AI Capital Allocation and Governance

A Portfolio Model

A model for allocating AI investment as a portfolio: steering capital to its highest return, subject to an agreed risk limit.

James E. Murphy

ORCID: https://orcid.org/0009-0001-2217-4306

Working draft for discussion · v0.70 · July 2026

Latest version at jamesemurphy.com/ai-capital-allocation/portfolio-model

Contents

Executive summary

Artificial intelligence has become a major call on capital, and in 2026 it is meeting a hard demand for proof. Organizations are spending heavily, the largest infrastructure providers alone on the order of several hundred billion dollars this year, while most can show little measurable return. The gap is not mainly a technology problem, or not only one. It is a problem of allocation, measurement, and governance: capital committed without the discipline that other large and uncertain investments routinely receive. The competing reading, that the value of AI simply takes time to arrive, is real and is taken up in Part I; this paper argues that the discipline gap is the larger and the more fixable part, without denying that some of the shortfall is a matter of timing.

This paper sets out a model for that discipline. It treats a firm's AI activity as a portfolio: a set of distinct investments, each valued, fully costed, and judged by what it adds to the firm's total risk, allocated and governed as a whole rather than approved one at a time. The model has one objective and one constraint, met by two halves that share a single spine: one half to steer capital to its best-returning use, and one to bound the risk that use adds so the firm can back its strongest bets with conviction rather than hedge thinly across all of them.

The objective is return: to admit the firm's best risk-adjusted additions and refuse the rest, so capital flows to the opportunities that earn it rather than to the loudest sponsors. The first half, allocation, is a consistent way to value, cost, and risk-adjust each candidate. Cost here is not the sticker price but the fully burdened figure, adjusted for how much of the capacity is actually used, which can reverse which option looks cheaper. The test is marginal, not a single score. Value is kept multidimensional and risk is not collapsed into one dollar figure, so a candidate is judged on whether it earns its place against the firm's hurdle and how it compares with peers in its class, rather than ranked on one number across the whole book.

The constraint exists to protect that objective rather than compete with it: an explicit limit on the total AI risk the firm will carry, grounded in the board's risk appetite and translated by the risk committee into an operating budget. The second half, governance, sets that budget as a portfolio-level control and gates new investment against it. Because the acceptance criteria are published in advance, the work of meeting them moves earlier, into design; this is the shift left, and it means teams build only what can pass and engineer it to be flexible and low-risk from the start. That works only when incentives reward approvable outcomes delivered within budget rather than raw activity, which is what turns governance from a brake into an enabler. The portfolio framing itself is now widely urged, by the major consultancies, the analyst firms, and the largest private-equity owners of software alike, and is not what distinguishes this model. What is new is what the model governs at the portfolio level: the correlation between investments, a budget denominated in money the firm can afford to lose, an absolute floor under the worst case, and the shape of the book itself, whether its capital is maturing toward scale or scattering across many small bets that answer to no thesis. That aggregate view is what single-system and single-investment methods leave out. Put plainly, most AI governance asks whether a system is safe enough to run; this model asks whether a workload is worth funding, how much risk it adds to the book, and whether the whole book stays inside the limit the firm has agreed to carry.

Three commitments run throughout. Where a quantity cannot be known precisely, the model prefers a defensible relative measure, with a clear provenance and a stress test, to a falsely precise point estimate. Governance exists to enable, not to obstruct: a portfolio that knows its own risk can commit more and move faster, because it can tell a good risk from a dangerous one. And the model

is strict in process but deliberately adaptive in content. The cloud-era playbook does not fit a setting where the model layer keeps producing new risks, agentic systems make cost variable rather than fixed, and a growing share of payoffs behave like options, so the model treats flexibility, the ability to change course without heavy cost, as something to measure and preserve.

The field is converging, slowly and unevenly, on pieces of this discipline. What follows is a foundation on which those pieces can converge, and how a serious organization can put it to work. A firm need not adopt all of it at once: Part V sets out the smallest viable start, six concrete moves that begin the practice.

A note on scope. This paper sets out the model and the reasoning behind it, not a turnkey specification. The particular thresholds, weights, and scoring rules that an implementation requires are matters of calibration to a firm's own appetite, data, sector, and maturity, noted here only where the argument depends on them. Making that calibration possible, and governable, is the point.

Part I. The state of play: a field still converging

After three years of pilots, 2026 has become a reckoning for enterprise AI. Spending is large and still climbing, with the major infrastructure providers alone guiding to roughly $700 billion in capital spending this year, more than sixty percent above the year before, with some estimates approaching $725 billion, even as the returns remain hard to find. A widely cited 2025 study from the MIT NANDA initiative reported that ninety-five percent of generative AI pilots produced no measurable profit impact. The figure is contested and narrowly defined, and it matters that it does not stand alone: the corroboration is telling precisely because the other readings measure different quantities and point the same way. A February 2026 study from the National Bureau of Economic Research found that roughly nine in ten firms reported no productivity effect even as executives projected gains, a contemporary echo of the productivity paradox; and S&P Global found that roughly two in five companies had abandoned most of their AI projects, up from one in six a year earlier. Pilot profitability, firm productivity, and project survival are three different lenses, and all three point the same way. Markets have begun to reprice the gap, and boards are now expected to say what their AI is for and what it returns. Patience for spending on faith has run out.

This impatience is sharpened by a wider argument about whether the spending itself is rational. The capital behind the build-out flows in a tight circle among a few firms that are at once each other's suppliers, customers, and investors; capital expenditure is running far ahead of the revenue it is meant to produce; and systemic-risk authorities have begun to question whether the assets being built will hold their value. Markets are already pricing the doubt, and the cost of capital is expected to rise for firms seen as spending on AI without showing a return. One distinction matters here, and it fixes the scope of this paper. For the handful of frontier builders doing the largest spending, the logic can genuinely favor overbuilding, because the cost of underinvesting, should the technology prove as transformative as hoped, is existential. For the enterprise buyer this paper addresses, that asymmetry does not hold: money spent on AI that does not pay off is simply waste. The right posture for the buyer is therefore discipline rather than a race, and that is what the rest of this paper sets out.

The reckoning has produced a scramble of improvised practice rather than a settled answer. Finance leaders are inventing stage gates and kill criteria, investment envelopes set as a share of revenue, weighted scorecards, caps on single-vendor commitment and on variable usage cost, and dedicated AI investment committees. Risk and compliance functions are standing up governance committees

and aligning to a small set of emerging standards. Engineering and finance together are building a new cost discipline for AI. There is real signal in all of this. Much of what is emerging is sound, and in places the separate efforts are beginning to resemble one another. But the practice has not converged, it differs widely from one firm to the next, and it is weakest exactly where the decisions are hardest: measuring value, measuring true cost, and seeing risk in aggregate. The distance left to travel also varies with maturity, and not only between the leaders and the rest: even among the most mature, firms running real pieces of the discipline differ from one another on the machinery underneath, and inside a single advanced firm the rigor often governs one part of the book and not the rest. The sections that follow locate those gaps. The rest of the paper proposes a foundation on which the better answers can converge.

1.1 From a technology lens to a capital lens

The first gap is one of framing. AI has been treated, understandably, as a technology question: which models, which platforms, which use cases, which vendors. That framing is not wrong so much as incomplete, because the decision in front of a board is a capital decision: scarce money and management attention committed under deep uncertainty for an uncertain return, while risk accumulates across everything the firm is building at once. Widening the lens does not discard the technology view; it places it inside a capital and risk view, where two groups who rarely share a vocabulary, the people who can judge whether a model will work and the people who decide how the firm allocates capital and how much risk it will carry, have to meet on common ground. The argument of this paper is that the common ground is the portfolio.

1.2 Where value measurement can improve

The second gap is in how value is measured, and it is more consequential than it looks. Under pressure to justify a commitment, most teams reduce a use case to a single number, most often a cost saving, because a single number is what a business case seems to want. But the value of an AI investment is rarely a single thing. The same deployment can lower a cost, protect a revenue stream, generate new revenue, avoid a loss, or relieve a capacity constraint, often several at once and in different proportions for different firms. The improvement available here is to keep that blend intact rather than collapse it too early.

Reducing that blend to one figure invites a specific and expensive error. A cost-per-outcome metric is only as good as the outcome it names, and naming the wrong outcome can destroy value while appearing to improve it. The clearest case is the contact center. Modeled as a cost center, its governing metric is deflection or cost per contact, and the obvious move is to handle more contacts with less labor. But for many firms the contact center is also where retention, recovery, and cross-sell happen. Optimizing the cost metric on what is partly a revenue function can cut the very interactions that produce the revenue, and the saving will look entirely real on the one dimension anyone is measuring.

The state of play confirms the gap. Most organizations measure activity rather than outcomes, tokens processed and queries answered rather than revenue influenced or cost avoided, and the same token carries the same line on the invoice whether it ran a throwaway experiment or a revenue-bearing production system. Surveys through early 2026 found a large majority of firms wanting AI to drive revenue and only a small share able to show that it had.

1.3 Where cost measurement can improve

The third gap is on the other side of the ledger. Cost is hard to see clearly because attention naturally fixes on the most visible line, the price of the model or the accelerator, which is rarely where the money actually goes. For owned infrastructure the purchase of hardware is a minority of the total cost of ownership over its life; power, cooling, facility, networking, and staff dominate, and the single largest driver is utilization. A cluster that is busy is a different investment from one that is idle, and the difference is usually larger than any discount negotiated on the hardware itself. The improvement is to cost the whole thing and adjust it for how busy the capacity really is.

Two effects in particular are routinely missed. The first is that cloud billing hides what owned infrastructure exposes. A rented endpoint priced per unit folds occupancy and efficiency into a single number; an owned cluster reveals that the effective cost of useful work is the nominal cost divided by how much of the time the hardware is doing anything and how efficiently it does it when it is. The cost of a unit of genuinely useful output can be several times the apparent rate. The second is that depreciation and useful life, which together drive the largest capital component, are a management judgment rather than an engineering fact, and reasonable firms have recently chosen lives that differ by years for similar hardware. A model that does not surface and test that assumption is not measuring cost so much as asserting it.

The live version of this problem in 2026 is variable cost that outruns the budget. Usage-based spending on tokens can climb overnight, the retrospective tooling built for cloud bills is poorly matched to the real-time economics of token consumption, and the cost discipline now forming around AI, though nearly universal as a concern, still lacks the granularity to attribute spend to the system, team, or product that incurred it. That the measurement problem is real, and until recently unsolved, is confirmed by the move to establish an open standard for AI billing. That standard has since matured quickly. FOCUS, the open specification for cloud and AI cost, reached in mid-2026 a version that normalizes billing across providers and carries token economics directly, with adoption among large spenders already substantial. Its arrival changes what is scarce: once spend is normalized and comparable, the binding constraint moves from the accounting to the allocation, from measuring the cost to judging whether it was worth incurring, which is the work the rest of this paper is about. Agentic systems sharpen the problem from variable to genuinely stochastic. An agent decides at run time how many steps to take, how many tools to call, and how often to retry, so the cost of a single task is not a fixed price but a distribution with a tail, and a small fraction of tasks can dominate the bill. Costing an agent means modeling that distribution, including its tail, rather than quoting a number.

The scale of this change is now visible in the aggregate, and three features of it bear directly on capital decisions.

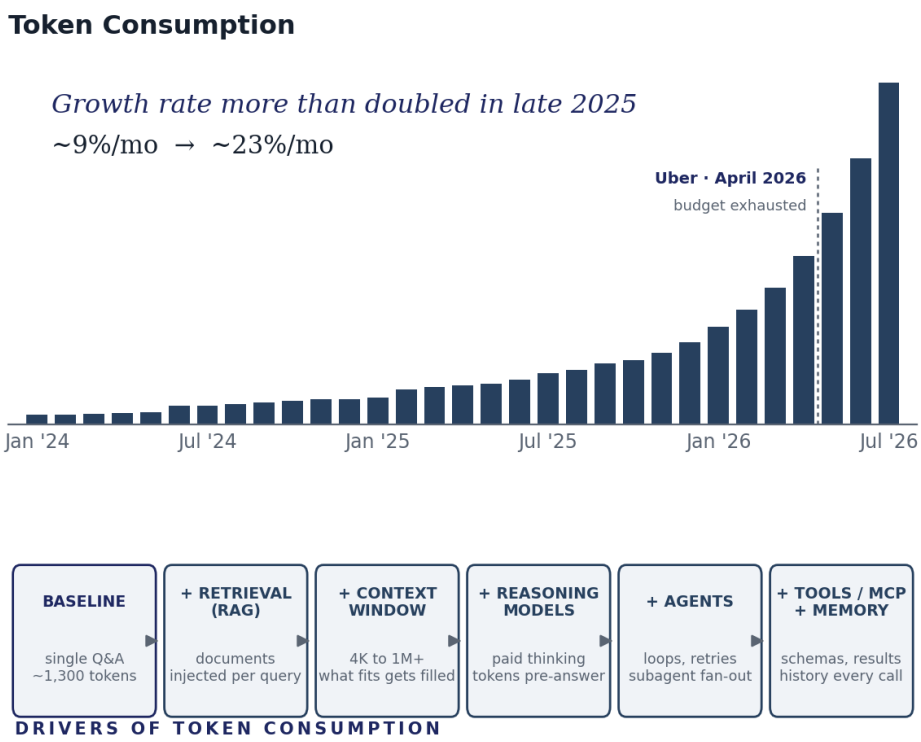

Figure 1. The consumption inflection and its drivers. Top: aggregate token consumption crossed from linear to nonlinear growth in late 2025, with the month-over-month rate roughly doubling. Bottom: the additive drivers that compound the bill, from a single chat exchange through retrieval, longer context, reasoning, agent loops, and tools. Source: the author's analysis (OpenRouter consumption data; FinOps X 2026 keynote). The decision implication: budgets framed for linear consumption fail once consumption goes nonlinear.

The first is that consumption has gone nonlinear. The unit cost of machine intelligence fell about as fast as anyone predicted, on the order of ninety-nine percent at constant capability over three years, but cost is rate times usage, and usage per task grew by as much as a hundredfold over the same period. The two moved in opposite directions and usage won: a token gets cheaper while tokens, in aggregate, do not. Budgets framed for the linear cloud era were already strained before the inflection; after it they break.

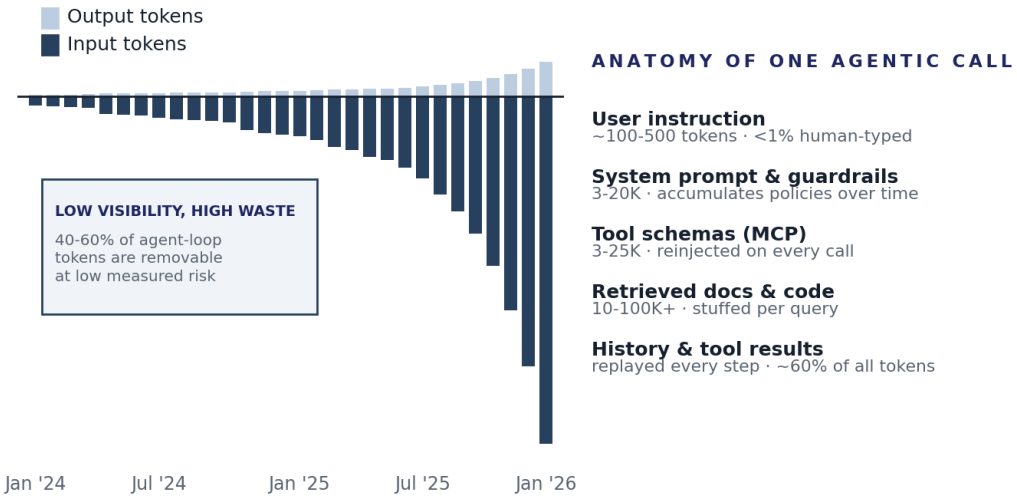

Figure 2. The growing iceberg and the anatomy of one call. Left: the tokens that drive the bill are overwhelmingly input rather than output, and machine-assembled rather than human-typed, with the typed share falling from about a fifth of a 2023 chat to under one percent of a 2026 agentic task. Right: the composition of that input in one agentic call, from the sub-one-percent human instruction up to history and tool results at roughly sixty percent of the total. Source: the author's analysis (SWE-bench context-engineering measurements, 2026). The decision implication: the bill is mostly machine-assembled context the business owner never types or sees, so waste hides below the visible prompt.

The second is that most of the spend is invisible to the people nominally accountable for it. What drives the bill is machine-assembled context, not anything a person typed: system prompts and accumulated guardrails, tool schemas reinjected on every call, retrieved documents stuffed into each query, and a running history replayed at every step. Benchmark measurements suggest tool results alone can be around sixty percent of the tokens, and that forty to sixty percent of the tokens consumed inside agent loops are removable at low measured risk. A large share of the spend, in other words, is waste that no one is watching.

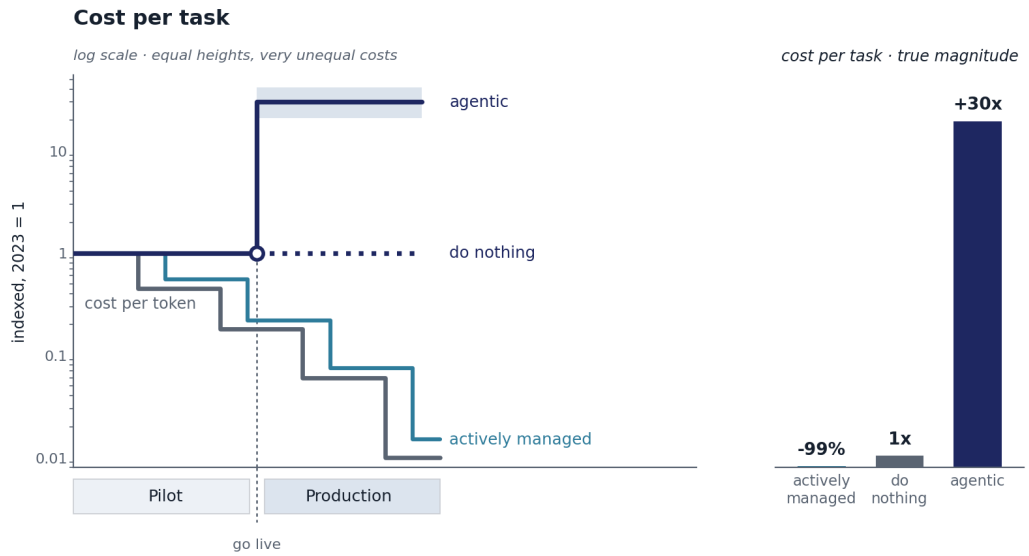

Figure 3. The forecast problem, two views. Left: cost per task splits at go-live into three paths while cost per token keeps falling. Agentic jumps to about thirtyfold at go-live, do-nothing holds flat at the 2023 level, and actively managed (proactively switching models as prices drop, for a given level of performance) tracks the rate down with a short lag to roughly minus ninety-nine percent; log scale, indexed to 2023 = 1. Right: the same three outcomes with the log scale removed, where the agentic path dwarfs the do-nothing baseline and the actively managed floor. Source: the author's analysis (observed deployments). The decision implication: neither the falling token price nor the rising task cost decides anything on its own; the fully burdened cost per resolved outcome does.

The third is that cost per token and cost per task have come apart. A customer-service interaction that was a single chat exchange in 2023, rearchitected as a fully orchestrated agent with reasoning, retrieval, tool calls, and subagents, runs on the token line alone at many times its predecessor, before the infrastructure, oversight, and failure recovery layered on top. Falling token prices and rising task costs are both real; they are different lines pointing in opposite directions. The decision-grade figure is therefore neither the price of a token nor the cost of a task but the fully burdened cost per resolved outcome, measured against what the same outcome costs a person, and a cost model that reads a falling token price as the trend in spending is measuring the wrong quantity.

None of this is hypothetical. In 2026 Uber exhausted its full-year budget for agentic coding tools within four months after ranking engineering teams by the tokens they consumed, with a senior executive noting the spending did not yet map to features shipped. Appendix C sets the episode out in full; it is the clearest recent case of the cost problem and the incentive problem arriving together, and the practical reason the model insists both on costing the whole distribution and on aligning incentives with the outcomes that justify the spend.

1.4 The missing portfolio view of risk

The fourth gap is the one most often misdiagnosed, because on the surface AI risk is being addressed energetically. It is. Single-system risk management is an active and maturing field, converging on a recognizable stack: the NIST AI Risk Management Framework as the working methodology, ISO 42001 as a certifiable management system, and the EU AI Act as binding, risk-tiered regulation, increasingly used together through published crosswalks, with cross-functional governance committees and, in the most recent guidance, dedicated bodies for agentic systems; in regulated sectors it

is denser still, with sector regimes such as the United States Treasury's 2026 financial-services AI risk framework layered on top. Even this ground is moving: the EU AI Act's high-risk timeline is itself under revision as this is written. Boards are increasingly expected to attest to the firm's AI risk posture. This paper does not duplicate any of it.

The gap is at the level of the portfolio, not the individual system. The maturing stack governs systems one at a time; it does not aggregate. A firm can run a hundred well-governed AI systems and still have no view of its total exposure, and the danger in a large program is precisely the aggregate: correlated failures, shared dependencies, and concentration no single-system review is built to see. That this aggregate risk is real is no longer only the argument of this paper: through late 2025 and early 2026, systemic-risk authorities including the European Systemic Risk Board and the Bank for International Settlements warned that model monoculture, the reliance of many institutions on the same few systems, could turn a single shock into correlated failures across all of them, which is the same concentration risk seen from above. Without a portfolio measure there is also no scarcity, and so no forcing function: each request looks reasonable alone, every reasonable request is approved, and risk accumulates with no one accountable for the total. Because the model layer keeps producing new failure modes faster than any catalog can close them, the portfolio view must treat its own risk taxonomy as provisional. Supplying that aggregate view, feeding it back into the existing governance regime, and keeping it current is this paper's contribution on the risk side.

The exposure can also shift from outside the firm, faster than any internal process. In June 2026 a United States export-control directive required Anthropic to disable two of its most capable models, Mythos 5 and Fable 5, for all customers including inside the United States within hours; its other models were unaffected. Merits aside, any firm built on those two models saw a dependency that was stable on Friday gone over the weekend, by act of government rather than technical fault. This is risk the cloud era rarely had to price: a model's availability is not a constant but a nonstationary variable exposed to regulation, geopolitics, and a vendor's fortunes, and it is correlated, because one directive can remove every workload sharing the dependency at once. A portfolio view therefore treats the model layer as something that can be withdrawn, not only something that can fail, and asks where shared model and vendor dependencies quietly concentrate that exposure.

1.5 The ground is moving, and the cloud-era playbook does not fit

The deepest gap is in the frame itself: the discipline many firms reach for is the one they learned in the cloud era, account for usage after the fact, optimize unit cost, and govern each system for compliance; even firms with mature risk functions tend to apply it one system at a time. That playbook is insufficient here, for three reasons that compound. First, the locus of action is moving up the stack, from infrastructure to models to agents, and each step makes behavior less deterministic, so a method built on stable unit costs and predictable usage misstates both the bill and the risk. This is the shift from systems that talk to systems that act, the change every major research firm now treats as the central one of 2026, and it is what makes a bounded-worst-case screen and a modeled cost distribution non-optional rather than refinements; the established governance standards were not built for it. Second, a growing share of AI investments behave less like fixed assets than like options, and where they do, their payoffs are contingent, asymmetric, and path-dependent, and much of their value lies in the choices they keep open, the right to scale a pilot, switch a model, or abandon a line cheaply, which takes a different toolkit from depreciating an asset. Third, the risk and market environment changes fundamentally and fast, not at the pace of a planning cycle, as new model-layer risks appear, vendor economics move, and capabilities shift under the portfolio; a model rigid in the wrong places will be precise about a world that no longer exists.

The implication is a deliberate tension the rest of the paper carries: disciplined enough to allocate capital and bound risk, yet built to absorb rapid change in what the risks and economics even are. And it must price one quantity the cloud era could largely ignore, manageability, the ability to change course, switch, throttle, or exit without heavy cost. In finance this is liquidity, measured and priced; in AI it is mostly unmeasured, and in a fast-moving regime it is what a portfolio needs most. It is an allocation variable, not an engineering preference.

1.6 Where this sits in existing work

The model does not invent its parts. It integrates established ones and adds the layer they leave out. Three bodies of work stand most directly behind it.

The first is real-options theory, which holds that under uncertainty the right to act later can be worth more than acting now: there is value in waiting for favorable signals, and value in the option to expand once they arrive. The foundational statements are McDonald and Siegel's 1986 analysis of the value of waiting to invest and Dixit and Pindyck's 1994 treatment of investment under uncertainty. This is the root of the decision vocabulary developed in Part III, where a workload funded as cheap exposure is an option bought, one held against a named trigger is the value of waiting made explicit, and one exercised now is an option whose signals have already resolved. The contribution here is not the theory, which is settled, but its operationalization for AI workloads.

The second is agency theory, founded in Jensen and Meckling's 1976 analysis of what happens when decisions are delegated to agents whose information and incentives differ from the owners', and of the monitoring and bonding that delegation then requires. Part IV is that theory applied to AI capital. The sponsor who owns the value and risk assumptions is the informed agent; the recorded claim and the ex post loop are monitoring; the published acceptance criteria are a bonding device, letting a team demonstrate in advance that it is building what the firm would fund; the evaluator is kept independent because monitoring by an interested party is worth little; and the concession that no incentive is gameproof is the residual loss the theory predicts, the agency cost that survives even well-designed governance. Here too the contribution is not the theory but the observation that AI capital currently flows with almost none of this machinery attached, inside organizations that apply all of it to every other form of delegated capital.

The third is the fast-growing buyer-side literature on evaluating AI investments, which has matured quickly through 2026. Recent treatments score a single investment through strategic readiness, a value-driver decomposition, staged gates, and a risk-adjusted hurdle, and treat vendor dependency and organizational friction as first-class risks rather than afterthoughts: the survey-based account of what drives returns on AI by Davenport and Srinivasan in the Harvard Business Review, the Wharton Executive Education work on why most AI investment is not yet paying off, and a recent working paper proposing a risk-adjusted AI return framework that integrates a management standard and regulatory exposure and presents returns as distributions rather than point estimates. This model shares their premises: that AI return is a structured argument rather than a single number, that organizational readiness is decisive, and that build-versus-buy is a risk trade rather than a sticker-price choice. The difference is one of altitude. The per-investment treatments stop at the single decision, and several say so, leaving portfolio correlation, a loss-denominated budget, and the valuation of option-like investments to later work. This model begins there: it takes the portfolio as the unit of governance and treats each single investment as its marginal contribution to the book. One methodological choice follows from that difference. A per-investment framework that builds a discount rate by adding risk premia lets a sufficiently strong cash flow clear a high rate, so a high enough return

can in principle buy its way past almost any risk; this model keeps the worst-case screen separate and absolute, because a fiduciary cannot let strong economics purchase an unbounded tail.

The broader practitioner landscape has moved as well, and fast, and by mid-2026 the convergence is explicit. BCG now tells boards directly that their job is to make sure AI investment is treated as a portfolio rather than a collection of projects, with the mix across deployment, reshaping, and invention held deliberate. Gartner's finance practice tells CFOs the same thing in allocation language: AI is a portfolio of very different bets, no single ROI formula spans them, and progress should be measured by realized value rather than deployment volume. The largest private-equity owners of software run a live version, most visibly Vista Equity Partners, reported in late 2025 to score every portfolio company on AI adoption velocity and to tie those scores to future capital allocation, while its peers push portfolio-wide adoption through central coordination and pooled hyperscaler terms. The useful sponsor distinction is between AI acceleration and AI allocation: shared engineering capacity, preferred model access, and pooled infrastructure accelerate every portfolio company at once, and they compound the same dependencies across the portfolio, so the allocation question survives the factory intact: which workloads earn funding, where concentration is accumulating across companies, and what evidence shows value maturing toward exit-relevant performance rather than scattering across demonstrations.

So the portfolio framing is no longer the distinctive claim it once was; it is the consensus position of the consultancies, the analysts, and the most sophisticated owners. What that consensus still omits is the portfolio-level layer named in the summary and built in Part IV: the correlations between investments, a loss-denominated budget and hurdle, and an absolute floor under the worst case. The sharpest way to see the gap is the failure budget now circulating in practitioner guidance, an assumed miss rate on the order of forty to fifty percent across initiatives. That is a useful number, and the model absorbs it, but it is a statement of frequency: how often the book is expected to miss. A risk budget is a statement of severity: how much loss capacity the book may consume when it does. A portfolio can run exactly at its assumed failure rate and still breach the firm's capacity on a single unbounded miss, which is why the two are carried separately here, the frequency in the ex post loop that compares outcomes with claims, the severity in the budget and the floor, and why neither substitutes for the other.

More specific than the general convergence is Gartner's, which sorts AI initiatives by strategic intent into three cases it names defend, extend, and upend: spend that holds competitive parity, investment that differentiates an existing process for return, and bets that aim to reshape an industry. The cut is a sound one, and this paper adopts it under sharper names, Parity, Advantage, and Frontier, because the intent of an initiative is the first thing a board should fix and it shapes how the initiative ought to be judged. What an intent label does not do is set the terms of the commitment. It does not say that the three cases call for three different kinds of limit, a capped spend pool for parity, a loss-denominated risk budget for advantage, and a bounded option premium for frontier, which is the distinction Part IV draws and the one most firms miss when they run all of their AI spend against a single ceiling. The taxonomy tells a board which game it is playing; the model tells it how much to commit, against which limit, and on what terms.

One further body of work sits beneath all of this rather than beside it: the cost-standardization movement in FinOps. Through 2026 the open specification for cloud and AI cost, FOCUS, extended to normalize AI billing and token economics across providers, and the FinOps discipline broadened its remit from the value of cloud to the value of technology. This model depends on that layer and does not duplicate it. A cost standard makes spend comparable; it does not aim to express what a

workload is worth, what risk it carries, or how those aggregate across a book, which is what Parts III and IV supply. The place of this paper is the value-and-risk layer above a normalized cost substrate: it takes the accounting as increasingly given and concerns itself with the allocation and exposure the accounting cannot express. The substrate makes the spend legible; this layer decides whether the spend deserves capital. Figure 4 places the layer visually: a normalized cost substrate at the base, the system-level control frameworks above it, board oversight at the top, and between controls and board the portfolio allocation layer this paper supplies.

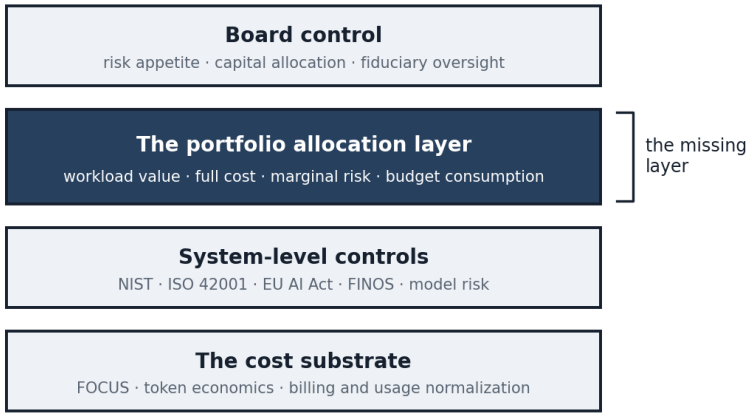

Figure 4. The missing layer. The stack as it exists and the gap this paper fills. At the base, the cost substrate: FOCUS and token economics normalizing what was consumed, by whom, through which provider, at what unit cost. Above it, the system-level controls: NIST AI RMF, ISO/IEC 42001, the EU AI Act, FINOS, and internal model risk, governing whether each system is fit to run. At the top, board control: risk appetite, capital allocation, fiduciary oversight. Between controls and board sits the portfolio allocation layer, where workloads are valued, fully costed, assessed for marginal correlation-aware risk, and admitted against a budget. The layers below and above are necessary; the decision layer between them is what has been missing.

1.7 What it costs to get this wrong

These gaps are not academic. Capital is misallocated when the loudest sponsor rather than the best risk-adjusted return wins the funding. The mechanism is an old one: Jensen's 1986 free cash flow argument holds that firms overinvest exactly where cash is abundant and monitoring is weak, and AI today reproduces that condition at scale, ample cash, an exciting use for it, and almost no portfolio-level check on what the spending earns. Risk accumulates unseen when no one holds the total. Value is claimed but never realized when the wrong outcome is measured. And the board, which carries the duty, cannot answer the two questions it will eventually be asked: what its AI is actually returning, and how much risk the firm is running to earn it.

One objection deserves a direct answer. A serious view holds that it is simply too early to demand returns at all, that AI follows the familiar lag of a general-purpose technology, in which productivity trails investment for years, and that revenue is in any case now accelerating quickly. There is truth in it. But the conclusion does not follow, because the model does not demand a payback date. It is built precisely for value that is contingent and slow: it funds option-like workloads as cheap exposure, stages commitments so that capital follows evidence, and treats the option to wait as something worth paying for. What it refuses is not patience but undisciplined allocation and an unbounded downside, which are the wrong posture whether the returns arrive early or late. Discipline and

patience are not opposites.

The model that follows is built so that a board can answer both, and defend the answers.

Part II. Principles

Before the mechanics, the model rests on a small set of commitments. They are the core of what the field can converge on, and everything in the operating model is an attempt to honor them. They describe the discipline the model aims at, and several are easier to state than to compute; Part VI is explicit about which of them remain on the open frontier rather than settled, so that the commitments are read as direction and not as solved problems.

- **Treat AI spend as capital under uncertainty.** AI investment is not an IT operating line. It is capital deployed for an uncertain return and should face the discipline other capital faces.
- **Direct capital to its best risk-adjusted use, within the limit.** The limit is not only a cap but an enabler: a firm that knows the bound it must stay within can back its strongest opportunities with conviction rather than hedge thinly across the board.
- **Manage the portfolio, not the project.** The unit of allocation, governance, and risk is the portfolio of AI investments, not any single use case.
- **Separate the activity from how it pays, and respect the shape of the payoff.** Value accrues through distinct pathways. Keep them distinct rather than collapsing them into one number, so that the shape of each payoff, including the convexity and optionality that define a growing share of these investments, survives into the decision.
- **Cost the whole thing, not the sticker.** A workload's cost is fully burdened and adjusted for utilization, or it is wrong, and the assumptions that drive it, above all useful life, are surfaced and tested.
- **Budget risk explicitly, and at the top.** The board sets the firm's appetite for AI tail risk; the risk committee expresses it as a bounded budget anchored to the firm's capacity to absorb loss, allocated across workloads and consumed visibly.
- **Prefer acknowledged uncertainty to false precision.** Where a quantity cannot be known precisely, use a defensible relative measure with a stated provenance and a stress test, not a falsely precise point estimate.
- **Build for a moving target, and for different organizations.** Because the risks and economics keep changing, and because a regulated bank and a fast-moving software company need different settings, the governance half is built in two layers: an invariant core that holds for every organization, and an adjustment layer that calibrates that core to the organization's own characteristics. It absorbs both regime change and organizational difference without being rebuilt.
- **Value flexibility, not only return.** Treat manageability, the firm's liquidity in its AI positions, as a quantity to be measured and priced rather than left implicit, recognizing that a measure comparable across workloads is still being sharpened. In a fast-moving regime, flexibility is worth paying for.
- **Shift governance left.** When teams know in advance what makes a workload fundable, they spend design effort only on what can pass and build the qualities that earn approval, flexibility and low risk, in from the start rather than retrofitting them.
- **Align incentives with the criteria.** Publishing the standard changes behavior only if people are rewarded for meeting it. Tie the incentives of the teams that build and run workloads to approvable outcomes delivered within budget, not to raw activity, consumption, or velocity, so

that shifting governance left works in practice rather than only on paper. No reward is immune to gaming, including this one, so it is the ex post check of promises against results, not the incentive alone, that ultimately holds the standard in place.

- **Keep the evaluator independent.** The judgment of value and risk should have no stake in which way any recommendation falls; a model for allocating capital and governing risk is only as trustworthy as the independence of the party applying it.

Part III. The model, part one: allocation

Allocation is where the return is won or lost: it is the half of the model that decides which uses of capital earn their place and how much to commit to each. Figure 5 traces the model as a whole: one workload's path from candidate to the ex post loop, inside the class mix and risk budget the board sets. This part builds the allocation half of that path; Part IV builds the governance half.

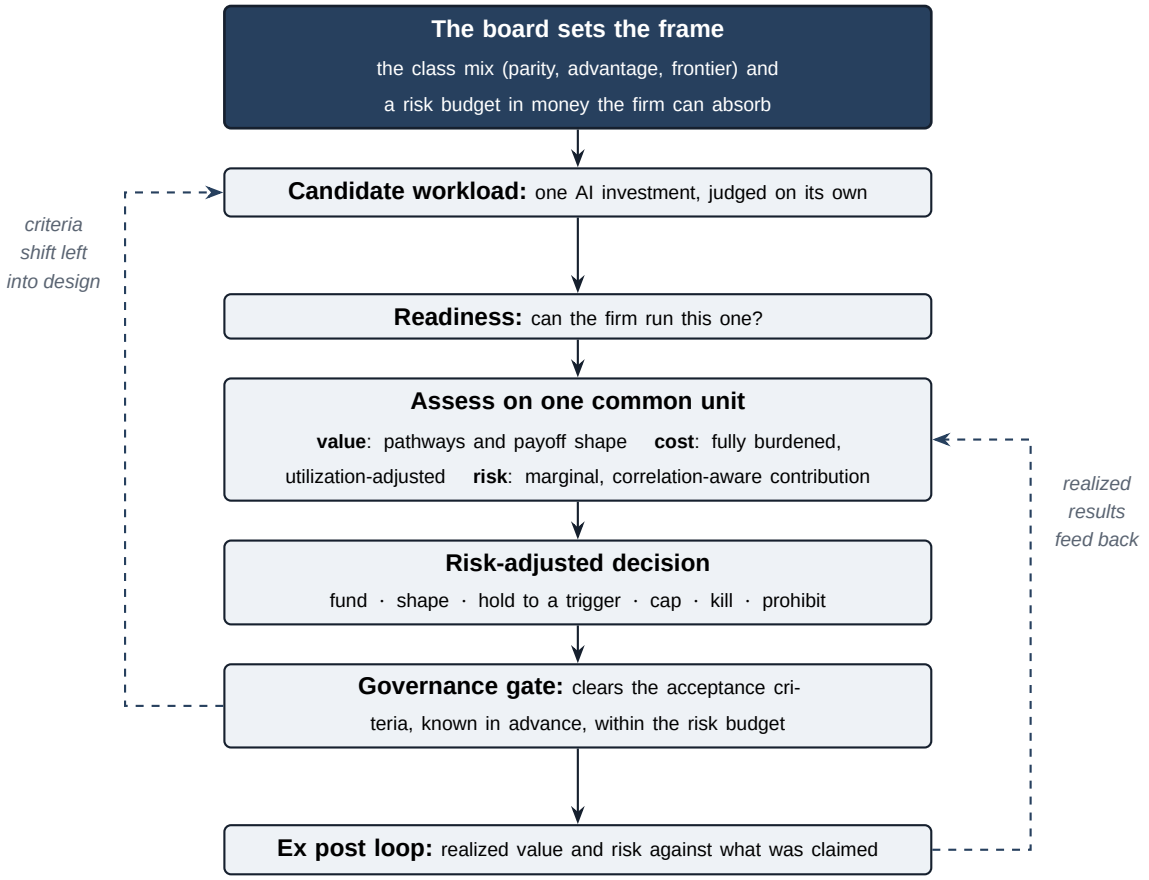

Figure 5. The model end to end: one workload's path through the allocation half and the governance half, within the class mix and risk budget the board sets.

Start with a single decision, because the portfolio is built from many of them. You already hold a set of AI investments. In front of you is one more candidate. Three questions decide it.

What does it pay, and in what shape? Not a single saving but the blend of pathways it runs, with the convex and the linear kept apart rather than averaged into one figure. What does it truly cost per

useful outcome, fully burdened and adjusted for how busy the underlying capacity actually is, and carried as a range when the usage is uncertain rather than as a single point? And what does it do to the risk you already carry: does it diversify the book, or does it lean harder on a model, a vendor, or a failure mode you are already exposed to? A cheap, high-value workload that concentrates an exposure you already hold can be worth less to you than a dearer one that spreads it, because what matters is the contribution to the whole, not the workload seen alone.

Answer those three for one candidate and you have an allocation move: fund it, shape it, or set it aside. Answer them the same way for every candidate, and bound the sum against what the firm can afford to lose, and you have the portfolio and its budget. The rest of this part first screens each candidate for readiness, then builds the three answers, value, cost, and risk, on a common unit, and ends in a decision; the next part bounds their sum.

3.1 The unit of analysis

The unit is the workload: a discrete AI investment that can be funded, measured, and governed on its own, whether that is a customer-service assistant, a document-processing pipeline, or an autonomous agent embedded in a business process. Each workload is characterized four ways: by the value pathways through which it pays, by its fully burdened and utilization-adjusted cost, by its contribution to the firm's portfolio risk, and by its manageability, how easily it can be changed, switched, throttled, or unwound if conditions move. The discipline is to capture the workload broadly, recording the several ways it might create value and consume risk, and then to project only the dimensions that matter for the decision at hand. The boundary is fixed by rule rather than by convenience: one purpose with one accountable owner is one workload, a shared system serving several purposes splits into several, and a multi-step pipeline serving one purpose stays one. The rule matters beyond tidiness, because the counts and concentration gauges Part IV builds on this unit are only as firm as the unit itself, and a book should not be able to look more or less diversified by redrawing its lines. A workload is not a model and not a vendor. It is a use of capital.

3.2 Readiness: can the firm run this one?

Before a workload is valued or costed, it faces a question the economics cannot answer: can this organization actually execute and absorb this particular investment? The evidence that this is the binding constraint is by now strong, and it is organizational rather than technical. The dominant reasons AI investments fail are that the data the workload needs is not ready, the process it touches is automated rather than redesigned, no one owns the outcome, or the people whose work must change were never brought along. A workload with excellent projected economics and no organizational footing is not a good investment; it is a good investment that some other firm is positioned to make. The readiness screen asks, for this workload specifically, whether the data it depends on exists and is usable, whether the process around it has been redesigned rather than merely overlaid with a model, whether a named owner is accountable for the outcome against a recorded baseline, and whether the firm can absorb the change at the pace the workload assumes. A workload that fails the screen is not killed. Its readiness becomes the first thing to fund, and the workload itself is sequenced behind it. The cheapest way to avoid a failed investment is to decline to build it until the organization can carry it, and to say so before design begins rather than discover it at the gate.

3.3 Valuing the workload

Value is expressed as the net contribution per unit of work, assembled across the workload's live pathways rather than reduced to one. A controlled vocabulary of pathways, set out in Appendix A, keeps this disciplined: direct cost reduction, revenue generation, revenue protection, loss or risk avoidance, capacity or throughput unlock, and capital efficiency. Each pathway carries its own outcome metric, its own direction, and, importantly, its own shape, because a linear cost saving and a convex revenue option behave differently as volume grows and should not be averaged into a single figure that hides the difference. Kept distinct does not mean left uncashed. Every pathway ultimately moves one of a small set of financial drivers, revenue, margin, capital efficiency, or the cost of risk, and each claim names the driver it moves and the translation that connects the system's measured performance to it, a deflection rate to a cost per contact, a lift to a margin, an error rate to a loss frequency. The refusal is of a single blended score, not of cash: a pathway that cannot state its route to a financial statement is not a value claim but a hope.

Two rules keep the valuation grounded. The first is that the pathways are kept distinct through the analysis and combined, if at all, only at the very end, so that the payoff shape and any optionality survive into the decision. The second is that the sponsor of the workload owns the value assumptions and states them explicitly. Putting the burden of proof on the requestor, and recording what was promised, is what later makes it possible to compare what was claimed with what was realized. Some workloads are best understood not as a fixed stream of value but as an option: a small outlay that buys the right, not the obligation, to scale later if the technology or the market cooperates. Where that is so, the value is in the optionality, and the model preserves it rather than discounting it to a single expected figure. The label carries an obligation, though. An option claim is admissible only with its discipline attached, a capped premium the firm is prepared to lose and a dated trigger at which the position is rescored or closed, the terms the decision vocabulary in 3.6 and the frontier limit in Part IV enforce, so that optionality names a bounded position rather than sheltering whatever a sponsor prefers not to justify. One more test guards the label. Option value accrues to differential position: an option written on capabilities every competitor can buy from the same vendors is an option everyone holds, and an option everyone holds is worth little to anyone. The premium is worth paying only where the firm brings something rivals cannot copy to the upside, proprietary data, distribution, an installed base, a regulatory position, and the claim states what that something is.

One further discipline turns a value estimate into something a single team can act on: naming the assumption the case rests on. A workload's projected value usually turns on one or two load-bearing assumptions, a churn reduction, a deflection rate, an adoption level, a volume, and the rest of the model is comparatively inert around them. The discipline is to identify those assumptions explicitly, size how much of the case depends on each, and test the most consequential one first, before the full build is committed. Doing so tells the team what to prove and in what order, and it is what later lets the ex post loop ask not merely whether the workload paid off but whether the assumption it rested on held.

3.4 Costing the workload

Cost is built as a chain, not a quote. It runs from the total cost of ownership, through a fully burdened hourly rate for the underlying capacity, to an effective cost adjusted for utilization, and finally to a cost per unit of useful work. The two haircuts that cloud pricing conceals are made explicit at the third step: how much of the time the capacity is occupied, and how efficiently it runs when it

is. For owned infrastructure this is the difference between a defensible number and a fiction. For rented capacity the same logic applies in reverse, since a per-unit price already embeds someone else's occupancy and margin, and the question becomes whether the firm can do better at its own expected utilization. Build versus rent is then a calculation on the firm's own numbers rather than an industry rule of thumb, with the useful-life assumption surfaced and tested for sensitivity, because it moves the answer more than almost anything else. For agentic workloads the unit cost is itself a distribution rather than a point, because the agent decides at run time how much work to do, so the model carries the expected cost and the tail of the cost together and prices the tail rather than hoping it away. The same energy data that drives these costs also drives carbon, so the chain is carbon-aware by construction: every cost figure has an emissions counterpart, available wherever a reader wants it.

Two cost profiles sit inside that chain and behave differently enough that the model keeps them apart. The front-loaded build, the training or fine-tuning, the integration, and the data and evaluation work, is lumpy and committed in advance of any value, and it is largely sunk if the workload is later abandoned; it behaves like a capital investment. The ongoing inference is usage-driven, scales with adoption, and can be shed by throttling or stopping; it behaves like an operating expense. Treating the two as interchangeable obscures the real cost drivers and misleads both the forecast and the benchmark, so the model gives them different units and reads the workload's cost two ways, once for each decision it informs. The fully loaded cost per unit of work, which carries the amortized front-loaded build alongside the running cost, answers whether the workload should exist at all and whether the whole investment cleared. The marginal cost, the usage-driven part alone, answers whether to serve more of it. Because the front-loaded component is never negative, the fully loaded figure sits at or above the marginal one, so the two can part in only one direction: a workload can be cheaper than its human alternative at the margin and still be underwater once the build is counted. That is the case a low per-call number quietly talks a firm into scaling, and naming the two figures separately keeps the decision from resting on the wrong one. The utilization adjustment carries one caution of its own. An allocated fixed cost divided by falling utilization produces a rising effective rate, which discourages use and lowers utilization further, the death spiral cost allocation is known for. The guard is the division of labor just drawn: the utilization-adjusted, fully loaded figure informs whether the workload should exist and how the capacity was judged, never the marginal decision to serve more, which reads only the usage-driven cost. It also puts the front-loaded build where it belongs, as committed capital in the risk view rather than a line smeared across a run-rate, which is why the cheapest moment to stop a workload is before that capital is committed.

3.5 Risk-adjusting the workload

A workload's value and cost are not enough. What it adds to the firm's risk has to enter the comparison. Before value and risk are even weighed against each other, the workload faces a separate screen the economics cannot buy down: what is the worst it can do unnoticed, and is that bounded? One whose worst case is unbounded, an irreversible action taken autonomously at systemic scale with no floor beneath it, is set aside until a floor exists, however strong its return, because strong economics do not purchase an unbounded tail. That hard limit under the worst case is the floor, the one line no return is permitted to buy through. The screen is not run once, either. What a workload may do without supervision is a boundary that moves with evidence, widening as reliability is demonstrated in production and narrowing on drift or a changed blast radius, so autonomy is held as a monitored variable in the register, re-examined on the same triggers that reopen the decision, rather than a box ticked at intake. Past that screen, the model expresses a workload's contribution to portfolio tail risk

relatively and with a stated provenance rather than as a falsely precise loss figure, for the simple reason that a credible point estimate of tail risk usually does not exist. The decision then turns on a marginal test: the workload must clear a risk-adjusted-value hurdle, earning enough expected value for the risk it adds, rather than merely fitting inside whatever budget remains. A cheap, high-value workload that nonetheless concentrates the firm dangerously can fail this test; a more expensive one that diversifies the portfolio can pass it. Manageability enters here too: a commitment that is hard to unwind, a deep vendor lock-in, a large sunk build, an irreversible process change, carries a penalty that a flexible alternative does not, because in a fast-moving regime the ability to change course is itself worth paying for. This is what it means to allocate as a portfolio rather than a queue.

3.6 The decision

The output is a decision, recorded with the assumptions behind it so the governance half can later check what was claimed against what was realized. But a simple yes or no is too blunt for investments that increasingly behave like options, so the model decides in a vocabulary that fits their shape:

Decision	When it applies	What it commits, and the discipline attached
Exercise now	Strong, measured economics with the controls already in place	Full commitment at scale; the claim recorded, the envelope set
Fund as cheap exposure	Value real but contingent on the technology, the market, or evidence still to arrive	A capped premium the firm is prepared to lose, a dated trigger, and the rate of learning tracked
Hold to a trigger	Worth making only once a dependency lands, a baseline is recorded, or capacity frees up	Nothing yet; the trigger written down in advance rather than vaguely deferred
Cap	Good economics with a control still missing	Current scale only; the control precedes any growth
Kill	Fully burdened economics that do not clear the bar	The budget and capacity returned to the book; capital already spent has no vote
Prohibit	A worst case that is unbounded	Nothing, regardless of economics; the floor is not for sale
Fund as readiness	The main return is raising the value of every other option	A reusable capability underwritten on that basis rather than on its own cash

The recommendations are not one-time labels. An investment funded as cheap exposure, or held against a named trigger, is by construction staged: a small first commitment that buys information, rescored when the information arrives, which is the value of waiting under uncertainty made operational. Each stage is gated by what the prior stage was meant to prove, the load-bearing assumption first, and the recommendation is rerun at each gate rather than fixed at the outset, because a workload's readiness, economics, and risk all move as it is built. This is the discipline of stage-gated development applied to a single investment: the conditions that would advance a workload to its next stage, and the conditions that would instead cap it, kill it, or send it back to a trigger, are written down in advance, so that the decision is a guide the team builds toward rather than a recommendation delivered at the end.

A board does not need the full method to see when a workload has stopped earning its place. A small number of observable conditions reliably signal that a workload should be sent back to the decision, and naming them in advance, before any one of them trips, is what keeps the eventual recommendation from being relitigated under pressure from a sponsor who has grown attached to the work. This is a commitment problem as much as a measurement one. A limit the firm can be argued out of at the moment it binds was never a limit, and the trigger published in advance is the device by which the board holds, under pressure, the line it set while calm. Each is a trigger for review rather than an automatic sentence; what the review concludes still depends on the workload's value, its readiness, and its place in the book.

- Automation that does not absorb the work: when the share of cases still handled by a person stays high and does not fall across successive iterations, the workload is not taking on the volume it was funded to take on, and the modeled value is not being realized. A high share that is still falling is a question of readiness and trajectory; a high share that has gone flat is a structural weakness in the value case.
- Inverted unit economics: when the fully counted cost of completing a unit of work with the system, including review, exception handling, infrastructure, and retraining, sits above the cost of the pre-AI baseline, the workload is destroying value at the margin. The only reason to continue is deliberate, to hold cheap exposure to a falling cost curve, in which case the workload is funded as an option and labeled as one rather than carried as though it were paying its way.
- Revealed rejection by users: when the people a workload is meant to serve route around it, or use it only when required to, the workflow fit is weak, and capability that is not adopted returns nothing however good the model is. Where the friction is fixable the recommendation is readiness; where the work simply does not want the system, it is a kill.
- Controls that erase the benefit: when the oversight required to make a workload safe consumes the speed or cost advantage that justified it, the net value for this use case is gone, even though the same model may earn its place in a lower-severity use elsewhere. This is the severity axis doing its work: the more consequential the failure, the heavier the control, and past some point the control cost overtakes the benefit.
- Value that does not survive production: when performance is strong in the pilot and weak in live operation, the gap is the cost of the edge cases the pilot never saw. If the gap can be closed, the recommendation is readiness; if it is structural, it is a kill.
- Capital without movement: when a workload has absorbed repeated rounds of funding without moving its value or its readiness, continued funding is no longer an investment in the outcome but a payment on the sunk cost. This is the trap that keeps weak workloads alive, and it is the clearest case for a kill.

A kill, in a portfolio, is not a failure. It is the release of scarce capacity, the talent, integration bandwidth, and governance attention the workload was consuming, back to the parts of the book that can return more on it. The discipline that names these signals in advance is the same discipline that makes the release defensible when it comes. The commonest return trip through this machinery, though, is the opposite case: a workload that is succeeding and asking to grow. Illustration 6 in Appendix C works that expansion decision through, from the unit figure everyone quotes to the margin, the claim, the boundary, and the book.

3.7 Reconciling the two readings

Two readings of the same investment have to agree. Bottom-up, the workload stands or falls on its own readiness, value, fully burdened cost, and bounded worst case. Top-down, it is judged by what it adds to the book: a workload with strong standalone economics that leans on a model, a vendor, or a failure mode the firm is already exposed to faces a higher hurdle than a dearer one that diversifies the portfolio. When the two readings disagree, the disagreement is information rather than noise. A workload that looks excellent on its own but fails on its marginal contribution is a concentration the firm has not yet priced. A workload that looks marginal on its own but materially diversifies the book may be worth funding for what it protects. Neither reading is sufficient without the other, which is why the model runs both and requires them to reconcile before capital moves. Illustration 2 in Appendix C makes this concrete, including why diversification that holds in calm periods cannot be banked against the tail.

Part IV. The model, part two: governance

The governance half exists so the firm can act on the allocation half with confidence: put real capital behind its best AI investments while keeping the total risk inside a bound the board has set. It does not replace the firm's existing AI risk regime; it supplies the portfolio-level layer that regime lacks, and feeds the result back into it. It has two jobs: to make the firm's total AI risk visible and bounded, and to gate new investment against that bound, so that the capital which clears is capital the firm can afford to commit.

4.1 Core and fit: one model, many organizations

The firms that will run this model are not alike. A globally systemic bank, a fast-moving software company, and a private-equity owner managing a portfolio of both sit in genuinely different places, and a single prescribed process calibrated for one of them will be wrong for the others: a brake on the fast mover, or a dangerous looseness for the regulated one. The governance half is therefore built in two layers. A core holds for every organization and does not move. An adjustment layer calibrates the settings of that core to the organization's own characteristics. The discipline of the split is that the flexibility lives entirely in the settings and never in the core, because a model that lets an organization tune away its fiduciary obligations is not a model worth having.

The core is small, and it is stated plainly so that it can be held to. For every organization, whatever its type: a risk budget exists, grounded in board-owned appetite and anchored in loss terms to the firm's capacity to absorb loss; new investment is allocated by its marginal, correlation-aware contribution to the portfolio rather than by order of arrival; the worst-case screen that prohibits an unbounded tail is absolute; the party that applies the method is independent of the outcome it measures; the acceptance criteria are published before design rather than revealed at the gate; and realized value and risk are checked against what was claimed. These are the invariants, and the adjustment layer cannot reach them.

What the adjustment layer sets is the calibration: how conservative the budget is, how low the prohibit threshold sits, how large the central reserve is, how heavy the approval path is, how often the portfolio is stress tested, and how much weight the hurdle places on the ability to change course. Two characteristics of the organization drive most of this, and they are kept separate because they pull in different directions. The first is the severity of failure: how regulated, how systemic, and how irreversible the consequences of a bad outcome are. High severity argues for a lower prohibit thresh-

old, a larger reserve, a more conservative budget, and more formal stress testing and independent challenge. The second is the clock speed of the organization's environment: how fast capability and competition move, and how cheaply the firm can reverse a commitment. A fast clock argues for a lighter approval path, a hurdle that rewards optionality and manageability, more granular staged commitments, and a more frequent but lighter review rhythm. A regulated bank sits at high severity and a slow clock; a growth-stage software company at lower severity and a fast clock; a frontier-model builder at high severity and a fast clock at once, which is exactly the combination a single axis would miss, and the reason two are needed rather than one.

The settings are derived from the organization's own characteristics rather than chosen from preference, in the same spirit in which the cost chain is built on the firm's own numbers. The budget level is anchored to the firm's capacity to absorb loss; the prohibit threshold tracks its regulatory and systemic exposure; the review cadence tracks how fast its portfolio is actually moving. A board that sets its parameters this way can defend why they are what they are, which is the point of deriving them rather than declaring them. For an organization that wants a starting point rather than a blank parameter sheet, a small set of named profiles bundles sensible settings for the recognizable positions, to be tuned from there.

Two cautions keep the layer from sprawling. The organization's position on this spectrum is a matter of its enduring type, not of how far along its adoption journey it has traveled; maturity is a path over time, handled in the operating model, and collapsing the two into a single notion of sophistication would blur both. And the difference between buying and building AI, or how centralized an organization's decision rights are, is mostly handled already at the level of the individual workload, in the build-versus-rent calculation and the dependency and manageability scoring, so it does not need to become a third axis. Two characteristics that visibly move different settings are the right amount of structure; more becomes a matrix that impresses on a slide and helps no one.

One organization type deserves a note, because it does not sit at a single point. A private-equity owner runs this model across a portfolio of companies, each at its own position on both axes, with a fund-level aggregation above them. That is the same model applied as a portfolio of portfolios, and it is a strength of the structure rather than a special case to apologize for. Finally, the precondition beneath the rest: an organization with no recorded baselines and no figure for its capacity to absorb loss is not yet ready to run even the core, which is the organization-level form of the readiness question the allocation half asks of every workload.

4.2 The portfolio risk budget

At the center is a risk budget: a firm-wide limit on AI tail risk, grounded in the board's risk appetite, anchored to the firm's capacity to absorb loss, and operationalized by the risk committee as a hard cap the existing governance regime can adopt as its portfolio-level control. A word is needed on what denomination in loss can and cannot mean here. The budget is a ceiling expressed in loss terms because that is the language in which a board sets appetite and a firm knows its capacity to absorb loss; it is not a claim that the portfolio's tail loss can be computed to a precise figure, which Part III has already said it usually cannot. The ceiling is anchored to loss-absorbing capacity and to stress scenarios rather than to a point estimate, and workloads consume shares of it by their relative, correlation-aware contribution rather than by a summed dollar figure. The construct is not new to finance. It borrows openly from risk budgeting and economic capital, where an institution decides in advance how much risk it will run and allocates that allowance across activities, and from the engineering practice of the error budget, where a team is given an explicit allowance for failure

precisely so that it can move quickly within it. The point of a budget is not to minimize risk. It is to make risk a deliberate, bounded choice. A fair question from finance is why the firm should bound this risk at all, when diversified shareholders can carry it. The answer is the established one from the corporate risk-management literature that runs through Smith and Stulz's 1985 analysis: hedging and limits earn their place through the frictions, the costs of financial distress, the investment a damaged balance sheet forgoes, the claims of employees, customers, and regulators who cannot diversify, and the franchise value a tail event destroys. The budget is not a taste for safety. It protects the firm's capacity to keep investing, which is the same objective the allocation half serves.

Where the capacity number comes from varies with the firm, and the variation is ordinary rather than a weakness. A regulated financial reads it off machinery it already runs, the economic-capital and capital-adequacy work that states risk-bearing capacity today, with the AI budget carved largely from its operational and strategic slice. A non-financial firm has no such artifact, but it holds the parts: covenant headroom restated as the loss that would breach it, liquidity above the minimum cash operations require, the buffer above a rating threshold or the loss the dividend could bear, with capacity set by whichever constraint binds first over the horizon, the same quantity enterprise-risk practice calls risk capacity. And where none of that analysis exists yet, the construct still works at its crudest, as a question a board can answer in one sitting: what is the largest loss from this program that would not change the firm's plans? That figure, named conservatively and sanity-checked against yardsticks the firm already watches, months of free cash flow, a share of covenant headroom, is capacity enough to start, because consumption is measured in relative shares, and an approximate anchor disciplines allocation long before it is precise. The translation belongs to whichever seat the adjustment layer assigns it, a risk committee where one exists, the audit committee or the CFO's office where one does not, and the number refreshes on the planning cycle, since capacity moves when earnings, buybacks, or refinancing move it.

Figure 6 draws the construct as a load line: the budget a small slice of the firm's capacity, the standalone book as cargo, the charge for shared dependencies as the water itself, rising from calm toward stress inside the hull, and headroom as what remains below the gunwale. The gap between what the book already consumes and the cap is the firm's headroom for AI: how much more capital it can commit within its own appetite, and what the allocation half works with.

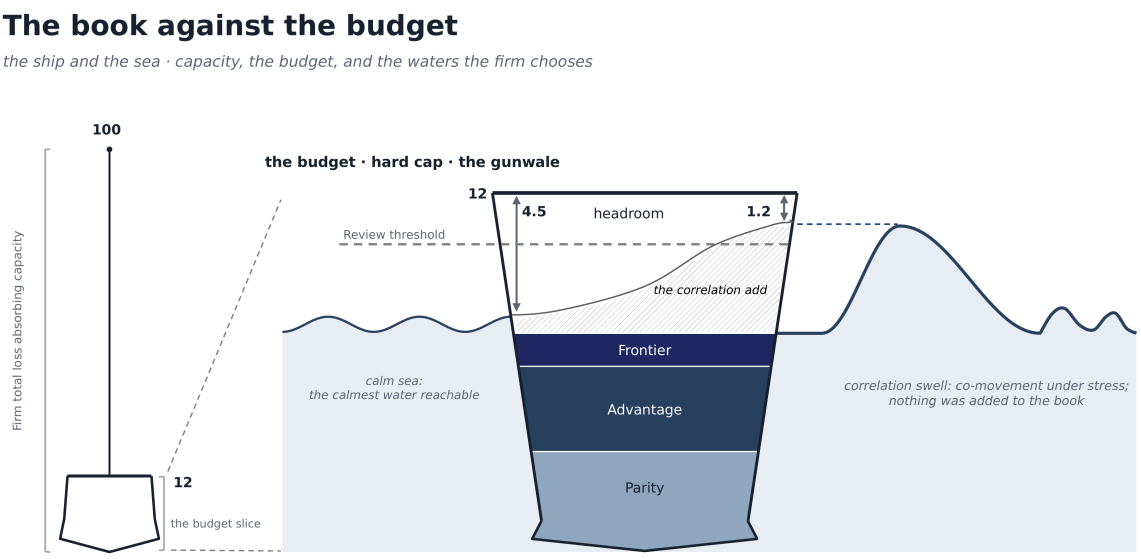

Figure 6. The book against the budget. The load line view. Left, the firm: loss-absorbing capacity indexed to 100, with the budget a 12-point slice of it. Right, that slice magnified into a vessel: the gunwale is the hard cap at 12, and the cargo is the standalone book, Parity, Advantage, Frontier, loaded to 6.9. The water is the aggregation environment. Even the calmest reachable sea ripples above the standalone level, the standing charge for what the book shares; inside the hull the same water rises left to right as correlations move toward their stressed values, the correlation add, lifting consumption from 7.5 to 10.8 while the cargo never moves. Headroom is what remains below the gunwale, 4.5 in calm and 1.2 under stress, and the review threshold is crossed where the rising water says it is. Illustrative; the swell at right is the sea state the stress test assumes.

4.3 Allocating the budget

Workloads consume the budget in proportion to their relative contribution to portfolio risk. The hard part, and the technical heart of the whole model, is that standalone charges do not by themselves describe the book: a ledger must add up, but the question is what is being added, and a sum of standalone figures misses what the workloads share. Two workloads that fail independently are a different exposure from two that fail together, and a portfolio's tail risk is dominated by its dependencies and correlations, shared models, shared data, shared infrastructure, shared failure modes, shared orchestration patterns, even a shared human review process, far more than by any single workload's standalone risk. The model makes that dependency layer explicit.

The intuition is worth making plain, because this is not how most executives are used to adding risk. Each workload's standalone figure prices its own private storm, a plausible bad outcome for that workload considered alone, so summing those figures quietly assumes the storms are private, that one workload's bad week says nothing about another's. A book built on shared foundations does not behave that way. An insurer knows the difference immediately: ten houses in ten towns and ten houses on one street can carry identical policies, yet the two books are entirely different exposures, because the street floods as one. Shared dependencies are the street. When several workloads ride the same model family, provider, data source, or review process, the event that hurts one tends to arrive for all of them in the same week, so bad cases the standalone view counted as separate draws land together. And landing together costs more than the sum of landing separately: fallbacks are contended when everyone reaches for them at once, the shared review and incident capacity each plan assumed saturates, unwind actions queue behind each other, and customers and regulators read five simultaneous failures more harshly than five scattered ones. The first mechanism raises

how often the book has a bad week; the second raises how bad that week is. Both are properties of what the book shares rather than of any workload alone, which is why the aggregate carries a charge for its dependencies even on an ordinary day.

Worse, these correlations are unlikely to be stable: by analogy with financial correlations, the working assumption is that they are mild in calm periods and rise toward one under stress, exactly when the firm can least afford it. That analogy is a precaution rather than an established law for AI workloads, which is itself the reason the dependency layer has to be examined under stress rather than estimated on an average day. It is also the most fragile part of the model and the part most in need of judgment and stress testing, which is why the paper does not pretend to a precision it does not have. Illustration 2 in Appendix C works the mechanism through on a two-workload example.

What the correlation work buys is therefore not a more precise number, since the budgeted figure is deliberately conservative, but three things the standalone view cannot give. It diagnoses which workloads concentrate the firm's existing exposures and which diversify them, so capital can be steered toward the diversifiers and away from the concentrators. It drives the shaping moves that lower a workload's contribution before it is funded, the proxy and the second-source fallback that turn a concentrating candidate into a tolerable one, which is where most of the value is realized. And it surfaces the shared models, vendors, and failure modes that quietly bind the book together, the exposures a workload-by-workload review never sees, so they can be managed before a single directive or outage removes several workloads at once. The sophistication is in the diagnosis and the design leverage it creates, not in the last digit of the budget.

In practice the firm estimates these correlations without pretending to measure them. The primary proxy is a shared dependency map: for each pair of workloads, how many and how material are the models, data sources, vendors, clouds, and failure modes they hold in common, on the principle that workloads leaning on the same things tend to fail together. Formally this is a factor structure rather than a pairwise estimate: the shared dependencies, provider, model family, data source, failure mode, are the factors, each workload's reliance on them its exposures, and comovement is read through the factors rather than estimated pair by pair from operating histories far too short and too nonstationary to support it. That map is refined by asking which workloads move together in the stress scenarios the portfolio is tested against, and it is expressed as coarse bands, low, medium, or high, set by the people who run the systems and then pushed upward under stress rather than trusted at their calm-period level. None of this is precise, but it is enough to separate a concentrating workload from a diversifying one and to size the budget conservatively, which is what the model asks of it. And when even the banded read is unavailable, the design already carries the fallback: the floor needs no correlation estimate at all, which is why it, and not the arithmetic, is the constraint that binds when the inputs are least trustworthy.

4.4 Three classes of investment, three kinds of limit

The risk budget is one limit, but it is not the only kind a portfolio needs, because not every AI investment is the same kind of decision. It helps to separate the book into three classes by intent, which a board can fix before any single workload is scored, and the test that places an initiative is plain: if it works, does it create a durable edge a rival cannot quickly copy, and if so, is that edge in the game the firm already plays or a new one.

The three classes call for three different kinds of limit, and conflating them is a frequent error:

Class	Intent	The kind of limit	How positions are judged
Parity	Competitive table stakes; diffuse productivity that rarely reaches the income statement	A capped spend pool, set deliberately and held against the sheddable run-rate	The cheapest adequate option bought, the rest declined; not underwritten for return
Advantage	Differentiating an existing process for a measurable, near-term return	The loss-denominated risk budget of this part	Each position earning its place by risk-adjusted return against the firm's hurdle
Frontier	A contingent, long-dated bet that could reshape the business or its market	A bounded option premium the firm is prepared to lose, managed as a venture book	Many small positions, each carried only to a dated trigger

For parity, the ceiling is held against the ongoing run-rate, the usage-driven cost that scales with adoption and can be shed, while the front-loaded build is committed capital, judged once at the build decision and kept in the capital view, which keeps two questions, what the firm may spend and what it may commit, from collapsing into one. A capped spend pool, a loss-denominated risk budget, and a bounded option premium are three different ceilings, and a firm that runs all of its AI spend against just one of them, usually by treating everything as either a cost to minimize or a bet to be made, will misgovern two classes out of three.

The highest-level allocation choice is therefore not which workloads to fund but how to split the firm's AI commitment across the three classes, and whether that split is deliberate. Most organizations are heavily weighted toward parity and have never decided to be. Setting the mix, and revisiting it, is a board decision that precedes every workload-level recommendation, and it is the point at which capital allocation for AI becomes a portfolio question rather than a procurement one.

4.5 The approval process

New investment is gated by a process designed to put the burden of proof where the knowledge is, and to do so without becoming a bottleneck. The sponsor owns and states the value and risk assumptions and records the status of baseline measurement, whether the firm has captured the before-state against which improvement will be judged. What happens next is where the adjustment layer does its work, because approving every workload through a committee would be sound governance for a bank and a tone-deaf blocker for a software company shipping daily.

The resolution is to approve an envelope rather than each decision. The committee defines, in advance, the zone within which a sponsor clears a workload itself: the kinds of value and risk, the spend, and the reversibility a workload may carry and still proceed without formal review. A workload inside the envelope clears in real time against those published rules. A workload that would leave it escalates to the committee. The board owns the firm's risk appetite, which cannot be delegated; the committee translates that appetite into the outer loss limit, the envelope, and the real-time rules; the sponsor operates inside it. Appetite set at the top, limits and rules applied below.

What sends a workload to the committee is a small set of boundaries rather than severity alone, because severity alone misses the kind of failure this paper opened with. The exhausted budget at

the technology company in Appendix C was not a safety event; it was a low-severity, high-spend one, and a gate that watched only for danger would have waved it through. A workload therefore self-clears unless it crosses a severity or blast-radius line, commits spend above a line, leans materially on a model, vendor, or failure mode the portfolio already concentrates, or is hard to unwind. Most workloads in a fast-clock organization cross none of these and clear at once. The worst-case prohibit screen still applies to every workload, but for low-severity work it is a fast self-check rather than a meeting, because the answer to what a workload can do unnoticed, and whether that is bounded, is usually an evident yes.

The price of clearing oneself, and the thing that keeps the core intact while the gate stays light, is a recorded claim and a review in arrears. A workload that self-clears still records what it promised, its value, its baseline, and the assumption the case rests on, so that the ex post loop can later check the promise against the result. That record is a lightweight entry, not a pack and a meeting, and it is what lets the committee govern by sampling and exception rather than by prior approval. The envelope is not fixed: the committee widens it for teams whose self-cleared workloads come in as promised and tightens it for those whose do not, so that autonomy is earned by demonstrated discipline rather than granted once. And the boundaries are drawn on a workload's total committed spend and its marginal contribution to the book, not on the size of the latest increment, so that a large commitment cannot be sliced into many small self-clearing pieces, which is the gate-gaming the operating model warns against in any case.

For the fast-clock corner this produces an approval process that feels like spending within an allowance rather than asking permission: a wide envelope, most workloads clearing in real time, continuous lightweight monitoring with alerts when a boundary is approached, and a frequent but light review rather than an infrequent heavy gate. For a high-severity, slow-clock organization the same machine sets a narrow envelope that routes most workloads to the committee and a heavier, less frequent review. The division of labor and the liability firewall are unchanged in both: the sponsor owns the claims, the committee owns the comparison and the budget and the envelope, and the independent method owns the rules. It is the width of the envelope, not the presence of the discipline, that moves across the spectrum.

4.6 Shifting governance left: the acceptance criteria as a design input

A gate that reveals its standard only at the moment of judgment wastes effort on both sides. Teams build to a guess, learn at the end whether it was the right guess, and the firm pays for the workloads that fail the test as surely as for the ones that pass. Much of the wasted pilot spend described in Part I has this shape: not bad ideas, but good ideas built without knowing what would make them fundable. The remedy is to move the standard forward. The model publishes its acceptance criteria, what makes a workload's value credible, its cost defensible, its risk contribution acceptable, and its manageability sufficient, so that they are known before design begins rather than discovered at the gate.

This is a communication layer as much as a control. Its job is to carry the criteria, and the reasoning behind them, upstream to the people making architecture decisions, so that approvability becomes a design input alongside the functional and technical requirements. The effect on incentives is immediate and healthy. Teams spend scarce design and build effort only on candidates that look likely to pass, which is the cheapest possible way to attack the problem of wasted spend, and they architect deliberately to be attractive to approvers, building in the very qualities the model rewards, management flexibility and risk reduction, from the first design rather than retrofitting them later.

Publishing the criteria is necessary but not sufficient, because a standard that is known but not rewarded loses to whatever the organization actually measures. Shifting left therefore carries an incentive condition: teams have to be judged on the outcomes the model values, approvable results delivered within budget, rather than on raw activity, consumption, or velocity. Decision rights, performance measurement, and rewards are one design rather than three. The envelope assigns the rights, the recorded claim and the ex post loop measure against them, and the incentive pays on the result; each leg fails without the others, since rights without measurement are license, and measurement without reward is decoration. Where that fails, the criteria become a poster on the wall while behavior follows the real reward. The Uber case in Appendix C is the stark case: a leaderboard that ranked teams by tokens consumed produced exactly that, consumption, and the caps and dashboards added afterward treat the symptom rather than the cause. Guardrails bound the damage; aligned incentives keep the damage from being the rational choice in the first place. That is the hinge on which governance-as-enablement turns: shift the standard left and align the reward to it, and teams build fundable, flexible, low-risk workloads because that is what they are rewarded for, not because they were inspected into it. The hinge is not a solved mechanism, and the paper does not pretend otherwise. Rewarding approvable outcomes within budget can itself be gamed, by sandbagging the baseline or defining approvable too kindly, which is why the ex post loop and the earned-autonomy envelope sit behind it as checks rather than the incentive standing alone. Aligning incentives without inviting new gaming is, as Part VI records, still open; the claim here is the modest one that an aligned reward beats a misaligned one, not that any reward is gameproof.

Appendix C sets out a concrete case: a thin proxy placed in front of a model call so a workload can switch providers without rebuilding and fail over if one is down, chosen not as governance overhead but as better engineering, because the criteria made its value legible at design time. This is the discipline software learned when it began designing for testability rather than testing at the end. Designing for approvability turns the gate from a recommendation into a guide: a firm gets safer, more flexible workloads not by inspecting them harder but because the people building them knew, from the start, what fundable looked like. The approval envelope is the operational form of this: the published criteria become the boundary within which a workload clears itself.

4.7 Keeping the inputs reliable

Because the sponsor owns the inputs, the process has to defend against optimism and against gaming. Four mechanisms do most of the work: provenance tiers that record how good each input is and weight it accordingly; an ex post loop that compares realized value and risk with what was claimed and feeds the result into the next round; a central reserve held outside the allocated budget for the surprises a tail-risk model exists to anticipate; and independent challenge of the assumptions that matter most. The challenge is aimed where manipulation predictably concentrates: at the load-bearing assumption, stated generously before funding, and at the baseline, understated so the later comparison flatters the result. Defending those two points buys most of what defending everything would. History sharpens the prior for one pathway in particular: technology's record of moving measured unit costs is weaker than its record of adding capability, with gains leaking into complexity and rents along the way, so cost reduction claims carry the heaviest ex post burden of all. None of these makes the inputs perfect. Together they make the inputs sound enough to act on, and they improve with every cycle. The case for the fourth has only strengthened. By 2026 roughly seven in ten chief executives describe themselves as their company's main decision maker on AI, which concentrates the bet and its evaluation in a single seat; whatever that concentration buys in speed, it is precisely the condition independent challenge exists to offset, because no sponsor, however senior,

can grade their own book.

4.8 Allocating by merit, not arrival

A hard cap alone is too blunt an instrument. Three refinements keep the budget a portfolio rather than a first-come ration. The marginal hurdle ensures that capacity is spent on the best risk-adjusted additions rather than the earliest applicants. A rebalancing and reassessment mechanism prevents incumbent workloads from locking up the budget indefinitely simply because they arrived first; existing consumption is periodically re-earned, not grandfathered. And a clean separation between the inviolable hard cap and softer thresholds that merely trigger review keeps the cap meaningful while allowing the firm to manage its approach to the cap deliberately rather than by surprise. That separation is the portfolio-level form of the approval envelope: the line past which a workload is reviewed rather than cleared.

4.9 The shape of the book

Concentration is one way a book fails in aggregate; dispersion is the other, and the risk budget set out earlier in this part governs only the first. A portfolio can clear every workload-level screen, sit inside all three class ceilings, and still fail as a whole by scattering its capital across many small, bottom-up pilots that never reach scale, answer to no stated thesis, and carry no measured outcome. This is the failure sometimes called death by a thousand cuts, and it is not the same problem as a single oversized commitment. The approval gate already stops one large bet from being sliced into self-clearing pieces; it does nothing about a hundred genuinely separate small ones, each reasonable alone, accumulating into a book that is busy everywhere and decisive nowhere. The remedy is the one the class mix already uses: give the board a few more views of the portfolio's shape, and let it decide whether that shape is a choice. Each view is a plain count kept as its own gauge rather than folded into a composite score, in keeping with the model's refusal to rank a book on a single number, and each limit is the same envelope-and-displacement device the approval process and the budget refinements already rely on, pointed at a new denominator.

Figure 7. The same matrix, two books: scattered against focused. Each bubble is one initiative, placed by its lifecycle stage (horizontal) and its strategic intent, the Parity, Advantage, and Frontier classes (vertical), with area proportional to committed capital; filled bubbles are tied to a stated thesis and open ones are off-thesis. The left book is scattered, many small and largely off-thesis bets crowded into the early stages with none reaching scale; the right book is focused, fewer and larger and mostly on-thesis bets maturing into validation and scale, with a bounded frontier pilot and a small off-thesis allowance. The readout beneath each panel is computed from the bubbles shown. Illustrative schematic; the two books are stylized.

The first view is maturity. Each workload carries a lifecycle stage, from exploration through pilot, validation, and scale to retirement, set beside its class and orthogonal to it, since the class fixes intent while the stage fixes how far the bet has actually travelled. Read as a grid of class against stage, the book shows at a glance whether it is maturing or smearing: a healthy one moves workloads steadily to the right, a failing one piles them in pilot and leaves them there. Three readings follow. A graduation rate, the share of piloted work that reaches production, turns the pattern into a number that a low value makes undeniable. The share of capital still parked in pre-scale stages says how much of the commitment is not yet earning. And a cap on the number of pilots in flight at once forces a queue, so that a new pilot waits for an existing one to graduate or be killed rather than adding to an ever-deeper stack; this is the most effective single check on sprawl, because it bites before the spending rather than after it. The stalled pilot, one sitting past a set age with no movement, is the portfolio-level form of the capital-without-movement signal named in Part III, raised from a per-workload trigger to a standing metric the board can watch.

The second view is orchestration. Class records intent but not provenance: it does not say whether a workload belongs to a deliberate, top-down thesis or simply arrived from the bottom up. The model therefore adds a thesis layer above the workloads, each one rolling up to a named strategic theme or sitting unaffiliated. The unaffiliated set is the off-thesis bucket, the organic experimentation a healthy firm wants but an undisciplined one lets quietly become the whole book, and the board

bounds it with an envelope expressed as a share of capital or a count. The displacement rule does the rest: once the envelope is full, a new organic bet must attach to a thesis or displace another, and the envelope does not widen on request. Each theme is then read on its own, so the pathology inside a thesis is visible too, not only the orphans outside one. A theme that is all small pilots and no scaled bet is a thesis the firm has funded but never committed to, and the readout names it as such.

The third view is accountability, and it closes the loop the paper opened with, where most firms were found measuring activity rather than outcomes. At the portfolio level the model reads the share of spend tied to a named outcome metric, where one is expected, against the share tied to none, the dark spend a board cannot defend when it is asked what the money bought. Naming the metric yields two further checks at no extra cost. A workload whose declared payoff is the movement of a key metric, but which names no metric, is undefendable on its own stated terms, and is flagged. And when several workloads claim the same metric, the book is either rebuilding one capability in several places, which is duplication to consolidate, or counting one benefit several times in its aggregate value, which is a double count that inflates the apparent return; either way the collision is something no workload-by-workload review can see, and the model surfaces it.

Two gauges sit beneath these views and read the dispersion directly. Where the correlation work reads concentration through shared dependencies, an effective number of bets reads it through capital itself, computed from how evenly committed capital is spread across the book, so that a portfolio of a few real positions is told apart from one smeared across many token ones. It is concentration's mirror image, and it is read beside the correlation picture rather than inside it. And a sub-scale line, a minimum commitment below which a bet is too small to move the firm, turns the crumbs into a cohort the board can actually see: how many there are, how much capital they hold between them, and what they return as a group. The purpose is not to forbid small bets, which are how options are bought and how the frontier class is meant to work, but to make their sum visible and bounded, so that the firm chooses to carry the small stuff rather than discovering after the fact that the book has become the sum of bets none of which was ever meant to be the strategy.

None of these views changes what a workload is or how it is scored; they are tags laid over the same book, and none enters the recommendation. As with the class mix, they exist so that a board sees the portfolio's shape and decides whether it is deliberate. Most books drift into scatter the same way they drift into parity, by accumulation rather than choice, and the discipline is the same in both cases: name the shape, set the envelopes, and revisit them on the same cadence as the budget.

4.10 Stress testing

Where a trustworthy point estimate of tail risk does not exist, the better substitute is to interrogate the number rather than to trust it. The model is stress tested on a cadence that ranges from quarterly to nearly continuous as the portfolio grows and as the organization's place on the clock-speed axis demands. Scenarios are run, correlations are pushed to their plausible extremes, and the tests reach beyond today's parameters to fundamental regime shifts: a new class of model-layer failure, a sharp move in vendor economics, a capability change that rewrites the risk, or the abrupt loss of access to a model the portfolio is built on, whether from an export-control action of the kind seen in 2026, a provider withdrawing or deprecating a model, a licensing change, or a sustained outage, since those, not an average bad day, are what a fast-moving field actually delivers. The tests also probe the portfolio's manageability, whether, if a scenario arrived, the firm could change course, switch, or unwind quickly enough, or whether it is locked in. The portfolio's behavior under stress, not its behavior on an average day, is what informs the budget. A risk number that has survived hard

questions is worth more than a more precise number that has never been asked any.

4.11 Scope: the model alongside risk and ESG

It helps to state plainly what the model does and does not do where it meets adjacent domains.

On risk, the model is a contributor, not a replacement. The firm already has a risk function and a maturing single-system governance regime; the model's job is the narrow one of producing the portfolio-level outputs those functions do not generate, the aggregate exposure, the concentration and correlation picture, and the consumption against the budget, handed over in a form the existing regime can act on. The practice supplies the measurement; the firm's risk governance owns the decision. How tightly the portfolio layer docks with that regime is itself an adjustment-layer setting. Supervisors, meanwhile, are beginning to expect the posture the model quantifies. Through 2026 prudential guidance moved from principle toward instruction: the Australian regulator wrote to its regulated entities in April expecting boards to align AI strategy with risk appetite and to measure AI's impact across risk classes including third-party dependencies, and the Reserve Bank of India's June draft guidance makes the board responsible for approving model risk appetite, with independent validation through the model lifecycle. Those expectations are stated in the language of controls; a risk budget of the kind set out here is one way to make them quantitative, a board-set appetite translated into a limit the allocation process consumes against and reports on.

On ESG, the scope is narrower still and is mostly about carbon. Because the cost chain already models energy, the model is carbon-aware by construction: a directional grid-intensity factor gives every cost figure an emissions counterpart. Where the firm prices carbon, through an internal price or a disclosure-driven cost under a regime such as the CSRD or California SB 253, that price folds into the fully loaded cost and is weighed inside the same risk-adjusted value as any other input. Where it does not, the model surfaces carbon as an optional co-benefit, flagging the placements and schedules, a cleaner region or lower-intensity hours, that cut emissions and cost together at constant output. Carbon informs the allocation; it does not gate it, and the intensity stays directional unless the firm has measured data. Beyond that the model stays out of the wider ESG apparatus: it is not a compliance regime, and it does not try to become a carbon-accounting or assurance service.

4.12 The controls layer: aligning to the FINOS AI Governance Framework

The method in this paper is a capital-allocation discipline. It decides whether an AI initiative earns committed capital, how much, and against which ceiling. It is deliberately not a controls framework. Whether a given deployment is safe to run, whether its data handling is sound, whether an agent that acts is bounded, and whether the regulatory obligations are met are controls questions, and they are answered one layer down.

That layer increasingly has an open, financial-services-specific home. The FINOS AI Governance Framework, built by members of the Fintech Open Source Foundation including major banks, is a catalogue of the operational, security, and regulatory risks of generative and agentic AI in financial services, each paired with preventative and detective controls and mapped to the horizontal standards a board already recognizes: the NIST AI Risk Management Framework, the certifiable ISO/IEC 42001 management-system standard, and the EU AI Act. It answers the question this paper does not, which is how to run a deployment safely and demonstrably once the decision to deploy has been made.

The two layers are complementary, and they meet at a single seam. The allocation decision sits on top: fund, cap, kill, or prohibit, and at what committed capital. The controls layer sits underneath:

the mitigations that make a funded workload safe to operate. The onboarding and approval gate is where they join, because the same risk profile that sets an initiative's allocation recommendation also determines which controls it must carry. A workload that acts on a customer-facing path implicates the framework's agent-authorization and model-overreach risks and must carry its firewalling and acceptance-testing controls; one with no firm consumption floor implicates its availability and Denial-of-Wallet risk; one touching regulated decisions implicates its bias, explainability, and regulatory-oversight risks. The capital recommendation and the control set are two reads of one profile.

The instrument makes the link explicit. For each initiative it surfaces, alongside the allocation call, the specific framework risks the profile implicates and the mitigations to require, citing the framework's own risk and control identifiers. A prohibited structure is the sharpest case: the controls are necessary but never sufficient, and the workload does not proceed on controls alone. The intent is interoperability rather than duplication. A firm that has adopted the framework can read this model's output in the vocabulary its risk and control functions already use, and the capital-allocation lens supplies the one thing the controls frameworks leave out, which is whether the investment should be made at all and at what size.

Part V. Operating the model

A model is only as good as a firm's ability to run it. This part describes the roles, the rhythm, and the failure modes, so that the discipline survives contact with an organization.

5.1 Roles

Three roles carry the model. The portfolio manager or AI investment committee owns the comparison across workloads and the budget they consume. The sponsor owns each workload's value and risk claims and the baseline behind them. And the independent advisor owns the method and its mechanics, deliberately as a toolmaker rather than as a certifier of any particular workload's risk, both to keep the method credible and to keep responsibility where it belongs. Independence here is structural and specific: the party that owns the method has no stake in which way any particular recommendation falls, which is achievable, rather than a claim to having no commercial relationship with the firm at all, which is not. The toolmaker owns the rules; the firm owns the decisions and the outcomes they produce.

5.2 Cadence and artifacts

The model runs on a small set of artifacts:

Artifact	What it does
The published acceptance criteria	Let teams design to the gate before they reach it
The approval pack	Carries the sponsor's claims, with their translations and grades, to the decision
The portfolio dashboard	Consumption against the budget: the shape of the book and its headroom
The stress-test scenario set	Pushes correlations toward their stressed values and refines the bands
The decision record	Each decision with the assumptions behind it, the claim the ex post loop later checks

The cadence is regular rather than continuous at first, a portfolio review on a quarterly rhythm, tightening as the portfolio grows and as more of the firm's value and risk run through it.

A portfolio review is only as good as the questions it forces. For any workload at any review, a board or investment committee should be able to answer the following, ordered, as this paper is, to put return before risk. They are, by design, the questions a workload's sponsor would prefer not to be asked.

- What is this workload returning now, measured at the operating outcome and net of the full cost of running it, rather than at the model's output?
- Is this still the best use of the next dollar, set against the other workloads competing for the same scarce capacity?
- Is this workload on a path to scale, tied to a stated thesis, and accountable to a named outcome, or is it one of many small bets the book is quietly accumulating?
- What is the worst plausible outcome, how much of the risk budget does this workload consume, and is any part of the downside uncapped?
- What else in the book shares this workload's model, vendor, or failure mode, and what happens if they fail together?
- Are we funding a commitment or buying exposure, and if exposure, what would make us exercise, hold, or walk away?
- What happens to cost and to risk when volume rises or conditions turn adversarial, given that an agentic system's cost is a distribution rather than a fixed price?
- What cost has been left out of the case: the review labor, the exception handling, the retraining, the integration, the change management?
- Which human work does this remove, and which new human work, oversight and exception handling, does it create?
- What signal, agreed in advance, would tell us to hold or kill this, has it tripped, and are we continuing because the evidence supports it or because we have already spent?

A review that can answer these is governing the portfolio. A review that cannot is, in this paper's terms, reporting on it.

5.3 The adoption path

No firm adopts this all at once, and none should. The sequence matters: understand the portfolio, then allocate against it, then govern it, and only then optimize it. The most common and most expensive inversion is to optimize first, chasing unit-cost savings before the firm understands which workloads are worth running at all. Savings won before allocation tend to erode, because they are spent efficiently on the wrong things. The discipline is to earn the right to optimize by first knowing what the portfolio is and what it is for.

The smallest viable start. A firm need not build the whole model to begin, and the first version of each piece can be deliberately crude. The minimum is six moves. Set the split of AI spend across the three classes as a board decision, even a rough one, so the mix is chosen rather than inherited. Name a risk budget in money the firm can afford to lose, even as a single conservative figure anchored to loss-absorbing capacity. Inventory the AI workloads already running and the dependencies they share, the models, vendors, and failure modes, since that concentration is what the portfolio view exists to surface. Route every new candidate through one consistent screen, its value pathways, its fully burdened cost, its marginal contribution to risk, and a bounded-worst-case veto, before it is

funded. Publish the acceptance criteria so teams design toward them rather than meeting them for the first time at the gate. And close the loop once, on a single funded workload, comparing what it returned and cost against what was claimed. Each move is useful on its own; together they form a standing practice the firm can deepen, and they are enough to stop the largest waste, capital committed before anyone has asked what it is for.

5.4 Pitfalls

A handful of failure modes recur often enough to name. Fixating on the hardware sticker while the real cost sits in power, utilization, and staff. Confusing high utilization with efficiency, when a cluster can be fully occupied and still doing little useful work. Optimizing before allocating. Applying a single depreciation life across a whole fleet of differently used hardware. Importing another firm's break-even thresholds instead of deriving the firm's own. Grandfathering incumbent workloads so that the budget ossifies. Scattering the book across many small, bottom-up pilots that never reach scale, until the portfolio is the sum of bets none of which was meant to be the strategy. And gaming the gate, which is why the safeguards against it are not optional. Each pitfall is a place where the appearance of discipline quietly substitutes for the real thing.

Part VI. What is settled and what is open

A model offered at the start of a field's convergence should be clear about its own maturity. Some of this is durable, some is a snapshot of 2026, and some is still being sharpened, and saying which is which is part of being trustworthy.

The durable core is the structure: AI as a portfolio, valued through distinct pathways, fully costed, risk-adjusted at the margin, and governed against an explicit budget that feeds the existing regime. We expect that structure to age well, because it rests on disciplines, capital allocation, risk budgeting, options thinking, and liquidity management, that long predate AI. The commitment to stay adaptive, and to treat flexibility as a measured quantity, is part of that durable design rather than a hedge against it.

The perishable layer is the numbers. The spending figures, the pilot-failure rates, the percent-of-revenue benchmarks, and the prevailing depreciation lives cited in this paper are a snapshot of 2026 and should be refreshed as a matter of course. None of the model depends on them.

What remains open is not the structure, which is set, but the calibration and the hard quantitative frontier, and these are the live agenda of this work rather than embarrassments to be hidden. Several pieces are a matter of fitting the model to a particular firm: the unit in which to denominate the risk budget, how the board's appetite is translated into a cap against the firm's capacity to absorb loss, the exact form of the marginal hurdle, and the mechanics of rebalancing and of the ex post loop that turns claims into learning. Others are genuinely hard and improve with research and data: how to model dependency and correlation, the most consequential piece and the one most dependent on stress testing; how to price the stochastic, heavy-tailed cost of agentic systems; how to value option-like workloads without overstating them; how to measure manageability in a way that is comparable across workloads; and how to align the incentives of the teams that build and run workloads with portfolio outcomes without inviting new forms of gaming.

Two notes on this open agenda. The hardest items named above, dependency and correlation, the valuation of option-like workloads, and a measure of manageability comparable across workloads,

are precisely the items that the most rigorous per-investment frameworks also leave open and assign to future work. That convergence is some assurance that the frontier is drawn correctly rather than idiosyncratically to this model. A second question is the model's applicability below the large-enterprise tier. A full value decomposition assumes a level of financial disclosure and baseline measurement that smaller and private firms often lack, and adapting the method to that setting, where much of the adoption will actually happen, is itself a live question rather than a solved one.

Naming these is not a weakness in the model. It is the difference between a method and a sales pitch.

Coda: the fiduciary case

The fiduciary case is simple, and it is the point of the whole model. A board carries the duty to allocate the firm's capital wisely and to understand the risks the firm is running. As AI becomes a material call on both, a board should be able to say, and defend, what that AI is returning, how much risk the firm is carrying to earn it, and why each answer is sound. The duty is not a brake: it is to know what kind of risk is being authorized and what evidence would change the answer, since limits discovered by accident are the expensive kind. A firm that can answer those questions is not flying blind, and so it can commit more and move faster than one that cannot, this governance exists to let the firm invest with confidence, and it works partly by aligning what the firm rewards with the flexible, low-risk, fundable choices it should want. The judgment that applies the model carries no stake in which way any particular recommendation falls, because a model for allocating capital and governing risk is only as trustworthy as the independence of the party applying it. That is the test this model is built to pass.

Appendices

Appendix A. The value pathway vocabulary

The model uses a small, fixed vocabulary of value pathways so that workloads can be compared consistently. Each pathway names a distinct way an AI investment pays, with its own outcome metric, its own financial driver, and its own payoff shape:

Pathway	Driver it moves	Measured as	Payoff shape
Direct cost reduction	Margin	Cost against a recorded baseline	Linear in volume
Revenue generation	Revenue	Incremental sales or margin attributable to the workload	Convex with adoption
Revenue protection	Revenue	Retained revenue or reduced churn	Defensive; a floor under erosion
Loss or risk avoidance	Cost of risk	Avoided cost from errors, fraud, or failures	Episodic; tail-shaped
Capacity or throughput unlock	Margin or revenue	Additional output or served demand	Stepped at the binding constraint
Capital efficiency	Capital efficiency	Freed working capital or deferred investment	A one-time release, then carried

A single workload commonly pays through several of these at once. The model records the blend and preserves each pathway's shape rather than averaging them into one figure.

Appendix B. The cost taxonomy and the burdened-rate chain

Fully burdened cost is assembled from the major components of total cost of ownership rather than from a headline price. For owned infrastructure these include hardware acquisition amortized over its useful life, power and cooling, facility and space, networking, software, and the staff who run it. The burdened-rate chain converts these into a decision-grade unit cost in four steps: total cost of ownership over the asset's life; a fully burdened hourly rate for the capacity; an effective rate adjusted for occupancy and efficiency; and a cost per unit of useful work. The detailed component library and the optimization levers that act on each component lie beyond the scope of this paper. In practice the cost and usage inputs to this chain increasingly arrive already normalized, under the FOCUS specification, which makes the components comparable across providers; the chain is agnostic to the source of the numbers and treats a normalized cost feed as the substrate it operates on.

Appendix C. Worked illustrations

These illustrations are deliberately simplified and use round, hypothetical numbers. They show how the model changes a decision, not real client figures.

Illustration 1. How full costing can reverse a build-versus-rent decision. A firm runs an internal assistant at 3.5 million requests a month and weighs two ways to serve it: a managed endpoint at $9.00 per thousand requests, or a reserved two-node cluster of its own. The cluster costs $40,000 a month in fixed capacity (hardware amortized over its useful life, reservation, power, cooling, and space) and $18,000 a month in staff and oversight, and it draws about $1.20 of compute and energy per thousand requests served. At full load it would serve 10 million requests a month.

On the headline number, building looks cheap: the next thousand requests, served on capacity the firm already holds, cost only the marginal $1.20, far below the $9.00 rented rate. On the decision-grade number it is the opposite. At today's 3.5 million requests the cluster is 35 percent used, so its fully burdened, utilization-adjusted cost is the monthly total over the volume: fixed $40,000, staff $18,000, and variable $4,200 (the $1.20 per thousand across 3,500 thousand requests) sum to $62,200, which over 3,500 thousand requests is $17.77 per thousand. That is nearly double the rented rate and about fifteen times the marginal figure. The workload that looked far cheaper to build is, as built and used, almost twice as expensive as renting.

The reversal hinges on utilization. Fill the cluster to 10 million requests and the variable cost rises to $12,000, the monthly total to $70,000, and the burdened cost falls to $7.00 per thousand, now below the rented rate: full, the build wins; one-third full, renting wins. The two readings govern different decisions. The marginal $1.20 says serving one more request on capacity already held is nearly free, so throttling existing demand off the cluster would waste it; the burdened $17.77 says that absent a credible path to fill the cluster, this was the wrong way to have built the workload, and the rented endpoint is cheaper. The headline number hides both truths, and only the decision-grade number should drive capital.

Illustration 2. Why a workload's risk depends on what the firm already holds. This is the correlation mechanism at the centre of the portfolio risk view, made concrete, and it carries a warning about its own arithmetic that is part of the lesson. The portfolio already holds Workload A, whose loss in a bad but plausible year is about $10 million. The firm is weighing Workload B, whose comparable standalone loss is $6 million. Naively, B adds $6 million. It does not: what B adds depends on

whether its bad year tends to coincide with A's, which is to say on their correlation, written ρ.

Combining the two losses with the textbook portfolio rule $\sqrt{A^2 + B^2 + 2\rho AB}$ shows the mechanism the example exists to demonstrate. A workload that leans on A's own model, vendor, or failure mode ($\rho = 0.8$) adds \$5.2 million, most of its standalone figure, since $\sqrt{100 + 36 + 96} = 15.2$ against A's \$10 million. A workload that fails for unrelated reasons ($\rho = 0$) appears to add only \$1.7 million, since $\sqrt{100 + 36} = 11.7$, and one that fails in the opposite conditions ($\rho = -0.5$) appears to lower the firm's loss, since $\sqrt{100 + 36 - 60} = 8.7$. So the quantity that should set B's hurdle and consume its budget is its marginal, correlation-aware contribution, not its \$6 million standalone number.

The warning is in the word *appears*. That rule holds only when losses are light-tailed and bound by a stable correlation, and AI losses, as 1.5 and 4.3 argue, are neither: they are heavy-tailed, and their correlations are mild in calm periods but climb toward one under stress, the exact regime a tail-risk budget exists to survive. As ρ approaches one the rule collapses to $A + B$, the full \$16 million, so the diversification the calm correlation showed is the first thing to vanish on the worst day, when the shared models, vendors, and failure modes that looked independent bind together. The model therefore does not bank calm-period diversification. It sizes a workload's contribution at a stressed correlation, so a candidate that looks diversifying at $\rho = 0$ is budgeted nearer its standalone \$6 million, and the formula is treated as a conservative starting point to be stress-tested rather than a figure to rely on. The illustration leaves two lessons at once: marginal, correlation-aware risk is the right quantity, and the convenient way to compute it understates the very tail it is meant to guard.

Illustration 3. Designing for approvability. A team building a workload that calls a large language model would, by default, wire the application straight to one provider's API. Knowing the acceptance criteria in advance, it instead places a thin proxy in front of the model call, able to route the same request to the original provider or to the same model on a major cloud. The change is small in engineering terms and large in portfolio terms: the workload can switch providers without rebuilding, so its lock-in falls and its manageability rises, and it can fail over between providers, so its standalone risk falls. Because both gains are visible against the published criteria, it reaches the gate already more attractive than its naive version.

Illustration 4. What misaligned incentives look like. In 2026 Uber set out to accelerate its engineering with agentic coding tools and, to drive adoption, ranked teams on a leaderboard by the tokens they consumed. Adoption climbed steeply, to the large majority of its engineers, and so did the bill: the firm exhausted its entire full-year budget for these tools in the first four months, and a senior executive said the link to shipped features was not yet there. The firm then reached for controls: a cap near \$1,500 per tool per month, spending dashboards, and an approval step to exceed the cap, the guardrail layer this model treats as table stakes. The episode maps onto the model at two points. It is an incentive failure before it is a cost failure: the firm measured and rewarded activity, so activity is what it got, the trap the model avoids by rewarding approvable outcomes within budget. And it shows the absence of two controls in practice, a consumption budget with stops on agent loops and retries, and the screening question asked of every workload before funding, namely what is the worst it can do unnoticed and is it bounded. An uncapped leaderboard pointed at a metered resource has no such floor, and is precisely what the model would decline. The figures are as reported by The Information and TechCrunch in 2026.

Illustration 5. One workload through the whole model. The earlier illustrations isolate single mechanisms; this one carries a single workload along the path in Figure 5, through both halves, to show them acting together. A bank is weighing a customer-service assistant that drafts replies for its

support agents.

Readiness. The firm has the conversation data, the integration into its case system, and the oversight capacity to review the assistant's output. It passes the screen.

Value, cost, risk. On value, the assistant pays mainly through one linear pathway, deflecting routine contacts, worth about $4 million a year in handling time, with a smaller convex upside if faster resolution lifts retention. On cost, its fully burdened, utilization-adjusted figure, the managed model calls plus oversight and integration, is about $1.5 million a year. On risk, its standalone bad-year loss, a spell of mass mis-handling or a compliance lapse, is about $3 million; but it runs on the same foundation model three of the bank's other workloads already depend on, so under stress it fails when they fail, and its marginal contribution to the portfolio's tail is close to the full $3 million rather than diversified away.

Decision. The value clears the cost comfortably, and $4 million against a $1.5 million cost and a $3 million risk contribution would ordinarily fund. But that risk is a concentration the bank already carries, so the decision is not a plain fund: it is fund on condition of shaping. Before scaling, the team puts a thin proxy in front of the model and adds a fallback to a second one, the move Illustration 3 describes, which lets the assistant fail over and cuts its marginal contribution from about $3 million toward $1.5 million.

Budget and gate. At that reduced contribution the workload consumes a share of the risk budget the bank can afford, and because it now carries the fallback, a bounded worst case, and the agreed oversight, it clears the published acceptance criteria at the gate and is funded to scale.

Ex post loop. Six months on, realized handling savings are about $3.5 million, modestly below the $4 million claimed, so the value estimate is recalibrated down; no incident occurs, and the fallback is exercised once during a provider outage, confirming the manageability the shaping bought. Those recalibrated figures re-enter the next round, which is the loop closing. One workload has moved from candidate to funded and reviewed, valued, costed, risk-adjusted, shaped, gated, and learned from, with the allocation half and the governance half acting on the same object.

Illustration 6. Expanding what is working: the unit figure and the margin. The commonest return trip through the decision is not a failing workload but a succeeding one asking for more. A bank runs an agentic review of periodic client-file refreshes on its lowest-risk tier. Last quarter it completed 240,000 reviews at a fully burdened cost of about $310,000, and the figure leadership quotes is the unit one: $1.30 per completed review against roughly $9.00 fully loaded for the human path it displaced. The proposal on the table is to expand the workload to the next tier of files.

The average and the margin. $1.30 is the average of the tier the workload was pointed at first, and the expansion is decided at the margin, on both sides of the comparison. The next tier's files are longer, retrieve more, call more tools, and escalate more often, so the marginal agent cost is about $3.20 per completed review, and only around 55 percent of files complete without a person at the new tier against 85 percent on the old, so the effective figure per truly completed review is nearer $5.80 once escalations are carried. The displaced human cost moves too: the easy tier cost about $6 by hand while the complex tier costs about $14, so the decision-grade comparison is $5.80 against $14 at the margin, not $1.30 against $9.00 on the averages. The expansion still pays, and the margin is roughly four times thinner than the quoted unit figure suggests.

The claim under challenge. The load-bearing assumption is the 55 percent completion rate, measured on a pilot of four hundred files rather than asserted, and the $14 baseline is reconciled with finance rather than taken from the sponsor. The improving-trajectory claim, completion rising as the tooling

tunes, is recorded as a claim with its triggers named in advance: completion below 45 percent for two consecutive months, a quality-assurance overturn rate above 6 percent, or a marginal cost above $8 sends the expansion back to the decision.

Risk and the boundary. The new tier raises the blast radius, higher-risk clients and heavier regulatory consequence for a wrong clearance, so the autonomy boundary does not widen with the volume: the agent drafts and a person approves on the new tier until reliability is demonstrated in production, and the floor is unchanged, no autonomous adverse action at any tier.

The book. The expansion adds roughly $0.9 million of annual run rate on the same provider and model family the bank's largest workloads already lean on, so its marginal, correlation-adjusted claim on the risk budget runs well above its standalone read. Headroom permits it, but the shape worsens, so the top-down reading attaches the condition the bottom-up reading cannot see: the new tier runs behind the proxy with a qualified second model before scale, the move Illustration 3 describes. And the incremental capital is not grandfathered in: under the allocation rule it competes at the gate on merit against the book's other claims, where success to date buys the hearing rather than the funding.

Decision. Expand, conditionally and staged: a first tranche of about a quarter of the new tier's volume inside a widened envelope, human approval retained, second-model fallback wired, triggers live, with full expansion returning to the gate on the tranche's measured completion, overturn, and marginal cost. The two readings reconcile as 3.7 requires, a margin that clears the hurdle bottom-up and a concentration priced top-down, and the unit figure that started the conversation, $1.30 a review, turns out to have been the least decision-relevant number in the room.

Appendix D. The model, stated compactly

For readers who ask whether this is a metaphor or a model. The statement is deliberately hybrid: it does not claim that AI tail risk can be estimated with the precision of market risk in liquid instruments. It separates what is denominated in loss terms at the portfolio level, where the board's appetite and the firm's capacity to absorb loss are real quantities, from what is estimated comparatively at the workload level, where bands, grades, and stress-consistent factor exposures replace point estimates. The budget creates scarcity; the estimation method refuses false precision; the floor binds when the estimates deserve the least trust.

Objects. The unit is the workload, indexed $i = 1, \dots, N$: one purpose with one accountable owner (Section 3.1), so the unit cannot be moved by redrawing lines. Each workload carries a value vector, a cost, a risk contribution, and a manageability score.

Value. $v_i = (v_i^{rev}, v_i^{mgn}, v_i^{cap}, v_i^{rsk})$: the pathways of Appendix A, each cashing out to a named financial driver through a stated translation, each carrying a provenance grade, never blended into one scalar. Expected value above the hurdle, $E[v_i] - h_i$, uses class-specific hurdles h_i.

Cost. c_i is the fully burdened, utilization-adjusted chain of Appendix B, used for the existence and capacity judgments; the marginal cost mc_i alone informs the decision to serve more.

Risk. Exposures β_{ik} load workload i on shared factors k: provider, model family, data source, orchestration pattern, failure mode, review process. Portfolio stress loss L is evaluated with correlations pushed toward their stressed values, and each workload's consumption is its marginal contribution RC_i to that stress loss, carried in bands rather than decimals.

Constraints. The budget: $\sum_i RC_i \le B$, where $B = \theta \cdot C$, a board-chosen share θ of loss-absorbing

capacity C, with a soft review threshold below the hard cap. The floor: for every workload, the unsupervised worst case must be bounded, an absolute screen no expected value can buy through. Class limits cap Frontier and govern the Parity, Advantage, Frontier mix; envelopes bound what runs without re-approval.

Objective. Admit, size, and keep the set of workloads that maximizes expected value above hurdle subject to those constraints, with candidates compared by the marginal ratio $(E[v_i] - h_i)/RC_i$ within class, allocation by merit rather than arrival.

Options. A Frontier claim is admissible only as a bounded position: premium capped at $p_i \leq \bar{p}$, a dated trigger at which the position is rescored or closed, and a stated differential claim on the upside, since an option every competitor holds is worth little to anyone.

Agency. The sponsor owns the claims because the information is theirs; the recorded claim and the ex post loop are the monitoring; published acceptance criteria are the bonding device; the evaluator is independent because interested monitoring is worth little; and the residual loss that survives is expected and priced, not denied.

Calibration. θ and C come from the board's appetite and the firm's capital analysis where one exists; elsewhere C is assembled from the anchors the firm already holds, covenant headroom, liquidity above minimum operating cash, rating and dividend buffers, or named at its crudest as the board's answer to the largest loss that would not change the firm's plans; the factor set and band levels from the shared dependency map, pushed up under stress; hurdles from the firm's cost of capital by class; and the failure budget, an assumed miss frequency across initiatives, is carried beside the risk budget and answers a different question, how often the book misses rather than how much loss it may consume.

Falsifiable expectations. The model predicts, testably: that workloads with baselines fixed before deployment show smaller gaps between claimed and realized value than those measured after; that incidents at a shared provider or model family produce comovement across dependent workloads well above their calm-period bands; that cost-reduction claims under-realize relative to claim more often than revenue claims of similar grade; and that firms operating published acceptance criteria kill fewer workloads after launch, because fewer unfundable ones get built. Where these fail, the loop says so, which is the property that makes this a model rather than a metaphor.

Appendix E. Glossary

- **Workload.** A discrete AI investment that can be funded, measured, and governed on its own.
- **Value pathway.** A distinct mechanism through which a workload creates value, with its own metric and payoff shape.
- **Fully burdened cost.** The total cost of delivering a unit of work, including all components of ownership, not just the headline price.
- **Effective cost.** The fully burdened cost adjusted for how much of the time capacity is used and how efficiently.
- **Loss-absorbing capacity.** The resources a firm could use to absorb realized losses over a horizon without breaching a constraint that matters to it, solvency or regulatory minima, covenants, the rating, the dividend, funded strategy. The denominator the risk budget is set against: read from capital machinery where it exists, assembled from covenant, liquidity, and buffer anchors where it does not.
- **Portfolio risk budget.** A firm-wide limit on AI tail risk, grounded in board-set appetite and

operationalized by the risk committee, allocated across workloads and consumed visibly.

- **Risk-adjusted-value hurdle.** The marginal test a workload must clear, earning enough value for the risk it adds to the portfolio.
- **Provenance tier.** A recorded indication of how reliable an input is, used to weight it.
- **Ex post loop.** The comparison of realized value and risk against what was claimed, fed back to improve future decisions.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Claude (Anthropic) to draft and edit the text, to assist in locating and checking sources, and to help develop and check the paper's model and its supporting computations. After using this tool, the author reviewed and edited the content as needed, independently verified the sources and the claims they support, validated the model and its outputs, and takes full responsibility for the content of this publication.

Sources and notes

This paper names its sources in the text rather than in footnotes; this section collects the principal ones and adds publication detail so a reader can trace them. As Part V notes, the empirical figures are a snapshot of 2026 and the most perishable part of the paper. None of the model depends on them, and they should be refreshed on a regular cadence. The exhibits in Part I are illustrative, showing magnitude and direction rather than exact levels, and should be refreshed from the firm's own data.

The scale of AI spending. The 2026 hyperscaler capital-expenditure figures are the major providers' own guidance through the first-quarter 2026 earnings season, which put combined spending for the largest hyperscalers in the region of $700 billion for the year, with some estimates approaching $725 billion, more than sixty percent above 2025 (reporting by CNBC, Yahoo Finance, and sector analysts, February to May 2026).

Pilot-failure and return statistics. The ninety-five-percent figure is from the MIT NANDA initiative's *The GenAI Divide: State of AI in Business 2025* (August 2025), a widely cited study whose narrow definition of impact and small sample have drawn criticism and which is directionally corroborated by the others here. The abandonment figures are from S&P Global Market Intelligence's *Voice of the Enterprise: AI and Machine Learning, Use Cases 2025* (forty-two percent abandoning most initiatives, up from seventeen percent the prior year). The quarter-of-initiatives figure is from IBM's early-2026 enterprise study; the three-in-four figure from BCG's *AI at Scale 2026* survey (twenty-six percent reporting meaningful financial value); and the two-in-five enterprise-earnings figure from McKinsey's *The State of AI in 2025* (November 2025, thirty-nine percent reporting any enterprise-level EBIT impact). Gartner's estimate that roughly thirty percent of generative-AI projects will be abandoned by the end of 2026 is consistent with these.

The productivity paradox. The finding that roughly nine in ten firms reported no own-firm effect on productivity or employment while executives projected future gains is from *Firm Data on AI*, NBER Working Paper 34836 (February 2026; Yotzov, Barrero, Bloom, and collaborators), a survey of nearly six thousand executives across four countries.

Concentration as systemic risk. The warnings that reliance on a few shared models could turn a single shock into correlated failures draw on financial-stability work from the European Systemic Risk Board and the Bank for International Settlements through late 2025 and early 2026 on AI adoption and third-party concentration.

The single-system governance stack. The risk-and-controls references are the NIST AI Risk Management Framework, the ISO/IEC 42001 management-system standard, the EU AI Act (whose use-based high-risk compliance date was deferred by the mid-2026 Digital Omnibus amendments from August 2026 to December 2027), financial-services supervisory guidance issued in 2026, and the FINOS AI Governance Framework as the open, financial-services-specific controls layer (FINOS, licensed CC-BY-4.0), to which the model's controls crosswalk in Section 4.12 is mapped.

The 2026 portfolio convergence. BCG's board guidance is *Five Things Boards Need to Get Right with AI* (February 2026), which states the duty to treat AI investment as a portfolio rather than a collection of projects, and its *AI Radar 2026* survey (January 2026, roughly 2,400 executives) is the source for corporate AI spending doubling to about 1.7 percent of revenues and for 72 percent of chief executives naming themselves their company's main decision maker on AI. Gartner's allocation language is from its Finance Symposium statements of March and May 2026, on AI as a portfolio of very different bets and on measuring progress by realized value rather than deployment volume. The Vista Equity Partners practice of scoring portfolio companies on AI adoption and tying the scores to capital allocation was reported by the Financial Times in November 2025; the portfolio-wide programs and hyperscaler partnerships at Vista and Thoma Bravo are from company announcements and reporting by CNBC and Bloomberg through the first half of 2026. The assumption of a forty to fifty percent failure rate across initiatives is from practitioner guidance on AI capital allocation published in 2026.

Supervisory expectations on AI risk appetite. The Australian Prudential Regulation Authority's letter to regulated entities on AI governance and risk management is from April 2026; the Reserve Bank of India's draft guidance on regulatory principles for model risk management, which covers AI and machine-learning models and places approval of model risk appetite with the board, was released for public comment in June 2026.

The strategic-intent taxonomy. The defend, extend, and upend cases, with their return-on-employee, return-on-investment, and return-on-the-future framing, are Gartner's, adopted here under the names Parity, Advantage, and Frontier (Gartner research on setting enterprise AI ambition and on calculating GenAI business value).

The buyer-side evaluation literature. The per-investment treatments referenced are the Harvard Business Review work on what drives returns on AI by Davenport and Srinivasan, the Wharton executive-education material on why much AI investment is not yet paying off, and a recent working paper proposing a risk-adjusted AI return framework that integrates a management standard and regulatory exposure and presents returns as distributions.

Cost and token economics, and the exhibits. The consumption, context-composition, and cost-per-task claims, and Figures 1 to 3, are the author's analysis, drawing on public sources including OpenRouter consumption data, the FinOps X 2026 keynote, SWE-bench context-engineering measurements, and observed deployments. The exhibits are directional constructions from that data, built to show order-of-magnitude behavior and its decision implications, not any single firm's spend. The open standard for AI billing named in Part I and 1.6 is FOCUS (the FinOps Open Cost and Usage Specification), ratified at version 1.4 in June 2026, which normalizes AI, cloud, and SaaS cost including token economics; the context is the Foundation's 2026 shift to the value of technology and a token-economics effort announced at FinOps X 2026.

The two episodes. Uber's exhaustion of its 2026 agentic-coding budget within four months, and the per-engineer spending cap it then imposed, were reported by The Information and TechCrunch. The June 2026 export-control action requiring Anthropic to disable Fable 5 and Mythos 5 is set out in the company's 12 June statement and coverage by Fortune and CNBC.