Glossary

Glossary

Terminology used across the mavai project family. Terms are defined in the context of the distributional contract methodology; framework-specific annotations and APIs may vary by implementation.

Term	Definition
Baseline Selection	The process of choosing the most appropriate baseline for a probabilistic test based on footprint match and covariate conformance. When multiple baselines exist, the one with the best covariate match is selected.
Categorical Clause	A service-contract clause stating an obligation or prohibition that holds with certainty, not in distribution — e.g. a prohibition on producing illegal content, or a requirement that every response carry an audit identifier. Discharged architecturally (interposed components, structural guarantees) rather than by rate estimation; not promotable from a rate-bounded clause by letting $p_c^* \to 1$. The rate-bounded machinery does not apply. Recognised in exactly one form — **zero-failures** — evaluated observationally. See Statistical Companion §"Clause Forms".
Compliance Testing	The **compliance / assurance procedure**: an affirmative test that a system meets a normative threshold p_{req} given by contract, SLA, SLO, policy, or regulation. The hypothesis is $H_0 : p \leq p_{\text{req}}$ vs $H_1 : p > p_{\text{req}}$; the controlled error is the probability of falsely affirming compliance. Distinct in error semantics from the regression procedure (see **Conformance Testing**). See Statistical Companion §3.6, §1.4.5a.

Term	Definition
Conformance	The overall assessment of whether a service contract's outcome satisfies its configured criteria across all dimensions. The functional dimension partitions into per-criterion verdicts (each criterion hosting one or more postconditions); the latency dimension carries percentile constraints. The contract's composite verdict is assembled from the per-criterion verdicts under the rules of §1.4.6 of the Statistical Companion.
Conformance Testing	The **regression / monitoring procedure**: a reference-control test that detects degradation from a measured baseline $\hat{p}_{\text{baseline}}$. The hypothesis is $H_0 : p \geq p^*$ vs $H_1 : p < p^*$, decided by an integer cutoff c_c chosen so that $P_{\text{ref}}(K < c_c) \leq \alpha$; the controlled error is the false-degradation-alarm rate under the reference. Distinct in error semantics from the compliance procedure (see **Compliance Testing**). See Statistical Companion §3, §1.4.5a.
Confidence	Probability of a correct verdict; equals 1 minus the false positive rate. Part of the parameter triangle.
Confidence-First Approach	User specifies confidence + power + minDetectableEffect; framework computes required samples.
Contract Reference	Human-readable string identifying the document or clause defining a test threshold (e.g., ”SLA v3.2 §5.1.2”).
Covariate	A contextual factor that drives variance in system behavior. Covariates are declared on a use case to indicate which environmental or configuration variables should be tracked for baseline matching and statistical comparison. Unlike functional inputs (factors), covariates represent conditions that affect outcomes but are often outside direct control.
Covariate Category	Classification of a covariate by its nature: temporal (time-based), configuration (deliberate choices), external dependency (third-party services), infrastructure (execution environment), operational (runtime conditions), or data state (data context).
Covariate Conformance	The degree to which a test's covariate values match those of the baseline. Full conformance means all covariates match; non-conformance indicates the test ran under different conditions than the baseline was established.
Covariate Profile	An immutable record of covariate values captured at a specific point in time, used to characterize the conditions under which an experiment or test was executed.

Term	Definition
Criterion	The partition unit of the functional dimension of a service contract. Hosts one or more postconditions, declares a mode (inferential or observational), a denominator policy, the experiment or test in which it is exercised, and — where inferential — a threshold and confidence level. Applied to the run's sampling, each criterion yields its own per-trial Bernoulli stream and is reduced to its own verdict. See Statistical Companion §1.4.2.
Denominator Policy	A criterion's declaration of how unevaluable trials (malformed output, timeout, refusal, missing required material) are counted. Values: `CONDITIONAL_ON_EVALUABLE` (denominator excludes unevaluable trials) or `MARGINAL_COUNT_UNEVALUABLE_AS_FAIL` (denominator includes them, counting each as a failure). Two criteria over the same postcondition set under different policies estimate genuinely different quantities. See Statistical Companion §1.4.5a.
Duration Constraint	A timing requirement in a service contract specifying the maximum allowed execution duration. Evaluated as an independent aspect of conformance alongside postconditions and expected-output matching.
Effect Size	The minimum degradation the test is designed to detect (same as minimum detectable effect).
Empirical Baseline	Machine-generated record of observed behavior from a measure experiment.
Empirical Clause (deprecated)	Former name for **Rate-Bounded Clause**; the word empirical is now reserved for the threshold **origin** (see Threshold Origin, Form vs Origin). See **Rate-Bounded Clause**.
Empirical Percentile	A quantile estimated directly from observed data without assuming a parametric distribution. In the mavai methodology, latency percentiles are computed using the nearest-rank method from the order statistics of successful-response latencies. See also: Nearest-Rank Method, Order Statistic.
Expected-Output Match	Comparison of a use case's actual response against a known expected result using a configurable matcher strategy (exact, case-insensitive, JSON-structural, or custom). The strongest form of correctness check: not ”is the response valid?” but ”is the response this specific value?” One independent aspect of conformance.
Experiment	A controlled execution of a use case to characterise system behaviour. Three modes exist: explore (compare configurations), measure (establish baselines), and optimize (tune parameters).

Term	Definition
Experiment Configuration	One concrete combination of factor values — the unit of execution in an explore experiment.
Experiment Design	Declarative description of what is explored (factors + levels).
Explore Experiment	Experiment mode that compares multiple configurations with fewer samples each.
Factor	One independently varied dimension in an explore experiment (e.g., model, temperature).
Factor Level	Design-of-Experiments term for one setting of a factor. **Not canonical mavai terminology** — kept here only as a nod to the DoE roots of the methodology. In mavai prose, code, and documentation, use ”factor value” instead. A Factor has values; an Experiment Configuration is one combination of factor values.
False Negative (Type II Error)	A test pass when the system has degraded.
False Positive (Type I Error)	A test failure when the system has not degraded.
Feasibility Gate	Pre-execution check that determines whether the configured sample size can in principle support a positive verdict at the required confidence level. If infeasible, the test is rejected before any samples are collected.
Footprint	A stable hash identifying the combination of use case identity, functional parameters, and covariate declarations. Two baselines with the same footprint are candidates for covariate-based selection.
Form vs Origin	Two orthogonal axes describing a clause. **Form** — rate-bounded vs categorical — is what kind of proposition the clause states. **Origin** — normative vs empirical — is where a rate-bounded threshold came from. The word empirical names a value on the origin axis only; it is not a form. See Statistical Companion §”Clause Forms”.
Indicative Result	A latency percentile computed from fewer samples than the minimum recommended for that percentile level. The result is reported but explicitly qualified as a directional signal, not a statistically reliable estimate. Analogous to the smoke/verification asymmetry for pass-rate testing.
Inferential Mode	Criterion mode that estimates the true rate of the criterion via Wilson construction and decides against a threshold p_c^* at confidence α_c. Produces a three-valued verdict (PASS, FAIL, or INCONCLUSIVE). Appropriate for criteria whose contractual question is ”what is the true rate of behaviour, with what confidence, and does it clear the demanded threshold?” See Statistical Companion §1.4.5.

Term	Definition
Input Source	The population from which per-sample inputs are drawn, cycled round-robin across samples.
Instance Conformance	Expected-output matching against a golden dataset of known-correct results, using a configurable matcher strategy (exact, case-insensitive, JSON-structural, or custom). One aspect of overall conformance. See: Expected-Output Match.
Latency Assertion	A set of percentile constraints $\{(p, \tau)\}$ specifying that the empirical percentile $Q(p)$ of successful-response latencies must not exceed threshold τ milliseconds. The overall latency assertion passes if and only if all individual constraints pass.
Latency Enforcement Mode	Controls whether a latency assertion breach fails the test (enforced) or produces a warning (advisory). Advisory is the default because baselines may have been recorded on different hardware. Enforced mode is appropriate for environments with controlled hardware consistency.
Latency Population	The conditional distribution of execution times given success: $T \mid X = 1$. Only successful samples contribute to latency analysis. Failed samples (fast rejections, timeouts) are excluded because their execution times are not comparable with successful-response latencies.
Latency Threshold Derivation	The process of computing a latency threshold from a baseline percentile as the exact binomial order-statistic upper confidence bound: $\tau = t_{(k)}$ where $k = \text{qbinom}(1-\alpha, n\_s, p) + 1$, clamped to $[\lceil p \cdot n\_s \rceil, n\_s]$. Distribution-free and exact for i.i.d. samples from a continuous latency distribution. The non-parametric counterpart of the Wilson lower bound used for pass-rate thresholds.
Measure Experiment	Experiment mode for precise estimation of one configuration with many samples, establishing an empirical baseline.
Minimum Detectable Effect	Smallest drop from baseline worth detecting; required for the confidence-first approach to compute sample size.
Minimum Pass Rate	The threshold pass rate the system must meet to pass the test. Part of the parameter triangle.
Minimum Sample Size (Latency)	The minimum number of successful samples required for a percentile estimate to be non-degenerate. Values: p50 → 5, p90 → 10, p95 → 20, p99 → 100. Below these thresholds the percentile collapses to the maximum (or minimum) of the sample and provides no distributional information.

Term	Definition
Nearest-Rank Method	Percentile computation method where the index is ⌈p·n⌉ − 1 (clamped to [0, n−1]). Produces integer-valued percentile estimates from the order statistics, aligning naturally with latency thresholds expressed in integer milliseconds. Preferred over linear interpolation methods (e.g., R's Type 7) for this reason.
Non-Parametric	A statistical method that makes no assumptions about the shape of the underlying distribution. The mavai latency analysis is entirely non-parametric: percentiles are estimated directly from order statistics rather than by fitting a parametric model (e.g., normal, log-normal) to the data.
Normative Threshold	A threshold origin that represents a requirement (SLA, SLO, or policy). With verification intent, insufficient sample sizes cause the test to be rejected before execution; with smoke intent, the test runs but verdicts include an explicit caveat.
Observational Mode	Criterion mode that reports whether any failure of the criterion was observed in the run. No population estimation, no confidence interval, no threshold parameter. Appropriate for criteria where any failure is consequential — typically the observational zero-failures criterion that provides evidence for a categorical clause discharged architecturally. See Statistical Companion §1.4.5.
One-Sided Lower Bound	Statistical threshold below which the true success rate is unlikely to fall at a given confidence level.
Optimize Experiment	Experiment mode that iteratively tunes a single factor to find the optimal value.
Order Statistic	The k-th smallest value in a sorted sample. The notation $t_{(k)}$ denotes the k-th order statistic. Empirical percentiles are specific order statistics selected by the nearest-rank index formula.
Parameter Triangle	The three interdependent parameters in probabilistic testing: samples, confidence, and minimum pass rate. You control two; statistics determines the third.
Per-criterion Bernoulli Stream	The sequence of per-trial pass/fail indicators that a criterion yields when applied to the sampling. Modelled as i.i.d. Bernoulli with parameter p_c under the methodology's working approximation; the substrate of all per-criterion inference. See Statistical Companion §1.4.3.

Term	Definition
Per-Sample Outcome vs Run-Level Verdict	A criterion's outcome on a **single sample** is two-valued — PASS or FAIL (a FAIL carrying a condition or transform/no-value reason). Only the **run-level verdict**, aggregated over the whole sampling, is three-valued and may be **INCONCLUSIVE** — when the run cannot support a determination (sample below the feasibility minimum, or an empirical-origin threshold with no baseline rate for the covariates in force). INCONCLUSIVE is a statement about the run, never about any one sample. See Statistical Companion §1.4.5, §1.4.5a.
Percentile Constraint	A single (percentile level, threshold) pair within a latency assertion — e.g., (p95, 500ms). The constraint passes when the empirical percentile at that level does not exceed the threshold.
Population Claim	A criterion's declaration of which population its rate p_c generalises to, recorded on the verdict as the `populationClaim` field. Three values: **finite-corpus** (the rate is a closed-form statement about a fixed enumerated corpus — no binomial confidence interval applies); **superpopulation** (the rate generalises to a named external population, e.g. production traffic — generalisability is argued, not validated, by the operator); **no-generalisation** (the rate describes the run only). A mismatch between baseline and test on this field is a §8.4.5 hard invalidator. See Statistical Companion §8.4.6.
Postcondition	A named predicate over a single trial's output, deciding pass or fail for one observable property. A postcondition carries no threshold and no statistical configuration of its own — it belongs to a criterion (which hosts one or more postconditions), and the criterion supplies the threshold, confidence level, denominator policy, and mode under which the postcondition's per-trial verdicts are aggregated. See Statistical Companion §1.4.2.
Power	Probability of catching a real degradation; equals 1 minus the false negative rate.
Provenance	The chain of artefacts from definition to enforcement, ensuring every verdict can be traced to its empirical foundation.

Term	Definition
Rate-Bounded Clause	A service-contract clause stating a rate-bounded proposition: the criterion's true pass rate p_c must clear a threshold p_c^* at confidence $1 - \alpha_c$. Discharged statistically by the Wilson construction and integer-pass-cutoff machinery (§3, §3.4). Its threshold's **origin** — normative or empirical — is a separate axis (see Threshold Origin, Form vs Origin). Not promotable to a categorical clause by letting $p_c^* \to 1$. See Statistical Companion §"Clause Forms".
Regression Threshold	Statistically-derived minimum pass rate for regression tests, computed from baseline data using the Wilson lower bound.
Right-Skewed Distribution	A distribution with a long right tail, where the mean exceeds the median. Service latency distributions are typically right-skewed due to cache misses, garbage collection pauses, and cold starts. This is the primary reason the mavai methodology uses non-parametric percentiles rather than parametric models.
Sample	A single execution of the system under test. Distinct from **Sampling**, which is the list of $N \geq 1$ samples comprising one experiment or test. Methodology prose (Statistical Companion, Statistical Model Overview) calls this a **trial** in the Bernoulli-trial sense; trial and sample are synonyms in mavai's vocabulary, partitioned by register: methodology prose uses trial, implementation code and per-project docs use sample. See also **Trial**.
Sample Size	Number of test executions; controls cost and time. Part of the parameter triangle.
Sample-Size-First Approach	User specifies samples + confidence; framework computes achievable threshold.
Sampling	The list of $N \geq 1$ sample inputs posted to the service in a single experiment or test. Each sample entry is presented once, producing N responses. Shared by every criterion of the run: a contract with multiple criteria produces a per-trial vector of per-criterion observations over a single sampling, not one sampling per criterion. See Statistical Companion §1.4.2.
Sentinel	A lightweight runtime agent that evaluates the distributional contract continuously against a live system, detecting degradation in situ rather than after the fact.

Term	Definition
Service Contract	A formal definition of what ”success” means for a use case. Decomposes into criteria, each of which hosts one or more postconditions and declares the mode (inferential or observational), threshold, confidence level, and denominator policy under which its per-trial outcomes are aggregated. Clauses take one of two **forms**: **rate-bounded** (statistically evaluated) or **categorical** (an obligation or prohibition, discharged architecturally). Used by both experiments (to measure behaviour) and probabilistic tests (to enforce correctness).
SLA (Service Level Agreement)	Contractual commitment to external customers defining minimum service quality.
SLO (Service Level Objective)	Internal target for service quality, often more stringent than SLAs.
Smoke (Intent)	Test-intent value declaring a sentinel posture: undersized configurations are admitted but verdicts are caveated. Smoke verdicts run as directional signals rather than evidential claims; they do not satisfy normative-threshold assurance requirements. See **Test Intent**; Statistical Companion.
Standard Error	The standard deviation of a statistic's sampling distribution, measuring the precision of the estimate. Used in pass-rate inference (Wilson construction); the latency dimension uses non-parametric order-statistic methods rather than parametric standard errors.
Statistical Power	The probability of correctly detecting a real degradation ($1 - \beta$); same as power.
Stipulated Threshold	Synonym for **Normative Threshold**; normative is the primary term. See **Normative Threshold**, **Threshold Origin**.
Test Intent	Umbrella declaration of the developer's posture for a probabilistic test; two values, **Verification (Intent)** and **Smoke (Intent)**. Verification is the evidential posture (normative-threshold assurance, feasibility-gated). Smoke is the sentinel posture (directional signal, undersized configurations admitted with caveat). See those entries.
Threshold-First Approach	User specifies samples + threshold; framework computes implied confidence.
Threshold Origin	Classification indicating the origin of a probabilistic test's threshold (e.g., SLA, SLO, policy, empirical). Origins marked as normative trigger stricter enforcement rules for verification intent.
Threshold Provenance	Metadata documenting where a test's threshold originated and the conditions under which it was established.

Term	Definition
Trial	Methodology-register synonym for **Sample** (Bernoulli trial). Used in the Statistical Companion and the Statistical Model Overview, in keeping with the statistical literature. The frameworks (punit, feotest) use sample in code and per-project documentation. See **Sample**.
Upper Confidence Bound	A one-sided bound above which the true value is unlikely to lie at a given confidence level. Used in latency threshold derivation to set a conservative ceiling on baseline percentiles, absorbing ordinary sampling variation. The latency counterpart to the Wilson lower bound used for pass-rate thresholds.
Use Case	An artefact defining a service operation and its service contract. It ensures that experiments and tests refer to the same expression of correctness. In implementation, a use case invokes production code and returns an outcome.
Use Case ID	A unique string identifier for a use case (e.g., `shopping.product.search`).
Verification (Intent)	Test-intent value declaring an evidential posture: the configuration must be statistically feasible to deliver the affirmative claim demanded by a normative threshold, or the test is rejected by the feasibility gate before execution. Verification verdicts carry the procedure's full error semantics; they are the form of verdict admissible for compliance against SLA, SLO, or policy thresholds. See **Test Intent**; Statistical Companion.
Verdict	The outcome of a per-criterion or composite evaluation, three-valued: **PASS**, **FAIL**, or **INCONCLUSIVE**. INCONCLUSIVE arises when the criterion's gate (e.g. integer pass cutoff at the configured confidence level) cannot be satisfied as PASS or FAIL by the observed evidence — typically too few trials, or an unevaluable run under a denominator policy that disallows scoring. Contract composite verdicts are assembled from per-criterion verdicts under the rules of §1.4.6. See Statistical Companion §1.4.5, §1.4.6.
Wilson Score Bound	Robust confidence bound for binomial proportions. Used exclusively throughout the mavai methodology for all confidence interval and threshold derivation calculations. See the Statistical Companion for the full treatment.

Term	Definition
Zero-Failures Clause	The single form in which a categorical clause is recognised: the developer declares the expectation while explicitly abandoning any statistical lower bound. Evaluated observationally — PASS if no failure is observed, FAIL on any, INCONCLUSIVE only if there were no trials. Authored as `zeroFailures()` (punit) / `zero_failures()` (feotest). See **Categorical Clause**, **Observational Mode**.