GUIDE G3

TURN8

HOW TO **MONITOR**, **OPTIMIZE**, AND **REPORT** ON AI AGENT PERFORMANCE

AI Agent Deployment, Integration & Performance Management

WHERE THIS IS USED

- AI Studio agent deployments
- Venture Studio programs with AI-enabled ventures
- Corporate Incubators
- Foundry-as-a-Service engagements
- CVC portfolio companies with AI product components

AUDIENCE

- AI Studio Agent Leads
- BU Operations Owners
- Tech Leads
- Venture Studio Program Managers
- EIRs / GMs
- Venture Board and Executive Sponsors

PHASE

Phase Three: Build and Launch → Ongoing AI Agent Operations
(Week 10 onwards — continuous)

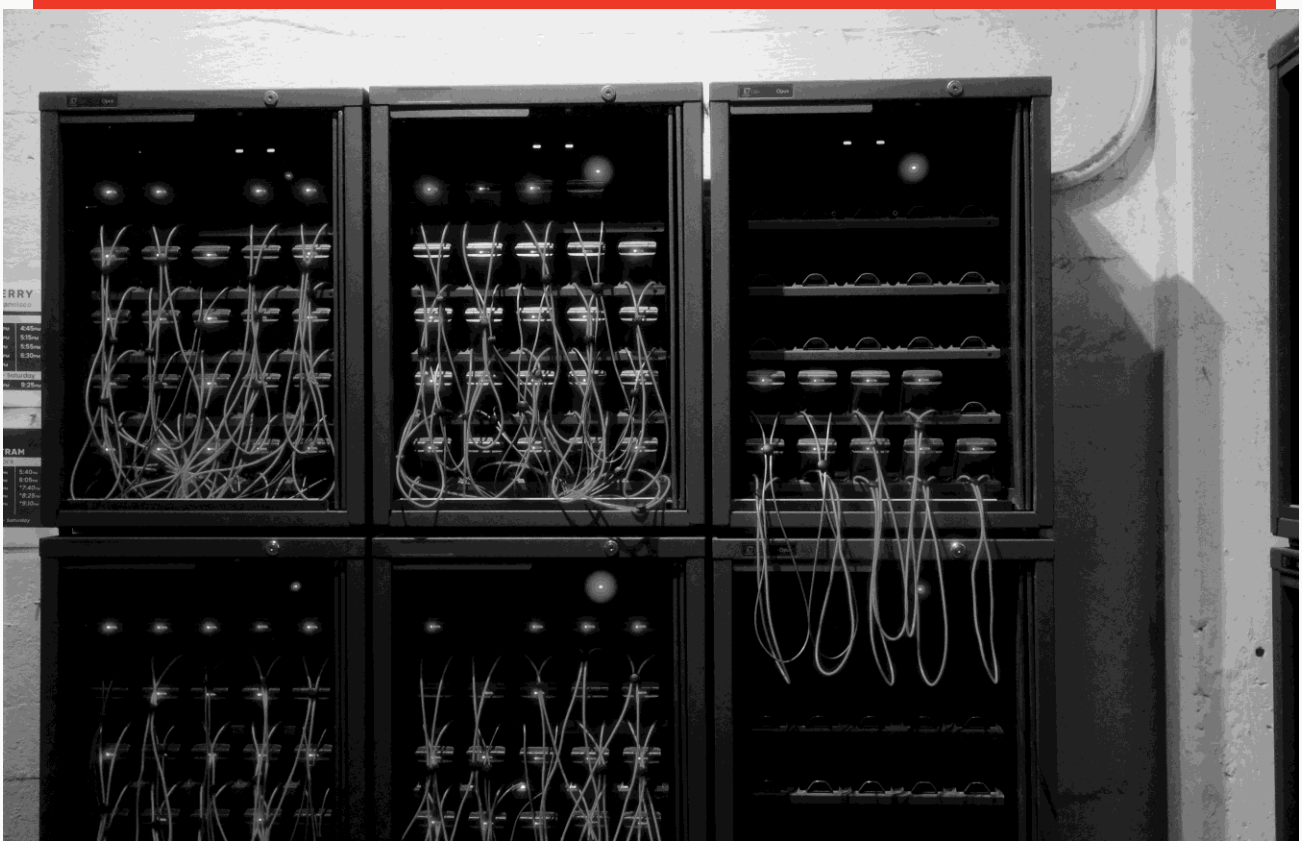
1

EXECUTIVE SUMMARY

An AI agent in production is not a deployed feature — it is a running experiment. The inputs it receives change over time. User behavior evolves. The underlying model may be updated by the API provider without notice. Business context shifts. Any of these can degrade performance without triggering a single alert — because degradation is often gradual, not sudden.

This guide covers the full ongoing operations cycle for a live AI agent: the weekly monitoring cadence that detects drift before it becomes a problem, the monthly review that translates operational data into improvement decisions, the prompt iteration protocol that safely updates system prompts in a live production environment without breaking running workflows, the three-condition decision framework for custom model training that prevents premature investment in infrastructure the use case does not yet justify, the quarterly Venture Board reporting format that presents agent performance in investment-case language rather than operational metrics, and the retirement framework that defines when an agent should be wound down rather than continuously patched.

The monitoring and optimization cycle begins on the first day of full production deployment — not after handover. G3 runs continuously alongside G2 from go-live. The BU Operations Owner is a participant in this cycle after handover acceptance, not a passive recipient of reports.



2

THE CORE PROBLEM

A deployed agent that is not actively monitored is degrading. Not because anyone made a mistake — because the world it was trained on is no longer identical to the world it now operates in. This is the defining characteristic of AI systems in production that distinguishes them from traditional software: a web application does not change its behavior when its users change their habits. An AI agent does.

The failure patterns in live agent operations are consistent:

- Accuracy drifts gradually — no single week triggers an alert, but over 8 weeks the agent is performing 15% below its deployment baseline. No one noticed because the monitoring was threshold-based, not trend-based.
- The system prompt is updated to fix a specific failure pattern, but the update is not regression-tested. It resolves the targeted pattern and silently degrades performance on a different input category. Both versions are now in production simultaneously — the change was not version-tagged.
- The escalation rate rises slowly. The Operations Owner interprets this as a user behavior change rather than a signal worth investigating. Six weeks later, the BU head notices that 30% of requests are being handled manually and asks why the agent is not working.
- The off-the-shelf API provider updates the underlying model without announcement. The agent's output format shifts. Users and the Operations Owner do not recognize this as a system-level change. The system prompt was written for the previous model version. No regression check was run.
- Custom model training is initiated because "the accuracy is not good enough." The three conditions have not been evaluated. The use case has not been validated at scale. The API ceiling has not been confirmed. The labeled dataset is insufficient. Six months of engineering effort produces a model that performs identically to the API it replaced.
- The quarterly Venture Board report presents agent uptime and escalation rates. The board asks how the agent is affecting business outcomes. No one can answer because business outcome metrics were never connected to the monitoring cadence.

The underlying issue:

Deployment was treated as completion. Monitoring was treated as maintenance. Both framings are wrong. Deployment is the beginning of the evidence-generation cycle. Monitoring is the mechanism that determines whether the agent continues to justify its operational cost and strategic purpose. This guide runs that cycle.

PREREQUISITES

Must Be Complete Before Starting:

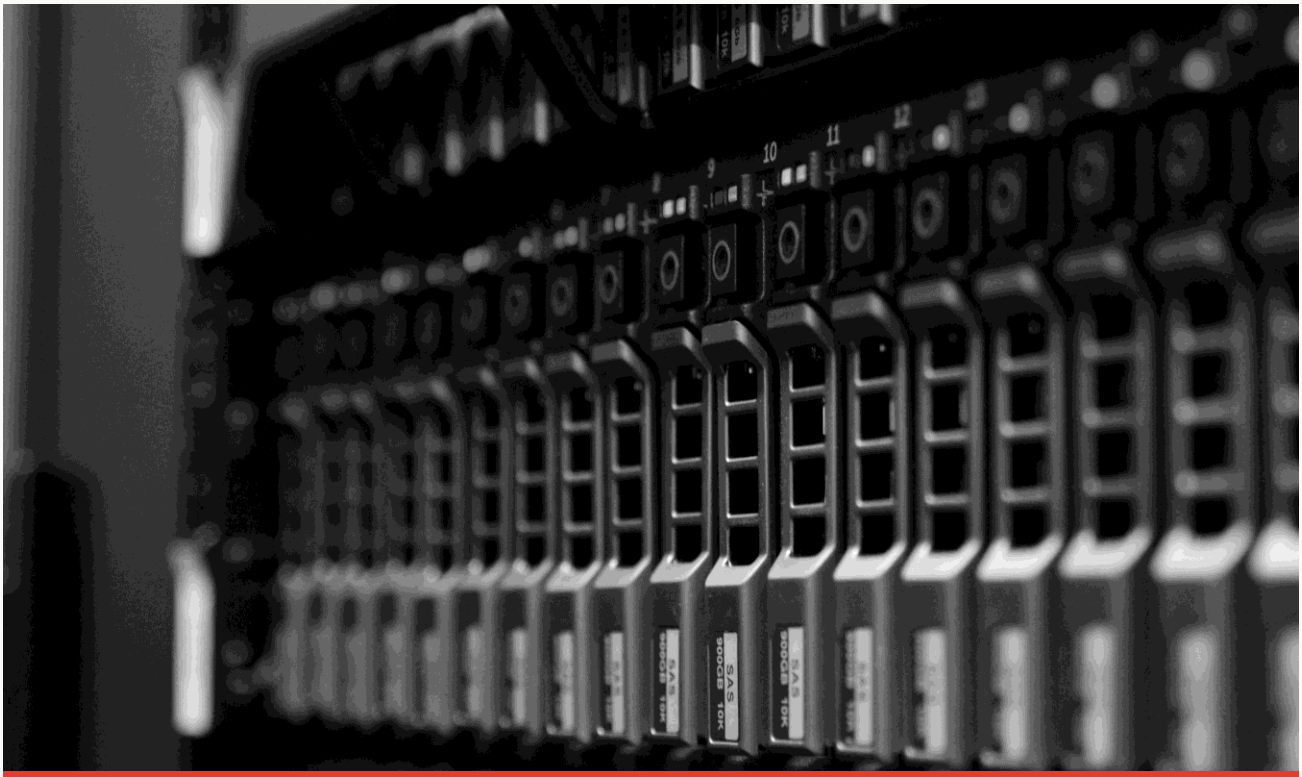
- G1 Go-Live Report filed — production monitoring dashboard live and verified, logging confirmed for all required fields including token usage, tool call traces, and safety flags
- G2 Handover Acceptance Certificate signed — BU Operations Owner has passed the readiness gate and is actively monitoring the daily dashboard
- Feedback mechanism active — users can log good/bad/change signals with minimal friction; feedback is captured in the logging system
- Agent configuration version-controlled — every system prompt is tagged; rollback is a version restore, not a manual re-edit
- For multi-agent or tool-using agents: monitoring must also cover tool-call intent, execution, and results, and cross-agent dependencies. A failure in a downstream agent or an unexpected tool action may not appear in the primary agent's accuracy metrics — it will appear in tool-call traces and downstream output quality.
- For regulated or high-risk agents: ongoing monitoring must satisfy traceability requirements — every AI-generated output is logged, attributable to a specific agent version, and auditable on request. Confirm with Legal that the logging configuration meets any applicable regulatory traceability standard before G3 begins.

4

EXPECTED OUTPUT/ SUCCESS CRITERIA

This guide is not completed — it runs continuously.
You are operating correctly when the following are always true:

- Weekly monitoring review completed every week by the Operations Owner — no week skipped
- Monthly performance review completed every month with the AI Studio Agent Lead — drift signals reviewed, improvement actions documented
- Every system prompt change processed through the prompt iteration protocol — no informal edits in production
- Quarterly Venture Board report produced for every board cycle — in the standard format, referencing the agent's contribution to business outcomes
- Three-condition gate evaluated before any custom model training investment is proposed
- Retirement criteria reviewed annually — agent either confirmed as continuing or a retirement timeline agreed



STEP-BY-STEP INSTRUCTIONS

STEP 1 RUN THE WEEKLY MONITORING REVIEW

The weekly monitoring review is the Operations Owner's primary operational responsibility. It takes 15–30 minutes. It uses the monitoring dashboard configured in G1. Its purpose is not to confirm that the agent is working — it is to detect early signals that performance is changing before those changes become problems.

The critical discipline: compare this week's metrics to the previous week AND to the 4-week moving average. Threshold-based monitoring catches sudden failures. Trend-based monitoring catches gradual drift. Both are required.

1.1 Review the seven core metrics in sequence — every week, in the same order

METRIC	WHAT DRIFT LOOKS LIKE	THRESHOLD ALERT	TREND ALERT (4-WEEK)	FIRST ACTION IF ALERT TRIGGERED
Accuracy rate (human-reviewed sample)	Outputs that were correct last month are now partially correct or require more user correction. As call volume grows, consider supplementing human sampling with automated eval scoring on stratified samples — lightweight judge-model scoring can extend coverage without proportionally increasing review time.	Below 85% on any weekly sample	> 5% decline vs 4-week average	Log the specific input patterns associated with the declined outputs. Do not revise the prompt until the pattern is identified.
Escalation rate	More inputs are being flagged as uncertain. Or fewer are — which may indicate the agent is becoming overconfident.	Above 25% or below 5% in any week	Upward trend > 3 consecutive weeks	Review 20 recent escalations. Are they genuine edge cases or inputs the system prompt should now handle? Classify before acting.
Response latency (p95)	Calls are taking longer to complete. Often the first signal of an upstream API change or infrastructure issue.	Above 3 seconds (SaaS) / 10 seconds (B2B complex)	Upward trend > 3 consecutive weeks	Check the API provider status page first. If no provider issue: review whether input length or tool call count has increased.

METRIC	WHAT DRIFT LOOKS LIKE	THRESHOLD ALERT	TREND ALERT (4-WEEK)	FIRST ACTION IF ALERT TRIGGERED
Cost per call	Each call is costing more. Often indicates longer contexts, more tool calls, or a model tier change by the provider.	3x above the established baseline	Upward trend > 2 consecutive weeks	Review token usage logs. Identify whether the cause is user input patterns (longer prompts) or agent behavior (more tool calls or longer outputs).
User feedback volume	Users are logging fewer corrections or good/bad signals. May indicate declining engagement with the agent.	Below 20% of sessions generating any feedback signal	Downward trend > 4 consecutive weeks	Flag to Operations Owner for user communication review. Review G2 feedback reinforcement cadence.
Safety / policy flag rate	More outputs are being filtered by content safety systems. Or the pattern of flagged content has changed.	Any week with > 1% of calls flagged	Any increasing trend	Review the specific inputs and outputs associated with flags. Classify: user input pattern change, or agent output pattern change?
Tool-call health (tool-using agents)	For agents that call external tools: tool calls failing at higher rates, unexpected tools being called, or action anomalies such as a write operation firing when only reads are expected.	Tool-call failure rate > 5% in any week; any unexpected tool invocation	Any upward trend in tool-call errors or anomalous action patterns	Review tool-call traces in the logging system. Distinguish between API failures (infrastructure issue) and unexpected tool selection (prompt logic issue). Escalate to Tech Lead for infrastructure failures immediately.
Business outcome metrics	The BU-specific metrics the agent is supposed to move: average handling time, error rate, task completion rate, or cost per transaction.	BU-defined — agreed at handover	Negative trend > 4 consecutive weeks vs baseline	Escalate to Operations Owner and BU Manager. A business metric declining while technical metrics are stable indicates an adoption or workflow integration issue, not an agent performance issue.

1.2 **Classify every drift signal before acting on it** — Not every metric movement requires a system prompt change. Before any action: classify the signal

- **Type 1** — Input distribution shift: Real-world inputs have changed. Users are now submitting queries the original test set did not represent. The agent is handling them correctly given its current configuration — but the configuration is no longer calibrated for this input mix. Action: expand the test set. Evaluate whether a prompt update is needed.
- **Type 2** — System prompt gap: A specific input pattern is consistently producing incorrect or escalated outputs, and the pattern is within scope of what the agent should handle. The system prompt needs a targeted update. Action: follow the prompt iteration protocol (Step 3).
- **Type 3** — External change: The API provider has updated the underlying model. Behavior has shifted without any change to the configuration. Action: run the full regression test immediately. Pin the model version if the provider allows it.
- **Type 4** — User or workflow change: The business metric is declining but technical metrics are stable. Users have changed how they submit inputs, or the workflow context has changed. Action: review with the BU Operations Owner. This is a G2 training/adoption issue, not a G3 technical issue.

1.3 **Document the weekly review in the monitoring log** — Format: "Week of [date]. Metrics reviewed: [list]. Any alerts triggered: [yes/no — specify]. Drift signals identified: [describe or none]. Classification: [Type 1/2/3/4 or none]. Action taken or scheduled: [describe or none]." This log is the primary input to the monthly review.

1.4 **For higher-maturity setups: extend monitoring with input distribution awareness and anomaly detection** — Two additions that become practical once several months of production data exist. First: monitor whether the distribution of incoming inputs has shifted materially from the training and test set distribution — a pattern change in the types of queries the agent receives is a leading indicator of accuracy drift, appearing before accuracy metrics move. The Type 1 classification captures this directionally; organizations with higher monitoring maturity can add statistical distribution checks to confirm it. Second: beyond static thresholds and 4-week averages, basic anomaly detection on key metrics — accuracy, cost, tool-call error rate — can surface unusual patterns that neither threshold nor trend analysis would catch in time. Most observability platforms support this without custom development.

AI PROMPT — Weekly Monitoring Review Summary

```
I am completing the weekly monitoring review for an AI agent
in production. Agent: [name and function]. Week of: [date].
This week's metrics: accuracy rate [X%], escalation rate [X%],
latency p95 [X seconds], cost per call [X vs baseline],
feedback volume [X% of sessions], safety flags [X%], business
outcome metrics [describe]. Previous week: [same fields]. 4-
week moving averages: [same fields]. For each metric: (1) is
it within normal range, triggering a threshold alert, or
showing a trend alert? (2) If drift is detected: classify as
Type 1 (input shift), Type 2 (prompt gap), Type 3 (external
change), or Type 4 (user/workflow change). (3) What is the
recommended action? Output as a completed monitoring log entry
ready for the Operations Owner to file.
```

STEP 2 RUN THE MONTHLY PERFORMANCE REVIEW

The monthly review is a structured decision session between the BU Operations Owner and the AI Studio Agent Lead. It synthesizes four weeks of monitoring logs into three outputs: a performance trend assessment, a prioritized improvement backlog, and a decision on whether any action requires escalation to Level 2 support. It takes 60 minutes. It is not a status update — it produces decisions.

2.1 Prepare the monthly review inputs before the session

- **From the monitoring logs**: All seven metric trends across the past four weeks. Any threshold or trend alerts triggered. Any Type 1, 2, 3, or 4 classifications made.
- **From the user feedback log:** The top three user correction patterns from the past month. The frequency of each pattern. Whether each pattern has been seen before or is new.
- **From the escalation log:** The top three escalation trigger patterns. The resolution times. Whether any escalation patterns indicate a system prompt gap (Type 2) that has not yet been addressed.
- **From the business outcome metrics:** Current values vs. the baseline agreed at handover. Trend direction. The Operations Owner's assessment of whether the agent is delivering the business impact it was designed to deliver.
- **From bias/fairness indicators (regulated or high-impact agents):** For agents that affect different customer segments differently — for example, an agent used in credit assessment, hiring, or customer service prioritization — review segment-level performance data from the past month. Are there any customer cohorts experiencing meaningfully higher error rates or escalation rates than others? Flag for the monthly decision if so.

2.2 Structure the session around three decisions

DECISION	QUESTION	POSSIBLE OUTCOMES
Performance assessment	Is the agent performing better, the same, or worse than it was 30 days ago — across both technical metrics and business outcomes?	Stable: continue monitoring cadence. Improving: document what changed and replicate. Declining: classify the cause and move to Improvement Backlog.
Improvement backlog prioritisation	What are the top 1–3 changes that would most improve performance this month? For each: what evidence from the monitoring log and feedback log supports this change? What is the expected improvement?	Prompt update: proceed to the prompt iteration protocol (Step 3). Test set expansion: schedule before the next monthly review. Workflow or adoption issue: escalate to Operations Owner for G2 action. No changes needed: document and continue.
Escalation to Level 2	Is there any issue in the monthly review that exceeds the Operations Owner's authority to resolve — a system-level change, a governance issue, or an infrastructure problem?	Yes: the Operations Owner initiates a Level 2 request with the standard format from G2. No: all improvement actions are within Level 1 scope.

2.3 Update the improvement backlog and assign owners

— Every identified improvement from the monthly review is logged: description, evidence basis, expected impact, owner (Operations Owner or Level 2), and target completion date. The backlog is reviewed at the start of the next monthly session — not left open-ended. For Operations Owners managing multiple agents or reporting to a BU head who is not in the monthly review, maintain a one-page Agent Health Scorecard that aggregates the monthly outcome across 3–4 color-coded indicators — this feeds the quarterly board section directly and removes the need for the BU head to read the full review record.

AI PROMPT — Monthly Performance Review

```
I am preparing the monthly performance review for an AI agent
in production. Agent: [name]. Month: [month/year]. Four-week
metric summary: [paste weekly monitoring log entries]. Top 3
user feedback patterns: [list — description and frequency].
Top 3 escalation trigger patterns: [list]. Business outcomes
vs baseline: [describe]. For this review: (1) overall
performance assessment — better, stable, or declining vs 30
days ago, with evidence, (2) top 1–3 improvements to
prioritize, each with the evidence basis, expected impact, and
owner, (3) any items requiring Level 2 escalation with the
reason and the correctly formatted Level 2 request. Output as
a structured monthly review record ready for the Operations
Owner and AI Studio Agent Lead to sign off.
```

STEP 3 APPLY THE PROMPT ITERATION PROTOCOL

A system prompt update in a live production environment is a change to a running system that affects every subsequent interaction. It must be treated with the same discipline as any other production change: documented, version-controlled, tested against the full regression set before deployment, and deployed through the same configuration management process used in G1.

The most common failure in prompt management is the informal edit: the AI Studio Agent Lead updates the system prompt directly in the production configuration because a specific failure pattern is observed. The change fixes the pattern. It also changes the behavior of the agent for all other inputs in ways that are not immediately visible — and were not tested because the regression run was skipped.

3.1 **Confirm that a prompt update is the correct intervention** — Before writing a new prompt: is the identified issue a Type 2 signal — a system prompt gap where a specific input pattern is producing incorrect outputs that should be handled? If the issue is Type 1 (input distribution shift), Type 3 (external model change), or Type 4 (user/workflow change), a prompt update may not be the correct intervention. Clarify the classification before proceeding.

3.2 Write and evaluate the proposed prompt change in a staging environment

- **Scope of the protocol:** The protocol applies to any "prompt unit" in the agent configuration — not only the top-level system prompt. For agents with sub-agent prompts, tool descriptions, RAG instructions, or routing logic, each is treated as a separately version-controlled configuration item. A change to a tool description is subject to the same regression test requirement as a change to the system prompt, because tool descriptions shape how the agent decides to act.
- **Write the change:** Document the specific failure pattern the update addresses. Write the new system prompt element. Keep changes targeted: one concern per prompt iteration, not a wholesale rewrite.
- **Test in staging — including safety:** Run the full regression test set against the updated prompt. Target: > 90% accuracy on standard cases, no regression in any input category. Also run safety and policy tests: the updated prompt must not introduce new safety regressions — adverse input patterns, disallowed content scenarios, and prompt injection attempts from the test set must all be re-evaluated. A prompt update that improves accuracy but introduces a safety regression does not pass.
- **Compare against baseline:** The updated prompt must match or exceed the production prompt across all test categories including safety. An update that improves the targeted pattern while degrading any other category or any safety test has not passed.

3.3 Version-tag the updated prompt and record the change

- **Version tag:** Every production prompt unit is version controlled. The updated configuration receives a new version number. The change log records: the version number, the date, the specific failure pattern that triggered the update, the change made, the regression test result, the safety test result, and the approver.
- **Change record:** Logged in the Operational Handover Document (G2, Section 8) and the G3 improvement backlog.

3.4 Deploy the updated prompt through the production configuration management process

- **Deployment:** Via the same CI/CD or configuration management process used in G1. Rollback is a version restore to the previous tag.
- **Optional — A/B testing for high-impact agents:** For agents where the prompt change is material and the user population is large enough, consider A/B testing the new prompt variant on a small percentage of traffic before full rollout — after it passes offline regression. This provides production-level validation before the full canary increment. It is optional, not required, and should not delay rollout when the regression evidence is strong.
- **Post-deployment monitoring:** Monitor all eight core metrics for the first 48 hours after any prompt update. If accuracy degrades or any safety metric regresses: roll back immediately.

3.5 Never deploy a prompt update that has not passed the full regression test —

This rule has no exceptions for urgency. If a production failure requires an immediate response, the correct action is rollback to the last stable version — not a rapid untested prompt update. Fix in staging. Test. Then deploy.

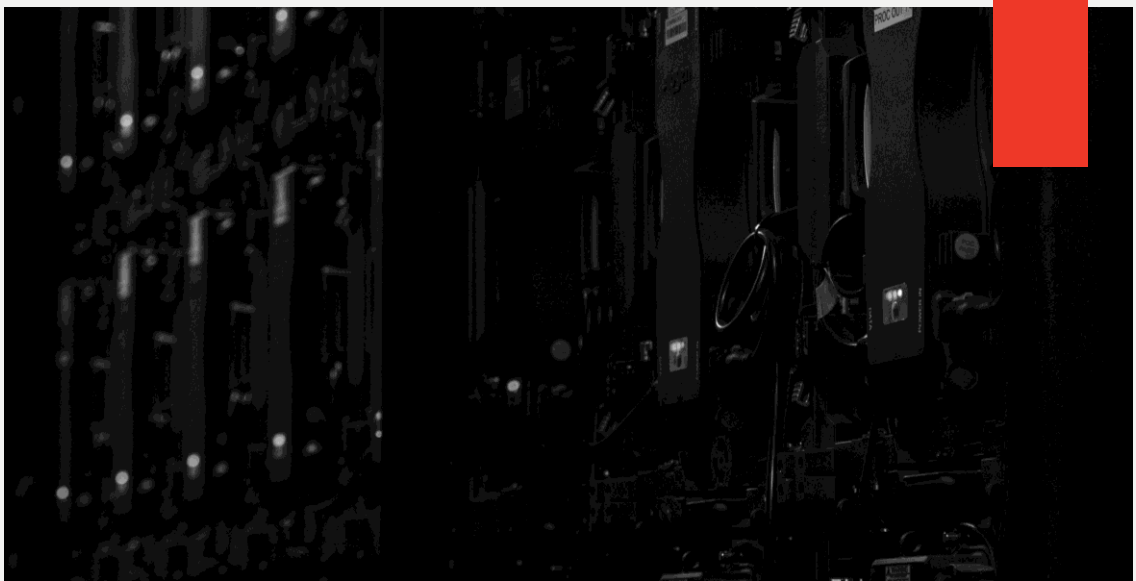
AI PROMPT — System Prompt Update

```
I am preparing a system prompt update for an AI agent in
production. Agent: [name]. Current prompt version: [version
tag]. Failure pattern identified: [describe — what specific
input pattern is producing incorrect or escalated outputs?].
Classification: Type 2 (system prompt gap). Current production
prompt: [paste]. For this update: (1) write the targeted
prompt change that addresses the failure pattern without
altering the handling of other input categories, (2) identify
the three test set categories most likely to be affected by
this change, (3) write the three regression test cases for the
targeted pattern that will confirm the fix, (4) write the
change log entry for the version control record. The updated
prompt must be testable against the full 50-input regression
set before deployment. Do not produce a wholesale prompt
rewrite — one targeted change per iteration.
```

STEP 4 APPLY THE CUSTOM MODEL TRAINING DECISION FRAMEWORK

Custom model training is the point at which the venture moves from operating an AI agent to building AI infrastructure. It is a significant investment of engineering time, data labeling effort, and ongoing maintenance. The decision to train a custom model must meet three conditions simultaneously — not just one or two. If any condition is not met, prompt iteration remains the correct intervention.

This framework was established in Guide F1 (the Build-Measure-Learn cycle for AI ventures) and applied throughout the sprint. In G3, the same three conditions are evaluated against accumulated production data rather than sprint-phase evidence. Production data is better evidence than pilot data — which means the conditions are now more rigorous, not less.



The Three-Condition Gate — all three must be confirmed before custom model training is proposed:

CONDITION	WHAT IT REQUIRES	HOW TO EVALUATE IT FROM PRODUCTION DATA	COMMON FAILURE — DO NOT TRAIN IF:
1. Use case validated at production scale	The agent is handling a real, recurring workflow that users depend on at meaningful volume. The use case is not a prototype — it is a production dependency.	Monthly active calls > [BU-specific threshold, typically 500+ calls per month]. Business outcome metrics are positive. At least 6 months of consistent production usage. The Operations Owner confirms the BU cannot operate the workflow without the agent.	The agent is used occasionally or only by a small number of users. The use case is still being refined. The BU has not fully adopted the agent as a workflow dependency.
2. Off-the-shelf API has hit a confirmed ceiling	The current API-based implementation has a specific, documented limitation that cannot be resolved through prompt iteration or through cheaper interventions. Before concluding that the API ceiling has been reached: exhaust the alternatives — better retrieval design (RAG), tool redesign, model routing to a different existing model, or prompt chaining. Only when these have been tried and documented does the API ceiling assessment carry weight.	Prompt iteration has been attempted at least 3 times for the specific failure pattern. Each iteration was regression-tested. Alternative interventions (retrieval, tool design, routing) have been evaluated. The accuracy on the targeted pattern has not improved beyond a stable ceiling. Or: cost per call at current volume is 2x+ the original model and trending upward.	The team believes "we can do better" without having exhausted cheaper options first. The failure pattern has not been addressed through prompt iteration or alternative interventions. The cost ceiling is an estimate, not a measured production value.
3. Sufficient labeled data exists for training — representative and unbiased	A custom model requires labeled training data. As a practical guide: at least 1,000 labeled examples for fine-tuning a general model. The labels must be high quality and representative across user segments. Check for skew: if labeled data comes primarily from override logs, it may over-represent certain user types or input patterns and under-represent others. Document any known representativeness gaps before training begins.	Review the override log (G2), feedback log (F1), and escalation resolution log (G1). Count labeled examples. Assess quality (human-reviewed, consistent) and representativeness (do the examples reflect the current production input distribution across all relevant user segments? are edge cases covered?).	The labeled dataset is < 500 examples. The labels are low quality or inconsistent. The examples do not reflect the current production distribution. The data is skewed toward a specific user segment or input type without documentation of that skew.

4.1 **Evaluate the three conditions formally at the monthly review — do not evaluate informally** — The decision to propose custom model training is a governance event. The three conditions are evaluated in writing. The assessment is filed in the improvement backlog. The AI Studio Agent Lead and the Operations Owner co-sign before escalating to the EIR/GM.

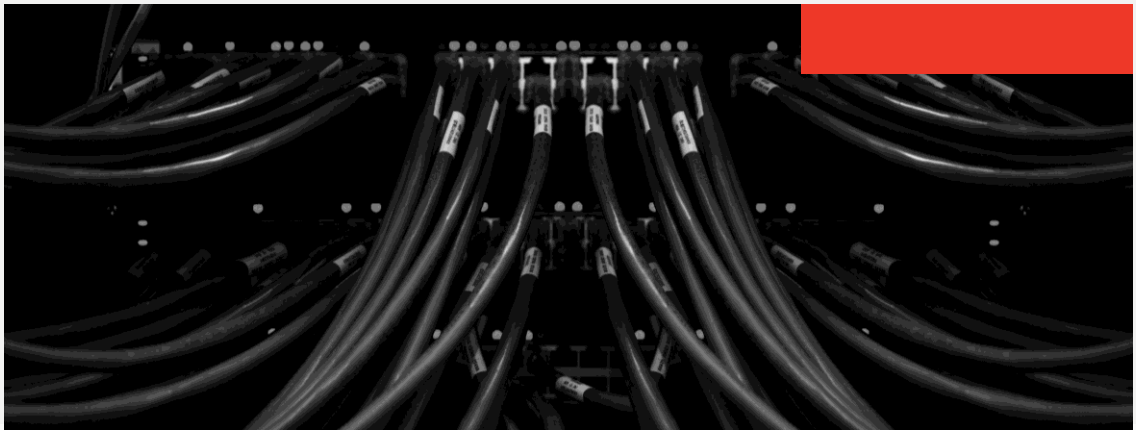
4.2 **If all three conditions are met: escalate to the EIR/GM as a capital and resource allocation decision** — Custom model training is a Phase Four investment decision — dedicated engineering time, data engineering work, compute budget, and ongoing maintenance. The EIR/GM presents the three-condition assessment to the Venture Board alongside an estimated cost, timeline, and expected performance improvement. For any approved custom model: produce a model card before deployment, documenting intended use, evaluation metrics, known risks and limitations, and training data provenance. Register the model in the enterprise AI catalogue if one exists. A model card is not optional for regulated sectors — it is the audit trail for the custom model decision.

4.3 **If any condition is not met: continue prompt iteration and defer the training decision** — Document the specific condition that was not met and the evidence that led to that assessment. Record the threshold that would trigger a re-evaluation. Continue monitoring. The three-condition gate is evaluated again at the next monthly review or when new production data materially changes any condition.

STEP 5 PRODUCE THE QUARTERLY VENTURE BOARD REPORT SECTION

The Venture Board receives a quarterly report on venture performance (from Guide E2). AI agent performance is one section of that report. The audience is the Venture Board — executive sponsors, BU heads, and investment decision-makers. They are not reading an operational dashboard. They are asking one question: is this agent contributing to the business case, and does its performance trajectory justify continued investment?

The quarterly report section translates operational metrics into strategic language. Every operational metric must connect to a business outcome. Uptime percentages and escalation rates are inputs. Cost savings, productivity gains, and revenue contribution are outputs. The board reads outputs. For ventures running several agents, the quarterly section may aggregate into a short portfolio view — the five active agents as a table, each with its current recommendation (continue / invest to improve / review for retirement) and its primary business impact metric. This allows the board to assess the AI program as a whole, not just individual agents in isolation.



5.1 Structure the quarterly AI agent performance section in four parts

PART	CONTENT	LENGTH	SOURCE
1. Performance Summary	One-paragraph assessment: is the agent performing better, the same, or worse than the previous quarter? State the primary metric trend and the direction of change. If the agent was updated during the quarter: what changed and what was the result?	1 paragraph	Monthly review records from the quarter
2. Business Impact	The specific business outcomes the agent contributed to this quarter. Use the business outcome metrics from the G2 Handover Document Section 3. State the baseline and the current value. Express impact in terms the board can connect to P&L.	3–5 bullet points with numbers	Business outcome metrics from the monitoring dashboard
3. Known Issues and Mitigations	Any performance issues identified during the quarter, the classification applied, and the mitigation taken. Explicitly include any material safety, bias, or compliance incidents — whether the incident occurred and how it was remediated and documented. A board that learns of a safety or compliance incident from outside the quarterly report loses confidence faster than one that was briefed proactively. No surprises.	1 paragraph or short table	Monthly review improvement backlog; incident log
4. Recommendation and Forward Outlook	One of three positions: Continue as deployed, invest to improve (specific change with cost and expected outcome), or Review for retirement. State the basis for the recommendation.	1 paragraph	EIR/GM + AI Studio Agent Lead assessment

5.2 **Apply the one-sentence board test to the section before including it** — Any board member who reads only the performance summary and the business impact should be able to answer: "Is this agent worth continuing to run?" If they cannot — if the section requires reading all four parts plus the monitoring logs to reach a conclusion — the section is too complex. Revise until the one-sentence answer is visible in the first two parts.

5.3 **For agents approaching retirement criteria: present the retirement recommendation alongside the performance section** — Do not wait for a separate agenda item. The quarterly report is the natural vehicle for a retirement recommendation. See Step 6 for the retirement criteria and the recommendation format.

AI PROMPT — Quarterly Venture Board Report Section

```
I am producing the quarterly AI agent performance section for the
Venture Board report. Agent: [name and function]. Quarter: [Q / year].
Monthly review summaries for the quarter: [paste or describe all three
monthly reviews]. Business outcome metrics at end of quarter vs
baseline: [describe]. Any prompt updates during the quarter: [list with
the change, the reason, and the result]. Custom model training
evaluation status: [not yet evaluated / evaluated — conditions not met
/ conditions met and recommendation pending]. Recommendation for next
quarter: [continue / invest to improve / review for retirement]. Write
the four-part quarterly report section in executive language — no
operational jargon. Every metric mentioned must be connected to a
business outcome or a business decision. The section must pass the one-
sentence board test: any board member who reads only Parts 1 and 2 can
answer "is this agent worth continuing to run?"
```

STEP 6 APPLY THE AGENT RETIREMENT AND REPLACEMENT FRAMEWORK

An agent that is retired deliberately is a success. It achieved what it was designed to achieve, the context changed, and the organization recognized the change and acted on it. An agent that is retired belatedly — after months of declining performance, increasing cost, and eroding user trust — represents a failure of the monitoring and governance cycle.

Retirement criteria must be defined at deployment — not at the moment the conversation about retirement begins. If the criteria are defined after problems emerge, the conversation is about whether to retire this specific agent at this specific time, which introduces judgement biases and organizational politics. When the criteria are defined at deployment and evaluated annually against production data, retirement becomes a governance decision, not a conflict.

6.1 Define the retirement criteria at handover — before the first quarterly report

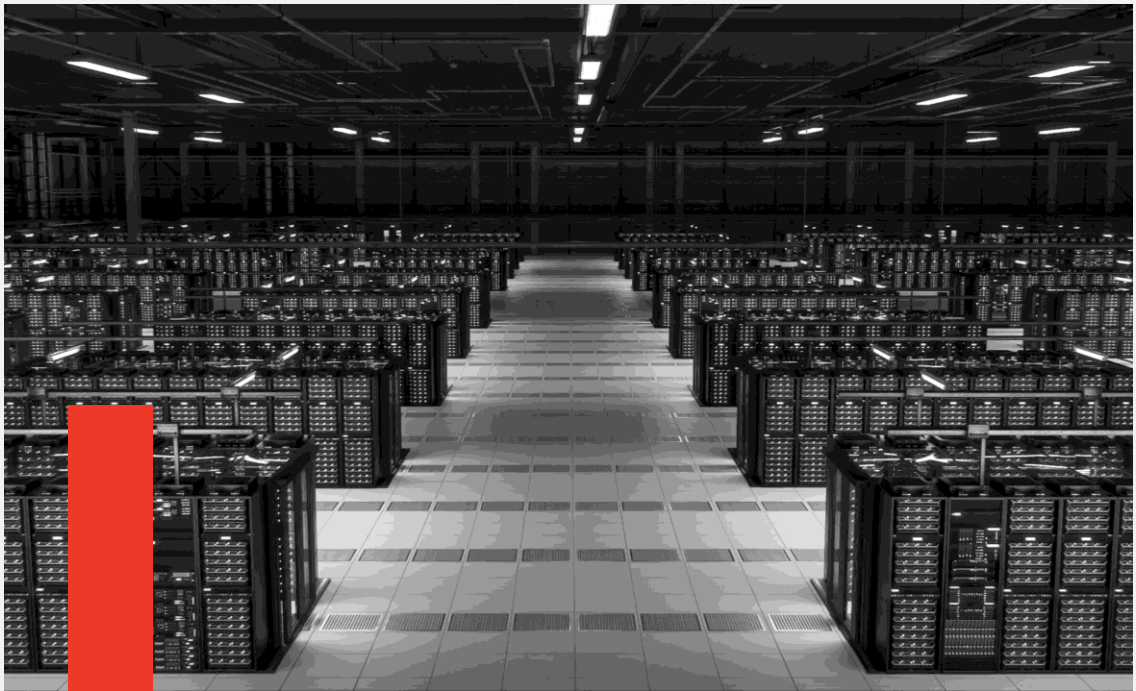
RETIREMENT SIGNAL	DESCRIPTION	EVALUATION FREQUENCY	DECISION OWNER
Sustained Decline	Accuracy rate has been below 80% for 3 consecutive monthly reviews despite targeted prompt iterations. No prompt update has reversed the trend. The off-the-shelf API ceiling has been confirmed.	Monthly review	AI Studio Agent Lead + Operations Owner recommend; EIR/GM decides
Business Outcome Reversal	The business metric the agent was designed to improve has returned to baseline or is now worse than baseline — measured over a minimum of 2 consecutive quarters.	Quarterly Venture Board review	BU Head decides, with Operations Owner and AI Studio Agent Lead input
Strategic Context Change	The workflow the agent automates has been changed, eliminated, or replaced by a different system. The agent's use case no longer exists in its original form.	Triggered by workflow change event	BU Head + EIR/GM jointly decide
Cost Unsustainability	Cost per call has reached a level where the operational cost of running the agent exceeds the measurable business value it generates. The three-condition custom model training gate has been evaluated and training is not justified.	Quarterly Venture Board review	EIR/GM decides with Venture Board awareness
Security or Governance Critical Issue	Persistent or unfixable security vulnerabilities (e.g., repeated successful prompt injection, tool misuse enabling unintended system actions), or a governance requirement change that makes the current deployment model non-compliant and where remediation cost exceeds the business value. Distinct from a single incident — this criterion applies when the issue is systemic and the architectural fix is disproportionate.	Triggered by security or governance event	Legal + EIR/GM + BU Head jointly decide
Governance Constraint	A regulatory, compliance, or governance requirement has changed in a way that makes the agent's current deployment model non-compliant. Remediation cost exceeds the agent's business value.	Triggered by governance event	Legal + EIR/GM + BU Head jointly decide

6.2 **Consider autonomy reduction before retirement** — When sustained issues appear — declining accuracy, rising escalation, or security concerns — the first response is not always retirement. Before the retirement process begins, evaluate whether stepping down the agent's autonomy level is sufficient: increase the HITL threshold so more outputs require human review, restrict the agent to lower-stakes workflow steps, or limit tool-use scope. Autonomy reduction preserves the agent's value in the parts of the workflow it still handles well, while protecting the BU from its failure modes. Retirement is the correct outcome when autonomy reduction cannot restore acceptable performance.

6.3 **Evaluate retirement criteria annually** — or when a Sustained Decline or Security/Governance Critical signal appears — The annual review is a 30-minute session between the Operations Owner, the AI Studio Agent Lead, and the EIR/GM. It reviews each criterion against the past 12 months of monitoring data. If no criterion is approaching: document and continue. If any criterion is approaching: move to the retirement recommendation process.

6.4 **Produce a retirement recommendation when any criterion is met** — The retirement recommendation is a one-page document: which criterion was triggered, the evidence, the recommendation (retire immediately or within a wind-down period), and the wind-down plan. The wind-down plan covers user notification and transition (minimum 4 weeks notice), the manual process or replacement system, and the data and configuration archival requirements. For multi-agent environments: the wind-down plan must also address upstream and downstream agents and any shared tools that depend on this agent. The BU Head signs the retirement recommendation before any decommissioning begins.

6.5 **Distinguish between agent retirement and agent replacement** — Retirement means the workflow is returned to a manual process or discontinued. Replacement means the agent is replaced by a different agent, a fine-tuned model, or an integrated AI system. A replacement is a new deployment — it goes through G1, G2, and G3 from the beginning, with the accumulated production data from the retiring agent as the starting test set.



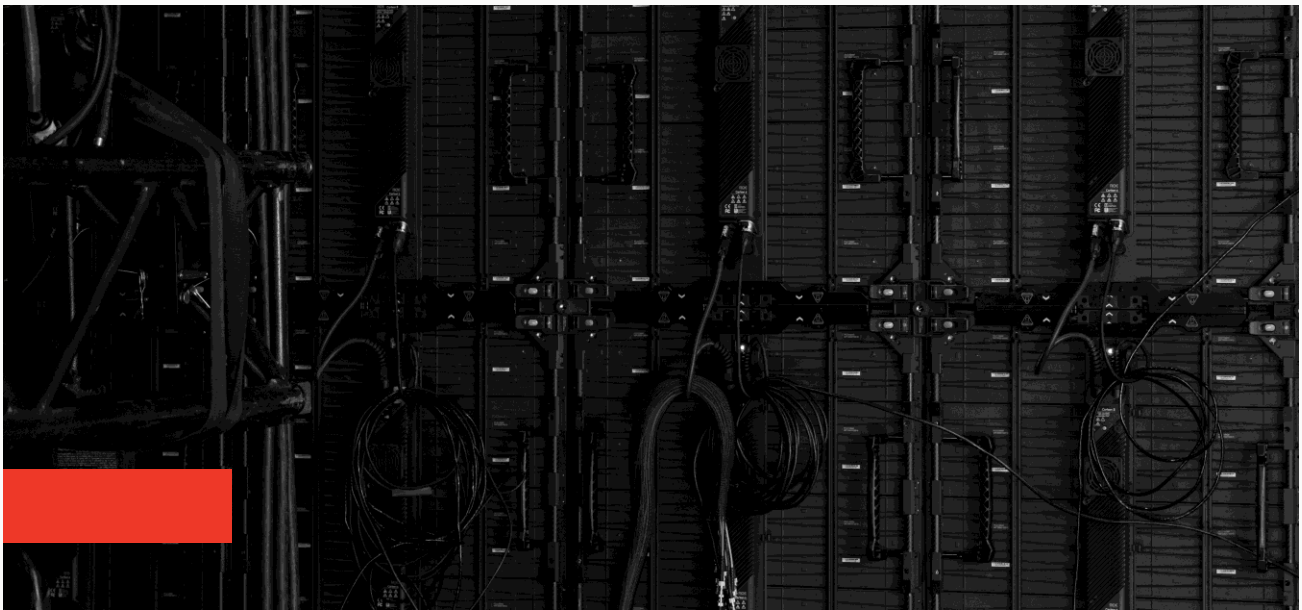
TROUBLESHOOTING

SYMPTOM	LIKELY CAUSE	FIX
Accuracy is declining week-on-week, but no single week triggered a drift alert	Drift is gradual — each week's drop is below the alert threshold individually, but cumulative decline is significant. No trend analysis is running.	Activate the 4-week trend comparison in the weekly review (Step 1). A 5% decline across any 4-week window triggers a review regardless of whether any individual week breached the threshold. Treat cumulative decline as seriously as a single-week spike.
Escalation rate is rising but the human review team is resolving escalations correctly	The escalation rate increase reflects genuine ambiguity in real-world inputs — not a system prompt failure. The agent is correctly identifying uncertainty.	Do not immediately revise the system prompt. First: review whether the escalating inputs represent a new input pattern the original test set did not cover. If yes: add them to the test set and extend the system prompt to handle them. The escalation rate target may need to be recalibrated for this real-world input distribution.
A system prompt update improved accuracy on the target pattern but degraded accuracy on a different input category	The prompt revision was not regression-tested against the full 50-input test set before deployment	Apply the change control protocol from Step 3 strictly: every system prompt change must pass a full regression test before production deployment. A prompt that improves one category while degrading another has not been validated — it has been changed. Version tag the new prompt and roll back if regression is confirmed.
The Venture Board section of the quarterly report cannot be completed because performance data is not in a reportable format	The monitoring dashboard tracks operational metrics but has not been structured to produce the quarterly report inputs — the data exists but is not aligned to the report template	Build the quarterly report template before the first quarterly cycle (Step 5). Every metric in the report must map to a specific dashboard field. If a report metric has no dashboard source, either add it to the dashboard or remove it from the report. Retrospective data extraction is avoidable with one hour of setup.
The custom model training decision is triggered prematurely — the team wants to train before the three conditions are met	Team members observe high escalation rates or specific failure patterns and conclude that a custom model will fix it. The three conditions have not been formally evaluated.	Apply the three-condition gate from Step 4: (1) has production usage validated the use case at scale? (2) has the off-the-shelf API demonstrably hit a confirmed quality or cost ceiling? (3) is the labeled dataset sufficient? If any condition is "no," prompt iteration is the correct intervention, not model training. A custom model is not a substitute for a well-designed system prompt.
The retirement decision is deferred indefinitely because no one wants to make it	No formal retirement criteria were defined at deployment. The agent is underperforming but no one has authority or mechanism to retire it.	Apply the retirement decision framework from Step 6. The Operations Owner and the AI Studio Agent Lead jointly review retirement criteria annually, or when a Sustained Decline signal appears in the monthly review. The BU Head makes the final retirement call — not the AI Studio team. A deferred retirement decision consumes operational cost and erodes user trust.
User feedback volume drops to near-zero after the first month of production	Users logged feedback in the first weeks, but the behavior was not reinforced — no one visibly acted on their feedback, so they stopped providing it.	Close the feedback loop visibly: in the monthly communication from the G2 Operations Owner, explicitly state what changed in the agent as a result of user feedback. Users who see their feedback acted on continue providing it. The feedback mechanism is the labeling pipeline for prompt improvement. Losing it degrades the monitoring and optimization cycle over time.

VALIDATION STEPS

Confirm the monitoring and optimization cycle is operating correctly when the following are true:

- [ ] Weekly monitoring review completed every week — monitoring log entry filed by the Operations Owner
- [ ] Trend analysis running: all seven metrics compared to both the previous week and the 4-week moving average
- [ ] Every drift signal classified (Type 1/2/3/4) before any action is taken
- [ ] Monthly performance review completed every month — three decisions documented: performance assessment, improvement backlog, Level 2 escalation decision
- [ ] Every system prompt change processed through the prompt iteration protocol — version-tagged, regression-tested, deployed through configuration management, post-deployment monitored
- [ ] Three-condition gate evaluated at the monthly review when custom model training is discussed — no training proposed without all three conditions met and formally documented
- [ ] Quarterly Venture Board report section produced for every board cycle — in the four-part format, passing the one-sentence board test
- [ ] Retirement criteria defined and on file — last annual review documented



8 NEXT STEPS

G3 runs continuously as the ongoing operations guide for the agent's entire production life. It connects to the other guides in the following ways:

- When a Type 2 drift signal requires a system prompt update: apply the prompt iteration protocol (Step 3) and record in the G2 Handover Document change log
- When the three-condition gate is met: escalate to the EIR/GM for a custom model training investment decision — these feeds into the H3 Quarterly Board Report as a capital request
- When a retirement criterion is met: produce the retirement recommendation and present at the next Venture Board quarterly review (H3)
- When the agent is to be replaced: the replacement goes through a new G1 deployment, using the accumulated production test set from the retiring agent as the starting regression set

The quarterly Venture Board report section (Step 5) feeds directly into the H3 quarterly board reporting guide. The AI agent performance section is one input to a broader venture performance picture that includes commercial milestones, financial burn rate, and strategic progress. The framing principle is the same across all sections: every operational metric must connect to a business outcome or a business decision before it reaches the board.

MASTER CHECKLIST

A. WEEKLY MONITORING REVIEW

☐ Weekly review completed every week — no skipped weeks

☐ All eight metrics reviewed: accuracy, escalation rate, latency (p95), cost per call, feedback volume, safety flags, tool-call health (tool-using agents), business outcome metrics

☐ Both threshold alert AND 4-week trend analysis applied to every metric

☐ Tool-call traces reviewed for tool-using agents: unexpected tool invocations and action anomalies classified

☐ Every drift signal classified: Type 1 (input shift), Type 2 (prompt gap), Type 3 (external change), Type 4 (user/workflow change)

☐ For regulated or high-risk agents: traceability confirmed — all outputs logged and attributable to a specific agent version

☐ Weekly monitoring log entry completed and filed before the next week's review

B. MONTHLY PERFORMANCE REVIEW

☐ Monthly review held every month — Operations Owner and AI Studio Agent Lead both present

☐ Five inputs prepared before the session: metric trends, user feedback patterns, escalation patterns, business outcome vs baseline, bias/fairness indicators (regulated/high-impact agents)

☐ Three decisions documented: performance assessment, improvement backlog prioritization, Level 2 escalation decision

☐ Agent Health Scorecard updated — 3–4 color-coded indicators for BU head visibility

☐ Improvement backlog updated with owner and target date for each item

☐ Any custom model training discussion accompanied by a formal three-condition evaluation

C. PROMPT ITERATION PROTOCOL

☐ Drift signal classified as Type 2 (prompt gap) before any prompt update is initiated

☐ Protocol scope confirmed: applies to all prompt units — system prompt, sub-agent prompts, tool descriptions, RAG instructions, routing logic

☐ Proposed change written and tested in staging — not drafted directly in the production configuration

☐ Full regression test run: all test categories including accuracy and safety/policy tests

☐ Safety regression check passed: adverse input patterns, disallowed content scenarios, and prompt injection attempts all re-evaluated

☐ Updated prompt matches or exceeds production prompt across all test categories and all safety tests

☐ Version tag assigned before deployment — change log entry completed with regression and safety test results

☐ Change recorded in G2 Handover Document (Section 8)

☐ Post-deployment monitoring window of 48 hours applied across all eight metrics

☐ No informal prompt edits in production — every change through the protocol

D. CUSTOM MODEL TRAINING DECISION

☐ Three-condition evaluation completed in writing: condition 1 (use case validated at production scale), condition 2 (API ceiling confirmed after exhausting alternatives), condition 3 (sufficient, representative, unbiased labeled data)

☐ Alternative interventions evaluated and documented before condition 2 is assessed: retrieval/RAG, tool redesign, model routing

☐ Labeled dataset representativeness reviewed: no significant skew toward specific user segments or input types; any known skew documented

☐ Evaluation signed by Operations Owner and AI Studio Agent Lead

☐ If all three conditions met: capital and resource request submitted to EIR/GM; model card produced for the custom model before deployment

☐ If any condition not met: training decision deferred, reason documented, re-evaluation threshold defined

E. QUARTERLY VENTURE BOARD REPORT

☐ Report section produced for every quarterly board cycle

☐ Four-part structure: performance summary, business impact, known issues and mitigations (including safety/compliance incidents), recommendation and forward outlook

☐ Every metric in the report connected to a business outcome — no standalone operational metrics

☐ Safety and compliance incidents from the quarter explicitly disclosed in Part 3 with remediation status

☐ One-sentence board test passed: board member reading only Parts 1 and 2 can answer "is this agent worth continuing to run?"

☐ For ventures running multiple agents: portfolio view included — each agent's recommendation and primary business impact metric

☐ Retirement recommendation included if any retirement criterion has been met or is being approached

F. AGENT RETIREMENT AND REPLACEMENT

☐ Retirement criteria defined and on file from the handover date — including Security/Governance Critical signal

☐ Annual retirement criteria review completed — findings documented

☐ Autonomy reduction evaluated before retirement is proposed for sustained performance issues

☐ Triggered by Sustained Decline or Security/Governance Critical signal: reviewed at the next monthly review

☐ Retirement recommendation produced when any criterion is met: criterion, evidence, recommendation, wind-down plan

☐ Wind-down plan addresses cross-agent dependencies for multi-agent environments

☐ BU Head signature obtained before any decommissioning begins

☐ Replacement distinguished from retirement: replacement initiates a new G1/G2/G3 cycle with accumulated production data as the starting test set