



An SHI perspective on AI value economics

AI cost and value management

A two-lens operating model to control AI spend and prove business value



Introduction:

Why AI economics has become a board-level issue

Enterprise AI moved from experimentation to everyday use with unusual speed. Since the launch of ChatGPT in late 2022, generative AI has shifted from a breakthrough technology story to an operational reality. Organizations are now deploying AI to improve productivity, accelerate decision-making, enhance customer experiences, automate work, and create entirely new capabilities.

As AI moves from pilots into production, many organizations are discovering that managing AI is fundamentally different from managing traditional software. Costs are no longer tied to a single licence or platform. Instead, they emerge across the entire lifecycle of an AI service, from infrastructure and model consumption to governance, oversight, and ongoing operations. As adoption grows, spending can spread across teams and business units faster than organizations can establish ownership, visibility, and accountability.

This has created a new challenge for technology and business leaders. The question is no longer simply whether AI is delivering useful outputs. It is whether organizations understand what AI really costs, who owns that spend, and whether the investment is creating measurable business value.

To date, much of the conversation around AI economics has focused on token consumption. Tokens remain important because they provide visibility into how AI services are being consumed. They can help identify inefficiencies, inform optimization efforts, and reveal how architecture choices influence spending. But token consumption represents only one part of the economic picture.

The emerging discipline of AI economics increasingly focuses on **Total Cost of AI (TCA)**. Rather than concentrating on a single consumption metric, TCA looks at the full cost of delivering AI capabilities. It captures not only what organizations spend to run AI, but also what they invest to build, govern, secure, and support it. Looking at token costs alone can therefore create a misleading picture of the true economics of AI.

Cost is only half of the equation. An AI initiative can be highly efficient and still fail to create meaningful value. Equally, a more expensive workload may be entirely justified if it delivers a trusted outcome that improves the business. Understanding cost is important, but understanding value matters just as much.

This is why the conversation is increasingly shifting away from individual consumption metrics and towards a broader question:



What is the total governed cost of producing a useful, trusted business outcome, and can we demonstrate that the value created justifies the investment?

Answering that question requires two perspectives. Organizations need a **financial view** that explains where costs originate, how they are managed, and whether spending is defensible. They also need an **enterprise view** that evaluates whether AI investments are trusted, adopted, aligned to business priorities, and producing measurable outcomes.

This ebook introduces SHI's two-lens operating model for AI cost and value management. Together, the financial and enterprise lenses provide a practical framework for understanding Total Cost of AI, connecting investment to outcomes, and making better decisions about where AI should be optimized, scaled, redesigned, or retired.

Table of contents

1. What is AI tokenomics	5
2. The two lenses to evaluate true AI value	6
3. The AI economics value ladder	8
4. The AI cost stack: where spend is created	9
5. The operating model mechanics: how the two lenses work together	10
6. Roles, decisions rights, and shared accountability	11
7. The AI economics maturity journey: practical steps to move from visibility to value	12
8. Common failure patterns and control tests	16
9. How SHI helps	17

1. What is AI tokenomics?

The [Tokenomics Foundation](#) – established in June 2026 – recently updated its original definition of AI token economics as “*the emerging discipline of converting energy and capital into AI capabilities and efficiently consuming that intelligence across the organization to realize measurable business value.*”

The latest work of the Tokenomics Foundation increasingly positions tokenomics as a framework for understanding Total Cost of AI and value realization, rather than focusing exclusively on token consumption.

[SHI is a founding member of the Tokenomics Foundation](#) and is contributing to the standards conversation. But enterprises cannot wait for definitions, benchmarks, and ownership models to fully mature before they act. AI is already becoming material enough to require formal cost governance before spending scales further.



The shift the market is living through

1. From licensing to consumption.

The good old days of per-user software as a service (SaaS) pricing is giving way to token-, request- and agent-based pricing. Your costs are now a direct consequence of your design choices, including AI model selection, routing, retrieval depth, context windows, and validation steps, rather than a line on a purchase order.

2. From IT spend to board-level scrutiny.

Some 98 percent of finance and engineering leaders surveyed in the [2026 State of FinOps report](#) now actively manage AI spend, and 78 percent of those teams report into the chief information officer or chief technology officer. This means AI cost is no longer a technology cost; it is a corporate cost with corporate-level accountability. AI has definitely got boardroom attention.

3. From market definition to governance decisions.

Industry groups, foundations, and emerging standards initiatives are helping establish common language, metrics, and operating practices. Definitions, benchmarks, and ownership models are continuing to evolve. Customers must make AI investment and governance decisions while that work is underway.

4. From adoption to visibility pressure.

[Flexera's 2026 State of ITAM](#) findings show that 84 percent of organizations cite AI adoption and tracking as a top challenge, while only 31 percent have visibility into AI software, resulting in “shadow AI,” where employees use unapproved AI tools without oversight from IT or security teams. In short, AI is scaling faster than the operating model around it.



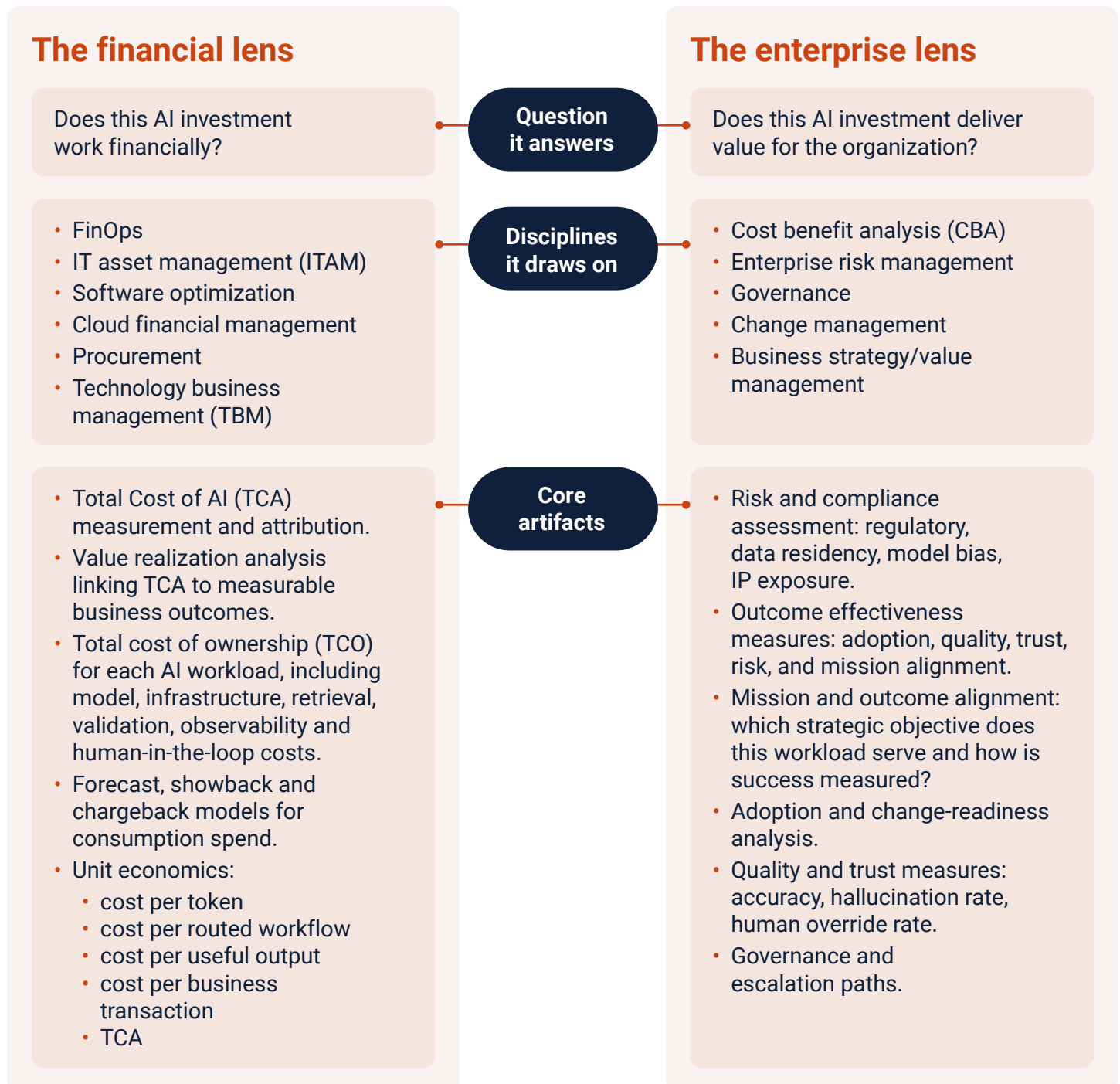
The result is a generation of AI investments being defended on the wrong metric of cost per token, measured against no agreed benchmark, and owned by no named business unit.

2. The two lenses to evaluate true AI value

AI value needs two lenses – a **financial lens** and an **enterprise lens**.

Cost discipline is necessary, but it cannot prove on its own whether an AI investment is useful, trusted, adopted, compliant, or aligned to business priorities.

- SHI defines **the financial lens** as the discipline of making the total cost, consumption, ownership, and unit economics of AI investments visible, attributable, optimizable, and financially defensible.
- SHI refers to **the enterprise lens** as AI Value Economics: the discipline of measuring and governing the business value created by AI investments to inform decisions about where to scale, redesign, or retire them.



The financial lens

- ROI and payback analysis at workload and portfolio level.
- Optimization mechanics: model routing, caching, prompt efficiency, vendor negotiation.
- Anomaly management

AI economics function (FinOps, ITAM, Procurement, or equivalent) with executive sponsorship.

Core artifacts

Primary owner

The enterprise lens

- Architecture decision review: which model, routing pattern, retrieval depth, validation steps, and human review points are needed for the workload to be useful, trusted and economically viable.

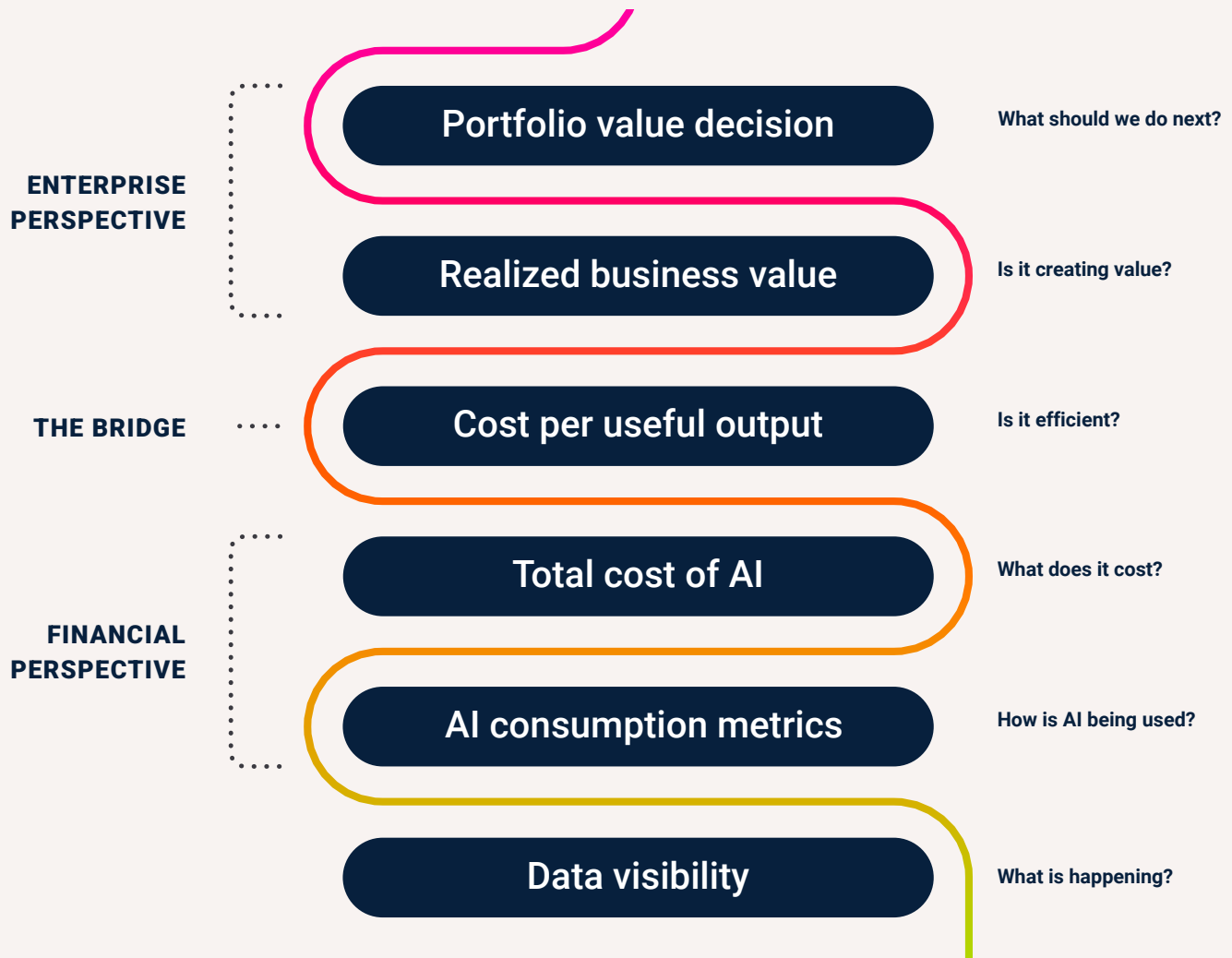
An AI governance or value-realization function with direct line of sight to the executive team.

The defining principle

Neither lens outranks the other. The financial lens prevents economically indefensible investments. The enterprise lens prevents financially defensible but strategically misaligned decisions. The board needs both answers in the same conversation, from people who can be held accountable for each.

3. The AI economics value ladder

Board-level decisions depend on the full ladder, not a single rung.



The value ladder shows how organizations move from measuring AI activity to measuring business value. Consumption metrics remain important because they provide visibility into how AI is being used, but they are only the starting point. The executive conversation happens further up the ladder, where organizations connect Total Cost of AI to useful outputs, realized business value, and portfolio decisions.

Value only counts when outputs meet an agreed quality threshold. A lower-cost output that fails quality expectations does not create value. The goal is not simply to reduce cost. It is to achieve the required outcome at the lowest defensible cost.

4. The AI cost stack: where spend is created

The AI cost stack is not just a list of technical cost drivers. It shows how AI spend moves from infrastructure and model consumption into telemetry, financial control, and enterprise decision-making. Cost is created in the technical layers, made visible through telemetry, governed through FinOps, and judged against business value through the enterprise lens.

Layer	What drives cost or control need	Why it matters
Infrastructure and platform	GPU usage, compute efficiency, workload placement, storage, networking, observability, security, and platform services.	Establishes the base cost of running AI workloads.
Model and token consumption	Model choice, token pricing, prompts, completions, context windows, embeddings, and routing complexity.	Makes AI consumption visible, but not sufficient as a value metric.
Data retrieval, grounding, and movement	Retrieval depth, vector search, data preparation, grounding sources, data quality, and repeated queries.	Poor data quality can increase cost and reduce usefulness.
Application and orchestration	Agent loops, tool calls, workflow automation, routing, caching, prompt design, validation steps, and retries.	Workflow design can multiply consumption and hidden costs.
Telemetry and attribution	Missing instrumentation, inconsistent usage signals, weak workload mapping, unclear ownership, and limited consumption-to-spend linkage.	Without telemetry, spend cannot be attributed, governed, or optimized.
Financial/FinOps	Forecasting, budgeting, showback, chargeback, unit economics, optimization rhythm, and cost-control decisions.	Without FinOps discipline, AI spend cannot be defended financially.
Enterprise decision	ROI uncertainty, risk, compliance, adoption, quality, trust, architecture decision quality, and outcome alignment.	Without enterprise review, teams may optimize spend without proving value.

The stack explains what needs to be managed before organizations assign ownership or move through the maturity journey. It also shows why cost per token is useful but incomplete. The economic question is not the price of one input. It is the total governed cost of producing a useful, trusted business outcome.

5. The operating model mechanics: how the two lenses work together

The operating model has four continuous mechanics: visibility, financial defensibility, enterprise decision quality, and integrated review. These are not separate projects or one-time phases. Visibility gives both lenses the evidence base. Financial defensibility tests whether spend can be justified. Enterprise decision quality tests whether the investment is useful, trusted, governed, and aligned. Integrated review brings both outputs together into a portfolio decision.

Visibility

Visibility starts with a workload-level inventory. For each priority AI workload, teams should know who owns it, what is driving consumption, and what business outcome it is expected to improve.

Financial defensibility

Financial defensibility applies cost discipline once visibility exists. This includes workload-level forecasting and budgeting, showback or chargeback, optimization opportunities, and unit economics over time.

Enterprise decision quality

Enterprise decision quality answers the questions financial data cannot answer on its own: whether the workload is aligned to business objectives, whether risks are being managed, whether adoption is sufficient, and whether the outcome justifies continued investment, redesign, scale, or shutdown.

Integrated review

Integrated review brings the financial and enterprise outputs together in a recurring executive review cycle. The output is a portfolio decision, not a cost report.



6. Roles, decision rights, and shared accountability

The organizational home for AI Economics varies by enterprise maturity and operating model. Some organizations may anchor ownership in FinOps, ITAM, Procurement, Enterprise Architecture, Strategy, or a dedicated AI Governance Office. Regardless of structure, a single executive sponsor must be accountable for integrated cost and value decisions.

The AI tokenomics operating model only works if accountability is named.
The minimum viable role map includes:

AI Economics Lead (FinOps, ITAM, Procurement or equivalent)

- Owns the financial perspective.
- Reports into CFO or CIO.
- Accountable for tokenomics discipline, forecasting, optimization.

AI Value Orchestrator

- Owns the enterprise perspective.
- Reports into CIO/CTO with a dotted line to risk or strategy.
- Accountable for the cost-benefit analysis, risk posture, adoption, architecture decision quality and outcome measurement of each workload.

Workload business owners

- Named for every AI workload.
- Accountable for the outcome the workload is supposed to deliver.

Executive sponsor

- Single name.
- Owns the integrated review and the portfolio decisions that come out of it.

Organizations that try to run AI economics without naming these four roles end up with dashboards no one acts on.



7. The AI economics maturity journey: practical steps to move from visibility to value

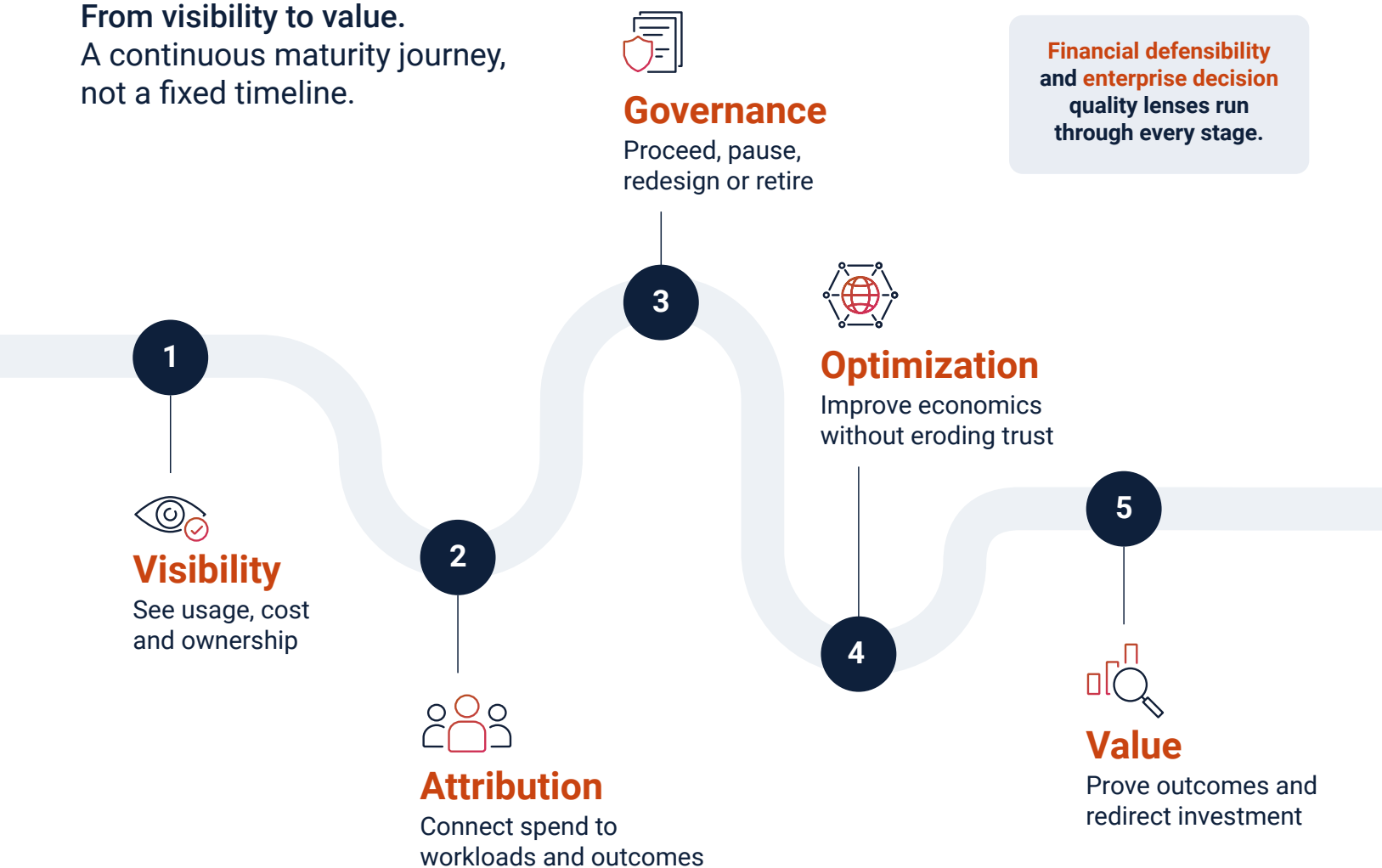
Organizations begin at different levels of AI maturity and progress at different speeds. The path to effective AI economics is therefore better understood as a practical maturity journey than as a fixed implementation timeline.

The journey moves from understanding where AI is being used and what it costs, to connecting that consumption with ownership, decisions, and measurable business outcomes. Each stage builds the capabilities needed for the next, but the journey is not strictly linear. As AI workloads, technologies, risks, and business priorities change, organizations will revisit earlier stages and refine the operating model.

Each stage follows the same pattern: what the stage means, what to do, and the decision it enables.

The AI economics journey

From visibility to value.
A continuous maturity journey,
not a fixed timeline.





1. Visibility

The journey begins with a clear view of the organization's AI environment. Before leaders can govern or optimize AI spend, they must understand where AI is being used, how consumption is occurring, and who is accountable for each workload.

What to do:

- Inventory sanctioned, embedded, and shadow AI workloads.
- Instrument consumption by workload, workflow, team, and business unit.
- Identify the models, platforms, infrastructure, data, applications, agents, and AI-enabled services supporting each workload.
- Assign named business and technical owners and establish the pre-AI baseline each workload is intended to improve.

Decision enabled:

Where is AI being used, what is it costing, and who is accountable for it?



2. Attribution

Visibility shows that consumption exists. Attribution explains what is causing it, who owns it, and what the organization expects to receive in return.

What to do:

- Map AI consumption and costs to workloads, teams, business functions, and owners. Unlike cloud, AI costs are influenced not only by consumption, but also by model choices, workflow design, and the quality of the outcomes produced.
- Connect each workload to its intended business objective or mission outcome.
- Establish unit economics, such as cost per request, workflow, resolved task, or useful output.
- Identify the architecture and design choices influencing cost, quality, latency, and risk.

Decision enabled:

What is creating the spend, who owns it, and what value is it intended to produce?



3. Governance

Optimization improves individual workloads. Governance creates a repeatable way to make decisions across the AI portfolio.

What to do:

- Establish decision rights across FinOps, architecture, governance, risk, security, and business ownership.
- Define intake, approval, exception, escalation, and retirement criteria.
- Apply appropriate financial, architectural, risk, compliance, and readiness guardrails.
- Run a recurring integrated review that uses both the financial and enterprise lenses to make portfolio decisions.

Decision enabled:

Which AI investments should the organization continue, change, scale, or stop?



4. Optimization

Once consumption can be attributed, the organization can identify where changes will produce the greatest benefit. Optimization should not be treated as cost reduction alone. A lower-cost workload is not better if the change reduces quality, trust, adoption, security, or business usefulness.

What to do:

- Evaluate the optimization levers available for the workload. For model-based AI this may include model selection, routing, caching, retrieval, context size, validation, and workload placement. For AI-native applications it may include licence management, feature adoption, application rationalization, and vendor optimization.
- Identify inefficient prompts, repeated requests, unnecessary model calls, and avoidable retries.
- Compare architecture alternatives using cost, quality, latency, security, and operational criteria.
- Validate that savings do not shift cost or risk into another layer of the AI environment.

Decision enabled:

Where can the organization improve AI economics without undermining usefulness, trust, or business outcomes?



5. Value

The final stage connects governed AI consumption with demonstrated business outcomes. Value is not defined solely by lower token costs or technical efficiency. It is determined by whether the workload produces a useful, trusted, adopted, and strategically relevant outcome at a defensible cost.

There are different categories of AI value:

- | | |
|--|---|
| 1) Direct financial impact <ul style="list-style-type: none">• New revenue• Revenue retained• Spend removed• Spend avoided• Capital freed | 3) Business outcomes <ul style="list-style-type: none">• Product quality• Speed to market• Risk reduction• New capabilities |
| 2) Labour impact <ul style="list-style-type: none">• Capacity gain• Augmentation• Automation | 4) Emerging <ul style="list-style-type: none">• Employee engagement |

What to do:

- Compare post-AI performance with the pre-AI baseline.
- Measure financial and non-financial benefits at workload and portfolio levels.
- Validate whether intended users are adopting the capability in real workflows.
- Assess whether the outcome meets agreed quality, trust, risk, and compliance expectations.

Decision enabled:

Is the AI investment producing enough measurable value to justify continued investment and scale?



The journey is continuous

Value is not the end of the operating model. It creates new information that informs the next cycle of visibility, attribution, optimization, and governance.

As models, pricing structures, architectures, regulations, business priorities, and usage patterns evolve at pace, organizations must continually reassess both cost and value. The objective is not to complete the journey once. It is to establish a repeatable management discipline that keeps AI investments financially defensible and aligned with enterprise priorities.

The goal is not to move through the stages as quickly as possible.

The goal is to build the evidence, accountability, and governance needed to make better AI investment decisions at every stage.

8. Common failure patterns and control tests

Even organizations that adopt this model can trip up in predictable ways. Most failures are not technical. They come from running one lens without the other, or from measuring activity instead of outcomes. AI coding assistants and agentic systems deserve particular attention because they can scale consumption, automation, and operational complexity faster than governance, ownership, and value measurement practices mature.

The patterns below are the ones we see most often, and each one is avoidable if accountability and the metric ladder are in place from the start.

Blind spot	Control test
FinOps-only blind spot. Beautiful dashboards, falling cost per token, but no improvement in outcomes. The enterprise lens was never considered.	If cost per token falls but cost per resolved task, completed workflow, or compliant response rises, the program is optimizing the wrong metric.
Governance-only blind spot. Strong risk and compliance posture, but no idea what anything costs, and the board cannot defend the spend.	If a workload passes governance review but cannot show total cost, owner, forecast, or unit economics, the program is managing risk without financial defensibility.
Unowned workloads. Visibility exists but no named owner ever has to decide anything. Reviews become reporting, not governance.	If a dashboard flags rising spend, falling adoption, quality issues, or duplicated usage and no named business owner is required to proceed, pause, redesign, scale, or retire the workload, accountability is missing.
Developer productivity assumptions. Organizations assume AI coding assistants are generating value based primarily on adoption, without confirming whether they improve development outcomes.	If AI coding assistant adoption rises but cycle time, defect rates, developer throughput, or release velocity do not improve, the organization is measuring usage rather than demonstrated value.
Cost per token as the headline metric. Reinforces a supply-side conversation and leaves outcomes invisible.	If executive reporting leads with cost per token but does not connect that metric to cost per useful output, resolved task, compliant response, or business outcome, the organization is measuring input efficiency instead of enterprise value.
Waiting for industry standards to settle. The standards conversation will run for years. Customers who wait will spend that time defending the wrong metric while impacting IT budgets.	If AI workloads are scaling before the organization has baseline outcomes, named owners, consumption attribution, and a recurring integrated review, the program is waiting for external standards instead of building internal control.
Activity mistaken for value. AI usage increases, but realized business value does not. The organization is measuring inputs and outputs rather than outcomes.	If AI usage, token consumption, or request volume increases but no measurable business outcome improves, the organization is measuring activity rather than value.

9. How SHI helps

SHI helps organizations put AI value economics into practice while industry standards are still forming. We connect cost visibility, spend attribution, workload optimization, governance, and value measurement so leaders can control AI spend and make better portfolio decisions.

Our role is to help customers move from scattered AI usage and cost data to a decision-ready view of where AI spend originates, who owns it, what value it is intended to create, and which actions should happen next. SHI helps:

- Establish AI cost visibility across models, platforms, workloads, teams, and business units.
- Attribute spend to named owners, intended outcomes, and measurable business context.
- Optimize consumption without degrading quality, trust, compliance, or adoption.
- Connect FinOps, ITAM, procurement, software optimization, cloud, AI architecture, governance, and business stakeholders into one operating rhythm.
- Connect Total Cost of AI to business outcomes and CFO-acceptable measures of value realization.
- Turn AI cost and value signals into portfolio decisions to proceed, pause, redesign, scale, or retire workloads.

Gaining control of AI costs requires an ecosystem of capabilities. SHI uniquely brings experience across SaaS, licensing, cloud, AI, FinOps, ITAM, software optimization, and partner technologies to help customers build the visibility, governance, and decision rhythm required for AI value economics.

Together, this gives teams a practical way to control AI spend, defend investment decisions, and focus resources on AI workloads that produce measurable enterprise value.

\$1 billion
saved for SHI's
Spend Optimization
Services clients
in 2025.



We can help no matter where you are on the AI economics journey

Start with our AI tokenomics briefing.

This no cost session helps organizations understand the economics of AI, including how token consumption drives cost, where value is created, and which metrics matter most. It's useful for anyone struggling with growing AI costs and wants an independent perspective.

Move to our AI tokenomics workshop.

Ready to dive into your specific AI challenges? The SHI tokenomics workshop is where the two-lens model becomes specific to your environment. Offered in half- and full-day formats, each expert-led session guides you through the areas where support is most needed, including readiness, governance, token management, value realization, and security. Instead of spending weeks debating where to start, stakeholders leave the session with a shared vocabulary, a prioritized set of focus areas, and a documented baseline of where you stand today and what to tackle next.

We're ready to help you get maximum value out of your AI journey.

► [Get in touch.](#)

