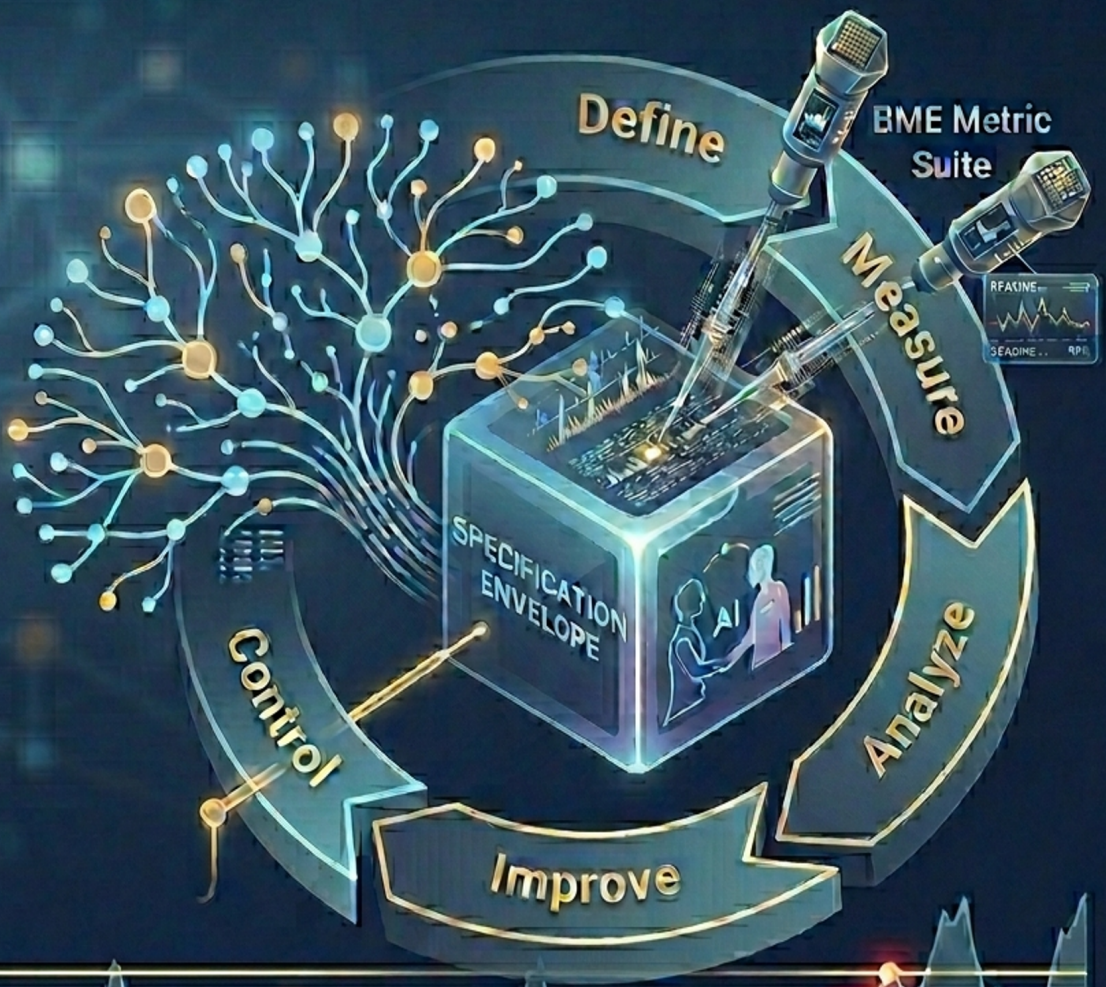
AI BEHAVIORAL ASSURANCE

A Lean Six Sigma Methodology for the Lifecycle Governance of Large Language Models and Agentic AI Systems



Dale Rutherford

The Center for Ethical AI

AI Behavioral Assurance

A Lean Six Sigma Methodology for the Lifecycle Governance of Large Language Models and Agentic AI Systems

First Edition

Dale Rutherford

The Center for Ethical AI
Little Rock, AR, USA

AI Behavioral Assurance

A Lean Six Sigma Methodology for the Lifecycle Governance of Large Language Models and Agentic AI Systems

First Edition

ISBN: 979-8-90452-913-0

dale.rutherford@thecenterforethicalai.com
The Center for Ethical AI, Little Rock, AR, USA

Preface

This work began with an observation: organizations deploying Large Language Models and agentic AI systems in consequential decision environments lacked a process-level methodology for governing those systems' behavioral outputs. Strategic governance frameworks existed. Risk management principles existed. What did not exist was the operational bridge: a structured, statistically grounded methodology that translates governance policy into measurable specifications, validates those specifications against empirical behavioral data, diagnoses the root causes of governance failures, designs and validates corrective interventions, and sustains conformance through continuous statistical monitoring.

That bridge is what classical quality management has provided to manufacturing for seven decades. Statistical Process Control, the DMAIC improvement cycle, and the Lean Six Sigma toolkit were designed to govern stochastic processes: systems whose outputs vary, whose variation has identifiable sources, and whose behavior responds to deliberate intervention. The structural claim of this work is that LLM and agentic AI systems are exactly such processes, and that the methodology developed to govern them in manufacturing can be systematically translated to govern them in AI deployment.

The translation is not trivial. LLM systems violate several assumptions that classical SPC takes for granted: their output distributions are non-normal, their operating conditions are non-stationary, their input spaces are unbounded, and their internal mechanisms are opaque. Each of these violations requires either adapting existing tools or building new ones. The Adopt, Adapt, Innovate taxonomy that structures this work is the instrument for making those decisions with formal precision rather than informal analogy.

The result is a comprehensive governance methodology: five DMAIC phases translated with operational protocols and tool specifications, a measurement system (the BME Metric Suite) purpose-built for behavioral process control, a governance architecture (ALAGF) that embeds DMAIC as a repeatable process within the AI lifecycle, and a complete taxonomy classifying 40 tools and constructs by their applicability to AI governance.

I wrote this work for three audiences. For standards bodies, it provides a process-level implementation methodology that existing AI governance frameworks require but do not themselves specify. For governance researchers, it provides a theoretical

framework connecting statistical process control to AI behavioral assurance. For practitioners deploying AI systems in regulated environments, it provides the operational discipline to transform governance aspirations into measurable, auditable conformance.

The framework is not complete. Chapter 11 identifies boundary conditions and open problems that require sustained research attention. But the structural claim is secured, and the methodology is specified with sufficient rigor to serve as a normative reference. I invite the governance community to test, refine, and extend it.

Dale Rutherford, PhD
Little Rock, Arkansas
March 2026

Acronyms and Abbreviations

Acronym	Definition
AAI	Adopt, Adapt, Innovate (tool classification taxonomy)
ALAGF	Adaptive Lifecycle Agentic Governance Framework
BAAGF	Behavioral Assurance and Agentic Governance Framework
BAR	Behavioral Assurance Rating
BME	Behavioral Measurement of Entropy
CBMES	Composite Behavioral Measurement Evaluation Score
CDSS	Clinical Decision Support System
COPQ	Cost of Poor Quality
CTG	Critical-to-Governance
CUSUM	Cumulative Sum (control chart)
DAST	Dynamic Adversarial Stress Testing
DMAIC	Define, Measure, Analyze, Improve, Control
DOE	Design of Experiments
ECI	Entropy Control Index
ECPI	Entropy-Calibrated Performance Index
EWMA	Exponentially Weighted Moving Average (control chart)
FMEA	Failure Mode and Effects Analysis
GRR	Gauge Repeatability and Reproducibility
LLM	Large Language Model
MIDCOT	Multi-Dataset IQ Drift and Cost Optimization Training
MSA	Measurement System Analysis
RAG	Retrieval-Augmented Generation
RACI	Responsible, Accountable, Consulted, Informed
RPN	Risk Priority Number
SIPOC	Supplier, Input, Process, Output, Customer
SPAR	Stochastic Process Adherence Ratio
SPC	Statistical Process Control
SymPrompt+	Symbolic Prompt Modulation
VOG	Voice of Governance
VSM	Value Stream Mapping

For the complete glossary including symbols and mathematical notation, see Appendix D.

Table of Contents

Chapter 1

Foundations: Process Thinking and Stochastic AI Systems

The argument of this work rests on a structural claim: that Large Language Model (LLM) and agentic AI systems satisfy the necessary and sufficient conditions for treatment as governable processes under Statistical Process Control (SPC). This claim is not self-evident. Classical SPC was developed for manufacturing contexts where processes are physically instantiated, inputs are enumerable, and outputs are measurable against deterministic specifications. LLM systems violate several of these assumptions: their internal states are opaque, their input spaces are effectively unbounded, and their outputs are sampled from probability distributions that shift with context, training state, and deployment configuration. The question is whether these violations disqualify LLMs from process governance or whether they instead demand a more general formulation of what constitutes a governable process.

This chapter establishes the theoretical foundations required to answer that question. It proceeds in six stages: process ontology, LLM characterization as stochastic systems, the complications of non-stationarity and emergent behavior, the redefinition of quality as behavioral conformance, the introduction of behavioral entropy as the governing process variable, and the formalization of the Adopt, Adapt, Innovate (AAI) taxonomy that governs every methodological classification in subsequent chapters.

1.1 Process Ontology: What Constitutes a Governable Process

The concept of a "process" in quality management is both specific and capacious. ISO 9001:2015 defines a process as a "Set of interrelated or interacting activities that uses inputs to deliver an intended result "[17]. Juran's formulation is more operational: a process is "a systematic series of actions directed to the achievement of a goal" whose outputs can be measured against specifications [22]. Deming's conceptualization, foundational to SPC, emphasizes that a process is any system of causes producing a result, and that the result's variability is the observable signature of the system's state [7]. These definitions share a common architecture: inputs, transformation logic, outputs, and measurable variability.

For a process to be *governable* under SPC, it must satisfy conditions beyond mere input-output structure. Drawing from Montgomery's formalization and Wheeler's operational criteria [28, 48], I propose five necessary conditions:

Definition 1.1 (Process Governability Conditions)**.** A process is *governable* under Statistical Process Control if and only if it satisfies the following five conditions:

G1. Observable outputs. The process produces outputs that can be observed and recorded. Observability does not require determinism; it requires that each execution yield a datum entering the measurement system.

G2. Measurable variation. The outputs exhibit variation quantifiable using a defined measurement system.

G3. Distributional characterization. The variation is characterized as a statistical distribution, even if non-normal, multimodal, or time-varying.

G4. Causal attribution. It is possible, at least in principle, to attribute variation to identifiable sources. This is the foundation of Shewhart's common-cause and special-cause distinction [41].

G5. Intervention responsiveness. The process responds to deliberate interventions in ways that alter its output distribution.

These conditions are deliberately minimal. They do not require linearity, stationarity, independence of observations, or normality. These stronger assumptions are often invoked in classical SPC practice, but they are simplifying assumptions that make certain tools tractable rather than foundational requirements of process governance itself. When they fail, tools must be extended or replaced, but governability is not revoked.

Remark. This distinction is critical: *the governability of a process is an ontological property* determined by the five conditions above, while *the applicability of specific SPC tools is a methodological property* determined by additional distributional assumptions. The AAI taxonomy (Section 1.6) operationalizes exactly this distinction.

The process approach codified in ISO 9001:2015 reinforces this position [17]. Nothing in this specification restricts the process approach to deterministic or physical processes. ISO/IEC 42001:2023 explicitly adopts a process approach for AI system governance [19], inheriting the process concept from ISO 9001 and extending it to AI lifecycle activities. This inheritance creates a direct bridge: if a process approach governs AI management, and if that approach is operationalized through SPC and DMAIC, then the only remaining question is whether AI systems satisfy the conditions under which SPC operates.

Before proceeding, it is worth distinguishing *process governance* from *model governance*. Model governance focuses on internal characteristics: architecture, training data, parameter configuration, and development lineage [26]. Process governance focuses on behavioral characteristics in operation: what the system produces, how outputs vary, whether variation is stable or drifting, and whether interventions can bring the system into conformance. The concrete problems in AI safety that Amodei et al. cataloged [1] are, in the framework of this work, behavioral failure modes amenable to statistical detection and governance response. This work is concerned with process governance: the statistical characterization, monitoring, and control of LLM behavioral outputs.

1.1.1 SPC Beyond Manufacturing: Precedent Domains

The application of SPC to stochastic, non-manufacturing processes is not unprecedented. Three domains provide instructive precedent for the extension proposed in this work. The broader context is significant: international bodies have established governance expectations for AI systems [32], and a global convergence of AI ethics principles has emerged [21], but operationalizing these principles into measurable, auditable process governance requires the kind of statistical methodology this work develops.

Financial risk monitoring. Value-at-Risk (VaR) and Expected Shortfall calculations use distributional characterization and threshold-based monitoring to govern portfolio risk. Financial return distributions are non-normal (heavy-tailed, skewed), non-stationary (volatility clustering, regime changes), and autocorrelated (GARCH effects). The financial risk community addressed these challenges by replacing parametric assumptions with empirical and semi-parametric methods: historical simulation, kernel density estimation, and extreme value theory. The parallel to AI governance is direct: BAR thresholds replace parametric control limits using the same empirical quantile logic that historical VaR replaced parametric VaR [38].

Biological process control. Pharmaceutical manufacturing and bioprocess monitoring apply SPC to inherently variable biological systems: fermentation yields, cell culture viability, and enzyme activity. These processes exhibit the same

characteristics that complicate AI governance: non-normal distributions, batch-to-batch variation analogous to model-to-model variation, and measurement systems that depend on subjective expert assessment (e.g., microscopy grading, organoleptic evaluation). The bioprocess community developed multivariate SPC methods (Hotelling T^2, principal component analysis-based monitoring) and batch-specific baseline protocols that accommodate legitimate process variation. The re-baselining protocol specified in Chapter 7 draws directly from this precedent [49].

Software quality assurance. Software defect detection applies SPC to processes whose "outputs" (code modules, test results) are discrete, variable in nature, and subject to human judgment in classification [40]. Software reliability models use stochastic process theory to characterize defect discovery rates and predict residual defect counts. The measurement challenges parallel AI governance: inter-rater reliability in defect-severity classification, ground-truth construction for test-oracle design, and the distinction between systematic defects (special causes) and residual complexity (common causes). The MSA/GRR adaptation for evaluator-based measurement (Chapter 4) addresses a measurement challenge that software quality research identified decades ago.

These precedents establish that extending SPC beyond manufacturing is a methodological tradition rather than a speculative proposal. Each precedent domain replaced parametric assumptions with domain-appropriate alternatives while preserving the interpretive framework of process governance. Deming's later systems thinking framework [8] provides the philosophical foundation: any system of interdependent components producing variable outputs is amenable to process thinking, regardless of whether its components are physical or computational. This work extends that tradition to AI behavioral processes.

1.2 LLMs as Stochastic Input-Output Systems

A Large Language Model is, at the level of its inference mechanism, a conditional probability distribution over token sequences. Given an input sequence, the model assigns a probability $P(y \mid x)$ to each possible output token y at each generation step, conditioned on the input and all previously generated tokens [43, 5]. The output sequence is generated via autoregressive sampling, so even with a fixed input, the model can produce different outputs across runs. This is not a defect; it is a structural property of the system's design.

The stochastic character of LLM inference has three governance-relevant consequences. First, single outputs are samples, not deterministic mappings: any individual output may or may not be representative of the system's "typical" behavior for the given input class. Governance must therefore operate on distributions of outputs rather than on individual instances. Second, the output distribution is conditioned on the full input context (system instructions, conversation history, retrieved content, user query), which means that the behavioral distribution shifts

with context in ways that may or may not be governed. Third, the sampling mechanism itself (temperature, top-k, top-p parameters) constitutes a governance-relevant configuration parameter: higher temperature increases output entropy, widening the behavioral distribution; lower temperature concentrates it. These parameters are not incidental; they directly control the trade-off between behavioral diversity and behavioral predictability that the specification envelope must accommodate.

Proposition 1.1 (LLM Governability)**.** LLM and agentic AI systems satisfy all five process governability conditions (Definition 1.1) and are therefore governable processes under SPC.

Proof sketch. Conditions G1 and G2 are satisfied by definition: every inference call produces an observable output, and repeated inference on identical or similar inputs produces outputs that vary. The remaining three conditions require demonstration.

Distributional characterizability (G3). For a given *input class* (a category of prompts with shared semantic characteristics), the distribution of outputs can be characterized along multiple dimensions: lexical diversity, semantic coherence, factual accuracy, tone consistency, and behavioral alignment with specified policies. These dimensions define a multivariate behavioral distribution estimable through repeated sampling [24, 3]. The distributions are typically non-normal and may exhibit heavy tails, but these properties complicate the use of specific tools without undermining characterizability.

The concept of an "input class" deserves formal attention, as it plays the role in LLM governance that a "process condition" plays in manufacturing SPC. Defining input classes precisely is a prerequisite for meaningful SPC; without them, behavioral variation conflates process drift with input diversity, rendering control charts uninterpretable.

Causal attributability (G4). Sources of variation in LLM outputs fall into categories mirroring the common-cause and special-cause distinction:

Model-intrinsic sources:
training data composition and biases, learned parameter distributions, attention pattern sensitivities, tokenization artifacts [46].

Configuration sources:
sampling parameters (temperature, top-k, top-p), system prompt content, retrieval augmentation quality, tool-use configurations for agentic systems.

Deployment sources:
model versioning and update cadence, quantization effects, hardware variation, load-dependent latency effects.

Input-population sources:
distributional shift in user query patterns, adversarial or out-of-distribution inputs, prompt injection attempts [34, 14].

Intervention responsiveness (G5). LLM systems are among the most intervention-responsive processes in engineering. Prompt modification, system instruction revision, fine-tuning , RLHF, retrieval corpus curation, sampling parameter adjustment, and output filtering each produce measurable shifts in the output distribution [33, 2]. The challenge for governance is not whether interventions work but whether their effects are predictable, reversible, and measurable. □

1.3 Non-Stationarity, Latent Variables, and Emergent Behavior

The five governability conditions are satisfied, but three characteristics of LLM processes complicate the application of specific SPC tools. Each triggers the Adapt or Innovate branches of the AAI taxonomy without revoking governability.

Non-stationarity. Classical SPC assumes stable mean and variance under normal conditions. LLM systems violate this assumption in a structured way: behavioral distributions shift due to both authorized changes (model updates, retrieval corpus maintenance, prompt revisions) and unauthorized drift (gradual degradation, input population shift, emergent adversarial exploitation).

Definition 1.2 (Non-Stationarity Classification)**.** *Legitimate non-stationarity* arises from authorized, documented process changes that are part of the governance lifecycle. *Unauthorized drift* arises from undocumented, unplanned, or adversarial changes that violate the governance charter.

Legitimate non-stationarity requires re-baselining; unauthorized drift requires investigation through the Analyze and Improve phases of DMAIC. This distinction has direct implications for control chart design: classical control charts must be extended by a change-management integration layer. Control charts are therefore classified as **[D ▷ Adapt]** under the AAI taxonomy.

The practical consequence of this classification is architectural: the monitoring system must maintain a change registry that records all authorized process changes with their effective dates and expected behavioral effects. When the monitoring system detects a distributional shift, it first consults the change registry. Shifts coinciding with registered changes are routed to the re-baselining protocol; shifts not attributable to any registered change are routed to the escalation protocol for investigation. This routing logic, implemented in MIDCOT's detection layer (Chapter 7), has no precedent in classical SPC because classical processes do not undergo intentional, periodic behavioral modification.

Four categories of unauthorized drift warrant distinct governance responses:

Gradual degradation

Slow decline in behavioral quality without a discrete triggering event. Detected by CUSUM charts; diagnosed in the Analyze phase as common-cause degradation requiring systemic intervention.

Input population shift
The distribution of user queries changes (new use cases, seasonal patterns, adversarial probing campaigns), altering the effective input class distribution. Detected as class-specific SPAR decline; addressed through input class registry revision and specification recalibration.

Retrieval corpus staleness
For RAG systems, the knowledge base becomes outdated relative to the domain's current state. Detected as grounding fidelity decline; addressed through corpus governance protocols within BAAGF's Phase 1 lifecycle operations.

Emergent adversarial exploitation
Novel prompt injection or jailbreak patterns that exploit behavioral vulnerabilities not anticipated during specification. Detected as sudden anomalous behavioral patterns; escalated to T3+ for immediate investigation.

Latent variables. LLM behavior depends on variables that are not directly observable: internal activation patterns, attention distributions, and learned representations. The governance response adopted in this work is a *behavioral-first governance posture*: it operates on externally observable behavioral measurements rather than on internal model states. The BME Metric Suite (Section 1.5) operationalizes this posture. This is a principled choice, not a workaround for a limitation: stakeholders care about behavioral conformance (what the system does), not internal mechanisms (how it does it). A system whose internal representations are opaque but whose behavioral outputs consistently satisfy governance specifications is governable in every sense that matters for deployment assurance.

Emergent behavior. LLM systems exhibit behaviors not explicitly trained, specified, or anticipated [45]. This work treats emergent behavior as *detectable and responsive, not preventable*: monitoring systems must have sufficient sensitivity to detect novel behavioral patterns, and escalation protocols must have sufficient authority to respond (Chapter 7). The specification envelope, by definition, cannot anticipate emergent behaviors; it defines the expected behavioral space. The monitoring system's role is to detect when behavior falls outside that space, regardless of whether the deviation was anticipated. The tiered escalation model ensures that the governance response is proportional to the severity of the deviation.

1.4 Defining Quality: From Defect Rate to Behavioral Conformance

Classical quality management defines output quality relative to specifications and quantifies it by defect rate. LLM outputs violate the assumptions underlying this framework: behavioral quality characteristics are multidimensional, subjective, and context-dependent; specifications are constructed from governance policy rather than engineering tolerances; and conformance is graded rather than binary.

Definition 1.3 (Behavioral Conformance)**.** *Behavioral conformance* is the degree to which a system's outputs, across a defined input class and measurement period, fall within the multidimensional governance specification envelope on all governed behavioral dimensions simultaneously.

Three constructs operationalize this framework:

Voice of Governance (VOG)
: The analog to Voice of the Customer (VOC). VOG aggregates governance requirements from the full stakeholder constituency: end users, deploying organizations, regulatory bodies, affected populations, and the public interest as instantiated in applicable standards.

Critical-to-Governance (CTG) characteristics
: The analogue to Critical-to-Quality (CTQ). Each CTG characteristic has a defined measurement method, target value, and specification limit.

Governance specification envelope
: The multidimensional region of acceptable behavioral output defined by the complete set of CTG characteristics and their specification limits. A system output is conformant if and only if it falls within specification on all governed dimensions simultaneously.

1.5 Behavioral Entropy as the Governing Process Variable

Behavioral entropy, as defined in this work, quantifies the degree of unpredictability, inconsistency, or disorder in a system's behavioral outputs across a defined input class. The governance objective is to maintain behavioral entropy within a specified range: excessively low entropy may indicate rigid behavior; excessively high entropy may indicate instability.

Behavioral entropy is operationalized through the Behavioral Measurement of Entropy (BME) Metric Suite, comprising four primary indices and one composite score:

Definition 1.4 (BME Metric Suite)**.** The BME Metric Suite comprises:

ecpi (Entropy-Calibrated Performance Index)
Process capability analogue. Compares observed behavioral entropy against specification limits using empirical quantiles:

$$\text{ECPI} = \frac{LSL_{\text{spec}} - Q_\alpha}{Q_{0.5} - Q_\alpha} \tag{1.1}$$

where $Q_{0.5}$ is the median, Q_α is the α-quantile (typically 0.135th percentile), and LSL_{spec} is the specification limit. Replaces classical C_p/C_{pk}.

bar (Behavioral Assurance Rating)
Control limit analogue. Empirical quantile-based thresholds defining expected behavioral variation boundaries. Replaces $\bar{x} \pm 3\sigma$.

eci (Entropy Control Index)
Sigma-level equivalent. Distance from current behavioral entropy to the nearest governance threshold in empirical standard deviation units. ECI ≥ 3.0 indicates Three Sigma equivalent stability.

spar (Stochastic Process Adherence Ratio)
Process yield analogue. Proportion of outputs conformant across all governed dimensions simultaneously. SPAR $= 0.997$ indicates 99.7% joint conformance.

cbmes (Composite Behavioral Measurement Evaluation Score)
Aggregate escalation metric combining ECPI, ECI, and SPAR into a single score calibrated to the tiered escalation model (T0 through T5) within the BAAGF framework.

These constructs are classified as **[I⋆Innovate]** under the AAI taxonomy: the classical constructs they replace (C_p/C_{pk}, control limits, sigma level, DPMO) embed distributional assumptions that LLM behavioral processes routinely violate. Full operational definitions and calculation procedures are developed in Chapter 4.

Remark. The term "behavioral entropy" operates at two levels in this work. In its *specific* sense, it refers to the Shannon entropy of a categorical behavioral distribution (e.g., the entropy of the grounding classification distribution used in the RAG scenario of Chapter 10). In its *general* sense, it refers to the governing conceptual variable: the degree of unpredictability or disorder in a system's behavioral outputs, as operationalized through the BME Metric Suite collectively. The individual BME metrics (ECPI, BAR, ECI, SPAR) each capture a facet of behavioral entropy. Not all CTG characteristics are naturally expressed in terms of Shannon entropy (e.g., the contradiction avoidance rate is a proportion rather than an entropy value). Still, all are interpretable as measures of the system's behavioral predictability relative to governance specifications. Where this monograph references behavioral entropy as a single calculable quantity (as in the RAG scenario's hallucination proxy), the Shannon formulation applies. Where it references behavioral entropy as the governing process variable, the BME Metric Suite is the operational instrument.

1.6 The Adopt, Adapt, Innovate (AAI) Taxonomy

The AAI taxonomy provides a principled classification scheme for every classical Six Sigma tool encountered in this work. It is the analytical lens through which every methodological claim is evaluated.

Definition 1.5 (AAI Classification Rules)**. [A · Adopt]**
Assigned when all classical assumptions hold (or deviations are immaterial). The tool's structural logic, input requirements, computational method, and interpretive framework apply without modification.

[D ▷ Adapt]
Assigned when structural logic applies, but one or more classical assumptions require specified modification. Each modification must be documented with justification and boundary conditions.

[I ⋆ Innovate]
Assigned when classical assumptions fail irrecoverably. A purpose-built construct replaces the classical tool with a theoretical justification for the replacement.

Table 1.1: *AAI Classification Decision Rules*

Classification	Decision Rule
Adopt	All classical assumptions hold, or deviations are immaterial. Method, validity conditions, and interpretive framework apply without modification.
Adapt	Structural logic applies, but one or more assumptions require specified modification. Each modification is documented with justification and boundary conditions.
Innovate	Classical assumptions fail irrecoverably. A purpose-built construct replaces the classical tool, with theoretical justification.

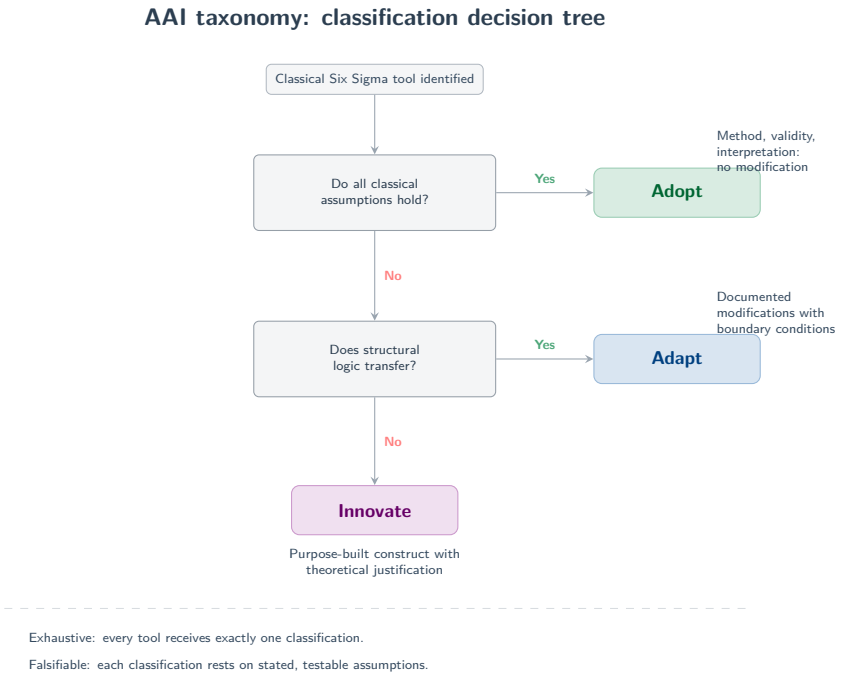


Figure 1.1: AAI taxonomy decision tree: classification proceeds through two decision gates. If all classical assumptions hold, the tool is Adopted without modification. If structural logic transfers but assumptions require specified modification, the tool is Adapted. If classical assumptions fail irrecoverably, a purpose-built construct is Innovated. The classification is exhaustive and falsifiable.

These rules produce a three-tier classification that is exhaustive (every tool receives exactly one classification) and falsifiable (each classification rests on stated assumptions testable against domain data). The rigor prevents two common errors: *uncritical adoption* (applying tools without verifying assumptions) and *premature innovation* (replacing tools that could have been adapted, abandoning validated methodology unnecessarily).

The complete AAI classification for all 40 tools and constructs is developed through Chapters 3–7 as each tool is encountered in an operational context, and consolidated in Chapter 8. With the theoretical foundations, the BME Metric Suite, and the AAI taxonomy established, Chapter 2 introduces the governance architecture within which DMAIC operates.

Chapter 2

The ALAGF Governance Architecture

The theoretical foundations of Chapter 1 established that LLM and agentic AI systems are governable processes amenable to Statistical Process Control. The DMAIC methodology provides the process improvement cycle. But DMAIC does not operate in an organizational vacuum: it requires a governance architecture that defines who has authority, what behavioral standards are enforced, how monitoring is conducted, and how escalation decisions are made. The Adaptive Lifecycle Agentic Governance Framework (ALAGF) provides that architecture.

This chapter specifies the ALAGF meta-framework: its tripartite subsystem structure, the performance loops that connect those subsystems, the escalation model that translates detected signals into proportional governance responses, and the mechanism through which DMAIC cycles are embedded as repeatable processes within the governance lifecycle. Understanding ALAGF is a prerequisite for the DMAIC phase chapters that follow (Chapters 3–7), because every phase operates within ALAGF's authority structure, uses its measurement system, and produces artifacts that ALAGF consumes for lifecycle governance.

2.1 The Governance Tripartite: Conscience, Voice, Memory

ALAGF rests on the premise that effective governance of stochastic AI systems requires the simultaneous operation of three functionally distinct but informationally coupled subsystems. This premise derives from a structural analysis of governance failure modes: organizations that monitor without intervening (memory without voice) detect drift but cannot correct it; organizations that intervene without monitoring (voice without memory) apply corrections without evidence;

and organizations that do both without architectural governance principles (voice and memory without conscience) lack the normative framework to determine what constitutes acceptable behavior.

Definition 2.1 (ALAGF Tripartite Architecture)**.** The ALAGF governance architecture comprises three subsystems:

baagf (Behavioral Assurance and Agentic Governance Framework)
The *architectural conscience*. Defines behavioral specifications, manages the escalation model, owns the BME Metric Suite definitions and calibration authority, and issues binding governance directives.

SymPrompt$^+$
The *operational voice*. Translates governance directives into prompt-level behavioral modulations, maintains the intervention version registry, and executes entropy-calibrated targeting.

midcot (Multi-Dataset IQ Drift and Cost Optimization Training)
The *performance memory*. Ingests behavioral metric data, maintains running control charts, detects signals, and triggers escalation events.

None of these subsystems is sufficient in isolation. Their governance power resides in their integration: MIDCOT detects that something has changed; BAAGF determines whether that change violates governance specifications and what response is warranted; SymPrompt$^+$ executes the warranted response under version control.

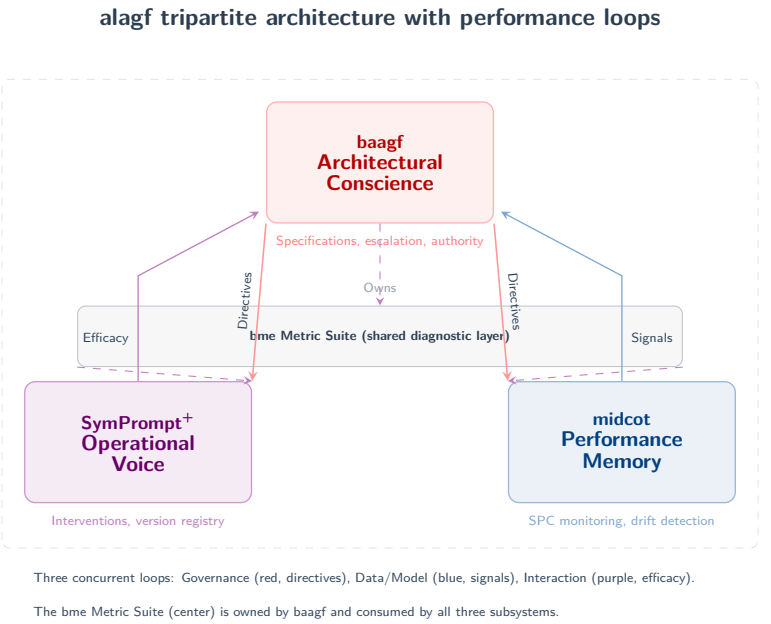


Figure 2.1: ALAGF tripartite architecture: BAAGF (architectural conscience) defines specifications and issues directives; SymPrompt$^+$ (operational voice) executes governed interventions; MIDCOT (performance memory) monitors behavioral metrics and detects drift. The BME Metric Suite operates as the shared diagnostic layer connecting all three subsystems. Red arrows show governance directives; blue/purple arrows show signal and efficacy feedback loops.

2.2 The BME Metric Suite as Shared Diagnostic Layer

The BME Metric Suite operates as the shared diagnostic infrastructure across all three ALAGF subsystems. Architecturally, the BME Metric Suite is housed within BAAGF, reflecting its status as a governance-owned measurement system, with definitions, thresholds, and calibration that are authoritative governance decisions.

This ownership separation prevents operational subsystems from adjusting the measurement system to improve their own metrics: a form of gaming that governance architectures must structurally prevent.

Table 2.1: *BME Metric Suite: subsystem data flows*

Subsystem	BME Data Usage
BAAGF	Owns metric definitions and calibration authority. Uses CBMES for escalation tier determination. Sets BAR thresholds.
MIDCOT	Consumes BME metric streams for SPC monitoring. Applies BAR thresholds to control charts. Detects and forwards signals.
SymPrompt$^+$	Uses BME metric values as targeting coordinates. Measures intervention efficacy through pre- and post-comparisons.

2.3 The Seven-Phase Recursive Governance Lifecycle

BAAGF organizes governance operations through a seven-phase recursive lifecycle. The lifecycle is recursive in two senses: it cycles continuously through all phases during normal operations, and escalation events can trigger re-entry into earlier phases.

Table 2.2: *BAAGF seven-phase governance lifecycle*

Phase	Function	Key Governance Artifacts
1	Data Governance	Data governance policy, corpus version registry, input class registry
2	Model Architecture	Model governance registry, capability boundary documentation
3	Training & Evaluation	Training governance protocol, DAST clearance report
4	Deployment & Integration	Deployment authorization, initial control plan, escalation protocol
5	Monitoring & Escalation	Escalation disposition log, signal investigation records
6	Feedback & Remediation	Updated FMEA register, intervention validation reports
7	Governance Reporting	Periodic governance reports, compliance evidence, audit trail

During stable operations (T0/T1), Phases 5 and 7 dominate. During escalation events (T3+), Phases 6 and possibly 1 through 4 activate as the DMAIC cycle re-examines specifications, measurement validity, and intervention design.

2.4 Performance Loop Topology

ALAGF operates through three concurrent performance loops, each connecting two subsystems through the shared BME diagnostic layer.

2.4.1 Data and Model Loop (MIDCOT to BME)

MIDCOT continuously monitors BME metric streams to detect drift in the behavioral distribution. When metrics shift, MIDCOT's detection layer evaluates whether the shift constitutes a signal (special-cause variation) or noise (common-cause variation within expected bounds). Signals are classified by severity and forwarded to BAAGF through the escalation pathway. This loop operates at the monitoring cadence specified in the governance charter.

2.4.2 Interaction Loop (SymPrompt+ to BME)

When SymPrompt$^+$ applies a modulation in response to a BAAGF directive, the BME Metric Suite measures the resulting behavioral distribution change. This measurement feeds back to SymPrompt$^+$ for efficacy assessment and to BAAGF for governance validation. The loop enforces the principle that no intervention is assumed effective until empirically confirmed.

2.4.3 Governance Loop (BAAGF Directives)

The authoritative control pathway through which BAAGF exercises governance authority. BAAGF issues two classes of binding directives:

Directives to midcot: rebaseline commands, threshold adjustments, monitoring cadence changes, and control chart reconfiguration.

Directives to SymPrompt$^+$: intervention orders (specifying target CTG characteristic, root cause, and permitted scope), modulation suspension orders, and scope restriction orders.

All directives are logged in the BAAGF governance registry with timestamps, authorization references, and expected behavioral effects. This registry constitutes the authoritative governance audit trail.

2.5 The CBMES Escalation Model

The CBMES escalation model translates composite behavioral scores into graduated governance responses. Table 2.3 defines six tiers (T0 through T5), each with a

CBMES score range, a classification label, and a mandatory governance response. The model ensures that the severity of the governance response is proportional to the magnitude of the behavioral deviation: minor drift triggers enhanced monitoring. In contrast, critical failures trigger automatic system suspension and full DMAIC re-initiation.

Table 2.3: *CBMES escalation tier model*

Tier	CBMES	Classification	Governance Response
T0	0.00–0.10	Optimal	Standard monitoring. No intervention required.
T1	0.11–0.25	Stable	Standard monitoring. Trend analysis. Advisory notation.
T2	0.26–0.50	Moderate Drift	Enhanced monitoring. Root cause investigation initiated.
T3	0.51–0.75	High Risk	Immediate review. SymPrompt^{+} intervention authorized. The system may be restricted.
T4	0.76–0.90	Critical	System suspended for affected classes. Full DMAIC Analyze-Improve. Committee convened.
T5	>0.90	Emergency	Automatic suspension. Emergency session. Full DMAIC from Define.

The model ensures proportional response. T0–T2 trigger monitoring adjustments; T3–T5 trigger operational restrictions and structured improvement cycles. Threshold values are governance decisions calibrated to the deployment risk profile and documented in the governance charter.

Remark. Threshold calibration requires: risk classification of the deployment context (high-risk per EU AI Act, safety-critical, enterprise, general-purpose); mapping from risk classification to threshold stringency; and a documented rationale that can withstand audit scrutiny. Response timeframes are also calibrated: T0/T1 per scheduled review cadence; T2 within 48- 72 hours; T3 within 24 hours; T4 immediate review with suspension within 4 hours; T5 instantaneous suspension with committee within 2 hours.

2.6 DMAIC as a Repeatable Process Within ALAGF

The relationship between DMAIC and ALAGF is not one of sequence but of embedding.

Procedure: alagf Initialization via dmaic

Step 1: Define. Governance charter, VOG analysis, CTG trees, behavioral specification envelope, input class registry, stakeholder RACI matrix.

Step 2: Measure. MSA/GRR validation, behavioral baseline, initial ECPI/ECI/SPAR/CBMES calculation.

Step 3: Analyze. Pre-existing capability gap diagnosis between baseline performance and specifications.

Step 4: Improve. SymPrompt^{+} modulation design, DOE validation, intervention deployment with rollback capacity.

Step 5: Control. MIDCOT monitoring regime with BAR thresholds, control chart architecture, escalation protocol, re-baselining triggers.

Step 6: DAST Clearance. Pre-deployment adversarial stress testing validates behavioral robustness before operational activation.

Subsequent DMAIC cycles are triggered by escalation events at T3 or above. The scope is defined by the escalation context: a targeted Analyze-Improve cycle for a single-dimension violation, or a full five-phase cycle for a systemic CBMES breach. Upon completion, the control plan is updated and MIDCOT transitions to a re-baselining window.

2.7 Re-Baselining Protocol

The re-baselining protocol governs how MIDCOT's monitoring parameters are updated following authorized process changes. Re-baselining is the governance mechanism that distinguishes legitimate non-stationarity (which updates baselines) from unauthorized drift (which triggers escalation). Without this protocol, every authorized change would appear as a drift signal, desensitizing the monitoring system to genuine degradation.

Procedure: Re-Baselining Protocol

Step 1: Trigger identification. Authorized change verified against the MIDCOT change management registry.

Step 2: Validation mode activation. MIDCOT enters enhanced sensitivity (80% of operational BAR values) for the re-baselining window.

Step 3: Post-change data collection. Behavioral metrics collected across all governed input classes at enhanced cadence.

Step 4: Metric recalculation. New ECPI, ECI, SPAR, CBMES values calculated. New BAR thresholds derived from empirical quantiles.

Step 5: Governance approval. BAAGF reviews results. If specifications are satisfied, updated thresholds are activated. If not, a targeted Improve cycle is initiated.

Step 6: Standard monitoring resumption. MIDCOT transitions from validation mode to standard monitoring with updated parameters.

2.8 DAST Pre-Deployment Clearance

Dynamic Adversarial Stress Testing (DAST) provides pre-deployment clearance gates. DAST evaluates behavioral response to adversarial inputs across all governed CTG characteristics, producing a pass/fail decision based on CBMES under stress conditions. DAST operates before initial deployment and before reactivation following $T4/T5$ suspension.

2.9 Structural Failure Modes

The ALAGF architecture has four characteristic failure modes, each representing a way in which the tripartite integration can break down. These failures are architectural rather than operational: they arise from structural disconnections between subsystems rather than from within any single subsystem's function.

Subsystem coupling failure
MIDCOT detects drift but BAAGF does not receive the signal, or BAAGF issues a directive that SymPrompt$^+$ does not execute. Remedy: a governance registry audit verifying that every escalation has a disposition and every directive has an execution record.

Monitoring atrophy
Alert fatigue desensitizes the governance team. MIDCOT's automated surveillance mitigates this, but T3+ responses require human engagement. Charter must specify minimum response requirements with audit verification.

Threshold ossification
BAR thresholds set at deployment and never revised. Scheduled reviews must explicitly assess the adequacy of thresholds.

Change management disconnection
System changes occur without triggering re-baselining. Integration between MIDCOT and the change registry prevents this.

2.10 Standards Traceability

Table 2.4 maps ALAGF's subsystems and governance operations to specific clauses and functions of the four primary international AI governance standards. This mapping demonstrates that ALAGF is not a standalone governance proposal but a process-level implementation architecture that satisfies existing standards requirements. Chapter 9 develops these alignments in full operational detail with jurisdiction-specific regulatory extensions.

Table 2.4: *ALAGF standards alignment summary*

Standard	alagf Implementation
ISO/IEC 42001 Cl. 4–5	BAAGF governance charter, authority structure, leadership commitment
ISO/IEC 42001 Cl. 6, 8	CTG trees, specification envelopes, FMEA, operational planning
ISO/IEC 42001 Cl. 9	MIDCOT continuous monitoring, BME measurement, governance registry audit
ISO/IEC 42001 Cl. 10	SymPrompt$^+$ interventions, DOE optimization, DMAIC cyclical improvement
NIST AI RMF: GOVERN	BAAGF charter, authority, escalation model
NIST AI RMF: MAP	CTG trees, input class registry, specification envelope
NIST AI RMF: MEASURE	BME Metric Suite, MSA validation, ECPI/SPAR/CBMES
NIST AI RMF: MANAGE	MIDCOT monitoring, T0–T5 escalation, SymPrompt$^+$ interventions
EU AI Act Art. 9	BAAGF risk management, FMEA, DAST clearance
EU AI Act Art. 72	MIDCOT post-market monitoring, automated escalation
EU AI Act Art. 61–62	T5 automatic suspension, post-incident review
IEEE P2863 / 7010	Decision authority structure; wellbeing-relevant CTG dimensions

Example 2.1: alagf Initialization for a CDSS

A healthcare organization deploys an LLM-based clinical decision support system under ALAGF governance. BAAGF establishes a governance charter with quarterly reviews and real-time escalation triggers. The VOG analysis aggregates requirements from four stakeholder classes: clinicians (evidence-based recommendations with uncertainty quantification), patients (no deviation from standards of care), the regulatory body (conformity with the EU AI Act for high-risk applications), and the deploying organization (operational efficiency without increased liability). The CTG tree produces four measurable characteristics: citation accuracy, guideline concordance, contradiction avoidance, and uncertainty disclosure. MIDCOT collects a four-week baseline: SPAR = 0.943, ECPI for uncertainty disclosure = 0.92 (incapable), ECI = 2.8. The gap between current SPAR and target SPAR (0.997) defines the Analyze and Improve scope developed in Chapters 5 and 6.

With the ALAGF governance architecture established, the next five chapters develop the DMAIC phases in full operational detail. Each phase operates within ALAGF's authority structure, produces governance artifacts consumed by BAAGF, and executes monitoring through MIDCOT and interventions through SymPrompt$^+$.

Chapter 3

Define Phase: Establishing the Governance Charter

In classical Six Sigma, the Define phase establishes the scope, objectives, and stakeholder alignment for a process improvement initiative. Its primary artifacts are the project charter, the Voice of the Customer (VOC) analysis, and the Critical-to-Quality (CTQ) tree that translates customer needs into measurable process specifications [35, 28]. The Define phase answers three foundational questions: What process is being governed? What constitutes acceptable performance? Who are the stakeholders whose requirements define acceptability?

The Define phase is often treated as administrative rather than analytical. This is a mistake in any domain, and a critical one in AI governance. The difficulty of defining behavioral specifications for stochastic systems is the single largest source of governance failure in AI deployment. Systems are deployed with vague objectives ("be helpful and harmless"), unmeasurable constraints ("don't hallucinate"), and stakeholder alignment that extends no further than the development team's intuitions. The Define phase, properly executed, prevents these failures by requiring explicit, measurable, stakeholder-validated behavioral specifications before any measurement or intervention activity begins.

This chapter presents the Define phase protocol, specifies each tool and construct with its AAI classification, and demonstrates full application through three parallel worked examples: a clinical decision support system (CDSS), an insurance claims triage system (ICT), and an emergency department triage agent (EDT). Each example produces the complete set of Define phase governance artifacts with populated data, demonstrating both the methodology's generalizability across domains and the specificity required for operational governance. Complete use-case data profiles for the ICT and EDT scenarios are provided in Appendix G and Appendix H, respectively.

3.1 Translation Logic: Define for Stochastic AI Systems

The translation to AI governance requires three structural modifications. First, the Voice of the Customer becomes the Voice of Governance (VOG), expanding the stakeholder constituency from end users to the full governance set: end users, deploying organizations, regulatory bodies, affected populations, and the public interest as instantiated in applicable standards. Second, Critical-to-Quality (CTQ) becomes Critical-to-Governance (CTG), retargeting the hierarchical decomposition from physical product characteristics to behavioral dimensions with correlated, multidimensional specification targets. Third, the project charter becomes a governance charter: a lifecycle governance commitment with continuous authority, adaptive specifications, and real-time escalation protocols.

These modifications reflect the structural reality that the logic of governance initialization transfers from manufacturing to AI, but the specific instruments require adaptation or replacement to accommodate stochastic, non-stationary AI behavior.

3.2 The Define Phase Protocol

The Define phase protocol (PM-001) proceeds through eight steps, from governance trigger assessment through phase completion and handoff to the Measure phase. The protocol enforces a strict sequencing discipline: each step's output serves as a required input for subsequent steps, and the phase completion gate (Step 8) validates cross-artifact traceability before authorizing the initiation of PM-002. The eight steps below constitute the minimum viable Define phase; governance teams may extend but may not abbreviate the protocol.

Procedure: Define Phase Protocol (PM-001)

Step 1: Governance Trigger Assessment. Identify the initiating condition: (a) new system deployment, (b) scheduled specification review, or (c) escalation-initiated DMAIC cycle at T3 or above. For escalation triggers, retrieve CBMES value, escalation tier, and affected CTG characteristics from MIDCOT. Record in the governance audit trail.

Step 2: Governance Charter Development. Establish the lifecycle governance commitment: system scope, governance authority structure, escalation protocol (referencing BAAGF T0–T5 tiers), specification review cadence, measurement cadence, and conditions for suspension or decommissioning. Submit for governance authority approval.

Step 3: Voice of Governance Analysis. Aggregate governance requirements from all four canonical stakeholder classes (end users, deploying organization, regulatory bodies, affected populations). Identify inter-class conflicts. Resolve

or prioritize every conflict with documented disposition. Produce the prioritized VOG requirement set.

Step 4: CTG Tree Construction. Decompose each prioritized governance requirement into measurable behavioral characteristics. Each leaf-level characteristic must satisfy four completeness criteria: measurable via a defined BME instrument, target value specified, specification limit defined, and measurement method documented. Set specification limits based on regulatory requirements, stakeholder negotiations, and empirical baselines.

Step 5: Behavioral Specification Envelope. Combine all CTG characteristics into the multidimensional conformance region. Define the joint conformance requirement: a system output is conformant only if it satisfies all specification limits simultaneously. Set the SPAR target (default: ≥ 0.997).

Step 6: Input Class Registry. Define the input classes for which specifications apply. Specify classification dimensions, class-specific specification limits, boundary conditions, out-of-scope handling rules, and the input classification method.

Step 7: Stakeholder RACI Matrix. Assign governance decision authority across the DMAIC lifecycle: specification changes, escalation responses at each tier, system suspension, re-baselining approval, SymPrompt^{+} modulation authorization.

Step 8: Phase Completion and Handoff. Validate completeness and cross-artifact consistency. Every CTG characteristic must trace to a VOG requirement. Every specification limit must trace to a governance rationale. Every escalation action must have an RACI assignment. Authorize PM-002 (Measure phase) initiation.

3.3 Define Phase Tool Specifications

This section specifies each tool and construct employed in the Define phase, following the AAI classification framework (Definition 1.5). Two tools are adopted without modification, three are adapted with defined extensions, and three are purpose-built innovations. For each tool, the specification is followed by fully populated worked examples from all three governance scenarios.

3.3.1 Adopted Tools

Two Define-phase tools transfer directly from classical Six Sigma without structural modification: SIPOC for system boundary and component mapping, and the RACI matrix for stakeholder responsibility assignment. Both operate on domain-independent structural logic (component enumeration and responsibility assignment, respectively) that applies to AI governance decisions without adaptation.

SIPOC for AI Governance [A · Adopt]

Module: TM-T1-01

Classical Definition

SIPOC (Supplier-Input-Process-Output-Customer) maps the high-level components of a process: who supplies inputs, what inputs are consumed, what transformation the process performs, what outputs are produced, and who receives them. It establishes the process boundary and identifies the key interfaces that governance must address.

Why Adopt

The structural logic of component mapping applies directly to AI system architectures. An LLM system has identifiable suppliers (training data providers, retrieval corpus sources, prompt template authors), inputs (user queries, system instructions, retrieved context), a transformation process (model inference), outputs (generated text, tool calls, decisions), and customers (end users, downstream systems, governance stakeholders). No modification to the SIPOC method is required; only the content of each category changes.

Operational Protocol

Identify the AI system boundary. Enumerate suppliers by category (data, model, configuration, infrastructure). Map inputs by type and source. Define the process as the inference pipeline from input reception through output delivery. Classify outputs by type and destination. Identify all customer classes, including governance stakeholders. The completed SIPOC provides the scope foundation for developing the governance charter.

Example 3.1: CDSS SIPOC

Table 3.1: *CDSS SIPOC map*

Element	Specification
Suppliers	Foundation model provider; Clinical guideline publishers; EHR vendor; Prompt governance team
Inputs	Clinician queries; Patient EHR context; Retrieved clinical guidelines; SymPrompt^{+} instructions (SP-CDSS-v2.3)
Process	RAG-augmented LLM inference: query interpretation $\rightarrow$ guideline retrieval $\rightarrow$ evidence synthesis $\rightarrow$ recommendation generation
Outputs	Clinical recommendations with citations; Uncertainty classifications (high/moderate/low); Guideline concordance flags
Customers	Clinicians (operational users); Patients (affected population); Clinical AI Governance Committee; EU AI Act conformity body

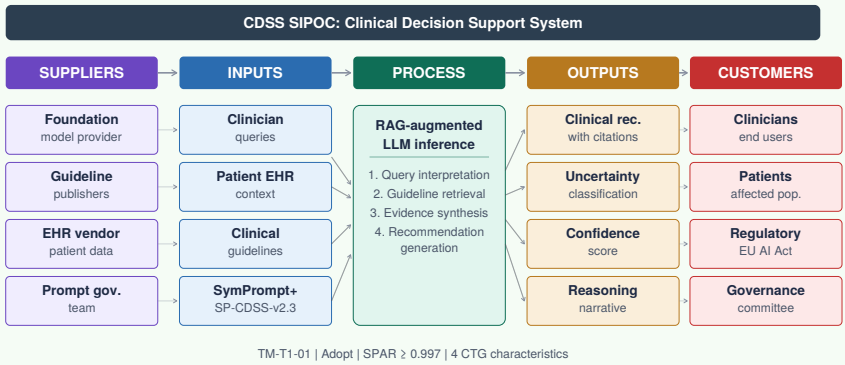


Figure 3.1: CDSS SIPOC diagram: suppliers, inputs, RAG-augmented inference process, outputs, and governance customers.

Example 3.2: Insurance Claims Triage SIPOC

Table 3.2: *ICT SIPOC map*

Element	Specification
Suppliers	Foundation model provider (API); Policy wording team; Underwriting (endorsements); Fraud intelligence unit; Prompt governance team
Inputs	FNOL submissions (PDF/digital); Policy wording documents (7,150); Endorsements and coverage guides; Fraud indicator reference sheets; SymPrompt^{+} instructions (SP-CT-v2.3.1)
Process	4-step agentic RAG workflow: (1) Document intake → (2) Coverage assessment via policy retrieval → (3) Fraud screening → (4) Pathway recommendation
Outputs	Structured JSON: coverage determination; Fraud risk score (low/medium/high); Pathway recommendation (fast-track/standard/specialist/SIU); Confidence score; Reasoning narrative
Customers	Claims handlers (42 motor, 28 property, 14 liability); Senior claims managers (8); Fraud investigation unit (6); Policyholders (affected population); FSCA/FAIS Ombud; Claims AI Governance Committee

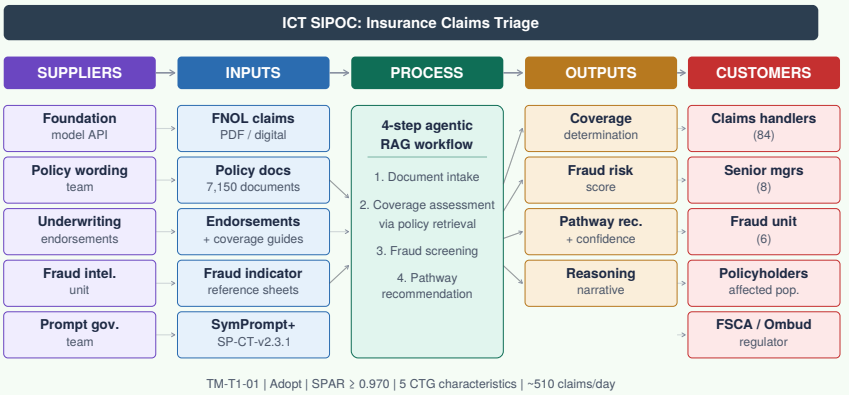


Figure 3.2: ICT SIPOC diagram: five supplier streams feeding the 4-step agentic RAG workflow, with six customer classes including FSCA regulatory oversight.

Example 3.3: ED Triage Agent SIPOC

Table 3.3: *EDT SIPOC map*

Element	Specification
Suppliers	Foundation model provider (fine-tuned 8B parameter); Clinical guideline publishers (ACEP, ACS); EHR vendor (Epic); Vitals feed (Philips IntelliVue); Prompt governance team
Inputs	Free-text triage notes; EHR patient history (FHIR R4); Real-time vitals stream; Chief complaint; Clinical protocols (4,280 documents); SymPrompt^{+} instructions (SP-ED-v3.1.2)
Process	4-step sequential pipeline: (1) Intake Parser → (2) Acuity Classifier (ESI 1–5 + screening) → (3) Resource Allocator → (4) Disposition Advisor (admit/observe/discharge)
Outputs	ESI level; Sepsis/stroke/STEMI screening alerts; Treatment area recommendation; Resource bundle; Disposition recommendation with reasoning chain; Uncertainty flag (confidence <0.80)
Customers	Triage nurses (112); ED physicians (38); Hospital administration; Patients (incl. pediatric, geriatric, non-English-speaking); CMS/Joint Commission; Clinical AI Committee

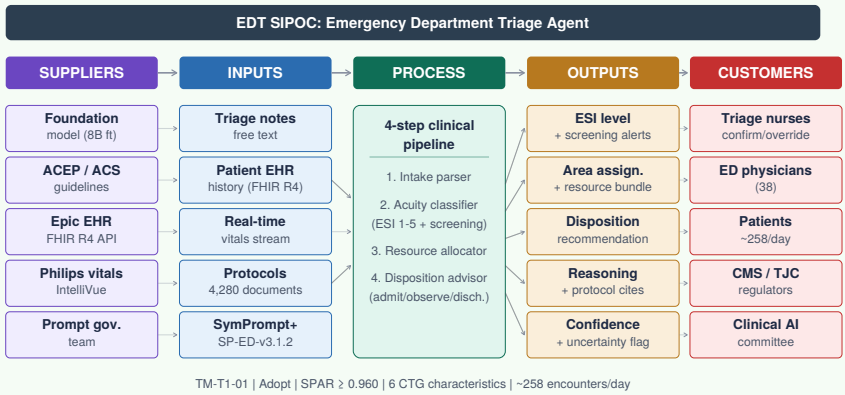


Figure 3.3: EDT SIPOC diagram: five supplier streams (including real-time vitals) feeding the 4-step clinical pipeline, with five customer classes under dual EMTALA and EU AI Act regulatory oversight.

Cross-scenario observation: All three systems share the RAG-augmented architecture pattern: a retrieval corpus mediates between the query and the foundation model. The SIPOC makes this visible (Figures 3.1–3.3): in all three cases, the "Suppliers" column includes a function for retrieval corpus maintenance. The drift events documented in each use-case data profile (Appendix G and Appendix H) trace to retrieval corpus currency failures, confirming that the SIPOC correctly identified this as a critical governance interface.

Stakeholder RACI Matrix [A · Adopt]

Module: TM-T1-02

Classical Definition

The RACI matrix assigns four responsibility levels to stakeholders for each decision category: Responsible (performs the work), Accountable (has final decision authority), Consulted (provides input before the decision), and Informed (notified after the decision). Every decision must have exactly one Accountable party.

Why Adopt

Stakeholder responsibility assignment is domain-independent. The RACI framework applies to AI governance decisions (specification changes, escalation responses, system suspension) without structural modification. The matrix ensures that governance authority is explicit and unambiguous, preventing the "diffusion of responsibility" failure mode where no individual has clear authority to act during escalation events.

Operational Protocol

Enumerate all governance decision categories (specification changes, escalation responses at each BAAGF tier, system suspension, re-baselining approval, SymPrompt$^+$ modulation authorization, scheduled review conduct, audit trail review). For each category, assign R, A, C, and I roles. Validate against the governance charter: every escalation action must have a corresponding RACI assignment.

Example 3.4: CDSS RACI Matrix (Excerpt)

Table 3.4: *CDSS governance RACI matrix*

Decision Category	Gov. Analyst	CMIO	Clin. Panel	Legal	IT Ops
Specification change	R	A	C	C	I
T2 escalation response	R	A	C	I	I
T4 system suspension	I	A	C	C	R
T5 emergency halt	I	A	C	C	R
Re-baselining approval	R	A	C	I	I
SymPrompt$^+$ modulation	R	A	C	I	I

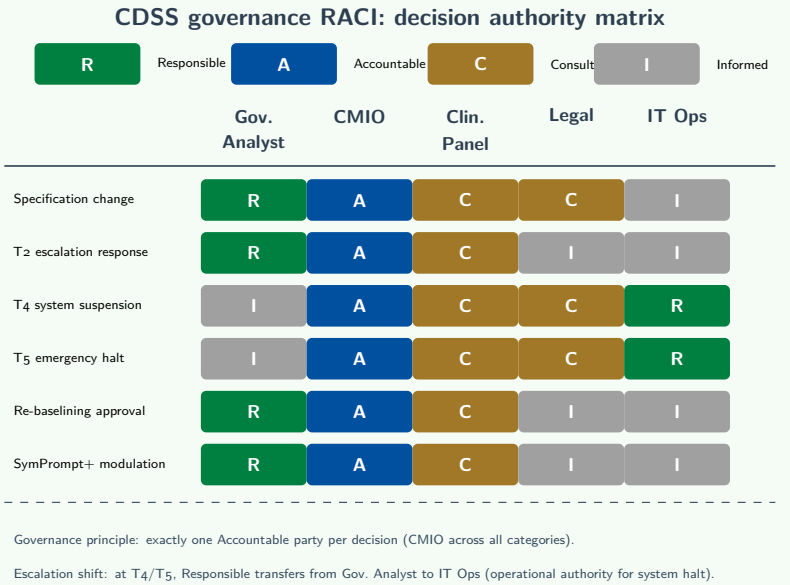


Figure 3.4: CDSS RACI diagram: color-coded responsibility assignments across six governance decision categories, showing the accountability chain (CMIO) and the responsibility shift from governance analyst to IT operations at critical/emergency tiers.

Example 3.5: Insurance Claims Triage RACI Matrix

Table 3.5: *ICT governance RACI matrix*

Decision	Op. Gov.	Claims Ops	CRO	Compl.	IT	Board
Spec change	R	C	A	C	I	I
Advisory investigation	R	I	I	I	C	I
Warning: restrict	R	R	A	C	R	I
Critical: suspend line	C	C	A	C	R	I
Emergency: full halt	I	C	A	R	R	I
Re-baselining	R	C	A	I	C	I
SymPrompt^{+} modulation	R	C	A	C	I	I
Regulatory notification	I	I	A	R	I	I

Example 3.6: ED Triage Agent RACI Matrix

Table 3.6: *EDT governance RACI matrix*

Decision	Op. Mon.	ED Med Dir	CMO	Pt. Safety	CIO	Counsel
CTG spec change	C	R	A	C	I	I
Advisory investigation	R	I	I	I	C	I
Warning: double-review	R	R	A	C	I	I
Critical: suspend class	C	R	A	R	R	C
Emergency: full halt	I	C	A	R	R	R
Re-baselining	R	R	A	C	C	I
SymPrompt^{+} modulation	R	R	A	C	I	I
CMS/TJC notification	I	C	A	R	I	R

3.3.2 Adapted Tools

Four Define-phase tools require defined adaptations for AI governance. The Voice of Governance analysis expands stakeholder scope from end users to the full governance constituency. The CTG tree retargets hierarchical decomposition from physical product attributes to behavioral dimensions. The governance charter extends from a time-bounded project instrument to a lifecycle governance commitment. The governance value stream map adds a closed-loop monitoring flow and a corpus supply chain to the classical material and information flows. In each case, the structural logic of the classical tool is preserved, but its content, scope, or

assumptions are modified to accommodate stochastic AI systems.

Voice of Governance (VOG) Analysis **[D ▷ Adapt]**

Module: TM-T1-03

Classical Definition

Voice of the Customer (VOC) captures end-user requirements through interviews, surveys, complaint analysis, and focus groups, translating them into process specifications.

What Is Preserved

The structured elicitation methodology: systematically gathering requirements from stakeholders, documenting them with priority and measurability assessments, and resolving conflicts before proceeding to specification.

What Is Modified

The stakeholder scope expands from end users to the full governance constituency. Four canonical stakeholder classes are required: end users, deploying organization, regulatory bodies, and affected populations. This expansion introduces *inter-class conflicts* absent from single-class VOC: regulatory explainability requirements may conflict with user preferences for conversational naturalness; organizational efficiency objectives may conflict with fairness requirements of affected populations. The VOG protocol requires that every identified conflict receive a documented disposition (resolution, prioritization, or conditional prioritization) before the Define phase proceeds.

Operational Protocol

Identify all four stakeholder classes. Elicit requirements per class through appropriate methods (user interviews, policy review, statutory analysis, impact assessment). Document each requirement with source class, priority, and measurability status. Conduct systematic conflict identification across all class pairs. Resolve or prioritize every conflict with a documented rationale. Compile the prioritized VOG requirement set.

Example 3.7: CDSS VOG Analysis

Four stakeholder classes identified. **Clinicians:** evidence-based recommendations, current with clinical guidelines, presented with uncertainty quantification. **Patients (affected population):** no contradiction of established standards of care; limitations disclosed. **Regulatory body (EU AI Act):** conformity documentation, post-market monitoring, human oversight. **Deploying organization:** operational efficiency without increased liability. **Conflict C-1.** Clinician preference for concise, confident recommendations conflicts with regulatory and patient requirements for uncertainty disclosure. *Disposition:* uncertainty disclosure is a primary governance requirement (reg-

ulatory and patient-safety priority); conciseness is a secondary optimization target. Resolution authority: Clinical AI Governance Committee.

Example 3.8: ED Triage Agent VOG Analysis

Table 3.7: *EDT Voice of Governance analysis*

Stakeholder	Auth. Wt.	Governance Requirement	Conflict
Patients (affected)	Highest	Timely, accurate triage; no discrimination; appropriate acuity	—
ED clinicians (users)	High	Reliable recommendations; clear reasoning; appropriate escalation	C-1
Regulators (CMS, TJC)	Mandatory	Patient safety; EMTALA compliance; QI documentation	—
Hospital admin.	Moderate	Compliance; reduced LWBS rate; efficient resource utilization	C-1
Accrediting bodies	Moderate	Proactive risk management evidence; FMEA documentation	—

Conflict C-1 Resolution. The administration's objective of resource efficiency conflicts with the clinician's requirement for appropriate escalation. *Disposition:* EMTALA creates an absolute floor: the AI cannot under-triage to manage capacity. Under-triage is severity 10 in FMEA. The under-triage rate CTG (target <0.020) takes precedence over resource efficiency. Resolution authority: Clinical AI Committee with ED Medical Director as lead.

CTG Tree Construction **[D ▷ Adapt]**

Module: TM-T1-04

Classical Definition

The CTQ tree hierarchically decomposes customer needs into measurable quality characteristics: Customer Need → Quality Driver → Quality Characteristic → Specification Limit.

What Is Preserved

The hierarchical decomposition structure requires that every leaf-level characteristic be measurable with a defined specification limit.

What Is Modified

Two adaptations. First, decomposition targets behavioral dimensions rather than physical product attributes. Second, specification limits cannot be derived from engineering tolerances; they must be constructed from governance

policy, regulatory requirements, and empirical baselines. Additionally, CTG trees must explicitly represent *dependencies between correlated characteristics*: optimizing factual accuracy may reduce conversational fluency. Primary governance requirements are enforced with specification limits; secondary optimization targets are given preferred ranges.

Operational Protocol

For each VOG requirement, initiate hierarchical decomposition: Governance Need → Governance Driver → CTG Characteristic. Each leaf node must satisfy four completeness criteria: (a) measurable via a defined BME instrument, (b) target value specified, (c) specification limit defined, (d) measurement method documented. Set specification limits from regulatory requirements, stakeholder-negotiated targets, and empirical baselines.

Example 3.9: CDSS CTG Tree

Table 3.8: *CDSS CTG tree decomposition*

Gov. Need	Driver	CTG Characteristic	Target	LSL
No harmful advice	Evidence-based	Citation accuracy rate	≥0.95	≥0.90
No harmful advice	Current guidelines	Guideline concordance rate	≥0.95	≥0.90
No harmful advice	Consistent	Contradiction avoidance rate	≥0.98	≥0.95
No harmful advice	Transparent	Uncertainty disclosure freq.	≥0.90	≥0.85

SPAR target: ≥ 0.997 (joint conformance across all four dimensions).

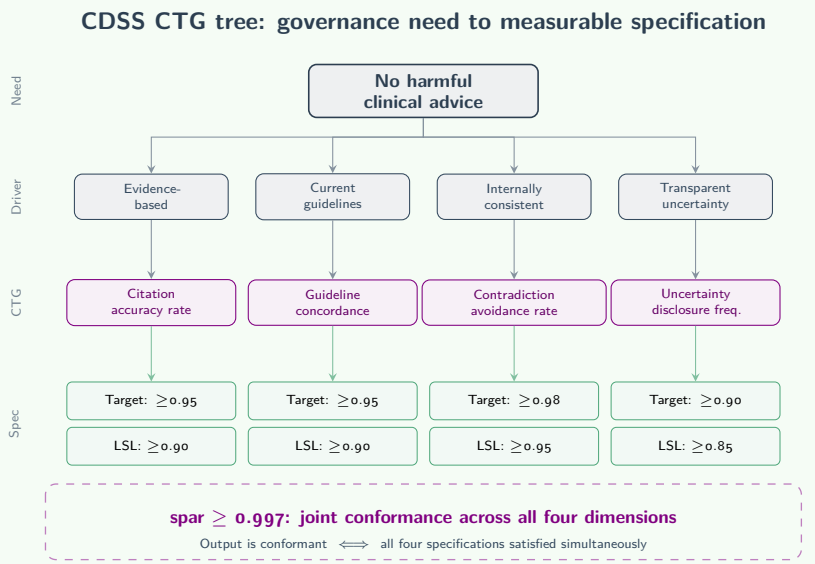


Figure 3.5: CDSS CTG tree: hierarchical decomposition from governance need through four drivers to measurable CTG characteristics with specification limits, converging on the SPAR joint conformance requirement.

Example 3.10: ED Triage Agent CTG Tree

Table 3.9: *EDT CTG tree decomposition*

Gov. Need	Driver	CTG Char.	Target	LSL	USL	Measurement Method
Pt. safety	Correct acuity	Acuity accuracy	0.940	0.900	—	3 board-cert. EP consensus
Pt. safety	No under-triage	Under-triage rate	<0.020	—	0.035	Retrospective attending review
Pt. safety	Screening catch	Screening sensitivity	0.970	0.950	—	Protocol match
Clin. trust	Valid reasoning	Reasoning trans-parency	0.950	0.900	—	Clinician panel evaluation
No discrim.	Equitable triage	Demographic fairness	<0.025	—	0.040	Segment analysis (5 dims.)
Alignment	Appropriate AI	Override concordance	<0.080	—	0.120	Override log analysis

SPAR target: ≥ 0.960 (workflow-level). Six CTG characteristics reflect the dual regulatory burden (EMTALA + EU AI Act). Per-step SPAR must exceed 0.990 for each pipeline module to maintain the workflow floor.
Cross-scenario analysis: The CDSS (4 characteristics, SPAR ≥ 0.997) has the tightest target but fewer dimensions. The ICT (5, ≥ 0.970) adds fairness under TCF. The EDT (6, ≥ 0.960) has the broadest set but the lowest target: six-dimensional joint conformance at 0.997 would be operationally unrealistic. The SPAR target is a governance decision balancing aspirational quality with achievability.

Governance Charter **[D ▷ Adapt]**

Module: TM-T1-05

Classical Definition

The Six Sigma project charter scopes a bounded improvement initiative with defined start/end dates, objectives, team composition, and expected deliverables.

What Is Preserved

The discipline of explicit scope definition, stakeholder alignment, and documented authority before work begins.

What Is Modified

The charter is extended from a time-bounded project instrument to a lifecycle governance commitment. It specifies: the AI system under continuous governance; the governance authority with power to suspend, retrain, or decommission; the BAAGF escalation protocol with tier-specific response actions and timeframes; the specification review cadence; the measurement cadence for BME metric collection; and suspension/decommissioning conditions. The governance charter is a living document revised through scheduled reviews.

Operational Protocol

Define governance scope. Establish governance authority. Specify escalation protocol (T_0–T_5). Define review cadence. Specify measurement cadence. Document lifecycle commitments. Submit for governance authority approval.

AI Governance Value Stream Map **[D ▷ Adapt]**

Module: TM-T1-09

Classical Definition

Value Stream Mapping (VSM) traces the complete flow of materials and information required to deliver a product or service, from the supplier through

the transformation process to the customer. It identifies value-adding and non-value-adding activities, handoff points, wait times, and information flows, producing a visual representation that reveals systemic inefficiencies invisible in activity-level analysis [35].

What Is Preserved

The end-to-end flow visualization methodology: tracing every transformation, handoff, and information dependency from input to output to governance response. The distinction between value-adding activities (those that directly contribute to behavioral conformance) and non-value-adding activities (delays, redundant evaluations, information gaps). The identification of critical handoff points where information quality degrades or governance authority is ambiguous.

What Is Modified

Three adaptations. First, the "material flow" becomes the *data-to-behavior flow*: the stream traces how input data enters the AI system, how it is transformed through the inference pipeline (retrieval, generation, post-processing), how the behavioral output reaches the end user, and how governance monitoring observes and responds to that output. Second, the "information flow" becomes the *governance information flow*: the stream traces how BME metric data flows from the measurement system through MIDCOT detection, BAAGF escalation, and SymPrompt$^+$ intervention back to the behavioral output. This creates a closed-loop governance value stream that classical manufacturing VSM does not require, because manufacturing processes do not dynamically modulate their own behavior in response to quality signals. Third, the VSM must represent the *retrieval corpus supply chain* as a parallel value stream feeding the inference pipeline: the drift events documented in all three scenarios (Chapters 7, Appendix G, Appendix H) trace to failures in this supply chain, confirming that it is a governance-critical pathway.

AI Governance VSM Components

The governance value stream comprises three concurrent flows:
Behavioral flow (left to right): Input query → retrieval corpus lookup → context assembly → LLM inference → post-processing/guardrails → behavioral output → end user. Each step is annotated with: processing time, the configuration version (SymPrompt$^+$ ID, corpus version, model version), and the CTG dimensions affected.
Governance monitoring flow (bottom): Behavioral output → evaluator/automated scoring → BME metric calculation → MIDCOT SPC charts → signal detection → BAAGF escalation → governance disposition. Annotated with: measurement cadence, detection latency, escalation response time, and authority at each decision point.
Corpus supply chain (top): External source (guideline publisher, policy underwriter, knowledge owner) → publication/update event → propagation to

retrieval corpus → version registration in MIDCOT change registry. Annotated with: update frequency, propagation delay, and currency validation mechanism (or its absence).

Governance-Specific VSM Metrics

Four metrics annotate each step in the governance value stream, extending the classical VSM metrics of cycle time and wait time:

Behavioral latency
Time from input receipt to behavioral output delivery. Governance constraint: must not exceed the operational SLA.

Monitoring latency
Time from behavioral output to governance signal detection. The interval during which nonconforming outputs may reach end users undetected. The EDT drift event (Section H.9) demonstrated 18-day monitoring latency for gradual sepsis screening drift.

Response latency
Time from signal detection to governance disposition (investigation initiated, restriction applied, or suspension executed). Governed by the escalation protocol, tier-specific timeframes.

Propagation delay
Time from external source update to retrieval corpus integration. The ICT drift event (Section G.9) demonstrated a 16-day propagation delay for motor endorsement updates. The CAPA targets reducing this to < 48 hours.

Operational Protocol

Map the behavioral flow for the governed AI system from input to output. Map the governance monitoring flow from output through MIDCOT to BAAGF disposition. Map the corpus supply chain from each external source through propagation to version registration. Annotate each step with the four VSM metrics. Identify non-value-adding steps (redundant evaluations, information gaps, authority ambiguities, manual handoffs that could be automated). Identify the critical path for monitoring latency (the longest delay between nonconforming output and governance detection). Document the governance value stream map as a Define-phase artifact feeding the Control-phase monitoring architecture design.

Example 3.11: CDSS Governance Value Stream Map

Behavioral flow: Clinical query → retrieval lookup (clinical guidelines corpus, ∼0.3s) → context assembly with patient history (∼0.2s) → LLM

inference (SP-CDSS-v2.3, ∼1.8s) → uncertainty classification post-processing (∼0.1s) → clinical recommendation to clinician. Total behavioral latency: ∼2.4s.

Governance monitoring flow: Clinical recommendation → automated scoring on citation accuracy and contradiction avoidance (real-time) → expert panel evaluation on uncertainty disclosure and guideline concordance (daily batch, 200 samples) → BME metric calculation (daily) → MIDCOT SPC chart update (daily) → signal detection → BAAGF escalation. Monitoring latency for gradual drift: ∼14 days (CUSUM accumulation period). Response latency: Advisory 72 hours, Warning 24 hours, Critical 4 hours.

Corpus supply chain: Clinical guideline publishers (NICE, WHO, specialty societies) → publication event → *manual review by clinical team* → corpus update → version registration. **Critical finding:** propagation delay from publication to corpus update was uncontrolled (no SLA, no automated monitoring). The Month 3 drift event confirmed this as the governance-critical gap. CAPA: automated publication monitoring pipeline.

Non-value-adding steps identified: Manual corpus update process (no automated trigger from publisher feeds); expert panel evaluation backlog during holiday periods (evaluation latency spikes from 1 day to 4 days); redundant scoring of high-confidence outputs that consistently score above BAR upper threshold.

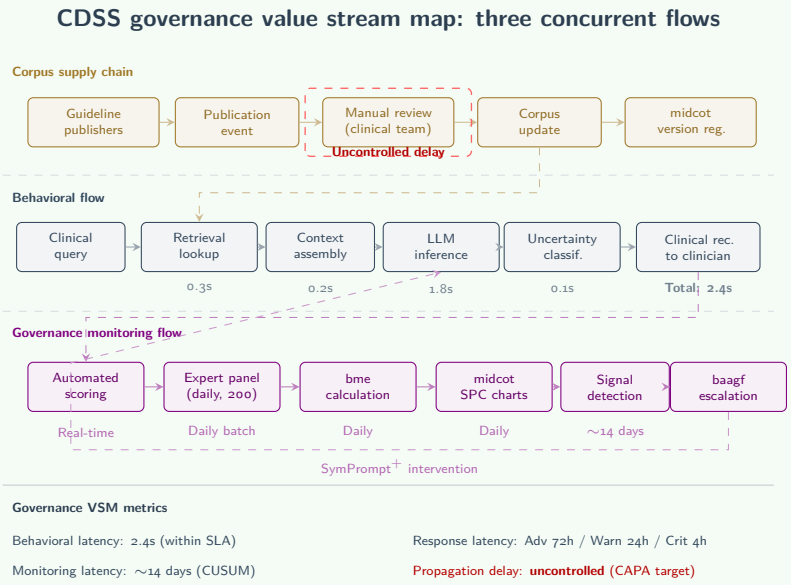


Figure 3.6: CDSS governance value stream map: three concurrent flows (corpus supply chain, behavioral inference pipeline, governance monitoring) with timing annotations and the governance-critical propagation delay identified.

Example 3.12: Insurance Claims Triage Governance Value Stream Map

Behavioral flow: Claim submission → Step 1 document intake (∼1.2s) → Step 2 coverage assessment with policy retrieval (∼2.1s) → Step 3 fraud screening (∼0.8s) → Step 4 pathway recommendation (∼1.1s) → triage output to claims handler. Total behavioral latency: ∼5.2s (within 95th percentile SLA of 8.0s).

Corpus supply chain (4 streams):

1. Policy wordings: Product development → annual release → corpus ingestion. Propagation delay: 2–4 weeks (acceptable; annual cycle).
2. Policy endorsements: Underwriting → quarterly + ad hoc release → corpus ingestion. Propagation delay: **16 days (uncontrolled)**. Root cause of the drift event MIDCOT-2026-0073.
3. Coverage guides: Claims technical team → semi-annual release → corpus ingestion. Propagation delay: 1–2 weeks.
4. Fraud indicators: Fraud unit → monthly intelligence update → corpus ingestion. Propagation delay: 3–5 days (acceptable).

Critical path: The endorsement supply chain has the highest propagation delay and the highest update frequency variability (ad hoc releases triggered by regulatory changes). The VSM identifies this as the governance-critical pathway, confirmed by the drift event.

Example 3.13: ED Triage Agent Governance Value Stream Map

Behavioral flow: Patient presentation → Step 1 intake parsing (EHR pull + vitals stream + free text, ∼1.4s) → Step 2 acuity classification with protocol retrieval (∼2.2s) → Step 3 resource allocation (∼0.6s) → Step 4 disposition with reasoning (∼1.8s) → triage recommendation to nurse. Total behavioral latency: ∼6.0s. *Clinical constraint:* triage decision must not delay patient flow; the 6.0s pipeline runs in parallel with the nurse's initial assessment.

Governance monitoring flow: Per-encounter automated scoring (real-time, 4 dimensions) → hourly batch aggregation → daily SPC chart update → CUSUM accumulation → signal detection. Monitoring latency for the sepsis screening drift: 18 days (CUSUM Advisory triggered Day 63, drift began Day 45). Response latency: Advisory investigation initiated within 24 hours of signal; corpus update deployed Day 65 (2 days post-detection).

Corpus supply chain: ACEP, ACS, Surviving Sepsis Campaign, institutional protocol committee → guideline publication event → *no automated monitoring* → manual review by Clinical AI Committee → corpus update → version registration. **Critical finding:** 18-day propagation delay for ACEP sepsis criteria update. CAPA targets: automated guideline publication monitoring, reducing propagation delay to <7 days.

Unique VSM element: The EDT value stream includes a *clinical override feedback loop* not present in the other scenarios: nurse override → override log → override concordance CTG calculation → fed back to governance

monitoring. High override rates are a leading indicator of AI-clinician misalignment, detectable before SPAR degradation.

Example 3.14: Insurance Claims Triage Governance Charter

Table 3.10: *ICT governance charter*

Charter Element	Specification
System	ClaimsTriage AI v2.3. RAG-augmented 4-step agentic claims triage.
Trigger	New deployment. High-risk: EU AI Act Annex III, §5(ca); TCF-regulated (FSCA).
Scope	Motor (52%), property (31%), liability (17%). ∼510 claims/-day, peak ∼780/day.
Authority	Claims AI Governance Committee: CRO (chair), Claims Ops Mgr, Head of IT, Compliance Officer, external actuarial advisor.
Escalation	T1 Advisory: 48-hr investigation. T2 Warning: restrict AI autonomy, 4 hr. T3 Critical: suspend affected lines, 1 hr. T4 Emergency: full halt, 15 min; regulatory notification.
Review cadence	Quarterly. Unscheduled on T3+ or regulatory directive.
Measurement	Daily batch. Real-time for 30 days post-intervention and re-baselining.
Suspension	T3 suspends affected lines. T4 full suspension. SPAR <0.940 for >48 hr.
Decommission	SPAR <0.960 after two DMAIC cycles. Regulatory directive.

Example 3.15: ED Triage Agent Governance Charter

Table 3.11: *EDT governance charter*

Charter Element	Specification
System	TriageAssist ED v3.1. 4-step agentic pipeline. Fine-tuned 8B model on 2.4M ED encounters.
Trigger	New deployment. EU AI Act high-risk Annex III, §5(c). EMTALA-regulated. Joint Commission accredited.
Scope	All ED triage: ∼258/day, peak ∼340/day. All ESI levels and complaint categories.
Authority	Three-tier: Executive Oversight (CMO chair); Clinical AI Committee (ED Med Dir, CMIO, Pt Safety, Bioethicist); Operational Monitoring (2 attendees, 2 charge nurses, AI analyst).
Escalation	Advisory: CUSUM signal; 48-hr. Warning: BAR violation or under-triage >0.030; 4 hr. Critical: 2+ violations or under-triage >0.040; suspend, 1 hr. Emergency: safety event or EMTALA violation; HALT, 15 min; CMS/TJC 24 hr.
EMTALA constraint	**ABSOLUTE:** Never recommend deferral/deprioritization/-denial of medical screening. Non-negotiable; not subject to committee override.
Suspension	Any Critical. Under-triage >0.040. Missed sepsis/STEMI/stroke with adverse outcome. SPAR <0.940.
Decommission	Safety event after two improvement cycles. Regulatory directive.

3.3.3 Innovated Constructs

Four Define-phase constructs require purpose-built innovation because they address governance problems with no manufacturing antecedent. The behavioral specification envelope defines multidimensional joint conformance across correlated behavioral metrics. The input class registry formally stratifies an unbounded input space into governable categories. The VOG weighting model resolves authority-asymmetric multi-class stakeholder conflicts through auditable composite priority scoring. The COPQ framework translates the economics of quality failure to AI systems where harm is probabilistic, delayed, and non-linear. Each of these constructs is introduced with a formal definition, an operational protocol, and fully populated worked examples.

Behavioral Specification Envelope **[I⋆Innovate]**

Module: TM-T1-06

Why Innovation Is Required

Classical tolerance specifications are unidimensional. LLM behavioral governance requires *multidimensional joint conformance*: an output must simultaneously satisfy specification limits on all governed behavioral dimensions. There is no classical analog for a conformance region defined by the intersection of limits across correlated, non-normally distributed behavioral metrics.

Construct Definition

The behavioral specification envelope is the multidimensional region of acceptable behavioral output defined by the complete set of CTG characteristics and their specification limits. A system output is conformant if and only if it falls within specification on all governed dimensions simultaneously. The SPAR metric quantifies the proportion of outputs achieving joint conformance.

Operational Protocol

Construct the envelope as the joint conformance region from all primary CTG specification limits. Document the joint conformance requirement explicitly. For correlated characteristics, define conditional specification structures. Set the SPAR target (default: ≥ 0.997). Record with version control; all changes require governance charter-level approval and trigger re-baselining.

Example 3.16: ICT Behavioral Specification Envelope

Table 3.12: *ICT specification envelope by input class*

Input Class	Cov. Acc.	Path. Con.	Fraud Pr.	Fair.	Compl.	spar
Motor – Low – Simple	0.930	0.910	0.620	0.045	0.960	0.975
Motor – High – Complex	0.940	0.920	0.650	0.040	0.960	0.960
Property – Low – Simple	0.920	0.900	0.600	0.050	0.950	0.970
Property – High – Complex	0.935	0.915	0.630	0.040	0.955	0.955
Liability (all)	0.940	0.920	0.640	0.040	0.960	0.960

High-value complex claims receive tighter limits reflecting higher downstream consequences. A triage output is conformant if and only if it falls within all five limits simultaneously for its assigned input class.

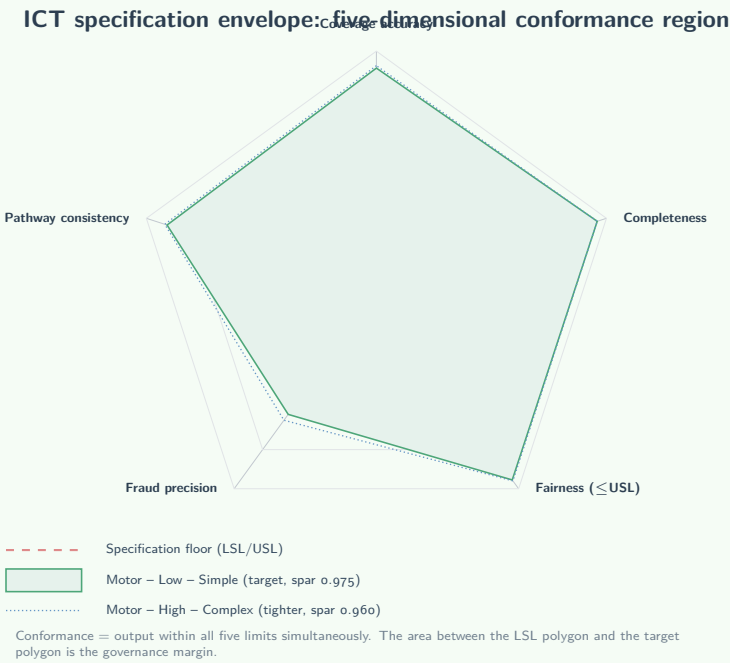


Figure 3.7: ICT specification envelope: five-dimensional radar chart showing the joint conformance region. The dashed polygon marks the specification floor (LSL/USL); the solid polygon marks the Motor–Low–Simple target profile. An output is conformant only within all five dimensions simultaneously.

Example 3.17: EDT Per-Step Specification Envelope

Table 3.13: *EDT per-step and workflow specification envelope*

Pipeline Step	spar Tgt	Floor	Primary CTG	Safety CTG	Cascading Risk
Step 1: Intake	0.995	0.990	Extraction acc.	—	Low→High
Step 2: Acuity	0.990	0.985	Acuity acc.	Under-triage	Critical
Step 3: Resource	0.990	0.985	Resource approp.	—	Medium
Step 4: Disposition	0.990	0.985	Reasoning transp.	—	Medium
Workflow (product)	0.960	0.940	Joint 6-CTG	Under-tri.+screen.	Multiplicative

The EDT requires both per-step and workflow-level specifications reflecting the cascading failure model. Workflow SPAR is the product of per-step conformance: $0.995 \times 0.990 \times 0.990 \times 0.990 \approx 0.965$.

Input Class Registry **[I⋆Innovate]**

Module: TM-T1-07

Why Innovation Is Required

Manufacturing SPC operates on physically bounded input conditions. LLM input spaces are effectively unbounded. Without formal input stratification, behavioral variation conflates process drift with input diversity, rendering control charts uninterpretable. No classical SPC construct addresses how to define conditional process populations in an open-ended input space.

Construct Definition

The input class registry formally stratifies the LLM input space into defined categories under which the behavioral distribution is expected to remain stable. Each class specifies defining characteristics, expected behavioral profile, class-specific specification limits, and boundary conditions. The registry also defines the out-of-scope category with specified handling rules.

Operational Protocol

Identify classification dimensions relevant to the deployment. Define each input class with assignment criteria. Specify class-specific specification limits where warranted. Define out-of-scope handling. Establish the classification method and document accuracy expectations.

Example 3.18: ICT Input Class Registry

Table 3.14: *ICT input class registry*

Dimension	Classes	Rationale
Line of business	Motor (52%), Property (31%), Liability (17%)	Different policy structures require separate baselines.
Claim value	Low (<R50K, 64%), Med (R50K–R250K, 28%), High (>R250K, 8%)	Value affects pathway distribution and fraud sensitivity.
Channel	Broker (68%), Digital (24%), Call center (8%)	Channel affects submission quality.
Complexity	Simple (71%), Complex (multi-peril/disputed, 29%)	Complex claims are the primary SPAR gap source.

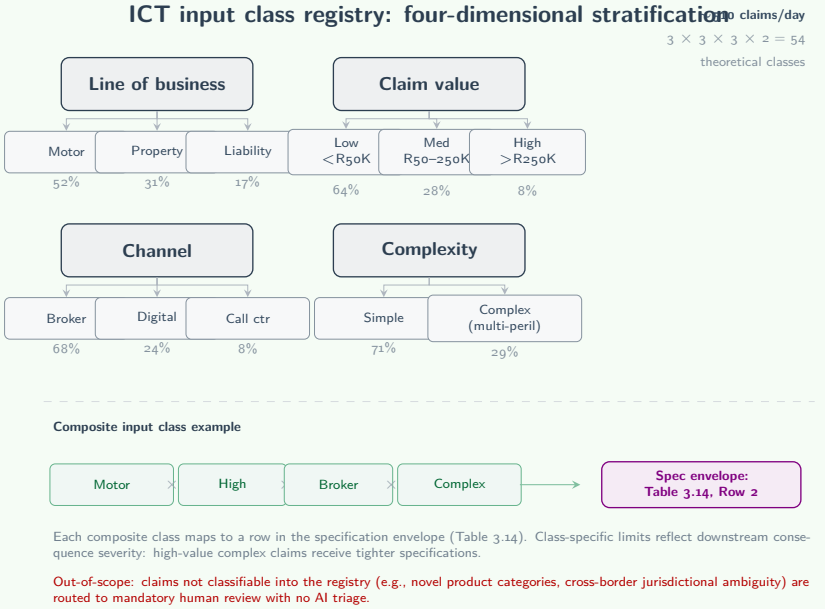


Figure 3.8: ICT input class registry: four classification dimensions with their categories and volume distributions. Each composite input class (e.g., Motor–High–Broker–Complex) maps to a specific row in the specification envelope, enabling class-stratified governance monitoring.

Example 3.19: EDT Input Class Registry

Table 3.15: *EDT input class registry*

Dimension	Classes	Rationale
Acuity	ESI-1 (2%), ESI-2 (18%), ESI-3 (42%), ESI-4/5 (38%)	Higher acuity = tighter specs. ESI-1/2 time-critical.
Chief complaint	Chest pain (14%), Abdominal (12%), Injury (18%), Resp. (11%), Neuro (8%), Psych (6%), Peds (9%), Other (22%)	Determines screening protocol applicability.
Complexity	Standard (68%), Complex (32%)	Complex presentations: SPAR 0.908–0.914 vs. 0.944–0.966.
Time of day	Day (38%), Evening (40%), Night (22%)	Night: thinner staffing, more altered patients.

VOG Multi-Stakeholder Weighting Model **[I ⋆ Innovate]**

Module: TM-T1-08

Why Innovation Is Required

Classical VOC prioritization operates within a single stakeholder class using importance ratings and business impact scores. AI governance involves multiple classes with different levels of authority and fundamentally different relationships to the governed system. Regulatory requirements must take precedence regardless of "importance." No classical method handles authority-asymmetric, multi-class conflict resolution.

Construct Definition

The VOG Weighting Model combines three dimensions into a Composite Priority Score (CPS): stakeholder authority weight (regulatory > affected populations > deploying organization > end users), requirement urgency, and governance impact severity. The CPS resolves inter-class conflicts through explicit, auditable priority calculation.

Operational Protocol

Assign authority weights for each stakeholder class based on the deployment risk classification. Score each requirement on urgency and impact. Calculate CPS. Rank requirements. Review for face validity. Document methodology and overrides.

AI Governance Cost of Poor Quality (COPQ) Framework **[I⋆Innovate]**

Module: TM-T1-10

Classical Definition

Cost of Poor Quality quantifies the total cost an organization incurs when its processes produce outputs that fail to meet quality specifications. Classical COPQ comprises four categories: internal failure costs (rework, scrap), external failure costs (warranty, returns, liability), appraisal costs (inspection, testing), and prevention costs (training, quality planning). Juran's economics of quality establishes that investment in prevention reduces total COPQ by shifting expenditure from reactive failure costs to proactive prevention [22].

Why Innovation Is Required

Classical COPQ assumes that quality failures produce countable, cost-assignable defects: a defective part has a known scrap, rework, or warranty-claim cost. AI behavioral nonconformance has consequences that differ fundamentally in three respects. First, *harm is often probabilistic and delayed*: an under-triaged ED patient may or may not experience an adverse outcome, and the causal connection between AI recommendation and patient outcome is mediated by clinical decision-making. Second, *regulatory consequences are threshold-dependent*: a single EMTALA violation has different cost implications than a pattern of discriminatory triage. Third, *reputational and trust costs are non-linear*: the first publicized AI governance failure at an organization may cost more in stakeholder trust than the next hundred nonconforming outputs combined.
Classical COPQ categories require translation:

COPQ Category	Classical (Manufacturing)	AI Governance Translation
Internal failure	Rework, scrap, downtime	Human review/override of nonconforming outputs; reprocessing costs; investigation labor; targeted DMAIC cycle costs
External failure	Warranty, returns, liability, lost customers	Stakeholder harm (clinical, financial, legal); regulatory penalties; litigation costs; policyholder complaints; trust erosion
Appraisal	Inspection, testing, audit	BME measurement system operation; evaluator panel costs; MSA/GRR validation; MIDCOT infrastructure; governance reporting
Prevention	Training, quality planning, process design	Define-phase governance design; SymPrompt$^+$ architecture; retrieval corpus governance; evaluator calibration; preventive MIDCOT monitoring

AI-COPQ Calculation Framework

The AI governance COPQ is calculated per governance period (typically quarterly) across four cost streams:

Internal failure costs. For each nonconforming output detected before reaching the end user (or detected and corrected through human-in-the-loop override): the cost of human review time, the cost of re-processing, and the cost of investigation when the nonconformance triggers an Advisory or Warning escalation.

$C_{\text{internal}} = \sum_i n_i \cdot c_{\text{review},i} + \sum_j T_j \cdot c_{\text{investigation},j}$

where n_i is the count of nonconforming outputs in input class i requiring human review, $c_{\text{review},i}$ is the per-output review cost, T_j is the investigation time for escalation event j, and $c_{\text{investigation},j}$ is the hourly investigation cost.

External failure costs. For each nonconforming output reaching the end user: the expected cost of downstream consequences, calculated as the product of nonconformance frequency and expected consequence cost, stratified by severity tier.

$C_{\text{external}} = \sum_k f_k \cdot E[C_{\text{consequence},k}]$

where f_k is the frequency of severity-tier-k nonconformances reaching end users, and $E[C_{\text{consequence},k}]$ is the expected consequence cost (regulatory penalty, litigation exposure, clinical harm cost, complaint resolution cost).

Appraisal costs. The operating cost of the governance measurement and monitoring system: evaluator panel compensation, automated scoring

infrastructure, MIDCOT compute and storage, governance reporting labor, and periodic MSA revalidation.
Prevention costs. The investment in governance design and maintenance: Define-phase labor, SymPrompt$^+$ architecture and maintenance, retrieval corpus governance, evaluator training and calibration, and preventive monitoring enhancements.

COPQ-Driven Governance Investment Justification

The AI-COPQ framework provides the economic argument for governance investment. The Juran principle applies directly: investment in prevention (governance design, corpus governance, monitoring infrastructure) reduces total COPQ by preventing external failures whose individual costs far exceed the cost of prevention. The COPQ calculation is included in the governance charter as an economic specification alongside the behavioral specifications, and is reported quarterly to governance authorities.

Operational Protocol

During the Define Phase: estimate the four COPQ streams for the governed system using historical data, regulatory penalty schedules, and domain-specific consequence models. Establish the COPQ baseline. Set the COPQ reduction target (typically aligned with the SPAR improvement target). During the Measure phase: collect actual COPQ data alongside BME metrics. During the Control phase: report COPQ trends alongside SPC charts. The COPQ trend validates the governance investment: declining external failure costs and stable or declining total COPQ confirm that prevention investment is effective.

Example 3.20: CDSS COPQ Analysis

Internal failure: Baseline SPAR of 0.943 on diagnostic queries means 5.7% of outputs are nonconforming. At 500 diagnostic queries/day, ∼29 outputs/day require enhanced clinical review. Estimated review cost: 15 minutes per output at $130/hr clinical time = $94/day = $34,310/year.
External failure: Uncertainty disclosure omission (the primary governance gap) means patients may receive overly confident clinical recommendations. Estimated external failure: 1 malpractice-relevant incident per 10,000 nonconforming outputs (conservative estimate based on clinical literature). At ∼10,585 nonconforming outputs/year, expected incidents ∼1.1/year. Average malpractice settlement cost: $302,000. Expected annual external failure cost: $332,000.
Appraisal: Clinical evaluator panel (3 experts, 0.2 FTE each): $78,000/year. MIDCOT infrastructure: $19,500/year. MSA revalidation (annual): $8,600. Total appraisal: $106,100/year.
Prevention: SymPrompt$^+$ modulation design and DOE validation: $27,000 (one-time). Corpus currency monitoring pipeline: $16,200/year. Evaluator calibration: $6,500/year. Total prevention: $49,700/year (after first-year

setup).
Total baseline COPQ: $522,110/year. Post-intervention (SPAR 0.943 → 0.981): internal failure costs reduce by ∼67%; external failure expected incidents reduce to ∼0.3/year. **Projected post-intervention COPQ:** $257,000/year. **Annual COPQ reduction: $265,000.** Prevention investment payback: <3 months.

Example 3.21: Insurance Claims Triage COPQ Analysis

Internal failure: Baseline SPAR of 0.921 means 7.9% of outputs nonconforming. At ∼510 claims/day, ∼40 claims/day require enhanced human review. Senior handler review at $47/hr, 20 min/claim: $31/day = $11,400/year. Targeted DMAIC cycle (drift event MIDCOT-2026-0073): investigation + re-baselining labor: $4,700 per event.
External failure: Coverage misinterpretation on complex claims (FMEA #1, RPN 252) produces incorrect triage pathways. Estimated TCF-reportable unfair outcome rate: 0.3% of nonconforming outputs. At ∼14,600 nonconforming outputs/year, ∼44 potentially unfair outcomes. FAIS Ombud complaint resolution cost: $2,500 per complaint. Regulatory penalty exposure (systemic TCF non-compliance): $110,000 per finding. Expected annual external failure: $109,000 + $110,000 regulatory exposure = $219,000.
Appraisal: Expert evaluator panel (3 at 0.15 FTE): $22,300/year. Automated scoring infrastructure: $6,600/year. MIDCOT operations: $5,200/year. Total: $34,100/year.
Prevention: Corpus governance (including automated endorsement pipeline post-CAPA): $9,900/year. SymPrompt$^+$ maintenance: $4,700/year. Evaluator calibration: $2,500/year. Total: $17,100/year.
Total baseline COPQ: $286,300/year. Post-intervention (SPAR 0.921 → 0.974): internal failure reduces by ∼68%; external failure reduces by ∼75% (unfair outcome rate drops proportionally with complex-claim coverage accuracy improvement). **Projected post-intervention COPQ:** $106,000/year. **Annual COPQ reduction: $180,300.** ROI on governance program: ∼340%.

Example 3.22: ED Triage Agent COPQ Analysis

Internal failure: Baseline workflow SPAR of 0.932 means 6.8% of encounters nonconforming. At 258 encounters/day, ∼18 patients/day receive AI recommendations that require a senior clinician's override or review. Override review at $185/hr, 8 min/encounter: $444/day = $162,060/year. Drift event MIDCOT-2026-ED-0041: investigation + re-baselining labor: $42,000.
External failure (safety-critical): Under-triage rate of 0.031 means ∼8 patients/day assigned lower acuity than warranted. For the subset involving

time-critical conditions (STEMI, stroke, sepsis): estimated 2.1 patients/day with clinically significant delay. Adverse outcome probability per delayed patient: 3.2% (based on sepsis bundle delay mortality data). Expected adverse outcomes: ∼24.5/year. *Average cost per adverse outcome* (including extended ICU stay, litigation, regulatory investigation, sentinel event response): $380,000. Expected annual external failure cost: $9,310,000.
During the 18-day drift window specifically: 54–72 elderly patients with delayed sepsis screening; 8–12 confirmed sepsis; estimated 0.6–0.9 excess mortality events attributable to delayed bundle initiation. *Single drift event external failure cost*: $228,000–$342,000.
Appraisal: Evaluator panel (3 physicians + 2 nurses, 0.1 FTE each): $225,000/year. Automated scoring infrastructure: $65,000/year. MIDCOT real-time monitoring: $48,000/year. Total: $338,000/year.
Prevention: Corpus currency monitoring pipeline (post-CAPA): $35,000/year. SymPrompt$^+$ architecture and maintenance: $28,000/year. Evaluator calibration: $18,000/year. Total: $81,000/year.
Total baseline COPQ: $9,933,060/year. Post-intervention (under-triage 0.031 → 0.018): external failure reduces by ∼42% (adverse outcomes reduce from 24.5 to ∼14.2/year). **Projected post-intervention COPQ:** $5,930,000/year. **Annual COPQ reduction: $4,003,060.** Prevention investment payback: <1 month.
Cross-scenario observation: External failure costs dominate COPQ in all three scenarios (64% for CDSS, 77% for ICT, 94% for EDT), confirming the Juran principle: governance investment in prevention and appraisal is economically justified because external failure costs dwarf prevention costs by one to two orders of magnitude. The EDT scenario makes this starkest: a $81,000/year prevention investment reduces external failure costs by $9.3M/year by $4.0M.

3.4 Governance Artifacts Summary

The Define phase produces ten governance artifacts, each assigned to a specific tool module and consumed by one or more subsequent DMAIC phases. Table 3.16 registers the complete artifact set with its producing tool, AAI classification, and downstream function. The traceability requirement is absolute: every CTG characteristic must trace to a VOG requirement, every specification limit must trace to a governance rationale, and every escalation action must have a RACI assignment. An artifact set that fails any of these traceability checks is incomplete and must not proceed to the Measure phase.

Table 3.16: *Define phase governance artifact register*

Artifact	Produced By	Function in Subsequent Phases
Governance Charter	TM-T1-05 (**Adapt**)	Authority structure for Control-phase escalation; review cadence
VOG Analysis	TM-T1-03 (**Adapt**)	Validation criteria for Improve-phase interventions; conflict register
CTG Tree	TM-T1-04 (**Adapt**)	Measurement targets for Measure phase; improvement objectives
Governance VSM	TM-T1-09 (**Adapt**)	Critical path analysis; corpus supply chain monitoring targets; propagation delay SLAs
Specification Envelope	TM-T1-06 (**Innovate**)	Conformance standard for Analyze-phase; BAR derivation basis
Input Class Registry	TM-T1-07 (**Innovate**)	Stratification for all measurement, analysis, and monitoring
RACI Matrix	TM-T1-02 (**Adopt**)	Decision authority map for all governance actions
SIPOC	TM-T1-01 (**Adopt**)	System boundary and interface map; critical supplier dependencies
VOG Weighting	TM-T1-08 (**Innovate**)	Auditable priority register; conflict resolution documentation
COPQ Baseline	TM-T1-10 (**Innovate**)	Economic justification for governance investment; quarterly COPQ reporting

The artifact register reveals the Define Phase's structural function within the DMAIC lifecycle. The four Adapted artifacts (charter, VOG, CTG tree, governance VSM) establish the governance scope, stakeholder alignment, behavioral targets, and process flow that the Measure phase must operationalize. The four innovative artifacts (specification envelope, input class registry, VOG weighting, COPQ baseline) provide the multidimensional conformance standard, input stratification, conflict-resolution methodology, and economic framework, all of which have no manufacturing antecedents. The two adopted artifacts (SIPOC, RACI) provide the system boundary and decision authority structures that transfer directly from classical Six Sigma. Together, the ten artifacts constitute the governance specification package that authorizes the initiation of PM-002 (Measure phase).

3.5 Failure Modes of the Define Phase

The Define phase has nine characteristic failure modes, each representing a way in which the governance specification package can be incomplete, imprecise, or structurally flawed. These failures propagate forward: a vague specification produces an unmeasurable characteristic in the Measure phase; a deferred conflict produces an unresolvable trade-off in the Improve phase; an under-specified charter produces an ad hoc escalation response in the Control phase. Recognizing these

failure modes during the Define-phase execution is the primary defense against downstream governance dysfunction.

Specification vagueness
The most common failure mode. Requirements stated as governance aspirations ("be fair, be accurate") rather than measurable specifications with defined methods, targets, and limits. A Define phase producing vague specifications has deferred the definition problem to the Measure phase, where it manifests as unmeasurable characteristics and arbitrary thresholds.

Stakeholder exclusion
VOG analysis conducted without representation from all canonical classes. Affected populations are most commonly omitted. Consequence: a governance framework that serves the deploying organization while violating the standards to which it is subject.

Dimensional incompleteness
CTG trees decomposing governance needs along too few dimensions. A system governed solely by factual accuracy may exhibit an unacceptable tone, confidence, or framing while remaining "conformant" on the measured dimension.

Conflict deferral
Identifying inter-class VOG conflicts but deferring resolution. Deferred conflicts manifest as immeasurable characteristics, unresolvable Improve-phase trade-offs, or escalation disputes in Control.

Input class under-generalization
Specifications covering a narrow set of anticipated inputs, but not the actual deployment distribution. The input class registry must include explicit boundary definitions and out-of-scope declarations.

Specification rigidity
No provisions for revision, re-baselining, or adaptive refinement. AI systems operate in non-stationary environments; specifications assuming static conditions become misaligned. The charter must include a review cadence and revision process.

Charter underspecification
Governance charter lacking specificity on escalation protocols, review cadence, or decision authority. Leaves Control-phase decisions to ad hoc judgment rather than predefined protocol.

Corpus supply chain neglect
Failing to map the retrieval corpus supply chain in the governance value stream. All three scenarios' drift events are traced to uncontrolled propagation delays in this supply chain; VSM identifies the gap during the Define phase before it manifests as a Control-phase drift event.

Economic justification omission
Proceeding with governance design without establishing the COPQ baseline. Without COPQ, governance investment competes with other organizational priorities based on assertion rather than evidence.

3.6 AAI Classification Summary: Define Phase

Table 3.17 summarizes the AAI classification for all ten Define-phase tools. The distribution (2 Adopt, 4 Adapt, 4 Innovate) reflects the structural reality that governance initialization logic transfers from manufacturing to AI, but the instruments require substantial adaptation or replacement to accommodate stochastic, non-stationary AI behavior.

Table 3.17: *AAI classification profile: Define phase (2 Adopt, 4 Adapt, 4 Innovate)*

Tool	AAI	TM Ref	Rationale
SIPOC	**Adopt**	TM-T1-01	Component mapping applies directly to AI architectures
RACI Matrix	**Adopt**	TM-T1-02	Responsibility assignment is domain-independent
VOG Analysis	**Adapt**	TM-T1-03	Stakeholder scope expanded to full governance constituency
CTG Tree	**Adapt**	TM-T1-04	Decomposition retargeted from physical to behavioral characteristics
Governance Charter	**Adapt**	TM-T1-05	Extended from project instrument to lifecycle commitment
Governance VSM	**Adapt**	TM-T1-09	Material/information flows become behavioral/governance/corpus flows
Specification Envelope	**Innovate**	TM-T1-06	Multidimensional joint conformance; no classical analogue
Input Class Registry	**Innovate**	TM-T1-07	Formal stratification of unbounded input space
VOG Weighting	**Innovate**	TM-T1-08	Authority-asymmetric multi-class conflict resolution
COPQ Framework	**Innovate**	TM-T1-10	Probabilistic/delayed/non-linear AI failure economics; no classical cost model applies

The Define phase AAI profile (2 Adopt, 4 Adapt, 4 Innovate) reflects a structural reality: the logic of governance initialization transfers from manufacturing to AI, but the instruments require substantial adaptation or replacement to accommodate stochastic AI systems. The four innovations (specification envelope, input class registry, VOG weighting, COPQ framework) address governance problems that have no manufacturing antecedent: multidimensional behavioral conformance, unbounded input stratification, authority-asymmetric stakeholder conflict resolution, and probabilistic AI failure economics. The governance value stream map adapts

the classical VSM by adding a closed-loop governance monitoring flow and a corpus supply chain, which the three scenarios' drift events confirmed as the governance-critical pathway.

Chapter 4 presents the Measure phase: establishing the governance measurement system that operationalizes the specifications, targets, and stratification structures defined in this chapter.

Chapter 4

Measure Phase: Establishing the Governance Measurement System

The Measure phase serves three interdependent purposes: validating the measurement system that will generate all governance-grade data, collecting baseline behavioral data under controlled conditions, and calculating the process capability indices and control limit parameters that characterize the system's current governance posture. It operates on a foundational principle: *data used for process governance must itself be governed.* If the measurement system is unreliable, biased, or insufficiently precise, every conclusion drawn from its data is suspect.

In classical Six Sigma, the Measure phase validates physical measurement instruments through Measurement System Analysis (MSA), collects baseline data on current process performance, and calculates process capability indices (C_p, C_{pk}) that quantify the process's ability to meet specifications [28, 35]. The translation to AI governance encounters the most technically demanding challenges of any DMAIC phase: the evaluator reliability problem (measurement instruments are human raters, not calibrated devices), the ground truth problem (many behavioral dimensions lack objective reference standards), and the distribution problem (behavioral metrics routinely violate the normality assumption underlying classical capability indices).

This chapter extends the exposition of the Measure phase with fully populated worked examples from all three governance scenarios: CDSS, Insurance Claims Triage (ICT), and Emergency Department Triage (EDT). Each example demonstrates the complete measurement system validation, baseline collection, and BME metric calculation with real operational data drawn from the use-case data profiles (Appendix G and Appendix H).

4.1 Translation Challenges

The evaluator reliability problem. In manufacturing, measurement instruments have characterizable precision and accuracy. In AI governance, the "measurement instruments" are evaluators: human raters, automated scoring systems, or LLM-based judges. Human raters disagree, fatigue, and apply subjective criteria inconsistently. The measurement system must quantify evaluator reliability before any baseline data can be treated as valid.

The ground truth problem. Physical measurement has reference standards (a calibration weight is known to define precision). AI behavioral assessment frequently lacks ground truth. For some CTG characteristics (factual accuracy against a curated knowledge base), ground truth exists. For others (tone appropriateness, uncertainty disclosure quality), ground truth must be constructed through expert consensus and validated rubrics.

The distribution problem. Classical process capability indices assume normally distributed output. LLM behavioral metrics routinely violate this: they may be bounded, multimodal, heavily skewed, or exhibit heavy tails. The BME Metric Suite addresses this through distribution-free, entropy-based constructs that require no parametric assumptions.

4.2 The Measure Phase Protocol

The Measure phase protocol (PM-002) proceeds through seven steps, from measurement system design through phase completion and handoff. The protocol enforces a strict dependency chain: the GRR validation study (Step 2) must be completed before baseline data collection (Steps 3–4), because data collected with an unvalidated measurement system cannot be treated as governance-grade. The seven steps below are executed sequentially; shortcuts that bypass GRR validation are the most consequential procedural failure in the Measure phase.

Procedure: Measure Phase Protocol (PM-002)

Step 1: Measurement System Design. For each CTG characteristic, specify the measurement method (human evaluator panel, automated scorer, LLM judge, or hybrid), the evaluation rubric with anchor examples, the scoring scale, and the governance-grade GRR threshold. The threshold may differ from the conventional 10%/30% manufacturing standards where behavioral subjectivity warrants relaxation, per the governance charter.

Step 2: GRR Validation Study. Minimum three evaluators independently evaluate a minimum of 50 outputs per CTG characteristic spanning the full quality range, including borderline cases. Two evaluation rounds are separated

by at least two weeks. Calculate intra-rater agreement (repeatability) and inter-rater agreement (reproducibility via Krippendorff's alpha). Classify each characteristic as adequate, marginal (wider confidence intervals required), or inadequate (redesign before proceeding).

Step 3: Baseline Data Collection Protocol. Specify the baseline window (minimum four weeks for high-stakes systems), sample sizes per input class (minimum 200 observations per class per characteristic), input sampling method, configuration control requirements (frozen model version, stable corpus, fixed prompts), and evaluation protocol with disagreement resolution.

Step 4: Baseline Data Collection Execution. Verify configuration stability. Collect data per protocol. Monitor for configuration drift. Conduct interim sample adequacy checks. Compile the validated baseline dataset.

Step 5: bme Baseline Calculation. Calculate ECPI, BAR, ECI, SPAR, and CBMES for each CTG characteristic within each input class per MM-001 protocols. Assess the governance gap by comparing baseline values against specification targets. If all specifications are met with margin, proceed to PM-005 (Control). If gaps exist, proceed to PM-003 (Analyze).

Step 6: Ongoing Data Collection Plan. Specify ongoing measurement cadence, evaluator protocol, data retention requirements, and MSA refresh cadence (minimum annually or following any evaluator/rubric change). Where the Define phase established a COPQ baseline (TM-T1-10), specify the quarterly COPQ data collection protocol: internal failure counts (human review/override events), external failure incidents, appraisal costs, and prevention expenditures. COPQ data collection runs in parallel with BME metric collection and feeds the Control-phase governance reporting.

Step 7: Phase Completion and Handoff. Validate artifact completeness. Determine next phase (PM-003 or PM-005). Prepare Measure Phase Completion Report.

4.3 Measure Phase Tool Specifications

This section specifies each tool and construct employed in the Measure phase, following the AAI classification framework. Two tools are adopted without modification, two are adapted with defined extensions, and four are purpose-built innovations. The Measure phase exhibits the highest Innovate concentration of any DMAIC phase, reflecting the fundamental reality that classical SPC measurement constructs embed distributional assumptions that LLM behavioral outputs routinely violate.

4.3.1 Adopted Tools

Two Measure-phase tools transfer directly from classical Six Sigma: the data collection plan for systematic sampling design, and operational definitions for unambiguous measurement specification. Both operate on universal principles (planned collection and precise definition) that apply to AI behavioral metrics without structural modification. Additionally, histogram/distribution analyses, along with run charts for behavioral trends, are adopted as standard visualization tools that require no adaptation for AI governance data.

Data Collection Plan **[A · Adopt]**

Module: TM-T2-01

Classical Definition

A structured document specifying what data to collect, from what population, using what sampling method, at what frequency, and with what recording protocol. Ensures data collection is systematic rather than ad hoc.

Why Adopt

The structural logic of planned data collection applies directly to AI behavioral metrics. The content changes (behavioral metrics rather than physical dimensions, input classes rather than production lots), but the method of designing a collection plan requires no modification.

Operational Protocol

Specify: data elements per CTG characteristic, sampling frame (production traffic, synthetic inputs, or hybrid), sample size rationale (bootstrap-based for non-normal distributions), collection frequency, configuration control requirements, data storage and retention, and quality checks.

Example 4.1: CDSS Data Collection Plan

Four-week baseline window, frozen configuration (model version, retrieval corpus, SymPrompt$^+$ SP-CDSS-v2.3). 500 outputs per input class (diagnostic, treatment, drug interaction, patient communications), evaluated by two of three clinical domain experts, with the third as adjudicator. Configuration stability is verified daily through automated hash checks of model weights, the corpus index, and prompt templates.

Example 4.2: ICT Data Collection Plan

Table 4.1: *ICT baseline data collection parameters*

Parameter	Value
Collection period	February 1, 2026 to March 2, 2026 (30 calendar days)
Total claims	72,480 (Motor: 37,690 [52%]; Property: 22,469 [31%]; Liability: 12,321 [17%])
Complex claims	21,019 (29.0%); High-value (>R250K): 5,798 (8.0%)
Evaluation method	Automated scoring (coverage accuracy, completeness) + expert panel (pathway consistency, fairness, fraud precision)
Expert panel	3 senior claims handlers (>10 years experience); quarterly inter-rater calibration
Configuration control	Frozen: ClaimsTriage AI v2.3, SymPrompt$^+$ SP-CT-v2.3.1, retrieval corpus v2026.02. No model updates, corpus changes, or configuration modifications during the collection period

Example 4.3: EDT Data Collection Plan

Table 4.2: *EDT baseline data collection parameters*

Parameter	Value
Collection period	30 days (Days 1–30 post-deployment)
Total encounters	7,740 (258/day × 30 days)
Evaluation method	100% automated scoring on 4 dimensions (acuity accuracy, screening sensitivity, override concordance, demographic fairness) + 15% expert panel review on all 6 dimensions
Expert panel	3 board-certified emergency physicians + 2 senior charge nurses; each >10 years ED experience, >50,000 patient encounters
Configuration control	Frozen: TriageAssist ED v3.1, SymPrompt$^+$ SP-ED-v3.1.2, retrieval corpus v2024.11
Per-step collection	All four pipeline steps (Intake, Acuity, Resource, Disposition) scored independently to enable cascading failure analysis

Operational Definitions for AI Behavioral Metrics [A · Adopt]

Module: TM-T2-02

Classical Definition

An operational definition specifies the exact meaning of a measurement term: what is measured, the measurement method, and the decision criteria. Two evaluators using the same operational definition should produce the same result for the same object.

Why Adopt

The principle of unambiguous measurement definition is universal. AI behavioral metrics require particularly rigorous operational definitions because the measured characteristics are inherently more subjective. "Citation accuracy" must specify: what counts as a citation, what "accuracy" means (source exists, supports the claim, is current), and how edge cases are resolved.

Operational Protocol

For each CTG characteristic: define exactly what is measured, the evaluation procedure, the scoring scale, explicit decision criteria for edge cases, and at least three conformant and three nonconformant examples, including borderline cases. Pilot test with two or more evaluators on 20- 30 outputs. Refine until agreement meets the governance threshold. Train all evaluators before data collection.

Histogram and Distribution Analysis [A · Adopt]

Module: TM-T2-06

Why Adopt

Distributional visualization and characterization (histograms, density estimates, QQ-plots) apply without modification to behavioral metric data. The analysis reveals the distribution's shape, modality, skewness, and tail behavior: essential information for selecting appropriate BME calculation methods and identifying input-class-specific differences in distribution.

Run Charts for Behavioral Trends [A · Adopt]

Module: TM-T2-06

Why Adopt

Temporal trend visualization (plotting metric values over time against a center line) applies without modification to behavioral metric time series. Run charts detect trends, cycles, and shifts during baseline collection, providing visual evidence of process stability before formal control chart deployment.

4.3.2 Adapted Tools

Two tools require defined adaptations. Measurement System Analysis (MSA/GRR) must account for evaluator subjectivity, the absence of stable reference parts, and domain-calibrated reliability thresholds that differ from the 10%/30% manufacturing convention. The process capability index (ECPI) replaces the parametric C_p/C_{pk} calculation with an empirical, quantile-based computation that directly handles non-normal, multimodal, and heavy-tailed behavioral distributions.

MSA/GRR for AI Behavioral Metrics **[D ▷ Adapt]**

Module: TM-T2-03

Classical Definition

Measurement System Analysis decomposes total observed variation into process variation and measurement system variation. Gauge Repeatability and Reproducibility (GRR) quantifies: repeatability (same operator, same part, same conditions produce the same result) and reproducibility (different operators produce the same result for the same part). Classical acceptance: GRR $<$ 10% of total variation is excellent; 10–30% is marginal; $>$ 30% is unacceptable.

What Is Preserved

The decomposition logic (separating measurement variation from process variation) and the imperative to validate measurement before collecting governance data.

What Is Modified

Three adaptations. First, "operators" become evaluators (human raters, automated scorers, LLM judges) whose variability differs from that of physical instruments: human evaluators exhibit fatigue, subjective drift, and anchoring effects. Second, "parts" become AI system outputs whose behavioral characteristics may not be stable across evaluation occasions. Third, the 10%/30% thresholds require recalibration: research on inter-rater reliability for LLM output quality consistently reports moderate agreement (Cohen's κ in 0.4–0.7 for complex behavioral dimensions), which translates to GRR contributions that would be unacceptable in manufacturing. The governance charter must specify domain-calibrated thresholds.

Operational Protocol

Select at least 50 outputs spanning the full quality range for each CTG characteristic. Assign a minimum of three evaluators. Execute two evaluation rounds with a two-week separation. Calculate intra-rater agreement and inter-rater agreement (Krippendorff's α for ordinal scales; classical GRR decomposition for continuous scales). Classify each characteristic as adequate, marginal (document wider confidence intervals), or inadequate (redesign measurement before proceeding).

Krippendorff's Alpha Calculation

For ordinal and categorical scales, Krippendorff's alpha (α_K) is the recommended inter-rater reliability statistic because it handles missing data, accommodates any number of raters, and applies to all measurement levels. Values above 0.80 indicate strong agreement; 0.67–0.80 is adequate for governance use with documented uncertainty; values below 0.67 are marginal, requiring wider confidence intervals for all downstream calculations.

Evaluator Calibration Protocol

Before the GRR study, all evaluators undergo a structured calibration process. Phase 1: independent study of operational definitions (TM-T2-02) with worked examples spanning conformant, nonconformant, and borderline cases. Phase 2: calibration exercise on 20 pre-scored outputs with known reference classifications. Phase 3: calibration accuracy assessment; evaluators achieving less than 80% agreement with reference classifications receive additional training. Phase 4: structured discussion of disagreements to surface definitional ambiguities. This protocol prevents evaluators from entering the GRR study with systematically different interpretations of scoring criteria.

LLM-as-Judge Considerations

Where automated LLM-based evaluators replace or supplement human raters, the GRR study must account for the evaluator LLM's own stochastic variation. Multiple inference passes per evaluation (minimum five) are required to assess evaluator repeatability. Inter-evaluator reproducibility is assessed by comparing multiple LLM evaluator configurations. The LLM evaluator must be validated against human expert reference scores to establish measurement validity, not merely reliability.

Example 4.4: CDSS GRR Validation

100 CDSS outputs (40 diagnostic, 30 treatment, 30 drug interaction) were evaluated by all three experts twice, with a two-week interval between evaluations.

Table 4.3: *CDSS GRR validation results*

CTG Characteristic	Intra-Rater	Krippendorff's α	Classification
Citation accuracy	>90%	0.78	Adequate
Guideline concordance	88%	0.75	Adequate
Uncertainty disclosure	80%	0.55	Marginal
Contradiction avoidance	82%	0.58	Marginal

All four characteristics proceed with documented precision claims. Marginal characteristics receive wider confidence intervals on all subsequent governance calculations.

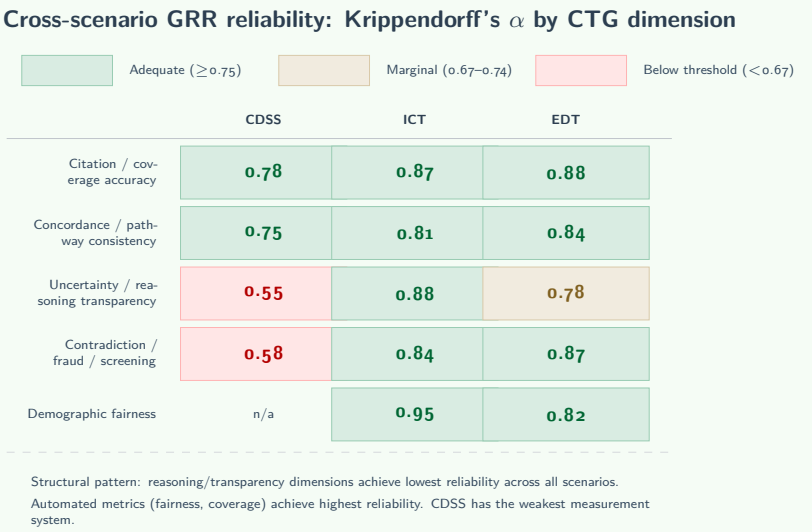


Figure 4.1: Cross-scenario GRR reliability heatmap: Krippendorff's α by CTG dimension. Green cells indicate adequate reliability ($\alpha_K \geq 0.75$); gold indicates marginal (0.67–0.74); red indicates below governance threshold (<0.67). The reasoning and transparency dimensions consistently exhibit lower reliability than the factual accuracy dimensions.

Example 4.5: Insurance Claims Triage GRR Validation

The ICT measurement system combines automated scoring (coverage accuracy, response completeness) with expert panel evaluation (pathway consistency, fraud precision, demographic fairness). The GRR study evaluated 150 claims spanning all lines of business and complexity levels.

Table 4.4: *ICT GRR validation results*

CTG Characteristic	Method	Intra-Rater	Krippendorff's α	Classification
Coverage accuracy	Automated + panel	94%	0.87	Adequate
Pathway consistency	Expert panel	86%	0.81	Adequate
Fraud flag precision	SIU feedback	91%	0.84	Adequate
Demographic fairness	Automated	98%	0.95	Adequate
Response completeness	Automated + panel	93%	0.88	Adequate

Overall GRR: 0.87, exceeding the 0.80 governance-grade threshold. All five characteristics are classified as adequate. The automated scoring components (coverage accuracy, completeness, fairness) achieve higher reliability than purely expert-evaluated dimensions, as expected. The insurance domain benefits from more structured output formats (e.g., JSON with enumerated fields) than clinical free-text scenarios.

Example 4.6: ED Triage Agent GRR Validation

200 AI triage outputs across all input classes, evaluated on all 6 CTG dimensions by 3 board-certified emergency physicians and 2 senior charge nurses. Two evaluation rounds with a 14-day separation.

Table 4.5: *EDT GRR validation results*

CTG Characteristic	Intra-Rater	Kripp. α	GRR Threshold	Classification
Acuity accuracy	94%	0.88	0.75	Adequate
Under-triage rate	91%	0.84	0.75	Adequate
Screening sensitivity	93%	0.87	0.75	Adequate
Reasoning transparency	86%	0.78	0.75	Borderline
Demographic fairness	89%	0.82	0.75	Adequate
Override concordance	92%	0.85	0.75	Adequate

Overall GRR: 0.86. Governance-grade. Reasoning transparency ($\alpha_K = 0.78$) is borderline: evaluators show moderate disagreement on what constitutes a "clinically valid reasoning chain." This dimension flagged for evaluator calibration improvement. All downstream ECPI and SPAR calculations

involving reasoning transparency carry wider confidence intervals reflecting this measurement limitation.
Cross-scenario observation: The three GRR studies reveal a consistent pattern: reasoning quality and transparency dimensions achieve the lowest inter-rater reliability across all three scenarios (CDSS uncertainty disclosure: 0.55; EDT reasoning transparency: 0.78). This reflects a structural measurement challenge: evaluating whether an AI system's reasoning chain is "adequate" requires clinical or domain judgment that resists operationalization into binary rubrics. The BME framework accommodates this through the marginal classification mechanism, which propagates measurement uncertainty into all downstream calculations rather than ignoring it.

Process Capability: ecpi **[D ▷ Adapt]**

Module: TM-T2-04

Classical Definition

Process capability indices (C_p, C_{pk}) quantify the relationship between process variation and specification limits. $C_{pk} = \min\left(\frac{USL-\bar{x}}{3\sigma}, \frac{\bar{x}-LSL}{3\sigma}\right)$. Assumes a normal distribution, a stable process, and known specification limits.

What Is Preserved

The interpretive framework: values above 1.33 indicate a capable process; 1.00–1.33 marginal; below 1.00 incapable. The comparative logic (ratio of specification space to process variation). The requirement for in-control data before capability calculation.

What Is Modified

The calculation method replaces parametric assumptions with empirical quantiles:

$$\text{ECPI} = \frac{LSL_{\text{spec}} - Q_\alpha}{Q_{0.5} - Q_\alpha} \tag{4.1}$$

where $Q_{0.5}$ is the median, Q_α is the α-quantile (typically 0.135th percentile), and LSL_{spec} is the specification limit. This handles non-normal, multimodal, and heavy-tailed distributions directly. ECPI is calculated independently per input class; the aggregate ECPI is the minimum across classes and dimensions.

Operational Protocol

Verify baseline stability through run chart analysis and configuration audit. For each CTG characteristic within each input class, calculate the empirical quantiles $Q_{0.5}$ and Q_α from the baseline dataset. Apply nonparametric bootstrap (minimum 2,000 resamples) to construct 95% confidence intervals for both quantile estimates and the resulting ECPI value.

Multi-Dimensional Capability Assessment

ECPI is calculated independently for each CTG dimension within each input class, producing a capability matrix. The aggregate ECPI is the minimum across all dimensions and classes, following the principle that the weakest link determines process capability.

Relationship to Specification Limits

ECPI interprets specification limits from the behavioral specification envelope (Chapter 3) as governance-set boundaries. For one-sided specifications (common for behavioral metrics with lower bounds), only the lower specification limit (LSL_{spec}) participates in the calculation.

4.3.3 Innovated Constructs

Four Measure-phase constructs require purpose-built innovation because they address measurement problems with no classical SPC antecedent. The BAR threshold derivation computes empirical quantile-based control limits that do not assume normality. The baseline distribution characterization replaces the classical ($\bar{x}$, σ) summary with a full distributional profile. The CBMES baseline calculation produces a composite escalation metric mapping to the T0–T5 tier model. The ongoing data collection protocol manages configuration-aware, evaluator-versioned measurement in a non-stationary deployment environment.

bar Threshold Derivation **[I⋆Innovate]**

Module: TM-T2-05

Why Innovation Is Required

Classical control limits ($\bar{x} \pm 3\sigma$) assume a normally distributed process output. LLM behavioral distributions are typically non-normal. Parametric limits applied to non-normal data produce incorrect false alarm rates: too wide for heavy-tailed distributions, too narrow for bounded distributions.

Construct Definition

BAR thresholds are empirical, quantile-based boundaries that define expected behavioral variation. Set at the 0.135th and 99.865th percentiles of the baseline behavioral distribution (corresponding to three-sigma equivalent boundaries under normality, but computed without the normality assumption). They define the region within which variation is attributed to common causes; points outside this region trigger a special-cause investigation.

Operational Protocol

From the validated baseline distribution, compute the empirical 0.135th and 99.865th percentiles. Bootstrap confidence intervals for threshold estimates. Verify that all thresholds fall within the specification envelope ($\text{BAR}_L > LSL_{\text{spec}}$). If $\text{BAR}_L \leq LSL_{\text{spec}}$, the process is not capable at

the desired sigma level, and an Improve-phase intervention is required. Configure thresholds into MIDCOT control charts. Thresholds are revised during re-baselining following authorized system changes.

Baseline Distribution Characterization **[I⋆Innovate]**

Module: TM-T2-06

Why Innovation Is Required

Classical baselines assume that, under normality, the mean and standard deviation characterize process output. LLM behavioral baselines require a full characterization of the empirical distribution, including shape, modality, skewness, tail behavior, and input-class-specific distributional differences.

Operational Protocol

Characterize the empirical distribution per CTG characteristic and input class using: kernel density estimation, empirical CDF, quantile summaries (5th, 25th, 50th, 75th, 95th, 99th, 99.865th percentiles), tests for multimodality, and distribution comparison across input classes. The characterized distribution serves as the reference for all subsequent drift detection.

cbmes Baseline Calculation **[I⋆Innovate]**

Module: TM-T2-07

Why Innovation Is Required

No classical SPC construct provides a composite escalation metric that aggregates multiple process capability dimensions into a single governance posture score calibrated to a tiered escalation model. CBMES fills this role, integrating ECPI, ECI, and SPAR into the T0–T5 escalation framework.

Operational Protocol

Calculate the component BME indices per MM-001. Apply the CBMES aggregation formula with governance-priority weights calibrated to the deployment context. Map the resulting score to the BAAGF escalation tier. Document the initial tier classification and the component contributions to the composite score.

Ongoing Data Collection Protocol **[I⋆Innovate]**

Module: TM-T2-08

Why Innovation Is Required

Classical ongoing data collection assumes a stable measurement infrastructure collecting fixed-format data at regular intervals. AI governance monitoring must accommodate: multiple evaluator types with periodic reliability revalidation, input-class-stratified sampling with adaptive rates, version-controlled collection tied to the system change registry, and MSA refresh triggers. This complexity exceeds the scope of the classical data collection plan.

Operational Protocol

Specify per-input-class monitoring cadence (real-time, daily, weekly). Define the evaluator rotation protocol and calibration maintenance schedule. Establish data retention and audit trail requirements. Specify MSA refresh triggers (annual minimum; immediate following evaluator changes). Link collection protocol to MIDCOT change registry for configuration-aware sampling.

4.4 Worked Examples: Baseline BME Metrics

This section demonstrates the complete Measure phase for all three governance scenarios, producing the baseline BME metric profiles that characterize each system's governance posture. Each worked example reports the full metric set (ECPI, ECI, SPAR, CBMES) stratified by input class, identifies the governance gap relative to specification targets, and determines the appropriate next phase (Analyze or Control). The cross-scenario comparison following the EDT example reveals a consistent baseline pattern, validating the input class registry's role as a critical Define-phase artifact.

Example 4.7: CDSS Measure Phase

Baseline bme Metrics.

Table 4.6: *CDSS baseline BME metric profile*

Metric	Diagnostic	Treatment	Drug Int.	Patient Comms
ECPI (aggregate)	1.35	1.52	1.41	1.78
ECI	2.8	3.1	2.9	3.5
SPAR	0.943	0.961	0.952	0.978
CBMES	0.31	0.22	0.28	0.14

Governance Gap. SPAR across all clinical input classes falls below the 0.997 target. Diagnostic queries show the largest gap (SPAR = 0.943; ECI = 2.8, below Three Sigma). The dimensional decomposition reveals the binding

constraint: ECPI for uncertainty disclosure within diagnostic queries is 0.92 (incapable). By contrast, citation accuracy (ECPI = 1.52) and contradiction avoidance (ECPI = 1.78) are comfortably capable, demonstrating that the governance gap is dimension-specific rather than systemic.
The CBMES calculation places diagnostic queries at T2 (Moderate Drift, CBMES = 0.31) and patient communications at T1 (Stable, CBMES = 0.14). The stratified analysis demonstrates why input-class-specific governance is essential: pooling across all input classes would yield a composite SPAR of approximately 0.956, masking the diagnostic query deficiency with the highest clinical consequence.
Authorization: proceed to Analyze phase (Chapter 5). Prioritized targets: (1) uncertainty disclosure in diagnostic queries (ECPI = 0.92, incapable), (2) guideline concordance in diagnostic queries (ECPI = 1.35, narrow margin), (3) overall SPAR gap from 0.943 to 0.997.

Example 4.8: Insurance Claims Triage Measure Phase

Baseline bme Metrics (Global).

Table 4.7: *ICT baseline BME metric profile (global)*

Metric	Value	Target	Status	Interpretation
SPAR	0.921	0.970	Below target	7.9% nonconformant on ≥1 CTG dimension
ECPI	1.18	≥1.33	Below target	Process capability insufficient for governance-grade
ECI	2.4	≥3.0	Below target	Below Three Sigma governance standard
BAR (coverage acc.)	0.948 / 0.992	—	Established	Empirical quantile control limits
BAR (pathway cons.)	0.912 / 0.978	—	Established	Wider range: higher variation in pathway decisions

Baseline by Input Class.

Table 4.8: *ICT baseline BME metrics by input class*

Input Class	spar	Cov. Acc.	Path. Cons.	Fraud Prec.	Fair. Δ
Motor – Simple	0.952	0.968	0.941	0.762	0.018
Motor – Complex	0.884	0.921	0.903	0.731	0.024
Property – Simple	0.944	0.959	0.932	0.748	0.021
Property – Complex	0.878	0.912	0.894	0.714	0.031
Liability (all)	0.908	0.938	0.916	0.722	0.027

Governance Gap. Global SPAR of 0.921 falls 0.049 below the 0.970 target. The input-class decomposition reveals the structural pattern: complex multi-peril claims across all lines show SPAR 0.878–0.884, substantially below both the target and the simple-claim baseline (0.944–0.952). Coverage assessment accuracy is the binding constraint for complex claims (ECPI for complex motor coverage accuracy: 0.94, marginal), while simple claims are comfortably capable (ECPI: 1.42). The ECI of 2.4 confirms the system operates below Three Sigma: behavioral variability is too high for governance-grade monitoring.
Property – Complex shows the worst demographic fairness delta (0.031), approaching the 0.050 USL. This dimension requires monitoring, but is not the primary improvement target.
Authorization: proceed to Analyze phase (Chapter 5). Prioritized targets: (1) coverage assessment accuracy on complex multi-peril claims (55% of SPAR gap per variation decomposition), (2) pathway recommendation consistency (30% of gap).

Example 4.9: ED Triage Agent Measure Phase

Baseline bme Metrics (Per-Step and Workflow).

Table 4.9: *EDT baseline BME metrics*

Metric	Baseline	Target	Gap	Interpretation
SPAR (workflow)	0.932	0.960	−0.028	6.8% nonconformant. ∼18 patients/day.
SPAR (Step 1: Intake)	0.991	0.995	−0.004	Errors in complex/multilingual presentations
SPAR (Step 2: Acuity)	0.974	0.990	−0.016	**Primary gap.** Under-triage at ESI-2/3 boundary
SPAR (Step 3: Resource)	0.986	0.990	−0.004	Inherits Step 2 errors
SPAR (Step 4: Disposition)	0.981	0.990	−0.009	Reasoning/citation gaps
ECPI	1.12	≥1.33	−0.21	Below governance-grade
ECI	2.6	≥3.0	−0.4	Below Three Sigma
Under-triage rate	0.031	<0.020	+0.011	∼8 patients/day under-triaged
Screening sensitivity	0.958	0.970	−0.012	Sepsis: 4.2% false negatives

Per-Input-Class Breakdown.

Table 4.10: *EDT baseline workflow SPAR by input class*

Input Class	Acuity	Under-tri	Screen	Reason	Fair	spar
ESI-1/2 Standard	0.951	0.024	0.968	0.961	0.018	0.952
ESI-1/2 Complex	0.928	0.043	0.944	0.938	0.032	0.914
ESI-3 Standard	0.946	0.028	0.962	0.955	0.020	0.944
ESI-3 Complex	0.921	0.039	0.948	0.931	0.034	0.908
ESI-4/5	0.962	0.015	0.974	0.968	0.016	0.966
Pediatric (all)	0.934	0.036	0.951	0.942	0.038	0.918

Governance Gap. The per-step decomposition identifies Step 2 (Acuity Classifier) as the primary bottleneck, contributing 57% of the workflow SPAR gap. The cascading failure model is confirmed: per-step SPARs of $0.991 \times 0.974 \times 0.986 \times 0.981 = 0.932$ demonstrate that the workflow is only as strong as its weakest step.

The safety-critical finding: Complex ESI-1/2 and pediatric presentations show under-triage rates of 0.043 and 0.036, respectively. In clinical terms, an ESI-2 patient triaged as ESI-3 may wait 45- 90 minutes longer for physician evaluation. For time-sensitive conditions (STEMI, stroke, sepsis), that delay is the difference between recovery and permanent disability or death.

The demographic fairness delta for pediatric patients (0.038) approaches the 0.040 USL, indicating potential age-based acuity bias that warrants

investigation.

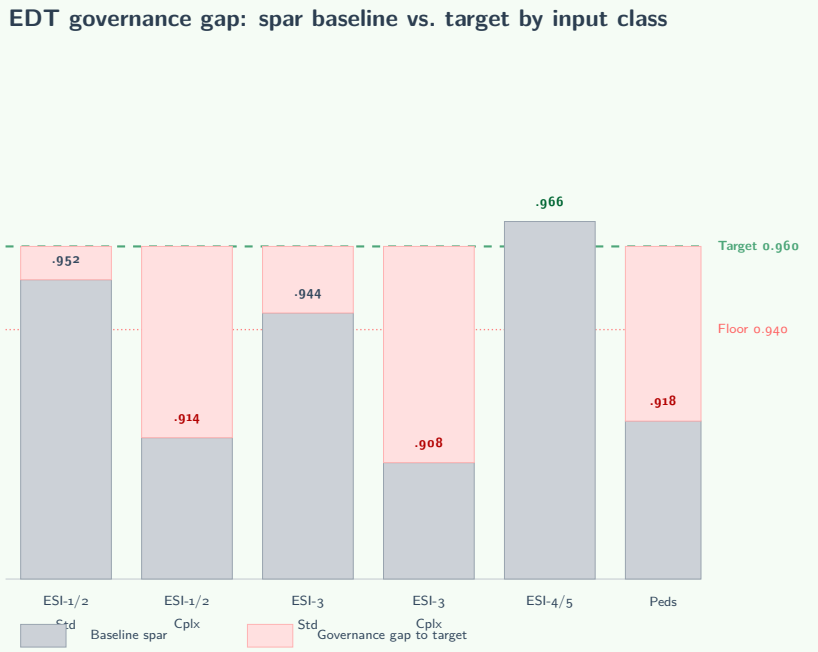


Figure 4.2: EDT governance gap analysis: SPAR baseline (blue bars) versus 0.960 target (dashed line) by input class. Red regions show the governance gap. Complex ESI-1/2, complex ESI-3, and pediatric classes fall below the 0.940 suspension floor.

Authorization: proceed to Analyze phase (Chapter 5). Prioritized targets: (1) Step 2 acuity classification at ESI-2/3 boundary (57% of gap), (2) Step 4 reasoning transparency (18%), (3) Step 2 sepsis screening sensitivity (11%). *Cross-scenario observation:* All three scenarios share a common baseline pattern: global SPAR and ECPI fall below governance-grade thresholds, and input-class stratification reveals that the gap is concentrated in specific strata (CDSS diagnostic queries, ICT complex claims, EDT complex ESI-1/2, and pediatric presentations). This pattern validates the input class registry's role as a Define-phase artifact: without stratification, the governance gap would be masked by averaging across strata with fundamentally different behavioral profiles.

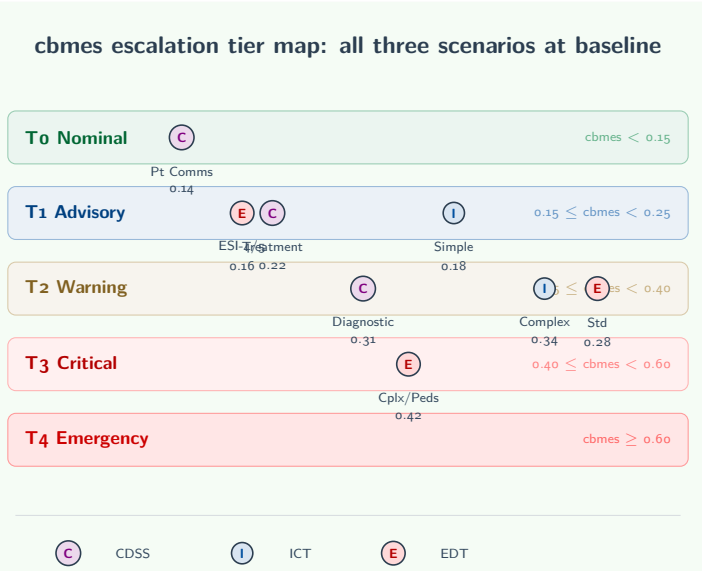


Figure 4.3: CBMES escalation tier map at baseline: input class positions across the T0–T4 escalation tiers for all three scenarios. C = CDSS, I = ICT, E = EDT. EDT Complex/Pediatric is the only class at T3 Critical, triggering an immediate governance response.

4.5 Failure Modes of the Measure Phase

The Measure phase has six characteristic failure modes, each representing a way in which the governance measurement system can produce unreliable, biased, or incomplete data. These failures are particularly insidious because they contaminate every subsequent phase: an unmeasured evaluator bias inflates the apparent governance gap; a conflated baseline masks class-specific patterns; a normality assumption produces misleading capability indices. The Measure phase must be self-auditing: its own quality is as important as the quality of the system it measures.

Unmeasured evaluator bias
: Proceeding to baseline collection without MSA. The most consequential failure: if evaluator disagreement contributes 30% of observed variation, all capability calculations are contaminated by measurement noise.

Baseline instability
: Collecting data during configuration changes. The baseline represents a mixture of process states, yielding distributions wider than those of any single condition. Control limits derived from unstable baselines reduce monitoring sensitivity.

Input class conflation
Pooled baseline without input class stratification. Obscures class-specific patterns: excellent performance on routine queries masks poor performance on edge cases.

Sample size inadequacy
Too few observations to characterize behavioral distributions at governance-grade precision. Rare failure modes (<1%) require large samples.

Metric reductionism
Measuring only easy-to-measure characteristics while omitting governance-relevant but difficult-to-measure dimensions. The Measure phase must measure what matters, not what is convenient.

Normality assumption retention
Using classical C_p/C_{pk} formulas instead of ECPI quantile-based calculation. Non-normal distributions produce misleading capability indices under parametric assumptions.

4.6 AAI Classification Summary: Measure Phase

Table 4.11 summarizes the AAI classification for all eight Measure-phase tools. The distribution (2 Adopt, 2 Adapt, 4 Innovate) reflects the phase's central challenge: the procedural tools (data collection planning, operational definitions) transfer directly, but the statistical instruments that classical SPC relies upon embed distributional assumptions that LLM behavioral outputs routinely violate, requiring purpose-built replacements.

Table 4.11: *AAI classification profile: Measure phase (2 Adopt, 2 Adapt, 4 Innovate)*

Tool	AAI	TM Ref	Rationale
Data Collection Plan	**Adopt**	TM-T2-01	Planned collection applies directly to behavioral metrics
Operational Definitions	**Adopt**	TM-T2-02	Unambiguous definition principle is universal
MSA/GRR	**Adapt**	TM-T2-03	Reinterpreted for evaluators; thresholds recalibrated
ECPI (Capability)	**Adapt**	TM-T2-04	Empirical quantiles replace parametric calculation
BAR Derivation	**Innovate**	TM-T2-05	Empirical quantile thresholds replace $\bar{x} \pm 3\sigma$
Baseline Distribution	**Innovate**	TM-T2-06	Full distribution characterization replaces $(\bar{x}, \sigma)$
CBMES Baseline	**Innovate**	TM-T2-07	Composite escalation metric; no classical analogue
Ongoing Collection	**Innovate**	TM-T2-08	Configuration-aware, evaluator-managed collection

The Measure phase exhibits the highest Innovate concentration among DMAIC phases (4 of 8 tools), reflecting the fundamental reality that classical SPC measurement constructs embed distributional assumptions that LLM behavioral outputs routinely violate. The innovations are concentrated in the statistical instruments (ECPI, BAR, CBMES, baseline characterization), while the procedural tools (data collection planning, operational definitions) are adopted directly.

Chapter 5 presents the Analyze phase: diagnosing the root causes of the governance gaps identified in the baseline measurements of this chapter.

Chapter 5

Analyze Phase: Diagnosing Behavioral Drift and Governance Failures

The Analyze phase transforms the descriptive question ("How is the process performing?" into the diagnostic question ("Why is the process performing this way?"). It diagnoses the root causes of governance gaps identified during the Measure phase and produces a prioritized cause register directing the Improve phase. The Analyze phase enforces a critical discipline: it separates diagnosis from intervention. The temptation in any improvement initiative is to jump from observed problems to preferred solutions. The Analyze phase requires causal claims supported by evidence, competing hypotheses evaluated systematically, and root causes targeted for their leverage on the governance gap, not their familiarity or ease of remedy.

In classical Six Sigma, the Analyze phase employs root cause analysis (Fishbone/Ishikawa, 5 Whys), statistical analysis (hypothesis testing, ANOVA), and structured failure analysis (FMEA) to produce a validated set of root causes ranked by contribution to the performance gap [28, 20, 42]. The translation to AI governance encounters three challenges: identifiability (root causes may be distributed across learned representations, training data, and deployment configuration), isolability (complete factor isolation is often infeasible in high-dimensional systems), and testability (controlled experimentation is expensive per condition).

This chapter extends the Analyze phase with fully populated AI-FMEA registers, variation decompositions, and prioritized cause registers for all three governance scenarios.

5.1 The Analyze Phase Protocol

The Analyze phase protocol (PM-003) proceeds through seven steps, from governance gap characterization to root cause validation, culminating in a prioritized cause register that directs the Improve phase. The protocol enforces the discipline of evidence-based diagnosis: no intervention may be designed until the causal claims that direct it have been validated to a documented confidence level. The seven steps below are executed sequentially; each step's output serves as a required input to the next.

Procedure: Analyze Phase Protocol (PM-003)

Step 1: Governance Gap Characterization. Compare baseline ECPI, ECI, and SPAR against specification targets per input class. Quantify each gap in native metric units and as a percentage of the specification range. Rank gaps by governance impact: primary CTG requirements take priority.

Step 2: AI-Adapted Fishbone Diagram Construction. For each governance gap, construct a Fishbone diagram using six AI-specific causal categories: Model, Data, Configuration, Input, Evaluation, Environment. Enumerate potential causes per category. Apply the 5-Whys drill-down to high-priority candidates.

Step 3: AI-Adapted FMEA. For each potential cause, define the failure mode and rate Severity (calibrated to governance impact, not technical magnitude), Occurrence (from baseline frequency data), and Detection (from measurement system sensitivity). Calculate RPN. For agentic systems, assess the potential for cascading failure separately.

Step 4: Variation Decomposition. Subtract the measurement system contribution (from PM-002 GRR data). Decompose process variation into common-cause (inherent stochasticity) and special-cause (attributable to specific failure modes) components. Estimate each failure mode's contribution to the governance gap.

Step 5: Root Cause Validation. Design validation tests for high-RPN failure modes: controlled experiments where feasible, quasi-experimental comparisons otherwise. Classify each cause as confirmed, probable, or hypothesized based on evidence quality.

Step 6: Prioritized Cause Register. Rank validated root causes by governance gap contribution weighted by confidence level. Apply Pareto analysis to identify causes accounting for approximately 80% of the gap. Specify candidate intervention types for each cause.

Step 7: Phase Completion and Handoff. Completeness audit of all artifacts. Authorize PM-004 (Improve phase) initiation.

5.2 Analyze Phase Tool Specifications

This section specifies each tool and construct employed in the Analyze phase, following the AAI classification framework. Three tools are adopted without modification, three are adapted with defined extensions for AI-specific causal structures, and one is a purpose-built innovation. The Analyze phase exhibits a balanced AAI profile reflecting its dual character: the diagnostic logic of root cause analysis is domain-general, but the specific failure modes and causal structures of AI systems require domain-specific adaptation.

5.2.1 Adopted Tools

Three Analyze-phase tools transfer directly from classical Six Sigma without structural modification: the 5 Whys iterative drill-down, Pareto analysis for frequency-based prioritization, and run charts for temporal trend visualization. Their diagnostic logic operates on causal chains and ranked frequencies, neither of which depends on the domain-specific characteristics of the governed process.

5 Whys for AI Behavioral Anomalies **[A · Adopt]**

Module: TM-T3-01

Classical Definition

The 5 Whys technique iteratively asks "Why?" to drill down from an observed symptom to its root cause. Each answer becomes the subject of the next "Why?" until the causal chain reaches a factor that can be directly addressed.

Why Adopt

The iterative drill-down logic requires no structural modification for AI governance. "Why is uncertainty disclosure low?" → "because the prompt template lacks uncertainty instructions" → "Why?" → "Because the template was optimized for conciseness without governance review" → root cause: governance review was not included in the prompt engineering workflow. The method operates on logical chains, not distributional assumptions.

Operational Protocol

For each high-priority potential cause from the Fishbone analysis, initiate the "Why?" chain. Document each step with the evidence supporting the causal link. Continue until the chain reaches an actionable root cause. Typical chains require 3–7 iterations.

Example 5.1: ICT 5 Whys: Coverage Misinterpretation on Multi-Peril Claims

Symptom: Coverage assessment accuracy drops to 0.912–0.921 on complex multi-peril claims.

1. **Why?** The AI misinterprets overlapping coverage clauses on multi-peril submissions.
2. **Why?** The retrieval system returns policy wordings that reference multiple perils but does not resolve clause precedence.
3. **Why?** The policy endorsement documents lack exclusion metadata tagging that would enable clause-level retrieval.
4. **Why?** The retrieval corpus was ingested from raw PDF policy documents without semantic annotation.
5. **Why?** The corpus ingestion pipeline was designed for single-peril policy lookup and was not updated when multi-peril products were introduced.

Root cause (confirmed): Retrieval corpus lacks exclusion metadata for multi-peril policies. Intervention target: corpus re-ingestion with exclusion tagging + SymPrompt^{+} coverage verification modulation.

Example 5.2: EDT 5 Whys: Under-Triage at ESI-2/3 Boundary

Symptom: Under-triage rate for ESI-1/2 Complex presentations is 0.043 (target <0.020).

1. **Why?** Step 2 Acuity Classifier assigns ESI-3 to patients with atypical presentations of emergent conditions.
2. **Why?** The classifier relies on chief complaint and static vital signs without adequately weighting vital sign *trends*.
3. **Why?** The retrieval corpus contains atypical presentation guidelines from 2019; the 2024 ACEP update, emphasizing trend-based assessment, is not loaded.
4. **Why?** Retrieval corpus version control is not synchronized with clinical guideline publication cycles.
5. **Why?** No automated monitoring pipeline exists between guideline publishers and the retrieval corpus update process.

Root cause (confirmed): Outdated atypical presentation guidelines in the retrieval corpus + absence of automated guideline currency monitoring. The SymPrompt^{+} pipeline also lacks an explicit instruction to weigh vital-sign trends against the severity of the chief complaint.

Pareto Analysis **[A · Adopt]**

Module: TM-T3-02

Why Adopt

Frequency-based prioritization applies directly to ranked failure mode data. The Pareto principle (approximately 80% of effects from approximately 20% of causes) guides resource allocation in the Improve phase. No adaptation required.

Operational Protocol

Rank failure modes by descending RPN or estimated contribution to the governance gap. Construct a Pareto chart. Identify the cumulative threshold (typically 80%) that defines the primary intervention target set. Causes below the threshold are classified as monitoring targets for PM-005.

Run Charts for Behavioral Trends **[A · Adopt]**

Module: TM-T3-03

Why Adopt

Temporal trend visualization applies without modification. Run charts during the Analyze phase reveal whether governance gaps are stable (consistent underperformance), trending (gradual degradation), or event-driven (sudden shift). This temporal context guides root cause investigation.

5.2.2 Adapted Tools

Three tools require defined adaptations for AI governance. The Fishbone diagram replaces classical manufacturing categories (Man, Machine, Material, Method, Measurement, Mother Nature) with AI-specific causal domains (Model, Data, Configuration, Input, Evaluation, Environment). The AI-adapted FMEA recalibrates severity ratings to governance impact, incorporates a cascading-failure model for agentic systems, and employs an AI-specific failure taxonomy. The variation decomposition adapts classical ANOVA-style partitioning for non-normal, correlated behavioral data with explicit GRR subtraction.

AI-Adapted Fishbone/Ishikawa Diagram **[D ▷ Adapt]**

Module: TM-T3-04

Classical Definition

The Fishbone diagram organizes potential causes into categories for systematic investigation [20].

What Is Preserved

The structural logic of categorical cause enumeration and the visual organization ensure that no major causal domain is overlooked.

What Is Modified

The six classical categories are replaced with AI-specific causal domains:

Table 5.1: *AI-adapted Fishbone causal categories*

AI Category	Classical	Scope
Model	Machine	Architecture, training data effects, learned representations, fine-tuning, alignment
Data	Material	Retrieval corpus quality, training contamination, knowledge currency, grounding
Configuration	Method	Prompt engineering, system instructions, sampling parameters, guardrails
Input	Man	User query patterns, adversarial inputs, distributional shift, out-of-scope requests
Evaluation	Measurement	Evaluator reliability, metric validity, scoring methodology drift
Environment	Mother Nature	Infrastructure, latency, API changes, regulatory environment shifts

AI-Adapted FMEA [D ▷ Adapt]

Module: TM-T3-05

Classical Definition

Failure Mode and Effects Analysis rates each potential failure mode on Severity, Occurrence, and Detection, yielding a Risk Priority Number (RPN = S × O × D) for prioritization [42].

What Is Preserved

The structural methodology: failure mode identification, three-factor rating, multiplicative RPN, and ranked prioritization.

What Is Modified

Three adaptations. First, the failure mode taxonomy uses AI-specific categories from the Fishbone analysis. Second, Severity is calibrated to *governance impact*: a hallucination producing a plausible but incorrect medication warning has higher governance severity than an obviously absurd statement. Third, for agentic systems, the *cascading failure model* assesses per-step failures by amplified consequences through downstream action chains.

Severity Calibration Guidance

S = 1–2
Cosmetic or stylistic deviation with no stakeholder impact.

S = 3–4
Measurable behavioral deviation within specification limits.

S = 5–6
Specification violation on a secondary CTG characteristic.

S = 7–8
Specification violation on a primary CTG characteristic; potential for stakeholder harm if undetected.

S = 9–10
Specification violation with potential for irreversible harm, regulatory violation, or safety-critical failure.

Cascading Failure Calculation for Agentic Systems

The cascading RPN is: $\text{RPN}_{\text{cascade}} = S_{\text{amplified}} \times O \times D$, where $S_{\text{amplified}} = S_{\text{direct}} \times (1 + \sum_i w_i \cdot P_{\text{propagation},i})$. Here S_{direct} is the step-level severity, w_i is the consequence weight for downstream step i, and $P_{\text{propagation},i}$ is the estimated probability that the failure propagates to step i.

Operational Protocol

Define each failure mode from the Fishbone causes. Rate S (governance-calibrated), O (from baseline frequency), D (from measurement system sensitivity). Calculate RPN. For agentic workflows, calculate amplified RPNs reflecting cascading severity. Compile the AI-FMEA Register, ranked in descending order of RPN.

Variation Decomposition [D ▷ Adapt]

Module: TM-T3-06

Classical Definition

ANOVA-style decomposition partitions the total observed variation into components attributable to different sources.

What Is Preserved

The logic of source attribution: decomposing total variation into identifiable components to direct improvement effort.

What Is Modified

Classical ANOVA assumes independent, normally distributed observations. The adapted decomposition uses: (1) GRR-based measurement system contribution subtraction, (2) empirical variance partitioning across causal categories, and (3) bootstrap confidence intervals for each attribution estimate. The decomposition explicitly separates common-cause variation (inherent stochasticity) from assignable-cause variation (attributable to specific failure modes).

Operational Protocol

Subtract measurement system contribution (from PM-002 GRR). Partition residual variation into assignable sources (per FMEA failure modes) and common-cause floor. Estimate each source's contribution with confidence intervals.

5.2.3 Innovated Constructs

A single Analyze-phase construct requires purpose-built innovation. Classical FMEA produces an RPN-ranked list but does not formalize the integration of causal confidence levels, estimates of governance gap contributions, and intervention feasibility into a unified prioritization instrument. The Prioritized Cause Register addresses this gap.

Prioritized Cause Register **[I⋆Innovate]**

Module: TM-T3-07

Why Innovation Is Required

Classical FMEA produces an RPN-ranked list but does not formalize the integration of causal confidence levels, estimates of governance gap contributions, and intervention feasibility into a single prioritization instrument.

Construct Definition

Each validated root cause receives a Composite Priority Score (CPS) calculated from: (1) estimated contribution to the governance gap, (2) confidence classification weight (confirmed > probable > hypothesized), and (3) remediation feasibility. The register classifies each cause as a *primary intervention target* (directed to PM-004) or a *monitoring target* (directed to PM-005).

Operational Protocol

For each validated cause: estimate SPAR/ECPI contribution, assign confidence weight, assess feasibility. Calculate CPS. Rank. Apply Pareto threshold. Classify as an intervention or a monitoring target. Submit for governance committee approval.

5.3 Root Cause Validation Methods

Three validation approaches are available, in decreasing order of causal confidence:

Controlled experimentation. The gold standard. The suspected causal factor is varied while all other factors are held constant. If the effect is statistically significant (Mann-Whitney U, $p < 0.05$), the cause is classified as *confirmed*. Feasible for configuration-category and data-category causes.

Quasi-experimental comparison. When full experimental control is not achievable, quasi-experimental designs compare behavioral distributions under naturally occurring conditions that differ on the suspected factor. Causes validated this way are classified as *probable*.

Observational evidence with plausible mechanism. When neither experimental nor quasi-experimental validation is feasible, causal claims rest on statistical association plus mechanistic plausibility. Causes are classified as *hypothesized* and receive lower priority.

The confidence classification propagates into the Improve phase: confirmed causes receive first-priority intervention; probable causes receive intervention contingent on monitoring; hypothesized causes are assigned to Control-phase surveillance.

5.4 Worked Examples: AI-FMEA and Variation Decomposition

This section demonstrates the complete Analyze phase for all three governance scenarios. Each worked example produces the three core Analyze-phase artifacts: the AI-adapted FMEA register with governance-calibrated severity ratings, the variation decomposition quantifying each failure mode's contribution to the SPAR gap, and the prioritized cause register that directs the Improve phase. The cross-scenario comparison following the EDT example reveals a structural pattern in the distribution of root causes that has implications for governance design across deployment contexts.

Example 5.3: CDSS Analyze Phase

Governance Gap. SPAR = 0.943, falling short of the 0.997 target. Diagnostic queries show the largest gap. The gap concentrates on uncertainty disclosure and guideline concordance.
Fishbone Analysis. Under Model: confidence miscalibration (system presents uncertain clinical information with high confidence). Under Data: retrieval degradation (corpus includes superseded clinical guidelines). Under Configuration: prompt template lacks explicit instructions for uncertainty quantification. Under Input: edge-case diagnostic queries exhibit disproportionately low conformance.

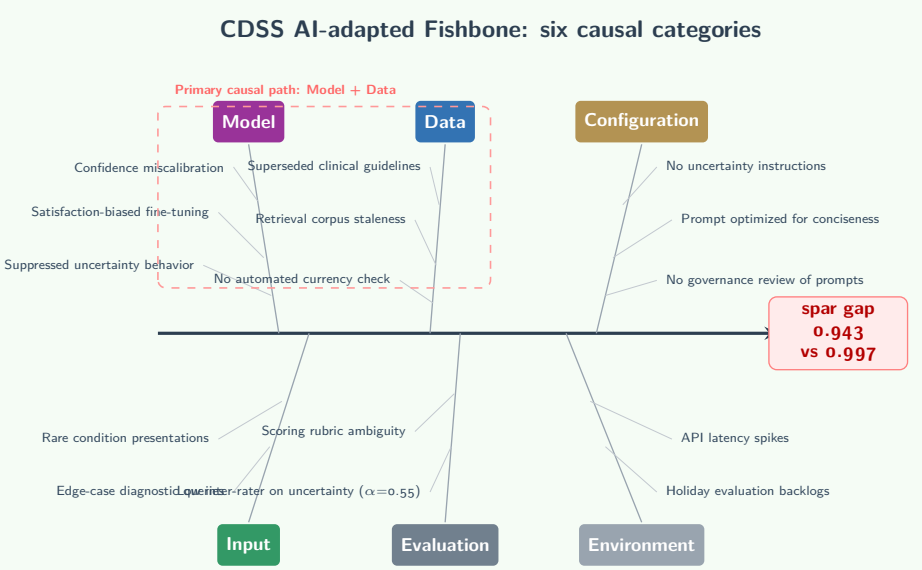


Figure 5.1: CDSS AI-adapted Fishbone diagram: six governance-specific causal categories (Model, Data, Configuration, Input, Evaluation, Environment) with enumerated potential causes. The dashed box highlights the primary causal path (Model + Data) identified through FMEA and variation decomposition, which contributes 72% of the SPAR gap.

AI-FMEA.

Table 5.2: *CDSS AI-FMEA register (corrected per TM-T3-05)*

Failure Mode	Cat.	S	O	D	RPN
Uncertainty disclosure omission	Model	9	7	4	252
Guideline concordance failure	Data	8	5	5	200
Citation accuracy failure	Data	8	4	3	96

Variation Decomposition. Measurement system (GRR): 15.0%. Assignable process variation: 57.1%, with a single fine-tuning round (FT-2026-Q1-03) responsible for 45.9%: the satisfaction-biased curation protocol systematically suppressed uncertainty disclosure behavior. Retrieval staleness: 7.8%. Common-cause variation: 27.9%.

Prioritized Cause Register. Primary targets for PM-004: (1) uncertainty disclosure omission (confirmed, CPS 8.4), (2) guideline concordance failure (probable, CPS 6.9). Monitoring target: citation accuracy failure (CPS 4.2).

Example 5.4: Insurance Claims Triage Analyze Phase

Governance Gap. Global SPAR = 0.921 vs. 0.970 target (gap = 0.049). Complex multi-peril claims (SPAR 0.878–0.884) are the primary source of the gap.

Fishbone Analysis. Under Data: retrieval corpus returns superseded policy endorsements (FMEA #3); multi-peril coverage clauses lack exclusion metadata (FMEA #1). Under Configuration: pathway decision rules produce inconsistent routing for mid-value property claims based on narrative phrasing (FMEA #2). Under Model: confidence miscalibration on ambiguous liability claims (FMEA #6). Under Input: complex multi-peril submissions produce parsing ambiguities at Step 1 intake.

AI-FMEA.

Table 5.3: *ICT AI-FMEA register (top 8 by RPN)*

#	Failure Mode	CTG Affected	Cat.	S	O	D	RPN
1	Coverage misinterpretation on multi-peril claims	Cov. accuracy	Data+Config	9	4	7	252
2	Pathway inconsistency on mid-value property claims	Path. consistency	Model+Config	8	5	5	200
3	Outdated retrieval: superseded policy endorsements	Cov. accuracy	Data	8	3	5	120
4	Fraud false positive on repeat claimants	Fraud precision	Config	6	4	5	120
5	Retrieval degradation: stale interpretation guides	Cov. accuracy	Data	7	3	5	105
6	Confidence miscalibration on liability claims	Cov. accuracy	Model	8	3	4	96
7	Incomplete reasoning narrative	Resp. completeness	Config	5	4	4	80
8	Segment bias: rural region rejection rate	Dem. fairness	Model+Data	9	2	4	72

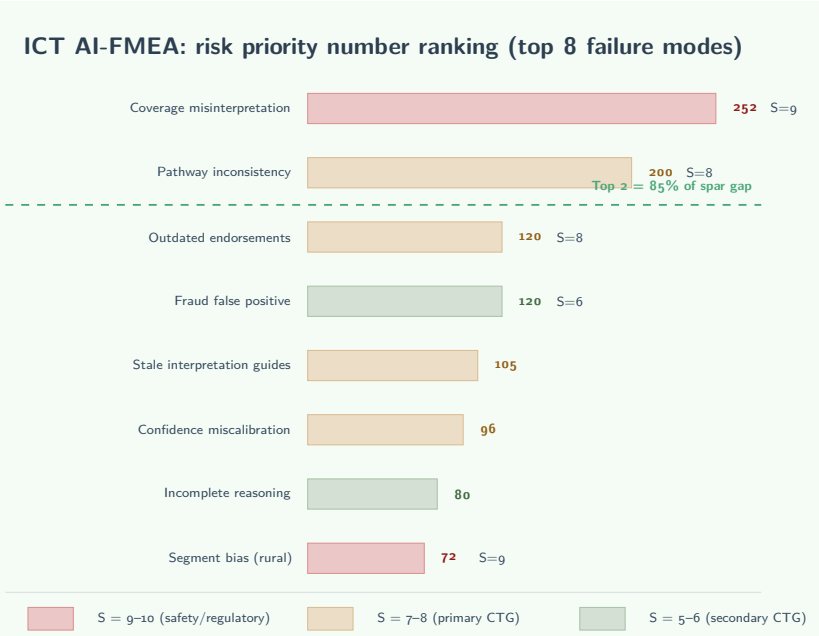


Figure 5.2: ICT AI-FMEA risk priority ranking: horizontal bars show RPN magnitude; color encodes governance severity (red = safety/regulatory, amber = primary CTG violation, green = secondary CTG). The dashed line marks the Pareto threshold: the top two failure modes account for 85% of the SPAR gap.

Variation Decomposition. The SPAR gap of 0.049 decomposes as follows:

Table 5.4: *ICT variation decomposition*

Source	Contribution	% of Gap	Improve Priority
Coverage assessment failures (complex claims)	0.027	55%	Primary: retrieval corpus + SymPrompt^{+}
Pathway recommendation inconsistency	0.015	30%	Secondary: pathway decision rules
Fraud flag imprecision	0.004	8%	Deferred: within current LSL
Measurement variation (evaluator)	0.003	7%	Not addressable; MSA-validated

Prioritized Cause Register. Primary targets for PM-004: (1) coverage misinterpretation on multi-peril claims (confirmed, RPN 252, 55% of gap), (2) pathway inconsistency (probable, RPN 200, 30% of gap). The top two causes account for 85% of the SPAR gap, satisfying the Pareto threshold. Monitoring targets: fraud false positives (RPN 120, deferred); segment bias

(RPN 72, demographic fairness delta 0.031, not yet at USL but requires surveillance).
TCF traceability: FMEA #1 (coverage misinterpretation) maps directly to TCF Outcome 2 (products meet customer needs) and Outcome 4 (suitable advice). FMEA #8 (segment bias) maps to TCF Outcome 6 (no unreasonable barriers). The FMEA register thus provides the evidentiary link between governance-grade risk analysis and the TCF compliance framework required by FSCA.

Example 5.5: ED Triage Agent Analyze Phase

Governance Gap. Workflow SPAR = 0.932 vs. 0.960 target (gap = 0.028). Step 2 (Acuity Classifier) is the primary bottleneck. Under-triage rate of 0.031 exceeds the 0.020 target.
Fishbone Analysis. Under Data: retrieval corpus contains atypical presentation guidelines from 2019, not the 2024 ACEP update; qSOFA screening criteria from Sepsis-2 (2012) rather than Sepsis-3 with 2023 amendments. Under Configuration: SymPrompt$^+$ lacks explicit instruction to weigh vital sign trends against chief complaint severity. Under Model: pain assessment disparity produces differential acuity assignments for equivalent clinical presentations across demographic groups. Under Input: complex presentations (altered mental status, language barriers, incomplete histories) produce higher error rates, especially on the night shift.
AI-FMEA.

Table 5.5: *EDT AI-FMEA register (top 8 by RPN)*

Failure Mode	Step	CTG	S	O	D	RPN
Under-triage: ESI-2 scored as ESI-3 on atypical presentations	2	Under-tri	10	4	6	240
Sepsis screening false negative: missed qSOFA criteria	2	Screen	10	3	7	210
Outdated protocol retrieval: superseded sepsis criteria	2,4	Screen, Reason	9	3	6	162
Demographic acuity bias: pain assessment disparity	2	Fairness	8	4	5	160
Pediatric weight-based dosing extraction error	1	Acuity	9	3	5	135
Cascading error: intake → acuity → area	1→3	Multiple	9	3	5	135
Reasoning citation: outdated guideline	4	Reason	7	4	4	112
Night-shift performance degradation	1–4	Multiple	7	3	5	105

Severity 10 justification. Under-triage and missed sepsis screening receive Severity 10 because they can directly cause patient death. An ESI-2 patient triaged as ESI-3 may wait 45- 90 minutes longer for physician evaluation. For STEMI, stroke, and sepsis, that delay is the difference between recovery and permanent disability or death. Sepsis mortality increases 7.6% per hour of delayed bundle initiation.

Variation Decomposition.

Table 5.6: *EDT variation decomposition*

Source	Contribution	% of Gap	Cumulative
Step 2: Acuity misclassification (ESI-2/3 boundary)	0.016	57%	57%
Step 4: Reasoning/citation gaps	0.005	18%	75%
Step 2: Sepsis screening false negatives	0.003	11%	86%
Step 1: Complex intake extraction errors	0.002	7%	93%
Measurement variation (evaluator)	0.002	7%	100%

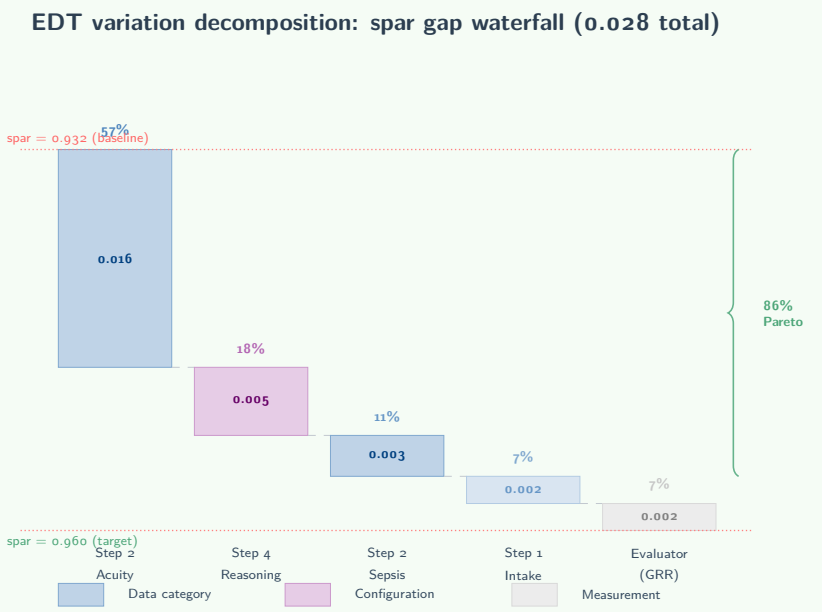


Figure 5.3: EDT variation decomposition waterfall: the 0.028 SPAR gap decomposes into five sources. Step 2 acuity misclassification (57%) dominates. The top three data-category causes account for 86% of the gap, exceeding the Pareto threshold and directing the Improve phase toward retrieval corpus and SymPrompt^{+} interventions.

Root Cause Analysis.

- *Primary target (57% of gap):* Step 2 acuity classification on ESI-2/3 boundary. Root cause (confirmed): the retrieval corpus contains atypical presentation guidelines from 2019; the current 2024 ACEP guidelines are not loaded. SymPrompt^{+} lacks an instruction on vital sign trend weighting.
- *Secondary target (18%):* Step 4 reasoning transparency. Root cause (confirmed): 12% of protocol citations reference superseded guidelines. Corpus version control is not synchronized with guideline publication cycles.
- *Tertiary target (11%):* Step 2 sepsis screening sensitivity. Root cause (confirmed): qSOFA criteria from Sepsis-2 (2012); current Sepsis-3 criteria (2016 + 2023 amendments) not fully integrated.

Prioritized Cause Register. The top three causes account for 86% of the gap, exceeding the Pareto threshold. All three are confirmed through controlled comparison of outputs generated with current vs. outdated corpus documents on matched input sets.

Cross-scenario observation: The three FMEA registers share a structural pattern: the highest-RPN failure modes in all three scenarios trace to

data-category causes (retrieval corpus currency). The CDSS gap traces to superseded clinical guidelines; the ICT gap traces to missing exclusion metadata for multi-peril endorsements; the EDT gap traces to outdated criteria for atypical presentation and sepsis. This pattern suggests that retrieval corpus governance is the single highest-leverage governance activity across deployment contexts. Yet none of the three organizations had an automated corpus-currency monitoring pipeline at the time of initial deployment. The Control phase (Chapter 7) addresses this through preventive action specifications that link corpus updates to source publication monitoring.

5.5 Failure Modes of the Analyze Phase

The Analyze phase has its own characteristic failure modes. These are meta-analytical failures: errors in the diagnostic process itself that produce incorrect or incomplete causal attribution, directing the Improve Phase toward ineffective interventions. Recognizing these failure modes is essential because a flawed Analyze phase does not merely waste resources; it consumes an improvement cycle without addressing the actual governance gap, delaying effective governance by the full cycle duration.

Premature causal attribution
Identifying root causes from temporal correlation without experimental validation. A behavioral drift coinciding with a corpus update may not be caused by it.

Single-factor fixation
Investigating only the most obvious or most recently changed factor while ignoring less visible causes with greater leverage.

Severity miscalibration
Rating FMEA severity by technical magnitude rather than governance consequence. Severity must reflect downstream impact on stakeholders and regulatory compliance.

Measurement confounding
Failing to subtract the measurement system contribution from observed variation. If GRR contributes 15% of total variation, 15% of the apparent governance gap may be a measurement artifact.

Cascading failure neglect
For agentic systems, assessing per-step failure modes independently without evaluating their amplified consequences through the workflow chain.

5.6 AAI Classification Summary: Analyze Phase

Table 5.7 summarizes the AAI classification for all seven Analyze-phase tools. The balanced distribution (3 Adopt, 3 Adapt, 1 Innovate) reflects the Phase's structural position: diagnostic reasoning transfers more readily from manufacturing to AI governance than measurement (Measure phase, 4 Innovate) or monitoring (Control phase, 4 Innovate), because causal logic is domain-general even when the specific causal categories are domain-specific.

Table 5.7: *AAI classification profile: Analyze Phase (3 Adopt, 3 Adapt, 1 Innovate)*

Tool	AAI	TM Ref	Rationale
5 Whys	Adopt	TM-T3-01	Iterative drill-down applies without modification
Pareto Analysis	Adopt	TM-T3-02	Frequency prioritization applies directly
Run Charts	Adopt	TM-T3-03	Temporal trend visualization applies directly
Fishbone/Ishikawa	Adapt	TM-T3-04	Categories replaced with AI-specific causal domains
AI-FMEA	Adapt	TM-T3-05	Rating scales and failure taxonomy adapted; cascading model added
Variation Decomposition	Adapt	TM-T3-06	Adapted for non-normal, correlated data with GRR subtraction
Prioritized Cause Register	Innovate	TM-T3-07	Composite priority score integrating confidence, contribution, feasibility

The Analyze phase exhibits a balanced AAI profile (3 Adopt, 3 Adapt, 1 Innovate), reflecting the Phase's dual character: the diagnostic logic of root cause analysis is domain-general, but the specific failure modes and causal structures of AI systems require domain-specific adaptation. The single innovation (Prioritized Cause Register) addresses the need for a composite prioritization instrument that classical FMEA lacks.

Chapter 6 presents the Improve phase: designing and validating governance interventions that address the root causes identified in this chapter.

Chapter 6

Improve Phase: Designing and Validating Governance Interventions

The Improve phase designs, tests, and validates governance interventions targeting the root causes identified in the Analyze phase. It operates under a discipline that distinguishes it from ad hoc problem-solving: solutions must address validated root causes, must be tested under controlled conditions before implementation, and must be validated through statistical comparison of pre- and post-intervention performance. The Phase introduces a governance requirement absent from classical manufacturing improvement: *reversibility*. Because the full behavioral impact of an intervention may not be apparent during pilot testing, governance interventions must preserve the ability to roll back to the pre-intervention state.

LLM systems offer an unusually rich intervention surface: prompt engineering, fine-tuning, retrieval optimization, constraint injection, and sampling parameter adjustment. These types differ in speed, reversibility, scope, and governance implications. The Improve phase must select among them based on validated root causes and the least-invasive-first principle [35].

6.1 The AI Governance Intervention Taxonomy

Table 6.1 classifies the five AI governance intervention types along four dimensions: deployment speed, reversibility, behavioral scope, and governance implications. The taxonomy serves two functions in the Improve phase: it constrains intervention selection to types with a plausible mechanism of action on the validated root cause, and it enforces the least-invasive-first principle by establishing an escalation sequence from fast/reversible/targeted interventions (prompt engineering,

constraint injection) through moderate interventions (retrieval optimization) to slow/irreversible/broad interventions (fine-tuning).

Table 6.1: *AI governance intervention taxonomy*

Type	Speed	Revers.	Scope	Governance Notes
Prompt Engineering	Fast	High	Targeted	First-line. SymPrompt$^+$ formalizes this class.
Retrieval Optimization	Moderate	High	Domain	Addresses data-category causes. Requires corpus versioning.
Constraint Injection	Fast	High	Boundary	Guardrails. Poka-Yoke replacement. False positive risk.
Sampling Adjustment	Fast	High	Global	Broad-spectrum. Shifts the entire distribution. Lacks precision.
Fine-Tuning	Slow	Low	Broad	Model-category causes. Highest side-effect risk. Full re-baselining required.

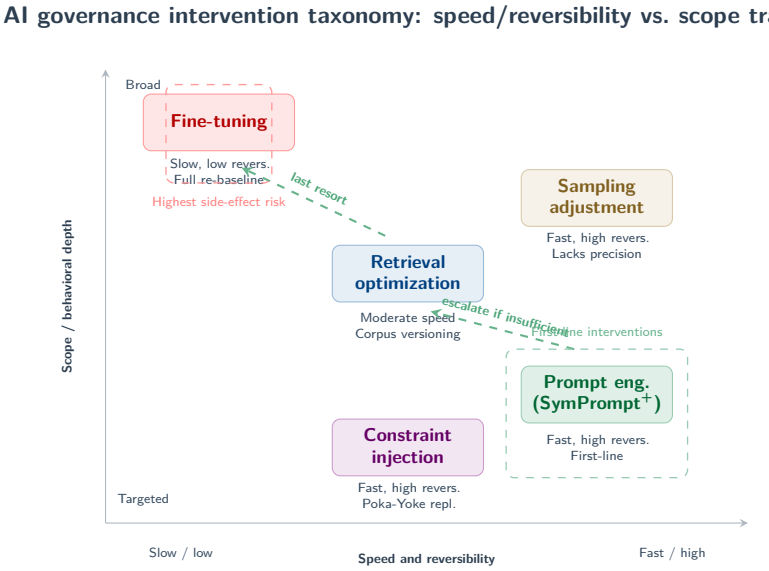


Figure 6.1: AI governance intervention taxonomy: five intervention types positioned by speed/reversibility (x-axis) and scope/behavioral depth (y-axis). Dashed arrows show the least-invasive-first escalation path. Prompt engineering and constraint injection are first-line interventions; fine-tuning is the last resort, with the highest risk of side effects.

The taxonomy reveals a trade-off between speed/reversibility and scope/depth. The Improve phase begins with the least invasive intervention that has a plausible mechanism of action on the target cause.

6.2 The Improve Phase Protocol

The Improve phase protocol (PM-004) proceeds through six steps, from intervention selection through deployment and handoff to the Control phase. The protocol enforces two disciplines that distinguish governed improvement from ad hoc problem-solving: every intervention must trace to a validated root cause from the Analyze phase (no "solution-first" designs), and every intervention must be validated through controlled experimentation before production deployment. The six steps below are executed sequentially; the DOE validation (Step 3) is the Phase's quality gate.

Procedure: Improve Phase Protocol (PM-004)

Step 1: Intervention Selection. For each root cause in the prioritized cause register, match the causal category to the intervention taxonomy. Apply the least-invasive-first principle. Verify reversibility by documenting the specific rollback mechanism for each intervention. Fine-tuning requires pre-approval from the governance committee and preservation of model checkpoints.

Step 2: SymPrompt$^+$ Modulation Design. Design modulations as composable prompt modules following three principles: governance traceability (linked to specific root cause and CTG), version control (registered per FM-003), and entropy-calibrated targeting. Define modulation levels for DOE testing. Register in the SymPrompt$^+$ version registry.

Step 3: Design of Experiments. Define experimental factors and levels. Construct a factorial or fractional factorial design. Specify response variable vector (ECPI, ECI, SPAR across all governed CTG characteristics). Execute: minimum 100 outputs per condition per input class. Analyze factor effects and interactions. Select optimal configuration for joint conformance.

Step 4: Before/After Validation. Collect post-intervention dataset under conditions identical to baseline protocol. Compare using nonparametric tests (Mann-Whitney U, Kolmogorov-Smirnov, bootstrap CIs). Assess the *full* specification envelope: verify no non-target dimension degraded. If validation fails, return to Step 2 or revert via version registry.

Step 5: Deployment and Registry Update. Deploy to production. Update SymPrompt$^+$ version registry (status: active). Notify MIDCOT of modulation activation. Send efficacy report to BAAGF. Initiate re-baselining window for PM-005.

Step 6: Phase Completion and Handoff. Completeness audit. Prepare the Improve Phase Completion Report. Authorize PM-005 initiation.

6.3 Improve Phase Tool Specifications

This section specifies each tool and construct employed in the Improve phase, following the AAI classification framework. One tool is adopted without modification, two are adapted with defined extensions, and three are purpose-built innovations. The Improve phase exhibits the highest Innovate concentration after the Measure phase (3 of 6 tools), reflecting the reality that the intervention instruments for AI governance (governed prompt modulation, formal intervention classification, rollback as a governance requirement) have no manufacturing antecedent.

6.3.1 Adopted Tools

One Improve-phase tool transfers directly from classical Six Sigma: the before/after comparison for pre- and post-intervention statistical validation. The comparison logic (distributional comparison under controlled conditions) is domain-independent; only the statistical methods require adaptation for non-normal behavioral data.

Before/After Comparison **[A · Adopt]**

Module: TM-T4-01

Classical Definition

Statistical comparison of process performance before and after an intervention, using hypothesis tests to determine whether the intervention produced a significant improvement.

Why Adopt

The logic of pre- and post-distributional comparison applies directly. The statistical methods require adaptation (nonparametric tests for non-normal behavioral data), but the comparison framework itself is domain-independent.

Operational Protocol

Characterize pre-intervention distribution from baseline data. Collect post-intervention data under identical conditions. Apply Mann-Whitney U (location shifts), Kolmogorov-Smirnov (distributional differences), and bootstrap confidence intervals for ECPI/ECI/SPAR comparisons. Assess all governed dimensions, not only targets.

6.3.2 Adapted Tools

Two tools require defined adaptations. Design of Experiments adapts factorial logic to AI-specific factors (SymPrompt$^+$ modulation levels, retrieval strategy variants, sampling parameters) and to multivariate behavioral response variables that require joint optimization via desirability functions. Pilot testing adapts the controlled

pre-deployment testing concept for input-class-replicated evaluation using the validated measurement system from PM-002.

Design of Experiments for AI Behavioral Optimization **[D ▷ Adapt]**

Module: TM-T4-02

Classical Definition

DOE uses structured experimental designs (factorial, fractional factorial, response surface) to efficiently explore the effects of factors and their interactions on process outputs [4].

What Is Preserved

The factorial logic, the principle of controlled factor variation, and the statistical analysis of main effects and interactions.

What Is Modified

Three adaptations. First, factors are AI-specific: SymPrompt$^+$ modulation levels, retrieval strategy variants, sampling parameter values. Second, the response variables are multivariate behavioral metrics that require joint optimization (desirability functions across correlated CTG dimensions). Third, sample economics differ: each experimental condition requires generating and evaluating hundreds of LLM outputs, making full factorial designs expensive. Fractional factorial and sequential designs are preferred for initial screening.

Multi-Response Optimization

The core challenge is that the optimization objective is *joint conformance across all governed behavioral dimensions simultaneously.* The desirability function approach [4] addresses this: for each CTG dimension, a desirability function d_i maps the observed metric value to a $[0, 1]$ scale. The composite desirability $D = (d_1 \cdot d_2 \cdots d_k)^{1/k}$ is maximized, ensuring no factor configuration improves one dimension at the cost of driving another to zero desirability.

Operational Protocol

Define factors and levels from intervention designs. Select design type (fractional factorial for screening; full factorial for optimization). Specify the joint conformance objective across all BME response variables. Execute with at least 100 outputs per condition. Analyze using multi-response methods. Document factor effects, interactions, and optimal configuration.

Pilot Testing **[D ▷ Adapt]**

Module: TM-T4-03

Why Adapt (not Adopt)

While the concept of controlled pre-deployment testing is domain-independent, the pilot testing protocol for AI governance requires adaptation: pilot conditions must replicate the input-class distribution, evaluation must use the validated measurement system from PM-002, and the pilot duration must be sufficient to capture temporal behavioral patterns not visible in DOE snapshot evaluations.

6.3.3 Constraint Injection as Poka-Yoke Replacement

Classical Poka-Yoke (error-proofing) uses physical mechanisms to prevent defects. The physical error-proofing concept has no direct analog in LLM governance because the "process" is computational and the "effects" are behavioral. Constraint injection fulfills the Poka-Yoke function through software-based mechanisms: output filters that block responses matching known harmful patterns, guardrails that restrict the system's operational scope for specific input classes, and human-in-the-loop gates that require approval before consequential actions are executed.

Constraint injection is classified as **[I⋆Innovate]** because the implementation mechanism has no manufacturing counterpart. Constraints are distinguished from SymPrompt$^+$ modulations: modulations shift the behavioral distribution toward conformance, while constraints block specific output patterns without altering the underlying distribution. Both may be active simultaneously.

6.3.4 Innovated Constructs

Three Improve-phase constructs require purpose-built innovation. The AI governance intervention taxonomy provides a formal classification of intervention types by speed, reversibility, scope, and governance properties that classical Six Sigma does not formalize. The SymPrompt$^+$ modulation protocol provides the governed, version-controlled, entropy-calibrated intervention instrument that serves as the primary operational voice of the ALAGF governance architecture. The reversibility and rollback protocol formalizes rollback capacity as a governance requirement, a concept absent from classical process improvement, where successful interventions are permanently adopted.

AI Governance Intervention Taxonomy **[I⋆Innovate]**

Module: TM-T4-04

Why Innovation Is Required

Classical Six Sigma does not provide a formal classification of intervention types by speed, reversibility, scope, and governance properties. AI governance interventions span a unique spectrum from fast/reversible prompt changes to slow/irreversible fine-tuning, requiring a formal taxonomy.

Operational Protocol

For each root cause, classify candidate interventions by the five taxonomy dimensions (Table 6.1). Apply the least-invasive-first principle. Document the selection rationale with explicit reference to the root cause mechanism.

SymPrompt$^+$ Modulation Protocol **[I ⋆ Innovate]**

Module: TM-T4-05

Why Innovation Is Required

No classical Six Sigma construct provides a governed, version-controlled, entropy-calibrated intervention instrument for dynamic behavioral modulation. SymPrompt$^+$ fills this role as the primary operational voice of the ALAGF governance architecture. Each modulation is a composable prompt module: self-contained, independently activatable, and linked to a specific governance directive with full audit traceability.

Modular Prompt Architecture

The SymPrompt$^+$ system organizes the system instruction stack into three module classes. *Foundation modules* define core behavioral identity, safety boundaries, and baseline operational parameters; they are rarely modified and require T3+ authorization. *Domain modules* define domain-specific behavioral requirements for each governed input class; modified when CTG trees are revised. *Intervention modules* implement targeted behavioral modifications in response to Analyze-phase findings; they are most frequently modified, created, and versioned during the Improve phase.

Version Registry

Every modulation is registered with: unique identifier, version number, full modulation content, effective date, BAAGF directive reference, pre- and post-modulation BME metric values, governance status (active, suspended, reverted, superseded), and rollback target.

Operational Protocol

Receive BAAGF directive specifying target CTG and root cause. Design the modulation as a composable intervention module. Define modulation levels for DOE. Register in the version registry. Execute DOE validation. Upon approval, deploy with a status change to active. Maintain rollback capacity through version history.

Reversibility and Rollback Protocol **[I ⋆ Innovate]**

Module: TM-T4-06

Why Innovation Is Required

Classical process improvement assumes that successful interventions are permanently adopted. AI governance requires explicit rollback capacity because: behavioral side effects may emerge after deployment; the non-stationary environment may change the intervention's effectiveness; and regulatory requirements may mandate the ability to revert.

Operational Protocol

For each intervention, identify the rollback mechanism (version reversion, configuration toggle, corpus restoration, model checkpoint). Verify rollback capacity before deployment. Specify rollback triggers (non-target degradation, CBMES escalation). Maintain rollback infrastructure for the duration specified in the governance charter (default: 90 days post-deployment for fine-tuning; indefinite for SymPrompt$^+$ modulations).

6.4 Worked Examples: Intervention Design, DOE, and Validation

This section demonstrates the complete Improve phase for all three governance scenarios. Each worked example produces the four core Improve-phase artifacts: the intervention design with taxonomy classification and reversibility documentation, the DOE specification with factorial structure and sample sizes, the before/after validation with statistical significance testing across the full specification envelope, and the deployment record with registry updates. The cross-scenario comparison following the EDT example reveals a consistent intervention architecture pattern with implications for governance investment strategy.

Example 6.1: CDSS Improve Phase

Intervention Selection. For uncertainty disclosure omission (RPN 252, confirmed): SymPrompt$^+$ modulation targeting uncertainty disclosure (prompt engineering, first-line, high reversibility). For guideline concordance failure (RPN 200, probable): retrieval optimization with corpus update and currency tagging (high reversibility via corpus versioning).
SymPrompt$^+$ Design. Modulation SP-CDSS-2026-003 introduces explicit uncertainty classification instructions: classify confidence as high, moderate, or low based on evidence strength and recency; present qualifiers proportional to classification. Three DOE levels: no instructions (baseline), brief classification, and detailed quantification with evidence grading.
DOE. 2×3 fractional factorial crossing SymPrompt$^+$ modulation (3 levels) with retrieval strategy (2 levels: pre/post corpus update). Six conditions, 200 diagnostic queries per condition. Results: retrieval optimization significantly improves guideline concordance; detailed uncertainty modulation significantly improves uncertainty disclosure without degrading accuracy or fluency.

Positive interaction: verification instructions are more effective when the retrieved data includes recency metadata.
Validation. Combined intervention (400 queries, full protocol): SPAR improves from 0.943 to 0.981. ECPI from 1.35 to 1.62. ECI from 2.8 to 3.4. All improvements were statistically significant ($p < 0.001$). No non-target dimension shows significant degradation. Intervention approved for deployment.
Deployment. SymPrompt$^+$ modulation registered as active. Corpus version updated. MIDCOT notified. BAAGF receives efficacy report. Re-baselining window initiated for the Control phase.

Example 6.2: Insurance Claims Triage Improve Phase

Intervention Selection.

Table 6.2: *ICT intervention design*

Intervention	Root Cause	Mechanism	Reversibility
Retrieval corpus update	Outdated policy templates (FMEA #1, #3, #5)	340 policy documents refreshed; multi-peril clauses tagged with exclusion metadata; superseded templates removed	Full: archived
SymPrompt$^+$ coverage verification modulation	Insufficient multi-peril reasoning (FMEA #1, #2)	Coverage verification step injected: system checks each clause against exclusion registry and states applicability; 3 modulation levels for DOE	Full: rollback to SP-CT-v2.3.0

Both interventions address the top two FMEA entries, which account for 85% of the SPAR gap. The least-invasive-first principle is satisfied: retrieval corpus update (data-category) and SymPrompt$^+$ modulation (configuration-category) are applied before any model-category intervention. Both are fully reversible via corpus versioning and the SymPrompt$^+$ version registry.
DOE Design.

Table 6.3: *ICT DOE specification*

Parameter	Value
Design type	Fractional factorial (3×2)
Factor A	SymPrompt$^+$ modulation: Level 1 (no verification, baseline); Level 2 (brief coverage check); Level 3 (detailed exclusion verification with citation)
Factor B	Retrieval corpus: pre-update (original); post-update (refreshed with exclusion metadata)
Cells/sample	6 cells; 300 complex claims per cell (deliberately oversampled from primary gap source)
Total evaluated	1,800 claims
Evaluation protocol	Full evaluator panel (3 senior handlers); blinded to experimental condition
Optimal combination	Level 3 SymPrompt$^+$ (detailed verification) + post-update corpus

Before/After Validation.

Table 6.4: *ICT before/after validation results*

Metric	Before	After	Change	Significance
SPAR (global)	0.921	0.974	+0.053	$p < 0.001$
ECPI	1.18	1.58	+0.40	$p < 0.001$
ECI	2.4	3.2	+0.8	Exceeds Three Sigma
Coverage accuracy (complex)	0.917	0.961	+0.044	$p < 0.001$
Pathway consistency	0.914	0.956	+0.042	$p < 0.001$
Fraud precision	0.738	0.741	+0.003	Not significant
Demographic fairness Δ	0.024	0.022	−0.002	Not significant
Response completeness	0.971	0.978	+0.007	Marginal (SymPrompt$^+$ side-effect)

No non-target dimension shows statistically significant degradation. Response completeness shows a marginal improvement as a beneficial side-effect of the SymPrompt$^+$ verification module (the coverage check instruction implicitly encourages more complete reasoning narratives). Intervention approved for deployment.

TCF compliance impact. The improvement directly addresses TCF Outcome 2 (products meet customer needs: coverage accuracy +0.044) and Outcome 5 (products perform as expected: pathway consistency +0.042). The SPAR improvement from 0.921 to 0.974 provides quantitative evidence of "embedded fair treatment culture" for TCF Outcome 1 reporting.

Example 6.3: ED Triage Agent Improve Phase

Intervention Selection. Three interventions targeting the top three FMEA entries (86% of gap):

Table 6.5: *EDT intervention design*

#	Intervention	Description	FMEA Target
I-1	Retrieval corpus update	840 clinical documents updated to current versions. Sepsis-3 (2023 amendments) integrated. 2024 ACS/ACEP atypical presentation guidelines added. Version metadata tagged.	FM #1, #2, #3, #7
I-2	SymPrompt$^+$ acuity verification module	New module between Steps 1 and 2: verify vital sign trends, evaluate against atypical criteria for age/sex, apply current screening criteria. Three modulation levels.	FM #1, #2, #4
I-3	Demographic fairness constraint	Pain assessment equalization: "Apply identical vital sign and symptom severity thresholds regardless of age, race, sex, language, or insurance status."	FM #4

The EDT scenario deploys three interventions (vs. two for CDSS and ICT), reflecting the higher-dimensional CTG set (six characteristics) and the need to address both the primary gap (acuity/screening) and the demographic fairness concern simultaneously. All three are configuration-category or data-category (no fine-tuning), preserving full reversibility.

DOE Design.

Table 6.6: *EDT DOE specification*

Component	Specification
Design	$3 \times 2 \times 2$ fractional factorial: 3 SymPrompt$^+$ levels $\times$ 2 retrieval conditions $\times$ 2 fairness constraint conditions
Sample	2,400 encounters (200/cell $\times$ 12 cells). Stratified: 40% ESI-2/3, 25% complex, 15% pediatric, 20% other
Evaluation	Full 6-CTG evaluation by calibrated panel. Automated + 25% expert overlay
Optimal combination	Updated corpus + detailed verification module + fairness constraint ON
Validation	600 encounters across all input classes. No non-target degradation ($p < 0.001$)

Before/After Validation.

Table 6.7: *EDT before/after validation results*

Metric	Before	After	Change	Status
SPAR (workflow)	0.932	0.968	+0.036	Exceeds 0.960 target
SPAR (Step 2)	0.974	0.992	+0.018	Primary gap closed
Under-triage rate	0.031	0.018	−0.013	Below 0.020 target
Screening sensitivity	0.958	0.978	+0.020	Exceeds 0.970 target
Demographic fairness	0.032	0.019	−0.013	Below 0.025 target
ECPI	1.12	1.48	+0.36	Approaching 1.33
ECI	2.6	3.3	+0.7	Above Three Sigma

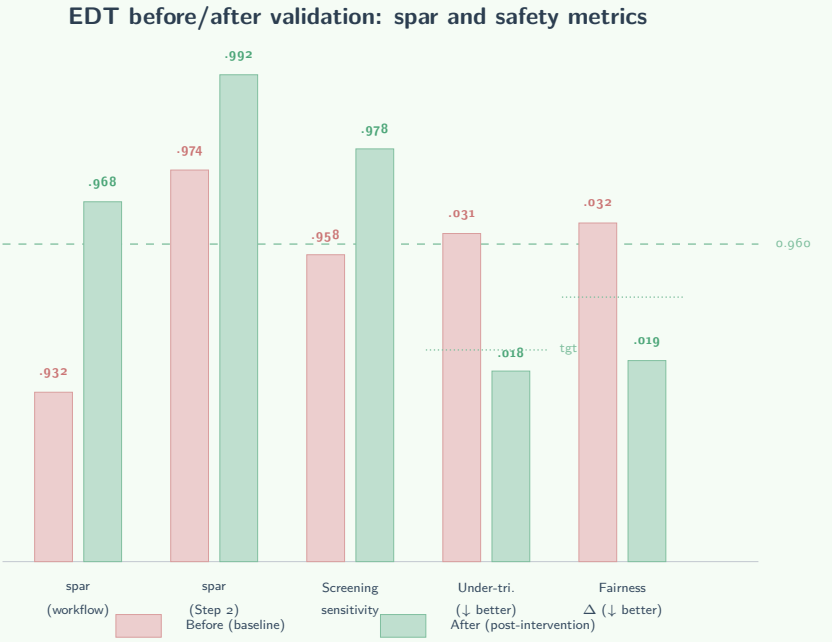


Figure 6.2: EDT before/after validation: paired comparison across five governance dimensions. Left three metrics: higher is better (bars scaled 0.90–1.00). Right two metrics: lower is better (bars scaled 0.00–0.05). All target dimensions show statistically significant improvement ($p < 0.001$); no non-target dimension shows degradation.

Clinical significance. Under-triage rate reduced from 3.1% to 1.8%. In operational terms, approximately 3 fewer patients per day receive delayed emergency evaluation (from ∼8 before to ∼5 after). For time-critical conditions (STEMI, stroke, sepsis), this improvement directly translates to reduced preventable harm.

Deployment. All three interventions were deployed as a coordinated package. SymPrompt^{+} SP-ED-v3.1.2 updated with acuity verification module (active) and fairness constraint (active). Retrieval corpus updated to v2026.04. MIDCOT notified. A 14-day re-baselining window has been initiated for the Control phase.
Cross-scenario observation: intervention patterns. All three scenarios share a common intervention architecture: retrieval corpus update (data-category) combined with SymPrompt^{+} modulation (configuration-category). No scenario required fine-tuning (model-category), confirming the empirical validity of the least-invasive-first principle: the root causes identified in the Analyze phase traced to data currency and prompt-configuration deficiencies, not to model-level behavioral inadequacy. This pattern has a governance implication: organizations that invest primarily in model fine-tuning to improve behavior may be targeting the wrong causal category. The FMEA registers across all three scenarios suggest that retrieval corpus governance and prompt architecture governance yield higher leverage per unit of intervention effort.

6.5 Failure Modes of the Improve Phase

The Improve phase has four characteristic failure modes, each representing a way in which the intervention design, testing, or deployment process can produce ineffective, fragile, or harmful governance changes. These failures are particularly consequential because they consume the opportunity to improve governance: a failed intervention cycle delays effective governance by the full cycle duration and may introduce new behavioral artifacts that complicate subsequent diagnosis.

Solution-first intervention
Implementing a preferred intervention (typically prompt engineering) without verifying it addresses the validated root cause. Produces fragile improvements that deteriorate as the underlying cause worsens.

Uncontrolled intervention stacking
Multiple interventions simultaneously without isolating effects. DOE discipline prevents this failure mode.

Insufficient validation scope
Validating only target dimensions while ignoring non-targets. The full specification envelope must be assessed.

Rollback capacity erosion
Failing to maintain rollback infrastructure. Once lost, negative consequences can only be addressed by designing new interventions, not restoring the prior state.

6.6 AAI Classification Summary: Improve Phase

Table 6.8 summarizes the AAI classification for all six Improve-phase tools. The distribution (1 Adopt, 2 Adapt, 3 Innovate) reflects the Phase's dual character: the experimental logic of DOE and before/after validation is domain-general, but the intervention instruments themselves are purpose-built for AI governance.

Table 6.8: *AAI classification profile: Improve Phase (1 Adopt, 2 Adapt, 3 Innovate)*

Tool	AAI	TM Ref	Rationale
Before/After Comparison	Adopt	TM-T4-01	Pre/post comparison logic applies directly
DOE	Adapt	TM-T4-02	Factorial logic applies; factors and response variables adapted
Pilot Testing	Adapt	TM-T4-03	Concept applies; protocol adapted for input class replication
Intervention Taxonomy	Innovate	TM-T4-04	Speed/reversibility/scope classification; no classical analogue
SymPrompt^{+} Modulation	Innovate	TM-T4-05	Governed behavioral modulation instrument; no classical analogue
Reversibility/Rollback	Innovate	TM-T4-06	Rollback as governance requirement; no classical formalization

The Improve phase AAI profile (1 Adopt, 2 Adapt, 3 Innovate) reflects the Phase's dual character: the experimental logic of DOE and before/after validation is domain-general, but the intervention instruments (SymPrompt^{+}, constraint injection, the intervention taxonomy itself) are purpose-built for AI governance. The three innovations address capabilities that have no manufacturing antecedent: governed prompt modulation, formal intervention classification, and rollback as a governance requirement.

Chapter 7 presents the Control phase: sustaining governance through statistical monitoring of the improvements validated in this chapter.

Chapter 7

Control Phase: Sustaining Governance Through Statistical Monitoring

The Control phase sustains the gains achieved in the Improve phase by implementing ongoing statistical monitoring to detect process drift, special-cause events, and behavioral degradation before they produce nonconforming outputs. It is the final phase of the DMAIC cycle and the phase that marks the transition from episodic improvement to continuous governance. When the Control phase detects signals escalating to T3 or above, it initiates a new DMAIC cycle, embedding process improvement as a repeatable mechanism within the ALAGF lifecycle [28, 48].

The Control phase translation confronts three challenges: non-stationarity (authorized changes require re-baselining. In contrast, unauthorized drift requires investigation), non-normality (BAR thresholds replace parametric $\bar{x} \pm 3\sigma$ limits), and autocorrelation (EWMA charts accommodate sequential dependence in conversational AI outputs).

7.1 The Control Phase Protocol

The Control phase protocol (PM-005) proceeds through seven steps, from control chart architecture selection through ALAGF lifecycle handoff. The protocol's central discipline is the separation of authorized non-stationarity (which triggers re-baselining) from unauthorized drift (which triggers escalation). Every signal detected by the monitoring system must be checked against the change registry before being routed to the escalation protocol; this single design decision prevents the most common Control-phase failure: false alarms from authorized changes that desensitize the governance team to genuine drift.

Procedure: Control Phase Protocol (PM-005)

Step 1: Control Chart Architecture Selection. For each CTG characteristic and input class, assess distributional characteristics (normality, autocorrelation, stationarity). Select primary charts: Shewhart I-MR with BAR thresholds (default), supplemented by CUSUM (gradual drift detection) and EWMA (autocorrelation tolerance). Configure multivariate Hotelling T^2 for joint behavioral monitoring. Specify Western Electric run rules.

Step 2: midcot Configuration. Configure MIDCOT's detection layer with chart specifications, BAR thresholds, and run rules per FM-004. Pre-register authorized process changes in the change registry. Configure monitoring cadence aligned with the governance charter. Map control chart signals to BAAGF escalation tiers (T0–T5). Execute configuration validation with baseline data.

Step 3: Escalation Protocol Operationalization. Implement the tiered escalation model from the governance charter. For each tier: specify trigger conditions (signal types, CBMES thresholds), mandatory response actions, responsible parties (from RACI matrix), maximum response timeframes, and the scope of the DMAIC cycle initiated (if applicable). For T5: configure automatic system suspension with BAAGF emergency notification.

Step 4: Re-Baselining Protocol Design. Specify triggers for re-baselining: authorized model updates, retrieval corpus changes, SymPrompt$^+$ modulation changes, specification revisions. Define the validation mode parameters: enhanced sensitivity (80% of the operational BAR), window duration (14 days by default for high-stakes), and post-change data-collection requirements. Specify the BAAGF approval process for updated thresholds.

Step 5: Control Plan Development. Compile the comprehensive control plan document: charts per CTG characteristic, BAR thresholds with derivation methodology, escalation protocol, re-baselining triggers and procedures, audit trail requirements, review cadence, and MSA refresh schedule. Where a COPQ baseline was established in the Define phase, the control plan includes quarterly COPQ trend reporting: internal and external failure cost trends alongside SPC charts, validating that governance investment produces declining total COPQ.

Step 6: alagf Lifecycle Handoff. The control plan becomes the operational specification for ALAGF's continuous monitoring function. MIDCOT executes the plan continuously. BAAGF maintains the governance registry. SymPrompt$^+$ maintains the intervention version registry. The DMAIC cycle is embedded within ALAGF as a repeatable process.

Step 7: Phase Completion. Document all Control phase artifacts. Record in ALAGF governance audit trail. The system enters production governance.

7.2 Control Phase Tool Specifications

This section specifies each tool and construct employed in the Control phase, following the AAI classification framework. One tool is adopted without modification, four are adapted with defined extensions, and four are purpose-built innovations. The Control phase exhibits the heaviest innovation load alongside the Measure phase (4 of 9 tools), reflecting the fundamental reality that continuous AI governance monitoring requires capabilities with no manufacturing SPC antecedent: production-grade automated monitoring, governance-gated threshold management, graduated escalation with DMAIC re-initiation, and systematic non-stationarity management.

7.2.1 Adopted Tools

One Control-phase tool transfers directly from classical SPC: the Western Electric run rules for pattern detection on control charts. These rules operate on sequential logic (consecutive points, runs above/below the center line) rather than distributional assumptions, making them applicable to any control chart regardless of the underlying behavioral distribution.

Western Electric Run Rules **[A · Adopt]**

Module: TM-T5-01

Classical Definition

Supplementary rules applied to control charts to detect non-random patterns before individual points violate control limits. Standard rules detect: one point beyond 3σ, two of three beyond 2σ, four of five beyond 1σ, eight consecutive on one side of the center line [47, 31].

Why Adopt

These pattern-detection rules operate on sequential logic rather than distributional assumptions. They apply to any control chart, including those with BAR thresholds. The rules detect emerging trends, oscillations, and sustained shifts before they escalate to full threshold violations.

Operational Protocol

Configure the standard Western Electric rule set on each Shewhart chart. Define the escalation tier for each rule violation (typically Advisory for single-rule violations and Warning for compound violations).

7.2.2 Adapted Tools

Four tools require defined adaptations. Shewhart charts replace parametric control limits with BAR empirical quantile thresholds. CUSUM charts use adaptive

reference values for non-stationary AI processes. EWMA charts accommodate the autocorrelation structure inherent in sequential AI outputs. The control plan adapts its content for AI-specific monitoring parameters, MIDCOT configuration specifications, and ALAGF lifecycle integration.

Shewhart Charts with bar Thresholds [D ▷ Adapt]

Module: TM-T5-02

Classical Definition

Shewhart control charts plot observations against control limits derived as $\bar{x} \pm 3\sigma$. Points outside limits signal special-cause variation.

What Is Preserved

The time-series plotting framework, the common-cause/special-cause interpretive logic, and the signal-investigation-response workflow.

What Is Modified

Parametric control limits are replaced with BAR thresholds: empirical quantile-based boundaries (0.135th and 99.865th percentiles) from the baseline behavioral distribution. This handles non-normal, multimodal, and heavy-tailed distributions. Thresholds are recalculated during re-baselining following authorized process changes.

Worked Calculation

Consider the CDSS uncertainty disclosure metric. The baseline distribution (500 diagnostic query outputs) has median $Q_{0.5} = 0.89$, 0.135th percentile $Q_{0.00135} = 0.82$, and 99.865th percentile $Q_{0.99865} = 0.96$. The BAR lower threshold is 0.82; upper 0.96. A daily observation of 0.80 falls below the lower threshold, triggering investigation. Classical $\bar{x} \pm 3\sigma$ on this left-skewed distribution would produce a lower limit of 0.767, below the 0.85 specification limit: the parametric limit would never trigger an alarm even when the system produces nonconformant output. The empirical approach detects what parametric limits miss.

Operational Protocol

Compute empirical quantiles at the governance-specified false alarm rate. Plot behavioral metric values over time against BAR thresholds. Points outside thresholds trigger investigation per the escalation protocol.

CUSUM Charts for AI Drift Detection [D ▷ Adapt]

Module: TM-T5-03

Classical Definition

Cumulative Sum charts accumulate deviations from a target value, detecting small sustained shifts that Shewhart charts may miss [28].

What Is Modified

The reference value and decision boundary require adaptation for non-stationary AI processes. Adaptive reference values accommodate legitimate non-stationarity. CUSUM is reset following authorized re-baselining events.

Sensitivity Configuration

CUSUM sensitivity is governed by k (reference value, typically 0.5σ) and h (decision interval, typically 4–5σ). For the CDSS, CUSUM on guideline concordance uses $k = 0.5$, $h = 4.5$, detecting a 0.02-point decline within approximately 15 daily observations.

Operational Protocol

Configure CUSUM with the post-improvement target value and governance-calibrated decision boundary. Apply to CTG characteristics susceptible to gradual drift. Reset following each re-baselining event.

EWMA Charts for Autocorrelated Data [D ▷ Adapt]

Module: TM-T5-04

What Is Modified

EWMA's exponential smoothing accommodates moderate autocorrelation in sequential observations. The smoothing parameter λ is calibrated to the deployment's autocorrelation structure. $\lambda = 0.10$–0.20 is the recommended default range.

When EWMA Is Preferred

EWMA is preferred when: the metric exhibits within-session autocorrelation ($\rho > 0.3$), the governance objective is detecting small sustained shifts (0.5–1.0σ), or the monitoring cadence produces observations at intervals shorter than the autocorrelation decay time.

Control Plan Design [D ▷ Adapt]

Module: TM-T5-05

What Is Preserved

The documented specification of what is monitored, how, what triggers investigation, and who responds.

What Is Modified

AI-specific content: per-CTG-characteristic chart configuration, BAR thresholds with empirical derivation, non-stationary baseline management, MIDCOT

configuration specifications, and ALAGF lifecycle integration points. The control plan is updated following each re-baselining event.

7.2.3 Innovated Constructs

Four Control-phase constructs require purpose-built innovation because they address governance requirements with no manufacturing SPC antecedent. The MIDCOT configuration protocol provides a production-grade three-layer monitoring system with integrated change management. The BAR threshold enforcement protocol manages the governance-authority-gated recalculation of thresholds. The tiered escalation protocol provides a graduated CBMES-driven response with DMAIC re-initiation. The re-baselining protocol manages the fundamental tension between continuous system evolution and monitoring stability.

midcot Configuration Protocol **[I⋆Innovate]**

Module: TM-T5-06

Why Innovation Is Required

No classical SPC construct provides a production-grade monitoring system with integrated change management, multi-chart architecture, CBMES-driven escalation, and subsystem directive interfaces.

Three-Layer Architecture

Monitoring layer. Ingests behavioral metric data at the governance-specified cadence. Maintains data with full provenance metadata. Computes running chart statistics for each governed CTG characteristic within each input class.
Detection layer. Evaluates statistics against thresholds and rules. Synthesizes concurrent signals across chart types and dimensions. Checks every signal against the change registry: signals coinciding with authorized changes are routed to re-baselining rather than escalation.
Escalation layer. Maps synthesized signals to BAAGF escalation tiers using CBMES-based classification. For T5 signals: executes automatic system suspension without waiting for human disposition.

Operational Protocol

Configure all three layers per the control plan. Pre-register authorized changes. Configure validation mode for re-baselining windows. Execute configuration validation by verifying zero false alarms on baseline data, correct detection of synthetic anomalies, and correct tier classification for test signals.

bar Threshold Enforcement [I⋆Innovate]

Module: TM-T5-07

Why Innovation Is Required

BAR enforcement must accommodate threshold recalculation during re-baselining, multi-chart threshold coordination, and governance-authority-gated changes (BAAGF approval required before updated thresholds are activated).

Tiered Escalation Protocol [I⋆Innovate]

Module: TM-T5-08

Why Innovation Is Required

Classical SPC escalation is binary. AI governance requires graduated response: behavioral deviations have varying consequences, restrictions must be proportional, and DMAIC re-initiation must be scoped to the escalation context. The T0–T5 tiered model with CBMES-driven tier assignment provides this capability.

Operational Protocol

Map signal types to initial tier classifications. Aggregate concurrent signals into CBMES-based determination. Execute mandatory response actions per tier. For T3+: initiate scoped DMAIC cycle. For T5: automatic suspension with BAAGF emergency notification. Log all escalation events in the audit trail.

Re-Baselining Protocol [I⋆Innovate]

Module: TM-T5-09

Why Innovation Is Required

AI systems under continuous development undergo authorized changes that legitimately shift the behavioral distribution. Re-baselining provides the governance mechanism for updating monitoring parameters while maintaining audit continuity.

Operational Protocol

Trigger identification (authorized change verified against the change registry). Validation mode activation (enhanced sensitivity for re-baselining window). Post-change data collection. Metric recalculation. Governance approval (BAAGF reviews updated thresholds). Standard monitoring resumption.

7.3 Worked Examples: Control Architecture and Drift Events

This section demonstrates the complete Control phase for all three governance scenarios, including the production drift events that validate the monitoring architecture's detection capability. Each worked example produces the control chart configuration, the escalation protocol, the re-baselining protocol, and a documented case record of a drift event with full CAPA (Corrective and Preventive Action) disposition. The cross-scenario comparison following the EDT example reveals a structural pattern in the causation of drift events that validates the Define-phase SIPOC analysis and confirms retrieval corpus currency monitoring as a standard governance requirement.

Example 7.1: CDSS Control Phase

Chart Architecture. Layered monitoring for each CTG characteristic:
Citation accuracy: Shewhart I-MR with BAR thresholds (lower: 0.87). Western Electric Rule 1 triggers the Warning tier. Daily cadence.
Guideline concordance: Shewhart I-MR (lower BAR: 0.88) layered with CUSUM ($k = 0.5$, $h = 4.5$). CUSUM is the primary sensitivity instrument: corpus staleness manifests as slow degradation.
Uncertainty disclosure: Shewhart I-MR (lower BAR: 0.82) layered with EWMA ($\lambda = 0.15$) for within-session autocorrelation. A post-improvement ECPI of 0.95 provides only a limited margin above the 0.85 specification limit.
Contradiction avoidance: Shewhart I-MR only (lower BAR: 0.94); strongest baseline performance, lowest drift susceptibility.
Joint monitoring: Multivariate Hotelling T^2 across all four dimensions simultaneously.
Drift Detection Event. Three months post-deployment, the CUSUM chart for guideline concordance signals sustained negative drift. The cumulative sum accumulates to 4.2 standard deviations over two weeks: Shewhart shows all individual points within BAR limits, but CUSUM detects the trend Shewhart misses. The investigation reveals that three clinical guidelines were updated but not refreshed in the retrieval corpus. The governance team initiates a targeted Improve cycle: corpus update, validation testing, and 14-day re-baselining with enhanced sensitivity (80% of operational BAR). Post-re-baselining: guideline concordance ECPI returns to 1.45 (from pre-drift 1.48). Complete incident documented in governance audit trail.

Example 7.2: Insurance Claims Triage Control Phase

midcot Configuration.

Table 7.1: *ICT MIDCOT control configuration*

Parameter	Configuration
Monitoring layer	Daily batch aggregation; real-time during first 30 days post-intervention and re-baselining windows
Shewhart charts	I-MR for each CTG characteristic per input class; BAR from post-improvement empirical quantiles
CUSUM charts	Layered for coverage accuracy and pathway consistency (highest drift susceptibility); adaptive reference values
EWMA charts	Session-level monitoring for sequential behavioral autocorrelation
Multivariate	Hotelling T^2 for joint behavioral vector across all five CTG dimensions
Run rules	Western Electric rules on all Shewhart charts (adopted without modification per AAI taxonomy)
Data retention	7 years (regulatory minimum: 5 years per PPR record-keeping requirements)

Tiered Escalation Protocol.

Table 7.2: *ICT escalation protocol*

Tier	Trigger	Response	Timeline
Advisory	Run rule violation; metric within 15% of BAR; CUSUM below warning	Investigation initiated; enhanced monitoring; system continues	48 hours
Warning	Single BAR violation; SPAR <0.960 for any input class	AI autonomy restricted for affected class; human review mandated; committee notified	4 hours
Critical	2+ BAR violations; SPAR <0.940; ECI <2.0	AI triage suspended for affected lines; committee convened; full Analyze-Improve cycle	1 hour
Emergency	Systematic denial of valid claims in protected segment; regulatory-reportable incident	Full system suspension; emergency session; regulatory notification; complete DMAIC cycle	<15 min

Re-Baselining Protocol.

Table 7.3: *ICT re-baselining triggers*

Trigger Event	Window	Approval	bar Update
Model version update	21 days	Governance Committee	Full recalculation
Retrieval corpus update	14 days	Op. Gov. Team (routine) / Committee (major)	Affected input classes
SymPrompt^{+} modulation change	14 days	Governance Committee	Full recalculation; before/after required
Specification revision	21 days	Committee + Board Risk	New targets/limits; full Define-through-Control
Regulatory change (PPR, FSCA)	Per directive	Committee + Compliance	Assess specification adequacy

Drift Event Case Record: MIDCOT-2026-0073.

Table 7.4: *ICT drift event timeline*

Day	Event	Metric/Action
Day 73	MIDCOT CUSUM signals sustained negative drift on motor coverage accuracy	CUSUM exceeds Advisory threshold; Shewhart within BAR
Day 74	Root cause: motor endorsements updated by underwriting 16 days prior but not loaded into retrieval corpus	Motor SPAR declined 0.974 $\rightarrow$ 0.958 over 16 days
Day 75	Targeted Improve cycle: corpus updated with current motor endorsements; SymPrompt^{+} unchanged	200 motor claims validated; coverage accuracy restored to 0.964
Days 76–89	Re-baselining: 14-day post-update data collection under enhanced monitoring	CUSUM resolves Day 80; SPAR stabilizes at 0.976
Day 90	Audit trail closed	Final: SPAR 0.976; ECPI 1.61; ECI 3.3

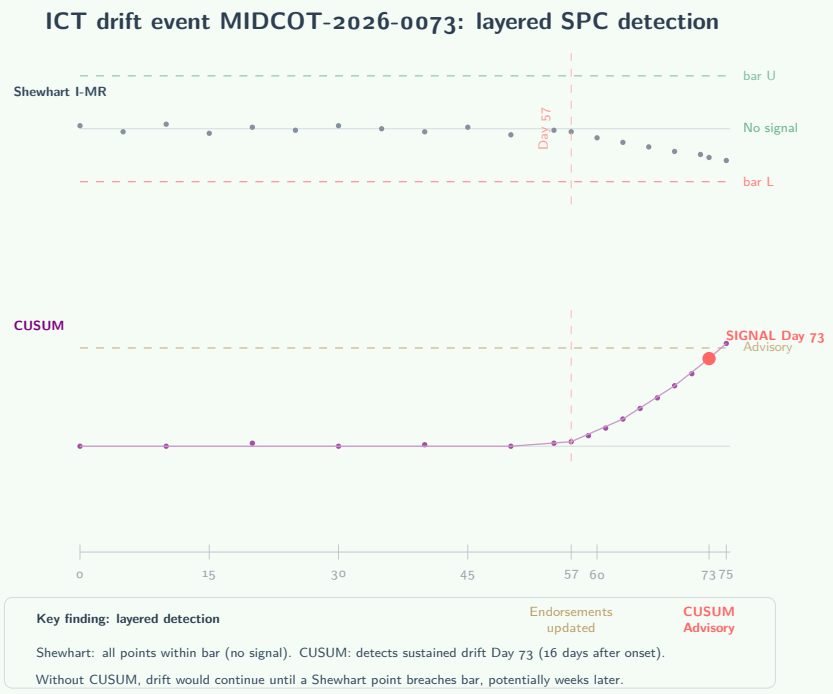


Figure 7.1: ICT drift event MIDCOT-2026-0073: layered SPC detection. Top: Shewhart I-MR chart shows all points within BAR thresholds (no signal). Bottom: CUSUM chart accumulates deviations and triggers Advisory on Day 73, detecting the gradual drift that Shewhart misses. This demonstrates why the layered chart architecture is a governance requirement.

Governance Audit Trail. Incident MIDCOT-2026-0073. Detection: CUSUM chart, motor coverage accuracy, and daily aggregation. Signal: sustained negative drift (gradual). Escalation tier reached: Advisory (Warning not breached). Root cause (confirmed): retrieval corpus data currency failure; 23 motor policy endorsements updated by underwriting but not propagated to the AI retrieval index. Time from endorsement change to detection: 16 days. Time from detection to resolution: 17 days (3 days investigation/intervention + 14 days re-baselining). SPAR impact: declined 0.974 → 0.958 (motor); restored to 0.976. Policyholder impact: no individual claim identified as materially harmed; all motor claims during the drift period flagged for retrospective review. Regulatory reportable: no (Advisory tier; PPR threshold not breached).

CAPA. Corrective: motor endorsements loaded into corpus (Day 75, complete); retrospective review of 1,280 motor claims during drift period (Day 105 target). Preventive: automated integration between underwriting endorsement release and AI corpus update pipeline (Day 120 target); retrieval corpus freshness monitoring added to MIDCOT with automated alert when any

document category exceeds the scheduled update date by >48 hours (Day 100 target).

Example 7.3: ED Triage Agent Control Phase

midcot Configuration.

Table 7.5: *EDT MIDCOT control configuration*

Parameter	Specification
Monitoring cadence	Real-time per-encounter scoring (Steps 1–4). Hourly batch aggregation. Daily SPC chart update. Weekly governance report
SPC charts	Shewhart I-MR for workflow SPAR (daily). CUSUM for under-triage rate and sepsis screening sensitivity. Hotelling T^2 for joint 6-CTG vector
BAR thresholds	Upper: 0.988. Lower: 0.948. Post-improvement empirical quantiles
CUSUM parameters	$k = 0.5\sigma$, $h = 4\sigma$. $\text{ARL}_0 = 370$ days (low false alarm). $\text{ARL}_1 =$ 14 days (detect 1σ shift within 2 weeks)

Tiered Escalation Protocol.

Table 7.6: *EDT escalation protocol*

Tier	Trigger	Response	Timeline
Advisory	CUSUM signal; metric within 15% of BAR; under-triage >0.025 for 3 consecutive days	Enhanced monitoring (hourly). Investigation initiated	48 hours
Warning	BAR violation; under-triage >0.030; workflow SPAR <0.955	Mandatory senior clinician double-review. Affected class isolated	4 hours
Critical	2+ BAR violations; under-triage >0.040; SPAR <0.940; missed sepsis/STEMI/stroke with adverse outcome	AI triage SUSPENDED for affected classes. Manual triage reinstated. Full investigation	1 hour
Emergency	Patient safety event attributable to AI; any EMTALA violation	FULL SYSTEM HALT. CMS/TJC/state notification. Sentinel event investigation	<15 min

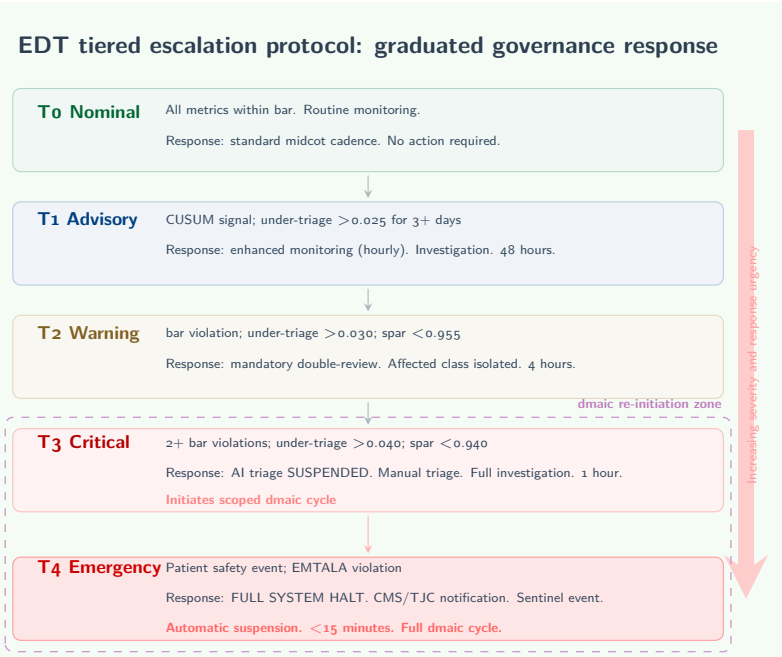


Figure 7.2: EDT tiered escalation protocol: five governance tiers (T0–T4) with trigger conditions, mandatory response actions, and maximum response timeframes. T3 and T4 (dashed box) trigger DMAIC cycle re-initiation. T4 executes automatic system suspension within 15 minutes.

Drift Event Case Record: MIDCOT-2026-ED-0041.

Table 7.7: *EDT drift event timeline*

Timeline	Event
Days 1–30	Post-improvement stability confirmed. Workflow SPAR stable at 0.968. All CTG metrics within specification
Day 45	ACEP publishes updated sepsis screening criteria emphasizing lactate trajectory over static qSOFA for patients >65
Days 45–62	Updated criteria NOT loaded. AI continues the previous criteria. No immediate degradation (most presentations still captured)
Day 58	CUSUM for sepsis screening sensitivity begins accumulating. Not yet at Advisory
Day 63	**ADVISORY triggered.** CUSUM crosses threshold. Screening sensitivity drifted 0.978 → 0.961 over 18 days. Under-triage >65 increased 0.018 → 0.028. Investigation initiated
Day 64	Root cause confirmed: ACEP 2026 criteria not loaded. Lactate trajectory criterion missing. 3–4 elderly sepsis patients/day receiving delayed screening
Day 65	Corpus updated with ACEP 2026 criteria. SymPrompt^{+} unchanged. 150 encounters validated: screening sensitivity restored to 0.976
Days 65–79	14-day re-baselining. Enhanced monitoring. New BAR calculated
Day 79	**RESOLVED.** Workflow SPAR 0.971. Screening sensitivity 0.979. Audit trail closed

Clinical Significance. During the 18-day drift window (Days 45–63), an estimated 54–72 elderly patients received delayed sepsis screening. Approximately 8 to 12 had confirmed sepsis. Each hour of delay in sepsis bundle initiation increases mortality risk by 7.6%. The 18-day detection lag, while fast compared with manual audits, remains an area for improvement.
CAPA. The drift event matches FMEA #3 (outdated protocol retrieval, RPN 162). Corrective: corpus updated (Day 65, complete). Preventive: automated guideline publication monitoring pipeline approved by Clinical AI Committee; target detection time reduced from 18 days to <7 days through automated corpus-currency checks linked to ACEP, ACS, and Surviving Sepsis Campaign publication feeds.
Cross-scenario observation: drift event structural pattern. All three drift events (CDSS guideline concordance, ICT motor endorsements, EDT sepsis criteria) share a common causal structure: an external authority publishes updated content, the update is not propagated to the AI system's retrieval corpus, the AI continues operating on outdated knowledge, and MIDCOT detects the resulting behavioral drift through CUSUM accumulation before Shewhart charts signal. In all three cases, the CAPA converges on the same preventive action: automated integration between the external publication source and the retrieval corpus update pipeline. This convergence validates the SIPOC analysis from the Define phase (Chapter 3), which identified the retrieval

corpus supply chain as a critical governance interface, and confirms that retrieval corpus currency monitoring should be a standard component of any MIDCOT configuration for RAG-augmented AI systems.

7.4 Failure Modes of the Control Phase

The Control phase has five characteristic failure modes, each representing a way in which the monitoring and response infrastructure can degrade over time, producing either false assurance (failing to detect genuine drift) or false alarms (detecting authorized changes as anomalies). These failures are uniquely consequential because the Control phase operates indefinitely: a manufacturing SPC implementation may run for months before a Measure phase failure manifests, but a Control phase failure can persist for years, silently eroding governance effectiveness.

Monitoring atrophy
Governance monitoring established with rigor but gradually receiving diminished attention. Alert fatigue from T1/T2 notifications desensitizes the team. MIDCOT's automated surveillance mitigates this, but T3+ responses require human engagement.

Threshold ossification
BAR thresholds set at deployment and never revised. Scheduled reviews must explicitly assess the adequacy of thresholds against the evolving operating context.

Change management disconnection
System changes occur without triggering re-baselining. Produces false signals (authorized changes detected as drift) and false assurance (monitoring against stale baselines).

Single-chart dependence
Relying on a single chart type when the behavioral distribution requires layered detection. Shewhart alone misses gradual drift; CUSUM alone misses sudden shifts. The layered architecture is a governance requirement.

Baseline instability
Re-baselining from data collected during an unstable post-change period. Validation mode with enhanced sensitivity addresses this.

7.5 AAI Classification Summary: Control Phase

Table 7.8 summarizes the AAI classification for all nine Control-phase tools. The distribution (1 Adopt, 4 Adapt, 4 Innovate) reflects the phase's position at the intersection of classical SPC methodology and AI-specific governance requirements:

the sequential pattern logic of run rules transfers directly, the chart architectures transfer structurally but require parametric replacements and non-stationarity accommodation, and the monitoring infrastructure, threshold management, escalation, and re-baselining protocols address governance requirements with no manufacturing antecedent.

Table 7.8: *AAI classification profile: Control phase (1 Adopt, 4 Adapt, 4 Innovate)*

Tool	AAI	TM Ref	Rationale
Run Rules	Adopt	TM-T5-01	Pattern detection applies to any control chart
Shewhart Charts	Adapt	TM-T5-02	BAR replaces parametric limits
CUSUM Charts	Adapt	TM-T5-03	Adaptive reference values for non-stationary processes
EWMA Charts	Adapt	TM-T5-04	Autocorrelation accommodation
Control Plan	Adapt	TM-T5-05	Content adapted for AI monitoring and lifecycle
MIDCOT Configuration	Innovate	TM-T5-06	Production-grade AI governance monitoring system
BAR Enforcement	Innovate	TM-T5-07	Governance-gated threshold management
Tiered Escalation	Innovate	TM-T5-08	Graduated CBMES-driven response model
Re-Baselining	Innovate	TM-T5-09	Non-stationarity management protocol

The Control phase exhibits a balanced but innovation-heavy profile (1 Adopt, 4 Adapt, 4 Innovate), reflecting its position at the intersection of classical SPC methodology and AI-specific governance requirements. The adopted tool (run rules) operates on sequential pattern logic. The adapted tools transfer structurally but require parametric replacements and non-stationarity accommodation. The innovations address governance requirements with no manufacturing antecedent: production-grade automated monitoring, governance-gated threshold management, graduated escalation with DMAIC re-initiation, and systematic non-stationarity management.

With the Control phase, the five DMAIC phases are complete. Chapter 8 consolidates the full AAI taxonomy. Chapter 9 provides the comprehensive standards alignment mapping. Chapter 10 demonstrates framework applicability across deployment classes. Chapter 11 identifies boundary conditions and research directions.

Chapter 8

The Tool Adaptation Layer: Comprehensive AAI Taxonomy

Chapters 3 through 7 translated each DMAIC phase into the AI governance domain, classifying every tool and construct using the AAI taxonomy and demonstrating each classification through three parallel worked examples: a clinical decision support system (CDSS), an insurance claims triage system (ICT), and an emergency department triage agent (EDT). This chapter consolidates those per-phase classifications into a comprehensive adaptation taxonomy, provides the theoretical justification for the classification boundaries, and addresses the foundational question: under what conditions do the classical assumptions of Six Sigma and SPC fail for LLM processes, and what mathematical remedies are required?

This is the intellectually differentiating contribution of the monograph. The claim is not merely that Six Sigma can be applied to AI governance, but that the conditions of applicability, extension, and replacement can be specified with formal precision.

8.1 The Comprehensive Tool Adaptation Taxonomy

Table 8.1 consolidates all AAI classifications from Chapters 3–7, reorganized by classification tier rather than by DMAIC phase. This reorganization reveals structural patterns that per-phase analysis obscures.

Table 8.1: *Comprehensive AAI tool adaptation taxonomy (40 entries mapping to 39 distinct tools and constructs)*

Tool / Construct	AAI	Phase	TM Ref
Adopt (9 tools)			
SIPOC Diagram	Adopt	Define	TM-T1-01
RACI Matrix	Adopt	Define	TM-T1-02
Data Collection Plan	Adopt	Measure	TM-T2-01
Operational Definitions	Adopt	Measure	TM-T2-02
5 Whys	Adopt	Analyze	TM-T3-01
Pareto Analysis	Adopt	Analyze	TM-T3-02
Run Charts	Adopt	Analyze	TM-T3-03
Before/After Comparison	Adopt	Improve	TM-T4-01
Western Electric Run Rules	Adopt	Control	TM-T5-01
Adapt (15 tools)			
VOG Analysis	Adapt	Define	TM-T1-03
CTG Tree Construction	Adapt	Define	TM-T1-04
Governance Charter	Adapt	Define	TM-T1-05
Governance Value Stream Map	Adapt	Define	TM-T1-09
MSA / GRR	Adapt	Measure	TM-T2-03
ECPI (Capability)	Adapt	Measure	TM-T2-04
Fishbone / Ishikawa	Adapt	Analyze	TM-T3-04
AI-Adapted FMEA	Adapt	Analyze	TM-T3-05
Variation Decomposition	Adapt	Analyze	TM-T3-06
DOE	Adapt	Improve	TM-T4-02
Pilot Testing	Adapt	Improve	TM-T4-03
Shewhart Charts	Adapt	Control	TM-T5-02
CUSUM Charts	Adapt	Control	TM-T5-03
EWMA Charts	Adapt	Control	TM-T5-04
Control Plan	Adapt	Control	TM-T5-05
Innovate (16 constructs)			
Behavioral Specification Envelope	Innovate	Define	TM-T1-06
Input Class Registry	Innovate	Define	TM-T1-07
VOG Weighting Model	Innovate	Define	TM-T1-08
COPQ Framework	Innovate	Define	TM-T1-10
BAR Threshold Derivation	Innovate	Measure	TM-T2-05
Baseline Distribution	Innovate	Measure	TM-T2-06
CBMES Baseline	Innovate	Measure	TM-T2-07

Table 8.1: *(continued)*

Tool / Construct	AAI	Phase	TM Ref
Ongoing Collection Protocol	Innovate	Measure	TM-T2-08
Prioritized Cause Register	Innovate	Analyze	TM-T3-07
Intervention Taxonomy	Innovate	Improve	TM-T4-04
SymPrompt^{+} Modulation	Innovate	Improve	TM-T4-05
Reversibility/Rollback	Innovate	Improve	TM-T4-06
MIDCOT Configuration	Innovate	Control	TM-T5-06
BAR Enforcement	Innovate	Control	TM-T5-07
Tiered Escalation	Innovate	Control	TM-T5-08
Re-Baselining Protocol	Innovate	Control	TM-T5-09

8.2 Structural Patterns in the Taxonomy

The comprehensive taxonomy reveals three structural patterns that illuminate the relationship between the classical Six Sigma methodology and AI governance requirements. The three parallel scenarios (CDSS, ICT, EDT) provide empirical grounding for each pattern.

Pattern 1: Diagnostic and organizational tools adopt; measurement and control tools require innovation. The Adopt category is dominated by tools whose function is structural organization (SIPOC, RACI), visualization (run charts), or logical analysis (Fishbone, 5 Whys, Pareto). These tools operate on the *logic* of process governance rather than on distributional assumptions. All three scenarios confirmed this: the SIPOC, RACI, and 5 Whys tools required no structural modification across clinical, insurance, and emergency medicine contexts. The Innovate category is dominated by measurement constructs (ECPI, BAR, ECI, SPAR, CBMES) and control instruments (MIDCOT, tiered escalation, re-baselining). These tools interact directly with the statistical properties of process data, and it is precisely those properties (non-normality, non-stationarity, autocorrelation, multidimensional correlation) that violate classical assumptions.

Pattern 2: The Adapt category bridges methodology and domain. Adapted tools are those whose methodological logic transfers but whose parametric assumptions, input formats, or interpretive conventions require defined modification. The three scenarios demonstrate that adaptation is deployment-context-sensitive: MSA/GRR thresholds were calibrated to domain-specific evaluator expertise levels (clinical domain experts for CDSS, senior claims handlers for ICT, board-certified emergency physicians for EDT), and FMEA severity scales were calibrated to

domain-specific harm profiles (Severity 10 reserved for irreversible clinical harm in EDT, systematic unfair outcomes in ICT).

Pattern 3: Innovation concentrates in Define and Measure/Control. The Define phase requires innovation because the governance constructs it produces (specification envelope, input class registry) have no manufacturing antecedent. The three scenarios confirmed this with different input class structures: 4 classes for CDSS, 4 stratification dimensions with combinatorial classes for ICT, and 4 dimensions including time-of-day for EDT. The Measure and Control phases require innovation because their statistical instruments are embedded in distributional assumptions that LLM processes violate. The Analyze and Improve phases are the most adoptable.

8.3 Phase-Level AAI Distribution

Table 8.2 and Figure 8.1 present the AAI distribution across the five DMAIC phases. The aggregate distribution (9 Adopt, 15 Adapt, 16 Innovate) confirms that classical Six Sigma methodology provides a substantial, transferable structure (24 of 40 tools are adopted or adapted) while requiring significant domain-specific innovation (16 constructs) to address the unique characteristics of stochastic AI systems. The per-phase breakdown reveals which phases are most adoptable and which demand the most innovation.

Table 8.2: *AAI distribution by DMAIC phase*

Phase	Adopt	Adapt	Innovate	Total	Dominant Pattern
Define	2	4	4	10	Largest phase; governance constructs and economic framework drive innovation
Measure	2	2	4	8	Highest innovation rate; distributional assumptions fail
Analyze	3	3	1	7	Most adoptable; logical tools dominate
Improve	1	2	3	6	Innovation in intervention instruments
Control	1	4	4	9	SPC adaptation plus governance innovation
Total	**9**	**15**	**16**	**40**	

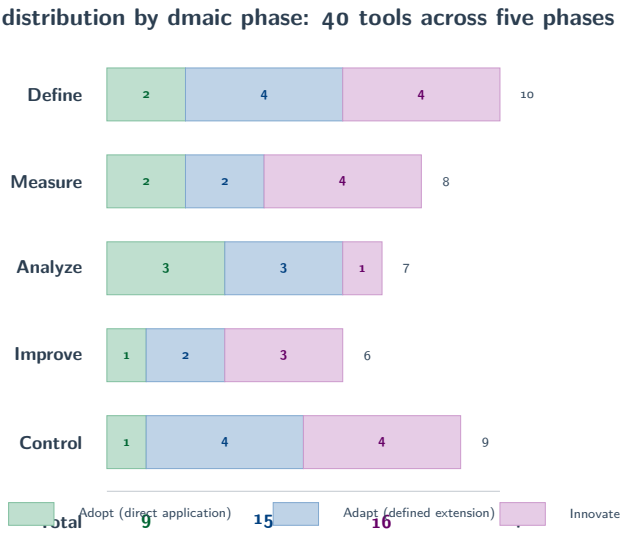


Figure 8.1: AAI distribution by DMAIC phase: stacked bars show the proportion of adopted, adapted, and innovated tools per phase. The Analyze phase is the most adoptable (6 of 7 tools transfer structurally); the Measure and Control phases require the most innovation (4 of 8 and 4 of 9, respectively). The aggregate 9/15/16 distribution confirms that classical Six Sigma provides substantial transferable structure while requiring significant domain-specific innovation.

8.4 When Classical SPC Assumptions Fail

The theoretical justification for the Innovate constructs rests on three specific assumption failures. Each failure is specified with the mathematical remedy that the corresponding BME construct provides.

8.4.1 Normality Failure

Classical process capability indices (C_p, C_{pk}) and Shewhart control limits are derived from the normal distribution [28]. LLM behavioral metrics are frequently non-normal: SPAR distributions are bounded on $[0, 1]$ and typically left-skewed; entropy-based metrics may exhibit multimodality; individual CTG characteristics may have heavy tails reflecting rare but extreme failures. All three scenarios confirmed non-normality in their baseline distributions, particularly for coverage accuracy (ICT) and under-triage rate (EDT), both of which exhibit strong left-skew with bounded support.

Remedy: Replace parametric capability calculations with empirical quantile-based methods. ECPI uses the ratio of specification width to empirical interquantile range. BAR thresholds are set at empirical quantiles (0.135th and 99.865th percentiles) rather than $\bar{x} \pm 3\sigma$.

8.4.2 Stationarity Failure

Classical SPC assumes a stable mean and variance. LLM processes exhibit legitimate non-stationarity: authorized model updates, corpus maintenance, and specification revisions produce authorized distribution shifts. The three drift event case records (Chapters 7, Appendices G and H) demonstrate this directly: all three trace to retrieval-corpora currency failures that are authorized changes by the publishing authority but are unauthorized from the AI system's governance perspective.

Remedy: Adaptive control chart architecture with change registry integration. Authorized changes are pre-registered in MIDCOT's change registry. Unauthorized shifts trigger an investigation.

8.4.3 Independence Failure

Classical Shewhart charts assume successive observations are independent. LLM outputs within a conversational session exhibit autocorrelation because each response is conditioned on conversation history. The EDT scenario demonstrates this most acutely: within a single patient encounter, the four-step pipeline produces sequentially dependent outputs where Step 2 is conditioned on Step 1.

Remedy: Sample independent observations (one per session), or use EWMA charts with exponential smoothing to accommodate moderate autocorrelation.

8.5 Theoretical Justification for the BME Metric Suite

The four primary BME constructs exist because the classical constructs they replace embed distributional assumptions that LLM behavioral processes violate:

ecpi replaces C_p/C_{pk}
Process capability assessed without normality or stationarity. Uses entropy-based measurement and empirical quantile-based calculation.

bar replaces $\bar{x} \pm 3\sigma$
Boundaries of expected behavioral variation derived from the empirical distribution rather than normal-theory parameters.

eci replaces sigma level
Distance to governance threshold measured in empirical behavioral standard deviations calculated from the actual distribution shape.

spar replaces DPMO yield
Conformance assessed jointly across correlated behavioral dimensions rather than independently across assumed-independent defect opportunities.

The common thread is *distribution-free governance*: the BME Metric Suite provides measurement infrastructure for process governance without requiring the governed process to conform to distributional assumptions it does not satisfy. The three scenarios validate this empirically: SPAR provided meaningful joint conformance measurement across 4 dimensions (CDSS), 5 dimensions (ICT), and 6 dimensions with per-step decomposition (EDT), none of which could have been assessed using classical DPMO yield calculations.

Chapter 9

Governance Integration: Alignment with AI Standards and Regulatory Frameworks

The DMAIC-AI framework achieves practical governance value only if it integrates with the standards and regulatory frameworks that govern AI deployment. This chapter maps the framework to four primary international governance references: ISO/IEC 42001:2023, the NIST AI Risk Management Framework 1.0, IEEE governance standards, and the EU AI Act. It then extends the alignment to jurisdiction-specific regulatory frameworks demonstrated through the three worked scenarios: South African financial conduct regulation (FAIS, TCF, PPRs) for the ICT scenario and US clinical regulatory requirements (EMTALA, CMS, Joint Commission) for the EDT scenario. The objective is not merely to demonstrate compatibility but to position DMAIC-AI as a process-level implementation methodology that existing strategic frameworks require but do not themselves provide.

9.1 ISO/IEC 42001:2023 Alignment

ISO/IEC 42001:2023 establishes requirements for an AI management system (AIMS), following the Annex SL high-level structure and complemented by ISO/IEC 23053's ML framework [18], which is common to ISO management system standards [19]. The standard requires organizations to establish, implement, maintain, and continually improve an AIMS but does not prescribe specific process-level methodologies. DMAIC-AI fills this gap.

Table 9.1: *ISO/IEC 42001:2023 clause-by-clause alignment*

Clause	Requirement	dmaic-AI Implementation
4.1–4.4	Context and AIMS establishment	Define phase: governance charter, VOG analysis, stakeholder identification. ALAGF lifecycle architecture.
5.1–5.3	Leadership, roles, authorities	BAAGF governance authority structure, RACI matrix, escalation ownership.
6.1	Risk assessment	Analyze phase: AI-FMEA with governance-adapted S/O/D ratings. Prioritized cause register.
8.1	Operational planning	Define: CTG trees, specification envelopes, input class registries. Measure: data collection plans.
8.2	AI risk assessment	BAAGF FMEA register, DAST pre-deployment clearance, adversarial testing.
9.1	Monitoring and measurement	Measure: BME Metric Suite, MSA validation. Control: MIDCOT monitoring, BAR thresholds.
9.2–9.3	Internal audit and review	Control plan review cadence, governance audit trail, MIDCOT reporting.
10.1–10.2	Continual improvement	Improve: SymPrompt$^+$ interventions, DOE optimization, before/after validation. DMAIC cyclical integration.

Organizations seeking ISO 42001 certification can use DMAIC-AI as the process methodology satisfying the standard's continual improvement and monitoring requirements, with ALAGF providing the management system architecture and BME providing the measurement infrastructure. The three parallel scenarios demonstrate this across deployment contexts: the CDSS satisfies Clause 9.1 through clinical panel-validated BME metrics; the ICT satisfies Clause 10.1 through DOE-validated SymPrompt$^+$ interventions with TCF traceability; and the EDT satisfies Clause 8.2 through the cascading-failure FMEA with Severity 10 clinical calibration.

9.2 NIST AI Risk Management Framework 1.0 Alignment

The NIST AI RMF organizes AI risk management into four core functions: GOVERN, MAP, MEASURE, and MANAGE [29].

Table 9.2: *NIST AI RMF function-level alignment*

Function	Core Activities	dmaic-AI Implementation
GOVERN	Policies, accountability, structures	Define: governance charter, RACI, BAAGF authority structure.
MAP	Context, capabilities, risk identification	Define + Analyze: VOG, input class registry, AI failure taxonomy, variation decomposition.
MEASURE	Quantitative risk assessment and metrics	Measure: BME Metric Suite (ECPI, BAR, ECI, SPAR), MSA validation.
MANAGE	Risk prioritization, response, monitoring	Analyze: FMEA prioritization. Improve: interventions. Control: MIDCOT, escalation, re-baselining.

The NIST AI RMF calls for "validated measurement approaches" but does not specify which tools to use. DMAIC-AI provides that specification. The NIST AI 600-1 [30] (Generative AI Risk Profile) complements the RMF by identifying risk categories specific to generative AI; each can be decomposed into measurable CTG characteristics with specification limits, enabling SPC-based monitoring.

9.3 IEEE Standards Alignment

IEEE P2863 (Organizational Governance of AI) addresses organizational structures, decision-making authorities, and accountability mechanisms [16]. The governance charter and RACI matrix from the Define phase operationalize the decision authority structures P2863 requires. IEEE 7010 (Well-Being Metrics) provides a complementary framework for societal impact; CTG characteristics can include well-being-relevant behavioral dimensions that connect process-level governance to societal assessment.

9.4 EU AI Act Considerations

The EU AI Act (Regulation 2024/1689) establishes risk-based classification with graduated requirements [9]. All three scenarios are classified as high-risk: the CDSS under health-related AI provisions, the ICT under Annex III Section 5(ca) (insurance risk assessment and pricing), and the EDT under Annex III Section 5(c) (emergency healthcare triage). The DMAIC-AI framework satisfies each Act requirement through identified governance artifacts (Art. 9 through the FMEA register, Art. 10 through corpus governance, Art. 13 through reasoning transparency

CTG, Art. 14 through RACI with human authority at T3+, Art. 72 through MIDCOT continuous monitoring).

9.4.1 Executive Order 13960 Alignment

Executive Order 13960 [10] establishes nine principles for trustworthy AI in federal agencies. DMAIC-AI addresses each through governance architecture: "Purposeful" and "Lawful" through VOG analysis; "Accurate" and "Safe" through CTG specifications and SPAR targets; "Regular monitoring" through MIDCOT; "Transparent" and "Accountable" through the governance audit trail and SymPrompt$^+$ version registry.

9.5 Jurisdiction-Specific Regulatory Alignment

The international standards mapped above provide the governance architecture. Jurisdiction-specific regulation imposes additional requirements that the framework must satisfy.

9.5.1 South African Financial Conduct Regulation (ICT Scenario)

The ICT scenario operates under four South African regulatory instruments:

Table 9.3: *ICT regulatory alignment: South African financial conduct frameworks*

Regulation	Authority	dmaic-AI Implementation
FAIS Act (Act 37 of 2002)	FSCA	VOG includes FAIS Ombud as stakeholder; audit trail provides Ombud-grade evidence for complaint adjudication
TCF Outcomes	FSCA	Six TCF outcomes mapped to specific CTG characteristics (Table 9.4); SPAR quantifies behavioral conformance as TCF evidence
PPRs	FSCA	Consistent handling monitored via pathway consistency CTG; tiered escalation satisfies intervention duty; 7-year data retention exceeds 5-year PPR minimum
Insurance Act (Act 18 of 2017)	Prudential Authority	Governance charter satisfies governance obligations; MIDCOT reporting feeds risk management; fit and proper requirements reflected in evaluator calibration

The TCF outcome mapping is the most governance-significant alignment for

the ICT scenario. Each TCF outcome is operationalized through specific CTG characteristics:

Table 9.4: *TCF outcome-to-CTG mapping*

TCF	Outcome	CTG Mapping	Monitoring Mechanism
1	Fair treatment culture	All five CTG characteristics (aggregate SPAR)	MIDCOT continuous monitoring; quarterly review
2	Products meet needs	Coverage accuracy; pathway consistency	Input-class-stratified SPAR; segment-level drift
3	Clear information	Response completeness	Automated completeness scoring
4	Suitable advice	Coverage accuracy (per individual terms)	Per-policy-type monitoring
5	Products perform as expected	Pathway consistency (temporal)	CUSUM temporal consistency charts
6	No unreasonable barriers	Demographic fairness ($\Delta \leq 0.030$); pathway appropriateness	Segment-stratified SPAR; barrier indicator monitoring

9.5.2 US Clinical Regulatory Requirements (EDT Scenario)

The EDT scenario operates under five US regulatory and accreditation frameworks:

Table 9.5: *EDT regulatory alignment: US clinical frameworks*

Framework	dmaic-AI Implementation
EMTALA	Absolute constraint in governance charter: AI must never recommend deferral of medical screening. Under-triage CTG (Severity 10) directly monitors EMTALA compliance. Demographic fairness CTG ensures non-discriminatory triage
CMS Conditions of Participation	DMAIC-AI provides the QAPI (Quality Assessment and Performance Improvement) structure CMS requires. SPC monitoring satisfies ongoing performance improvement. Audit trails available for CMS surveys
Joint Commission	FMEA maps to National Patient Safety Goals proactive risk assessment. BAR escalation satisfies leadership oversight standards. Control-phase monitoring is continuous performance improvement
State medical practice acts	Human-in-the-loop at every consequential decision point. AI recommends; clinician confirms or overrides. Override concordance is a monitored CTG dimension
HIPAA	PHI handling governed by data minimization in retrieval. All AI interactions logged for HIPAA audit trail. De-identification for governance analytics

The EMTALA alignment is the most governance-critical for the EDT scenario. EMTALA establishes an absolute regulatory floor: the AI system cannot, under any operational state, produce outputs that result in the denial of a medical screening examination. This constraint is encoded in the governance charter as non-negotiable and not subject to override by the governance committee. The under-triage CTG characteristic (target <0.020, Severity 10 in FMEA) operationalizes EMTALA compliance as a continuously monitored behavioral specification.

9.6 Toward a Unified Governance Reference Model

The alignment mappings reveal a consistent structural pattern: existing AI governance frameworks [11, 27] specify *what* must be done but not *how* to do it at the process level. DMAIC-AI provides the process-level implementation methodology. ISO 42001 provides the management system architecture. The NIST AI RMF provides the risk management functions. The EU AI Act provides the regulatory requirements. Jurisdiction-specific regulation (FAIS/TCF for South Africa, EMTALA/CMS for the US) provides deployment-context obligations. DMAIC-AI provides the statistical methodology, the measurement instruments, and the control systems that operationalize all of these [6, 36].

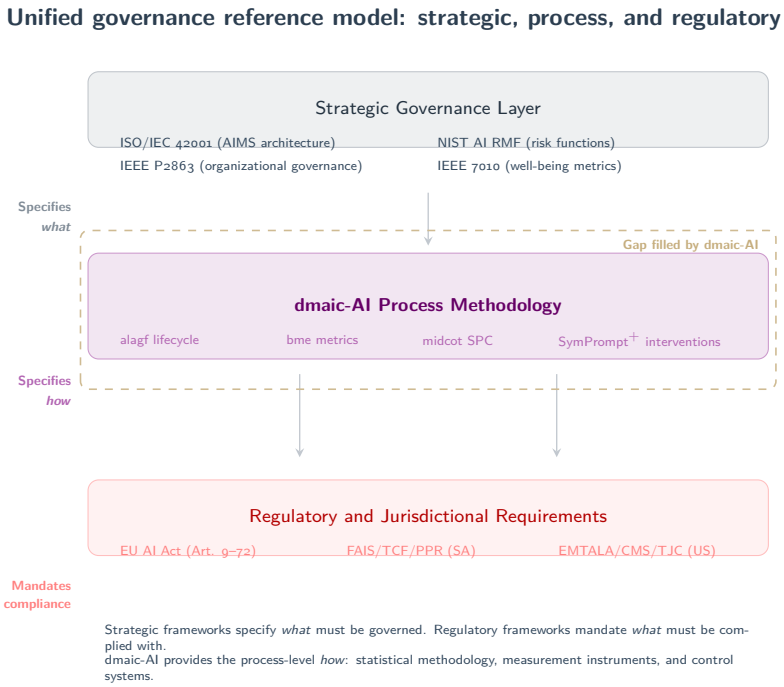


Figure 9.1: Unified governance reference model: three-layer architecture. Strategic frameworks (ISO 42001, NIST AI RMF, IEEE) specify governance requirements. Regulatory frameworks (the EU AI Act, jurisdiction-specific regulations) impose compliance obligations. DMAIC-AI fills the process-level gap between these layers, providing the statistical methodology, measurement instruments (BME), and control systems (MIDCOT) that operationalize both.

Three specific gaps in current standards that DMAIC-AI addresses:

No process capability requirement. ISO 42001 requires monitoring (Cl. 9.1) but does not specify capability indices. A future revision could require ECPI ≥ 1.00 on all governed dimensions as a minimum, with ≥ 1.33 for high-risk systems.

No measurement system validation requirement. The NIST AI RMF calls for "validated measurement approaches" (MEASURE 2.5) but provides no validation specification. MSA/GRR with domain-calibrated thresholds provides the protocol.

No post-market monitoring architecture. The EU AI Act mandates post-market monitoring (Art. 72) but does not specify architecture. MIDCOT's layered SPC regime provides a reference architecture that future regulations could adopt.

Chapter 10

Illustrative Examples: Architecture-Agnostic Governance Scenarios

Chapters 3–7 demonstrated the DMAIC-AI framework through three parallel worked examples, each developed with full phase-by-phase depth across all five DMAIC phases. The clinical decision support system (CDSS) exercises governance under the EU AI Act high-risk clinical classification. The insurance claims triage system (ICT) exercises governance under South African FAIS/TCF financial conduct regulation with a four-step agentic workflow. The emergency department triage agent (EDT) exercises governance under EMTALA and CMS clinical regulation with per-step and workflow-level SPAR decomposition and Severity-10 failure modes. Complete use-case data profiles are provided in Appendix G (ICT) and Appendix H (EDT).

This chapter broadens the demonstration of applicability with three additional scenario classes, each exercising aspects of the framework not fully surfaced in the primary scenarios. These compressed walkthroughs are deliberately architecture-agnostic: they describe governance challenges and DMAIC-AI responses in terms that apply to any LLM implementation meeting the governability conditions established in Chapter 1.

10.1 Autonomous Agentic Systems

This scenario extends the cascading failure model from the EDT's four-step clinical pipeline to longer action chains (up to eight steps) with consequential autonomous actions. The financial services context introduces governance challenges absent from the primary scenarios: autonomous client communications and portfolio updates that execute without human-in-the-loop confirmation.

10.1.1 Context

A financial services firm deploys an agentic AI system that autonomously executes multi-step workflows [44, 50]: researching market conditions, drafting investment analysis reports, and preparing client communications. Each workflow comprises four to eight sequential tool-use steps [39], with consequential actions such as sending client communications and updating portfolio records. Approximately 1,200 workflows execute daily across market analysis (five-step), client communication (four-step), and compliance reporting (eight-step) domains.

This scenario exercises the cascading failure model: per-step behavioral deviations compound through the action chain. While the EDT scenario in Chapters 3–7 also demonstrates cascading failure analysis (four-step clinical pipeline), this financial services scenario extends the model to longer action chains (up to eight steps) with consequential autonomous actions (client communications, portfolio updates) that the EDT scenario does not include because the EDT maintains human-in-the-loop at every decision point.

10.1.2 Compressed dmaic Walkthrough

Define. The governance charter establishes real-time monitoring. CTG tree decomposes into per-step and workflow-level characteristics. Per-step SPAR target: 0.995; workflow-level: 0.975. High-consequentiality workflows receive separate monitoring strata with tighter BAR thresholds.

Measure. Baseline across 2,400 workflows:

Table 10.1: *Agentic system baseline metric profile*

Metric	Market Analysis	Client Communication
Per-Step SPAR (mean)	0.987	0.979
Workflow-Level SPAR	0.936	0.871
ECI	2.6	2.3
ECPI (Reasoning Coherence)	1.41	1.18
ECPI (Comm. Appropriateness)	N/A	1.05

Per-step SPAR of 0.979 compounds to workflow-level 0.871 across a four-step chain. Compare with the EDT scenario, where per-step SPARs of $0.991 \times 0.974 \times 0.986 \times 0.981 = 0.932$ demonstrate the same multiplicative attrition at higher per-step performance.

Analyze. AI-FMEA identifies cascading failures as the dominant concern. Outdated market data retrieval (Step 1, RPN 160) propagates to reasoning from false premises (Step 2, amplified RPN 216). Variation decomposition: 60% Step 2, 25% Step 1, 10% Step 4, 5% measurement.

Improve. Three coordinated interventions: SymPrompt$^+$ inter-step verification, constraint injection for consequential actions (human-in-the-loop before client communications), and retrieval recency preference. 2×3 DOE confirms synergistic improvement.

Table 10.2: *Agentic system pre/post-intervention comparison*

Metric	Baseline	Post-Int.	Change
Per-Step SPAR (Client Comms)	0.979	0.994	+0.015
Workflow-Level SPAR (Client Comms)	0.871	0.962	+0.091
ECI (Client Comms)	2.3	3.1	+0.8
ECPI (Reasoning Coherence)	1.18	1.52	+0.34
Workflow-Level SPAR (Market Analysis)	0.936	0.971	+0.035

Workflow-level improvement (+0.091) substantially exceeds per-step improvement (+0.015), demonstrating multiplicative leverage.

Control. Layered architecture: Shewhart per-step, CUSUM workflow-level, EWMA for reasoning coherence drift. Emergency tier triggers automatic suspension for constraint-injection bypass.

10.2 Retrieval-Augmented Generation Systems

This scenario exercises multi-population governance under a shared RAG infrastructure. While the ICT scenario operates a RAG architecture within a single organizational context, this enterprise scenario demonstrates differentiated CTG trees, department-specific intervention designs, and behavioral entropy as a hallucination proxy across three user populations sharing a common retrieval corpus.

10.2.1 Context

An enterprise deploys a RAG knowledge management system [12, 23] serving 10,000 employees across Legal, Engineering, and HR departments. 500,000-document corpus with department-specific configurations [13]. Approximately 3,000 queries per day.

This scenario exercises department-specific CTG trees with shared and divergent governance dimensions, behavioral entropy as a hallucination proxy, and corpus-update-triggered re-baselining. The ICT scenario in Chapters 3–7 also operates on a RAG architecture, but with a single organizational context; this scenario demonstrates multi-population governance under a shared system.

10.2.2 Compressed dmaic Walkthrough

Define. Department-specific CTG trees share common dimensions (retrieval accuracy, grounding fidelity, completeness) with department-specific additions: citation precision for Legal, version specificity for Engineering, policy sensitivity for HR. Differentiated SPAR targets: 0.960 (Legal), 0.970 (Engineering), 0.950 (HR).

Measure. Grounding fidelity operationalized through claim-level classification. Behavioral entropy (Shannon entropy of the grounding distribution) serves as a hallucination proxy. Baseline across 42,000 queries:

Table 10.3: *Enterprise RAG system baseline metric profile*

Metric	Legal	Engineering	HR
SPAR	0.912	0.951	0.893
ECI	2.5	2.9	2.2
ECPI (Grounding Fidelity)	1.22	1.48	1.08
ECPI (Dept-Specific)	0.98	1.31	0.87
Behavioral Entropy	0.64	0.41	0.72

HR exhibits the highest behavioral entropy (0.72), indicating the least concentrated grounding distribution and the highest hallucination risk.

Analyze. Stratified root cause analysis. Legal: semantic mismatch in cross-referenced documents (55%). Engineering: version-ambiguous retrieval (65%). HR: corpus ambiguity from overlapping policy revisions (40%) plus missing policy-sensitivity instructions (35%). AI-FMEA prioritizes HR policy sensitivity gap (RPN 189).

Improve. Department-specific interventions. DOE confirms per-department improvements:

Table 10.4: *RAG system pre/post-intervention comparison*

	Legal		Eng.		HR	
Metric	Pre	Post	Pre	Post	Pre	Post
SPAR	0.912	0.968	0.951	0.982	0.893	0.957
ECI	2.5	3.2	2.9	3.5	2.2	3.0
ECPI (Ground.)	1.22	1.61	1.48	1.72	1.08	1.44
Beh. Entropy	0.64	0.38	0.41	0.22	0.72	0.41

All departments achieve or exceed targets. HR shows the largest absolute improvement.

Control. Department-specific monitoring strata. CUSUM for grounding fidelity. Behavioral entropy as an early-warning indicator. Differentiated re-baselining cadences.

10.3 Continuous Learning and Fine-Tuning Pipelines

This scenario exercises lifecycle governance of continuous learning, the most invasive intervention class in the AAI taxonomy. The three primary scenarios in Chapters 3–7 resolved their governance gaps without fine-tuning; this scenario demonstrates the governance methodology when the root cause is embedded in the training data curation protocol and is not addressable through retrieval or prompt interventions.

10.3.1 Context

A technology company operates a customer service LLM with monthly fine-tuning on curated interaction data [25]. 50,000 daily interactions. Eleven fine-tuning cycles completed.

This scenario exercises lifecycle governance of continuous learning: managing authorized process changes that modify behavioral distributions. It introduces the stability constraint and the MIDCOT dual-mode monitoring regime. The three primary scenarios in Chapters 3–7 did not require fine-tuning interventions; this scenario demonstrates the governance methodology for the most invasive intervention class in the taxonomy.

10.3.2 Compressed dmaic Walkthrough

Define. Each fine-tuning cycle is classified as an authorized process change requiring mandatory re-baselining. Five governed dimensions. SPAR target: 0.965.

The governance charter introduces the *stability constraint*: post-tuning metrics must remain within 10% of their pre-tuning values on all non-target dimensions; violations trigger an automatic reversion to the pre-tuning checkpoint.

Measure. Dual-mode protocol. Pre-Cycle 12 baseline:

Table 10.5: *Customer service LLM pre-tuning baseline (Cycle 12)*

Metric	Product Support	Billing	Account Mgmt
SPAR	0.958	0.971	0.943
ECI	2.8	3.1	2.5
ECPI (Response Accuracy)	1.38	1.52	1.25
ECPI (Tone Consistency)	1.45	1.41	1.48
ECPI (Escalation Approp.)	1.12	1.28	0.95

Three-cycle degradation trend in escalation appropriateness: ECPI declined from 1.18 (Cycle 9) to 0.95 (Cycle 11).

Analyze. Root cause: curation protocol selects high-quality resolutions, systematically excluding correctly escalated interactions. Training signal suppresses escalation behavior. Confirmed (CPS 8.1).

Improve. Training data correction (escalation-inclusive curation, 15%), SymPrompt^{+} escalation guidance, knowledge base update. DOE validation. Stability constraint satisfied on all non-target dimensions.

Table 10.6: *Post-intervention validation (Account Management)*

Metric	Pre-Tuning	Post-Int.	Stability
SPAR	0.943	0.974	Pass
ECI	2.5	3.2	Pass
ECPI (Escalation)	0.95	1.38	Target dim.
ECPI (Response Acc.)	1.25	1.34	Pass (+7.2%)
ECPI (Tone Cons.)	1.48	1.44	Pass (−2.7%)

Control. Dual-mode MIDCOT: validation mode (14-day post-tuning) and operational mode. Change registry records each cycle. Cycle 13 validation mode detects tone consistency violation (−12.5%, exceeding stability constraint). Automatic reversion. Curation protocol amended.

10.4 Cross-Scenario Synthesis

The six scenarios (CDSS, ICT, and EDT in Chapters 3–7, plus three in this chapter) exercise the DMAIC-AI framework across fundamentally different deployment contexts. Five structural observations emerge:

Retrieval corpus currency is the universal governance concern. All four RAG-based scenarios (CDSS, ICT, EDT, Enterprise RAG) experienced drift events traceable to retrieval corpus currency failures. The SIPOC analysis in the Define phase correctly identified this interface in every case. This convergence suggests that automated corpus-currency monitoring should be a default MIDCOT component for any RAG-augmented deployment.

The cascading failure model differentiates agentic from single-inference governance. The EDT and financial services scenarios demonstrate that per-step SPAR of 0.979 compounds to workflow-level SPAR of 0.871 across a four-step chain. The framework's per-step specification envelope and multiplicative SPAR decomposition are structural necessities for agentic governance.

Jurisdiction-specific regulation drives CTG tree complexity. The CDSS (4 CTG, EU AI Act), ICT (5 CTG, FAIS/TCF), and EDT (6 CTG, EMTALA/CMS/EU AI Act) demonstrate that the number of governed behavioral dimensions scales with the breadth of applicable regulation. The framework accommodates this through the VOG multi-stakeholder weighting model and the specification envelope's dimensional flexibility.

Continuous learning demands formal non-stationarity management. The fine-tuning scenario demonstrates governance constructs (stability constraint, dual-mode monitoring, automatic reversion) with no classical SPC analog, applicable whenever authorized process changes are part of the operational lifecycle.

The intervention taxonomy's least-invasive-first principle is empirically validated. All six scenarios resolved governance gaps with the least invasive available intervention. The three primary scenarios required only data-category and configuration-category interventions (no fine-tuning). The customer service scenario required only fine-tuning because the root cause lay in the training data curation protocol, which is not addressable through retrieval or prompt-based interventions.

10.5 Second-Cycle DMAIC: Demonstrating Repeatable Governance

Chapters 3–7 developed the initial DMAIC cycle for each scenario: the governance initialization cycle that establishes specifications, baselines, and monitoring. The drift events documented in Section 7.3 triggered targeted Improve cycles (corpus updates with re-baselining), but these were scoped interventions within the existing control plan, not full DMAIC re-initiations. This section demonstrates a complete second-cycle DMAIC triggered by a T3 Critical escalation in the EDT scenario, validating the "repeatable process" claim from Chapter 2.

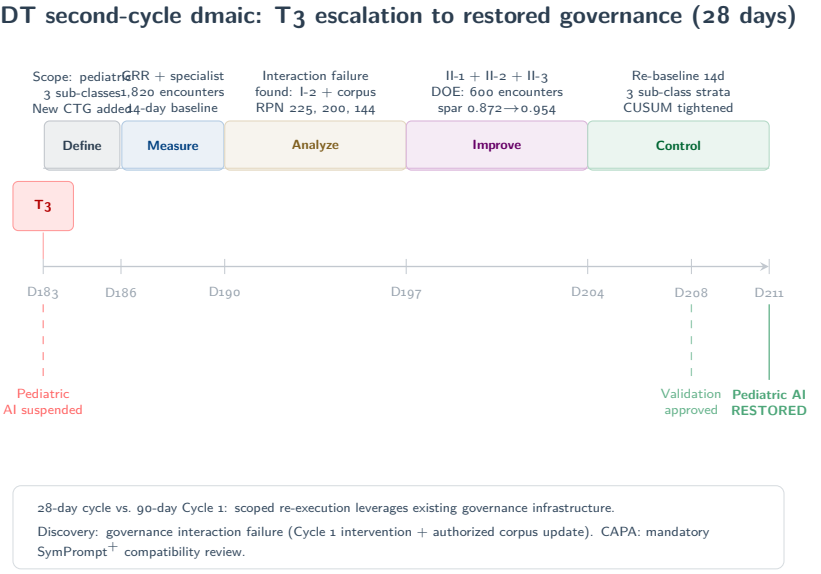


Figure 10.1: EDT second-cycle DMAIC timeline: T3 Critical escalation (Day 183) triggers pediatric AI suspension and a scoped five-phase governance cycle completed in 28 days (vs. 90 days for Cycle 1). The compressed timeline demonstrates that escalation-triggered cycles leverage existing governance infrastructure from prior cycles.

10.5.1 Triggering Event: T3 Critical Escalation

Six months post-deployment, the EDT system encounters a compound escalation event that exceeds the scope of a targeted Improve cycle:

Signal 1 (Day 182)
MIDCOT Shewhart chart for demographic fairness detects a BAR violation: fairness delta for pediatric patients exceeds the 0.040 USL at 0.047, driven by a pattern where pediatric patients presenting with abdominal pain receive systematically lower acuity scores than adults with equivalent symptom severity.

Signal 2 (Day 183)
CUSUM for under-triage rate crosses the Warning threshold for the pediatric input class: under-triage rate has drifted from 0.018 (post-intervention) to 0.034 over a six-week period.

Signal 3 (Day 183)
Multivariate Hotelling T^2 chart signals a joint behavioral shift across acuity accuracy, under-triage, and demographic fairness simultaneously.

cbmes assessment
Pediatric input class CBMES = 0.42 (T3: Critical Drift). Two BAR violations concurrent. Workflow SPAR for pediatric patients: 0.891 (below 0.940 Critical threshold).

The Clinical AI Committee convenes within 2 hours per the Critical escalation protocol. AI triage for pediatric patients is **suspended pending investigation**. Manual triage reinstated for the pediatric population. A full second-cycle DMAIC is authorized.

10.5.2 Second-Cycle Define

The second cycle's Define phase is scoped to the triggering event: pediatric patient triage governance. The existing governance charter, VOG analysis, and authority structure from Cycle 1 remain valid. The second-cycle Define phase focuses on three scoped outputs:

Scope refinement. The pediatric input class is decomposed into three sub-classes for targeted investigation: pediatric abdominal (the trigger pattern), pediatric respiratory, and pediatric other. This sub-classification was not required in Cycle 1 because the pediatric class performed adequately as a single stratum.

Specification review. The Clinical AI Committee reviews whether the Cycle 1 pediatric CTG specifications remain appropriate. Finding: the acuity accuracy target (0.940) and under-triage USL (0.035) are confirmed as clinically necessary. No specification change required. However, the committee adds a *new CTG dimension* for the second cycle: **age-adjusted symptom severity weighting**, which requires the system to demonstrate equivalent symptom-severity assessment across pediatric and adult populations with the same chief complaint. The first-cycle monitoring identified this dimension but did not formally specify it as a CTG characteristic.

COPQ update. The second-cycle COPQ analysis estimates: 6 weeks of degraded pediatric triage ($\sim$108 pediatric patients potentially under-triaged), with an estimated 3.5 adverse outcome events and $1,330,000 in external failure costs attributable to the degradation period. This COPQ estimate justifies the second-cycle governance investment.

10.5.3 Second-Cycle Measure

MSA revalidation. The evaluator panel is augmented with a pediatric emergency medicine specialist (not in the Cycle 1 panel). The GRR study is repeated on 100 pediatric triage outputs across the three new sub-classes. Result: Krippendorff's α for the new age-adjusted severity dimension is 0.81 (adequate). Existing CTG dimensions retain adequate GRR for the pediatric sub-classes.

Targeted baseline. A 14-day focused baseline is collected on pediatric encounters only (not the full 30-day cross-population baseline of Cycle 1). 1,820 pediatric encounters evaluated. Baseline confirms the degradation: pediatric abdominal SPAR = 0.872 (vs. 0.918 in Cycle 1 baseline). Fairness delta for pediatric abdominal: 0.052 (violating the 0.040 USL).

10.5.4 Second-Cycle Analyze

Fishbone analysis. Under Data: the retrieval corpus was updated with 2026 pediatric clinical practice guidelines (authorized change, Day 145), but the new guidelines use a different symptom severity scoring framework for children than the adult guidelines. Under Configuration: the SymPrompt$^+$ acuity verification module (Intervention I-2 from Cycle 1) contains the instruction "evaluate against atypical criteria for the patient's age and sex," but does not specify the pediatric-specific severity scale introduced in the 2026 guidelines. Under Model: the fine-tuned model was trained predominantly on adult ED encounters (2.4M total, of which only ∼216,000 [9%] were pediatric), creating a training-data-driven bias toward assessing adult symptom severity.

AI-FMEA.

Table 10.7: *EDT second-cycle AI-FMEA (pediatric)*

Failure Mode	Step	S	O	D	RPN	Root Cause
Pediatric abdominal pain severity underestimation	2	9	5	5	225	SymPrompt$^+$ lacks pediatric severity scale mapping; corpus contains conflicting adult/pediatric scales
Age-differential acuity: equivalent symptoms scored lower for children	2	10	4	5	200	Training data imbalance (9% pediatric); model learned adult symptom baseline
Outdated pediatric vital sign reference ranges in retrieval corpus	1	8	3	6	144	2026 guideline update changed pediatric vital sign thresholds; old and new coexist in corpus

Variation decomposition. The pediatric SPAR gap of 0.088 (0.960 target minus 0.872 observed) decomposes: SymPrompt$^+$ severity scale gap: 45%; training data imbalance: 30%; retrieval corpus conflict: 18%; measurement: 7%.

Critical finding: The primary root cause (45%) is a *configuration-category* failure created by the interaction between a Cycle 1 intervention (I-2 SymPrompt$^+$ acuity verification) and a subsequent authorized corpus update. The SymPrompt$^+$ module was validated against the pre-update corpus; the post-update corpus introduced a pediatric severity scale that the module does not reference. This is a *governance interaction failure*: two individually valid governance actions (the Cycle 1 intervention and the authorized corpus update) produced a compound behavioral effect that neither would produce alone.

10.5.5 Second-Cycle Improve

Three scoped interventions:

Table 10.8: *EDT second-cycle interventions*

#	Intervention	Mechanism	Reversibility
II-1	SymPrompt$^+$ pediatric severity module	New sub-module within I-2 that maps pediatric-specific severity scales from 2026 guidelines. Activates when patient age <18	Full: version registry
II-2	Retrieval corpus deconfliction	Remove superseded pediatric vital sign reference ranges; tag all pediatric guidelines with age-applicability metadata	Full: corpus versioning
II-3	Pediatric calibration prompt	Additional instruction: "For patients under 18, apply pediatric-specific vital sign thresholds and symptom severity scales. Do not apply adult baselines to pediatric presentations"	Full: version registry

DOE. 3 × 2 design: 3 SymPrompt$^+$ configurations (baseline, II-1 only, II-1 + II-3) × 2 corpus conditions (pre-deconfliction, post-deconfliction). 600 pediatric encounters (200/cell). Result: II-1 + II-3 + deconflicted corpus achieves pediatric SPAR = 0.954 (vs. 0.872 pre-intervention). Fairness delta: 0.021 (below 0.040 USL). Under-triage for pediatric abdominal: 0.019 (below 0.020 target).

Validation. 400 pediatric encounters across all sub-classes. No degradation on adult populations confirmed ($p < 0.001$). Intervention approved.

10.5.6 Second-Cycle Control

Control plan update. The pediatric input class monitoring is enhanced: three sub-class strata (abdominal, respiratory, other) replace the single pediatric stratum. The new age-adjusted severity CTG is added to the MIDCOT monitoring configuration. CUSUM parameters for pediatric under-triage are tightened: $k = 0.3\sigma$ (from 0.5σ), reducing ARL_1 from 14 days to 8 days for the pediatric class.

Re-baselining. 14-day window with enhanced sensitivity (80% of operational BAR). Post-re-baselining pediatric metrics: SPAR = 0.958 (approaching 0.960 target); fairness delta = 0.019; under-triage = 0.017. BAR thresholds recalculated from the post-Cycle 2 distribution.

Governance interaction prevention. The second cycle identifies a new failure mode class: *intervention-change interaction*, where a previously validated SymPrompt$^+$ module becomes inadequate following an authorized corpus update. The CAPA specifies: all future corpus updates that modify clinical guideline content must trigger a SymPrompt$^+$ compatibility review before the re-baselining window opens. This requirement is added to the re-baselining protocol in the control plan.

AI triage restored for pediatric patients. The suspension (Day 183) is lifted at Day 211 (28 days). The complete second-cycle DMAIC (Define through Control) was executed in 28 days, compared to the initial cycle's ∼90-day timeline. The compressed timeline reflects the scoped nature of escalation-triggered cycles: the governance infrastructure (authority, measurement system, monitoring architecture) from Cycle 1 remained operational, and only the targeted specifications, baselines, and interventions required re-execution.

10.5.7 Second-Cycle Governance Implications

The second-cycle walkthrough validates three architectural claims:

Repeatability. The DMAIC cycle re-initiated from a T3 escalation and produced a complete governance response (new specifications, targeted baseline, root cause analysis, validated interventions, updated control plan) within 28 days. The governance infrastructure from Cycle 1 provided the foundation; the second cycle added to rather than replaced it.

Scoping. The escalation context (pediatric fairness degradation) defined the cycle scope. Not every second-cycle DMAIC requires a full five-phase re-execution across all input classes. The scoped approach preserved operational governance for non-affected populations while concentrating resources on the triggered concern.

Governance interaction detection. The second cycle discovered a failure mode class (intervention-change interaction) that the first cycle could not have anticipated: it arises only when Cycle 1 interventions interact with subsequent authorized changes. This discovery validates the cyclical architecture: governance knowledge accumulates across cycles. The CAPA from Cycle 2 (mandatory

SymPrompt^{+} compatibility review for corpus updates) prevents recurrence of this failure class in future cycles.

Chapter 11

Limitations, Open Problems, and Future Research Directions

The DMAIC-AI framework provides a structured, rigorous methodology for LLM behavioral governance, but it operates within boundary conditions that must be stated explicitly. The three parallel scenarios developed in Chapters 3–7 and the three compressed scenarios in Chapter 10 provide empirical grounding for these limitations: the boundary conditions are not merely theoretical but are observable in the governance data.

11.1 Framework Boundary Conditions

Emergent behavior at scale [15]. Chapter 1 identified emergent behavior as a governance boundary condition. The framework can monitor for emergent behaviors (through MIDCOT anomaly detection), classify them (through the AI failure taxonomy), and respond (through tiered escalation). It cannot predict them. None of the six scenarios encountered emergent behavior during the documented governance periods; this absence does not constitute evidence that the framework prevents emergence, only that the monitoring infrastructure would detect and classify it if it occurred.

Foundation model opacity. The behavioral-first governance posture accepts that internal mechanisms are opaque and focuses on externally observable outputs. This means the Analyze phase can attribute root causes to the model category without identifying specific internal mechanisms. In practice, all three primary scenarios attributed their root causes to the data or configuration categories rather

than to the model category. This pattern may reflect the relative accessibility of data/configuration causes to investigation rather than their actual prevalence.

Multi-model architectures. The framework applies to each model individually, and SPAR's cascading failure model addresses multi-step workflows. The EDT scenario's four-step pipeline demonstrates this capability. However, governance of emergent behavior arising from model-to-model interactions remains an open problem.

Evaluator scalability. The ICT scenario processes ∼510 claims per day with automated scoring supplemented by expert panel evaluation. The EDT scenario processes ∼258 encounters per day with 100% automated scoring and 15% expert overlay. These cadences are achievable. For systems processing millions of daily requests, the fundamental tension between measurement validity and scalability remains unresolved.

Irreducible measurement uncertainty. The CDSS GRR study demonstrated marginal reliability for uncertainty disclosure ($\alpha_K = 0.55$). The EDT study showed borderline reliability for reasoning transparency ($\alpha_K = 0.78$). These measurement imprecisions propagate through all downstream calculations. Standards bodies specifying measurement requirements should explicitly account for this irreducible contribution.

Jurisdiction-specific regulatory fragmentation. The three primary scenarios operate under three different regulatory regimes: EU AI Act (CDSS), South African FAIS/TCF (ICT), and US EMTALA/CMS (EDT). The framework accommodates this through jurisdiction-specific VOG requirements and CTG tree structures. However, for organizations operating across multiple jurisdictions simultaneously, the specification envelope must satisfy the most restrictive applicable regulation on each dimension, which may produce specification limits that are operationally infeasible. This cross-jurisdictional specification problem has no current solution within the framework.

11.2 Open Measurement and Methodological Problems

Ground truth for behavioral conformance. For dimensions with objective reference standards (factual accuracy, guideline concordance), measurement validity is established. For subjective dimensions (tone appropriateness, reasoning transparency), ground truth is constructed through evaluator consensus, introducing circularity. The EDT reasoning transparency dimension exemplifies this: evaluator

disagreement on what constitutes a "clinically valid reasoning chain" limits the precision of all downstream governance calculations.

Real-time SPC for streaming inference. The control chart architectures assume batch-mode monitoring. For safety-critical applications like the EDT scenario, real-time monitoring during inference would be preferable. Real-time SPC requires evaluation methods that operate with inference latency while maintaining governance-grade validity.

Causal governance modeling. The Analyze phase employs observational and quasi-experimental methods. Full causal modeling would provide stronger causal claims. The ICT and EDT scenarios both confirmed root causes through controlled comparison (outputs generated with current vs. outdated corpus on matched inputs); this quasi-experimental approach is sufficient for data-category causes but may be inadequate for model-category causes requiring mechanistic interpretability.

Formal verification integration. SPC provides probabilistic assurance, not deterministic guarantees. For the EDT scenario's EMTALA constraint (the absolute prohibition on under-triage recommendations), formal verification, providing mathematical proof of the constraint, would be more appropriate than statistical monitoring. Integration of formal verification with SPC-based governance is a promising research direction.

11.3 Research Agenda for the Governance Community

For standards bodies. Current AI governance standards require measurement but do not specify instruments. Future standards should incorporate explicit process capability requirements (ECPI ≥ 1.00 minimum, ≥ 1.33 for high-risk), MSA requirements with domain-calibrated thresholds, and specifications for post-market monitoring architecture. The three scenarios demonstrate that such requirements are operationally achievable: all three achieved ECPI ≥ 1.33 post-intervention.

For AI governance researchers. Priorities include: adaptive control charts with formal sensitivity guarantees for non-normal processes; multivariate capability indices for correlated behavioral dimensions; ground truth construction methodologies with formal uncertainty quantification; and the intersection of SPC and mechanistic interpretability. The three drift-event case records provide empirical data to validate the drift detection methodology.

For the quality management community. This application represents the first major extension of Six Sigma beyond manufacturing and service processes. The community should test the AAI classifications against additional process domains, validate the BME Metric Suite against empirical deployment data, and refine the methodology through practitioner feedback. The ICT and EDT use-case data profiles (Appendices G and H) provide the level of operational detail required for empirical validation exercises.

For regulatory bodies. The South African FSCA and US CMS/Joint Commission frameworks demonstrate that DMAIC-AI can operationalize jurisdiction-specific regulatory requirements through the CTG tree decomposition and specification envelope. Regulatory bodies developing AI-specific conduct requirements should consider specifying process capability thresholds as conformity conditions, with DMAIC-AI or equivalent methodologies providing the implementation pathway.

The monograph closes with the claim that opened it: LLM and agentic AI systems are stochastic processes with measurable behavioral outputs, identifiable drift characteristics, and classifiable failure modes. The DMAIC methodology, extended through the AAI taxonomy and the BME Metric Suite, provides a sufficient methodological scaffold for rigorous, auditable AI governance. The framework is not complete; the open problems identified in this chapter require sustained research attention [37]. But the structural claim is secured, and the three parallel scenarios developed in this monograph demonstrate that process intelligence, applied with discipline, extends to the governance of intelligent processes.

Appendix A

DMAIC-AI Phase Summary

This appendix provides a single-page reference for each DMAIC phase as applied to LLM and agentic AI governance.

Phase Transition Logic

Define → Measure: All six governance artifacts are complete and approved.

Measure → Analyze: Governance gaps identified (baseline below specification).

Measure → Control: No governance gaps (baseline meets all specifications with margin).

Analyze → Improve: Prioritized cause register approved by governance committee.

Improve → Control: Validated intervention deployed; re-baselining initiated.

Control → Define: T3+ escalation triggers new DMAIC cycle.

Table A.1: *DMAIC-AI phase summary reference*

Phase	Purpose	Key Artifacts	BME Instruments	AAI Profile
Define	Establish governance charter, behavioral specifications, stakeholder alignment	Governance charter, VOG analysis, CTG tree, specification envelope, input class registry, RACI matrix	CTG characteristics define BME measurement targets	2A / 3D / 3I
Measure	Validate measurement system, establish behavioral baseline, calculate capability indices	GRR validation report, baseline dataset, baseline BME metric report, ongoing data collection plan	ECPI, BAR, ECI, SPAR, CBMES baseline values	2A / 2D / 4I
Analyze	Diagnose root causes of governance gaps, prioritize causes for intervention	Fishbone diagrams, AI-FMEA register, variation decomposition, root cause validation, prioritized cause register	SPAR gap analysis, ECI trend analysis	3A / 3D / 1I
Improve	Design, test, and validate governance interventions targeting validated root causes	Intervention selection, SymPrompt^{+} modulation designs, DOE results, validation report, deployment records	Pre/post ECPI, ECI, SPAR comparison	1A / 2D / 3I
Control	Sustain gains through statistical monitoring; detect drift and escalate	Control chart architecture, MIDCOT configuration, escalation protocol, re-baselining protocol, control plan	BAR enforcement, CBMES escalation, re-baselining	1A / 4D / 4I

Appendix B

Complete Tool Adaptation Taxonomy

This appendix reproduces the complete 40-entry AAI taxonomy from Chapter 8 in a format optimized for practitioner reference. Each entry specifies the tool, its AAI classification, the DMAIC phase in which it is introduced, the Tool Module (TM) reference for detailed procedural guidance, and the rationale for the key adaptation or innovation.

Table B.1: *Complete AAI tool adaptation taxonomy (40 entries mapping to 39 distinct tools and constructs)*

Tool / Construct	AAI	Phase	TM Ref	Key Rationale
SIPOC Diagram	**Adopt**	Define	T1-01	Component mapping applies directly
RACI Matrix	**Adopt**	Define	T1-02	Responsibility assignment domain-independent
Data Collection Plan	**Adopt**	Measure	T2-01	Planned collection applies directly
Operational Definitions	**Adopt**	Measure	T2-02	Unambiguous definition principle universal
5 Whys	**Adopt**	Analyze	T3-01	Iterative drill-down domain-independent
Pareto Analysis	**Adopt**	Analyze	T3-02	Frequency prioritization applies directly
Run Charts	**Adopt**	Analyze	T3-03	Temporal visualization applies directly
Before/After Comparison	**Adopt**	Improve	T4-01	Pre/post logic domain-independent
Run Rules (Western Electric)	**Adopt**	Control	T5-01	Pattern detection applies to any chart

Table B.1: *(continued)*

Tool / Construct	AAI	Phase	TM Ref	Key Rationale
VOG Analysis	**Adapt**	Define	T1-03	Stakeholder scope expanded to full constituency
CTG Tree	**Adapt**	Define	T1-04	Decomposition retargeted to behavioral dimensions
Governance Charter	**Adapt**	Define	T1-05	Extended to lifecycle commitment
Governance VSM	**Adapt**	Define	T1-09	Behavioral/governance/corpus supply chain flows
MSA / GRR	**Adapt**	Measure	T2-03	Reinterpreted for evaluators; thresholds recalibrated
ECPI (Capability)	**Adapt**	Measure	T2-04	Empirical quantiles replace parametric calculation
Fishbone / Ishikawa	**Adapt**	Analyze	T3-04	AI-specific causal categories
AI-FMEA	**Adapt**	Analyze	T3-05	Cascading failure model; governance-calibrated severity
Variation Decomposition	**Adapt**	Analyze	T3-06	Non-normal, correlated data; GRR subtraction
DOE	**Adapt**	Improve	T4-02	AI factors; multi-response optimization
Pilot Testing	**Adapt**	Improve	T4-03	Input class replication required
Shewhart Charts	**Adapt**	Control	T5-02	BAR replaces parametric limits
CUSUM Charts	**Adapt**	Control	T5-03	Adaptive reference values for non-stationarity
EWMA Charts	**Adapt**	Control	T5-04	Autocorrelation accommodation
Control Plan	**Adapt**	Control	T5-05	AI monitoring and lifecycle content
Specification Envelope	**Innovate**	Define	T1-06	Multidimensional joint conformance
Input Class Registry	**Innovate**	Define	T1-07	Unbounded input space stratification
VOG Weighting	**Innovate**	Define	T1-08	Authority-asymmetric multi-class prioritization
COPQ Framework	**Innovate**	Define	T1-10	Probabilistic AI failure economics
BAR Derivation	**Innovate**	Measure	T2-05	Empirical quantile thresholds
Baseline Distribution	**Innovate**	Measure	T2-06	Full distribution characterization
CBMES Baseline	**Innovate**	Measure	T2-07	Composite escalation metric
Ongoing Collection	**Innovate**	Measure	T2-08	Configuration-aware evaluator-managed protocol
Prioritized Cause Register	**Innovate**	Analyze	T3-07	Composite priority score integrating confidence
Intervention Taxonomy	**Innovate**	Improve	T4-04	Speed/reversibility/scope classification

Table B.1: *(continued)*

Tool / Construct	AAI	Phase	TM Ref	Key Rationale
SymPrompt^{+} Modulation	**Innovate**	Improve	T4-05	Governed behavioral modulation instrument
Reversibility/Rollback	**Innovate**	Improve	T4-06	Rollback as governance requirement
MIDCOT Configuration	**Innovate**	Control	T5-06	Production-grade AI monitoring system
BAR Enforcement	**Innovate**	Control	T5-07	Governance-gated threshold management
Tiered Escalation	**Innovate**	Control	T5-08	CBMES-driven graduated response
Re-Baselining Protocol	**Innovate**	Control	T5-09	Non-stationarity management

Aggregate distribution: 9 Adopt (22%), 15 Adapt (38%), 16 Innovate (40%). Note: BAR appears in both Measure (derivation, TM-T2-05) and Control (enforcement, TM-T5-07), producing 40 entries mapping to 39 distinct tools and constructs.

Appendix C

BME Metric Suite Operational Definitions

This appendix provides the operational definitions for all BME Metric Suite constructs. Each definition specifies the metric name, formula or calculation method, interpretation, and governance function.

ecpi: Entropy-Calibrated Performance Index

Governance function: Process capability analogue. Replaces C_p/C_{pk}.

Definition: Quantifies the ratio of specification space to observed behavioral variation using empirical quantiles:

$$\text{ECPI} = \frac{LSL_{\text{spec}} - Q_\alpha}{Q_{0.5} - Q_\alpha} \tag{C.1}$$

where $Q_{0.5}$ is the median of the behavioral metric distribution, Q_α is the α-quantile (typically 0.00135, corresponding to the 0.135th percentile), and LSL_{spec} is the lower specification limit from the CTG tree.

Interpretation: ECPI $\geq$ 1.33: capable. 1.00–1.33: marginal. $<$ 1.00: incapable (behavioral distribution extends below specification).

Two-sided specification: For behavioral metrics with both lower and upper specification limits (where both excessive and insufficient behavioral expression are nonconformant), the upper ECPI is calculated as:

$$\text{ECPI}_U = \frac{USL_{\text{spec}} - Q_{1-\alpha}}{Q_{1-\alpha} - Q_{0.5}} \tag{C.2}$$

where USL_{spec} is the upper specification limit and $Q_{1-\alpha}$ is the $(1-\alpha)$-quantile (typically 99.865th percentile). The reported ECPI for a two-sided specification is:

$$\text{ECPI} = \min(\text{ECPI}_L,\ \text{ECPI}_U) \tag{C.3}$$

where ECPI_L is the lower formulation given above. This parallels the relationship between C_{pk} and its two-sided components in classical SPC, with empirical quantiles replacing parametric estimates.

Calculation scope: Per CTG characteristic, per input class. Aggregate ECPI is the minimum across all dimensions and classes.

bar: Behavioral Assurance Rating

Governance function: Control limit analogue. Replaces $\bar{x} \pm 3\sigma$.

Definition: Empirical quantile-based thresholds defining the boundaries of expected behavioral variation. Set at the 0.135th and 99.865th percentiles of the validated baseline behavioral distribution.

Interpretation: Points within BAR thresholds reflect common-cause variation. Points outside trigger special-cause investigation per the escalation protocol.

Recalculation triggers: Re-baselining following authorized system changes (model updates, corpus changes, SymPrompt$^+$ modulations). BAAGF approval required before updated thresholds are activated.

eci: Entropy Control Index

Governance function: Sigma-level equivalent. Replaces Z-score.

Definition: Distance from the current behavioral metric level to the nearest BAR threshold, expressed in units of empirical behavioral standard deviation:

$$\text{ECI} = \frac{\min(\tilde{x} - \text{BAR}_L,\ \text{BAR}_U - \tilde{x})}{s_{\text{emp}}} \tag{C.4}$$

where $\tilde{x}$ is the current period's behavioral metric value (or the median of the monitoring window), BAR_L and BAR_U are the lower and upper BAR thresholds respectively, and s_{emp} is the empirical standard deviation of the baseline behavioral distribution. For one-sided specifications (the typical case for behavioral metrics with lower bounds only), ECI reduces to:

$$\text{ECI} = \frac{\tilde{x} - \text{BAR}_L}{s_{\text{emp}}} \tag{C.5}$$

Interpretation: ECI $\geq$ 3.0: Three Sigma equivalent behavioral stability. ECI $\geq$ 6.0: Six Sigma equivalent. Values below 3.0 indicate insufficient margin between operating performance and governance thresholds.

spar: Stochastic Process Adherence Ratio

Governance function: Process yield analogue. Replaces DPMO.

Definition: Proportion of outputs, within a defined input class and measurement period, that fall within the behavioral specification envelope on all governed CTG dimensions simultaneously.

Interpretation: SPAR = 0.997 indicates 99.7% joint behavioral conformance (Three Sigma equivalent yield). The joint conformance requirement distinguishes SPAR from per-dimension pass rates: a system may score highly on each dimension independently while failing to achieve joint conformance.

Agentic extension: For multi-step workflows, per-step SPAR is calculated for each step, and workflow-level SPAR is calculated as the product across steps (multiplicative cascading model).

cbmes: Composite Behavioral Measurement Evaluation Score

Governance function: Aggregate escalation metric. Drives BAAGF tier classification (T0–T5).

Definition: Weighted composite of ECPI, ECI, and SPAR, calibrated to the deployment risk profile. The aggregation proceeds in two stages: component normalization and weighted combination.

Component Normalization. Each component is normalized to a $[0, 1]$ governance concern scale where 0 indicates no concern and 1 indicates maximum concern:

$$C_{\text{ECPI}} = \max\left(0,\ 1 - \frac{\text{ECPI}_{\text{min}} - 1.00}{1.33 - 1.00}\right) \tag{C.6}$$

$$C_{\text{ECI}} = \max\left(0,\ 1 - \frac{\text{ECI}_{\text{min}}}{6.0}\right) \tag{C.7}$$

$$C_{\text{SPAR}} = \max\left(0,\ 1 - \frac{\text{SPAR}_{\text{min}} - \text{SPAR}_{\text{floor}}}{\text{SPAR}_{\text{target}} - \text{SPAR}_{\text{floor}}}\right) \tag{C.8}$$

where ECPI_{min} is the minimum ECPI across all governed dimensions and input classes, ECI_{min} is the minimum ECI across dimensions, SPAR_{min} is the minimum SPAR across input classes, $\text{SPAR}_{\text{target}}$ is the governance-specified target (typically 0.997), and $\text{SPAR}_{\text{floor}}$ is the minimum acceptable yield (typically 0.900). Each component is clamped to $[0, 1]$.

Weighted Aggregation:

$$\text{CBMES} = w_{\text{ECPI}} \cdot C_{\text{ECPI}} + w_{\text{ECI}} \cdot C_{\text{ECI}} + w_{\text{SPAR}} \cdot C_{\text{SPAR}} \tag{C.9}$$

where $w_{\text{ECPI}} + w_{\text{ECI}} + w_{\text{SPAR}} = 1$. Weights are governance decisions specified in the governance charter, reflecting the relative priority of capability, stability, and yield in the deployment context. Default weights for high-risk deployments: $w_{\text{SPAR}} = 0.45$, $w_{\text{ECPI}} = 0.35$, $w_{\text{ECI}} = 0.20$, reflecting the primacy of joint conformance yield in safety-critical contexts.

Interpretation: CBMES 0.00–0.10: T0 (Optimal). 0.11–0.25: T1 (Stable). 0.26–0.50: T2 (Moderate Drift). 0.51–0.75: T3 (High Risk). 0.76–0.90: T4 (Critical). >0.90: T5 (Emergency). Higher values indicate greater governance concern.

Appendix D

Acronym and Symbol Glossary

Acronyms

Acronym	Definition
AAI	Adopt, Adapt, Innovate (tool classification taxonomy)
AHRS	Adaptive Harm Reduction Score
AIMS	AI Management System (ISO/IEC 42001)
ALAGF	Adaptive Lifecycle Agentic Governance Framework
ANOVA	Analysis of Variance
BAAGF	Behavioral Assurance and Agentic Governance Framework
BAR	Behavioral Assurance Rating (control limit analogue)
BME	Behavioral Measurement of Entropy
CBMES	Composite Behavioral Measurement Evaluation Score
CDSS	Clinical Decision Support System
CPS	Composite Priority Score
CTG	Critical-to-Governance (analogue to CTQ)
CTQ	Critical-to-Quality
CUSUM	Cumulative Sum (control chart type)
DAST	Dynamic Adversarial Stress Testing
DMAIC	Define, Measure, Analyze, Improve, Control
DOE	Design of Experiments
DPMO	Defects Per Million Opportunities
EChPI	Echo Chamber Propagation Index
ECI	Entropy Control Index (sigma-level equivalent)
ECPI	Entropy-Calibrated Performance Index (capability analogue)
EU AI Act	European Union Artificial Intelligence Act (Regulation 2024/1689)

EWMA	Exponentially Weighted Moving Average (control chart type)
FMEA	Failure Mode and Effects Analysis
GRR	Gauge Repeatability and Reproducibility
IQD	Information Quality Deviation
LLM	Large Language Model
MIDCOT	Multi-Dataset IQ Drift and Cost Optimization Training
MSA	Measurement System Analysis
NIST AI RMF	National Institute of Standards and Technology AI Risk Management Framework
PTDI	Prompt Traceability and Disclosure Index
RACI	Responsible, Accountable, Consulted, Informed
RAG	Retrieval-Augmented Generation
RPN	Risk Priority Number
SIPOC	Supplier, Input, Process, Output, Customer
SPAR	Stochastic Process Adherence Ratio (yield analogue)
SPC	Statistical Process Control
SymPrompt+	Symbiotic Prompt Modulation (operational governance instrument)
VOC	Voice of the Customer
VOG	Voice of Governance (analogue to VOC)

Key Symbols

Symbol	Definition
C_p, C_{pk}	Classical process capability indices (replaced by ECPI)
$D(t)$	Drift Index (MIDCOT-native metric)
LSL_{spec}	Lower specification limit from the CTG tree
$P(y \mid x)$	Conditional probability of output y given input x
Q_α	α-quantile of the empirical behavioral distribution
$Q_{0.5}$	Median of the empirical behavioral distribution
$\bar{x} \pm 3\sigma$	Classical control limits (replaced by BAR)
κ	Cohen's kappa (inter-rater agreement)
α	Krippendorff's alpha (inter-rater reliability); also quantile parameter
λ	EWMA smoothing parameter
T0–T5	BAAGF escalation tiers (Optimal through Emergency)

Appendix E

Standards Alignment Cross-Reference Tables

This appendix consolidates the standards traceability mappings from all DMAIC phase chapters into unified cross-reference tables organized by standard.

E.1 ISO/IEC 42001:2023 Cross-Reference

Table E.1: *ISO/IEC 42001:2023 clause-to-DMAIC-AI mapping*

Clause	Requirement	dmaic-AI Implementation
4.1	Organization context	Define: governance trigger assessment, deployment context documentation
4.2	Interested parties	Define: VOG stakeholder analysis (four canonical classes)
4.3	AIMS scope	Define: governance charter scope specification
4.4	AIMS and processes	ALAGF lifecycle architecture; seven-phase BAAGF governance lifecycle
5.1	Leadership commitment	BAAGF governance authority structure; governance charter approval
5.3	Roles and authorities	Define: RACI matrix; BAAGF escalation ownership
6.1	Risk assessment	Analyze: AI-FMEA; CBMES escalation tiers
6.2	AI objectives	Define: CTG tree specification limits; SPAR targets
7.5	Documented information	Define/Control: governance audit trail; version-controlled artifacts

Table E.1: *(continued)*

Clause	Requirement	dmaic-AI Implementation
8.1	Operational planning	Define: CTG trees, specification envelopes, input class registries
8.2	AI risk assessment	Analyze: FMEA register; DAST pre-deployment clearance
9.1	Monitoring and measurement	Measure: BME Metric Suite, MSA validation. Control: MIDCOT SPC monitoring
9.2–9.3	Internal audit and review	Control: plan review cadence; MIDCOT governance reporting
10.1	Nonconformity and correction	Improve: intervention design targeting validated root causes
10.2	Continual improvement	Improve: SymPrompt$^+$ modulations, DOE optimization. DMAIC cyclical integration

E.2 NIST AI RMF 1.0 Cross-Reference

Table E.2: *NIST AI RMF function-to-DMAIC-AI mapping*

Function	Category	dmaic-AI Implementation
GOVERN	1.1 Legal/regulatory	Define: regulatory requirements in VOG analysis
GOVERN	1.2 Accountability	Define: governance charter, RACI matrix; BAAGF authority
GOVERN	1.4 Documentation	Define/Control: governance audit trail; version-controlled registries
MAP	1.1 Purpose/context	Define: governance charter; deployment context documentation
MAP	1.5 Risk tolerances	Define: specification limits; SPAR targets; escalation thresholds
MAP	2.1 Risks mapped	Analyze: Fishbone analysis; AI-FMEA register
MAP	2.3 Scientific integrity	Define: input class registry; measurement methodology
MEASURE	1.1 Measurement approaches	Measure: BME Metric Suite; ECPI/BAR/ECI/SPAR
MEASURE	2.5 Validated approaches	Measure: MSA/GRR validation; baseline establishment
MEASURE	2.6 Performance criteria	Define: CTG specification limits; Measure: operational definitions
MANAGE	Risk prioritization	Analyze: FMEA-based RPN ranking; prioritized cause register
MANAGE	Risk response	Improve: intervention taxonomy; SymPrompt$^+$ modulations

Table E.2: *(continued)*

Function	Category	dmaic-AI Implementation
MANAGE	Monitoring	Control: MIDCOT continuous SPC; T_0–T_5 escalation; re-baselining

E.3 EU AI Act Cross-Reference

Table E.3: *EU AI Act article-to-DMAIC-AI mapping*

Article	Requirement	dmaic-AI Implementation
Art. 9	Risk management system	BAAGF escalation model; FMEA register; DAST clearance
Art. 9(2)	Risk identification	Analyze: AI failure taxonomy; variation decomposition
Art. 9(5)	Testing	Measure: baseline establishment; Improve: DOE validation
Art. 13	Transparency	Define: VOG transparency requirements; SymPrompt$^+$ version registry
Art. 61–62	Serious incidents	Control: T_5 automatic suspension; post-incident review
Art. 72	Post-market monitoring	Control: MIDCOT continuous monitoring; re-baselining

Appendix F

CDSS Worked Example: Consolidated Data

This appendix consolidates all quantitative data from the CDSS (Clinical Decision Support System) worked example that threads through Chapters 3–7.

F.1 System Profile

System identifier
CDSS-2026-001

Deployment context
Hospital clinical decision support

Regulatory classification
EU AI Act high-risk (health-related AI)

Governance authority
Clinical AI Governance Committee (Chair: CMIO)

Review cadence
Quarterly scheduled; unscheduled on T3+ escalation

Input classes
Diagnostic queries, treatment inquiries, drug interaction checks, patient communication assistance

F.2 CTG Tree

Table F.1: *CDSS CTG specification*

Governance Driver	CTG Characteristic	Target	LSL
Evidence-based	Citation accuracy rate	≥0.95	≥0.90
Guideline-current	Guideline concordance rate	≥0.95	≥0.90
Internally consistent	Contradiction avoidance rate	≥0.98	≥0.95
Uncertainty-transparent	Uncertainty disclosure freq.	≥0.90	≥0.85

SPAR target: ≥ 0.997 (joint conformance across all four dimensions).

F.3 Measurement System Validation

Table F.2: *CDSS GRR validation results*

CTG Characteristic	Intra-Rater	Krippendorff's α	Status
Citation accuracy	>90%	0.78	Adequate
Guideline concordance	88%	0.75	Adequate
Uncertainty disclosure	80%	0.55	Marginal
Contradiction avoidance	82%	0.58	Marginal

Governance-grade GRR threshold: 20% (relaxed from manufacturing 10% per governance charter).

F.4 Baseline and Post-Intervention Metrics

Table F.3 consolidates the baseline and post-intervention BME metric values for the CDSS diagnostic query input class. Baseline values were collected during a four-week stability window under frozen model version and retrieval corpus configuration (see Section 4.4 for the data collection protocol). Post-intervention values were

collected following the combined deployment of SymPrompt$^+$ modulation SP-CDSS-2026-003 (detailed uncertainty classification) and the retrieval corpus update (guideline currency tagging and supersession flagging), validated against 400 diagnostic queries using the full evaluator protocol described in Section 4.4.

Three observations warrant emphasis. First, SPAR improved from 0.943 to 0.981, narrowing the gap to the 0.997 governance target by 66%. The residual gap (0.016) is attributable to common-cause variation and edge-case input handling, both identified in the variation decomposition (Section F.6) as requiring longer-term structural intervention rather than prompt-level modulation. Second, the ECPI improvement is concentrated in the two dimensions targeted by the Improve phase: uncertainty disclosure (0.92 to 0.95) and guideline concordance (1.35 to 1.48). Non-target dimensions (citation accuracy, contradiction avoidance) show no statistically significant degradation, confirming that the interventions did not introduce unintended behavioral side effects. Third, ECI improved from 2.8 to 3.4, crossing the Three Sigma governance threshold (ECI $\geq$ 3.0) for the first time. This transition represents a qualitative shift in governance posture: the system moves from a state requiring active improvement to a state eligible for sustained monitoring under the Control phase.

The CBMES values reflect the escalation tier model (Table 2.3): the baseline score of 0.31 places the system in T2 (Moderate Drift), while the post-intervention score of 0.18 places it in T1 (Stable). The improvement reflects reduced entropy concentration in the uncertainty disclosure dimension, consistent with the SymPrompt$^+$ modulation's intended effect of distributing confidence classifications across the high/moderate/low spectrum rather than defaulting to unqualified assertions.

All improvements are statistically significant (Mann-Whitney U, $p < 0.001$; bootstrap 95% CI for SPAR difference: [0.028, 0.048]). Statistical methods follow the nonparametric validation protocol specified in Section 6.4.

Table F.3: *CDSS baseline and post-intervention BME metrics (diagnostic queries)*

Metric	Baseline	Post-Intervention
SPAR	0.943	0.981
ECPI (Uncertainty Disclosure)	0.92	0.95
ECPI (Citation Accuracy)	1.52	1.55
ECPI (Guideline Concordance)	1.35	1.48
ECPI (Contradiction Avoidance)	1.78	1.76
ECI	2.8	3.4
CBMES	0.31 (T2)	0.18 (T1)

F.5 AI-FMEA Register (Corrected)

Table F.4 presents the corrected AI-FMEA register for the CDSS system, constructed per the adapted FMEA protocol (TM-T3-05) specified in Section 5.2. The register reflects the Analyze phase root cause investigation documented in Section 5.4. Three failure modes were identified as primary contributors to the SPAR gap between the observed baseline (0.943) and the governance target (0.997).

The "corrected" designation indicates that severity, occurrence, and detection ratings have been calibrated against the measurement system validation results from Section F.3. Specifically, detection ratings incorporate the GRR findings: failure modes assessed by metrics with marginal measurement reliability (uncertainty disclosure, Krippendorff's $\alpha = 0.55$; contradiction avoidance, Krippendorff's $\alpha = 0.58$) receive detection ratings that reflect the reduced confidence in identifying these failure modes through routine measurement. This calibration prevents the FMEA from assigning artificially optimistic detection scores to failure modes that the measurement system identifies with limited reliability.

The Risk Priority Number (RPN) ranking establishes the intervention priority sequence for the Improve phase. Uncertainty disclosure omission (RPN 252) is classified as Priority 1 because it combines the highest severity ($S = 9$: overconfident clinical recommendations carry direct patient safety implications under the EU AI Act high-risk classification), the highest occurrence ($O = 7$: observed in approximately 8% of diagnostic outputs during baseline collection), and moderate detection difficulty ($D = 4$: the uncertainty disclosure metric captures this failure mode, but with marginal GRR reliability). Guideline concordance failure (RPN 200) is Priority 2, driven primarily by the detection challenge ($D = 5$) associated with assessing whether a recommendation aligns with the most current clinical guidance when the retrieval corpus contains guidelines of mixed vintage. Citation accuracy failure (RPN 96) is Priority 3; its lower occurrence ($O = 4$) and better detection ($D = 3$) reflect the relative maturity of citation verification as a measurable governance dimension.

The AI-specific causal categories assigned to each failure mode follow the adapted Fishbone taxonomy defined in Section 5.2 (TM-T3-01). Confidence miscalibration (uncertainty disclosure omission) is classified under the Model category; guideline concordance failure and citation accuracy failure are classified under the Data category, reflecting their root cause in retrieval corpus quality rather than model-level behavioral drift.

Table F.4: *CDSS AI-FMEA register*

Failure Mode	S	O	D	RPN	Priority
Uncertainty disclosure omission	9	7	4	252	1
Guideline concordance failure	8	5	5	200	2
Citation accuracy failure	8	4	3	96	3

F.6 Variation Decomposition (TM-T3-06)

The variation decomposition partitions total observed variation in uncertainty disclosure ECPI into four source categories, following the adapted ANOVA protocol (TM-T3-06) specified in Section 5.2. This analysis was triggered by the Analyze phase investigation into the SPAR gap and provides the quantitative basis for directing Improve phase intervention resources toward the highest-leverage variation sources. Bootstrap confidence intervals (10,000 resamples, bias-corrected and accelerated) are reported for each attribution estimate to account for the non-normal distributional characteristics of ECPI observations.

The decomposition reveals a dominant assignable-cause structure. Assignable process variation accounts for 57.1% (95% CI: 49.6–64.2%) of total observed variation, with a single fine-tuning round (FT-2026-Q1-03) responsible for 45.9% of the total. Root cause analysis identified the mechanism: the satisfaction-biased curation protocol used in FT-2026-Q1-03 systematically overrepresented user-rated-positive training examples, and these examples disproportionately lacked uncertainty qualifiers. The fine-tuning process consequently reinforced assertive response patterns at the expense of calibrated uncertainty disclosure. This finding illustrates a governance-critical failure mode: training data curation decisions propagating into behavioral drift on governed dimensions that were not explicitly represented in the curation criteria.

Measurement system contribution (GRR) accounts for 15.0% (95% CI: 11.2–19.4%) of total variation. This is consistent with the marginal GRR status reported in Section F.3 for the uncertainty disclosure metric (Krippendorff's $\alpha = 0.55$). The measurement contribution establishes a floor below which variation cannot be attributed to the process with confidence; governance decisions based on uncertainty disclosure ECPI must account for this measurement uncertainty through wider confidence intervals, as specified in the governance charter.

Retrieval staleness contributes 7.8% of total variation, with 3.4% remaining as unattributed residual within this category. The attributable portion reflects cases where outdated clinical guidelines in the retrieval corpus produced recommendations that conflicted with current practice, manifesting as uncertainty disclosure

failures when the system presented superseded guidance without appropriate qualification.

Common-cause variation accounts for 27.9% of the total, decomposed into inherent stochasticity (16.5%) and input variation (11.4%). This establishes the irreducible floor: the proportion of behavioral variation that governance must accommodate rather than eliminate. The inherent stochasticity component reflects the fundamental nondeterminism of autoregressive generation; the input variation component reflects the natural diversity within the diagnostic query input class. Together, these set a theoretical lower bound on achievable ECPI improvement through the intervention classes available in the Improve phase. Reducing variation below this floor would require either narrowing the input class definition (restricting the types of diagnostic queries the system accepts) or fundamentally altering the generation architecture, neither of which falls within the scope of the current DMAIC cycle.

The decomposition directs the Improve phase strategy: the two assignable-cause sources (fine-tuning artifact and retrieval staleness) account for 64.9% of total variation and are tractable through SymPrompt$^+$ modulation and retrieval corpus optimization respectively. The interventions deployed (Section F.7) target these sources. The post-intervention metrics (Section F.4) confirm that the targeted reduction in assignable-cause variation produced the expected improvement in SPAR and ECI.

Total observed variation in uncertainty disclosure ECPI:

Measurement system (GRR)
15.0% (95% CI: 11.2–19.4%)

Assignable process variation
57.1% (95% CI: 49.6–64.2%). Fine-tuning round FT-2026-Q1-03, responsible for 45.9%: satisfaction-biased curation protocol suppressed uncertainty disclosure.

Retrieval staleness
7.8% (3.4% unattributed residual)

Common-cause variation
27.9% (inherent stochasticity 16.5%; input variation 11.4%). Establishes the irreducible floor.

F.7 Interventions Deployed

SymPrompt$^+$ Modulation SP-CDSS-2026-003
Explicit uncertainty classification instructions (high/moderate/low confidence based on evidence strength and recency). AAI: **[I ⋆ Innovate]**.

Retrieval Corpus Update
Clinical guidelines tagged with publication date, issuing body, and supersession status. Outdated guidelines flagged as archived. AAI: **[D ▷ Adapt]**.

Validation: 400 diagnostic queries. All improvements statistically significant (Mann-Whitney U, $p < 0.001$; bootstrap 95% CI for SPAR difference: [0.028, 0.048]). No non-target degradation.

Appendix G

Insurance Claims Triage: Use-Case Data Profile

This appendix provides the complete use-case data profile for the Insurance Claims Triage (ICT) scenario that threads through Chapters 3–7 as a parallel worked example. All values are canonical and internally consistent. The profile is structured to mirror the DMAIC lifecycle: organizational context and system architecture (Sections G.1–G.2), regulatory alignment and behavioral specification (Sections G.3–G.4), baseline measurement (Section G.5), failure analysis (Section G.6), intervention and improvement (Section G.7), control and monitoring (Section G.8), drift event documentation (Section G.9), and standards alignment (Section G.10).

G.1 Organizational Profile

The deploying organization is a mid-size South African short-term insurer operating under FSCA regulatory oversight. This section profiles the organization, its claims operations structure, and the AI governance authority hierarchy established by the Define phase governance charter.

G.1.1 Deploying Organization

Organization type
Mid-size South African short-term insurer

Regulatory jurisdiction
South Africa: FSCA, FAIS Act (Act 37 of 2002), Insurance Act (Act 18 of 2017)

Insurance lines
Personal and commercial short-term: motor, property (homeowner/commercial), liability (public and professional)

Policyholder base
∼620,000 active policies across all lines

Annual claims volume
∼185,000 claims/annum (∼510/day average, peak ∼780/day)

Annual GWP
R4.2 billion (approximately USD 230 million)

Claims ratio
62.4% (industry average: 64.1%)

Geographic footprint
All nine South African provinces; Gauteng (38%), Western Cape (22%), KwaZulu-Natal (16%)

Distribution model
Intermediated (68% broker-originated) and direct (32% digital/call center)

Existing quality framework
ISO 9001:2015 certified; TCF self-assessment annually; no ISO 42001 certification

G.1.2 Claims Operations Structure

The claims operations team comprises 108 staff across eight functional roles, with authority levels tiered by claim value and complexity.

Table G.1: *ICT claims operations staffing*

Role	Headcount	Responsibility
Claims handlers (motor)	42	First-line assessment, coverage verification, authority up to R150,000
Claims handlers (property)	28	Property damage assessment, contractor coordination, authority up to R250,000
Claims handlers (liability)	14	Liability assessment, legal referral, authority up to R100,000
Senior claims managers	8	Escalated claims, authority up to R1,000,000, QA oversight
Fraud investigation unit	6	Fraud detection, investigation, SIU referral
Claims operations manager	1	End-to-end claims process governance
Quality assurance team	4	Monthly QA sampling (∼2%), TCF compliance monitoring
IT/data team (claims)	5	Claims system administration, reporting, data extraction

G.1.3 Governance Authority Structure

The AI governance structure is established by the Define phase governance charter and operates within the insurer's existing risk management framework.

Table G.2: *ICT governance authority hierarchy*

Body	Composition	Authority	Cadence
Claims AI Gov. Committee	CRO (chair), Claims Ops Mgr, Head of IT, Compliance Officer, external actuarial advisor	Suspend/restrict AI triage; approve specification changes; approve re-baselining; authorize new DMAIC cycles	Quarterly review; ad hoc on Critical/Emergency
Operational Gov. Team	2 senior claims managers, 1 QA analyst, 1 data scientist	Monitor MIDCOT dashboards; investigate Advisory/Warning signals; execute re-baselining protocols; maintain audit trail	Daily monitoring; weekly review
Board Risk Committee	NEDs, CRO, CEO	Oversight of AI governance program; risk appetite for AI-assisted claims; regulatory reporting	Semi-annual AI governance report

G.2 AI System Technical Profile

ClaimsTriage AI v2.3 is a RAG-augmented agentic system executing a four-step sequential workflow. This section documents the system architecture, the agentic workflow with cascading failure dynamics, and the retrieval corpus maintenance schedule that governs knowledge currency.

G.2.1 System Architecture

System designation
ClaimsTriage AI v2.3

Core LLM
Commercial API-accessed foundation model (GPT-4 class); temperature 0.2 (triage), 0.4 (fraud narrative)

Architecture
Retrieval-Augmented Generation (RAG) with structured policy knowledge base

Retrieval corpus
4,280 policy wordings, 1,640 endorsements, 890 coverage guides, 340 fraud reference sheets (total: ∼7,150 documents)

Retrieval method
Hybrid: dense vector similarity (top-k=8) + BM25; cross-encoder re-ranking

System prompt
SymPrompt^{+} SP-CT-v2.3.1 (versioned in governance registry)

Output structure
Structured JSON: coverage_assessment (enum), pathway_recommendation (enum), fraud_indicators (list), confidence_score (float), reasoning_narrative (text)

Integration
REST API to Guidewire ClaimCenter; async; 95th percentile latency <3.2s

Deployment
Azure South Africa North; POPIA-compliant data residency

Fallback mode
AI unavailable → direct human handler queue with priority flag

G.2.2 Triage Workflow (4-Step Agentic Process)

The AI executes a four-step sequential workflow for each incoming claim. Each step's output becomes input to the next, creating the cascading failure model addressed by SPAR's multiplicative attrition calculation.

Table G.3: *ICT four-step agentic workflow*

Step	Action	Inputs	Outputs	Cascading Risk
1	Document intake	Raw claim form (PDF/digital submission)	Structured claim data object	Low: extraction errors propagate to all downstream steps
2	Coverage assessment	Structured claim data + retrieved policy documents	Coverage determination (covered/excluded/partial/unclear) with reasoning	High: incorrect coverage → wrong pathway and potential unfair outcome
3	Fraud screening	Structured claim data + coverage assessment + fraud reference sheets	Fraud risk score (low/medium/high) with flagged indicators	Medium: false positives create policyholder barriers; false negatives miss fraud
4	Pathway recommendation	All prior step outputs	Pathway (fast-track / standard / specialist / SIU referral) with confidence score	High: final output directly determines policyholder experience and handling timeline

G.2.3 Retrieval Corpus Maintenance Schedule

The retrieval corpus is the primary knowledge source for Steps 2–4. Corpus currency is governed by the maintenance schedule below; the Month 3 drift event (Section G.9) confirmed that policy endorsement currency is the most common drift source.

Table G.4: *ICT retrieval corpus maintenance*

Document Category	Volume	Update Frequency	Responsible	Governance Implication
Policy wording (standard)	2,840	Annually (product cycle)	Product development	Re-baselining triggered per MIDCOT protocol
Policy endorsements	1,640	Quarterly + ad hoc	Underwriting	Most common drift source; motor endorsements identified as Month 3 root cause
Coverage interpretation guides	890	Semi-annually	Claims technical team	Critical for multi-peril reasoning accuracy
Fraud indicator reference	340	Monthly (fraud intelligence)	Fraud investigation unit	Affects fraud screening precision CTG

G.3 Regulatory and Compliance Profile

The ICT scenario operates under South African financial conduct regulation with international AI governance standards as reference frameworks. This section maps the applicable regulatory instruments and provides the TCF outcome-to-CTG mapping that links behavioral assurance monitoring to regulatory compliance evidence.

G.3.1 Applicable Regulatory Framework

Eight regulatory and standards frameworks govern AI-assisted claims triage in this deployment context.

Table G.5: *ICT regulatory framework alignment*

Regulation	Jurisdiction	Relevance	dmaic-AI Alignment
FAIS Act (Act 37 of 2002)	South Africa	Governs advisory/intermediary services; FAIS Ombud adjudicates complaints; fitness and propriety	Define: VOG includes Ombud; Control: audit trail provides Ombud-grade evidence
TCF	FSCA	Six fairness outcomes; outcomes-based assessment; embedded fair treatment culture	Measure: SPAR quantifies conformance as TCF evidence; Control: continuous monitoring demonstrates embedding
PPRs	South Africa	Consistent, fair handling; no unreasonable barriers; disclosure requirements	Analyze: drift detection addresses consistency; Control: escalation satisfies intervention duty
Insurance Act (Act 18 of 2017)	South Africa	Prudential and conduct regulation; governance and risk management	Define: charter satisfies obligations; Control: MIDCOT feeds risk management
POPIA (Act 4 of 2013)	South Africa	Personal information processing; purpose limitation; consent	Define: input class registry respects data scope; System: POPIA-compliant residency
ISO/IEC 42001:2023	International	AI management system; risk assessment, monitoring, improvement	Full DMAIC cycle maps clause-by-clause
NIST AI RMF 1.0	International	GOVERN, MAP, MEASURE, MANAGE	DMAIC phases map to all four functions
EU AI Act	International (ref.)	High-risk for insurance AI; post-market monitoring	Control: MIDCOT provides conformity evidence

G.3.2 TCF Outcome Mapping

Each TCF outcome is mapped to specific CTG characteristics and DMAIC-AI governance artifacts, providing the evidentiary link between behavioral assurance monitoring and regulatory compliance.

Table G.6: *TCF outcome-to-CTG mapping*

TCF	Outcome Description	CTG Mapping	Monitoring Mechanism
1	Fair treatment central to culture	All five CTG (aggregate SPAR)	MIDCOT continuous monitoring; quarterly review
2	Products meet customer needs	Coverage accuracy; pathway consistency	Input-class-stratified SPAR; segment drift
3	Clear information provided	Response completeness	Automated completeness scoring; panel review
4	Suitable advice for circumstances	Coverage accuracy (per individual terms)	Per-policy-type accuracy monitoring
5	Products perform as expected	Pathway consistency (temporal)	CUSUM temporal consistency charts
6	No unreasonable post-sale barriers	Demographic fairness ($\Delta \leq 0.030$); pathway appropriateness	Segment-stratified SPAR; barrier indicators

G.4 Behavioral Specification Profile

The behavioral specification profile defines the measurable governance requirements produced by the Define phase. It comprises the CTG characteristic definitions with operational measurement protocols, the input class registry for stratified monitoring, and the specification envelope defining the multidimensional conformance region.

G.4.1 CTG Characteristic Definitions

Each Critical-to-Governance characteristic is formally defined with an operational definition, measurement method, target value, and specification limits. The system is conformant only if within specification on all five dimensions simultaneously.

Table G.7: *ICT CTG specification*

CTG	Definition	Measurement	Target	LSL	USL
Coverage assessment accuracy	Proportion of claims where AI coverage determination agrees with expert panel consensus	Automated: policy DB cross-reference + expert panel validation (quarterly calibration)	0.960	0.920	—
Pathway rec. consistency	Proportion of claim-pairs with similar profiles receiving the same pathway recommendation	Matched-pair analysis: claims matched on peril type, value band, policy terms; pathway agreement rate	0.940	0.900	—
Fraud flag precision	Proportion of AI fraud flags confirmed as warranting investigation upon human review	SIU feedback loop: flags reviewed by fraud team within 5 business days; precision calculated monthly	0.750	0.600	—
Demographic fairness	Maximum SPAR differential between any two policyholder demographic segments within an input class	Segment-stratified SPAR: segments defined by geographic region, policy channel, and claim history	$\leq$0.030	—	0.050
Response completeness	Proportion of triage outputs including all required structured fields with non-empty, substantive content	Automated schema validation + narrative quality scoring (minimum reasoning length, exclusion citation)	0.980	0.950	—

SPAR target: ≥ 0.970 (joint conformance across all five dimensions).

G.4.2 Input Class Registry

The input class registry defines the stratification structure for all baseline measurement, monitoring, and analysis. Each combination of dimensions constitutes a distinct governance stratum requiring separate BAR thresholds.

Table G.8: *ICT input class registry*

Dimension	Classes	Rationale
Line of business	Motor (52%), Property (31%), Liability (17%)	Different policy structures, coverage complexity, and peril types require separate behavioral baselines
Claim value band	Low (<R50K, 64%), Medium (R50K–R250K, 28%), High (>R250K, 8%)	Value band affects pathway distribution and fraud screening sensitivity; high-value claims have tighter specifications
Policy channel	Broker-originated (68%), Direct digital (24%), Direct call center (8%)	Channel affects submission quality and documentation completeness, influencing AI intake accuracy
Claim complexity	Simple (single-peril, 71%), Complex (multi-peril/disputed, 29%)	Complex claims identified as primary SPAR gap source in Analyze phase; separate monitoring essential

G.4.3 Behavioral Specification Envelope Summary

The specification envelope defines the multidimensional region of acceptable output. A triage output is conformant if and only if it falls within specification limits on all five CTG dimensions simultaneously. Class-specific limits reflect the differentiated risk profiles identified in the input class registry.

Table G.9: *ICT specification envelope by input class*

Input Class	Cov. Acc. LSL	Path. Con. LSL	Fraud Pr. LSL	Fair. USL	Compl. spar LSL	Tgt
Motor – Low – Simple	0.930	0.910	0.620	0.045	0.960	0.975
Motor – High – Complex	0.940	0.920	0.650	0.040	0.960	0.960
Property – Low – Simple	0.920	0.900	0.600	0.050	0.950	0.970
Property – High – Complex	0.935	0.915	0.630	0.040	0.955	0.955
Liability (all)	0.940	0.920	0.640	0.040	0.960	0.960

G.5 Baseline Measurement Profile

Baseline data was collected over a 30-day stability window under frozen system configuration. This section documents the collection protocol, GRR-validated measurement system, and the resulting BME metric values at both global and input-class-stratified levels.

G.5.1 Data Collection Summary

Collection period
1 February 2026 to 2 March 2026 (30 calendar days)

Total claims evaluated
72,480

Motor claims
37,690 (52.0%)

Property claims
22,469 (31.0%)

Liability claims
12,321 (17.0%)

Complex claims (multi-peril/disputed)
21,019 (29.0%)

High-value claims (>R250K)
5,798 (8.0%)

Evaluation method
Automated scoring (coverage accuracy, completeness) + expert panel (pathway consistency, fairness, fraud precision)

Expert panel composition
3 senior claims handlers (>10 years experience), quarterly inter-rater calibration

GRR inter-rater agreement
0.87 (exceeds 0.80 governance-grade threshold)

Baseline stability verification
No model updates, retrieval corpus changes, or configuration modifications during collection period

G.5.2 Baseline BME Metrics (Global)

Table G.10: *ICT baseline BME metrics (global)*

Metric	Value	Target	Status	Interpretation
SPAR	0.921	0.970	Below	7.9% nonconformant on ≥1 CTG dimension
ECPI	1.18	≥1.33	Below	Insufficient for governance-grade operation
ECI	2.4	≥3.0	Below	Below Three Sigma governance standard
BAR (cov. acc.)	0.948/0.992	—	Established	Empirical quantile-based control limits
BAR (path. cons.)	0.912/0.978	—	Established	Wider range: higher variation in pathway decisions

G.5.3 Baseline BME Metrics by Input Class

Table G.11: *ICT baseline BME metrics by input class*

Input Class	spar	Cov. Acc.	Path. Cons.	Fraud Pr.	Fair. Δ
Motor – Simple	0.952	0.968	0.941	0.762	0.018
Motor – Complex	0.884	0.921	0.903	0.731	0.024
Property – Simple	0.944	0.959	0.932	0.748	0.021
Property – Complex	0.878	0.912	0.894	0.714	0.031
Liability (all)	0.908	0.938	0.916	0.722	0.027

Complex claims across all lines show SPAR 0.878–0.884, confirming the Analyze phase finding that complex multi-peril claims are the primary governance gap source.

G.6 AI-FMEA Register

The AI-adapted Failure Mode and Effects Analysis identifies 14 failure modes across all CTG dimensions. The top 8 by Risk Priority Number are presented below with full governance-adapted ratings. The variation decomposition partitions the SPAR gap by contribution source.

Table G.12: *ICT AI-FMEA register (top 8 by RPN)*

#	Failure Mode	CTG	Cat.	S	O	D	RPN
1	Coverage misinterpretation on multi-peril claims: overlapping clauses parsed incorrectly	Cov. acc.	Data+Config	9	4	7	252
2	Pathway inconsistency: similar mid-value property claims routed differently based on narrative phrasing	Path. cons.	Model+Config	8	5	5	200
3	Outdated retrieval: superseded policy endorsements returned	Cov. acc.	Data	8	3	5	120
4	Fraud false positive on repeat claimants: flagged on frequency alone	Fraud prec.	Config	6	4	5	120
5	Retrieval degradation: stale interpretation guides used	Cov. acc.	Data	7	3	5	105
6	Confidence miscalibration: high confidence on ambiguous liability claims	Cov. acc.	Model	8	3	4	96
7	Incomplete reasoning narrative: exclusion clauses cited but not explained	Resp. compl.	Config	5	4	4	80
8	Segment bias: marginally higher rejection rate for rural regions	Dem. fair.	Model+Data	9	2	4	72

G.6.1 Variation Decomposition

The SPAR gap of 0.049 (0.970 target minus 0.921 baseline) is decomposed by contribution source.

Table G.13: *ICT variation decomposition*

Source	Contribution	% of Gap	Improve Priority
Coverage assessment failures (complex claims)	0.027	55%	Primary: retrieval corpus + SymPrompt^{+}
Pathway recommendation inconsistency	0.015	30%	Secondary: pathway decision rules
Fraud flag imprecision	0.004	8%	Deferred: acceptable within current fraud precision LSL
Measurement variation (evaluator)	0.003	7%	Not addressable via process improvement; MSA-validated

G.7 Intervention and Improvement Profile

Two interventions target the top two FMEA entries, which account for 85% of the SPAR gap. Both satisfy the least-invasive-first principle: retrieval corpus update (data-category) and SymPrompt^{+} modulation (configuration-category) are applied before any model-category intervention. This section documents the intervention design, DOE specification, and before/after validation results.

G.7.1 Intervention Design

Table G.14: *ICT intervention design*

Intervention	Root Cause	Mechanism	AAI	Reversibility
Retrieval corpus update	Outdated templates (FMEA #1, #3, #5)	340 documents refreshed; multi-peril clauses tagged with exclusion metadata; superseded templates removed	N/A (data)	Full: previous version archived
SymPrompt^{+} coverage verification	Insufficient multi-peril reasoning (FMEA #1, #2)	Coverage verification step injected: system checks each clause against exclusion registry and states applicability; 3 modulation levels for DOE	Innovate	Full: rollback to SP-CT-v2.3.0

G.7.2 DOE Design and Results

A fractional factorial design tests the two interventions across six experimental cells with complex claims deliberately oversampled from the primary gap source.

Table G.15: *ICT DOE specification*

Parameter	Value
Design type	Fractional factorial (3 × 2)
Factor A: SymPrompt^{+}	Level 1: no verification (baseline); Level 2: brief coverage check; Level 3: detailed exclusion verification with citation
Factor B: Retrieval corpus	Pre-update (original); Post-update (refreshed with exclusion metadata)
Experimental cells	6
Sample per cell	300 complex claims (deliberately oversampled from primary gap source)
Total evaluated	1,800 claims
Evaluation protocol	Full evaluator panel (3 senior handlers); blinded to experimental condition
Optimal combination	Level 3 SymPrompt^{+} (detailed verification) + post-update corpus

G.7.3 Before/After Validation

Post-intervention validation confirms statistically significant improvement on all target dimensions with no non-target degradation.

Table G.16: *ICT before/after validation*

Metric	Before	After	Change	Significance
SPAR (global)	0.921	0.974	+0.053	$p < 0.001$
ECPI	1.18	1.58	+0.40	$p < 0.001$
ECI	2.4	3.2	+0.8	Exceeds Three Sigma
Coverage acc. (complex)	0.917	0.961	+0.044	$p < 0.001$
Pathway consistency	0.914	0.956	+0.042	$p < 0.001$
Fraud precision	0.738	0.741	+0.003	Not significant
Demographic fairness Δ	0.024	0.022	−0.002	Not significant
Response completeness	0.971	0.978	+0.007	Marginal (SymPrompt^{+} side-effect)

No non-target dimension shows statistically significant degradation. Intervention approved for deployment.

G.8 Control and Monitoring Profile

The Control phase establishes ongoing statistical monitoring through MIDCOT's layered SPC architecture with tiered escalation and governance-gated re-baselining. This section documents the MIDCOT configuration, the four-tier escalation protocol, and the re-baselining trigger conditions.

G.8.1 MIDCOT Configuration

Table G.17: *ICT MIDCOT configuration*

Parameter	Configuration
Monitoring layer	Ingests BME data via daily batch aggregation; real-time batch during first 30 days post-intervention and during re-baselining windows
Shewhart charts	I-MR for each CTG characteristic per input class; BAR thresholds from post-improvement empirical quantiles
CUSUM charts	Layered for coverage accuracy and pathway consistency (highest drift susceptibility); adaptive reference values
EWMA charts	Session-level monitoring for sequential behavioral autocorrelation
Multivariate chart	Hotelling T^2 for joint behavioral vector across all five CTG dimensions
Run rules	Western Electric rules applied to all Shewhart charts (adopted without modification per AAI taxonomy)
Data retention	All metric data, control chart signals, investigation records, and intervention logs retained for 7 years (regulatory minimum: 5 years per PPR record-keeping requirements)

G.8.2 Tiered Escalation Protocol

Four escalation tiers with graduated restrictions. The Emergency tier is triggered by systematic unfair outcomes or regulatory-reportable incidents.

Table G.18: *ICT escalation protocol*

Tier	Trigger	Response	Timeline	Authority
Advisory	Run rule violation; metric within 15% of BAR; CUSUM below warning	Investigation initiated; enhanced monitoring; system continues	48 hours	Op. Gov. Team
Warning	Single BAR violation; SPAR <0.960 for any class	AI autonomy restricted; human review mandated; committee notified	4 hours	Op. Gov. + Claims Ops
Critical	2+ BAR violations; SPAR <0.940; ECI <2.0	AI suspended for affected lines; committee convened; full Analyze-Improve cycle	1 hour	Gov. Committee
Emergency	Systematic denial in protected segment; regulatory-reportable incident	Full suspension; emergency session; regulatory notification; complete DMAIC cycle	<15 min	Committee + CRO

G.8.3 Re-Baselining Protocol

Five categories of authorized changes trigger re-baselining with differentiated windows and approval requirements.

Table G.19: *ICT re-baselining triggers*

Trigger Event	Window	Approval	bar Update
Model version update (API)	21 days	Governance Committee	Full recalculation from new distribution
Retrieval corpus update	14 days	Op. Gov. (routine) / Committee (major)	Recalculation for affected input classes
SymPrompt$^+$ modulation change	14 days	Governance Committee	Full recalculation; before/after required
Specification revision (CTG)	21 days	Committee + Board Risk	New targets/limits; full Define-through-Control
Regulatory change (PPR, FSCA)	Per directive	Committee + Compliance	Assess spec adequacy; revise if required

G.9 Drift Event Case Record: MIDCOT-2026-0073

This section documents the Month 3 drift event as a complete governance case record, demonstrating the DMAIC-AI control loop in operational practice. The event was detected by CUSUM before any individual Shewhart point violated BAR thresholds, confirming the value of the layered monitoring architecture.

G.9.1 Event Timeline

Table G.20: *ICT drift event timeline*

Day	Event	Signal / Action
Day 73	MIDCOT CUSUM signals sustained negative drift on motor coverage accuracy	CUSUM exceeds Advisory; Shewhart within BAR
Day 74	Root cause: motor endorsements updated by underwriting 16 days prior but not loaded into retrieval corpus	Motor SPAR declined 0.974 → 0.958
Day 75	Targeted Improve: corpus updated with current motor endorsements; SymPrompt^{+} unchanged	200 motor claims validated; coverage accuracy 0.964
Days 76–89	Re-baselining: 14-day post-update data collection under enhanced monitoring	CUSUM resolves Day 80; SPAR stabilizes at 0.976
Day 90	Audit trail closed	Final: SPAR 0.976; ECPI 1.61; ECI 3.3

G.9.2 Governance Audit Trail Record

The complete audit trail record provides Ombud-grade evidence for regulatory review and FSCA compliance demonstration.

Table G.21: *ICT governance audit trail*

Element	Content
Incident ID	MIDCOT-2026-0073
Detection method	CUSUM chart, motor coverage accuracy, daily aggregation
Signal type	Sustained negative drift (gradual)
Escalation tier reached	Advisory (Warning not breached)
Root cause (confirmed)	Retrieval corpus data currency failure: 23 motor policy endorsements updated by underwriting but not propagated to AI retrieval index
Detection lag	16 days from endorsement change to detection
Resolution time	17 days (3 investigation/intervention + 14 re-baselining)
Intervention	Retrieval corpus update (motor endorsements); no SymPrompt^{+} change required
Validation	200 motor claims post-intervention; coverage accuracy 0.964 (above 0.940 LSL)
Re-baselining	14-day window; new BAR thresholds approved by Operational Governance Team
SPAR impact	Declined 0.974 → 0.958 (motor); restored to 0.976
Policyholder impact	No individual claim materially harmed; drift-period claims flagged for retrospective review
Regulatory reportable	No (Advisory tier; PPR threshold not breached; no individual unfair outcome)
Preventive action	Automated corpus update trigger linked to underwriting endorsement release; >14-day lag reduced to <48 hours

G.9.3 Corrective and Preventive Actions

The CAPA register documents both immediate corrective actions and systemic preventive measures to eliminate recurrence.

Table G.22: *ICT CAPA register*

Type	Description	Owner	Status	Date
Corrective	Motor policy endorsements loaded into retrieval corpus	Claims IT	Complete	Day 75
Corrective	Retrospective review of motor claims processed during 16-day drift period (est. 1,280 claims)	Senior Claims Mgr	In progress	Day 105 (tgt)
Preventive	Automated integration between underwriting endorsement release system and AI retrieval corpus update pipeline	Claims IT + UW IT	Approved	Day 120 (tgt)
Preventive	Retrieval corpus freshness monitoring added to MIDCOT: alert when any document category exceeds scheduled update date by >48 hours	Op. Gov. Team	In progress	Day 100 (tgt)

G.10 Standards Alignment Summary

This section maps the use-case governance implementation to the primary standards frameworks, providing clause-level and function-level traceability.

Table G.23: *ICT standards alignment summary*

Standard	Requirement	Implementation
ISO 42001 Cl. 4–5	Context, leadership	Governance charter; VOG; 3-tier authority
ISO 42001 Cl. 6.1	Risk assessment	AI-FMEA; prioritized cause register
ISO 42001 Cl. 8.1–8.2	Operational planning; AI risk assessment	CTG tree; specification envelope; input class registry
ISO 42001 Cl. 9.1	Monitoring	BME Metric Suite; MIDCOT SPC; MSA validation
ISO 42001 Cl. 9.2–9.3	Internal audit; management review	Control plan review (quarterly); audit trail; MIDCOT governance reporting
ISO 42001 Cl. 10	Continual improvement	DMAIC cyclical integration; SymPrompt^{+} interventions; DOE optimization
NIST AI RMF: GOVERN	Governance structures, policies, accountability	Governance charter; RACI; tiered escalation authority
NIST AI RMF: MAP	AI context, capabilities, limitations, impact	Input class registry; CTG tree; behavioral specification envelope
NIST AI RMF: MEASURE	Quantitative and qualitative risk assessment	BME metrics (ECPI, BAR, ECI, SPAR); MSA/GRR validation
NIST AI RMF: MANAGE	Risk prioritization, response, monitoring	AI-FMEA risk prioritization; MIDCOT monitoring; tiered escalation
EU AI Act Art. 9	Risk management system	Full DMAIC cycle; AI-FMEA; control plan
EU AI Act Art. 72	Post-market monitoring	MIDCOT continuous SPC; re-baselining protocol; audit trail
FAIS Act	Fitness/propriety; Ombud	Audit trail provides Ombud-grade evidence; evaluator calibration
TCF Outcomes 1–6	Fair treatment culture	CTG-to-TCF mapping (Table G.6); SPAR as quantitative TCF evidence
PPRs	Consistent handling; records	Pathway consistency CTG; 7-year data retention

Appendix H

Emergency Department Triage Agent: Use-Case Data Profile

This appendix provides the complete use-case data profile for the Emergency Department Triage (EDT) scenario that threads through Chapters 3–7 as a parallel worked example. All values are canonical and internally consistent. The scenario exercises the cascading failure model, Severity-10 FMEA calibration, and per-step SPAR decomposition. This document parallels the insurance claims triage data profile (Appendix G). The profile is structured to mirror the DMAIC lifecycle: organizational context and system architecture (Sections H.1–H.2), regulatory alignment and behavioral specification (Sections H.3–H.4), baseline measurement (Section H.5), failure analysis (Section H.6), intervention and improvement (Section H.7), control and monitoring (Section H.8), drift event documentation (Section H.9), and standards alignment (Section H.10).

H.1 Organizational Profile

Metro Regional Medical Center is a Level II Trauma Center operating a high-volume emergency department. This section profiles the facility, its ED operations, and the three-tier AI governance hierarchy established by the Define phase governance charter.

Facility
Metro Regional Medical Center

Type
Level II Trauma Center, 420-bed community teaching hospital

ED annual volume
94,000 visits/year (∼258/day, peak 340/day)

Acuity distribution
ESI-1: 2%, ESI-2: 18%, ESI-3: 42%, ESI-4: 28%, ESI-5: 10%

ED staffing
38 physicians, 112 nurses, 24 PAs/NPs, 18 techs

Annual admissions
∼31,000 (33% admission rate)

Door-to-provider
28 minutes (target: <20 min ESI-2, <10 min ESI-1)

LWBS rate
4.2% (national benchmark <2%)

EHR
Epic Hyperspace with integrated clinical decision support

Governance structure
3-tier: Executive Oversight (CMO chair), Clinical AI Committee, Operational Monitoring Team

H.1.1 Governance Hierarchy

The AI governance structure operates within the hospital's existing quality and patient safety framework, with three tiers of escalating authority.

Table H.1: *EDT governance hierarchy*

Tier	Composition	Authority
Executive Oversight	CMO (chair), CNO, CIO, VP Quality, General Counsel	Suspend AI triage system. Authorize specification changes. Regulatory notification
Clinical AI Committee	ED Medical Director, ED Nurse Manager, CMIO, Clinical Informaticist, Patient Safety Officer, Bioethicist	Approve CTG specifications. Review FMEA. Authorize Improve-phase interventions. Quarterly governance reviews
Operational Monitoring	2 ED attending physicians, 2 charge nurses, Clinical AI analyst, Quality data analyst	Daily monitoring. Escalation initiation. Investigation execution. Re-baselining recommendations

H.2 AI System Technical Profile

TriageAssist ED v3.1 is a RAG-augmented multi-step agentic system executing a four-step sequential pipeline for each patient encounter. This section documents the system identity, the agentic workflow with cascading failure dynamics, and the integration architecture.

H.2.1 System Identity

System name
TriageAssist ED v3.1

Architecture
RAG-augmented multi-step agentic workflow

Foundation model
LLM (8B parameter, fine-tuned on 2.4M de-identified ED encounters)

Retrieval corpus
4,280 clinical documents: ED protocols, clinical practice guidelines, drug references, institutional policies, sepsis/stroke/STEMI pathways

Prompt architecture
SymPrompt^{+} SP-ED-v3.1.2 (four-module sequential pipeline)

Integration
Epic Hyperspace via FHIR R4 API; real-time vitals feed from Philips IntelliVue

Human oversight
AI recommendation displayed to triage nurse; nurse confirms or overrides before assignment

EU AI Act classification
High-risk (Annex III, Section 5(c): emergency healthcare patient triage systems)

H.2.2 Agentic Workflow: Four-Step Sequential Pipeline

Each patient encounter triggers a four-step sequential workflow. Each step's output becomes input to the next, creating the cascading failure dynamics described in Chapter 5.

Table H.2: *EDT four-step agentic pipeline*

Step	Module	Function	Output
1	Intake Parser	Extracts structured clinical data from free-text triage notes, EHR history pull, and real-time vitals stream. Identifies chief complaint, symptom onset, relevant history, current medications, allergies	Structured patient clinical profile (JSON) with confidence scores per extracted field
2	Acuity Classifier	Assigns ESI level (1–5) based on structured profile + retrieved clinical protocols. Applies sepsis, stroke, STEMI screening criteria. Calculates predicted resource utilization	ESI level, screening alerts (sepsis/stroke/STEMI flags), resource prediction, confidence score
3	Resource Allocator	Recommends treatment area assignment and initial resource set based on acuity + predicted needs	Area assignment, resource bundle recommendation, estimated wait time, priority ranking
4	Disposition Advisor	Generates preliminary disposition recommendation with reasoning chain and evidence citations from retrieval corpus	Disposition recommendation, reasoning narrative, cited protocols, uncertainty flag if confidence <0.80

Cascading failure model: a Step 1 extraction error (e.g., missing symptom onset time) propagates to Step 2 (incorrect acuity scoring), Step 3 (wrong area assignment), and Step 4 (inappropriate disposition). Per-step SPAR of 0.99 across four steps yields workflow-level SPAR of $\sim$0.96.

H.3 Regulatory and Standards Profile

The EDT scenario operates under US clinical regulatory requirements with international AI governance standards as reference frameworks. EMTALA creates an absolute governance floor that is non-negotiable and not subject to governance committee override.

Table H.3: *EDT regulatory framework alignment*

Framework	Requirement	dmaic-AI Response
EMTALA	Medical screening examination for all ED patients regardless of ability to pay. Appropriate transfer protocols	AI triage cannot refuse or deprioritize patients. Acuity scoring must be clinically appropriate. Demographic fairness CTG ensures non-discriminatory triage
CMS CoP	Conditions of Participation. Quality assessment and performance improvement programs	DMAIC-AI provides the QAPI structure. SPC monitoring satisfies ongoing performance improvement. Audit trails for CMS surveys
Joint Commission	National Patient Safety Goals. Leadership standards for patient safety culture	FMEA maps to NPSG proactive risk assessment. BAR escalation satisfies leadership oversight. Control-phase monitoring is continuous PI
State medical practice acts	AI cannot independently practice medicine. Physician/nurse must authorize clinical decisions	Human-in-the-loop at every consequential decision point. Override concordance CTG monitors alignment
EU AI Act (Art. 6, Annex III)	High-risk classification. Conformity assessment. Post-market monitoring. Human oversight	DMAIC-AI provides conformity evidence. MIDCOT satisfies post-market monitoring. Escalation protocol satisfies human oversight
ISO/IEC 42001:2023	AI management system. Risk assessment. Performance monitoring. Continual improvement	Clause-level alignment through full DMAIC cycle
NIST AI RMF 1.0	GOVERN, MAP, MEASURE, MANAGE functions	DMAIC phases map to all four functions
HIPAA	Protected health information safeguards. Minimum necessary standard. Audit controls	PHI handling governed by data minimization in retrieval. All AI interactions logged. De-identification for governance analytics

Critical regulatory note: EMTALA creates an absolute floor: the AI cannot under any circumstances deprioritize a patient based on non-clinical factors. Any governance failure that produces discriminatory acuity scoring is simultaneously a clinical safety event AND a federal regulatory violation. This dual consequence drives the Severity 10 ratings in the FMEA.

H.4 Behavioral Specification

The behavioral specification profile defines the measurable governance requirements produced by the Define phase. It comprises the VOG analysis identifying stakeholder governance needs, six CTG characteristics with operational definitions, the input class registry for stratified monitoring, and per-step specification requirements unique to the agentic pipeline architecture.

H.4.1 Voice of Governance (VOG) Analysis

Five stakeholder groups contribute governance requirements, translated into measurable CTG characteristics through the VOG weighting model.

Table H.4: *EDT Voice of Governance analysis*

Stakeholder	Governance Need	CTG Translation
Patients	Timely, accurate triage. No discrimination. Appropriate acuity assignment	Acuity accuracy, demographic fairness, time-to-triage
ED clinicians	Reliable recommendations. Clear reasoning. Appropriate escalation. Low false alarm rate	Acuity accuracy, reasoning transparency, screening sensitivity/specificity
Hospital administration	Regulatory compliance. Reduced LWBS rate. Efficient resource utilization. Liability protection	EMTALA compliance, throughput efficiency, audit completeness
Regulators (CMS, TJC, state)	Patient safety. EMTALA compliance. QI documentation. Adverse event reporting	Under-triage rate, safety event detection, audit trail availability
Accrediting bodies	Evidence of proactive risk management. FMEA documentation. Performance improvement trends	FMEA completion, SPC trend documentation, governance cycle records

H.4.2 Critical-to-Governance (CTG) Characteristics

Six measurable behavioral dimensions with specification limits. The system is conformant only if within specification on all six dimensions simultaneously.

Table H.5: *EDT CTG specification*

CTG	Target	LSL/USL	Measurement	Description
Acuity accuracy	0.940	0.900	Expert consensus	Proportion of ESI assignments matching consensus of 3 board-certified emergency physicians
Under-triage rate	<0.020	<0.035	Retrospective review	Proportion of patients assigned lower acuity than warranted. Most dangerous failure mode: under-triage of ESI-1/2 to ESI-3+
Screening sensitivity	0.970	0.950	Protocol match	Proportion of true sepsis/stroke/STEMI cases correctly flagged by Step 2 screening. False negatives are potentially fatal
Reasoning transparency	0.950	0.900	Clinician panel	Proportion of disposition recommendations with clinically valid reasoning chains and accurate protocol citations
Demographic fairness	<0.025	<0.040	Segment analysis	Maximum difference in acuity accuracy between any two demographic segments (age, race, sex, language, insurance status)
Override concordance	<0.080	<0.120	Override log	Proportion of AI recommendations overridden by triage nurses. Low override = high trust alignment. High override = AI misaligned with clinical judgment

SPAR target: ≥ 0.960 (workflow-level). Per-step SPAR must exceed 0.990 for each of the four pipeline modules to maintain workflow-level SPAR above the 0.960 governance floor.

H.4.3 Input Class Registry

Patient encounters are stratified across four dimensions for separate monitoring. Each combination constitutes a distinct governance stratum.

Table H.6: *EDT input class registry*

Stratification	Classes	Rationale
Acuity level	ESI-1 (resuscitation, 2%), ESI-2 (emergent, 18%), ESI-3 (urgent, 42%), ESI-4/5 (less/non-urgent, 38%)	Higher acuity = tighter specifications. ESI-1/2 patients have time-critical conditions where under-triage is potentially fatal
Chief complaint	Chest pain (14%), Abdominal pain (12%), Injury/trauma (18%), Respiratory (11%), Neurological (8%), Psychiatric (6%), Pediatric (9%), Other (22%)	Complaint category determines screening protocol applicability and expected resource profile
Complexity	Standard (68%): single chief complaint, clear history. Complex (32%): multiple complaints, altered mental status, language barrier, incomplete history	Complex presentations are the primary under-triage risk
Time of presentation	Day shift (7a–3p, 38%), Evening (3p–11p, 40%), Night (11p–7a, 22%)	Night shift has thinner staffing and higher proportion of altered/intoxicated patients; separate monitoring detects time-dependent drift

Per-step specification: Unlike the insurance claims scenario, this scenario requires per-step AND workflow-level CTG specifications because the four-step pipeline creates cascading failure dynamics. Step-level SPAR must exceed 0.990 for each module.

H.5 Baseline Measurement

Baseline data was collected over a 30-day stability window under frozen system configuration with 100% automated scoring supplemented by 15% expert panel review. This section documents the MSA validation, collection parameters, and resulting BME metric values at both workflow and input-class-stratified levels.

H.5.1 Measurement System Analysis

The evaluator panel comprises five clinicians with extensive ED experience. All six CTG dimensions exceed the 0.75 governance threshold, with reasoning transparency borderline at 0.78.

Table H.7: *EDT MSA/GRR validation*

MSA Component	Result
Evaluator panel	3 board-certified emergency physicians + 2 senior charge nurses. Each evaluator: >10 years ED experience, >50,000 patient encounters
Evaluation protocol	200 AI triage outputs across all input classes, evaluated on all 6 CTG dimensions. Evaluated twice with 14-day separation
Intra-rater agreement	Acuity accuracy: 0.94. Under-triage: 0.91. Screening sensitivity: 0.93. Reasoning transparency: 0.86. Demographic fairness: 0.89. Override concordance: 0.92
Inter-rater agreement	Krippendorff's α: Acuity 0.88, Under-triage 0.84, Screening 0.87, Reasoning 0.78, Fairness 0.82, Override 0.85. All above 0.75 governance threshold except Reasoning at 0.78 (borderline)
GRR result	Overall GRR: 0.86. Governance-grade. Reasoning transparency flagged for evaluator calibration improvement

H.5.2 Baseline Collection

Baseline data was collected under frozen system configuration to ensure distributional stability.

Table H.8: *EDT baseline collection parameters*

Parameter	Value
Baseline period	30 days (Days 1–30)
Total encounters evaluated	7,740 (258/day × 30 days)
Evaluation method	100% automated scoring on 4 dimensions + 15% expert panel review on all 6 dimensions
System configuration	TriageAssist ED v3.1 with SymPrompt^{+} SP-ED-v3.1.2. Retrieval corpus v2024.11. Frozen during baseline

H.5.3 Baseline Metrics

Table H.9: *EDT baseline BME metrics*

Metric	Baseline	Target	Gap	Interpretation
SPAR (workflow)	0.932	0.960	−0.028	6.8% nonconformant. ∼18 patients/day
SPAR (Step 1: Intake)	0.991	0.995	−0.004	Errors concentrated in complex/multilingual presentations
SPAR (Step 2: Acuity)	0.974	0.990	−0.016	**Primary gap contributor.** Under-triage on ESI-2/3 boundary for atypical presentations
SPAR (Step 3: Resource)	0.986	0.990	−0.004	Resource allocation mostly correct but inherits Step 2 errors
SPAR (Step 4: Disposition)	0.981	0.990	−0.009	Reasoning transparency gaps. Protocol citations occasionally outdated
ECPI	1.12	≥1.33	−0.21	Process capability below governance-grade
ECI	2.6	≥3.0	−0.4	Below Three Sigma
Under-triage rate	0.031	<0.020	+0.011	3.1% under-triage. ∼8 patients/day assigned lower acuity than warranted
Screening sensitivity	0.958	0.970	−0.012	Sepsis screening missed 4.2% of true positives. Stroke/STEMI at 0.971 (acceptable)

H.5.4 Per-Input-Class Breakdown (Workflow SPAR)

Table H.10: *EDT baseline workflow SPAR by input class*

Input Class	Acuity	Under-tri	Screen	Reason	Fair	spar
ESI-1/2 Standard	0.951	0.024	0.968	0.961	0.018	0.952
ESI-1/2 Complex	0.928	0.043	0.944	0.938	0.032	0.914
ESI-3 Standard	0.946	0.028	0.962	0.955	0.020	0.944
ESI-3 Complex	0.921	0.039	0.948	0.931	0.034	0.908
ESI-4/5	0.962	0.015	0.974	0.968	0.016	0.966
Pediatric (all ESI)	0.934	0.036	0.951	0.942	0.038	0.918

Critical finding: Complex ESI-1/2 and pediatric presentations show the widest governance gaps. Under-triage rates of 0.043 and 0.036 respectively represent patients who should have been classified as emergent but were not. In an ED, under-triage of ESI-2 to ESI-3 can mean a 45-minute delay in evaluation for a patient having a STEMI.

Multiplicative attrition demonstrated: Per-step SPARs of $0.991 \times 0.974 \times 0.986 \times 0.981 = 0.932$ workflow SPAR. The workflow is only as strong as its weakest step. Step 2 (acuity classification) is the primary bottleneck.

H.6 AI-FMEA Risk Register

The AI-adapted FMEA identifies failure modes across all six CTG dimensions with governance-calibrated severity ratings. Two failure modes receive Severity 10 (maximum) because they can directly cause patient death. The variation decomposition partitions the SPAR gap by step-level contribution, and the root cause analysis identifies three confirmed causal targets for the Improve phase.

H.6.1 Variation Decomposition

Of the total workflow SPAR gap (0.932 vs. 0.960 target = 0.028 gap):

Table H.11: *EDT variation decomposition*

Source	Contribution	% of Gap	Cumulative
Step 2: Acuity misclassification (ESI-2/3 boundary)	0.016	57%	57%
Step 4: Reasoning/citation gaps	0.005	18%	75%
Step 2: Sepsis screening false negatives	0.003	11%	86%
Step 1: Complex intake extraction errors	0.002	7%	93%
Measurement variation (evaluator)	0.002	7%	100%

H.6.2 FMEA Risk Register (Top 8 by RPN)

Table H.12: *EDT AI-FMEA register (top 8 by RPN)*

Failure Mode	Step	CTG	S	O	D	RPN	Governance Consequence
Under-triage: ESI-2 scored as ESI-3 on atypical presentations	2	Under-tri	10	4	6	240	Delayed evaluation. Potential preventable morbidity/mortality
Sepsis screening false negative: missed qSOFA criteria	2	Screen	10	3	7	210	Delayed sepsis bundle. Mortality increases 7.6% per hour
Outdated protocol retrieval: superseded sepsis criteria	2,4	Screen, Reas.	9	3	6	162	AI applies old Surviving Sepsis Campaign criteria
Demographic acuity bias: pain assessment disparity	2	Fairness	8	4	5	160	Lower acuity for specific demographics. EMTALA risk
Pediatric weight-based dosing extraction error	1	Acuity	9	3	5	135	Wrong weight cascades to wrong resource/medication recs
Cascading: intake → acuity → area	1→3	Multiple	9	3	5	135	Step 1 error amplified through Steps 2 and 3
Reasoning citation: outdated CPG	4	Reason	7	4	4	112	Superseded guideline cited. Clinician trust erosion
Night-shift performance degradation	1–4	Multiple	7	3	5	105	Higher errors on altered/intoxicated patients at night

Severity 10 justification: Under-triage and missed sepsis screening receive Severity 10 because they can directly cause patient death. An ESI-2 patient triaged as ESI-3 may wait 45–90 minutes longer for physician evaluation. For time-sensitive conditions (STEMI, stroke, sepsis), that delay is the difference between recovery and permanent disability or death.

H.6.3 Root Cause Analysis: Causal Targets for Improve Phase

Three root causes are confirmed through controlled comparison of outputs generated with current vs. outdated corpus documents on matched input sets.

- *Primary target (57% of gap):* Step 2 acuity classification on ESI-2/3 boundary. Root cause: retrieval corpus contains atypical presentation guidelines from 2019; current guidelines (2024 update) not loaded. SymPrompt$^+$ lacks explicit instruction to weigh vital sign trends against chief complaint severity.
- *Secondary target (18% of gap):* Step 4 reasoning transparency. Root cause: 12% of protocol citations in disposition recommendations reference superseded guidelines. Retrieval corpus version control not synchronized with guideline publication cycles.
- *Tertiary target (11% of gap):* Step 2 sepsis screening sensitivity. Root cause: qSOFA screening criteria in retrieval corpus are from Sepsis-2 (2012); current Sepsis-3 criteria (2016 update with 2023 amendments) not fully integrated.

H.7 Intervention Profile

Three interventions target the top three FMEA entries, which account for 86% of the SPAR gap. All are configuration-category or data-category, preserving full reversibility. No fine-tuning is required. This section documents the intervention design, the DOE specification, and the before/after validation results with clinical significance.

H.7.1 Interventions

Table H.13: *EDT intervention design*

#	Intervention	Description	FMEA Target
I-1	Retrieval corpus update	840 clinical documents updated to current guideline versions. Sepsis-3 criteria with 2023 amendments integrated. Atypical presentation recognition guidelines (2024 ACS/ACEP updates) added. Version metadata tagged for currency validation	FM #1 (under-triage), FM #2 (sepsis screening), FM #3 (outdated protocols), FM #7 (citation currency)
I-2	SymPrompt$^+$ acuity verification module	New instruction module between Steps 1 and 2: "Before assigning ESI level, verify: (a) vital sign trends consistent with proposed acuity, (b) chief complaint evaluated against atypical criteria for age and sex, (c) sepsis/stroke/STEMI screening uses current (2023+) criteria." Three modulation levels for DOE	FM #1 (under-triage), FM #2 (sepsis screening), FM #4 (demographic bias)
I-3	Demographic fairness constraint	Pain assessment equalization: "Do not adjust acuity assessment based on patient demographics. Apply identical vital sign and symptom severity thresholds regardless of age, race, sex, language, or insurance status"	FM #4 (demographic bias)

H.7.2 Design of Experiments

A fractional factorial design tests the three interventions across 12 experimental cells with stratified sampling targeting the primary gap sources.

Table H.14: *EDT DOE specification*

DOE Component	Specification
Design	$3 \times 2 \times 2$ fractional factorial: 3 SymPrompt$^+$ modulation levels $\times$ 2 retrieval conditions (old vs. updated) $\times$ 2 fairness constraint conditions (off vs. on)
Sample	2,400 encounters (200 per cell $\times$ 12 cells). Stratified: 40% ESI-2/3, 25% complex, 15% pediatric, 20% other
Evaluation	Full 6-CTG evaluation by calibrated panel. Automated scoring + 25% expert review overlay
Optimal combination	Updated corpus + detailed verification module + fairness constraint ON
Validation	600 encounters across all input classes. No non-target degradation confirmed ($p < 0.001$)

H.7.3 Improvement Results

Post-intervention validation on 600 encounters confirms statistically significant improvement on all target dimensions with no non-target degradation.

Table H.15: *EDT before/after validation*

Metric	Before	After	Change	Status
SPAR (workflow)	0.932	0.968	+0.036	Exceeds 0.960 target
SPAR (Step 2)	0.974	0.992	+0.018	Primary gap closed
Under-triage rate	0.031	0.018	−0.013	Below 0.020 target
Screening sensitivity	0.958	0.978	+0.020	Exceeds 0.970 target
Demographic fairness	0.032	0.019	−0.013	Below 0.025 target
ECPI	1.12	1.48	+0.36	Approaching 1.33
ECI	2.6	3.3	+0.7	Above Three Sigma

Clinical significance: Under-triage rate reduced from 3.1% to 1.8%. In practical terms, approximately 3 fewer patients per day receive delayed emergency evaluation. For time-critical conditions, this improvement directly translates to reduced preventable harm.

H.8 Control Profile

The Control phase establishes real-time per-encounter monitoring through MIDCOT's layered SPC architecture with CUSUM tuned for clinical drift detection and a four-tier escalation protocol with EMTALA-specific Emergency provisions.

H.8.1 MIDCOT Configuration

The clinical context requires real-time per-encounter scoring with layered SPC architecture tuned for low false alarm rates and rapid drift detection.

Table H.16: *EDT MIDCOT configuration*

Parameter	Specification
Monitoring cadence	Real-time per-encounter scoring (Steps 1–4). Hourly batch aggregation. Daily SPC chart update. Weekly governance report
SPC charts	Shewhart I-MR for workflow SPAR (daily). CUSUM for under-triage rate and sepsis screening sensitivity. Multivariate Hotelling T^2 for joint 6-CTG vector
BAR thresholds	Upper: 0.988. Lower: 0.948. Calculated from post-improvement empirical quantiles (not parametric)
CUSUM parameters	Reference value $k = 0.5\sigma$, decision interval $h = 4\sigma$. Tuned for $\text{ARL}_0 = 370$ days (low false alarm), $\text{ARL}_1 = 14$ days (detect 1σ shift within 2 weeks)

H.8.2 Escalation Protocol

Four tiers with escalating restrictions. The Emergency tier is distinguished by its regulatory dimension: any EMTALA violation triggers full system halt with mandatory regulatory notification.

Table H.17: *EDT escalation protocol*

Tier	Trigger	Response	Timeline
Advisory	CUSUM signal or any metric within 15% of BAR. Under-triage >0.025 for 3 consecutive days	Enhanced monitoring (hourly review). Operational team investigation. Root cause initiated	48 hours to initial finding
Warning	BAR violation on any CTG. Under-triage >0.030. Workflow SPAR <0.955	AI recommendations flagged for mandatory double-review by senior clinician. Affected input class isolated	4 hours for senior review. Clinical AI Committee within 24 hours
Critical	2+ BAR violations. Under-triage >0.040. SPAR <0.940. Any missed sepsis/STEMI/stroke with adverse outcome	AI triage SUSPENDED for affected input classes. Manual triage reinstated. Full incident investigation	Suspension within 1 hour. Executive Oversight within 2 hours
Emergency	Patient safety event directly attributable to AI. Any EMTALA violation linked to AI triage	FULL SYSTEM HALT. All AI triage suspended. Regulatory notification (CMS, TJC, state). Sentinel event investigation	Halt within 15 min. CMO immediate. Regulatory within 24 hours

H.9 Drift Event Case Record: MIDCOT-2026-ED-0041

This section documents the post-deployment drift event as a complete governance case record, demonstrating the DMAIC-AI control loop in operational clinical practice. The event was detected by CUSUM before any individual Shewhart point violated BAR thresholds, confirming the value of the layered monitoring architecture for gradual clinical drift.

Table H.18: *EDT drift event timeline*

Timeline	Event
Days 1–30	Post-improvement stability confirmed. Workflow SPAR stable at 0.968. All CTG metrics within specification. BAR thresholds established
Day 45	American College of Emergency Physicians (ACEP) publishes updated sepsis screening criteria. New criteria emphasize lactate trajectory over static qSOFA score for patients >65
Days 45–62	Updated criteria NOT loaded into retrieval corpus. AI continues applying previous screening criteria. No immediate metric degradation because most sepsis presentations are still captured
Day 58	CUSUM for sepsis screening sensitivity begins accumulating. Not yet at Advisory threshold
Day 63	**ADVISORY triggered.** CUSUM crosses Advisory threshold. Sepsis screening sensitivity drifted from 0.978 to 0.961 over 18 days. Under-triage rate for patients >65 increased from 0.018 to 0.028. Investigation initiated
Day 64	Root cause confirmed: ACEP updated sepsis screening criteria published Day 45 not loaded into retrieval corpus. AI applies old criteria for elderly patients. Specifically: lactate trajectory criterion missing, causing 3–4 elderly sepsis patients per day to receive delayed screening
Day 65	Retrieval corpus updated with ACEP 2026 sepsis criteria. SymPrompt^{+} SP-ED-v3.1.2 unchanged (root cause is data, not instructions). 150 encounters validated: sepsis screening sensitivity restored to 0.976
Days 65–79	14-day re-baselining period. Enhanced monitoring. New BAR thresholds calculated from post-update distribution
Day 79	**RESOLVED.** Re-baselining complete. Workflow SPAR 0.971. Sepsis screening sensitivity 0.979. BAR thresholds updated. Audit trail closed. CAPA: automated guideline publication monitoring pipeline approved by Clinical AI Committee

Resolution: Detection to resolution: 16 days. Drift was detected before any sentinel event. Root cause (data currency) identified within 24 hours. Matches FMEA #3 (outdated protocol retrieval, RPN 162). CAPA prevents recurrence through automated guideline publication monitoring.

Clinical significance. During the 18-day drift window (Days 45–63), an estimated 54–72 elderly patients received delayed sepsis screening. Of these, approximately

8–12 had confirmed sepsis. Each hour of delay in sepsis bundle initiation increases mortality risk by 7.6%. The monitoring system's 18-day detection lag, while fast relative to manual audits, represents an area for improvement: the CAPA targets reducing detection time to <7 days through automated corpus-currency checks.

CAPA. Corrective: corpus updated (Day 65, complete). Preventive: automated guideline publication monitoring pipeline approved by Clinical AI Committee; target detection time reduced from 18 days to <7 days through automated corpus-currency checks linked to ACEP, ACS, and Surviving Sepsis Campaign publication feeds.

H.10 Standards Alignment Summary

This section maps the EDT governance implementation to primary standards frameworks. The clinical context requires alignment with both international AI governance standards and US clinical regulatory requirements.

H.10.1 ISO/IEC 42001:2023

Table H.19: *EDT alignment: ISO/IEC 42001:2023*

Clause	Requirement	dmaic-AI Evidence
4.1–4.4	Context, leadership, AIMS planning	Define phase: governance charter, VOG, CTG tree, 3-tier oversight hierarchy
6.1	Risk assessment	Analyze phase: AI-FMEA, variation decomposition, cascading failure model
8.2–8.4	AI risk assessment, treatment	Define + Analyze: FMEA severity calibrated to clinical impact. EMTALA risk integrated
9.1	Monitoring, measurement, analysis	Measure + Control: MSA validation, baseline, MIDCOT SPC monitoring
10.1–10.2	Continual improvement, corrective action	Improve + Control: DOE-validated interventions, drift detection, CAPA

H.10.2 NIST AI RMF 1.0

Table H.20: *EDT alignment: NIST AI RMF 1.0*

Function	Requirement	dmaic-AI Evidence
GOVERN	Policies, accountability, oversight	Define: 3-tier governance hierarchy, RACI, escalation protocol
MAP	Context, capabilities, risk identification	Define + Analyze: VOG, input class registry, FMEA, cascading failure model
MEASURE	Quantitative assessment, metrics	Measure + Control: BME metrics (SPAR, ECPI, ECI), MSA, MIDCOT SPC
MANAGE	Risk treatment, monitoring, improvement	Improve + Control: validated interventions, 4-tier escalation, CAPA, re-baselining

H.10.3 EU AI Act (High-Risk Requirements)

Table H.21: *EDT alignment: EU AI Act*

Article	Requirement	dmaic-AI Evidence
Art. 9	Risk management system	Full DMAIC cycle as risk management. FMEA, SPC, escalation
Art. 10	Data governance	Retrieval corpus version control, currency validation, HIPAA-compliant handling
Art. 13	Transparency	Step 4 reasoning transparency CTG. Protocol citation requirements
Art. 14	Human oversight	Nurse override at every decision point. Override concordance CTG
Art. 72	Post-market monitoring	MIDCOT continuous SPC monitoring. Automated drift detection

Bibliography

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, et al. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley, 2nd edition, 2005.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[6] Miles Brundage, Shahar Avin, Jasmine Wang, et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

[7] W. Edwards Deming. *Out of the Crisis*. MIT Press, 1986.

[8] W. Edwards Deming. *The New Economics for Industry, Government, Education*. MIT Press, 2nd edition, 1994.

[9] European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, 2024.

[10] Executive Office of the President. Executive order 13960: Promoting the use of trustworthy artificial intelligence in the federal government. *Federal Register*, 85(236):78939–78943, 2020.

[11] Luciano Floridi and Massimo Chiriatti. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[12] Yunfan Gao, Yun Xiong, Xinyu Gao, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024.

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[14] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, et al. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on AI and Security*, pages 79–90, 2023.

[15] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*, 2023.

[16] IEEE Standards Association. IEEE P2863: Recommended practice for organizational governance of artificial intelligence. Technical report, IEEE, 2020.

[17] International Organization for Standardization. *ISO 9001:2015: Quality Management Systems—Requirements*. ISO, 2015.

[18] International Organization for Standardization. *ISO/IEC 23053:2022: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*. ISO/IEC, 2022.

[19] International Organization for Standardization. *ISO/IEC 42001:2023: Information Technology—Artificial Intelligence—Management System*. ISO/IEC, 2023.

[20] Kaoru Ishikawa. *What Is Total Quality Control? The Japanese Way*. Prentice Hall, 1985.

[21] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[22] J. M. Juran and A. B. Godfrey. *Juran's Quality Handbook*. McGraw-Hill, 5th edition, 1999.

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[24] Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[25] Grégoire Mialon, Roberto Dessì, Maria Lomeli, et al. Augmented language models: A survey. *Transactions on Machine Learning Research*, 2023.

[26] Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. Model cards for model reporting. In *Proceedings of FAT**, pages 220–229, 2019.

[27] Brent Mittelstadt. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507, 2019.

[28] Douglas C. Montgomery. *Introduction to Statistical Quality Control*. Wiley, 8th edition, 2019.

[29] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical report, U.S. Department of Commerce, 2023.

[30] National Institute of Standards and Technology. NIST AI 600-1: Generative artificial intelligence profile. Technical report, U.S. Department of Commerce, 2024.

[31] Lloyd S. Nelson. The Shewhart control chart: Tests for special causes. *Journal of Quality Technology*, 16(4):237–239, 1984.

[32] OECD. OECD principles on AI. Technical report, OECD Publishing, 2019.

[33] Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

[34] Ethan Perez, Saffron Huang, Francis Song, et al. Red teaming language models with language models. In *Proceedings of EMNLP*, pages 3419–3448, 2022.

[35] Thomas Pyzdek and Paul A. Keller. *The Six Sigma Handbook*. McGraw-Hill Education, 4th edition, 2014.

[36] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, et al. Closing the AI accountability gap. *Proceedings of FAT**, pages 33–44, 2020.

[37] Dale Rutherford. *Governing BME Amplification in Large Language Models: A Standards-Aligned Framework for LLM Data Lifecycle Risk Detection and Ethical AI Governance*. PhD thesis, University of Arkansas at Little Rock, 2026.

[38] Thomas P. Ryan. *Statistical Methods for Quality Improvement*. Wiley, 3rd edition, 2011.

[39] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[40] D. Sculley, Gary Holt, Daniel Golovin, et al. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28, pages 2503–2511, 2015.

[41] Walter A. Shewhart. *Economic Control of Quality of Manufactured Product*. Van Nostrand, 1931.

[42] D. H. Stamatis. *Failure Mode and Effects Analysis: FMEA from Theory to Execution*. ASQ Quality Press, 2nd edition, 2003.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[44] Lei Wang, Chen Ma, Xueyang Feng, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[45] Jason Wei, Yi Tay, Rishi Bommasani, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[46] Laura Weidinger, John Mellor, Maribeth Rauh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[47] Western Electric Company. *Statistical Quality Control Handbook*. Western Electric Company, 1956.

[48] Donald J. Wheeler and David S. Chambers. *Understanding Statistical Process Control*. SPC Press, 2nd edition, 1992.

[49] William H. Woodall. Controversies and contradictions in statistical process control. *Journal of Quality Technology*, 32(4):341–350, 2000.

[50] Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of ICLR*, 2023.

Index

This body of work establishes that Large Language Model and agentic AI systems are stochastic processes amenable to Statistical Process Control, and translates the DMAIC improvement cycle into a rigorous, auditable governance methodology for behavioral assurance.

Through 40 tools classified by a formal Adopt, Adapt, Innovate taxonomy, three parallel worked examples across clinical, insurance, and emergency medicine domains, and the original BME Metric Suite for distribution-free behavioral measurement, the work provides standards bodies, governance researchers, and practitioners with a complete process-level implementation methodology.

ORIGINAL CONTRIBUTIONS

ALAGF	Adaptive Lifecycle Agentic Governance Framework
BME Metric Suite	ECPI, BAR, ECI, SPAR, CBMES
AAI Taxonomy	40-tool classification across five DMAIC phases
SymPrompt+	Governed behavioral modulation protocol
MIDCOT	SPC-driven drift detection and escalation
AI-FMEA	Cascading failure model for agentic systems

STANDARDS ALIGNMENT

ISO/IEC 42001:2023 · NIST AI RMF 1.0 · EU AI Act
IEEE P2863 / 7010 · ISO/IEC 23053 · ISO 9001:2015

WHO IS THIS BOOK FOR

- Standards bodies seeking process-level AI governance specifications
- AI governance researchers building measurement frameworks
- Practitioners deploying LLMs in regulated environments
- Quality management professionals extending Six Sigma to AI

Dale Rutherford, PhD *is an Ethical AI Strategist and researcher specializing in the integration, deployment, and governance of agentic AI systems. His work applies Lean Six Sigma methodologies, statistical process control, and behavioral measurement theory to produce operationalizable governance frameworks aligned with international AI standards for LLM and agentic AI systems deployed in regulated, high-stakes environments.*

ISBN 979-8-90452-913-0

The Center for Ethical AI
Little Rock, AR, USA