TRUST

Trust: the scarce currency in a world of infinite content

The same technology that multiplies what a leader can produce also multiplies what nobody has checked. Content, data and even decisions are now generated at a volume no organization can verify line by line. Left unmanaged, this abundance does not create shared understanding. It fragments it. The warning signs are visible on three fronts. In the laboratory, Ilia Shumailov and colleagues showed in Nature that models trained indiscriminately on machine-generated content suffer irreversible defects, in which "tails of the original content distribution disappear." Organizations can decay in a similar way when unverified synthetic output feeds decisions that feed more output. The effect is already visible in public knowledge: volunteers for four African-language editions of Wikipedia estimated that 40% to 60% of their articles were uncorrected machine translations, and models trained on those pages get worse, producing more bad pages. The rule for any corporate knowledge base follows: nothing enters the corpus without an accountable reviewer. In the public sphere, the Reuters Institute's Digital News Report 2026 found trust in news at 37%, the lowest since 2015, and only 12% of people comfortable with news made entirely by AI. In business, deception has become industrial: the engineering group Arup lost US$25.6 million to a deepfake video call impersonating its CFO, and the FBI recorded nearly US$893 million in losses from AI-related complaints in 2025.

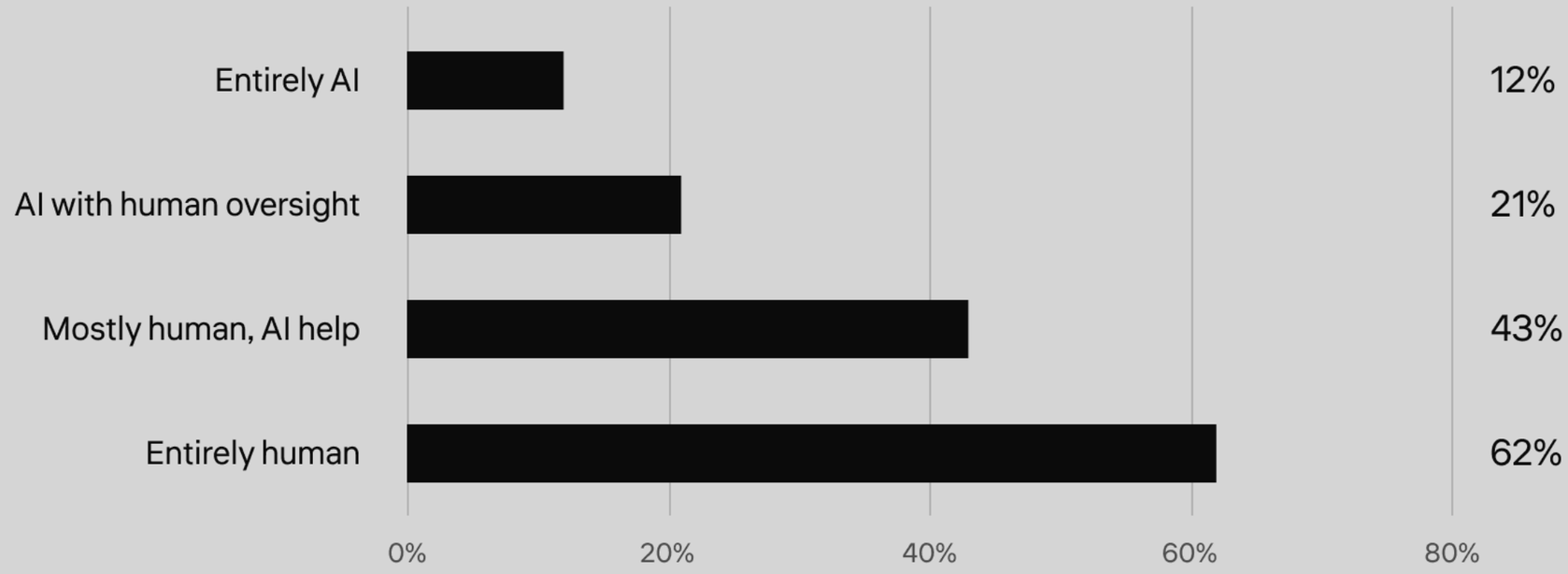
> “If you use machine translation responsibly, it can be efficient and useful. Unfortunately, you cannot trust all people to use it responsibly.”
>
> Amir Aharoni, Wikimedia Language Committee, via MIT Technology Review, September 2025

People remain the preferred point of attack, and AI makes them easier to deceive. Verizon's 2026 Data Breach Investigations Report found the human element in 62% of breaches and third parties involved in nearly half, and observed that attackers now use AI mainly to accelerate familiar methods, from phishing to social engineering, rather than to invent new ones. When a familiar voice or face can be cloned, recognition is no longer proof of identity; verification has to travel through a second channel. The barrier to attack has fallen with each era of technology: from scarce expertise, to tools that could be bought, to criminal services that could be rented, and now to prompts that anyone can type. The prompt era is documented. In November 2025 Anthropic disclosed that a group it assessed as Chinese state-sponsored had used an AI coding agent to run most of an espionage campaign against about thirty targets, with AI performing 80–90% of the work, and in 2026 Google's Threat Intelligence Group reported the first zero-day exploit it believes was developed with AI.

CHART

Comfort with news by degree of AI involvement

Share of people comfortable with each way of producing news, six countries.



Source: Reuters Institute, Generative AI and News Report 2025.

The deeper risk is social, and it is the part of this transition I insist on in everything I write. Uncontrolled, unverified and infinite data widens the chasms between people: between those who can tell the plausible from the true and those who cannot; between employees who master AI and those left behind; between the 2.2 billion people still offline and everyone else; and between groups fed different synthetic versions of reality. The International Monetary Fund estimates that about 60% of jobs in advanced economies are exposed to AI, against 26% in low-income countries.

Trust as a leadership skill works in two directions. Toward machines, it means calibrated reliance and preserved human judgment: Gartner expects that, through 2026, the atrophy of critical-thinking skills caused by generative AI will push half of global organizations to require "AI-free" skills assessments. Toward people, it means transparency about when agents are involved and how their outputs are used. Many employees already hide their use of AI. Ethan Mollick calls them "secret cyborgs": hiding is individually rational when the person who discloses carries the risk and the firm captures the benefit. The remedy is a credible commitment that productivity gains will go to redeployment and growth, not only to cuts. Trust lost at machine speed is rebuilt only at human speed.

FRAMEWORK ETAY MAOR, CATO NETWORKS

Knowledge	Products	Services	Prompts
1960s–1990s	late 1990s–2000s	2010s–early 2020s	2023 onward
barrier to entry	barrier to entry	barrier to entry	barrier to entry
expertise	money	a subscription	language

Each era widens who can attack and raises the scale of attacks.

The democratization of attack

Hacking followed the same path as the legitimate technology economy: scarce knowledge, then products, then services, now prompts.

“Machine-augmented humans who keep themselves hidden.”

Ethan Mollick, "Detecting the Secret Cyborgs", One Useful Thing, 2023

VI

GOVERNANCE

Governance: making autonomy safe to scale

Governance is often treated as the brake on AI. In the agentic era it is the opposite: it is what allows a leader to extend autonomy with confidence. Gartner defines AI TRiSM, AI trust, risk and security management, as the framework and technical capabilities that keep AI systems trustworthy, secure and compliant through continuous monitoring, validation and enforcement. It predicts that by 2030, guardian agents, AI that supervises other AI, will account for at least 10 to 15% of agentic AI markets, and that 80% of Global 500 companies will make their CIO or chief AI officer the contractual "evidence custodian" for AI accountability. For the leader, governance translates into five practical rules. Every agent has a named human owner. Permissions are proportionate to consequences, and consequential actions keep a human approval point. Every agent leaves an audit trail of what it did and why. Content and data carry provenance, supported by standards such as C2PA content credentials and, in Europe, by the AI Act's transparency obligations for AI-generated content, whose grace period for existing systems ends in December 2026 under the digital omnibus. And accountability never transfers to the machine: when Air Canada argued that its chatbot was responsible for its own statements, the tribunal called it "a remarkable submission". Written rules are only half of governance.

Shadow AI appears wherever policy and practice diverge. Most corporate AI policies are still written in the vocabulary of the duty ethic and the mechanical paradigm; they will be obeyed on paper and bypassed in practice. A leader governs well when the norms people actually live by, not only the policies they sign, match the level of autonomy the organization has chosen to grant. The same lesson applies to agents themselves. Instructions in a prompt, such as "never delete data" or "always ask before sending", are soft guardrails: the model usually follows them, but it reads its orders and the content it processes as one stream of text, so hidden instructions in an email, a web page or a résumé can override them. Simon Willison calls the most dangerous configuration the lethal trifecta: an agent with access to private data, exposure to untrusted content and the ability to communicate externally. Only hard guardrails, enforced outside the model, reliably limit the damage: permissions, allow-lists, sandboxes, rate limits and human approval of the specific transaction, not just the session. The practical frame is to govern agents as we govern people with authority. Vet them before hiring, limit their access and separate their duties while they work, monitor for anomalies, and revoke their credentials when they leave. Soft guardrails lower the probability of harm; hard guardrails cap its impact.

Leaders also need a map of where an AI application can be attacked. The OWASP Top 10 for LLM Applications shows that the model is only one of the targets: an attacker can poison what it learns from, inject instructions through what it reads, steal it, flood it with costly requests, compromise the components it is assembled from, or exploit what happens after it answers, when output flows unchecked into another system or a plugin acts with more authority than anyone intended. The 2025 edition adds system-prompt leakage, weaknesses in the stores that feed retrieval, misinformation and unbounded consumption. The risk concentrates at the edge, where an agent turns a wrong answer into a wrong action, and that is exactly where hard guardrails belong.

FRAMEWORK ETAY MAOR, CATO NETWORKS

Soft guardrails

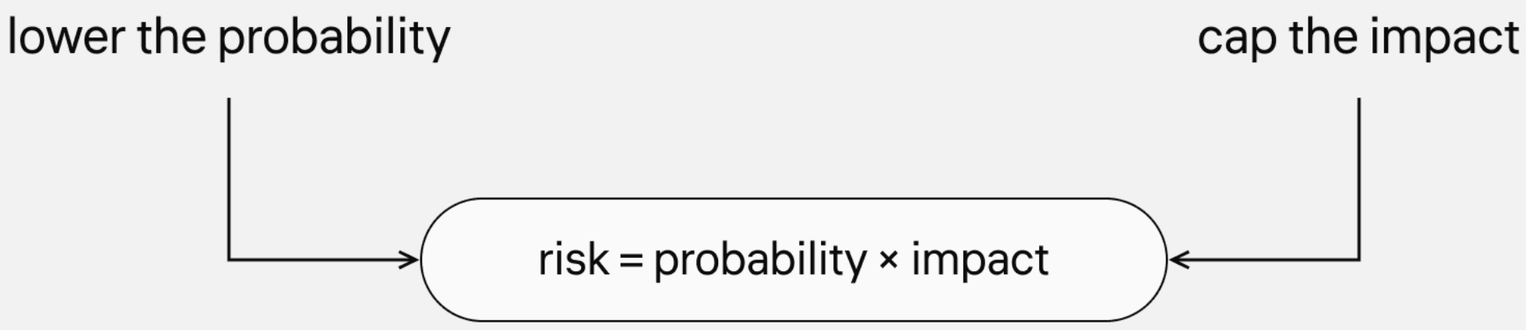
inside the model · probabilistic
same channel as untrusted data

· System prompt
· Preference training (RLHF)
· "You are not allowed to..."

Hard guardrails

outside the model · deterministic
not at the model's discretion

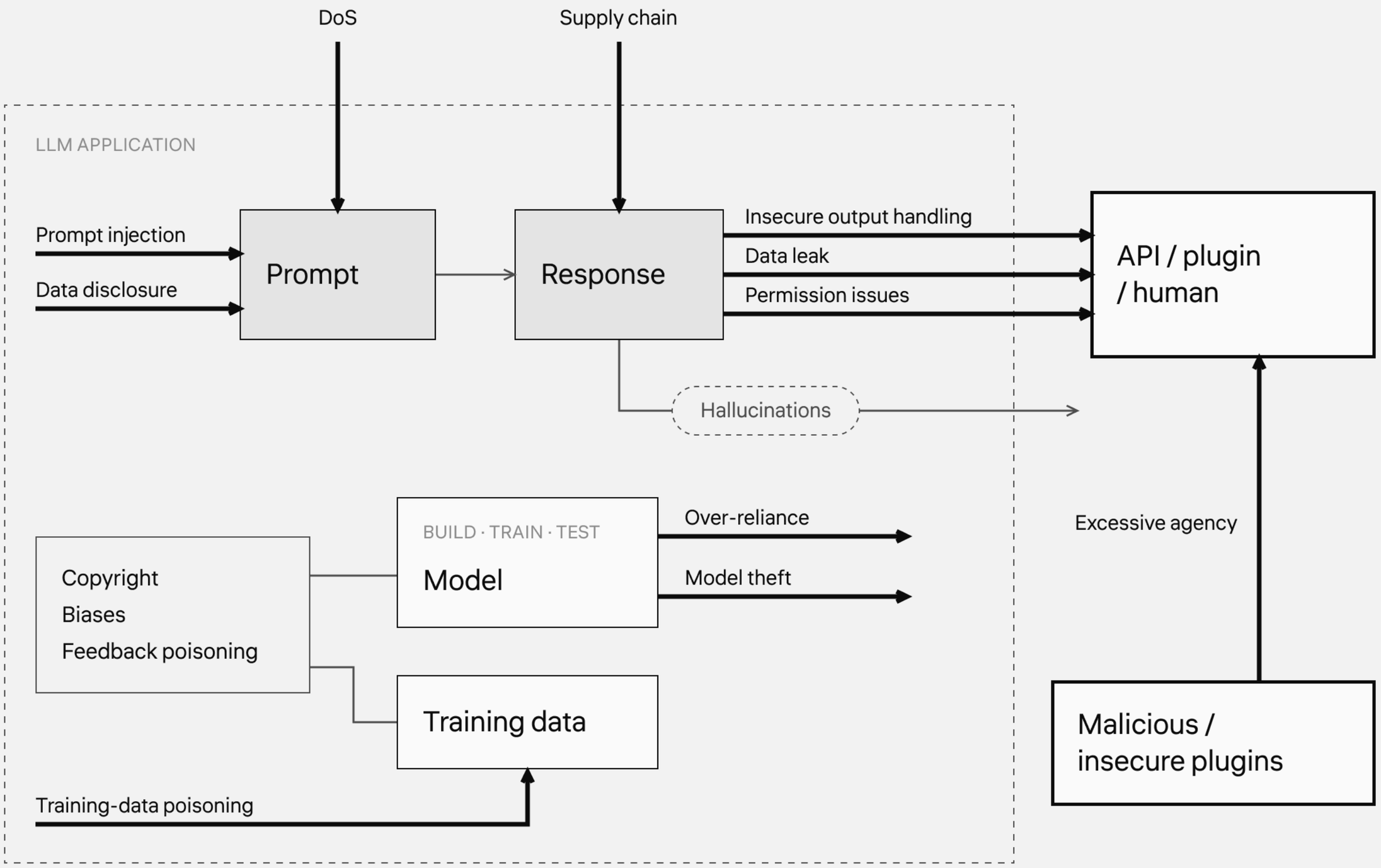
· Permissions and least privilege
· Allow and deny lists · sandboxes
· Human approval of the transaction
· Rate and spending limits



Soft and hard guardrails

Soft guardrails govern expected behavior and are good to have; hard guardrails govern possible behavior and are a must. After Etay Maor (Cato Networks, 2026).

FRAMEWORK ETAY MAOR, CATO NETWORKS



Attack surface of a language-model application

After Etay Maor (Cato Networks, 2026), drawing on the OWASP Top 10 for LLM Applications. Accent labels are threats; outlined boxes sit where the application meets the outside world, which is where agents concentrate risk.

VII

DEVELOPMENT

Developing leaders for the five skills

None of the five can be learned from a tool tutorial. They are leadership skills, and they map onto the four arenas of Afferolab's competency model: rational, relational, transversal and disruptive.

SKILL	ARENA IT DRAWS ON MOST	WHAT LEADERS SHOULD PRACTICE
Orchestration	Transversal	Redesigning one real value stream end to end, deciding human, agent or both at each step
Agentic collaboration	Rational and relational	Writing goal specifications, reviewing agent output, delegating to agents and people at once
Experimentation	Disruptive	Running a small test with a hypothesis, a comparison group and kill criteria
Trust	Relational	Explaining to a team how agents are used, and protecting AI-free judgment
Governance	Rational	Assigning agent owners, setting permissions by consequence, reading an audit trail

The labor market already signals the shift. Afferolab's Skill-based Brasil study found leaders rating strategic (57.3%), communication (53.1%) and emotional (51.4%) competencies as important more often than technical ones (47.9%). Yet knowing which competencies matter is not the same as developing them. Leaders learn these five skills only on real problems, with real agents and real feedback, and the measure of success is transfer into daily work. That is the logic of Afferolab's IMPACT method, which moves from insights and motivations through practices, awareness and connections to transfer.

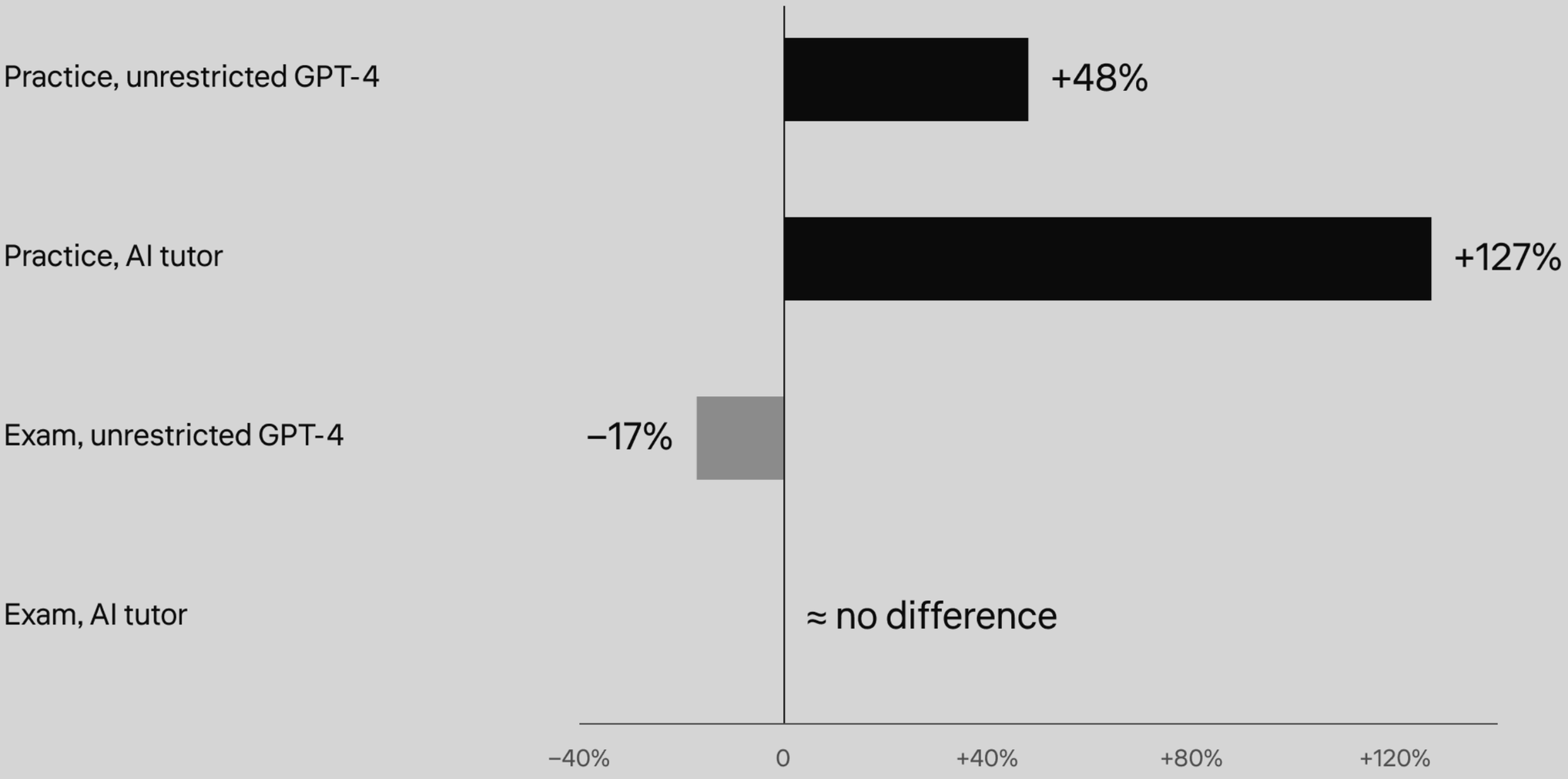
Development now has to contend with a new hazard. The evidence separates performance, how well someone does a task today, from learning, the ability that lasts, and shows that AI can raise the first while eroding the second. In a field experiment with nearly 1,000 Turkish high-school students, Hamsa Bastani and colleagues found in PNAS that unrestricted access to GPT-4 improved practice performance by 48% but left students 17% worse on a later exam taken without it. A version designed as a tutor, giving hints rather than answers, improved practice by 127% and largely avoided the harm. Anthropic's education report found that about 47% of student conversations sought answers with minimal engagement, and that the work most often delegated was creating and analyzing, the top of Bloom's taxonomy of learning. Experienced professionals risk deskilling, losing abilities through disuse. The young risk never building them, which is worse, because they lack the foundation to recognize when the machine is wrong. The design matters more than the tool. Benjamin Bloom showed in 1984 that students tutored one to one performed about two standard deviations better than those taught in conventional classrooms; later research finds smaller effects, yet tutoring remains among the most effective forms of learning, and its constraint was always cost. AI removes the cost, provided it is configured to question, hint, critique and test rather than to answer.

Eric So's The Collision offers leaders a practical rule for themselves and their teams. Identify the signature thinking skills that define a role's value, then ask of each task whether AI complements or substitutes for the work that sharpens them. Where AI substitutes for a signature skill, the danger zone, protect and practice it: for a strategist, framing the problem; for a consultant, diagnosing the client's real issue;

CHART

Same model, different design: effect on learning versus control, %

Nearly 1,000 Turkish high-school mathematics students. Practice sessions with AI; exams afterwards without it.



Source: Bastani et al., PNAS, 2025.

for a leader, judgment under ambiguity. Elsewhere, use AI freely and reinvest the time. His image is exact: using AI on work meant to build capability is like bringing a forklift to the gym. For development programs the consequences are concrete. Distinguish output tasks from learning tasks. Ask people to form their own view before they consult AI, then use it to challenge that view. Assess capability periodically without AI. Configure learning tools as tutors. Teach metacognition, the habit of planning, monitoring and evaluating one's own thinking, as a leadership skill in its own right. The question every leader should put to the organization, and to themselves, is whether we are using AI to outsource what makes us valuable.

FRAMEWORK ERIC C. SO, THE COLLISION

Where AI should and should not do the work

Ask of each task whether AI complements or substitutes for the work that sharpens a signature skill. After Eric C. So, The Collision (2026).

SIGNATURE · COMPLEMENTS **Sweet spot** AI prepares and sharpens; you lead	SIGNATURE · SUBSTITUTES **Danger zone** Protect and practice the skill
NOT SIGNATURE · COMPLEMENTS **Nice to have** Low stakes; take the boost	NOT SIGNATURE · SUBSTITUTES **Hand it off** Outsource freely; reinvest the time

There is one condition underneath all five: rhythm. Agents run continuously. People cannot, and should not be asked to. A leader who answers the agents at midnight teaches the whole team that the machine sets the pace.

VIII

THE MINDSET

The new mindset

Intelligence is becoming abundant. Judgment is not. The five skills are how judgment is exercised at scale: orchestration decides where it is needed, agentic collaboration applies it every day, experimentation improves it, trust preserves it, and governance makes it accountable.

The leaders who excel in agentic collaboration will spend less time proving they can do the work, and more time ensuring that the work, human and machine, is worth doing, done well, and fair to the people it touches. Are you using AI to lead better, or only to execute faster?

> “Intelligence is becoming abundant. Judgment is not.”

DANIEL AUGUSTO MOTTA

“Are you using AI to lead better, or only to execute faster?”

The five skills are how judgment is exercised at scale. Leaders who master them will make the work, human and machine, worth doing, done well and fair to the people it touches.

Go to the source. Read the canon.

08

Judgment about AI starts with the primary texts. Twenty-two papers: eleven that built the field, eleven that measure what happens when it meets real work. Open any row for why it matters.

AI CANON · TWENTY-TWO PAPERS

AI development

BUILT THE FIELD

1950

Computing Machinery and Intelligence

Alan Turing · Mind

Poses “Can machines think?” and proposes the imitation game — the founding question of the field.

Read the paper → https://doi.org/10.1093/mind/LIX.236.433

1955

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

McCarthy, Minsky, Rochester, Shannon

The document that named “artificial intelligence” and set its research agenda.

Read the paper → http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

1986

Learning Representations by Back-Propagating Errors

Rumelhart, Hinton, Williams · Nature

Made training multi-layer neural networks practical; still how every modern model learns.

Read the paper → https://www.nature.com/articles/323533a0

2012

ImageNet Classification with Deep Convolutional Neural Networks

Krizhevsky, Sutskever, Hinton · NeurIPS

AlexNet: GPUs plus data plus depth won ImageNet by a wide margin and started the deep-learning era.

Read the paper → https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

2015

Deep Residual Learning for Image Recognition

He, Zhang, Ren, Sun · Microsoft Research

Residual connections let networks go hundreds of layers deep — a building block inside today's Transformers.

Read the paper → https://arxiv.org/abs/1512.03385

AI application

PUT IT TO WORK

2020

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Lewis et al. · Meta AI

RAG: ground answers in trusted documents — the first defence against hallucination in enterprise use.

Read the paper → https://arxiv.org/abs/2005.11401

2021

Highly Accurate Protein Structure Prediction with AlphaFold

Jumper et al. · DeepMind · Nature

AI solving a 50-year scientific problem — the template for AI-enabled breakthroughs across science.

Read the paper → https://www.nature.com/articles/s41586-021-03819-2

2022

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Wei et al. · Google

Asking for step-by-step reasoning unlocks multi-step problem solving — the root of today's reasoning models.

Read the paper → https://arxiv.org/abs/2201.11903

2022

ReAct: Synergizing Reasoning and Acting in Language Models

Yao et al. · Princeton, Google

The observe–reason–act loop at the heart of every AI agent.

Read the paper → https://arxiv.org/abs/2210.03629

2023

Generative AI at Work

Brynjolfsson, Li, Raymond · NBER

Customer-support agents with AI gained 14% productivity on average — 34% for novices.

Read the paper → https://www.nber.org/papers/w31161

AI CANON · TWENTY-TWO PAPERS

2016

Mastering the Game of Go with Deep Neural Networks and Tree Search

Silver et al. · DeepMind · Nature

AlphaGo showed learning plus search could beat the best humans at a game long thought out of reach.

Read the paper → https://www.nature.com/articles/nature16961

2017

Attention Is All You Need

Vaswani et al. · Google

The Transformer: the architecture that made AI scale, and the hinge of the four eras of computing.

Read the paper → https://arxiv.org/abs/1706.03762

2020

Scaling Laws for Neural Language Models

Kaplan et al. · OpenAI

Performance improves predictably with data, compute and parameters — turning AI progress into capital allocation.

Read the paper → https://arxiv.org/abs/2001.08361

2020

Language Models are Few-Shot Learners

Brown et al. · OpenAI

GPT-3 showed a single large model could perform new tasks from a few examples in the prompt.

Read the paper → https://arxiv.org/abs/2005.14165

2022

Training Language Models to Follow Instructions with Human Feedback

Ouyang et al. · OpenAI

InstructGPT and RLHF: the post-training step that turns a text-continuer into an assistant — and shapes its character.

Read the paper → https://arxiv.org/abs/2203.02155

2022

Constitutional AI: Harmlessness from AI Feedback

Bai et al. · Anthropic

Aligns models to an explicit set of written principles, using AI feedback to supervise AI behavior.

Read the paper → https://arxiv.org/abs/2212.08073

2023

Navigating the Jagged Technological Frontier

Dell'Acqua et al. · Harvard Business School, BCG

758 BCG consultants: large gains inside AI's frontier, worse results just outside it.

Read the paper → https://www.hbs.edu/faculty/Pages/item.aspx?num=64700

2023

Experimental Evidence on the Productivity Effects of Generative AI

Noy & Zhang · Science

Professional writing tasks were done 40% faster and rated higher with ChatGPT.

Read the paper → https://doi.org/10.1126/science.adh2586

2023

GPTs are GPTs: Labor Market Impact Potential of LLMs

Eloundou, Manning, Mishkin, Rock

About 80% of the US workforce could see at least 10% of their tasks affected by LLMs.

Read the paper → https://arxiv.org/abs/2303.10130

2023

Can Generalist Foundation Models Outcompete Special-Purpose Tuning?

Nori et al. · Microsoft (Medprompt)

Prompting alone took general GPT-4 past 90% on MedQA — exhaust prompting before fine-tuning.

Read the paper → https://arxiv.org/abs/2311.16452

2023

Large Language Models as Simulated Economic Agents

Horton, Filippas, Manning · NBER

Homo silicus: using LLMs as stand-in respondents for research, policy and negotiation rehearsal.

Read the paper → https://www.nber.org/papers/w31122

2025

Measuring AI Ability to Complete Long Tasks

Kwa et al. · METR

The time-horizon series: how long a task an agent completes at 50% reliability, doubling every ~7 months.

Read the paper → https://arxiv.org/abs/2503.14499

Sources

01 McKinsey, "The state of AI in 2026: On the road to ROI" (Aug 2026)

02 Menache, Pathuri, Simchi-Levi, Linton, "How Generative AI Improves Supply Chain Management", HBR (Jan–Feb 2025); OptiGuide paper (2023)

03 Gosline et al., "Nudge Users to Catch Generative AI Errors", MIT SMR (2024)

04 Bloomberg, Kodak Chapter 11 (Jan 2012); NPR, Facebook–Instagram (Apr 2012)

05 Anthropic, AI-orchestrated espionage campaign (Nov 2025); Google Threat Intelligence Group (May 2026); Recorded Future H1 2026

06 OWASP, Top 10 for LLM Applications (2023 edition and 2025 edition)

07 Baker et al., OpenAI, "Monitoring Reasoning Models for Misbehavior" (Mar 2025); Simon Willison, "The lethal trifecta" (Jun 2025)

08 Google DeepMind, "AI Agent Traps" (2026); Cato Networks (HashJack, Immersive World); NewsGuard (Pravda network); Aim Security (EchoLeak)

09 Fortune, Replit database deletion (Jul 2025); Gartner on AI browsers, via The Register (Dec 2025)

10 Salem, Desai, Gupta, "Don't let Ricci v. DeStefano hold you back", FAccT 2022; Gupta et al., "Too Many Fairness Metrics"; MIT Sloan AI Expert Spotlight; Kleinberg, Mullainathan, Raghavan (2016)

11 Cui et al., Management Science (2026); Dell'Acqua et al., Organization Science (2026); "The Cybernetic Teammate", Organization Science; Bick, Blandin, Deming, Management Science (2026)

12 Narayanan and Kapoor, "AI as Normal Technology" (2025)

13 Duncan, Anderson, Saviano, HBR (Sep 2025); Source Global Research (Feb 2026)

14 Eric So, The Collision: What AI Does to Us (W. W. Norton, October 2026); MIT IDE book launch

15 Noy and Zhang, Science (2023)

16 Bratsberg and Rogeberg, PNAS (2018); Dworak, Revelle and Condon, Intelligence (2023)

17 Anthropic Education Report (Apr 2025)

18 Strömberg, Lei and Wu, "The Generative AI Learning Penalty", CEPR DP21577 (Jun 2026); Bastani et al., PNAS (2025)

19 Lee et al., CHI 2025; Kosmyna et al., "Your Brain on ChatGPT" (preprint, 2025)

20 Brynjolfsson, Chandar and Chen, "Canaries in the Coal Mine?" (Aug 2025); August 2026 update

21 Autor et al., "Does AI Assistance Enhance or Erode Expertise?", NBER WP 35720 (Sep 2026)

22 Bloom, "The 2 Sigma Problem", Educational Researcher (1984); Polanyi, The Tacit Dimension (1966)

23 Google Quantum AI, Willow, Nature (2024); Quantinuum Helios (Nov 2025); Bluvstein et al., Harvard/QuEra (Nature, 2025)

24 Forrester, practical quantum computing and Q-Day by 2030 (2026); DARPA Quantum Benchmarking Initiative

25 McKinsey Quantum Technology Monitor 2026 (Apr 2026)

26 Executive Order 14412 (22 Jun 2026); NIST PQC standards (Aug 2024); NIST IR 8547 (draft); UK NCSC timelines; Google; Microsoft; Expansión on Mexico's council

27 Mosca, "Cybersecurity in an era with quantum computers: will we be ready?" (2015)

28 D. A. Motta, Anthesis (2019): Disruptive Innovation and Institutional Framework

29 D. A. Motta & M. Pontuschka, "As organizações do futuro" (2023)

30 MIT Sloan Teaching & Learning Technologies, AI hub

31 Vaswani et al., "Attention Is All You Need" (2017); Uszkoreit, Google Research blog

32 Rich Sutton, "The Bitter Lesson" (2019)

33 Li et al., GPT-4 vs BloombergGPT (EMNLP 2023); Nori et al., Medprompt (2023)

34 Andrej Karpathy, "Power to the People" (2025)

35 Moffatt v. Air Canada, 2024 BCCRT 149

36 Horton, Filippas & Manning, NBER WP 31122; Manning & Horton, NBER WP 34937

37 America's AI Action Plan (July 2025)

38 Xi Jinping, WAIC keynote (July 2026)

39 Christopher Smart, "The Most Important AI Race" (Aug 2026)

40 Bill Gates, "The turbulent AI era is here" (Aug 2026)

41 "Pacing the Frontier" statement (July 2026)

42 Genewein, Legg et al., "From AGI to ASI", Google DeepMind (June 2026)

43 Ethan Mollick, Co-Intelligence (2024)

44 D. A. Motta, proprietary frameworks (House of Brains, Sept 2026): Organizational Dynamics; technological-paradigm and social-ethics transitions; megatrends of structural change; Creation and Destruction of Value; Digital Competitive Arena; Digital Drivers of Organizational Culture; Value Chain for Enterprise Modelling

45 Motta, Anthesis: uma jornada sobre a essência (Galáxia, 2019)

46 Federal Reserve Bank of St. Louis, "Tracking AI's contribution to GDP growth" (Jan 2026); Federal Reserve Board FEDS Note (Jul 2026); ING THINK

47 Gartner AI spending forecast (Sep 2026); McKinsey, "The cost of compute" (2025)

48 Stanford AI Index 2026, ch. 4

49 IEA, Energy and AI; BloombergNEF via Bloomberg (Jul 2026)

50 The Economist (Sep 2025); CNBC, Fortune, Reuters, BBC for Altman, Bezos, Georgieva, Pichai quotes

51 MIT News, Generative AI Impact Consortium (Feb 2025); MIT Everest Lab (G5)

52 Gartner, agentic AI forecast (Jun 2025); McKinsey, agentic AI at scale (Apr 2026)

53 Anthropic Claude pricing documentation (checked 24 Sep 2026)

54 Jacob Judah, MIT Technology Review (Sep 2025)

55 David Rotman, "What's at stake in AI's trillion-dollar gamble", MIT Technology Review (Sep 2026)

56 Kissinger, Schmidt & Huttenlocher, "The Metamorphosis", The Atlantic (2019); The Age of AI (2021)

57 Brynjolfsson, Li, Raymond, "Generative AI at Work", QJE (2025)

58 Ethan Mollick, "Detecting the Secret Cyborgs" (2023)

59 Castro, Gao, Martin, "Does GenAI Impose a Creativity Tax?", MIT SMR

60 Terwiesch, Meincke, Girotra, Mollick, Nave, Ulrich, Production and Operations Management (2026)

61 Sun, Li, Foo, Zhou et al., Journal of Applied Psychology (2025); Rice News

62 Goldman Sachs, "Global AI Investment Is Forecast to Exceed $1 Trillion in 2026" (Briggs, 7 Aug 2026)

63 J.P. Morgan Private Bank, "AI use is exploding. So are the bills." (Jul 2026)

64 J.P. Morgan Asset Management, "How is AI being monetized?" (Jan 2026); Morgan Stanley, AI financing gap (Aug 2025); Man Group, "The AI Bubble" (2025)

65 Torsten Slok, Apollo, "The AI Capex Boom Is Building Twice as Fast as the Housing Boom" (6 Aug 2026)

66 Epoch AI, open vs closed ECI gap (2026); Epoch AI, US vs China (Jan 2026)

67 Nvidia Q4 FY22 and Q4 FY26 releases; Nvidia Q2 FY27; Platformonomics on 2025 capex

68 Motta, "Big Tech and the Global Top 50: Financial Analysis, 2021–2026" (Sep 2026) and the Big Tech Five-Year Index data room (AI Finance Battle, Part III)

69 Steven Eppinger, MIT Sloan presentation (2026) of six functional productivity figures (HubSpot, SHRM, Deloitte, Choi and Xie, Cui et al.)

The source canon

MONITORING BASE · 73 SOURCES IN 7 GROUPS

This edition was checked against a wider canon of research and reporting, listed below by type. Items cited in the text appear in the Sources list above; the canon is the base I monitor to keep the argument current.

INFORMATION SYSTEMS RESEARCH

MIS Quarterly
Information Systems Research
Journal of Management Information Systems
Journal of the AIS
European Journal of Information Systems
Information Systems Journal
Journal of Strategic Information Systems
Journal of Information Technology

MANAGEMENT, STRATEGY AND INNOVATION RESEARCH

Management Science
Strategic Management Journal
Organization Science
Academy of Management Journal
Research Policy
Journal of Product Innovation Management
Technovation
Technological Forecasting and Social Change
IEEE Transactions on Engineering Management

MANAGEMENT REVIEWS

MIT Sloan Management Review
Harvard Business Review
California Management Review
Knowledge at Wharton
McKinsey Quarterly
strategy+business

BUSINESS AND TECHNOLOGY PRESS

MIT Technology Review
The Economist
Financial Times
IEEE Spectrum
Bloomberg Businessweek
Fortune
Wired
Fast Company
Forbes
The Verge
TechCrunch

INDEPENDENT ANALYSIS AND NEWSLETTERS

Stratechery
Simon Willison
Import AI (Jack Clark)
Interconnects
Lilian Weng
Chip Huyen
Latent Space
SemiAnalysis
Epoch AI
Benedict Evans
Exponential View
One Useful Thing
Understanding AI
The Pragmatic Engineer
AI Snake Oil
Gary Marcus
Dwarkesh Podcast
Platformer
a16z

INDUSTRY ANALYSTS AND ADVISORY RESEARCH

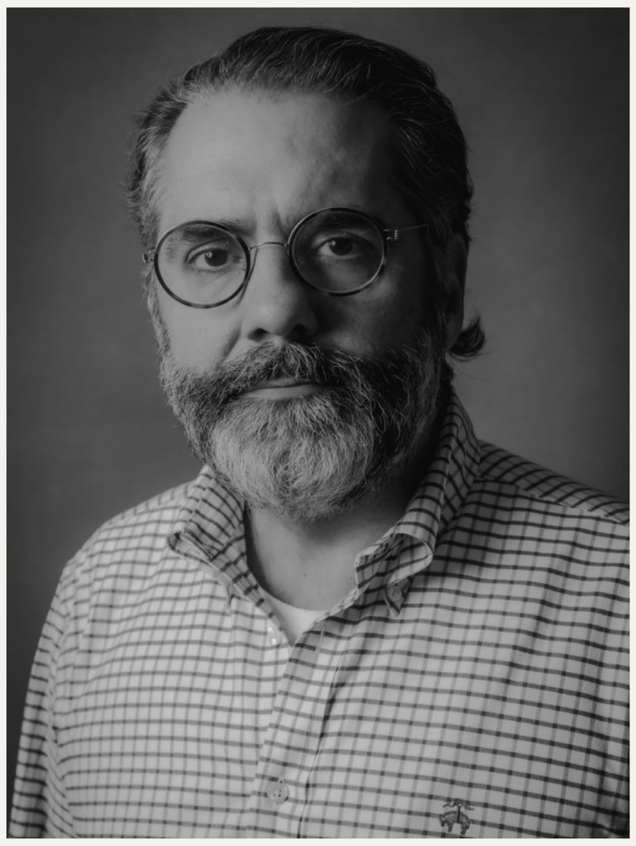
Gartner
Forrester
IDC
Constellation Research
Futurum
Omdia
HFS Research
Everest Group
ISG
Source Global Research
McKinsey QuantumBlack
Deloitte Insights
Accenture Technology Insights

RESEARCH INSTITUTES AND INVESTOR REPORTS

Stanford HAI (AI Index)
MIT CISR
State of AI Report
Bond
Menlo Ventures
CB Insights
Institute for the Future

Daniel Augusto Motta, PhD

Managing Partner HOB



Daniel Augusto Motta is Managing Partner of House of Brains and Founding Partner of BMI Blue Management Institute. He is also Managing Partner of WhiteFox, a corporate venture capital firm based in San Francisco, and a member of Afferolab's Partners Committee. He holds a PhD in Economics from the University of São Paulo (USP), a Master's in Economics from FGV-EAESP and a Bachelor's in Economics from USP. He is an alumnus of Harvard Business School's Owner/President Management (OPM) program and completed executive specializations in management at MIT Sloan School of Management, Singularity University and the Center for Creative Leadership. He teaches Corporate Strategy at IBGC and has taught in the MBA programs of Fundação Dom Cabral, Insper, FGV, ESPM and PUC-SP. He is a member of the Deliberative Council of MASP, served as Director of Strategic Planning at UNIBES, and was a founding member of the Brazilian Finance Society (Sociedade Brasileira de Finanças). His articles have appeared in Valor Econômico, EXAME and Folha de S.Paulo, and he has published three articles in Harvard Business Review Brasil. He is the author of the books A Liderança Essencial, Anthesis, Data Insights and Coleções de Ideias.

House of Brains

houseofbrains.com.br

bmi. BLUE MANAGEMENT INSTITUTE

AfferoLab

Intelligence is becoming abundant. Judgment is not.

House of Brains perspectives on AI and the enterprise. Nine theses, twenty figures and seventeen chapters on where AI value is created, where it leaks away, who is paying for the race, and what it means for boards, leaders and the firms paid to advise them. Every figure carries its source.

IN THIS EDITION

(01) The argument in brief

(02) AI Highlights

(03) AI Stats

(04) AI Deep Dive

(05) AI Finance Battle

(06) AI Quotes

(07) AI Articles

(08) AI Canon

HOUSE OF BRAINS

HOUSEOFBRAINS.COM.BR

Daniel Augusto Motta, PhD, Managing Partner HOB