AI COGNIFIT

Judgment Metrics Playbook

A Framework for Measuring Human Judgment in AI-Augmented Workflows

→ Anchor AI performance to human judgment and risk

→ Detect when AI silently degrades quality or increases load

→ Create shared vocabulary for cross-functional governance

→ Practical metrics for engineering, legal, HR, and finance

Version 1.0 • aicognifit.com

How to Use This Playbook

This playbook is designed to be both a reference guide and an implementation manual. You can read it cover-to-cover for a comprehensive understanding, or jump directly to the sections most relevant to your role.

If you're new to judgment metrics: Start with Section 1 to understand why traditional productivity metrics fail, then read Section 2 to grasp the five-pillar framework. This foundation will make the specific metrics in Section 3 more meaningful.

If you're ready to implement: Jump to Section 4 to find the metric bundle for your function, then use Section 5's roadmap to plan your rollout. Return to Section 3 for detailed metric specifications as you instrument each one.

If you're building organizational buy-in: Use Section 1 to frame the problem for leadership, Section 4 to show role-specific applications, and the FAQ in Section 6 to address common objections.

> QUICK START
>
> For the fastest path to value: Pick one high-stakes workflow, instrument TTPR + DER + AOC, baseline for one cycle, then introduce AI changes and compare. You can do this in under a month with existing tools.

Why traditional productivity metrics fail in AI-augmented work

The Productivity Paradox

AI has made it cheap to generate code, content, and decisions at scale. The bottleneck is no longer producing artifacts—it's **evaluating** them: deciding what is correct, safe, compliant, and strategically aligned.

Traditional productivity dashboards focus on volume metrics: tokens generated, prompts sent, lines of code written, tickets closed. In AI-heavy workflows, these numbers go up automatically—even when quality, risk, and human judgment are deteriorating.

Consider a software team that adopts AI coding assistants. Their "lines of code per developer" metric doubles. Leadership celebrates the productivity gains. But six months later, production incidents are up 40%, code reviews take twice as long, and senior engineers are burning out trying to catch subtle bugs in AI-generated code. The dashboard showed success while reality deteriorated.

> **KEY INSIGHT**
>
> AI commoditizes generation. Human judgment becomes the scarce, critical resource. Your metrics should reflect that shift.

What Goes Wrong

Organizations that measure only velocity encounter predictable failure modes. These aren't edge cases—they're systematic consequences of measuring the wrong things:

- **Rubber-stamping:** Reviewers are measured on throughput, not catch rate. They learn to approve quickly rather than scrutinize carefully. AI-generated content gets the same cursory glance as human work, despite potentially hiding different kinds of errors.
- **Escaped defects:** Review quality collapses under volume. Issues that would have been caught in lower-volume workflows slip through because reviewers are cognitively overloaded.
- **Hidden burnout:** Experts burn out from the mental load of reviewing AI output while dashboards show "productivity gains." By the time attrition spikes, the damage is done.
- **Compliance drift:** Humans bypass guardrails to meet speed targets. "I'll just approve this without the full review—we're behind on our metrics."
- **Technical debt:** "Lines of code" went up but understanding went down. The team ships faster but can't maintain what they shipped.

The Solution

Judgment metrics re-center the scoreboard on what actually matters: whether your teams can reliably oversee AI, catch issues before they escape, and make sound decisions under pressure. Instead of measuring "how much did we produce?" they measure "how well did we evaluate what we produced?"

This isn't about slowing down or rejecting AI. It's about adopting AI with visibility. You want to know when AI is genuinely helping versus when it's creating hidden risks. You want early warning when review quality degrades, before defects escape. You want evidence that your governance actually works.

Three Core Purposes

1. **Anchor to Risk** — Connect AI performance to human judgment and business risk, not vanity metrics. When leadership asks "how is AI working for us?", you have substantive answers about quality, risk, and sustainability—not just volume.
2. **Detect Silent Drift** — Spot when AI is quietly increasing error rates or compliance gaps before incidents occur. Many AI problems are gradual: review quality slowly degrades, escape rates creep up, burnout accumulates. Judgment metrics catch these trends early.
3. **Create Shared Language** — Give engineering, legal, HR, and leadership a common vocabulary to compare AI experiments. When everyone uses the same metrics, you can have substantive conversations about trade-offs across functions.

The Business Case

Judgment metrics aren't just risk management—they're competitive advantage. Organizations that measure judgment can:

- **Adopt AI faster:** With visibility into risks, you can move quickly without flying blind. Guardrails let you experiment boldly.
- **Retain experts:** When you track cognitive load and prevent burnout, you keep your best people. Attrition is expensive.
- **Satisfy auditors:** For regulated industries, judgment metrics provide the audit trail regulators want to see.
- **Build trust:** Teams that can demonstrate their AI governance earn trust from customers, partners, and leadership.

> TIP
>
> The goal isn't to slow AI adoption—it's to adopt AI sustainably, with visibility into when you're trading judgment for speed. Organizations that measure judgment can move faster because they know where the risks are.

A comprehensive model for measuring judgment in AI-augmented work

The five pillars provide a comprehensive framework for measuring how well your organization handles AI-assisted work. Each pillar addresses a different aspect of human-AI collaboration, from raw processing capacity to governance compliance.

Think of the pillars as answering five essential questions about your AI workflows:

1. Can we process the volume of work AI creates? (Throughput)
2. Are our decisions correct and well-calibrated? (Quality)
3. Are defects caught before they cause damage? (Risk)
4. Is the cognitive load sustainable for our people? (Burden)
5. Do people actually follow our oversight processes? (Governance)

A healthy AI-augmented organization should have acceptable metrics across all five pillars. Excelling in one while failing another creates hidden vulnerabilities. The framework helps you maintain balance as you scale AI adoption.

1 Evaluative Throughput

How much can your experts review without quality collapse?

Measures the volume of review and decision work your experts can process per unit time while maintaining acceptable quality.

WHY IT MATTERS

Rising AI output without review capacity leads to rubber-stamping and missed defects.

KEY METRICS

TTPR • RIC

2 Judgment Quality

Are decisions correct, robust, and aligned with policy?

Assesses how often human decisions are correct, context-appropriate, and aligned with organizational standards.

WHY IT MATTERS

Overconfident reviewers pass flawed work. Early detection prevents downstream failures.

KEY METRICS

OCPR • JCI

3 Risk & Defect Dynamics

Where do errors surface, and how severe are escapes?

Tracks where errors appear across the lifecycle and how severe they are when they reach production.

WHY IT MATTERS

Rising escape rates signal review isn't keeping pace—early warning before incidents.

KEY METRICS

DER • CIR

REAL-WORLD EXAMPLE

A consulting firm found that Pillar 1 (throughput) looked great—analysts were reviewing 3x more AI-drafted reports. But Pillar 3 (risk) showed DER climbing from 4% to 12%. The speed gains were illusory; they were shipping more defects to clients.

4 Evaluative Burden

How demanding is oversight, and is it sustainable?

Measures the cognitive load required for humans to provide reliable oversight of AI-assisted work.

WHY IT MATTERS

Burnout is a lagging indicator. Tracking burden catches problems before quality collapses.

KEY METRICS

eTLX • RL

5 Governance & Behavior

Do people actually follow the human-in-the-loop process?

Monitors whether people follow intended guardrails and approval gates when AI is involved.

WHY IT MATTERS

Shadow AI usage and undocumented overrides create audit and compliance risk.

KEY METRICS

AOC • BOR

Pillar Interactions

The five pillars don't operate in isolation—they interact in predictable ways. Understanding these interactions helps you interpret metric movements and avoid unintended consequences.

Throughput vs. Burden: There's a natural tension between processing more work and the cognitive load it creates. Sustainable operations balance both. If your team hits throughput targets but evaluative burden spikes, you're on a path toward quality degradation and burnout. The goal is to find the sustainable frontier: the highest throughput your team can maintain without burden becoming unsustainable.

Quality vs. Risk: High judgment quality should translate into lower defect escape rates. If your JCI and OCPR look healthy but DER is rising, investigate the gap—perhaps reviewers are well-calibrated but review criteria don't match real-world failure modes. The metrics should tell a coherent story.

Governance as the Wrapper: Pillars 1-4 measure how well judgment happens. Pillar 5 measures whether it happens at all. Strong governance ensures the other metrics are meaningful—if 30% of work bypasses review entirely, even perfect quality metrics for reviewed work don't protect you.

Reading the Signals

When metrics move, they tell a story. Here are common patterns and what they indicate:

Pattern	What It Signals	Action
TTPR ↓, DER ↑	Speed at cost of quality	Slow down, add review steps
eTLX ↑, OCPR ↓	Overwhelmed reviewers	Reduce volume or add staff
AOC ↓, BOR ↑	Process breakdown	Investigate bypasses, reinforce training
All metrics stable	Sustainable equilibrium	Consider expanding AI usage

DESIGN PRINCIPLE

Never optimize a single pillar in isolation. Use balanced scorecards that track multiple pillars simultaneously. Single-metric targets invite gaming.

10 essential metrics with formulas and targets

This section provides detailed specifications for each core metric. You don't need to implement all ten at once—start with the most relevant to your context and expand over time. Each metric includes a definition, calculation formula, interpretation guidance, and recommended targets.

The metrics are grouped by their primary pillar, though many inform multiple pillars. For example, Defect Escape Rate (a Risk metric) also reflects Judgment Quality—if reviewers miss issues consistently, DER rises. Use these interdependencies to triangulate signals and validate findings.

Quick Reference: Metric Summary

Metric	Pillar	Measures	Good Direction
DER	Risk	Post-release defects	Lower ↓
CIR	Risk	Critical incidents	Lower ↓
TTPR	Throughput	Review cycle time	Lower ↓
RIC	Throughput	Revision rounds	Lower ↓
OCPR	Quality	Issue catch rate	Higher ↑
JCI	Quality	Calibration accuracy	Higher ↑
eTLX	Burden	Cognitive load	Stable/Lower ↓
RL	Burden	Queue wait time	Lower ↓
AOC	Governance	Review coverage	Higher ↑
BOR	Governance	Bypass frequency	Lower ↓

Defect Escape Rate DER

RISK & DEFECT

DEFINITION

Percentage of defects found only after approval or deployment.

FORMULA

`post_approval / total × 100`

INTERPRETATION

Lower is better. Rising DER = review not keeping pace.

TARGET

<5% for high-stakes workflows

Critical Incident Rate CIR

RISK & DEFECT

DEFINITION

Frequency of high-severity failures per unit of output.

FORMULA

`incidents / artifacts × 1000`

INTERPRETATION

Should stay flat or decrease. Any increase is a red flag.

TARGET

No increase tolerated

Time-to-Passed-Review TTPR

THROUGHPUT

DEFINITION

Elapsed time from first submission to final approval.

FORMULA

`time_approved - time_submitted`

INTERPRETATION

Shorter is better, but watch for quality trade-offs.

TARGET

20-30% reduction with AI

Review Iteration Count RIC

THROUGHPUT

DEFINITION

Number of review-revision cycles before acceptance.

FORMULA

`count(review_cycles)`

INTERPRETATION	TARGET
Fewer = cleaner first drafts. AI should reduce RIC.	**1-2 routine; 2-3 complex**

Oversight Challenge Pass Rate OCPR

JUDGMENT

DEFINITION

Percentage of seeded issues caught by reviewers.

FORMULA

`caught / seeded × 100`

INTERPRETATION

Below 70% indicates rubber-stamping.

TARGET

>80% critical; >90% regulated

Judgment Calibration Index JCI

JUDGMENT

DEFINITION

Correlation between confidence and actual correctness.

FORMULA

`calibration(confidence, outcomes)`

INTERPRETATION

Near 1.0 = well-calibrated.

TARGET

0.8 - 1.0

Evaluative NASA-TLX eTLX

BURDEN

DEFINITION

Self-reported mental workload for oversight tasks.

FORMULA

`avg(mental, temporal, effort)`

INTERPRETATION

Rising eTLX predicts burnout.

TARGET

Flag >15% increase

Review Latency RL

BURDEN

DEFINITION

Time artifacts wait before review begins.

FORMULA

`review_start - artifact_ready`

INTERPRETATION

High RL = bottleneck or overload.

TARGET

<24h std; <4h urgent

AI Oversight Coverage AOC

GOVERNANCE

DEFINITION

Percentage of AI artifacts passing human review.

FORMULA

`reviewed / total × 100`

INTERPRETATION

Should be near 100%. Gaps = policy failure.

TARGET

>95% regulated; >90% std

Bypass/Override Rate BOR

GOVERNANCE

DEFINITION

Percentage of undocumented safeguard bypasses.

FORMULA

`bypasses / decisions × 100`

INTERPRETATION

High unexplained BOR = governance failure.

TARGET

<2% critical; <5% std

Choosing Your Starting Metrics

You don't need all ten metrics from day one. The right starting set depends on your context, available data, and primary concerns. Here's how to choose:

If risk is your main concern (regulated industries, high-stakes decisions): Start with DER, CIR, and AOC. These tell you whether defects escape, incidents occur, and whether governance actually happens. Add BOR once you have governance instrumented.

If efficiency is your main concern (throughput pressure, competitive timelines): Start with TTPR, RIC, and RL. These show whether AI accelerates your workflows and where bottlenecks emerge. Monitor DER as a check that speed isn't degrading quality.

If burnout is your main concern (overloaded teams, high turnover): Start with eTLX, RL, and RIC. These reveal cognitive load patterns and where work piles up. Track TTPR to understand throughput pressures driving the burden.

Starter Bundles by Priority

Priority	Core Metrics	Add Later
Risk-First	DER, CIR, AOC	BOR, OCPR
Efficiency-First	TTPR, RIC, RL	DER, eTLX
Sustainability-First	eTLX, RL, RIC	TTPR, JCI
Balanced (Recommended)	TTPR, DER, AOC	eTLX, OCPR, JCI

RECOMMENDATION

For most organizations, the minimum viable set is: TTPR (throughput), DER (risk), and AOC (governance). These three metrics provide a basic view across the framework. Add metrics as you mature.

Tailored metric bundles for different teams

Different teams experience AI differently. A software engineer using Copilot faces different risks than a compliance analyst reviewing AI-drafted policies. The metrics matter across all contexts, but emphasis and thresholds vary.

This section provides recommended metric bundles organized by function. Each bundle specifies which metrics to prioritize, how to apply them in context, and suggested alert thresholds. Use these as starting templates and adjust based on your specific workflows.

Software Engineering & Data Science

AI coding tools (Copilot, Cursor, Claude) can dramatically accelerate development—but code is unforgiving. A subtle bug in AI-generated code can cause production incidents, security vulnerabilities, or data corruption. Engineering teams need metrics that verify AI-assisted code doesn't compromise quality.

Metric	Application	Alert Threshold
TTPR	Per pull request / merge request	>50% increase
DER	AI-assisted vs non-AI code	>5% or rising trend
CIR	Production incidents by origin	Any increase
OCPR	Seeded bugs in code review	<75%
eTLX	Reviewer workload surveys	>20% above baseline

CASE STUDY: ENGINEERING TEAM

A fintech engineering team tracked DER for AI-assisted PRs versus manual PRs. Initial data showed AI PRs had 2x the defect rate. Investigation revealed reviewers were spending less time on "AI-verified" code. After adjusting review practices, DER normalized.

Consulting, Product & Strategy

AI is transforming how knowledge workers draft analysis, recommendations, and client deliverables. The risk here is subtle: AI-generated content often sounds confident and polished even when the underlying analysis is shallow or incorrect. Strategy teams need metrics that verify AI assistance actually improves output quality, not just quantity.

Metric	Application	Alert Threshold
TTPR	Client deliverables, memos	>40% increase
RIC	Iterations per deliverable	>3 for routine work
DER	Issues from clients/seniors	>3% or rising
JCI	Calibration on recommendations	<0.75
AOC	AI-assisted analyses	<90%

CASE STUDY: CONSULTING FIRM

A management consulting firm introduced AI for market analysis drafts. JCI tracking revealed analysts were accepting AI conclusions without verification. Their calibration score dropped from 0.85 to 0.62. Training on "AI as draft, not answer" restored calibration within 6 weeks.

Customer Support & Operations

Support teams were early AI adopters—chatbots, suggested replies, and automated ticket classification are now common. The efficiency gains are real, but so are the risks: incorrect information, tone-deaf responses, and frustrated customers. Support needs metrics that verify AI assistance doesn't degrade customer experience or escalate issues.

Metric	Application	Alert Threshold
TTPR	KB updates, escalations	>30% increase
DER	Ticket re-open rate	>8% or rising
CIR	Severe mishandling cases	Any increase
AOC	High-risk ticket categories	<95%

HR, Legal, Compliance & Finance

These functions handle decisions where errors have severe consequences: compensation disputes, legal liability, regulatory violations, financial misstatements. AI can accelerate drafting and analysis, but the review requirements are non-negotiable. These teams need metrics focused on governance, auditability, and zero tolerance for unreviewed high-stakes decisions.

Metric	Application	Alert Threshold
DER	Audit findings	Any increase
CIR	Legal/regulatory incidents	Zero tolerance
BOR	Bypass rate for regulated	>1%
AOC	Sensitive decisions	<98%

REGULATED INDUSTRIES

For industries with regulatory oversight (financial services, healthcare, pharmaceuticals, legal), add explicit documentation requirements for any override. Maintain audit-ready logs with justification trails. Many regulations require demonstrable human oversight of AI-influenced decisions—these metrics provide that evidence.

Adapting Thresholds to Your Context

The alert thresholds above are starting points, not universal standards. Your specific thresholds depend on factors like: error severity in your domain, regulatory requirements, team maturity, and baseline performance.

To calibrate thresholds: First establish baseline metrics without AI (or with minimal AI). Then run AI changes as experiments with generous thresholds initially. Tighten thresholds as you accumulate data and understand your team's patterns. Different workflows within the same function may need different thresholds—a routine PR doesn't need the same scrutiny as a security-critical deployment.

CALIBRATION TIP

Start with loose thresholds and tighten over time based on data. Overly strict initial thresholds create false alarms that erode trust in the metrics.

A practical 6-step guide

Implementing judgment metrics doesn't require a massive tooling investment or organization-wide rollout. The most successful implementations start small: one team, one workflow, a handful of metrics. This section provides a practical roadmap for getting started, proving value, and scaling what works.

The key insight: you don't need to instrument everything at once. Most organizations already have the underlying data—timestamps, defect logs, review records. The work is primarily joining and relabeling existing data, not building new infrastructure. Start with what you have, prove value, then invest in more sophisticated instrumentation.

Step 1: Choose Judgment-Critical Workflows

Map workflows on error-cost × tacitness. Start with high-stakes work: production deploys, high-value deliverables, compensation decisions, compliance sign-offs.

> **TIP**
>
> Avoid starting with simple tasks—they rarely reveal judgment failures.

Step 2: Define the Unit of Evaluation

Specify what counts as an "artifact," "review," and "approval." Artifacts: PRs, memos, contracts, tickets. Events: first submission, review start, final sign-off.

> **TIP**
>
> Consistency matters more than perfection. Pick definitions you can instrument reliably.

Step 3: Wire Up Existing Data Sources

Use current systems before buying new tools. Source TTPR, RIC, RL, AOC from code review tools and ticket systems. Link artifacts to defects for DER and CIR.

> **TIP**
>
> Most organizations have 80% of the data—it just needs joining and relabeling.

Step 4: Baseline Pre-AI Performance

Capture at least one full cycle of "as-is" data before AI changes. Use historical data if available, or create a control period by pausing AI tool rollout for select workflows. The baseline should be long enough to capture normal variation—typically one to two sprint cycles or one full project cycle.

> TIP
>
> Without a baseline, you can't distinguish AI impact from seasonal variation or other confounding changes.

Step 5: Run AI Changes as Experiments

Define explicit hypotheses before rolling out AI changes: "TTPR decreases 20%, DER stays flat, CIR doesn't increase." Set guardrails with clear thresholds—if metrics exceed acceptable levels, pause and adjust before continuing. This experimental mindset protects you from hidden degradation.

> TIP
>
> Treat AI adoption like a clinical trial. Hypothesize, measure, adjust. Don't just "try AI and see what happens."

Step 6: Standardize Dashboards and Governance

Once you've validated patterns, codify them into standing dashboards and operating rules. Create metric bundles (not single targets) for each workflow to avoid gaming. Establish regular review cadences where teams examine their metrics and discuss anomalies.

> TIP
>
> Single-metric targets get gamed. Balanced scorecards that track multiple pillars are harder to manipulate.

Common Implementation Patterns

Pattern A: Engineering-Led (Bottom-Up) — Start with one engineering team tracking DER and TTPR on their AI-assisted PRs. Prove value locally, then expand to adjacent teams. Leadership gets involved once you have data showing impact.

Pattern B: Risk-Led (Top-Down) — Compliance or risk function mandates AOC tracking for AI-assisted decisions in regulated workflows. Engineering implements instrumentation. Start with audit requirements, expand to operational metrics.

Pattern C: Pilot Program — Select a cross-functional pilot group representing different functions. Each team picks their most relevant metric bundle. Share learnings across the pilot group weekly. Scale what works.

Avoiding Common Pitfalls

Teams implementing judgment metrics often encounter similar obstacles. Here's how to navigate them:

Pitfall: "We don't have the data." You probably do—it's just not labeled as judgment data. Review timestamps exist in your code review tool. Defect data exists in your bug tracker. Customer complaints exist in your support system. The work is joining and relabeling, not building from scratch.

Pitfall: "This feels like surveillance." Frame metrics at the workflow level, not individual level. The question isn't "which reviewer is slowest?" but "is our review process sustainable?" Team-level metrics drive process improvement; individual metrics create anxiety and gaming.

Pitfall: "Leadership only cares about velocity." Translate judgment metrics into risk language. "DER is up 40%" becomes "we're catching 40% fewer issues before customers see them." "eTLX is spiking" becomes "our expert reviewers are 3x more likely to burn out this quarter." Connect to business outcomes leadership cares about.

Pitfall: "The metrics look fine but something feels wrong." Trust qualitative signals too. If team members report feeling overwhelmed but eTLX looks stable, your eTLX measurement may be miscalibrated. Metrics complement—they don't replace—human judgment about human judgment.

Pitfall: "We implemented metrics but nothing changed." Metrics without action are just overhead. Every metric review should end with "what will we do differently?" If metrics consistently show problems but nobody adjusts workflows, you've built a dashboard, not a governance system.

> SUCCESS CRITERIA
>
> A successful implementation means: teams discuss their metrics regularly, anomalies trigger investigation, and AI workflows get adjusted based on what the data shows. The metrics should drive decisions, not just fill dashboards.

Answers to frequently asked questions

Implementing judgment metrics raises practical questions. This section addresses the most common concerns we hear from teams adopting this framework. If your question isn't covered here, visit aicognifit.com for discussion forums and additional resources.

Q: How is this different from velocity dashboards?

Velocity measures how fast you produce. Judgment metrics measure how well you evaluate. Traditional metrics can improve even while error rates and fatigue worsen. Velocity tells you "we shipped 50 PRs"; judgment metrics tell you "we caught 95% of issues before production." Both matter—but most organizations only track velocity.

Q: Do I need new tooling?

Usually not at first. Most metrics derive from existing data: code review timestamps, ticket logs, QA records. Start by joining and relabeling what you have. Only invest in new tooling when you've validated the value of specific metrics and need better instrumentation.

Q: How should I interpret a rising Defect Escape Rate?

Rising DER is an early warning that review isn't keeping pace with volume. It doesn't mean AI is "bad"—it means you may need different workflows, more reviewers, or better reviewer training. Treat it as a signal to investigate root causes, not as a verdict on AI adoption.

Q: What's the relationship between Throughput and Burden?

There's natural tension: pushing for more throughput increases cognitive burden. Sustainable workflows balance both. If throughput rises but eTLX degrades, you're borrowing against future productivity—heading toward burnout and quality collapse.

Q: Are these metrics research-validated?

Yes. DER is standard in quality management (ISO/QMS frameworks). NASA-TLX is a gold-standard workload measure used for decades in aviation and healthcare. JCI builds on Brier score and calibration research from forecasting. This playbook adapts proven measures for AI-augmented contexts.

Q: How do I get leadership buy-in?

Frame it as risk management, not productivity policing. Position judgment metrics as the "AI guardrail" that provides visibility into strategic risk while enabling faster adoption. Connect to outcomes leadership cares about: reduced incidents, lower attrition, audit readiness.

Q: Can I use this in regulated industries?

Regulated industries are exactly where judgment metrics matter most. They provide the auditable trail regulators want: who reviewed what, how long it took, what escaped, and whether bypasses are documented and justified.

Q: What's the minimum viable implementation?

Pick one high-stakes workflow. Instrument TTPR, DER, and AOC using existing tools. Baseline for one sprint or cycle. Then introduce or adjust AI usage and compare. That's enough to validate the approach and build evidence for broader adoption.

Q: How often should we review these metrics?

Weekly for operational teams, monthly for leadership. Weekly reviews catch issues quickly. Monthly reviews track trends and inform strategic decisions. Avoid daily reviews—metrics need time to show meaningful patterns.

Q: What if different teams have different AI maturity levels?

That's expected. Start each team where they are. A team new to AI might focus on AOC (is governance happening at all?). A mature team might focus on JCI (is judgment quality improving?). The framework scales—use the metrics appropriate to each team's stage.

About AI CogniFit

AI CogniFit is built around the same principle as this playbook: the main constraint in AI adoption is human judgment, not model performance. As AI tools become more powerful, the bottleneck shifts from "can AI do this task?" to "can humans reliably evaluate AI output?"

Our platform helps individuals and teams measure, calibrate, and improve their judgment under uncertainty. We provide assessments that reveal blind spots, training that builds calibration, and metrics that track improvement over time. Whether you're an individual wanting to improve your AI collaboration skills or a leader implementing judgment metrics across teams, we have tools to help.

Explore Our Platform

Research Quality Standards

aicognifit.com/research

Understand how we grade evidence behind recommendations. Our A/B/C research quality framework ensures transparency about what's well-established versus emerging.

Human Judgment Arena

aicognifit.com/arena

Practice judgment with calibrated feedback. Compete on real-world scenarios across domains like forecasting, analysis, and risk assessment. Track your improvement over time.

AI Literacy Assessments

aicognifit.com/literacy-hub

Measure AI collaboration readiness for yourself or your team. Identify blind spots in prompt engineering, output evaluation, and effective human-AI teaming.

Judgment Metrics Dashboard

aicognifit.com/judgment-metrics

Interactive exploration of this framework with calculators, examples, and implementation templates you can adapt for your organization.

Quick-Start Checklist

Use this checklist to guide your first implementation. Each item builds on the previous one—complete them in order for the smoothest path to value.

1. **Identify one high-stakes workflow** where AI is already in use or being considered. Choose something where errors have real consequences—production deploys, client deliverables, compliance reviews. Avoid starting with low-stakes work that won't reveal meaningful patterns.
2. **Define your units clearly:** What counts as an "artifact"? What constitutes a "review"? When is something "approved"? Write these definitions down. Consistency matters more than perfection—pick definitions you can instrument reliably.
3. **Instrument TTPR, DER, and AOC** using existing tools. Pull timestamps from your code review system or ticket tracker. Link defects to their originating artifacts. Track which AI-assisted work goes through human review.
4. **Baseline current performance** over one sprint or project cycle. You need to know where you started to measure progress. If AI is already in use, consider a control period or use historical pre-AI data.
5. **Set hypotheses and guardrails** before making changes. "We expect TTPR to decrease 20%, DER to stay flat, AOC to stay above 95%." Define what would cause you to pause and investigate.
6. **Review metrics weekly** with your team. Discuss anomalies. Adjust AI usage based on what the data shows. The goal is continuous improvement, not just measurement.

Ready to implement?

Visit aicognifit.com/judgment-metrics for interactive tools, downloadable templates, and community discussion. Join thousands of organizations building sustainable AI practices with visibility into human judgment.

Final Thought

AI isn't going away, and neither is the need for human judgment. The organizations that thrive will be those that measure both—using AI's speed where appropriate while maintaining rigorous oversight where it matters. This playbook gives you the framework. The implementation is up to you.

Start small. Prove value. Scale what works. And remember: the goal isn't perfect metrics—it's better decisions, sustainably made.

AI CogniFit

Measuring human judgment in the age of AI

aicognifit.com

© 2025 AI CogniFit. All rights reserved.