AIMONGER
WHITE PAPER

AI Discovery

Durable Advantage When Models Are Commodities

When every serious competitor can rent comparable model capability, advantage migrates to assets that are hard to copy: proprietary feedback loops, customer retention economics, institutional judgement, and the operating system that turns intelligence into shipped work. Where many firms compete with thinner AI teams and closer customer scrutiny than US platforms, strategy now means building what the API cannot sell you.

David Saliba . 2026-07-21 . aimonger.com/whitepapers/durable-advantage-commodity-models/

The strategy shift

For a brief period, access to frontier models felt like strategy. It is no longer.

In the boardroom, the question is no longer whether the firm can access a capable model. It can. As of July 2026, closed flagships (GPT-5.x, Claude Opus-class, Gemini 3.x) and competitive open weights (DeepSeek-V4, Zhipu GLM-5.2, Alibaba Qwen3.x) are widely rentable or self-hostable. The question is what remains proprietary once a competitor, a cloud vendor, or a global software suite offers the same capability band next quarter.

What you see globally right now is a commodity curve forming in public. US hyperscalers are spending at infrastructure scale, open and closed models keep leapfrogging, and enterprise software vendors are bundling AI into existing seats. Gartner projected that by 2026 more than 80 per cent of enterprises would have used generative AI APIs or deployed GenAI-enabled applications in production, up from less than 5 per cent in 2023. McKinsey's State of AI 2025 survey finds roughly 88 per cent of organisations regularly using AI in at least one function, while only 39 per cent report any enterprise-level EBIT impact, and about 6 per cent qualify as high performers with material EBIT contribution from AI.

Those numbers describe a market where capability is widely rented and value is narrowly captured. In that market, "our model" is a weak story. Durable advantage sits one layer up: in how the firm learns from reality, keeps customers, makes judgements, and absorbs intelligence into operations.

This paper is a corporate strategy briefing. It deliberately avoids personality-driven business folklore. The frameworks are operational: what a leadership team can inspect, fund, and manage.

> **Commodity models, plainly:** foundation models that are good enough for many enterprise tasks from multiple vendors, with falling price per unit of capability, not identical models, but interchangeable enough that model brand alone is a weak moat. **In practice:** your three largest competitors can call GPT-5.x or Claude Opus-class APIs, or self-host DeepSeek-V4 / GLM-5.2 next quarter; strategy must live in workflows and data, not in a logo on the API invoice.

What commodity means here

Commodity does not mean "all models are identical." It means:

1. **Multiple adequate options.** Multiple adequate options exist for common enterprise tasks such as drafting, classification, extraction, summarisation, and coding assistance, so vendor selection is hygiene not strategy. When five vendors can all summarise a contract to a usable standard, the vendor choice is a procurement decision, not a competitive position.
2. **Falling model switching cost.** Switching cost at the model layer is falling as interfaces standardise, which means today's bake-off winner is next quarter's interchangeable supplier. A firm that built its roadmap around one model's edge now has a roadmap built around a depreciating asset.
3. **Declining unit capability price.** Price per unit of capability trends down over time for a given quality band, rewarding absorbers who expand coverage rather than those who celebrate access alone. The strategic question is whether you convert cheaper tokens into more coverage or simply let them sit as margin headroom someone else will take.
4. **Peer parity on purchase.** Competitors can buy the same capability band you can buy, so advantage must live above the API in workflows, data, and trust rather than in exclusive model contracts. Purchase parity is the fact that dissolves "we have a better model" as a moat.

If your advantage requires the other side not to have access to a similar model, your advantage is on a timer.

Open weights sharpen the same point. DeepSeek-V4, Zhipu GLM-5.2, and Alibaba Qwen3.x (as of July 2026; see vendor release pages in References) mean peers can self-host near-frontier capability without a closed US API contract. CNBC journalism in June–July 2026 reported Chinese open models gaining share as closed token prices stayed high, and as US access limits briefly constrained some Anthropic and OpenAI rollouts. That journalism is colour on market pressure, not primary proof of quality or adoption. The durable point is the geopolitical fork: US closed meters versus Chinese-lab open weights you can serve indoors. Advantage still does not live in the download. It lives in proprietary evaluation fixtures, workflow attachment, and alpha that never leaves your tray. Firms that treat Chinese open weights as free strategy without an origin policy and supply-chain check are buying a new supplier risk under a sovereignty label.

Eurostat's 2025 reading, 20.0 per cent of EU enterprises with 10+ employees using AI, shows EU adoption is still uneven at the firm-count level. That creates a temporary tempo window for absorbers. It does not create a permanent moat from model access alone. As adoption normalises inside your peer set, only hard-to-copy assets remain.

This matters more for the mid-market than for the global platform tier. Most firms will not win by outspending hyperscalers or hiring a research lab. They can win by making better use of their own workflows, customer knowledge, and delivery reputation.

Advantage map: five assets the API cannot sell you

1. Reality feedback loops

> **Reality feedback loop, in short:** routing real-world corrections, rejected drafts, fixed extractions, won/lost outcomes, back into evaluation sets, retrieval corpora, and process design so the system learns from operations. **Worked case:** every wrong clause extraction in a Dutch legal-services workflow becomes a regression fixture within two weeks, owned by the domain lead.

A model without a proprietary feedback loop is a rented brain with amnesia.

Reality feedback means the organisation captures outcomes from the real world, won/lost deals, corrected extraction fields, accepted/rejected drafts, incident postmortems, customer complaints, and routes those signals back into prompts, retrieval corpora, evaluation sets, and process design.

Questions for the board:

- **Correction routing.** Where corrections go after a wrong answer determines whether the firm learns or repeats mistakes, so every production workflow needs a named destination for fixes and fixtures. A wrong answer with no destination is paid for twice: once in the error, once in the repeat.
- **Evaluation set ownership.** Who owns the evaluation set for each production workflow must be written before scale, because ownerless eval sets become stale the moment the champion rotates. An eval set without an owner is a set nobody updates, which means it stops measuring the system the firm actually runs.
- **Failure-to-fixture speed.** How fast a repeated failure becomes a regression fixture measures learning rate, which is often a stronger strategic indicator than which model brand you rented this quarter. Speed here is the operational definition of "are we getting better."

Firms that close this loop learn faster than firms that only chat with a generic model. Learning rate becomes strategy.

2. Retention and switching costs that are earned, not cosmetic

> **Switching costs means:** friction that keeps a customer or operator on your workflow because history, training, audit trails, and trust are embedded, not because you hid the export button. **On Monday:** a Belgian SaaS vendor's renewal reason is "our exception queue lives here", not "we had the same chat UI first."

AI features do not automatically retain customers. Retention comes from workflows embedded in the customer's operating rhythm, data gravity, and trust.

Earned switching costs look like:

- **Structured customer history.** The system holds the customer's structured history in a form competitors cannot instantly recreate, so migration means rebuilding exception queues and audit trails not exporting a prompt file. The customer stays because leaving means re-living the painful implementation they just finished.
- **Workflow-trained operators.** Operators are trained on your workflow rather than only on a chat box, which raises switching cost because retraining is operational downtime not a UI preference. A trained operator is a retention asset that does not appear on the balance sheet.
- **Regulator-grade provenance.** Provenance and audit trails are part of how the customer satisfies their own regulators, making your path load-bearing for compliance rather than optional convenience. When your system is part of the customer's compliance evidence, switching is a regulatory event, not a procurement one.

Cosmetic switching costs look like:

- **Public model skin.** A skin on a public model with no proprietary data or workflow attachment, which a competitor can replicate in a quarter by signing the same API contract. The skin is a UI choice, and UI choices are the cheapest thing to copy.
- **Prompt library only.** A prompt library with no proprietary data, evaluation sets, or customer history embedded, so switching vendors costs little more than retyping instructions. A prompt library is a file, and files travel by email.
- **Pilot discount expiry.** A pilot discount that expires into a blank contract, leaving no operational dependence, audit trail, or trained operators who prefer your workflow over a rival assistant. When the discount ends, so does the only reason the customer was there.

If a competitor can replicate your product experience in a quarter with the same model API, you built a feature, not a moat.

3. Judgement under uncertainty

Models compress the cost of first drafts and first passes. They do not own accountability.

Judgement advantage shows up when the organisation can:

- **Intelligence versus automation choice.** Decide which workflows deserve intelligence versus deterministic automation, so model spend lands where messiness pays and rules engines suffice elsewhere. The choice is judgement because it determines where scarce model budget produces the most operating change.
- **Weak-evidence refusal rules.** Set refusal rules when evidence is weak, preventing confident drafts from becoming customer or regulator commitments the firm cannot defend. A refusal rule is judgement encoded as policy, and it is what stops the model from making promises the firm has to keep.
- **Consequential human gates.** Place human gates on consequential actions with named reviewers, so accountability stays visible when models propose faster paths. The gate is where the firm decides which proposed actions become real ones.
- **Metric-based project kills.** Kill projects that cannot show a metric inside an agreed window, stopping immortal pilots from consuming capital better spent on absorption. A kill decision is judgement with a date, and it is how the firm protects its portfolio from hope.

McKinsey's high-performer minority redesigns workflows and invests with intent. That is judgement institutionalised, not a better demo.

4. Distribution and trust

In mid-markets, trust travels through professional networks, sector reputation, and proof under scrutiny. A louder model launch does not substitute for a referenceable operating result.

Meanwhile, the global narrative rewards visible AI announcements. Customers in dense professional markets often reward something quieter: a system that works in their actual operating context, with someone accountable when it does not.

Trust compounds when:

- **Provenance on consequential answers.** You can show provenance for consequential answers with version IDs and passage locators, which lets customers and regulators audit what the system asserted and why. Provenance is how a firm turns "trust us" into "verify us," which is a stronger position in a dense network.
- **Honest failure modes.** You can explain failure modes without theatre, which builds referenceability when something goes wrong instead of destroying dense professional-network reputation. A firm that explains a failure keeps the reference; a firm that hides it loses the network.
- **Client team training.** You train client teams on gates and overrides, not only sell licences, so operators stay when the model API becomes interchangeable next quarter. Trained client operators are a retention moat the API cannot touch.
- **Invitation-quality delivery.** You stay within invitation-quality delivery rather than spray marketing, because mid-market trust travels through references more than launch-post volume. One strong reference in a dense sector outperforms a hundred posts in a market that buys on reputation.

This is slow. It is also difficult to copy quickly, which is the point.

5. The operating system around the model

> **Absorption capacity, put simply:** the firm's ability to integrate model capability into systems of record, permissions, evaluation, owners, and trained operators (see the companion AIMonger paper on that topic). **Operating case:** two Austrian manufacturers buy the same model; only one ships document intelligence into procurement within a quarter because data access and owners were already decided.

Call it absorption capacity (see the companion AIMonger paper on that topic). The operating system includes:

- **Grantable data permissions.** Data permissions that can be granted without indefinite delay, so integration tickets do not become the silent killer of every absorption pilot. A permission that takes six months is a permission that kills the pilot it was meant to enable.
- **Systems-of-record integration.** Integration with systems of record where outcomes must land, because intelligence that stops at a chat window does not change cycle time finance recognises. The system of record is where the work actually happens, and a model that cannot read or write it is a spectator.
- **Evaluation and escalation.** Evaluation and escalation paths with named owners, so wrong outputs become fixtures and fixes rather than anecdotes shared in hallway complaints. An escalation path is how a wrong answer becomes an asset instead of a recurring cost.
- **Named outcome owners.** Named owners accountable for workflow metrics, not merely for procurement of another seat bundle or vendor renewal. An owner accountable for the metric is who changes the workflow when the metric demands it.
- **Operator training.** Training for operators who run production daily, since tools without trained operators become shelfware or shadow-IT workarounds within weeks. A tool with no trained operators is a tool that reverts to email the moment the champion is on holiday.

Two firms can buy the same model. Only one may be able to put it into claims, procurement, or discovery in a quarter. That difference is strategy.

Thin wrappers: the value trap

> **Thin wrapper (plain English):** a product whose only scarce ingredient is a public model API plus UI, with no proprietary data loop, workflow ownership, or retention logic. **Board reading:** a competitor clones your "AI assistant" in ninety days because your moat was a prompt library, not operating attachment.

A recurring capital allocation error: funding products or internal platforms whose only scarce ingredient is a public model.

Symptoms:

- **Next-model roadmap.** The roadmap is "wait for the next model release" instead of shipping workflow metrics this quarter, which signals strategy lives at the vendor keynote not in operations. A roadmap that depends on someone else's release schedule is not a roadmap the firm owns.
- **Model-brand differentiation slides.** Differentiation slides mention model brands more than proprietary workflows, telling investors and boards the moat is a logo on the API invoice. A slide that names the vendor more than the workflow is a slide with no moat on it.

- **Token arbitrage margin.** Gross margin depends on token arbitrage competitors can also perform, so price cuts from hyperscalers flow straight through with no proprietary loop to absorb the shock. Arbitrage is a margin position that lasts exactly as long as the price gap does.
- **No domain evaluation set.** No evaluation set ties releases to your domain propositions, which means every upgrade is a faith exercise rather than a measured promotion decision. Without a domain eval set, the firm cannot tell improvement from regression on the cases that matter.
- **UI-only retention story.** No retention story exists beyond UI familiarity, so customers can leave when a suite vendor bundles a similar assistant into software they already pay for. Familiarity is a retention story that ends the day a cheaper familiar alternative appears.

Gartner's agent-related forecasts sharpen the same point from another angle: by the end of 2026, 40 per cent of enterprise applications are expected to feature task-specific AI agents, up from less than 5 per cent. When agent features become a default checkbox in enterprise software, "we added an agent" stops being a strategy announcement. It becomes table stakes, unless the agent sits on proprietary loops, data, and judgement.

That is why seat-count adoption theatre is such a weak signal. A firm can look current in the board pack and still own no learning loop, no workflow attachment, and no reason for a customer or operator to stay.

A value equation boards can actually use

For any AI-facing offer, external product or internal capability, inspect four terms:

Term	Board question
Dream outcome	What measurable end-state does the user want (time, risk, revenue, compliance)?
Perceived likelihood of achievement	Why should they believe we can deliver: proof, provenance, references, pilots with metrics?
Time delay	How long until the first trusted win in their workflow?
Effort and sacrifice	What integration, training, and process change do they must accept?

Advantage improves when you raise the first two and reduce the second two, using assets competitors cannot buy. Improving only the demo (perceived likelihood without reality feedback) creates churn.

Time horizon: sticky growth versus theatre growth

Theatre growth:

- **Launch-post signups.** Signups spike after a launch post but do not convert to weekly workflow use, which flatters marketing dashboards while renewal conversations stay empty. A signup that never becomes a weekly workflow is a number that costs money to acquire and keep.
- **Pilot logos without production.** Pilot logos accumulate without production usage or metrics, creating a slide deck asset that evaporates when procurement asks for reference calls. A pilot logo is not a customer; it is a trial that has not yet decided.
- **Novelty usage spikes.** Usage spikes die when novelty ends because no workflow attachment or training made the tool load-bearing in daily operations. Novelty is a usage pattern with a half-life, and the half-life is short.

Sticky growth:

- **Weekly production workflows.** Workflows run weekly without heroics from the project team, proving the system survives champion vacation and reorganisation. A weekly workflow is the operational proof that the system is load-bearing.

- **Expanding proposition coverage.** Proposition coverage in the evaluation set expands as corrections become fixtures, showing the asset learns from reality rather than from demo scripts. An eval set that grows with real corrections is an asset that compounds.
- **Operational net retention.** Net retention is driven by operational dependence and trust, so customers renew because exception queues and audit trails live in your path, not because the chat UI was first. Operational retention is the kind that survives a competitor's price cut.

Boards should ask for the sticky metrics: weekly active workflows, exception rates, time-to-resolution, and renewal reasons, not only seat counts.

Hiring and capability without unicorn theatre

Commodity models change what "talent" means.

You still need strong engineers and domain experts. You do not need to wait for mythical people who "know every model." You need people who can:

- **Workflow placement.** Map a workflow and place a model where it changes cost or quality with a ninety-day metric, not where a demo looks impressive in a steering meeting. Placement is a judgement skill, and it is what turns a model into a workflow change.
- **Evaluation and provenance.** Build evaluation and provenance instrumentation that survives vendor swaps, so model churn does not reset your trust controls every quarter. Instrumentation that travels across vendors is how the firm keeps its trust controls when the model changes.
- **Systems-of-record integration.** Integrate systems of record where outcomes must be recorded, because intelligence that cannot write or read the authoritative store changes little finance recognises. Integration is the unglamorous work that makes the model matter to the P&L.
- **Operator teaching.** Teach operators on production paths with failure-mode drills, turning awareness slides into capacity that survives the next model release. A drill teaches an operator what to do when the system is wrong, which a slide never does.
- **Trade-offs under EU rules.** Make trade-offs under EU operating constraints such as residency, consultation, and sector rules without using them as excuses for permanent sandbox culture. The skill is making constraints design parameters, not reasons to defer.

Eurostat's adoption figures imply many peers are still early. Capability advantage can be built with disciplined mid-level teams, clear ownership, and operator training, not only with scarce research celebrities.

What to build when models get stronger

Assume models continue to improve. Strategy still holds if you invest in:

1. **Proprietary corpora with governance.** Proprietary corpora include temporal validity and permissions, so answers cite the signed version rather than the draft still indexed from a shared drive. A governed corpus is what makes a citation mean something, because the firm knows which version the system read.
2. **Evaluation as a product.** Evaluation becomes a product with proposition-level tests for your domain, making promotion decisions reproducible instead of argued from vendor benchmarks alone. Evaluation as a product is how the firm stops trusting someone else's benchmark and starts trusting its own.
3. **Human gates by design.** Human gates are designed into consequential workflows rather than apologised for afterward, especially where actions affect customers, employees, records, or regulated processes. A gate designed in is a gate that works; one added after an incident is a gate that arrived too late.

4. **Operator training capacity.** Operator training creates capacity through workshops on real systems, not awareness theatre that leaves teams reverting to email when the project team steps away. Trained operators are the capacity that absorbs the next model without a fresh project.
5. **Feedback capture loops.** Feedback capture routes corrections into evaluation sets and fixtures within weeks, so the firm learns faster than competitors who only chat with a generic model. A feedback loop is the asset that makes the firm smarter than its model choice.

If models get much stronger and you own none of the above, you become a reseller of someone else's improvement curve.

Capital allocation checklist (this year)

Fund:

- **Workflow redesign with owners.** Workflow redesign with named owners and ninety-day metrics, so capital connects to cycle time or margin movement finance can audit. A workflow with an owner and a metric is an investment; one without either is a bet.
- **Trust instrumentation.** AI discovery systems with holdouts and provenance before scale, because trust controls are cheaper before consequential traffic than after a visible failure. Trust instrumentation bought early prevents the incident that costs ten times more.
- **Work-changing training.** Training programmes that change how work is done on production paths, not town halls that celebrate licences without operator practice. Training that changes work is an asset; training that celebrates a launch is an event.
- **Absorption-enabling governance.** Data governance that unblocks absorption with bounded fast paths, so legal and security become tempo enablers rather than indefinite deferral machines. Governance that enables absorption is governance that pays for itself in tempo.

Defund or tightly cap:

- **Undifferentiated wrappers.** Undifferentiated wrappers whose only scarce ingredient is a public API plus UI, replicable by any peer within a quarter. A wrapper with no proprietary loop is a feature a competitor can buy the same ingredients for.
- **Immortal pilots.** Permanent pilots without graduation or kill criteria, which tax attention while producing no workflow metric movement for the CFO pack. An immortal pilot is a budget item that never produces a decision.
- **Ownerless bake-offs.** Model bake-offs disconnected from a production workflow and named owner, optimising vendor logos while process design stays untouched. A bake-off without an owner produces a winner and no workflow.
- **Tool sprawl without retirement.** Tool sprawl with no retirement plan, leaving operators juggling five assistants while none reach operating capacity. Sprawl without retirement is how a firm pays for five tools and gets zero capacity.

McKinsey's split between widespread use and scarce EBIT impact is a capital allocation warning. Spend that does not change workflows is consumption, not investment.

Competitive scenarios for strategy offsites

Use these scenarios to force choices. They are not predictions; they are decision drills.

Scenario 1: Peer parity on models

Your top three competitors gain access to the same model band within one quarter. What remains of your advantage in twelve months? If the answer is "our UI," revisit the advantage map.

Scenario 2: Price collapse

Inference costs for your main task class fall by half. Do you expand coverage, drop price, improve margin, or do nothing because the workflow was never absorbed? Commodity shocks reward absorbers.

Scenario 3: Operating constraint under EU rules

A sector rule, customer requirement, or internal risk decision forces human gates on a workflow you auto-ran. Does your operating system already support propose-and-wait, or does the product break?

Scenario 4: Vendor lock narrative

A cloud vendor bundles "agents" into the suite you already buy (consistent with Gartner's forecast of agents becoming common in enterprise apps by end-2026). Is your differentiation above that bundle, or inside it?

Portfolio design: where to place bets

Allocate AI-related investment across four buckets with explicit percentages:

Bucket	Purpose	Healthy signal
Absorb	Redesign priority workflows	Metric movement in 90 days
Trust	Evaluation, provenance, security	Scorecard green before scale
Learn	Feedback loops and training	Correction-to-fixture latency falling
Explore	Bounded bets on new model capabilities	Kill dates honoured

A portfolio that is 80 per cent Explore and 20 per cent Absorb is how organisations buy demos. Invert it until absorption produces cash or capacity.

What "judgement" looks like as a managed asset

Judgement is not mystique. It is a set of written decision rights:

- **New workflow approval.** Who may approve a new AI workflow with documented residual risk acceptance, so autonomy expands only where consequence and reversibility are understood. Approval authority is how the firm decides which experiments become operating dependencies.
- **Tool permission expansion.** Who may expand tool permissions for agents or discovery systems, preventing "temporary" write access from becoming permanent privilege creep. Permission expansion is the most common path from a bounded pilot to an unbounded incident.
- **Pilot retirement authority.** Who may retire a pilot with a written reason, stopping immortal demos from consuming budget because no one wants to be the person who said no. Retirement authority is how the firm keeps its portfolio honest.
- **Refusal-rate risk acceptance.** Who may accept residual risk on refusal rates when the business prefers speed over coverage, documenting the trade-off instead of hiding it in averages. A documented trade-off is a decision; an undocumented one is a drift.
- **Incident board voice.** Who speaks to the board when an incident occurs, so accountability is named before the first customer-visible failure rather than debated under pressure. A named voice is how the firm responds in days rather than weeks.

Document decision rights the way you document banking authorities. Commodity models increase the number of possible actions; judgement decides which actions exist.

Retention economics without vanity metrics

Track:

- **Workflow attachment.** Share of target process volume touching the AI path, because seats without attachment are deferred churn waiting for a cheaper rival assistant. Attachment is the metric that distinguishes a customer from a subscriber.
- **Time-to-first-trusted-win.** Time-to-first-trusted-win for new teams measures how fast operators trust the system on real cases, not how fast they logged in after a launch email. A trusted win is the moment the system becomes load-bearing for that team.
- **Renewal reason codes.** Reason codes on renewal or expansion separate operational dependence from novelty, so sales hears “our exception queue lives here” instead of “we liked the demo.” Reason codes are how the firm learns why customers actually stay.
- **Training completion signal.** Training completion is a leading indicator of attachment, since teams that skip operator enablement revert to manual chase work when champions rotate. A trained operator is the difference between a renewal and a quiet departure.

Seat licences without attachment are deferred churn.

Buildversus-buy when the model is rented

Buy when:

- **Generic capability need.** The capability is generic such as email drafting or meeting notes, where proprietary feedback loops are unlikely to differentiate you within a year. A generic capability is a commodity; treat it as one.
- **Adequate isolation and evaluation.** Isolation and evaluation from the vendor are adequate for the risk tier, meaning permissions, logging, and export paths meet your minimum control bar. Adequate is a risk decision, not a hope.
- **Acceptable switching cost.** Switching cost is acceptable if the vendor changes terms, so you are not betting the operating model on a single non-portable control plane. An acceptable switching cost is what keeps the firm from being held hostage by a vendor roadmap.

Build (or tightly configure) when:

- **Operating uniqueness.** The workflow is your operating uniqueness, where mistakes or leakage would damage customer trust or regulatory standing in ways a generic tool cannot understand. Uniqueness is the signal that a generic tool will fail in ways specific to your firm.
- **Domain-matched provenance.** Provenance and evaluation must match your domain propositions, because public benchmarks do not prove accuracy on your clauses, policies, or case patterns. Domain-matched provenance is what makes the system trustworthy on your cases, not on a benchmark’s.
- **Feedback as product.** Feedback loops are themselves the product, meaning corrections, fixtures, and routing policy compound into an asset competitors cannot buy on a price sheet. When feedback is the product, the firm owns an asset that appreciates with use.

Most mid-market firms should buy commodity layers and build absorption on top, not rebuild foundation models, and not confuse a thin wrapper with a build strategy.

Communication discipline

Internal and external messaging should prefer:

- **Workflow and training metrics.** Workflow metrics and training outcomes that finance and customers can verify, rather than model-brand announcements that any peer can copy next quarter. A metric a customer can verify is stronger than a claim a peer can match.
- **Trust controls and gates.** Trust controls and human gates described plainly, so buyers know how you fail safely instead of hearing only capability superlatives. A plainly described gate is how the firm shows it has thought about failure, not just capability.
- **Peer tempo comparisons.** Tempo relative to peers on cycle time and exception rates, grounded in operating evidence rather than viral anecdotes about AI failure or success. Tempo comparisons turn strategy from a claim into a measurable position.

Avoid:

- **Model brand as strategy.** Model brand announcements as strategy, which sound current in a board pack and leave no asset when the API price sheet changes. A brand announcement is a press release, not a position.
- **Unverified viral statistics.** Unverified viral failure statistics that cannot be tied to a named study, sample, and question wording finance would accept. An unverified statistic is a story, and stories do not survive a sceptical board.
- **Client identity implications.** Claims that imply client identities or confidential systems, which destroy invitation-quality trust in dense professional networks. An implication of client identity is a trust cost paid in the network the firm depends on.

AIMonger's public posture remains invitation-quality and precise. Strategy documents inside the firm can be sharper; public pages stay falsifiable and non-disclosing.

Appendix: strategy one-pager for the board pack

Fill the Entry column for the next board pack. Starter text is a prompt, not a finished answer.

Field	Entry
Thesis	Models are widely available; advantage is absorption, feedback, retention, judgement, and trust.
Evidence	Gartner GenAI mainstreaming projection; McKinsey use-versus-EBIT gap; Eurostat EU enterprise adoption.
Where we will invest	List three workflows, trust systems, or training programmes.
Where we will not invest	Thin wrappers, permanent pilots, bake-offs without owners.
Leading indicators (next two quarters)	Workflow attachment, time to trusted win, evaluation coverage, training completion.
Lagging indicators	EBIT or capacity metrics attributable to redesigned workflows.
Risks	Evaluation theatre, agentwashing purchases, underestimated regulatory gates.
Ask of the board	Approve portfolio percentages; require kill criteria; sponsor data-access decisions.

Advantage decay tests (run twice a year)

1. **Clone test.** Could a well-funded competitor copy our AI-facing experience in ninety days using public models, and if yes, which proprietary loops must we fund this quarter? The clone test is how the firm measures how much of its position is actually hard to copy.
2. **API test.** If our model vendor vanished tomorrow, which assets would still be ours including corpora, fixtures, workflows, and trained operators? The API test separates owned assets from rented ones in a single thought experiment.

3. **Staff test.** If three early champions left next month, would the workflows still run with documented gates and operator coverage rather than tribal knowledge? The staff test measures whether the system depends on heroes.
4. **Trust test.** Can we show provenance and evaluation for our top consequential AI path today, not after the next release is promised? The trust test is whether the firm can prove its position now, not after a roadmap.
5. **Retention test.** Would customers stay if a rival offered a similar assistant at half price because exception queues and audit trails live in our workflow? The retention test is whether the firm's moat survives a price attack.

Failing two or more tests means the portfolio is overweight Explore and underweight Absorb/Trust/Learn.

Closing the loop with training and workshops

Durable advantage is partly pedagogical. Tools change quarterly. Operators who understand failure modes, gates, and metrics can absorb the next model without restarting strategy. That is why executive workshops and team training sit inside the advantage map, not beside it as optional culture work.

Commodity intelligence rewards firms that teach faster than they shop.

Evidence base: capability diffuses, complementary assets do not

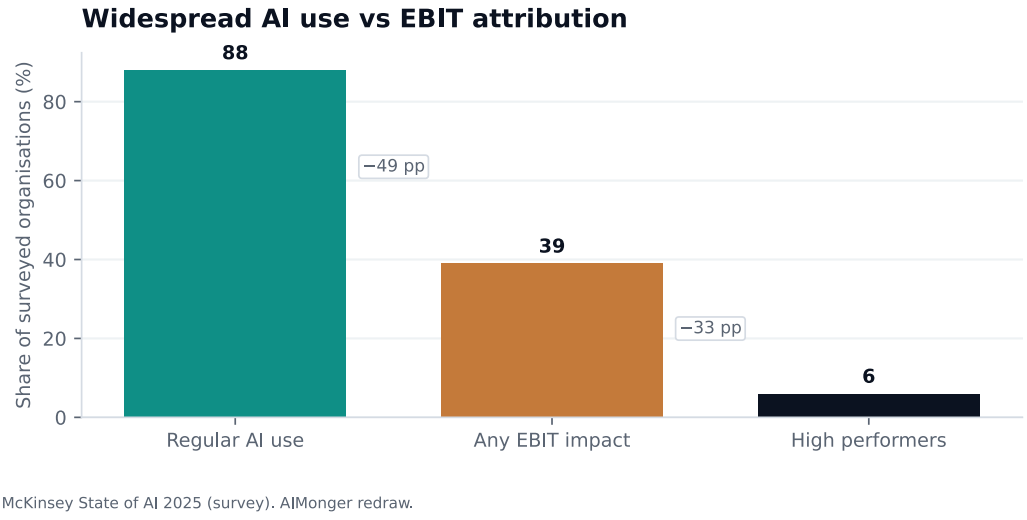

Figure 1. Regular AI use vs enterprise EBIT attribution. Plain read: access is common; financial impact is not. **On the P&L:** reallocate from model bake-offs to workflow metrics that move margin. Source: McKinsey State of AI 2025 (survey). AIMonger redraw.

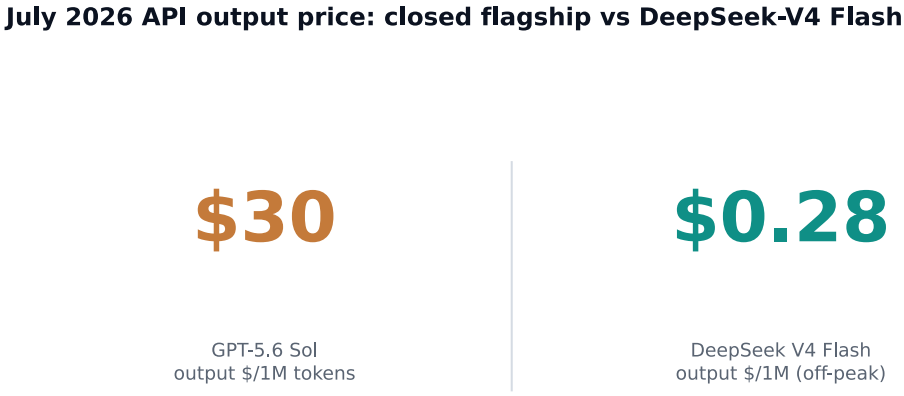

Figure 2. July 2026 list prices for output tokens: OpenAI GPT-5.6 Sol at $30 per million versus DeepSeek V4 Flash off-peak at $0.28 per million - roughly 100× cheaper on the open/API commodity band. In short: current closed flagships (GPT-5.5 / GPT-5.6 Sol-Terra-Luna) and near-frontier open weights (DeepSeek-V4, GLM-5.2, Qwen3.x) already make model access a buyer's market; proprietary workflow attachment still is not. **On Monday:** route extraction to the cheap open or mid tier, reserve GPT-5.6 Sol / Claude Opus-class for judgement, and measure margin on the workflow - not on which SKU wins a bake-off this week. Source: OpenAI API pricing page (Jul 2026); DeepSeek V4 Flash off-peak list (Jul 2026). Peak-hour DeepSeek rates double. AIMonger redraw.

Evidence	Finding	Strategic implication
Gartner public forecast (2023)	More than 80% of enterprises expected to have used GenAI APIs/models or deployed GenAI-enabled applications by 2026	Model access cannot remain a durable differentiator
Stanford AI Index (2026) + July 2026 list prices	Industry produced over 90% of notable frontier models in 2025; July 2026 output list prices still span roughly $30/M (GPT-5.6 Sol) to $0.28/M (DeepSeek V4 Flash off-peak), with GPT-5.5/5.6, Claude Opus/Fable-class, Gemini 3.x, DeepSeek-V4, GLM-5.2, and Qwen3.x all in the live market	Enterprise strategy depends on external model markets and routing; differentiation migrates upward into workflows
McKinsey State of AI (2025)	About 6% of respondents met the high-performer definition; workflow redesign and faster scaling were among their reported practices	Complementary operating assets correlate with value, although causality is not proven
PwC AI Jobs Barometer (2025)	Analysis of nearly one billion job ads and company financials found revenue-per-employee growth three times higher in industries most exposed to AI; AI-skilled roles carried a 56% wage premium	Skills and work design remain scarce complements; sector-level association is not proof that AI alone caused the productivity difference

Methodology and limits

"Commodity" is used here in an economic, not technical, sense: multiple adequate suppliers, falling switching friction at the API layer, and broad buyer access. Frontier models still differ materially by task, cost, latency, language, and governance. Firms should evaluate them. The argument is narrower: a model-selection advantage decays faster than proprietary workflow, feedback, distribution, and trust advantages.

PwC's analysis is observational. Industries differ for reasons beyond AI, and job-ad data captures employer demand rather than realised worker capability. McKinsey's high-performer practices are self-reported correlations. These limitations make the strategic prescription more conservative, not less: do not bet a durable position on a benchmark lead controlled by a supplier.

Counter-position: superior model access can still matter

Exclusive access, fine-tuning rights, scarce inference capacity, or task-specific performance can create temporary advantage. The appropriate response is to value that advantage with a half-life. Ask how long until peers can buy an adequate substitute, then invest the temporary surplus in assets with slower decay: proprietary evaluation data, customer workflow attachment, trained operators, and trusted distribution.

Competitive advantage register

For each claimed AI advantage, the strategy team records:

Field	Question
Asset	What exactly do we own or control?
Replication time	How long for a competent peer to reproduce it?
Supplier dependency	What disappears if the model vendor changes?
Feedback velocity	How does real-world use improve the asset?
Retention effect	Why does a customer or operator stay?
Evidence	Which operating metric proves the claim?

Claims without an owned asset and evidence belong in product backlog, not corporate strategy.

Research addendum: cost collapse and benchmark convergence

Stanford AI Index 2026 documents industry concentration in frontier model production and a closed US-China performance gap on public arenas. July 2026 API lists make the buyer's market concrete: GPT-5.5 and GPT-5.6 Sol/Terra/Luna sit beside Anthropic and Google flagships, while DeepSeek-V4, GLM-5.2, and Qwen3.x undercut closed meters by large multiples (and can be self-hosted). Exact SKUs and list prices churn monthly; the strategic point is stable: **current-generation capability is widely available at competing prices**, so strategy migrates to proprietary evaluation data, workflow integration, routing policy, reliability, and learning rate.

As public leaderboards converge, enterprise differentiation should be proven on **internal task sets** with latency, residency, safety behaviour, and cost-per-accepted-outcome, not on a single Elo or MMLU lead controlled by suppliers.

Closing position

When models are commodities, strategy is everything you build above the API.

Durable advantage is reality feedback, earned retention, institutional judgement, trust, and an operating system that absorbs intelligence into work. Those assets are slower to assemble than a prototype, and that is why they still matter for mid-market firms competing on reputation, delivery quality, and tempo.

Leadership teams that keep announcing model brands will sound current and remain replaceable. Teams that build hard-to-copy operating capacity will still have a position when the next model ships.

References

1. Gartner, "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026," press release, 11 October 2023. https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-

of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026

2. Gartner, "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025," press release, 26 August 2025. https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025

3. McKinsey & Company / QuantumBlack, "The State of AI: Global Survey 2025." https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

4. Stanford Institute for Human-Centered Artificial Intelligence, "The 2026 AI Index Report," Economy chapter. https://hai.stanford.edu/ai-index/2026-ai-index-report/economy

5. PwC, "The Fearless Future: 2025 Global AI Jobs Barometer," 3 June 2025. https://www.pwc.com/gx/en/news-room/press-releases/2025/ai-linked-to-a-fourfold-increase-in-productivity-growth.html

6. Eurostat, "20% of EU enterprises use AI technologies," 11 December 2025 (20.0% in 2025; 13.5% in 2024). https://ec.europa.eu/eurostat/en/web/products-eurostat-news/w/ddn-20251211-2

7. Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

8. Stanford Institute for Human-Centered Artificial Intelligence, "The 2026 AI Index Report," Technical Performance / Research chapters. https://hai.stanford.edu/ai-index/2026-ai-index-report

9. CNBC, "China's Zhipu is booming with Anthropic and OpenAI held back" (26 June 2026; journalism). https://www.cnbc.com/2026/06/26/china-zhipu-z-ai-open-source-anthropic-openai.html

10. CNBC, "Chinese AI models gain ground with U.S. companies as costs surge" (7 July 2026; journalism). https://www.cnbc.com/2026/07/07/chinese-ai-models-costs-us-openai-anthropic.html

11. DeepSeek, "DeepSeek V4 Preview Release" (24 April 2026). https://api-docs.deepseek.com/news/news260424/

12. Z.ai / Zhipu, "GLM-5.2: Built for Long-Horizon Tasks" (16 June 2026). https://z.ai/blog/glm-5.2

Published by AIMonger . https://aimonger.com . Malta . Europe