PRESENTED BY PACE CAPITAL ↗
AI PEREZ

A WORKING THEORY OF AI VALUE ACCRUAL OVER TIME

AI GETS CHEAPER. WHAT STAYS *VALUABLE?*

As capable models get cheaper,[1] pricing power shifts toward companies that run agents reliably, connect them to customer data, and own the trusted interface where work gets done. Leading model labs won't automatically control those businesses.

THE THESIS

Pricing power follows what customers need and can't easily replace. In AI, our hypothesis is that the main source of that power shifts from **intelligence** to **deployment**, then to **context** and ultimately **distribution**. These advantages coexist. Later ones gain bargaining power only as earlier inputs become more commoditized. Adding compute can be faster than getting permission to use customer data or earning trust to act on it.

THE IDEA BEHIND THE CURVE

AI can spread while its suppliers lose pricing power. When customers can switch to an adequate alternative, the premium on a scarce input comes under pressure. That can happen even as the frontier advances.

Even if frontier development slows, compute demand may keep rising as more work goes into validating outputs, running evals and auditing agent actions. That work may favor different chips and systems from those used for frontier training. Coordinating multiple models and harnesses can also squeeze more useful work from existing models by choosing the right setup for each task. Lower costs per task can coexist with higher total compute use.

Carlota Perez's framework ↗ separates infrastructure build-out from widespread use. The curve tracks adoption, not profits. Applied to AI, it suggests that pricing power can move from models and compute to customer context and distribution, with the same company competing across several layers.

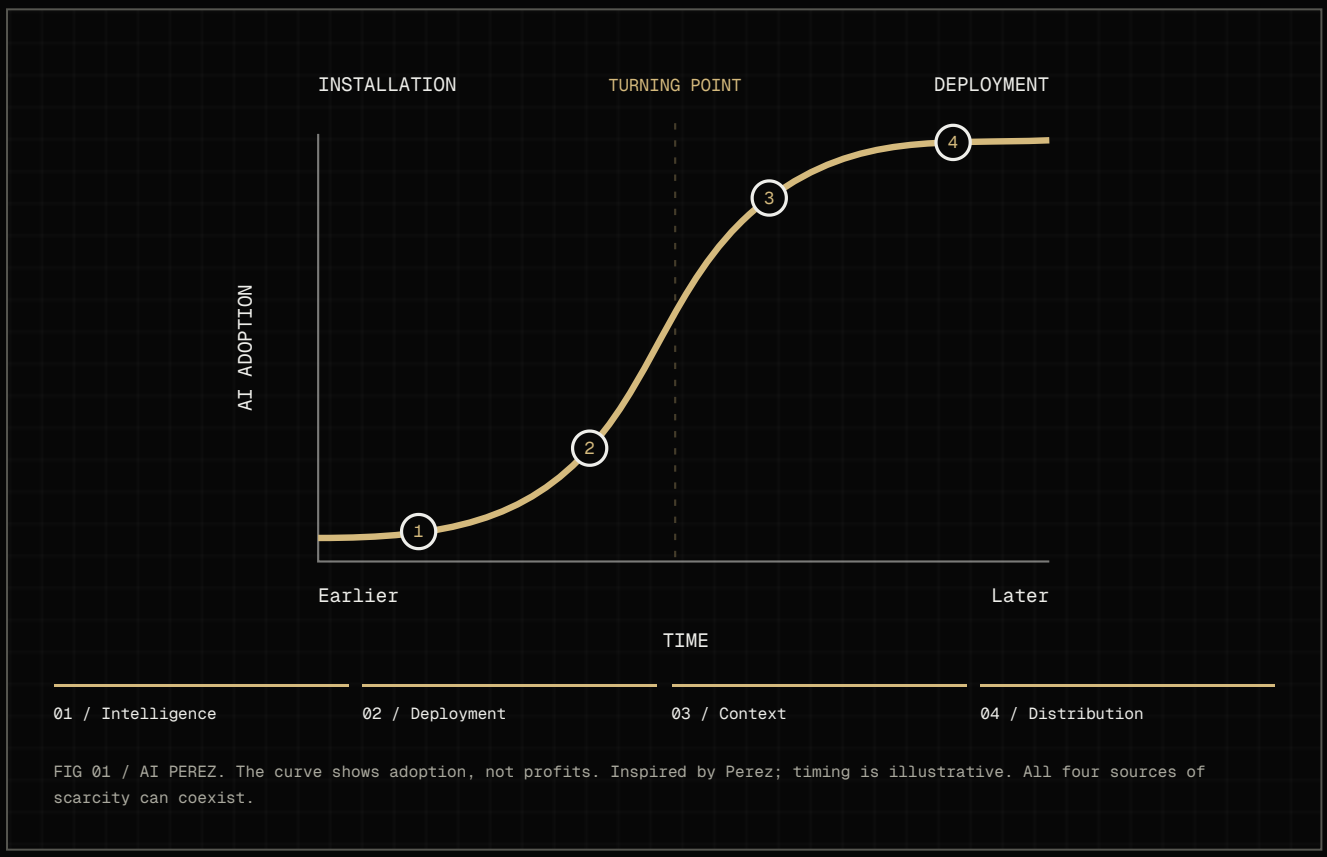


FIG 01 / AI PEREZ. The curve shows adoption, not profits. Inspired by Perez; timing is illustrative. All four sources of scarcity can coexist.

SOURCE OF PRICING POWER

CAPABILITY SETS THE PRICE

INTELLIGENCE SCARCITY

At the beginning of the cycle, only a handful of companies could produce the strongest models. That gap in capability, and the hardware required to sustain it, supported exceptional pricing power.

HOW AI IS BUILT AND USED — INTELLIGENCE

TYPICAL SETUP

CLOSED-CLOUD FRONTIER; CENTRALIZED BY NECESSITY

The model is the product. Frontier capability is reached through a proprietary cloud API, while open and local models remain useful but materially behind on the tasks that command the highest willingness to pay.

WHERE AI RUNS

Large GPU clusters, HBM, interconnect and power; consumer hardware is largely an endpoint, not an inference substrate.

MODEL DEPLOYMENT — ROLE IN THIS SOURCE OF PRICING POWER

CLOSED CLOUD — DOMINANT

The only practical route to frontier capability. The lab owns the weights, the serving stack and the meter.

OPEN CLOUD — EMERGING

Open weights broaden access, but most serious workloads still require rented cloud hardware and lag the frontier.

HYBRID — SPECIALIST

Some data remains private while difficult reasoning is sent to the cloud; integration is bespoke rather than a default architecture.

LOCAL / EDGE — CONSTRAINED

Small models run on customer hardware, but memory, thermals and capability limit the economically important use cases.

HARNESS / AGENT LAYER

THIN WRAPPERS AROUND SCARCE MODELS

Market shape: Thousands of applications; very few own the underlying capability

Where the power law sits: The power law sits below the harness: a handful of frontier labs and compute suppliers collect most of the leverage.

WHO HOLDS PRICING POWER?

Many wrappers; power stays with model and chip suppliers.

EACH BAR = A PROVIDER · TALLER = MORE POWER

App / harness providers

Leading providers — Others

Illustrative market structure · not measured data

“Harness” means the layer that turns a model into a usable agent: interface, instructions, tools, memory and policy.

Few companies had the research expertise or hardware to train frontier models. Training depended on scarce accelerators, memory, interconnect, power and engineering talent. Model labs charged for access to a capability few others possessed; NVIDIA, TSMC and the hyperscalers charged for the means of producing and serving it.

Those two forms of scarcity need not clear together. The premium attached to model quality can contract while power, memory and data-center capacity remain constrained.[2] Intelligence may become substitutable before the infrastructure built to provide it has finished depreciating. Treating all of this as ‘compute’ obscures the most important timing difference in the cycle.

Frontier labs have so far captured most enterprise model spending. ↗ Coding shows why: better models can take on more of a large, expensive labor pool’s work. Our hypothesis is that the premium for each new capability gain grows more slowly as cheaper models catch up. Demand for the frontier can still grow wherever better performance makes new work possible, including in defense, cybersecurity and scientific research.

Open models can put downward pressure on prices wherever their capability is adequate. A cheaper alternative makes a closed lab’s premium harder to sustain in that use case, but frontier advantages can reopen.[1] Distillation can accelerate the transfer of capability. The commercial territory reserved for frontier access may therefore shrink even while the frontier itself advances.

SCARCE		CONTROLLED BY		VALUE ACCRUES TO
Frontier models, chips, interconnect, power	→	**Labs + compute substrate**	→	**Frontier access and the physical capacity behind it**

> THE CONSEQUENCE
>
> For a closed lab, the relevant measure is not absolute capability but the distance from the cheapest adequate alternative. If that distance cannot be maintained, value has to be recovered elsewhere: by selling completed work, entering end markets or controlling distribution. RSI is the theoretical exception, but only if capability gains are economically productive faster than they are copied.

WHERE PRICING POWER MAY SETTLE

FRONTIER MODEL LABS

OpenAI · Anthropic · Google DeepMind · xAI · Meta · Mistral · DeepSeek · Cohere

WHY IT MATTERS

Frontier training remains concentrated among organizations with exceptional research talent, capital and cluster access.

HOW VALUE ACCRUES

They can charge for capabilities that customers cannot reproduce or obtain from a cheaper model. The rent is the performance gap—not intelligence in the abstract.

WHAT COMPRESSES THE RENT

Open models, distillation and diminishing willingness to pay narrow that gap use case by use case.

AI ACCELERATORS + SYSTEMS

NVIDIA · AMD · Google TPU · AWS Trainium · Intel Gaudi · Cerebras · SambaNova

WHY IT MATTERS

Accelerator supply and integrated systems remain difficult to add quickly, even as model access becomes less scarce.

HOW VALUE ACCRUES

Training and high-volume inference are bounded by usable compute, not chip counts alone. Vendors that combine silicon, networking and software capture more of the system economics.

WHAT COMPRESSES THE RENT

Custom silicon, better utilization and a shift from training to cheaper inference weaken the premium on general-purpose frontier hardware.

FOUNDRY, PACKAGING + MEMORY

TSMC · Samsung Foundry · SK hynix · Micron · Samsung Memory · ASE Technology

WHY IT MATTERS

Leading-edge fabrication, advanced packaging and high-bandwidth memory expand on industrial rather than software timelines.

HOW VALUE ACCRUES

These suppliers control capacity that cannot be replicated with code or purchased at short notice. Their bottlenecks can persist after model quality begins to converge.

WHAT COMPRESSES THE RENT

Committed capacity eventually arrives; when it does, volume may remain high while scarcity margins fall.

NETWORKING + INTERCONNECT

Broadcom · Arista Networks · Marvell · NVIDIA Networking · Astera Labs · Cisco

WHY IT MATTERS

Ever-larger clusters require more bandwidth within and between racks, making network architecture part of model economics.

HOW VALUE ACCRUES

A costly accelerator is useful only when data can reach it. Suppliers capture value by removing communication bottlenecks that otherwise leave scarce chips idle.

WHAT COMPRESSES THE RENT

Standards, merchant silicon and slower growth in frontier cluster size can transfer value back to buyers.

POWER, COOLING + ELECTRICAL SYSTEMS

Vertiv · Eaton · Schneider Electric · GE Vernova · Siemens Energy · Caterpillar

WHY IT MATTERS

Grid connections, generation, cooling and power-management equipment have lead times measured in years.

HOW VALUE ACCRUES

The constraint is increasingly energized capacity rather than server procurement. Suppliers sell into a build-out that cannot proceed without them and is difficult to accelerate.

WHAT COMPRESSES THE RENT

Their order books can outlast peak scarcity, but returns normalize when data-center construction slows or designs become more efficient.

HYPERSCALE + SPECIALIST CAPACITY

AWS · Microsoft Azure · Google Cloud · Oracle Cloud · CoreWeave · Crusoe · Nebius

WHY IT MATTERS

Few operators can finance, provision and run large clusters across regions with enterprise-grade availability.

HOW VALUE ACCRUES

They monetize balance sheet, procurement access and operating expertise by turning scarce physical infrastructure into capacity customers can consume immediately.

WHAT COMPRESSES THE RENT

Large fixed costs make this category vulnerable once capacity is interchangeable and utilization becomes the central problem.

SIGNALS OF TRANSITION

Compare capability gains with changes in willingness to pay, rather than benchmarks alone. Further signs of transition include labs selling outcomes instead of tokens, sovereign buyers supporting a high-cost frontier, and chipmakers backing open ecosystems to protect demand for their hardware.

SOURCE OF PRICING POWER

EXECUTION BECOMES THE PRODUCT

DEPLOYMENT SCARCITY

As models become more available, the practical constraint shifts from producing intelligence to delivering it reliably, at the required speed and cost.

HOW AI IS BUILT AND USED — DEPLOYMENT

TYPICAL SETUP

MULTI-MODEL CLOUD; ROUTING BECOMES THE ARCHITECTURE

Models multiply faster than production systems can absorb them. Closed and open models share the cloud; the scarce product is a dependable execution layer that chooses where a request runs and keeps it running.

WHERE AI RUNS

Hyperscale and specialist inference fleets dominate, while open models begin to run on developer workstations and private servers.

MODEL DEPLOYMENT — ROLE IN THIS SOURCE OF PRICING POWER

CLOSED CLOUD — STRONG

Frontier APIs remain the easiest route to difficult tasks, but applications can switch among several credible suppliers.

OPEN CLOUD — RAPID GROWTH

Open models are served by competing inference clouds, turning weights into a portable workload rather than a proprietary destination.

HYBRID — FORMING

Routers send routine work to cheaper or private models and escalate harder requests to frontier clouds.

LOCAL / EDGE — PRACTICAL

Quantization and better runtimes make local inference credible for development, privacy-sensitive work and bounded tasks.

HARNESS / AGENT LAYER

FRAMEWORK PROLIFERATION; PLATFORM CONSOLIDATION

Market shape: Hundreds of frameworks and libraries; a much smaller set of production runtimes, gateways and sandboxes

Where the power law sits: Moderate and unsettled. Order flow can concentrate in a few gateways, but open standards and low switching costs keep the contest open.

WHO HOLDS PRICING POWER?

A few gateways may pull ahead; the market is still open.

EACH BAR = A PROVIDER · TALLER = MORE POWER

Gateways / agent runtimes

Leading providers — Others

Illustrative market structure · not measured data

“Harness” means the layer that turns a model into a usable agent: interface, instructions, tools, memory and policy.

The center of activity moves from training to inference. Customers must choose among models, chips and clouds while managing latency, availability and cost.[3] Inference providers, gateways and agent runtimes turn that complexity into a service. Developer distribution matters because the default route can capture order flow even when the underlying suppliers change.

Deployment scarcity may erode especially quickly. Routing makes inference services more commoditized; optimization extracts more work from each chip; open software lowers switching costs. The economics still depend on how much of the rented capacity is put to use.[4] The firms solving deployment scarcity are also helping to eliminate it. That does not make them poor businesses, but it does place a burden on them to build workflow, distribution or scale advantages before basic inference becomes a utility.

The physical layer presents a separate risk. Data centers are financed and built in parallel, with each operator planning against similar demand. Once the capital is committed, the incentive is to bid aggressively for utilization. The result has been familiar across railways, canals and fiber: a useful network, followed by disappointing returns for a portion of those who financed it.

The risk is that economic returns deteriorate before GPU assets finish depreciating. Contracted revenue can cushion that adjustment, but renewal prices and uncontracted capacity remain exposed.[5] If inference efficiency continues to improve while new capacity arrives, a shift in bargaining power from deployment to context could combine falling unit prices with an overbuilt asset base. Consolidation would then be a consequence of the build-out, not evidence that demand had disappeared.

A SCENARIO

What if frontier development slows?

Suppose labs agree to slow frontier development while deployment and optimization continue. Labs could spend less on the race, and existing models and hardware might remain commercially useful longer. Suppliers could face weaker growth in training demand, with financing commitments still due. Contracts and inference usage would help determine which capacity stays profitable. Incumbents might retain their advantages longer; productivity gains requiring new capabilities could be delayed.

Without an agreement, power bottlenecks, financing limits and weak returns can still restrain expansion. Competition and efficiency can reduce costs per task as the frontier advances. Lower prices may encourage enough use for total compute spending to grow.[6] A shift toward context or distribution still requires adequate alternatives and customers able to switch.

SCARCE		CONTROLLED BY		VALUE ACCRUES TO
Serving capacity, routing, latency, reliable execution	→	**Inference + orchestration layers**	→	**Reliable throughput across a fragmented supply base**

THE CONSEQUENCE

The deployment layer can produce important companies while still becoming economically harsher. The strongest positions will use today’s fragmentation to acquire durable developer relationships, proprietary workflow or operating scale. Pure capacity is more exposed: high fixed costs and falling prices are an unforgiving combination.

WHERE PRICING POWER MAY SETTLE

INFERENCE + MODEL CLOUDS

Together AI · Fireworks AI · Baseten · Modal · Groq · Cerebras · DeepInfra · Replicate · Cloudflare Workers AI

WHY IT MATTERS

Supply is fragmented across chips, clouds and model architectures while production demand is rising faster than internal serving expertise.

HOW VALUE ACCRUES

They capture the spread between raw accelerator time and reliable, model-specific throughput. Optimized kernels, batching and capacity management let them deliver more useful tokens per dollar.

WHAT COMPRESSES THE RENT

That spread compresses as optimization diffuses, capacity grows and customers treat inference providers as interchangeable.

ROUTING + MODEL GATEWAYS

OpenRouter · Portkey · LiteLLM · Cloudflare AI Gateway · Kong · Helicone · Martian

WHY IT MATTERS

Rapid model releases create meaningful differences in price, latency, capability and availability from one request to the next.

HOW VALUE ACCRUES

The gateway sees the order flow and can choose the supplier. That position compounds through usage data, policy controls, failover and one integration across many models.

WHAT COMPRESSES THE RENT

Routing succeeds by commoditizing what it routes. Clouds and application platforms can absorb the feature once the decision logic becomes standard.

OPEN + LOCAL INFERENCE RUNTIMES

Ollama · vLLM · llama.cpp · SGLang · LM Studio · Open WebUI · BentoML

WHY IT MATTERS

Capable open models and smaller hardware footprints make private and on-device deployment practical for a growing set of workloads.

HOW VALUE ACCRUES

These projects set the deployment standard and reduce dependence on proprietary clouds. Commercial value can accrue through managed hosting, enterprise support and control of the local model workflow.

WHAT COMPRESSES THE RENT

Adoption alone does not create a business; agents also reduce the labor premium customers once paid to avoid self-deployment.

MODEL + DEVELOPER DISTRIBUTION

Hugging Face · GitHub Models · AWS Bedrock · Vertex AI Model Garden · Azure AI Foundry · Replicate

WHY IT MATTERS

Developers need discovery, testing, governance and deployment across an unusually fast-changing supplier base.

HOW VALUE ACCRUES

Platforms capture value by becoming the place a model is first found, evaluated and integrated. Distribution can persist after any individual model loses its premium.

WHAT COMPRESSES THE RENT

The largest clouds can bundle the catalog, while open standards keep migration costs lower than in traditional platform markets.

AGENT EXECUTION + SANDBOXES

Daytona · E2B · Temporal · Modal · Browserbase · Fly.io · Cloudflare

WHY IT MATTERS

Agents are moving from single responses to long-running jobs that use browsers, code, files and external systems.

HOW VALUE ACCRUES

They sell the controlled environment in which unreliable model output becomes resumable, observable work. Security isolation and state management matter before agents can receive broader permissions.

WHAT COMPRESSES THE RENT

Basic sandboxing and orchestration may consolidate into clouds, operating systems or agent frameworks.

INFERENCE OPTIMIZATION

NVIDIA TensorRT-LLM · Red Hat vLLM · Modular · SambaNova · CentML · BentoML

WHY IT MATTERS

Unit economics depend heavily on compiler choices, quantization, caching and hardware-aware scheduling.

HOW VALUE ACCRUES

Small efficiency gains multiply across enormous token volumes. Optimizers capture value while expertise is scarce and serving stacks remain heterogeneous.

WHAT COMPRESSES THE RENT

The economics are self-eroding: successful techniques enter open runtimes, chip libraries and cloud platforms.

SIGNALS OF TRANSITION

The hand-off is under way when inference capacity is marketed as interchangeable, routing is absorbed into larger platforms, management teams emphasize utilization rather than expansion, and distressed assets begin to consolidate.

SOURCE OF PRICING POWER

CONTEXT BECOMES CAPABILITY

CONTEXT SCARCITY

Once generic capability is widely available, performance depends increasingly on what a system knows about a particular organization, what it is permitted to see and what it is authorized to do.

HOW AI IS BUILT AND USED	CONTEXT
TYPICAL SETUP **HYBRID BY POLICY; MODELS ARE SELECTED INSIDE THE WORKFLOW**	No single model topology wins. Workload placement follows context: sensitivity, latency, policy and the value of the task. Hybrid systems become normal because the enterprise boundary matters more than allegiance to one model.
WHERE AI RUNS	Public cloud, private cloud, on-premise servers and edge devices operate as one governed estate; locality becomes a permission decision.

MODEL DEPLOYMENT	ROLE IN THIS SOURCE OF PRICING POWER
CLOSED CLOUD	SELECTIVE
Used where frontier performance justifies cost and data policy permits an external service.	
OPEN CLOUD	ESTABLISHED
Open weights provide control and portability without requiring the customer to own every piece of serving infrastructure.	
HYBRID	DEFAULT
Policy-aware routing keeps sensitive context private, uses local or open models for routine work and buys frontier capability when needed.	
LOCAL / EDGE	STRATEGIC
Private inference supports continuous memory, low latency and regulated workloads; hardware ownership becomes part of the trust model.	

HARNESS / AGENT LAYER

DOMAIN-SPECIFIC AGENTS AROUND SYSTEMS OF RECORD

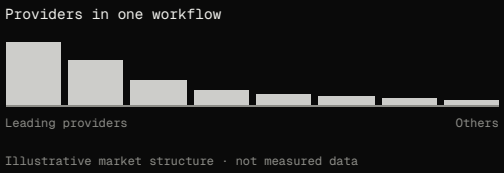
Market shape: Many enterprise and vertical harnesses; a small number can dominate each workflow or industry

Where the power law sits: The power law fragments by domain. The winners are attached to live data, identity and outcomes—not necessarily to the most popular base model.

WHO HOLDS PRICING POWER?

Leaders emerge within each workflow, not across all agents.

EACH BAR = A PROVIDER · TALLER = MORE POWER



Illustrative market structure · not measured data

“Harness” means the layer that turns a model into a usable agent: interface, instructions, tools, memory and policy.

Context is more than data retrieval. It includes current state, institutional memory, identity, policy, secrets and the authority to take a consequential action. A model may know how to approve an invoice; it still needs to know which invoice, under whose policy and with whose liability.

This constraint clears at institutional speed. Security reviews, procurement, data migration and regulation do not improve at the rate of model inference. Context scarcity should therefore last longer than deployment scarcity and favor businesses already embedded in systems of record and operational workflows.

The adjective ‘proprietary’ is not enough to make a dataset defensible. Static information can be exported, reproduced or lose relevance. A stronger position is a live feedback loop: software participates in the work, observes the result, improves from it and earns broader permission over time. The scarce asset is not simply knowledge. It is authorized participation.

The next shift is conditional. If customers can carry their records and permissions into another service, a trusted interface may choose among competing workflows. If context cannot travel safely, the workflow remains hard to replace. Ownership of the customer relationship only becomes the stronger advantage when it brings a credible ability to redirect work.

This also explains why open-source software can support valuable companies. Enterprises do not pay only for access to code. They pay for integration, accountability and the transfer of operational risk. Agents may reduce the effort required to deploy software, but they do not eliminate the need for someone to be responsible when it fails.

SCARCE		CONTROLLED BY		VALUE ACCRUES TO
Current data, identity, permissions, evaluation	→	**Systems embedded in proprietary workflows**	→	**Access to live context and permission to act on it**

> THE CONSEQUENCE
>
> Cheaper intelligence need not make responsibility cheaper. The margin pool could shift toward systems that can establish identity, govern access, evaluate behavior and accept liability. The commercial model may resemble risk transfer as much as conventional software licensing.

WHERE PRICING POWER MAY SETTLE

SYSTEMS OF RECORD + WORKFLOW

Salesforce · ServiceNow · SAP · Oracle · Workday · Epic · Guidewire · Atlassian

WHY IT MATTERS

Enterprises can access capable models before they can safely connect them to the systems where consequential work is recorded.

HOW VALUE ACCRUES

These incumbents already hold current state, permissions and workflow. They can place agents inside an approved operating environment rather than asking customers to reconstruct one elsewhere.

WHAT COMPRESSES THE RENT

Possession of data is not enough if a new interface owns the user and treats the system of record as a replaceable back end.

DATA + RETRIEVAL PLATFORMS

Databricks · Snowflake · MongoDB · Confluent · Elastic · Pinecone · Weaviate · Redis

WHY IT MATTERS

Generic intelligence is abundant, but useful answers depend on fresh, permission-aware access to proprietary state.

HOW VALUE ACCRUES

They sit between models and changing enterprise data. Value comes from governed access, lineage and real-time retrieval—not from attaching the word ‘AI’ to stored data.

WHAT COMPRESSES THE RENT

Static datasets move easily; durable value requires staying inside the live read-and-write path of the workflow.

CONNECTORS + TOOL PROTOCOLS

Model Context Protocol · Zapier · Workato · Boomi · MuleSoft · Merge · Pipedream · Composio

WHY IT MATTERS

The immediate obstacle is not reasoning but the fragmented work required to connect agents to thousands of software tools and data sources.

HOW VALUE ACCRUES

Connection layers can become the authorized route through which agents discover and use enterprise tools. Breadth, maintenance and policy enforcement create the commercial advantage.

WHAT COMPRESSES THE RENT

Open protocols reduce lock-in, and systems of record may expose their own first-party agent interfaces.

IDENTITY, SECRETS + PERMISSION

Okta · Auth0 · Microsoft Entra · CyberArk · HashiCorp Vault · Infisical · Aembit

WHY IT MATTERS

Agents need identities distinct from their human sponsors and permissions bounded by task, time and resource.

HOW VALUE ACCRUES

Autonomy cannot expand without a control plane that can authenticate an agent, limit its authority and revoke it. Existing trust relationships and integrations are difficult to reproduce quickly.

WHAT COMPRESSES THE RENT

Identity suites can bundle agent controls; point solutions must own a genuinely new policy layer rather than a renamed service account.

EVALUATION + OBSERVABILITY

LangSmith · Braintrust · Arize · Langfuse · Galileo · Patronus AI · Humanloop

WHY IT MATTERS

Non-deterministic agents are reaching production before organizations have adequate evidence of how they behave.

HOW VALUE ACCRUES

These platforms turn traces and outcomes into tests, audit records and feedback. That evidence lets customers grant more autonomy without accepting unknowable risk.

WHAT COMPRESSES THE RENT

Tracing is easy to commoditize; durable value depends on proprietary evaluation data, workflow integration or becoming part of the approval process.

AGENT SECURITY + GOVERNANCE

Credo AI · HiddenLayer · Protect AI · Noma Security · Lakera · Robust Intelligence

WHY IT MATTERS

Prompt injection, tool misuse and data leakage become budgeted problems once agents can take actions rather than merely generate text.

HOW VALUE ACCRUES

They capture value by reducing the expected cost of autonomy and by producing the controls regulators, insurers and security teams require before deployment.

WHAT COMPRESSES THE RENT

Security platforms can absorb these capabilities, and fear alone will not sustain a category without measurable risk reduction.

VERTICAL WORKFLOW OWNERS

Harvey · Abridge · Glean · Sierra · Decagon · Hebbia · Cursor · Ambience Healthcare

WHY IT MATTERS

Narrow domains can support agents earlier because the workflow, vocabulary, data and acceptable outcome are more clearly defined.

HOW VALUE ACCRUES

They capture value by owning the completed task and the feedback it generates. Deep integration and domain liability matter more than access to any one model.

WHAT COMPRESSES THE RENT

A vertical product is vulnerable if customers can reproduce the workflow with a general assistant and a small number of connectors.

DATA + ENVIRONMENT SUPPLIERS

Scale AI · Turing · Mercor · Surge AI · Handshake · AfterQuery · Invisible Technologies

WHY IT MATTERS

Continuous learning requires specialized human feedback, task environments and outcome data after the easy public corpus has been consumed.

HOW VALUE ACCRUES

They organize scarce expertise and generate the environments in which models learn to perform economically useful work. The value lies in repeatable production, not a one-off dataset.

WHAT COMPRESSES THE RENT

Margins fall when data production is undifferentiated or when customers internalize the workflow; privileged supply and quality control must persist.

SIGNALS OF TRANSITION

Look for agent identity to separate from human identity, permissions to become specific to a task and duration, and evaluation to move from demonstrations to replayable evidence. For a shift toward distribution, watch whether customers can change service providers while retaining context, permissions and audit history. If those remain locked to one workflow, its bargaining power persists.

SOURCE OF PRICING POWER

FROM ATTENTION TO DELEGATION

DISTRIBUTION SCARCITY

Distribution matters from the start. It becomes more powerful when a trusted interface can choose among several capable suppliers on the customer's behalf. That shift requires alternatives and permission to act. It does not require free intelligence or local computing.

HOW AI IS BUILT AND USED	DISTRIBUTION
TYPICAL SETUP **CUSTOMER-LED SERVICE, FLEXIBLE COMPUTING**	A trusted interface receives the request and selects services behind it. It may use paid cloud models, open models or local inference. The defining feature is authority over supplier choice, not a particular computing architecture.
WHERE AI RUNS	Cloud, private servers or customer devices, depending on capability, cost and privacy. Local execution is optional.

MODEL DEPLOYMENT	ROLE IN THIS SOURCE OF PRICING POWER
CLOSED CLOUD	PREMIUM INPUT
Frontier models remain valuable for difficult work, but increasingly appear as an invisible supplier inside someone else's service.	
OPEN CLOUD	ALTERNATIVE SUPPLY
Open models can give the interface more supplier choice. They need not be free or replace the frontier for every task.	
HYBRID	POSSIBLE MIX
The agent decides whether work stays on the device, moves to a private service or purchases frontier reasoning. The user need not see the choice.	
LOCAL / EDGE	POSSIBLE ROUTE
Capable local models could support private memory and low-latency tasks. Distribution can remain valuable even if most work stays in the cloud.	

HARNESS / AGENT LAYER

A FEW TRUSTED AGENTS ABOVE A VAST MARKET OF SKILLS

Market shape: Single-digit daily agent relationships per user; potentially millions of downstream services, tools and specialist agents

Where the power law sits: A barbell power law: winner-take-most at the customer interface, extreme fragmentation among suppliers behind it.

WHO HOLDS PRICING POWER?

A few agents own demand; many suppliers compete behind them.

EACH BAR = A PROVIDER · TALLER = MORE POWER

Illustrative market structure · not measured data

"Harness" means the layer that turns a model into a usable agent: interface, instructions, tools, memory and policy.

A workflow owner knows how to complete a task inside its system. A customer interface receives the broader request and may decide which systems do the work. The balance changes when reliable connections, portable context and compatible permissions let that interface replace a supplier without forcing the customer to rebuild the process. The workflow is still necessary, but it has less control over the next purchase.

Three conditions make that hand-off plausible: several suppliers can perform the task well enough; the interface can access the context and authority needed to use them; and customers trust it to choose. Without those conditions, distribution may deliver leads while the workflow owner still sets the terms. A large audience alone does not transfer bargaining power.

Context and distribution often belong to the same company. An accounting platform could keep both the records and the assistant that receives the request. In that case, the advantage deepens inside the incumbent rather than migrating to a new firm. The test is who can replace whom: can the assistant switch workflows while keeping the customer, or can the workflow replace the assistant without losing demand?

Expensive cloud inference is compatible with this argument. An assistant can own the customer relationship while purchasing every model call. If it can switch model suppliers, it has negotiating power; if only one model can do the job, that supplier can still claim a large share of the economics. Distribution can gain power without eliminating intelligence scarcity.

Local-first computing is one possible route, not a necessary condition. Capable models on customer hardware could lower serving costs, keep more data private and support continuous use. Paid cloud models or a hybrid architecture could support the same customer relationship. Where inference runs affects cost and privacy; it does not by itself determine who owns demand.

Incumbents have formidable advantages here: installed distribution, capital, data and default placement. Their victory is not automatic. Rebuilding a product around a new interface is harder than adding AI features to an existing one; Microsoft owned GitHub yet did not monopolize coding agents. Platform shifts have a long record of making secure positions look less secure in retrospect.

Consumer demand is also likely to divide. Convenience products will accept extensive automation because the desired outcome is the service itself. Leisure, culture and status operate differently. Authorship and human participation remain part of what is being consumed. As synthetic production becomes abundant, verified human origin may command a premium precisely because it is scarce.

The likely result is a barbell: inexpensive, automated utility at one end and scarce human craft or prestige at the other. The middle—competent but undifferentiated production—faces the greatest pressure. Enterprise buyers are less sentimental, except when the reputation of the provider is itself part of the risk decision.

SCARCE		CONTROLLED BY		VALUE ACCRUES TO
Customer trust and authority to choose suppliers	→	**Default interfaces + direct relationships**	→	**Control of the customer relationship and the allocation of demand**

> THE CONSEQUENCE
>
> The customer relationship becomes more valuable when it carries the authority to choose and replace suppliers. Cheap inference can help, but substitution and trust do the economic work. Where context remains locked inside an indispensable workflow, that workflow may retain the stronger position.

WHERE PRICING POWER MAY SETTLE

OPERATING SYSTEMS + DEVICES

Apple · Google · Microsoft · Meta · Samsung · Amazon · emerging AI-native hardware

WHY IT MATTERS

Capability differences become less visible to mainstream users while assistants are embedded into phones, PCs, glasses and home devices.

HOW VALUE ACCRUES

The operating surface receives intent before a model is selected. Defaults, hardware integration and installed base let the platform route demand and make upstream suppliers invisible.

WHAT COMPRESSES THE RENT

A default is valuable only if people use it. Weak products can still cede the relationship to an application that users actively choose.

GENERAL-PURPOSE ASSISTANTS

ChatGPT · Claude · Gemini · Microsoft Copilot · Meta AI · Perplexity

WHY IT MATTERS

Assistants are building direct consumer habits before the underlying models become fully substitutable.

HOW VALUE ACCRUES

Repeated use creates memory, preference data, identity and trust. If users delegate decisions to the assistant, it becomes the buyer of downstream models and services rather than a reseller of answers.

WHAT COMPRESSES THE RENT

The relationship can be displaced by the operating system, browser or specialist product where the user's intent actually originates.

WORK SURFACES

Microsoft 365 + Teams · Google Workspace · Slack · Notion · Zoom · Atlassian · Canva

WHY IT MATTERS

Enterprise agents first appear inside products employees already open each day and that already carry organizational identity.

HOW VALUE ACCRUES

These products own both attention and workflow context. They can choose the model, surface the agent at the moment of intent and retain the feedback generated by completed work.

WHAT COMPRESSES THE RENT

They lose leverage if the agent becomes the primary interface and reduces the underlying application to a tool call.

SEARCH + BROWSERS

Google Search + Chrome · Safari · Microsoft Edge · Perplexity · Dia · Brave

WHY IT MATTERS

Research, navigation and commercial discovery are being compressed into conversational and agentic interfaces.

HOW VALUE ACCRUES

Browsers and search products observe high-intent queries at the point where users choose what to read, buy or do. That position can direct traffic, transactions and model selection.

WHAT COMPRESSES THE RENT

Answer interfaces can weaken the open-web supply that makes them useful, while operating systems can move discovery one layer higher.

COMMERCE + PAYMENT RAILS

Amazon · Shopify · Stripe · Visa · Mastercard · Instacart · DoorDash · Booking Holdings

WHY IT MATTERS

Delegation becomes economically consequential when agents can compare, purchase and settle on a user’s behalf.

HOW VALUE ACCRUES

These networks combine merchant access, identity, payment credentials and dispute handling. An agent needs all four to convert a recommendation into a trusted transaction.

WHAT COMPRESSES THE RENT

Standards may separate the agent relationship from fulfillment and payment, forcing each layer to compete on price.

VERTICAL CONSUMER RELATIONSHIPS

Abridge + Epic · Duolingo + Khan Academy · Intuit · Harvey · WhatsApp · travel and health incumbents

WHY IT MATTERS

Users delegate sooner in domains where a product already understands the task, history and acceptable boundary of action.

HOW VALUE ACCRUES

Domain trust and longitudinal context make the product a plausible representative rather than a generic interface. The strongest positions own a recurring decision, not a single query.

WHAT COMPRESSES THE RENT

The relationship is less durable where a general assistant can import the history and reproduce the service without regulatory or workflow friction.

LOCAL + PERSONAL AGENT ECOSYSTEMS

Ollama · LM Studio · Open WebUI · on-device Apple + Android models · personal-agent entrants

WHY IT MATTERS

Capable local models make it possible to keep private context and continuous memory on hardware the user controls.

HOW VALUE ACCRUES

If local models become capable and economical enough, applications could use private context without a metered cloud call for every task. A personal agent could combine local inference with paid cloud services. This is one possible architecture for owning the customer relationship, not a prerequisite.

WHAT COMPRESSES THE RENT

Convenience may outweigh sovereignty for most consumers, leaving the category influential technically but difficult to monetize directly.

AI-NATIVE SERVICES + SOFTWARE

Cursor · Harvey · Abridge · Sierra · Decagon · Replit · new consumer and small-business services

WHY IT MATTERS

Falling inference costs allow products to deliver completed work in markets that could not support the cost of a human service or repeated frontier-model calls.

HOW VALUE ACCRUES

These companies can turn labor into a repeatable software product: research completed, claims processed, customers served or code shipped. Cheap intelligence creates the supply; workflow ownership, brand and distribution capture the margin by packaging it into an outcome someone will pay for.

WHAT COMPRESSES THE RENT

Cheaper intelligence can lower the barrier for competitors too. Without proprietary workflow, trust, data or a low-cost route to customers, the service becomes as interchangeable as the model beneath it.

AUTHORED MEDIA + SCARCE ATTENTION

YouTube · TikTok · Spotify · Substack · Patreon · Live Nation · creator and luxury brands

WHY IT MATTERS

Synthetic production becomes abundant while human attention remains fixed and verified authorship becomes easier to value by contrast.

HOW VALUE ACCRUES

Platforms and brands capture value by aggregating scarce audiences, identity and cultural signal. Utility content can automate; participation, status and human provenance cannot be manufactured at the same marginal cost.

WHAT COMPRESSES THE RENT

The undifferentiated middle is exposed. Only products with genuine audience ownership, trusted curation or scarce human origin retain the premium.

SIGNALS OF TRANSITION

The transition becomes visible when model brands recede behind product brands, agents transact with other agents, and changes in defaults move demand more than changes in benchmarks. In consumer markets, watch automated utility separate from work whose value depends on human authorship.

THE INVESTMENT QUESTION

WHAT REMAINS SCARCE AFTER THE BUILD-OUT?

The shifts in bargaining power are conditional. When models become more commoditized, reliable execution can command a premium. When execution becomes easier to buy, authorized context can matter more. A trusted interface gains power over those workflows only if it can access their context, choose alternatives and keep the customer. These advantages may accumulate in one company, and the hand-off may never happen in some markets.

This is why the hand-off from deployment to context deserves particular attention. It may coincide with the largest financial risk in the cycle: inference prices falling, infrastructure still arriving and enterprise demand constrained by permissions rather than capacity. The assets will remain useful. Their owners may not all earn the returns they expected.

WHAT WOULD MAKE THIS WRONG?

01 Model quality keeps widening willingness to pay faster than open alternatives narrow it.

02 Context and permissions remain tied to indispensable workflows, so interfaces cannot redirect demand.

03 Agents never earn enough trust to receive meaningful delegated authority.

04 Incumbents absorb the interface shift without surrendering their old economics or their distribution.

05 Demand for intelligence proves effectively unbounded, keeping compute scarce despite the build-out.

Look beyond the best model. Ask **what customers pay extra for today, how long that advantage will last, and what will keep them from switching.**

FOOTNOTES

Sources and calculations behind the economic claims. Prices are dated snapshots in USD; worked examples use stated assumptions.

---

1. Cheaper at a given capability

Stanford reports that inference prices at GPT-3.5-level performance fell **over 280×** between November 2022 and October 2024. That measures the price of reaching a capability threshold, not the expense of every frontier workload. Its 2026 report also finds that the open-closed model performance gap reopened in 2025, with closed models ahead as of March 2026. Cheaper adequate intelligence can put pressure on prices even while a frontier premium survives; these observations do not establish an irreversible loss of pricing power.

Stanford HAI · AI Index 2025, takeaway 7 ↗ (opens in a new tab)

November 2022–October 2024

Stanford HAI · AI Index 2026, technical performance ↗ (opens in a new tab)

2025–March 2026

Verified 14 September 2026

---

2. Two different build-out clocks

The IEA's Electricity 2026 report puts planning, permitting and building new grid infrastructure at **5–15 years**, compared with **1–3 years** for a data center. This mismatch helps explain why physical constraints can outlast changes in model availability. The ranges describe infrastructure projects, not a universal wait for every grid connection. They support the timing distinction; they do not establish that all regions will face shortages, or that an eventual capacity glut is inevitable.

IEA · Electricity 2026, Grids ↗ (opens in a new tab)

2026 assessment of construction lead times

Verified 14 September 2026

---

3. What one model call costs

Anthropic lists Claude Sonnet 5 at **$2 per million input tokens** and **$10 per million output tokens** on its standard API. For an illustrative invoice attempt using 5,000 input and 1,000 output tokens:

> (5,000 × $2 + 1,000 × $10) / 1,000,000 = **$0.02 per attempt**
> **$20 per 1,000 attempts**

This is our calculation using assumed token counts, not an invoice benchmark or a cost per successful payment. It excludes caching and batch discounts, taxes, retries, tools and human review. The relevant service metric is total expense divided by successfully completed tasks; API prices do not reveal the model provider's production costs.

Anthropic · Claude API pricing ↗ (opens in a new tab)

Standard global API rates, 14 September 2026

Verified 14 September 2026

4. Idle time changes the unit cost

Lambda lists a single **H100 SXM, 80 GB, at $4.29/hour**. Suppose it remains rented continuously, including idle time. Treat utilization here as the share of rented hours spent productively serving work:

> $4.29 / 25% = **$17.16**
> $4.29 / 75% = **$5.72**
> Rental expense per productive GPU-hour

The utilization levels are illustrative assumptions, not measured fleet statistics. This calculation shows why filling rented capacity matters; it is not a measure of provider margins. It excludes taxes and additional operating costs. Translating GPU-hours into token costs would require measured throughput for a specific model, workload and latency target.

Lambda · GPU instances, 1× H100 SXM ↗ (opens in a new tab)

On-demand rental price, 14 September 2026

Verified 14 September 2026

5. Accounting life and economic life

CoreWeave’s 2025 annual filing assigns technology equipment a **six-year accounting life**. It also describes committed contracts as take-or-pay: customers owe payment regardless of utilization. Those commitments can protect contracted revenue from low customer usage. The risk is that achievable prices and returns deteriorate before assets finish depreciating, particularly on uncontracted capacity or at renewal. An accounting estimate does not establish when a GPU becomes uneconomic, and this filing does not prove that an industry-wide write-down is coming.

CoreWeave · 2025 Form 10-K, Note 2 ↗ (opens in a new tab)

Year ended 31 December 2025

Verified 14 September 2026

6. Efficiency, demand and a possible slowdown

The IEA’s 2026 assessment links AI energy demand to “improvements in efficiency, surging uptake, and changing model capabilities.” It also identifies power, chip and financing constraints on expansion. Cheaper tasks need not mean lower total demand; energy consumption does not establish industry revenue or profitability. The effects of a coordinated frontier slowdown described here are our scenario analysis, assuming deployment and optimization continue. The report does not model that agreement.

IEA · Key Questions on Energy and AI, executive summary ↗ (opens in a new tab)

2026 assessment of efficiency, demand and investment constraints

Verified 14 September 2026