APPENDIX B

Unified Product and AI Quality Monitoring

Connecting Engagement Metrics, Failure Mode Analysis, and the Product Improvement Loop

Supplement to: AI Evaluation in the SDLC
Joint Brief from the Director of Engineering and the Director of Product

The Two Dashboards Problem

Today, most organizations that ship AI features operate with two completely disconnected monitoring systems. Product managers watch **engagement dashboards**: daily and monthly active users, feature adoption rates, session duration, task completion rates, Net Promoter Scores, and churn. Engineers and AI evaluation teams watch **quality dashboards**: trace failure rates, failure mode distributions, model latency, hallucination frequency, guardrail trigger rates, and automated evaluator pass rates.

These two systems describe the same product and the same users, but they rarely appear on the same screen or in the same conversation. The result is a dangerous blind spot: **you can have excellent AI quality metrics and still lose users, or you can have strong engagement metrics while accumulating AI quality debt that will eventually surface as a crisis.**

A product team watching only engagement metrics might see churn increasing and attribute it to competitive pressure, pricing, or feature gaps—when the actual cause is that the AI feature produces subtly wrong outputs often enough that users stop trusting it. An AI evaluation team watching only trace quality metrics might celebrate a hallucination rate below 2% without realizing that the 2% of hallucinations are concentrated in the highest-value use case, driving away the users who matter most.

> *The complete picture of whether an AI feature is delivering on its job to be done requires both lenses—engagement and quality—viewed together, correlated in time, and interpreted as a single story about whether users are being served.*

Two Measurement Traditions, One Product

Before describing how to unify these views, it is worth being explicit about what each tradition measures, who owns it, and what decisions it informs.

Product Engagement Metrics

These are the metrics product managers have used for decades to understand whether a product is delivering value and retaining users. They are *outcome metrics*—they tell you *what users did*, not why.

Metric	What It Measures	Signal Type	Decision It Informs
Daily / Monthly Active Users (DAU/MAU)	How many users interact with the feature in a given period	Adoption and reach	Is the feature being discovered and used? Is awareness a problem?
Feature Adoption Rate	Percentage of eligible users who use a specific AI feature	Adoption depth	Is the feature compelling enough to try? Is onboarding effective?
Task Completion Rate	Percentage of users who complete the intended workflow	Usability and value delivery	Does the feature actually help users accomplish their goal?
Session Duration / Depth	How long users spend and how deeply they engage per session	Engagement intensity	Are users getting value, or are they struggling?
Retention Rate (Day 7, 30, 90)	Percentage of users who return after first use	Stickiness and ongoing value	Does the feature deliver enough value to warrant repeat use?
Churn Rate	Percentage of users who stop using the feature over a period	Disengagement and dissatisfaction	Are we losing users? At what point in the lifecycle?
Net Promoter Score (NPS)	Likelihood that a user would recommend the feature	Satisfaction and advocacy	Are users satisfied enough to promote the product?
Customer Satisfaction (CSAT)	Direct satisfaction rating after interaction	Immediate experience quality	How did the user feel about this specific interaction?

AI Quality Metrics (from Trace Analysis)

These are the metrics that emerge from the trace capture and error analysis methodology described in Appendix A. They are *diagnostic metrics*—they tell you *why the AI behaved the way it did* and whether its behavior met quality standards.

Metric	What It Measures	Signal Type	Decision It Informs
Failure Mode Rate (by category)	Percentage of traces exhibiting each identified failure mode	AI output quality	Which failure modes need targeted fixes? Where to invest improvement effort?
Overall Pass Rate	Percentage of traces that pass automated evaluators and human review	Aggregate AI quality	Is the feature meeting quality standards? Is it production-ready?
Hallucination Rate	Percentage of outputs containing fabricated information	Factual reliability	Can users trust the AI's factual claims?
Guardrail Trigger Rate	How often safety or scope guardrails activate	Boundary effectiveness	Is the AI attempting out-of-scope actions? Are guardrails calibrated correctly?
User-Flagged Error Rate	Percentage of interactions where users report a problem	User-perceived quality gap	What is the gap between AI confidence and user satisfaction?
Model Latency (P50, P95, P99)	Response time distribution	Performance experience	Is the AI fast enough for the use case? Are there latency spikes?
Drift Score	Statistical measure of output distribution change over time	Model stability	Has the AI's behavior shifted? Do we need to re-evaluate?
Automated Evaluator Agreement	Agreement rate between automated judges and human reviewers	Evaluator reliability	Can we trust our automated quality checks?

> *Product engagement metrics tell you the user is leaving. AI quality metrics tell you why. You need both to take the right action.*

The Unified Monitoring View

The central argument of this appendix is that product engagement metrics and AI quality metrics must be **displayed side by side, correlated in time, and analyzed together.** This is not a theoretical aspiration—it is a specific design requirement for how we build monitoring dashboards and conduct product reviews for AI features.

Design Principles for the Unified View

1. Time-Aligned Panels

Engagement metrics and quality metrics should be displayed on the same time axis. When you see a spike in churn on Tuesday, you should be able to look directly below it and see whether failure mode rates changed on Tuesday. When you see a latency increase on Wednesday, you should be able to look above it and see whether task completion rate dropped on Wednesday. The visual correlation is what makes the causal investigation possible. Without time alignment, the two data sources remain parallel monologues instead of a dialogue.

2. Layered Drilldown

The unified view should support three levels of depth. The first is the **executive view**: a single screen showing aggregate engagement health (DAU trend, retention, churn) alongside aggregate quality health (overall pass rate, top failure mode rates, latency). This is the view for weekly product reviews and business stakeholder updates. The second is the **product view**: engagement metrics segmented by user cohort, feature, and channel alongside quality metrics segmented by the same dimensions. This is the view for product managers investigating a specific trend. The third is the **trace view**: the custom trace review interface described in Appendix A, where individual interactions are examined. This is where diagnostic work happens.

3. Correlation Alerts, Not Just Threshold Alerts

Traditional monitoring fires alerts when a single metric crosses a threshold: churn exceeds 5%, or hallucination rate exceeds 3%. A unified system should also fire **correlation alerts** when engagement metrics and quality metrics move together in ways that suggest a causal relationship. For example: retention drops 15% for a user cohort at the same time that failure mode rate for “scope violation” increases 8% for the same cohort. Or: CSAT drops 20 points in a segment while latency P95 doubles. These correlated movements are often invisible when the two dashboards live in different tools seen by different teams.

4. Shared Segmentation

The most important design choice: segment both metric sets by the same dimensions. If you can see churn by user cohort (new users, power users, enterprise users), you should be able to see failure mode rates by the same cohorts. If you can see adoption by feature, you should be able to see quality by feature. Shared segmentation is what allows the question “why are enterprise users churning?” to be answered with “enterprise users encounter the key-term-omission failure mode at 3x the rate of other cohorts because they use longer, more complex documents.”

The Diagnostic Framework: Reading the Signals Together

When engagement metrics and quality metrics are side by side, four diagnostic patterns emerge. Each pattern tells a different story about the product and demands a different response.

	Engagement: Healthy	Engagement: Declining
AI Quality: Healthy	**QUADRANT A: THRIVING** The feature is delivering on its job to be done. Users are engaged and the AI is performing well. **Action:** Maintain. Expand to new use cases. Invest in the next improvement cycle. Document current state as the quality baseline for future TSRs.	**QUADRANT B: PRODUCT PROBLEM** The AI works well, but users are leaving. The problem is not quality—it is product-market fit, discoverability, onboarding, workflow integration, or the feature is solving the wrong job. **Action:** Product investigation. User research. Jobs-to-be-done interviews. The fix is product design, not AI improvement.
AI Quality: Degraded	**QUADRANT C: HIDDEN DEBT** Users are engaged, but AI quality is degrading. This is the most dangerous quadrant. Users have not noticed yet—or are tolerating issues because the feature is otherwise valuable. Quality debt is accumulating and will eventually surface as a crisis. **Action:** Urgent error analysis cycle. Fix failure modes before they become visible to users. This is where the TSR-driven improvement cycle pays its highest dividends.	**QUADRANT D: ACTIVE CRISIS** Users are leaving and AI quality is degraded. The quality problems are likely driving the engagement decline. This is the scenario that generates compliance and reputational risk. **Action:** Triage. Identify the failure modes correlated with churn. Consider rollback or feature gating while the error analysis → fix → automate cycle runs. TSR documents the incident and remediation.

> *The diagnostic matrix is not a one-time assessment. It is the framework for every product review of an AI feature. Each review should begin by locating the feature in one of these four quadrants based on the most recent data, and the action plan should follow from that quadrant's guidance.*

Reading the Dashboard: Three Scenarios

To make this concrete, here are three scenarios that illustrate how the unified view changes the product team's ability to diagnose and respond to signals.

Scenario 1: Churn Spike with No Quality Change

WHAT THE DASHBOARD SHOWS

Week-over-week churn increases from 4% to 9% for first-time users of the document summarization feature. At the same time, all AI quality metrics remain stable: pass rate steady at 94%, hallucination rate flat at 1.8%, latency unchanged, no new failure modes detected. The feature is in Quadrant B.

WHAT IT MEANS

The AI is working correctly, but new users are leaving anyway. The problem is upstream of AI quality. Possible causes: the onboarding experience does not make the feature's value clear; users expect the feature to do something it does not do (a job-to-be-done mismatch); users find the feature but cannot integrate it into their existing workflow; the feature is discoverable but the entry point is confusing.

WHAT THE TEAM DOES

This is a product problem, not an AI problem, and the response should be product-led. The PM conducts user interviews with churned first-time users. Session recordings are reviewed to identify where users abandon the workflow. The engineering team checks whether onboarding is surfacing the right examples. The TSR for any resulting changes is lightweight (Tier 1)—because the changes are to UI and onboarding, not to the AI behavior.

Scenario 2: Quality Degradation with Stable Engagement

WHAT THE DASHBOARD SHOWS

Hallucination rate for the AI research assistant increases from 2% to 7% over three weeks. Guardrail trigger rate doubles. But DAU is unchanged, task completion rate is flat, and churn has not moved. The feature is in Quadrant C.

WHAT IT MEANS

The AI is getting worse, but users have not reacted yet. This is a grace period, not evidence that the quality degradation does not matter. Possible explanations: users have not encountered the hallucinations in their specific use cases yet; users are encountering them but assuming they are isolated incidents; the hallucinations are in low-consequence parts of the output that users are not relying on. Regardless of the explanation, this is accumulating risk. When users do notice, the engagement decline will be sudden rather than gradual.

WHAT THE TEAM DOES

Urgent error analysis cycle. The team pulls traces where hallucinations were detected and reviews them in the custom trace review interface to identify the root cause. Common causes for sudden quality degradation: an upstream model version change by the provider; a data distribution shift in user inputs; a prompt that worked well for common cases but breaks on a growing edge case category. The TSR for this remediation is Tier 2 at minimum, because the quality degradation, if unaddressed, will reach users.

Scenario 3: Correlated Engagement and Quality Decline

WHAT THE DASHBOARD SHOWS

For enterprise users, retention at Day 30 drops from 78% to 61%. At the same time, the “key term omission” failure mode rate for this cohort increases from 5% to 18%. The correlation is tight: the engagement decline began the same week the quality metric shifted. The feature is in Quadrant D for this segment.

WHAT IT MEANS

Enterprise users work with longer, more complex documents. The AI’s key term extraction is failing on these documents, and enterprise users—who rely on the summaries for real decisions—are losing trust. The segmentation is what makes this visible: the overall pass rate might still look healthy, but the enterprise segment has a severe quality problem that is directly driving churn.

WHAT THE TEAM DOES

Triage. The team considers whether to gate the feature for document types that exceed a complexity threshold until the fix is in place. Error analysis is focused specifically on enterprise-length documents. The retrieval pipeline is likely the root cause: chunk sizes and overlap settings that work for standard documents are insufficient for 50-page contracts. The TSR documents the incident, the root cause, the segmented failure analysis, the fix, and the monitoring plan that specifically tracks enterprise user quality and retention going forward.

Embedding the Unified View in the TSR

The main brief's TSR structure includes seven sections. This appendix proposes that engagement metrics be woven into three of those sections to create a TSR that tells the complete product story—not just the AI quality story.

TSR Section	Current Content (from Main Brief)	Enhanced with Engagement Context
1. Change Summary	What changed, why, and what job-to-be-done it serves.	Add: What engagement signal or quality signal triggered this change? Example: "Enterprise user retention declined 17 points over 3 weeks, correlated with a 13-point increase in key term omission rate for documents exceeding 20 pages."
3. Error Analysis Results	Failure modes discovered, categorized by type and severity.	Add: Engagement impact assessment for each failure mode. Example: "Key term omission: affects 18% of enterprise user traces. Enterprise cohort shows 3.2x higher churn rate than non-enterprise cohort. Estimated revenue impact: $X per quarter if unaddressed."
6. Production Monitoring Plan	What will be monitored post-deployment, thresholds, rollback criteria.	Add: Engagement metrics to monitor alongside quality metrics. Example: "Monitor enterprise Day-30 retention alongside key term omission rate. Success criterion: retention recovers to above 72% within 30 days of deployment. If retention does not improve despite quality improvement, escalate to product investigation."

The enhanced TSR does not replace the existing sections. It adds a thin layer of product context that transforms the TSR from a purely technical artifact into a business artifact. When a business stakeholder reads the Change Summary and sees the engagement signal that triggered the work, they understand immediately why the investment was made. When they see the engagement impact assessment in the Error Analysis, they can evaluate whether the failure mode justifies the remediation effort. When they see engagement metrics in the Monitoring Plan, they have confidence that the team will know whether the fix actually solved the business problem—not just the technical one.

The Product Review Cadence for AI Features

With the unified view in place, the product review cadence for AI features should follow a rhythm that keeps both engagement and quality visible to the right audiences at the right frequency.

Cadence	Audience	What Is Reviewed	Decisions Made
Daily	Engineering + AI evaluation team	Quality metrics: failure mode rates, evaluator pass rates, latency, guardrail triggers. Automated alerts for threshold breaches.	Triage incoming quality issues. Route to error analysis if patterns emerge.
Weekly	Product + Engineering	Unified dashboard: engagement trends alongside quality trends. Segmented by cohort and feature. Review any correlation alerts fired.	Locate feature in diagnostic quadrant. Decide whether to initiate error analysis, product investigation, or both. Prioritize improvement backlog.
Bi-weekly / Sprint	Product + Engineering + QA	TSR review for any changes in progress or recently deployed. Error analysis results from the trace review interface. Automated test suite status.	Go/no-go decisions on pending deployments. Review monitoring data for recently shipped changes. Decide if observation window can close.
Monthly	Product + Engineering + Business Stakeholders	Executive view: aggregate engagement health, aggregate quality health, trend lines, and notable incidents. TSRs completed in the period. Quadrant assessment for each AI feature.	Strategic product decisions: feature investment, expansion, pivot, or deprecation. Compliance and risk posture assessment. Resource allocation.
Quarterly	Product + Engineering + Compliance + Leadership	Portfolio view: all AI features assessed across both dimensions. Trend analysis across the quarter. Cumulative TSR record. Regulatory and audit readiness.	Portfolio-level investment decisions. Governance framework adjustments. Evaluation process improvements. Risk appetite calibration.

Note that engagement metrics and quality metrics appear at every level of this cadence. The difference is granularity: the daily review is operational and quality-focused; the weekly review is tactical and correlation-focused; the monthly review is strategic and outcome-focused. This ensures that no signal is orphaned in a silo that the wrong team is watching.

Closing the Loop: Metrics, Jobs to Be Done, and the Improvement Cycle

The Jobs to Be Done framework is the conceptual thread that ties engagement metrics and quality metrics together. A user hires an AI feature to accomplish a specific job. The engagement metrics tell you whether the user keeps hiring the feature (retention) or fires it (churn). The quality metrics tell you whether the feature is performing the job well (pass rate, low failure modes) or poorly (high failure modes, hallucinations, drift).

The unified monitoring view makes it possible to ask—and answer—the question that matters most: ***Is this AI feature worthy of being hired?***

A feature worthy of being hired is one where both engagement and quality are healthy (Quadrant A). It is a feature that users return to, that the automated evaluators validate, that the error analysis shows is stable, and that the TSR documents as fit for production. It is a feature that the business can point to and say: we deployed this with evidence, we monitor it with rigor, we improve it with discipline, and we can demonstrate all of that to anyone who asks—an auditor, a regulator, a customer, or a board member.

And when a feature is not in Quadrant A, the unified view tells you which direction to move: fix the product (Quadrant B), fix the AI quality before users notice (Quadrant C), or triage an active crisis (Quadrant D). The TSR documents each improvement cycle. The engagement metrics confirm whether the improvement worked. The quality metrics confirm whether it was the right fix. And the cycle continues.

THE COMPLETE PICTURE

Product managers bring the engagement lens: are users being served? Engineering brings the quality lens: is the AI trustworthy? The unified monitoring view brings these together on the same screen, correlated in time, segmented by the same dimensions. The TSR captures the synthesis. The improvement cycle responds to the signals. And the business gets what it needs: AI features that are not just technically sound, but genuinely useful—products that earn the right to be hired, keep earning it over time, and can prove it to anyone who asks.

Engagement tells you the verdict. Quality tells you the evidence. You need both.