Brandy McCarron
AI SYSTEMS STRATEGIST

SAMPLE DOCUMENT

DIAGNOSTIC FRAMEWORK

AI System *Diagnostic*

A practical framework for diagnosing broken AI systems — where outputs fail, why leadership doesn't trust them, and what needs to change before AI can scale.

6% → 81%	33	900+	Zero
Accuracy gain, real engagement	AI systems governed	Stakeholders aligned	Model changes required

© 2026 Brandy McCarron · Confidential Sample

RECOGNITION

When AI works technically but fails operationally

The most common AI failure isn't model failure — it's system failure. The model responds, but the surrounding context, evaluation, and governance are too weak for the business to trust what comes out.

AI outputs vary by user

Different teams get different answers because no shared context or standards exist across the organization.

Analysts still translate everything

AI returns data, but humans still convert it into recommendations before leadership can act on it.

Leadership doesn't trust the system

The system technically works, but no one actually relies on it for decisions that matter.

Failures can't be explained

There's no evaluation trail, failure log, or repeatable testing process.

Diagnostic goal: Identify the smallest system changes that create the largest gain in accuracy, trust, and usability — without replacing the model.

THE FRAMEWORK

The system behind the answer

A model answer is only the visible output. The diagnostic examines the full system that produced it: user intent, model behavior, business context, evaluation logic, and governance controls.

01 Context Architecture

Does the system understand the business — metrics, product logic, constraints, and the decision hierarchy that determines what a "correct" answer actually means?

02 Evaluation System

Can the team measure accuracy, reasoning quality, risk, and failure patterns over time? Without measurement, improvement is invisible and claims are anecdotal.

03 Governance Layer

Are standards, ownership, auditability, and escalation paths defined before outputs reach stakeholders? Governance built into the system — not bolted on after.

Working thesis: If the model has access to the same data but produces inconsistent or unusable results, the fix is almost always context and evaluation — not a new model.

PHASE 01

System Audit — Gap Assessment

Three gap categories, each scored 1–5. A score of 3 or below in any category indicates a systemic issue requiring intervention before the system can be trusted at scale.

GAP TYPE	WHAT IT MEANS	COMMON SYMPTOMS
Context Gap	The AI lacks the business logic, metric definitions, and decision context it needs to reason correctly about the company's reality.	Plausible but wrong outputs. Inconsistent answers across users. Results require human interpretation before anyone acts.
Evaluation Gap	No system to measure whether AI outputs are accurate or reliable. Problems go undetected until they surface as visible failures.	No one can explain why outputs fail. Errors recur without improvement. Accuracy claims are anecdotal.
Governance Gap	No shared standards for how AI is built, prompted, or evaluated across teams. Behavior is unpredictable and unauditable.	Inconsistency across users and teams. Compliance risk accumulating. Leadership can't audit what they can't predict.

How scoring works: Each category is scored 1–5 across four dimensions: existence of standards, consistency of application, measurability of outcomes, and organizational ownership. 1–2 = critical gap. 3 = functional but fragile. 4–5 = production-ready. Most organizations score 1–2 across all three on first audit.

PHASE 02

Failure Mapping — Classify Before You Fix

The most expensive mistake in AI remediation is fixing symptoms instead of root causes. Failure mapping classifies every failure by root cause first — the same bad output can trace back to three different problems, each requiring a different fix.

CONTEXT FAILURE

Root cause: The model doesn't know what it needs to know about the business.

Missing metric definitions, unexplained data relationships, business logic that lives only in analyst heads.

Fix: Context architecture — embed business knowledge into the system layer.

EVALUATION FAILURE

Root cause: No mechanism to detect when outputs are wrong.

Accuracy unmeasured. No rubric defines "correct." Failures surface only when a human notices.

Fix: Evaluation system — define correctness, build measurement, create improvement loop.

GOVERNANCE FAILURE

Root cause: No shared standards for how the system should behave.

Inconsistent prompting across teams. No accountability. Compliance risk accumulating unseen.

Fix: Governance layer — standards, ownership, auditable behavior.

COMPOUND FAILURE

Root cause: Multiple gap types reinforcing each other.

Context gaps corrupt output. Evaluation gaps hide it. Governance gaps prevent correction.

Fix: Sequenced remediation — context first, evaluation second, governance to lock in gains.

> **Example from a real engagement:** Low accuracy wasn't caused by the model or the data source. The model had data access but lacked metric definitions, domain language, product constraints, and examples of what a correct answer looked like. Once those were structured into context modules and scored through a repeatable evaluation set, accuracy improved from 6% to 81%.

REAL ENGAGEMENT · COMPOUND CONTEXT + EVALUATION FAILURE

Kitewing — 75-Point Accuracy Gain. Zero Model Changes.

AI reasoning system · Starting accuracy 6% · A compound failure, diagnosed and fixed at the system layer.

KEY RESULTS

6%	81%	+75 pts	0
Starting accuracy	Post-remediation accuracy	Accuracy improvement	Model changes required

AUDIT FINDINGS — WHAT THE SYSTEM LACKED

- No business context — the model had no knowledge of what the company did or how decisions were made
- No metric definitions — "correct" was never defined, so accuracy couldn't be measured
- No evaluation rubric — failures were invisible until a human happened to notice
- No domain vocabulary — model reasoned generically, not in business terms
- No output format standards — responses weren't structured for decision use

WHAT WAS FIXED — SYSTEM CHANGES ONLY

- Built complete business context layer — company structure, decision types, data relationships
- Defined accuracy rubric — 5-dimension scoring with documented criteria for "correct"
- Built evaluation pipeline — automated scoring of output samples against rubric
- Embedded domain vocabulary and reasoning constraints into context
- Redesigned outputs around the decision types leadership needed to act on

ACCURACY BEFORE VS. AFTER

> **The takeaway:** The model never changed. The system around it did. Every point of improvement came from giving the model what it needed — business context, a definition of "correct," and a mechanism to measure it. **All 75 points: system architecture only.**

EVALUATION FRAMEWORK

How I score whether AI is decision-ready

The rubric separates technically plausible answers from answers a business leader can act on. Scores reveal patterns across the system — not just whether a single output is right or wrong.

DIMENSION	SCORE	WHAT IT MEASURES	FAILURE SIGNAL
Factual Accuracy	4–5 · Pass	Is the answer factually correct based on available data and documented business logic?	Correct-looking response with wrong numbers, flawed logic, or unsupported assumptions.
Business Context	3 · Review	Does the system understand definitions, relationships, and the company's specific constraints?	Generic answer that ignores company-specific logic, terminology, or decision hierarchy.
Reasoning Quality	1–2 · Fail	Does the answer explain the why, not just the what? Interpretation, not just data retrieval.	Data retrieved without interpretation, context, or any recommendation on what to do next.
Decision Readiness	4–5 · Pass	Can leadership act on the output directly — no analyst translation required?	Requires a human to rewrite, qualify, or interpret before any real decision can be made.
Compliance Safety	4–5 · Pass	Does it avoid unsupported claims, PII exposure, and policy violations?	Confident but unverified, unsafe, or non-compliant output that would concern legal or audit.

Decision-ready threshold: A system is not treated as decision-ready until it scores consistently across all five dimensions — not just one. For Kitewing, the composite score moved from 1.4 to 4.1 over four months of remediation. No model changes.

BEFORE / AFTER

Clear findings. Prioritized fixes. Measurable results.

The diagnostic doesn't stop at recommendations. Every finding maps to a specific fix, and every fix maps to a measurable outcome — so teams know exactly what to do first and how to prove it worked.

BEFORE	SYSTEM FIX	AFTER
AI gives data, not insight.	**Add business context modules and answer patterns.**	AI explains what changed, why it matters, and what to do next.
Output varies by user.	**Standardize prompts, context inputs, and response rules.**	Output is consistent across teams and use cases.
Failures are anecdotal.	**Create evaluation set, scoring rubric, and failure categories.**	Failures are measurable, repeatable, and fixable.
Leadership doesn't trust the answer.	**Add QA loop, governance checks, and ownership model.**	Leadership has a documented basis for confidence and escalation.

DELIVERABLES

DELIVERABLE	PURPOSE
Failure map	Shows where outputs break and which failures repeat across prompts, users, and use cases.
Evaluation rubric	Defines how output quality is scored and creates a repeatable, trackable accuracy baseline.
Context architecture plan	Identifies missing business logic, definitions, ownership, and decision patterns.
Governance gap analysis	Surfaces risks in consistency, compliance, ownership, and escalation paths.
Prioritized fix roadmap	Sequences the highest-leverage fixes so teams act without trying to fix everything at once.

Best-fit use case This diagnostic is designed for post-pilot AI systems: the tool exists, people have tried it, but trust, adoption, consistency, or decision quality isn't where it needs to be yet.

Brandy McCarron
AI SYSTEMS STRATEGIST

NEXT STEP

If AI outputs aren't trusted, they won't *scale* — regardless of the model.

I work with companies that already have AI in motion but need the system around it to become reliable, measurable, and trusted enough to act on.

Best fit: teams with AI already in production but struggling with trust, consistency, or decision quality.

connect@brandymccarron.com →

Response within 24 hours. No pitch deck.

brandymccarron.com
linkedin.com/in/brandymccarron

ENGAGEMENTS

2–3 Week Diagnostic
System audit, failure map, fix roadmap. Fixed scope, fixed fee.

3–6 Month Fractional
Embedded AI sys