

# Agentic Systems at Scale: Reliability Economics for High-Frequency Workflows

*An Operating Framework for Outcome Reliability, Human Escalation and Unit Economics*

**Authored by Chennakeshav Adya, Independent Researcher**

September 2026

---

## Abstract

*Agentic systems can plan, retrieve information, call tools, modify state, recover from errors and escalate to people across many steps. These capabilities can improve automation coverage while increasing latency, variable call counts, permission exposure, correlated failure and human exception demand. Current technical guidance also emphasises simple composable architectures, complete trajectory evaluation, clear stopping conditions, well-designed tools, environmental feedback and risk controls for actions that extend beyond text generation. A high-frequency workflow therefore needs operating and economic evidence at the business-outcome level before production scale. This paper develops an evidence-controlled framework for agentic-system reliability economics across forty modules. It connects business outcomes, architecture choice, task graphs, frequency, consequence, permissions, tool contracts, state, identity, data, stopping, retries, reliability propagation, common causes, semantic success, evaluations, trajectories, tool use, escalation, guardrails, injection, irreversible actions, human operations, latency, compute and tool cost, failure and escalation cost, reliability-cost frontiers, service levels, observability, incidents, change control, provider routing, scale scenarios, enterprise value, ninety-day execution and accountable close. Five figures, five tables, eight frequently asked questions and twenty-six primary or authoritative references support workflow-, provider-, firm- and period-specific review. The framework does not establish safety, security, legal compliance, service performance, financial return, enterprise value or suitability and does not replace authorised technical, cybersecurity, legal, regulatory, accounting, tax, valuation or investment advice.*

JEL Classification: D24, L86, M15, O32, O33

Keywords: agentic AI, AI agents, reliability economics, tool use, evaluation, latency, human escalation, unit cost, operational resilience

## 1. Define the business outcome

Specify the user, trigger, desired result, consequence, service boundary and measurable completion condition.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an outcome charter. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when teams optimise agent activity without a verified business outcome. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for define the business outcome should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **2. Choose the simplest architecture**

Compare deterministic automation, single model, fixed workflow, routed workflow and autonomous agent.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an architecture decision record. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when unnecessary autonomy increases latency, cost and failure surface. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for choose the simplest architecture should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **3. Map the task graph**

Decompose planning, retrieval, reasoning, tool use, validation, escalation and completion into observable nodes.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an agent call graph. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when the end-to-end service is treated as one opaque model call. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for map the task graph should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 4. Measure task frequency

Forecast requests, bursts, seasonality, retries, concurrency and downstream events by workflow.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a frequency and capacity model. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when a pilot unit cost is multiplied by volume without production distribution. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for measure task frequency should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 5. Classify consequence

Segment actions by reversibility, financial effect, customer impact, data sensitivity and regulatory exposure.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a consequence-tier schedule. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when the same autonomy is applied to low- and high-consequence work. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for classify consequence should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 6. Define permissions

Map read, propose, approve, write, spend, communicate, delete and administer authority for each tool.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a least-authority matrix. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when the agent receives broad credentials for implementation convenience. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for define permissions should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **7. Design tool contracts**

Specify purpose, inputs, outputs, validation, errors, idempotency, side effects and examples.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a controlled tool catalogue. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when ambiguous tool interfaces generate systematic agent mistakes. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for design tool contracts should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **8. Map state and memory**

Separate session context, durable memory, business records, caches and external state with ownership rules.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a state architecture map. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when stale or contaminated memory propagates across high-frequency work. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

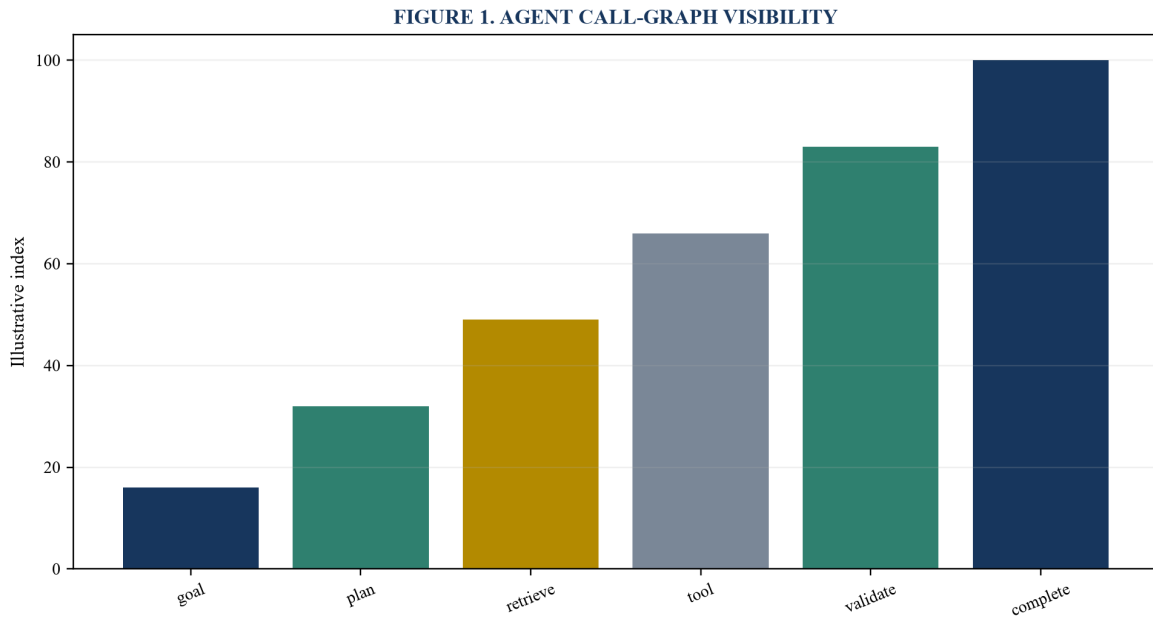
The decision pack for map state and memory should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

**Table 1. Agent workflow perimeter**

| Layer   | Core question            | Evidence        |
|---------|--------------------------|-----------------|
| outcome | was work completed       | business record |
| agent   | were decisions valid     | trace           |
| tool    | was action authorised    | tool log        |
| human   | was escalation effective | disposition     |

*Illustrative controls require workflow-, provider-, firm- and period-specific approval.*

**Figure 1. Agent call-graph visibility**



*Values are illustrative indices and require replacement with approved production evidence.*

## 9. Control identity

Bind user, agent, service, tenant, role, credential and delegated authority through every action.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an identity and delegation model. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when actions cannot be attributed to the correct principal. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for control identity should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 10. Control data access

Apply classification, minimisation, retrieval permissions, retention, encryption and audit to every node.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a data-access control map. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when workflow scale expands access beyond the initiating user's authority. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for control data access should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 11. Set stopping conditions

Define completion, uncertainty, iteration, time, cost, tool, failure and human-intervention limits.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a bounded-loop policy. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when open-ended loops consume resources and compound errors. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for set stopping conditions should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 12. Set timeout and retry policy

Assign timeouts, retryable errors, backoff, budgets, circuit breakers and compensation actions.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a retry and recovery specification. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when retry storms amplify provider or tool failure. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for set timeout and retry policy should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **13. Model call-chain reliability**

Calculate success across sequential, parallel, optional and redundant nodes using measured probabilities.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a reliability propagation model. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when high node accuracy is mistaken for high end-to-end completion. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for model call-chain reliability should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **14. Model failure dependence**

Identify shared models, providers, prompts, data, tools, networks and human queues.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a common-cause failure map. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when independent-failure assumptions understate correlated breakdown. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for model failure dependence should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **15. Define semantic success**

Measure correctness, completeness, groundedness, policy adherence and action validity by task.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a semantic acceptance rubric. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when transport success is counted as business completion. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for define semantic success should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **16. Build the evaluation corpus**

Collect representative, adverse, long-tail, multilingual, permission and recovery cases from approved work.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a versioned task suite. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when evaluation excludes the cases that dominate production loss. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for build the evaluation corpus should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

**Table 2. Reliability evidence**

| Dimension  | Measure           | Release gate     |
|------------|-------------------|------------------|
| semantic   | task success      | threshold        |
| trajectory | valid actions     | no critical path |
| recovery   | contained failure | exercise         |
| authority  | permitted action  | control proof    |

*Illustrative controls require workflow-, provider-, firm- and period-specific approval.*

**Figure 2. Reliability maturity**



*Values are illustrative indices and require replacement with approved production evidence.*

## 17. Evaluate trajectories

Score the sequence of decisions, tool calls, state changes and recovery, together with final outcome.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a trace-evaluation framework. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when final-answer grading misses unsafe or wasteful paths. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for evaluate trajectories should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 18. Evaluate tool use

Test selection, arguments, permissions, confirmation, idempotency and interpretation of results.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a tool-use conformance suite. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when correct reasoning triggers the wrong external action. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for evaluate tool use should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 19. Evaluate escalation

Test recognition of uncertainty, consequence, conflict, missing authority and anomalous state.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an escalation test pack. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when the agent continues when human judgement is required. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for evaluate escalation should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 20. Evaluate stopping

Test premature completion, looping, repeated failure, resource exhaustion and false success.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a stopping-condition test suite. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when the system terminates on activity rather than verified completion. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for evaluate stopping should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **21. Design guardrails**

Combine deterministic checks, model policies, environmental constraints and human approval at risk points.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a layered control design. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when guardrails exist only in natural-language instructions. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for design guardrails should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **22. Protect against injection**

Separate trusted instruction, untrusted content, tool output and external communication with containment tests.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an injection-resistance programme. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when hostile content redirects the agent into unauthorised action. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for protect against injection should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **23. Control irreversible actions**

Require preview, confirmation, dual control, staged execution, reconciliation and recovery evidence.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an irreversible-action gate. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when one model decision creates unrecoverable external change. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for control irreversible actions should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **24. Design human escalation**

Define triggers, queue, context package, authority, response time, fallback and disposition capture.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an escalation operating model. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when human review becomes an unmeasured operational bottleneck. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

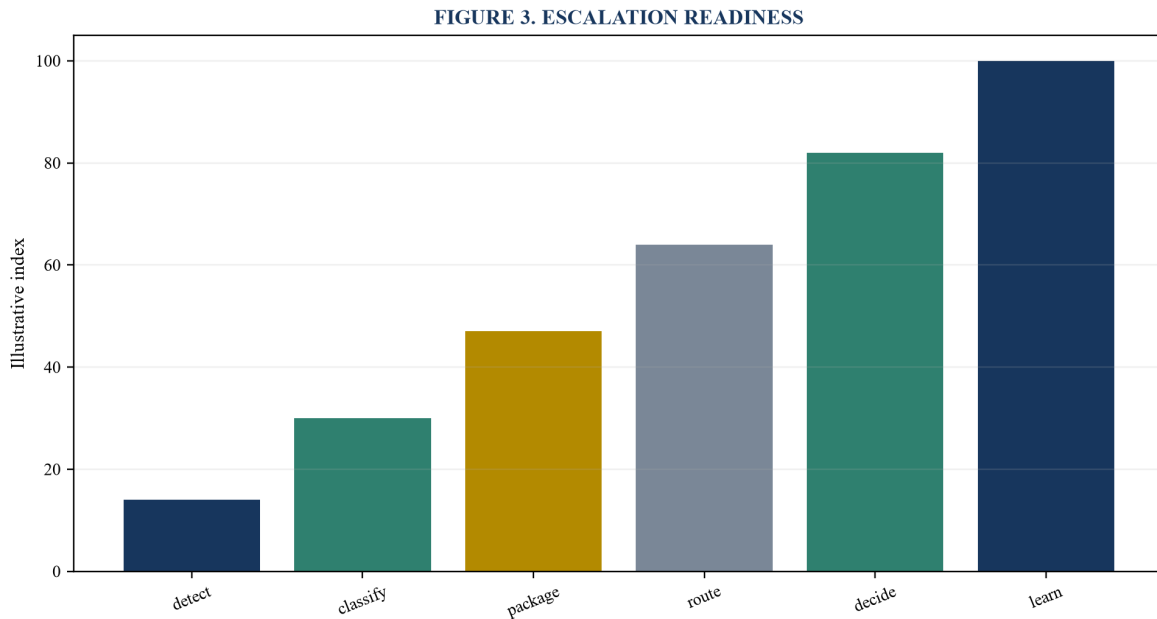
The decision pack for design human escalation should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

**Table 3. Escalation design**

| Tier  | Trigger              | Response            |
|-------|----------------------|---------------------|
| one   | low uncertainty      | agent recovery      |
| two   | bounded exception    | operator review     |
| three | material consequence | specialist approval |
| four  | control breach       | pause and contain   |

*Illustrative controls require workflow-, provider-, firm- and period-specific approval.*

**Figure 3. Escalation readiness**



*Values are illustrative indices and require replacement with approved production evidence.*

## 25. Measure human load

Forecast review minutes, exception rate, staffing, skill, coverage and fatigue at production volume.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a human-capacity model. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when automation shifts hidden work into an unaffordable queue. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for measure human load should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 26. Measure latency

Decompose queue, model, retrieval, tool, network, validation, retry and human time by percentile.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a latency budget. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when average response time conceals tail failure. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for measure latency should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 27. Measure token and compute cost

Track input, output, reasoning, cache, embedding, tool and retry consumption per completed outcome.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an outcome-level compute ledger. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when cost is reported per call while workflows contain variable call counts. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for measure token and compute cost should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## 28. Measure tool cost

Include search, database, API, browser, messaging, payment and third-party charges by trajectory.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a tool-cost schedule. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when non-model expenses erode the apparent automation margin. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for measure tool cost should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **29. Measure failure cost**

Quantify rework, refund, customer harm, remediation, incident, compliance and lost opportunity.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a failure-cost ledger. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when failed runs are excluded from unit economics. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for measure failure cost should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

## **30. Measure escalation cost**

Combine queueing, specialist time, delay, quality recovery and abandoned work.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an escalation cost model. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when human support is treated as free capacity. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for measure escalation cost should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### 31. Build reliability-cost curves

Compare architectures, models, evaluations, redundancy, controls and human review across outcome rates.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a reliability-cost frontier. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when teams pursue maximum technical performance without economic optimisation. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for build reliability-cost curves should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### 32. Set service levels

Define completion, quality, latency, availability, recovery and escalation objectives by consequence tier.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a tiered service-level schedule. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when one headline uptime target governs heterogeneous workflows. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

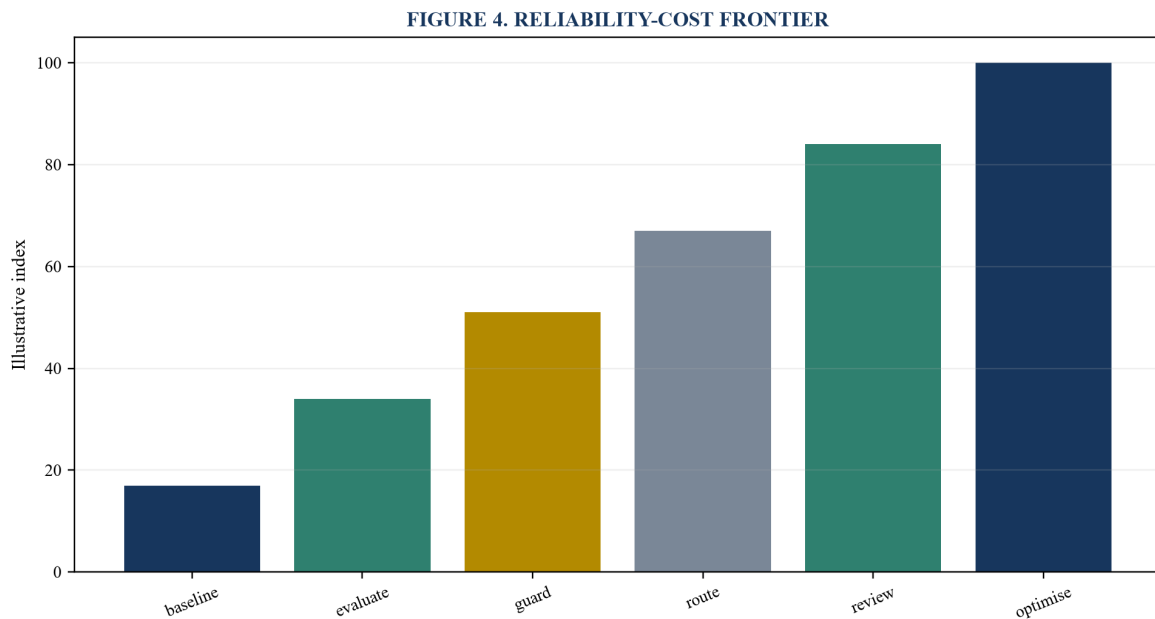
The decision pack for set service levels should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

**Table 4. Reliability economics**

| Driver  | Unit                    | Model treatment |
|---------|-------------------------|-----------------|
| compute | cost per outcome        | direct cost     |
| tools   | charges per trajectory  | direct cost     |
| human   | minutes per exception   | capacity cost   |
| failure | loss per failed outcome | risk cost       |

*Illustrative controls require workflow-, provider-, firm- and period-specific approval.*

**Figure 4. Reliability-cost frontier**



*Values are illustrative indices and require replacement with approved production evidence.*

### 33. Design observability

Join request, trace, model, prompt, state, tool, permission, cost, error, escalation and outcome identifiers.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an agent evidence ledger. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when teams cannot reconstruct why an agent acted. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for design observability should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### 34. Design incident response

Define detection, containment, authority revocation, customer communication, recovery and learning.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an agent incident playbook. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when agent incidents are handled as ordinary model defects. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for design incident response should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **35. Govern changes**

Require evaluation and approval for models, prompts, tools, permissions, data, routing and thresholds.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a controlled release process. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when silent changes invalidate reliability evidence. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for govern changes should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **36. Manage provider routes**

Test capacity, latency, quality, policy, data location, price and failover across permitted routes.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a route-equivalence matrix. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when provider switching restores service while changing control or economics. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for manage provider routes should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **37. Model scale scenarios**

Stress base, burst, degraded provider, tool outage, elevated escalation and adverse input mixes.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a scale stress model. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when production readiness relies on average pilot conditions. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for model scale scenarios should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **38. Translate into enterprise value**

Connect completion, labour leverage, margin, retention, liability, concentration and reinvestment to value.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a reliability-adjusted value bridge. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when agent reliability remains outside commercial and valuation decisions. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for translate into enterprise value should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **39. Execute the first ninety days**

Sequence outcome scope, graph, permissions, evaluations, observability, economics, exercises and controlled rollout.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is a 30-60-90-day scale roadmap. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when a broad platform programme delays evidence for the highest-value workflow. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for execute the first ninety days should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

### **40. Close through evidence gates**

Require reproduced outcomes, bounded authority, tested recovery, staffed escalation, viable economics and accountable residual risk.

The controlled record includes workflow; consequence tier; task node; model; prompt; state; tool; permission; call count; latency; result; error; retry; escalation; cost; evidence; owner; threshold and outcome. The immediate deliverable is an accountable scale decision. Measures should be reported per verified business outcome and by percentile or distribution where averages conceal operational risk.

The principal failure occurs when production scale is approved from demonstrations rather than evidence. Reviewers should reproduce the full trajectory, reconcile technical and commercial units, separate model, tool, state, provider and human causes, test recovery and quantify the effect on throughput, customer outcome, labour, cash cost, margin and risk.

The decision pack for close through evidence gates should show scope, baseline, architecture, observed trace, success criterion, failure mode, permission, control, escalation, capacity, latency, unit cost, downside, owner and approval. Material exceptions flow into product design, staffing, contracts, reserves, customer commitments, capital allocation, pricing and enterprise-value analysis.

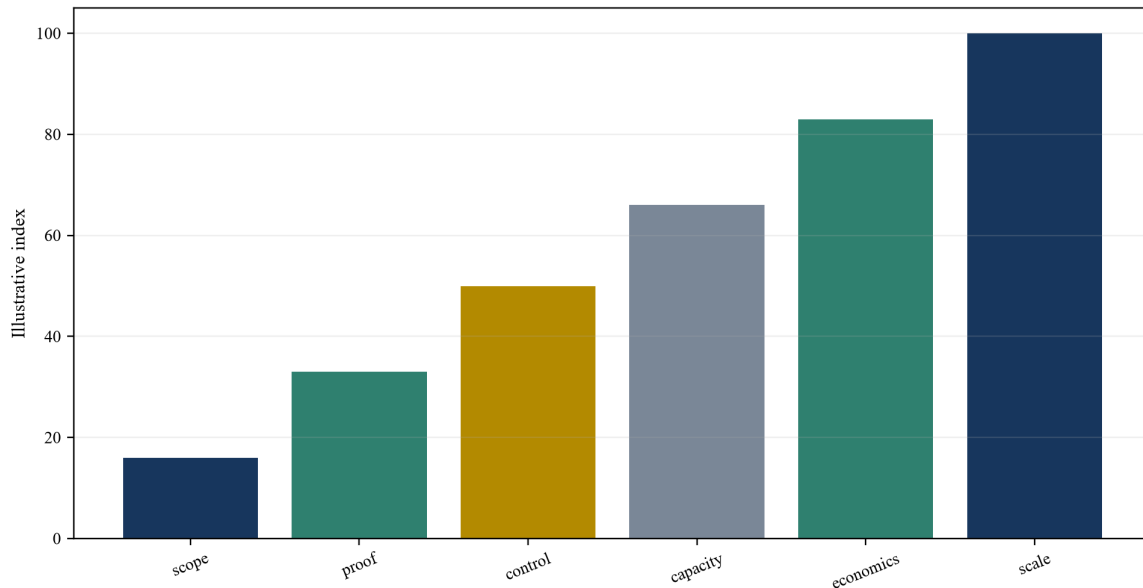
**Table 5. Scale close gates**

| Gate       | Required proof        | Decision             |
|------------|-----------------------|----------------------|
| outcome    | reproduced success    | scope accepted       |
| authority  | controls tested       | permissions accepted |
| operations | recovery and staffing | scale credible       |
| economics  | frontier approved     | investment signed    |

*Illustrative controls require workflow-, provider-, firm- and period-specific approval.*

**Figure 5. Production-scale gates**

**FIGURE 5. PRODUCTION-SCALE GATES**



*Values are illustrative indices and require replacement with approved production evidence.*

## References

- [1] National Institute of Standards and Technology, Agentic AI, <https://www.nist.gov/agentic-ai>
- [2] National Institute of Standards and Technology, AI Agent Standards Initiative, <https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative>
- [3] National Institute of Standards and Technology, Lessons Learned on Tool Use in Agent Systems, <https://www.nist.gov/news-events/news/2025/08/lessons-learned-consortium-tool-use-agent-systems>
- [4] National Institute of Standards and Technology, Strengthening AI Agent Hijacking Evaluations, <https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations>
- [5] National Institute of Standards and Technology, AI Risk Management Framework, <https://www.nist.gov/itl/ai-risk-management-framework>
- [6] National Institute of Standards and Technology, Generative AI Profile NIST AI 600-1, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
- [7] National Institute of Standards and Technology, AI Resource Center, <https://airc.nist.gov/>
- [8] Anthropic, Building Effective Agents, <https://www.anthropic.com/engineering/building-effective-agents>
- [9] Anthropic, Demystifying Evals for AI Agents, <https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents>
- [10] Anthropic, Trustworthy Agents in Practice, <https://www.anthropic.com/research/trustworthy-agents>

- [11] OpenAI, A Practical Guide to Building Agents,  
<https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf>
- [12] OpenAI, Agents SDK Documentation, <https://openai.github.io/openai-agents-python/>
- [13] OpenAI, Agent Evals Documentation, <https://platform.openai.com/docs/guides/agent-evals>
- [14] OpenAI, Trace Grading, <https://platform.openai.com/docs/guides/trace-grading>
- [15] OpenAI, Building Self-Improving Tax Agents with Codex,  
<https://openai.com/index/building-self-improving-tax-agents-with-codex/>
- [16] Google Cloud, Agent Development Kit Documentation, <https://google.github.io/adk-docs/>
- [17] Google Cloud, Agent2Agent Protocol, <https://cloud.google.com/discover/what-is-agent2agent-protocol>
- [18] Amazon Web Services, Generative AI Lens,  
<https://docs.aws.amazon.com/pdfs/wellarchitected/latest/generative-ai-lens/generative-ai-lens.pdf>
- [19] Microsoft, AI Agent Orchestration Patterns,  
<https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/ai-agent-design-patterns>
- [20] Cloud Native Computing Foundation, OpenTelemetry Documentation, <https://opentelemetry.io/docs/>
- [21] OWASP Foundation, Agentic AI Threats and Mitigations, <https://genai.owasp.org/>
- [22] MITRE, ATLAS Adversarial Threat Landscape for AI Systems, <https://atlas.mitre.org/>
- [23] European Union Agency for Cybersecurity, Cybersecurity Practices for AI,  
<https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai>
- [24] European Union, Regulation 2024/1689 Artificial Intelligence Act,  
<https://eur-lex.europa.eu/eli/reg/2024/1689>
- [25] ISO, ISO/IEC 42001 Artificial Intelligence Management Systems,  
<https://www.iso.org/standard/81230.html>
- [26] Google, Site Reliability Engineering Workbook, <https://sre.google/workbook/table-of-contents/>

## About the Author

Authored by Chennakeshav Adya, Independent Researcher

Chennakeshav (CK) is a corporate finance and investment banking executive with 25+ years of global experience in deal origination, structuring and execution across M&A, growth capital and corporate strategy. He has led value-creation mandates for founders, corporates and funds — bridging the boardroom view to hands-on execution and close.

His career spans Morgan Stanley, HSBC, Lloyds Banking Group, EWEC, ADQ portfolio companies and Emirates Growth Fund, across TMT, real estate, fintech, deeptech, cleantech, infrastructure and energy. He has partnered with C-suite leaders, private equity and venture funds, sovereign wealth funds and family offices to finance complex fund raises and scale-up ventures, and has led M&A due diligence, post-merger integration and business-transformation initiatives to create value.

At Matchpoint Partners he is Managing Partner, leading the firm's corporate finance, M&A and capital-raising practice. He holds an MBA from London Business School, an engineering degree from VTU and a Master of Laws (LLM, in progress) from UCL London.

An active start-up mentor, CK mentors at Techstars, DIFC FinTech Hive, Startup Grind, Founder Institute and IN5, serves as Entrepreneur Mentor in Residence (EMiR) at London Business School, and judges the Entrepreneurship World Cup.