The Reliability-First Standard (RFS)

Applying Industrial Metrology and MSA to Large Language Model Validation

Version 1.1 — Proposed Standard
January 2026

Author
Abdul Martinez

Prime Studios
Applied Research Division

Abstract

Current Large Language Model (LLM) benchmarks such as MMLU, HumanEval, and Chatbot Arena measure capability—the theoretical ceiling of what a model can achieve under optimal conditions. However, for AI systems to be integrated into mission-critical infrastructure, stakeholders require measures of reliability—the practical floor of what users can consistently expect. This gap between capability measurement and reliability assurance represents a fundamental obstacle to responsible AI deployment in regulated industries.

This paper introduces the Reliability-First Standard (RFS), a framework that transposes over a century of industrial metrology principles onto AI evaluation. Drawing from Gage Repeatability and Reproducibility (GR&R) studies, Measurement System Analysis (MSA), and Six Sigma process capability methodologies, we propose treating Large Language Models as measurement instruments subject to the same rigorous qualification requirements applied to physical gages in aerospace, medical device, and automotive manufacturing.

A critical clarification: RFS does not redefine correctness, nor does it assume language outputs must reduce to scalar values. Rather, it quantifies the uncertainty and repeatability of model responses relative to defined task references. This approach aligns with established metrology practice for non-scalar measurements including optical inspection, spectral analysis, and human-rated instruments—systems where consistency matters as much as absolute truth.

The core contribution is the L-MSA (LLM Measurement System Analysis) framework, which decomposes model output variance into Equipment Variation (stochastic repeatability), Appraiser Variation (linguistic prompt sensitivity), and temporal Stability. We introduce the concept of Linguistic %GR&R, propose process capability indices (Cp/Cpk) adapted for language models, and present the Digital Calibration Certificate as a standardized artifact for communicating model reliability to enterprise and regulatory stakeholders.

The RFS framework addresses a critical gap identified by the National Institute of Standards and Technology (NIST), which has called for methods to quantify uncertainty in AI evaluations and assess the validity of benchmark instruments. By shifting focus from single-shot accuracy scores to uncertainty-qualified reliability floors, this standard aims to transform LLMs from unpredictable oracles into calibrated industrial assets suitable for safety-critical deployment.

Keywords: Large Language Models, Metrology, Measurement System Analysis, GR&R, Process Capability, AI Safety, Reliability Engineering, Calibration, Uncertainty Quantification, NIST AI RMF

I. Introduction: The Metrology Gap

1.1 The Problem

The artificial intelligence industry has developed sophisticated benchmarks for evaluating Large Language Models. MMLU (Massive Multitask Language Understanding) tests knowledge across 57 academic subjects. HumanEval assesses code generation capabilities. Chatbot Arena employs crowdsourced human preference to generate Elo-style rankings. These instruments serve a valuable purpose: they provide standardized comparisons of model capability and track progress over time.

However, capability benchmarks answer only one question: What can this model achieve at its best? They do not answer the question that matters for deployment: What can users consistently expect?

In engineering terms, this is a question of measurement system capability, not task difficulty. When a manufacturing facility acquires a Coordinate Measuring Machine (CMM), engineers do not simply verify that it can measure a part correctly once. They conduct a Measurement System Analysis (MSA) to determine whether the instrument produces consistent results across multiple operators, multiple measurement cycles, and varying environmental conditions. The goal is not to establish the CMM's theoretical precision, but its practical reliability in production use.

Current AI benchmarks commit what industrial metrologists would recognize as fundamental malpractice: single-shot measurement with no variance decomposition, no uncertainty quantification, and no traceability to reference standards. A model that scores 87% on MMLU may achieve that score on average, but the benchmark reveals nothing about the variance around that mean. Does the model reliably produce correct answers for the same query? Does rephrasing the question change the result? Does performance drift over time as providers silently update model weights?

These questions are not academic. They determine whether an AI system can be trusted in contexts where consistency matters as much as capability.

1.2 The Measurement Philosophy

A potential objection must be addressed directly: language is not a physical measurement. Metrology traditionally assumes scalar ground truth—a part is either 25.4mm or it is not. Language outputs are high-dimensional, semantic, and often subjective.

This objection, while intuitive, misunderstands the scope of modern metrology. Industrial measurement science has long accommodated non-scalar measurements:

- Optical inspection systems evaluate surface quality against reference images

- Spectral analyzers compare emission patterns to reference spectra
- Human-rated instruments (e.g., color matching, texture grading) quantify inter-rater consistency
- Image analysis systems measure dimensional conformance from visual data

In none of these cases does metrology require reduction to a single scalar value. What metrology requires is: (1) a defined reference against which measurements are compared, (2) quantification of variance across repeated measurements, and (3) decomposition of variance sources to identify whether inconsistency stems from the instrument, the operator, or the item being measured.

RFS applies this same framework to language models. We do not claim to measure "truth" in any absolute sense. We measure response consistency relative to a defined task reference. The task reference may be a correct answer (for deterministic tasks), a policy specification (for constrained tasks), or a human rubric (for evaluative tasks). The contribution is uncertainty quantification, not truth determination.

1.3 The Thesis

This paper proposes that AI models must be treated as Measurement Systems subject to the same rigorous qualification standards applied to physical instruments in regulated industries. We introduce the L-MSA (LLM Measurement System Analysis) framework to bridge the gap between stochastic AI outputs and industrial reliability requirements.

The core insight is a change in perspective. In the L-MSA framework:

- The LLM is the Gage—the instrument performing evaluation
- The Prompt is the Appraiser—the operator using different techniques to obtain a result
- The Logical Problem or Task is the Part—the objective reference being measured against

From this framing, established metrology concepts map directly onto AI evaluation:

Metrology Concept	**AI Translation**	**Significance**
Repeatability	Stochastic Stability	Consistency on identical inputs across trials
Reproducibility	Linguistic Robustness	Consistency across different prompt phrasings
Bias	Systematic Offset	Consistent deviation from ground truth
Stability	Temporal Drift	Performance change over time
%GR&R	Total Linguistic Variance	Combined measurement system variation
Cp/Cpk	Process Capability Index	Ability to perform within specification limits

The Reliability-First Standard provides the methodology to calculate these metrics for language models, enabling stakeholders to make deployment decisions based on quantified reliability rather than capability claims alone.

1.4 Related Work and Differentiation

The application of metrology concepts to AI evaluation is an emerging area with several related efforts:

NIST's Industrial AI Management and Metrology (IAIMM) project, initiated in 2023, advances metrology principles for AI adoption broadly, focusing on manufacturing applications and general AI system characterization. The RFS framework extends this work by providing specific protocols for linguistic variance decomposition in LLMs and introducing the Digital Calibration Certificate as a standardized reporting artifact.

Recent work on applying MSA to LLM evaluation has focused primarily on annotator consistency—treating human raters as the measurement system. RFS inverts this framing: the model itself is the measurement system, and linguistic prompt variation constitutes appraiser variation. This shift enables reliability characterization of the model independent of human annotation variance.

The "Benchmarks as Microscopes" framework proposes "model metrology" with targeted evaluations for specific capabilities. RFS complements this by providing the statistical infrastructure for quantifying uncertainty in any such evaluation, regardless of the capability being measured.

The contribution of RFS is the synthesis: a complete framework that combines variance decomposition (from MSA), process capability metrics (from Six Sigma), ongoing monitoring protocols (from SPC), and a standardized reporting artifact (the Digital Calibration Certificate) into a coherent standard specifically designed for LLM reliability qualification.

1.5 The Regulatory Opportunity

The timing for this framework is propitious. Current benchmarks face well-documented challenges. MMLU has become saturated, with top models exceeding 90% accuracy, eliminating meaningful differentiation. Research has revealed data contamination in benchmarks like GSM8K, where models reproduce memorized answers rather than demonstrating reasoning. The Chatbot Arena has faced criticism for gaming vulnerabilities, with reports of providers testing dozens of private variants and selectively publishing only favorable results.

Simultaneously, regulatory bodies are calling for the capabilities this framework provides. NIST's Center for AI Standards and Innovation (CAISI) has identified key open questions including: How should evaluators identify, quantify, and communicate sources of uncertainty in

an evaluation setup? To what degree are benchmark results sensitive to prompt selection and task design? The NIST AI Risk Management Framework explicitly calls for rigorous testing with associated measures of uncertainty—yet no standard methodology exists to provide this.

The RFS framework offers regulated industries a familiar vocabulary and methodology for qualifying AI systems, while providing the AI community with rigorous tools for characterizing reliability beyond simple accuracy metrics.

II. Foundations: 100 Years of Measurement Science

2.1 Historical Context

The science of measurement has evolved from craft practice to rigorous discipline over the past century. In the early industrial era, measurement was the province of skilled artisans who calibrated instruments by feel and experience. The interchangeability requirements of mass production forced standardization: first of physical reference artifacts (gauge blocks, standard weights), then of measurement procedures, and finally of the statistical methods for characterizing measurement uncertainty.

The transformation accelerated through the quality revolution of the late twentieth century. When Japanese manufacturing demonstrated that statistical process control could achieve defect rates orders of magnitude lower than Western competitors, industries worldwide adopted methodologies codified in standards like the AIAG Measurement Systems Analysis Manual. Today, no regulated manufacturing environment deploys a measurement instrument without rigorous qualification studies.

AI evaluation currently occupies the position of measurement science circa 1920: valuable instruments exist, but the discipline lacks standardized protocols, uncertainty quantification methods, and qualification requirements that enable trustworthy deployment. The RFS framework proposes to advance AI evaluation by importing the hard-won lessons of physical metrology.

2.2 Measurement Without Physical Units

A common misconception holds that metrology applies only to measurements expressible in SI units—length, mass, time. In practice, industrial metrology has long addressed non-physical measurements:

- Visual inspection systems grade surface defects against reference standards
- Color measurement quantifies perceptual differences using standardized color spaces
- Texture analysis compares samples to reference textures via image processing
- Sensory evaluation panels rate products on trained perceptual scales

In each case, the measurement output is not a simple scalar in physical units. What makes these systems amenable to metrology is not the nature of the output, but the presence of: (1) defined reference standards, (2) repeatable measurement protocols, and (3) statistical methods for quantifying variance and uncertainty.

LLM evaluation fits this pattern. The output is linguistic rather than physical, but the metrology requirements are identical: defined task references, repeatable evaluation protocols, and statistical variance decomposition. RFS provides the methodology to satisfy these requirements.

2.3 The Six Sigma Legacy

The Six Sigma methodology, developed at Motorola in the 1980s and refined at General Electric, established that measurement system qualification is prerequisite to process improvement. The core insight: you cannot improve what you cannot reliably measure.

Gage Repeatability and Reproducibility (GR&R) studies became the standard tool for qualifying measurement systems. The methodology partitions observed variation into components attributable to the measurement instrument (repeatability), the operator (reproducibility), and the parts being measured (part variation). Industry standards establish clear thresholds:

- Less than 10% GR&R: Measurement system is capable
- 10-30% GR&R: May be acceptable depending on application criticality
- Greater than 30% GR&R: Measurement system is not adequate

These thresholds exist because measurement uncertainty directly impacts decision quality. If your measurement system contributes 30% of observed variation, you cannot reliably distinguish between parts that are truly different—the noise drowns the signal. The same principle applies to AI evaluation: if prompt phrasing contributes significant variance to a model's output, your benchmark cannot reliably distinguish between models that are truly different in capability.

2.4 Terminology Alignment

To apply metrology principles to AI evaluation, we must establish clear mappings between traditional and digital-linguistic contexts:

Traditional Term	L-MSA Definition	Operational Meaning
Gage	The LLM / Inference Engine	The system under evaluation
Appraiser	Linguistic Framing (Prompt)	The method of eliciting a response
Part	The Logic Problem / Task	The reference being measured against
Repeatability	Stochastic Stability	Same prompt, same model, multiple trials
Reproducibility	Linguistic Robustness	Same logic, different prompt phrasings
Bias	Systematic Offset	Consistent directional error

Stability	Temporal Consistency	Performance over time intervals
Linearity	Difficulty Scaling	Consistency across task complexity

This terminology alignment enables practitioners familiar with industrial metrology to immediately understand L-MSA concepts, while providing AI researchers with rigorous definitions grounded in established measurement science.

III. The L-MSA Framework: Core Methodology

The L-MSA framework consists of three primary study types, each addressing a distinct source of variation in LLM outputs. Together, these studies produce a comprehensive picture of model reliability that goes far beyond single-point accuracy metrics.

3.1 Equipment Variation: Repeatability Study

The Repeatability Study measures the internal noise of the model—variation that occurs even when all external factors are held constant. This is the AI equivalent of asking: if I measure the same part with the same gage multiple times, how much do the readings vary?

3.1.1 The Temperature Paradox

In traditional metrology, a gage is ideally deterministic—the same input produces the same output. LLMs introduce a complication: many deployments intentionally use temperature > 0 for reasoning diversity or creative applications. This raises a question: how does RFS distinguish between allowable stochasticity and unreliable variance?

The framework addresses this through a two-tier repeatability protocol:

Zero-Baseline Repeatability (T=0):

Characterizes the model's deterministic behavior with all sampling randomness eliminated. This establishes the theoretical floor of the model's consistency. Significant variance at T=0 indicates fundamental instability unrelated to intentional sampling.

Operational Repeatability (T=target):

Characterizes variance at the intended deployment temperature. This measures the combined effect of model behavior and intentional stochasticity. The difference between Zero-Baseline and Operational repeatability quantifies the variance contribution of the sampling strategy itself.

Both measurements are reported in the Digital Calibration Certificate. For safety-critical applications, Zero-Baseline repeatability may be required; for creative applications, Operational repeatability at higher temperatures may be acceptable if the variance remains within specification limits.

3.1.2 Protocol

- Select a representative set of test prompts with defined correct responses
- Execute each prompt N times (minimum N=30 for statistical validity) at T=0
- Repeat at operational temperature T=target
- Hold all other inference parameters constant (top_p, system prompt, context)

- Record all outputs and their relationship to the task reference

3.1.3 Metrics

- Response Consistency: Percentage of trials producing identical outputs
- Semantic Variance: For non-identical outputs, measure semantic distance using embedding similarity
- Answer Stability: For tasks with discrete correct answers, variance in correctness across trials

A high-reliability model should demonstrate minimal variance across repeated trials of identical prompts at T=0. Variance at operational temperature should be characterized and bounded within acceptable limits for the deployment context.

3.2 Appraiser Variation: Reproducibility Study

The Reproducibility Study measures how sensitive the model is to prompt phrasing—the linguistic equivalent of operator variation. This addresses a fundamental question: if two users ask the same logical question using different words, do they get consistent answers?

3.2.1 The Universal Appraiser Matrix

We propose a standardized set of linguistic transformations that allow systematic evaluation of prompt sensitivity. Each logical problem is presented through multiple linguistic framings:

Framing Style	Description	Example Transformation
Formal Academic	Technical, precise language	"Determine the derivative of f(x)..."
Casual Conversational	Natural, informal phrasing	"Hey, can you help me figure out..."
Minimal/Terse	Stripped-down, essential only	"Derivative: $f(x)=x^2$. Find."
Verbose/Contextual	Extensive background provided	"Given the following context..."
Implicit/Indirect	Meaning requires inference	"I'm stuck on this calculus thing..."
Adversarial/Edge	Unusual syntax, typos, ambiguity	"derivtive of x sqared??"
Chain-of-Thought	Explicit reasoning request	"Think step by step..."
Direct Answer	No reasoning scaffolding	"Answer only, no explanation..."
Few-Shot	Examples provided	"Here are examples: [ex1], [ex2]..."
Zero-Shot	No examples or scaffolding	Raw question only

The inclusion of Chain-of-Thought vs. Direct and Few-Shot vs. Zero-Shot framings is critical: these represent major sources of variance in frontier models, with accuracy swings of 20% or more observed between prompting strategies for identical logical problems.

3.2.2 Protocol

- Create semantically equivalent prompts across all framing styles
- Execute each variant multiple times to separate appraiser variation from equipment variation
- Compare outputs across framings for consistency of reasoning and conclusions

3.2.3 Metrics

- Cross-Framing Agreement: Percentage of framing pairs producing consistent answers
- Framing Sensitivity Index: Which styles produce the most variance
- Worst-Case Degradation: Maximum accuracy drop between best and worst framing

A model with high reproducibility produces consistent results regardless of how the question is phrased. A model with poor reproducibility may appear capable when prompted optimally but fail unpredictably in real-world use where users phrase questions in diverse ways.

3.3 Temporal Stability Study

The Stability Study measures whether model performance remains consistent over time. This is critical because, unlike physical gages that maintain calibration until explicitly recalibrated, LLMs deployed via API may be silently updated by providers.

3.3.1 Protocol

- Establish a baseline L-MSA characterization at time T_0
- Re-execute the same evaluation protocol at regular intervals (weekly, monthly)
- Track performance metrics over time, noting any model version changes
- Identify drift patterns and establish re-qualification triggers

3.3.2 Metrics

- Baseline Deviation: Current performance relative to initial characterization
- Drift Rate: Trend in performance change over time
- Volatility: Variance in performance across time periods

The Stability Study enables organizations to detect when a model has drifted outside its qualified operating envelope, triggering re-evaluation before the drift causes operational problems.

3.4 The %GR&R Calculation

The culminating metric of L-MSA is the Linguistic %GR&R—a single percentage representing the total measurement system variation attributable to the model (gage) and prompt (appraiser) combined, relative to the total observed variation.

Conceptually, the calculation partitions total variance:

- Total Variance = Part Variance + Equipment Variance + Appraiser Variance
- %GR&R = (Equipment Variance + Appraiser Variance) / Total Variance × 100

A low %GR&R indicates that most variation comes from actual differences in task difficulty (part variation), meaning the measurement system reliably distinguishes between tasks. A high %GR&R indicates that the model and prompt contribute significant noise, limiting the system's ability to make reliable distinctions.

3.4.1 Part Variance and Contamination

A critical assumption in MSA is that Part Variance reflects true variation in the items being measured. For LLM evaluation, this assumption can be violated by data contamination: if the model has memorized benchmark answers during training, observed "Part Variance" may actually reflect recall consistency rather than reasoning capability.

L-MSA studies should prioritize novel or held-out tasks to ensure Part Variance reflects true task difficulty, not memorization. Where contamination is suspected, fresh task generation or dynamic benchmarks (e.g., LiveBench) are required. The Digital Calibration Certificate should document whether contamination testing was performed and the provenance of evaluation tasks.

3.4.2 Industry Thresholds

The industry-standard thresholds apply:

- Less than 10%: System is capable of reliable measurement
- 10-30%: Marginal—may be acceptable for non-critical applications
- Greater than 30%: System is not capable—reliability improvements required

3.5 Task Taxonomy

The applicability of L-MSA metrics depends on the nature of the task being evaluated. Not all tasks have the same type of "ground truth," and the framework must accommodate this variation:

Task Type	Ground Truth Definition	L-MSA Application	Examples

Deterministic	Single correct answer exists	Full accuracy + variance	Math, code, factual extraction
Spec-Bound	Policy or rule set defines acceptability	Constraint violation rate + variance	Safety classification, content moderation
Evaluative	Human rubric defines quality	Inter-rater agreement + variance	Summarization, style, reasoning quality

For Deterministic tasks, L-MSA applies directly with binary correctness as the primary metric. For Spec-Bound tasks, the "correct answer" is compliance with the specification, and variance measures consistency of compliance decisions. For Evaluative tasks, L-MSA quantifies the consistency of model outputs relative to a defined rubric, acknowledging that multiple outputs may be acceptable.

The Digital Calibration Certificate must specify which task types were evaluated and the metrics applicable to each. A model qualified on Deterministic tasks should not be assumed reliable for Evaluative tasks without separate characterization.

3.6 Language as Appraiser Variable

The Universal Appraiser Matrix as presented assumes a single language (English). However, language itself constitutes a fundamental dimension of appraiser variation. Preliminary studies indicate that cross-lingual variance can exceed intra-lingual style variation by an order of magnitude.

For multilingual deployments, the L-MSA protocol must be extended to include semantically equivalent prompt sets across target languages. Any deployment claiming multilingual capability should document the languages against which L-MSA qualification was performed, and the cross-lingual variance metrics for each language pair.

Full specification of a Cross-Lingual Appraiser Matrix is reserved for future work, but we note that this extension is essential for global deployment of AI in safety-critical contexts where translation errors carry significant consequences.

IV. Calculating the Uncertainty Budget

Industrial metrology requires that every measurement be expressed with an associated uncertainty. The statement "this part measures 25.4mm" is incomplete without "±0.05mm (k=2, 95% confidence)." The RFS framework brings this discipline to AI evaluation.

A critical clarification: all uncertainty statements in RFS are conditional, not absolute. They are valid only within the qualified operating envelope—the specific model version, inference parameters, task types, and deployment conditions documented in the calibration certificate. This mirrors ISO 17025 practice for physical calibration laboratories.

4.1 Sources of Uncertainty

The L-MSA framework identifies and quantifies multiple sources of uncertainty in LLM outputs:

Uncertainty Source	Origin	Mitigation
Stochastic Sampling	Temperature, top-p, random seed	Fixed parameters, Zero-Baseline testing
Prompt Sensitivity	Linguistic framing variation	Universal Appraiser Matrix testing
Context Sensitivity	Preceding conversation, system prompt	Controlled context conditions
Temporal Drift	Model updates, API changes	Stability monitoring, version pinning
Environmental	Hardware, batch size, load	Controlled inference environment
Part Contamination	Memorized training data	Novel task generation, held-out sets

4.2 The Linguistic Noise Floor

Every LLM has a linguistic noise floor—the minimum variance introduced by reasonable variations in how questions are asked. This floor exists because natural language is inherently variable; there is no single "correct" way to phrase a question.

Quantifying this floor is essential for setting realistic reliability expectations. A model cannot be more reliable than its linguistic noise floor permits. If reasonable prompt variations introduce 5% variance in outputs, no amount of model improvement can reduce total variance below that floor without constraining acceptable input language.

4.3 Error Propagation in Multi-Model Systems

In many deployments, LLMs are components in larger systems: RAG pipelines, multi-agent workflows, tool-use chains. Uncertainty propagates and compounds through these systems.

Simple error compounding: A 5% error rate at each of four independent stages compounds to approximately 19% system error rate ($1 - 0.95^4 \approx 0.19$).

Process capability compounding: Two components with Cpk=1.33 do not yield a system with Cpk=1.33. The joint probability distribution of failures must be calculated, accounting for correlation between component failures. In general, system-level Cpk is lower than component-level Cpk, often significantly so.

The uncertainty budget must account for this propagation. System-level reliability cannot be inferred from component-level benchmarks alone. The RFS framework recommends that pipeline deployments conduct system-level L-MSA studies in addition to component characterization. Future work on A-MSA (Agentic Measurement System Analysis) will provide detailed protocols for multi-agent systems.

4.4 Measurement Resolution (ndc)

The Number of Distinct Categories (ndc) metric from traditional MSA answers a crucial question: how many meaningfully different values can this measurement system distinguish?

Applied to AI evaluation, ndc reveals whether benchmark scales are meaningful. If a model's response variance means it can only reliably distinguish "poor / acceptable / good" performance levels, then reporting results on a 100-point scale creates false precision. The ndc calculation grounds scoring systems in the actual measurement capability of the evaluation.

A minimum ndc of 5 is generally required for a measurement system to be considered capable of useful discrimination.

V. Process Capability for LLMs

Process Capability indices (Cp and Cpk) answer a fundamental question: can this process consistently produce outputs within specification limits? In manufacturing, these metrics determine whether a production line can reliably make parts that meet engineering tolerances. For LLMs, they determine whether a model can reliably produce outputs that meet application requirements.

5.1 Defining Specification Limits

Before calculating process capability, specification limits must be defined. For LLMs, these limits represent the boundaries of acceptable performance for a given application:

- Upper Specification Limit (USL): Maximum acceptable error rate, response time, or other metric
- Lower Specification Limit (LSL): Minimum acceptable accuracy, relevance, or quality score
- Target: Ideal performance level

Specification limits are application-dependent. A customer service chatbot may tolerate 10% unhelpful responses; a medical triage system may require less than 1% critical misclassifications. The RFS framework does not prescribe universal limits but requires that limits be explicitly defined and documented for each deployment context.

5.2 The Capability Index

The Cp index measures whether the process spread fits within specification limits, assuming the process is centered. It compares the tolerance width (USL - LSL) to the process spread (typically 6 standard deviations).

The Cpk index accounts for process centering—how close the mean is to the target. A process with good Cp but poor Cpk is capable but not centered; it could meet specifications if adjusted.

Standard interpretations:

- $Cpk \geq 1.33$: Process is capable (standard requirement)
- $Cpk \geq 1.67$: Process is capable with margin (required for safety-critical)
- $Cpk \geq 2.0$: Six Sigma level capability
- $Cpk < 1.0$: Process is not capable—outputs will exceed specification limits

5.3 From Accuracy to Reliability Maturity

Current AI discourse focuses on accuracy percentages: "Model X achieves 95% accuracy on benchmark Y." This framing obscures reliability. A 95% accuracy rate means 5 failures per 100 attempts—a defect rate of 50,000 per million.

Six Sigma processes achieve 3.4 defects per million opportunities (DPMO). This level serves as a reference point for reliability maturity, not an expectation for current foundation models. The gap between current AI performance and Six Sigma expectations quantifies the reliability improvement required for truly mission-critical deployment.

More importantly, the Cpk framework shifts the conversation from capability claims to reliability requirements: What defect rate is acceptable for this application? Can this model demonstrably achieve that rate with quantified variance? A model with 95% mean accuracy but high variance may have Cpk < 1.0, meaning its outputs will regularly exceed acceptable limits despite strong average performance.

For safety-critical applications in regulated industries, this distinction is essential. The L-MSA framework provides the tools to measure and communicate reliability in terms that quality engineers already understand.

VI. Failure Mode Analysis for AI Systems

Failure Mode and Effects Analysis (FMEA) is a systematic methodology for identifying potential failures, assessing their severity and likelihood, and prioritizing mitigation efforts. The RFS framework adapts FMEA for AI systems, providing structured analysis of how LLMs fail, not just how often.

6.1 Severity, Occurrence, and Detection

Traditional FMEA scores each failure mode on three dimensions:

- Severity (S): How serious is the impact if this failure occurs? (1-10 scale)
- Occurrence (O): How likely is this failure to occur? (1-10 scale)
- Detection (D): How likely is the failure to be detected before causing harm? (1-10 scale)

The Risk Priority Number (RPN) = S × O × D identifies which failure modes demand the most urgent attention. High RPN scores indicate failures that are severe, frequent, and difficult to detect—the most dangerous combination.

6.2 AI-Specific Failure Modes

LLMs exhibit characteristic failure modes that should be systematically analyzed:

Failure Mode	Description	Detection Challenge
Hallucination	Confident presentation of false information	Low—output appears authoritative
Prompt Injection	Adversarial input bypasses intended behavior	Medium—requires monitoring
Distribution Shift	Performance degrades on out-of-distribution inputs	Low—silent degradation
Linguistic Brittleness	Inconsistent outputs for equivalent inputs	Medium—requires L-MSA testing
Context Overflow	Performance degrades with long contexts	Medium—gradual degradation
Sycophancy	Model agrees with user rather than being accurate	Low—appears cooperative
Latent Saliency	Specific token sequences trigger outlier behavior	Low—appears random, hard to predict

The addition of Latent Saliency addresses failures that occur not because of prompt structure or task difficulty, but because specific token sequences trigger outlier behavior rooted in training

data artifacts. These failures are particularly dangerous because they appear random and are difficult to anticipate or test for systematically.

6.3 Root Cause Decomposition

The L-MSA framework enables root cause analysis of failures by decomposing variance sources. When a failure occurs, the framework asks: Is this failure attributable to the Gage (model weights, architecture), the Appraiser (prompt formulation, linguistic framing), or the Part (inherent task difficulty, ambiguity in ground truth)?

This decomposition is essential for effective remediation:

- High Equipment Variance: Model instability—consider different model, lower temperature, or ensemble approaches
- High Appraiser Variance: Prompt sensitivity—develop robust prompt templates, constrain input formats
- High Part Variance: Task ambiguity—clarify specifications, accept higher variance as inherent

Without systematic decomposition, improvement efforts may address symptoms rather than causes.

6.4 Bias and Equity Considerations

The FMEA framework should extend beyond technical failures to include bias and equity considerations. Systematic offset (bias in metrology terms) may manifest not just as overall accuracy degradation but as differential performance across demographic groups, cultural contexts, or language varieties.

L-MSA studies for deployments with diverse user populations should include:

- Demographic disaggregation of accuracy and variance metrics
- Cross-cultural prompt testing within the Appraiser Matrix
- Analysis of whether Part Variance differs systematically across user groups

This aligns with NIST AI RMF emphasis on fairness and equity as dimensions of trustworthy AI. The goal is not to replace specialized fairness audits, but to integrate equity considerations into the standard reliability characterization process.

VII. Statistical Process Control for Deployed Models

Qualification is not a one-time event. Physical measurement systems require periodic recalibration; AI systems require ongoing monitoring. Statistical Process Control (SPC) provides the methodology for continuous reliability assurance.

7.1 Control Charts for Inference

Control charts track process performance over time, distinguishing normal variation from signals of process change. For deployed LLMs, control charts monitor key metrics derived from L-MSA:

- Accuracy/correctness rate on ongoing evaluation samples
- Response consistency metrics
- User feedback signals (escalations, corrections, abandonment)
- Latency and throughput characteristics

Control limits are established from the baseline L-MSA characterization, typically at ±3 standard deviations from the mean. Points outside control limits, or patterns suggesting non-random variation, trigger investigation.

7.2 Out-of-Control Signals

Standard SPC rules (Western Electric rules) identify signals requiring attention:

- Single point beyond 3σ control limits
- Two of three consecutive points beyond 2σ
- Four of five consecutive points beyond 1σ
- Eight consecutive points on one side of center line (drift)
- Six consecutive points steadily increasing or decreasing (trend)

When these signals appear, the model should be flagged for investigation and potential re-qualification.

7.3 Re-Qualification Triggers

The RFS framework defines explicit triggers for re-qualification:

- Control chart signals indicating process shift or drift
- Known model version updates from the provider
- Changes to inference infrastructure or parameters

- Elapsed time exceeding the qualified calibration interval
- Expansion to new use cases or domains not covered by original qualification

Re-qualification does not necessarily require full L-MSA repetition. Abbreviated studies may suffice if changes are limited in scope. However, the decision to abbreviate should be documented and justified.

VIII. Safety-Critical Deployment Requirements

Regulated industries—medical devices, aerospace, automotive, financial services—operate under frameworks that demand demonstrated reliability, not just claimed capability. The RFS framework aligns AI qualification with these existing requirements.

Important disclaimer: RFS does not replace domain-specific certification requirements. Medical AI systems must still comply with FDA regulations; aerospace systems with DO-178C; automotive with ISO 26262. RFS provides the measurement infrastructure to support these certifications, not a substitute for them.

8.1 The Measurement Capability Gate

A foundational principle of industrial metrology: measurement system uncertainty must be substantially smaller than the tolerance being verified. The common rule requires measurement uncertainty to be no more than 10% of the tolerance.

For AI in safety-critical contexts, this principle translates to a gate: if the uncertainty in model performance (as quantified by L-MSA) exceeds 10% of the acceptable error tolerance, the model is not qualified for that application. A model might average excellent performance but, if variance is high, cannot be trusted for safety-critical use.

8.2 Regulatory Framework Alignment

The RFS framework maps to existing regulatory requirements:

Industry	Governing Standards	MSA Requirement	RFS Alignment
Medical Devices	FDA 21 CFR 820, IEC 62304	Required for production equipment	L-MSA for AI components
Aerospace	AS9100, DO-178C	Mandatory for measurement systems	Process capability for AI
Automotive	ISO 26262, IATF 16949	Cpk ≥ 1.33 minimum	Cpk calculation for LLMs
General Quality	ISO 9001	Suitable monitoring resources	SPC for deployed models
AI-Specific	NIST AI RMF, ISO/IEC 42001	Risk-based evaluation	Uncertainty quantification

Organizations already compliant with these frameworks have existing MSA infrastructure and expertise. The RFS framework enables them to extend that infrastructure to AI components using familiar methodologies.

8.3 Traceability Requirements

Regulated industries require traceability: the ability to trace any measurement back through an unbroken chain to reference standards. For AI systems, traceability means:

- Model version traceability: Exact identification of model weights, version, and configuration
- Prompt traceability: Documentation of system prompts, templates, and few-shot examples
- Evaluation traceability: Reference to specific benchmark versions and scoring criteria
- Calibration traceability: Link from L-MSA results to the specific studies that produced them

The Digital Calibration Certificate (Section IX) provides the artifact for documenting and communicating this traceability chain.

IX. Practical Implementation

9.1 The Digital Calibration Certificate

Physical measurement instruments come with calibration certificates documenting their qualified performance. The RFS framework proposes an analogous artifact for AI systems: the Digital Calibration Certificate.

A compliant certificate includes the following elements:

Section 1: Model Identification

- Model name, version, and unique identifier (weights hash if available)
- Provider and API version (if applicable)
- Inference configuration: temperature, sampling parameters, context limits
- System prompt and any fixed templates

Section 2: Qualification Scope

- Intended use case(s) and application domain(s)
- Task types covered (Deterministic, Spec-Bound, Evaluative)
- Languages qualified (with cross-lingual variance if multilingual)
- Operating envelope: acceptable input characteristics, context lengths, etc.

Section 3: L-MSA Results

- Repeatability study results: Zero-Baseline and Operational, N trials, consistency metrics
- Reproducibility study results: Appraiser matrix coverage, cross-framing agreement
- Stability study results: Baseline date, drift observations, stability classification
- Combined %GR&R calculation and classification
- Contamination testing status and task provenance

Section 4: Process Capability

- Specification limits for the qualified application
- Calculated Cp and Cpk indices

- Capability classification and any conditions

Section 5: Uncertainty Statement

- Combined uncertainty budget
- Major uncertainty contributors
- Confidence level and coverage factor

Section 6: Limitations and Conditions

- Conditions under which the qualification is void
- Known failure modes and mitigations
- Use cases explicitly excluded from qualification

Section 7: Validity

- Issue date and issuing authority
- Calibration interval: recommended re-qualification date
- Version control: certificate revision history

9.2 Automation and Tooling

Manual L-MSA studies are labor-intensive. Practical implementation requires automation:

- Prompt Generators: Tools to automatically generate Universal Appraiser Matrix variants
- Evaluation Harnesses: Integration with frameworks like HELM, DeepEval, or custom harnesses
- Statistical Engines: Automated calculation of L-MSA metrics, %GR&R, and capability indices
- Monitoring Dashboards: Real-time SPC visualization (integration with Prometheus, Grafana, or similar)
- Certificate Generators: Standardized reporting tools for calibration certificates

The RFS framework specifies methodologies and metrics; implementation tooling is expected to emerge from both commercial vendors and open-source community efforts.

9.3 Cost Considerations

Rigorous L-MSA studies require substantial evaluation runs. For expensive API-based models, cost is a practical constraint. The framework accommodates this through:

- Statistical power calculations to determine minimum viable sample sizes
- Staged qualification: screening studies to identify issues before full characterization
- Sampling strategies for large prompt spaces
- Hybrid approaches: use cheaper models for initial screening, reserve expensive models for final qualification
- Cached evaluation reuse where conditions permit

The cost of qualification must be weighed against the cost of deployment failures. For safety-critical applications, robust qualification is not optional regardless of evaluation expense. This parallels physical metrology: calibrating a CMM is expensive, but deploying an uncalibrated CMM in production is far more costly.

X. Integration with Governance Frameworks

10.1 Relationship to V.A.L.I.D.

The RFS framework provides measurement infrastructure that complements governance frameworks. The V.A.L.I.D. Framework (Value-Aligned Logic and Identity Determinism) addresses how AI systems should behave—specifying requirements for alignment, consistency, and governance. RFS addresses how to verify that systems meet these requirements reliably.

Governance without measurement is aspirational. Measurement without governance is directionless. Together, they form a complete system:

- V.A.L.I.D. specifies behavioral requirements → RFS specification limits
- L-MSA studies verify consistency with requirements
- Calibration certificates document compliance status
- SPC monitoring detects drift from alignment

RFS is the "how" to V.A.L.I.D.'s "what."

10.2 NIST AI RMF Alignment

The NIST AI Risk Management Framework (AI RMF) organizes AI governance around four functions: Govern, Map, Measure, and Manage. The RFS framework directly supports the Measure function:

- Govern: RFS provides standardized methodology for evaluation requirements
- Map: L-MSA identifies sources of variation and reliability risks
- Measure: Core RFS contribution—rigorous uncertainty quantification
- Manage: SPC enables ongoing monitoring and re-qualification triggers

NIST has explicitly called for methods to quantify uncertainty and assess benchmark validity. The RFS framework answers this call with concrete methodology grounded in established measurement science.

10.3 ISO/IEC 42001 Alignment

ISO/IEC 42001 specifies requirements for AI management systems. The RFS framework supports conformance by providing:

- Documented procedures for AI system evaluation (L-MSA protocol)
- Objective evidence of system capability (calibration certificates)
- Continuous monitoring mechanisms (SPC)

- Criteria for re-evaluation (qualification triggers)

Organizations pursuing ISO 42001 certification can incorporate RFS methodology as the technical foundation for their AI evaluation processes.

XI. Discussion: The Path to Standardization

11.1 Adoption Pathway

The RFS framework is proposed as a voluntary standard, following the model of quality standards that achieved widespread adoption through demonstrated value rather than regulatory mandate. The adoption pathway proceeds through stages:

- Pilot implementations by early adopters in regulated industries
- Refinement based on practical implementation experience
- Development of open-source tooling and reference implementations
- Industry consortium formation for standard governance
- Integration with existing quality management systems
- Potential regulatory recognition or incorporation

11.2 Open Questions and Future Work

This specification leaves several questions for future development:

Technical Extensions:

- Full specification of the Universal Appraiser Matrix with validated transformations
- Cross-Lingual Appraiser Matrix for multilingual qualification
- V-MSA: Extension to computer vision models
- A-MSA: Extension to agentic systems with compound uncertainty propagation
- Pipeline MSA for multi-model systems with joint Cpk calculations
- Formal mathematical specification of σ calculation for embedding-based semantic distance

Implementation Questions:

- Minimum sample sizes for statistical validity across study types (power analysis)
- Cost-optimized sampling strategies for expensive API models
- Automation requirements and reference tool architectures
- Inter-laboratory reproducibility of L-MSA results

Governance Questions:

- Accreditation requirements for L-MSA evaluation laboratories

- Certification body structure and governance
- Integration pathways with existing regulatory frameworks

11.3 The Standard as Living Document

Unlike physical metrology standards that may remain stable for decades, AI metrology operates in a rapidly evolving landscape. The RFS framework must be maintained as a living document, with regular review cycles to incorporate lessons learned, address emerging model architectures, and respond to evolving regulatory requirements.

Version control, change management, and stakeholder input processes are essential elements of standard governance. The specification presented here should be understood as a foundation for iterative refinement rather than a final specification.

11.4 Conclusion

The gap between AI capability claims and deployment reliability represents a fundamental obstacle to responsible AI adoption in mission-critical contexts. Current benchmarks measure what models can achieve at their best; stakeholders need to know what they can consistently expect.

The Reliability-First Standard addresses this gap by applying over a century of measurement science to AI evaluation. By treating LLMs as measurement instruments subject to rigorous qualification, the L-MSA framework enables uncertainty quantification, process capability assessment, and ongoing reliability monitoring that regulated industries require.

The transformation from "90% accuracy" to "qualified Cpk of 1.45 for specified applications, with documented uncertainty budget and calibration status" represents a fundamental maturation of AI evaluation practice. This transformation is necessary for AI systems to move from experimental curiosities to trusted industrial assets.

The conversation begins here. We invite practitioners, researchers, and regulators to engage with this framework, test its applicability, identify its limitations, and contribute to its refinement. The goal is not to constrain AI development, but to enable its responsible deployment by providing the measurement infrastructure that trustworthy systems require.

— End of Document —

Appendix A: Digital Calibration Certificate Template

DIGITAL CALIBRATION CERTIFICATE

LLM Measurement System Analysis (L-MSA)

RFS v1.1 Compliant

Field	Value
Certificate Number	[UNIQUE-ID]
Issue Date	[DATE]
Valid Until	[DATE + CALIBRATION INTERVAL]
Issuing Authority	[ORGANIZATION]
RFS Version	1.1

1. MODEL IDENTIFICATION

Attribute	Specification
Model Name	[e.g., GPT-4, Claude 3, Llama 3]
Model Version	[Exact version string]
Model Identifier/Hash	[If available]
Provider	[Organization]
API Version	[If applicable]
Inference Temperature (Operational)	[Value]
Inference Temperature (Zero-Baseline)	0
Sampling Parameters	[top_p, top_k, etc.]
Max Context Length	[Tokens]
System Prompt	[Reference to controlled document]

2. QUALIFICATION SCOPE

Attribute	Specification
Intended Application	[Description]
Task Domain	[e.g., Customer Service, Medical Triage]
Task Types Qualified	[Deterministic / Spec-Bound / Evaluative]
Qualified Languages	[List with cross-lingual variance]
Input Constraints	[Character limits, format requirements]
Excluded Use Cases	[Explicitly not qualified for]
Contamination Testing	[Performed / Not Performed]
Task Provenance	[Source of evaluation tasks]

3. L-MSA RESULTS

3.1 Repeatability Study

Metric	Zero-Baseline (T=0)	Operational (T=target)	Classification
Number of Trials (N)	[≥30]	[≥30]	—
Response Consistency	[%]	[%]	[Pass/Marginal/Fail]
Semantic Variance (σ)	[Value]	[Value]	—
Equipment Variation (%EV)	[%]	[%]	[<10% / 10-30% / >30%]

3.2 Reproducibility Study

Metric	Value	Classification
Appraiser Styles Tested	[N styles, including CoT/Direct, Few/Zero-shot]	—
Cross-Framing Agreement	[%]	[Pass/Marginal/Fail]
Worst-Case Degradation	[%]	—
Appraiser Variation (%AV)	[%]	[<10% / 10-30% / >30%]

3.3 Stability Study

Metric	Value	Classification

Baseline Date	[Date]	—
Study Duration	[Weeks/Months]	—
Drift Observed	[Yes/No]	—
Stability Classification	[Stable/Marginal/Unstable]	—

3.4 Combined Results

Metric	Value	Classification
Total %GR&R	[%]	[<10% / 10-30% / >30%]
Number of Distinct Categories (ndc)	[Value]	[≥5 required]
Overall L-MSA Assessment	—	[CAPABLE / MARGINAL / NOT CAPABLE]

4. PROCESS CAPABILITY

Metric	Value	Requirement
Upper Specification Limit (USL)	[Value]	[Application-specific]
Lower Specification Limit (LSL)	[Value]	[Application-specific]
Process Mean (μ)	[Value]	—
Process Std Dev (σ)	[Value]	—
Cp Index	[Value]	[≥1.33 standard]
Cpk Index	[Value]	[≥1.33 standard / ≥1.67 safety-critical]
Capability Assessment	—	[CAPABLE / NOT CAPABLE]

5. UNCERTAINTY STATEMENT

The reported performance metrics have a combined standard uncertainty of [U] with coverage factor k=2, corresponding to approximately 95% confidence. Primary contributors to uncertainty are: [list major sources]. All uncertainty statements are valid only within the qualified operating envelope documented in this certificate.

6. CONDITIONS AND LIMITATIONS

This qualification is valid only under the following conditions:

- Model accessed via [specified API/deployment]
- Inference parameters as documented in Section 1
- Input within specified constraints
- Use cases within qualified scope
- Task types as documented in Section 2

This qualification is VOID if:

- Model version changes from documented version
- Inference parameters are modified
- Use case extends beyond qualified scope
- Validity date is exceeded without re-qualification
- Control chart signals indicate process shift (see SPC requirements)

7. AUTHORIZATION

Role	Name	Signature	Date
Qualified By	[Name]	_____	[Date]
Reviewed By	[Name]	_____	[Date]
Approved By	[Name]	_____	[Date]

This certificate template conforms to the Reliability-First Standard (RFS) v1.1

References

[1] Automotive Industry Action Group (AIAG). Measurement Systems Analysis Reference Manual, 4th Edition. AIAG, 2010.

[2] National Institute of Standards and Technology. AI Risk Management Framework (AI RMF 1.0). NIST, 2023.

[3] National Institute of Standards and Technology. "Accelerating AI Innovation Through Measurement Science." CAISI Research Blog, December 2025.

[4] National Institute of Standards and Technology. Industrial AI Management and Metrology (IAIMM) Project. NIST, 2023.

[5] International Organization for Standardization. ISO/IEC 42001:2023 — Artificial Intelligence Management System. ISO, 2023.

[6] Hendrycks, D., et al. "Measuring Massive Multitask Language Understanding." ICLR 2021.

[7] Zheng, L., et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." NeurIPS 2023.

[8] Wheeler, D.J. and Lyday, R.W. Evaluating the Measurement Process, 2nd Edition. SPC Press, 1989.

[9] Montgomery, D.C. Introduction to Statistical Quality Control, 8th Edition. Wiley, 2019.

[10] International Organization for Standardization. ISO 17025:2017 — General requirements for the competence of testing and calibration laboratories. ISO, 2017.

[11] Stanford Center for Research on Foundation Models. "Holistic Evaluation of Language Models (HELM)." Stanford CRFM, 2022.

[12] Martinez, A. "The V.A.L.I.D. Framework: Value-Aligned Logic and Identity Determinism for AI Agent Governance." Didomi Research, 2026.

[13] Raji, I.D., et al. "AI and the Everything in the Whole Wide World Benchmark." NeurIPS 2021.

[14] Burnell, R., et al. "Rethink reporting of evaluation results in AI." Science, 2023.