The Precision Calibration Pipeline

Specification-Faithful, Multiple-Testing-Aware Strategy Calibration
with Independent Re-Derivation

Haneef R. Haqq, MBA[1]

[1]Quant7 Alpha, LLC

Technical Whitepaper · v1.0

June 30, 2026

Abstract

We describe the calibration pipeline used by Q7 AEGIS AI to take a systematic trading strategy from a single canonical specification to a signed, deployable configuration set. The pipeline is organized around one premise: the dominant failure mode in systematic trading is not a weak signal but a *specification–implementation divergence*—the calibrated and executed system silently differs from its design—and this divergence must be eliminated *before* optimization, not diagnosed after deployment. We formalize the integrity preconditions, the optimization objective, the human–AI gating ensemble, the out-of-sample protocol (purged combinatorial cross-validation with embargo, evaluated under a deflated performance statistic), and the preventive risk construction. We then describe the producer/verifier separation that makes every shipped result independently reproducible. The pipeline issues residual-risk-qualified verdicts; it never certifies optimality, and it states plainly what it does not address.

1. The Failure We Engineer Against

Let `S` denote a strategy specification (the canonical charting source), `C` its calibration implementation, and `E` its execution implementation. Three pathologies dominate live underperformance, and conventional pipelines detect none of them with high power.

Functional divergence (`C ≠ S`). The calibration backtester implements a strict subset of the specification's logic—a reduced exit ladder, an absent stop-governance layer, an engine whose entry condition was replaced by a proxy. The optimizer then converges on the global optimum of the *wrong* objective surface. Because `E ≈ S`, the deployed system is out-of-distribution relative to everything that was calibrated: train≠serve at the structural level.

Non-convergence under structural defect. A defective `C` does not converge within a sane evaluation budget; it either exhausts the budget without shipping or returns an overfit artifact. Additional compute is a symptom-level response. The relevant prior: a strategy in this program finalized as a hard block after 1,478 trials because engine×direction attribution was mis-wired upstream—a Stage-0 defect masquerading as a Stage-2 budget problem.

Presence-checking masquerading as fidelity-checking. A certification that verifies parameters are *declared and consumed* will pass an implementation that is a stripped shadow of its specification, because subset-implementation preserves the input signature. Structural-presence GREEN is necessary but not sufficient for formula faithfulness.

The pipeline is constructed so that these are caught with fail-closed gates prior to trial #1.

2. Operating Axioms

Specification authority. `S` is the sole source of truth. The admissible calibration set is exactly the declared input set of `S`, less a signed exclusion list; derived artifacts carry no authority and cannot expand the search space.

Fail-closed evaluation. A fail, an unknown, or a missing artifact is terminal for the dependent check. Absence of evidence is never treated as evidence of correctness.

No silent narrowing. Full calibratable coverage, single mode, no top-k / sampling / freezing. Per-configuration evaluation is mandatory; cross-configuration uniformity is a shortcut signature, not a calibration result.

Producer ≠ verifier. The agent that generates a result does not grade it. Verification is performed under separate credentials and is externally auditable.

Honest-empty. Unevaluable checks are reported as such with a named missing artifact; no value, attribution, or metric is inferred or fabricated to clear a gate.

3. Stage 0: Structural-Integrity Cornerstone

Before any search, an agent establishes `C ≡ S ≡ E` at the formula level across the tri-implementation. The load-bearing checks are: lint and compile of the canonical source; *entry/engine liveness* (every engine emits signals end-to-end, not merely declares them); *searched-dimension consumption* (a

searched-but-unconsumed dimension is a dead dimension and a fail—it inflates the multiple-testing burden while contributing nothing); *engine×direction attribution* (the single most defect-prone wire); and—decisively—*formula-footprint parity*, in which exit-ladder cardinality, stop-governance layers, the sizing model, and gating logic are compared function-by-function against S, not by input-signature coverage. A runtime ship-smoke confirms the model imports, trades, and ships at least one live engine end-to-end. Any fail or unknown renders the pipeline RED; all downstream calibration is advisory until the cornerstone is GREEN. The economic argument is direct: a structurally clean model converges within the evaluation budget, so the cheapest path to a trustworthy result is to pay the structural cost once, up front.

4. Stages 1–2: Population, Data, and Concerted Search

Population & data. The calibratable population is enumerated directly from S (per configuration), minus signed exclusions. The market-data snapshot is content-addressed and frozen; every downstream result is reproducible against a known hash, and any contaminated or stale snapshot forces re-certification rather than reuse.

Search. Every admissible dimension is optimized jointly. High dimensionality is addressed with a sequential model-based optimizer (Bayesian / tree-structured Parzen estimators) under a *return-within-budget* objective—a risk-adjusted functional that rewards utilization of the drawdown budget rather than minimization of variance, which prevents the degenerate collapse to near-zero exposure a naive ratio objective induces. The total evaluation count is capped ($\leq$ 1000 per configuration); the cap is a convergence discipline, not a sampling shortcut. Crucially, the evaluation budget is fixed and disclosed, because it is an input to the multiple-testing correction applied downstream.

5. Stage 3: The HAI Tandem (Train = Serve)

Calibration is performed against the gated system that will actually be deployed. The Honest-AI overlay is a three-model ensemble evaluated per bar: a gradient-boosted *directional* classifier, a transformer-based *sentiment* estimator, and a *market-regime* classifier. Each is trained in-cycle, per configuration, and the objective scores the full HAI-gated equity curve—not the raw-signal curve. A neutral-stubbed gate (trained but not consulted, or consulted ML-only with sentiment/regime stubbed) is a train≠serve violation and is forbidden. The gate is admitted only if it passes health constraints—an adversarial discriminability bound and per-direction discriminability floors—and the model hash that gates live trades must equal the hash calibrated in-cycle. A configuration without a fresh, health-passing, hash-matched model is blocked, never traded gate-off.

6. Stages 4–5: Robustness and the Honest Charter Gate

Out-of-sample protocol. Robustness is assessed on the HAI-gated system using combinatorial purged cross-validation with embargo: folds are constructed combinatorially, observations whose labels overlap the test window are *purged*, and a temporal *embargo* removes post-test observations that could leak through serial correlation. This yields a distribution of out-of-sample paths; the lower confidence bound of that distribution—not the point estimate—is what gates.

Multiple-testing correction. Because the search evaluates a large, disclosed number of configurations, the in-sample optimum is upward-biased by selection. We therefore evaluate against a *deflated* performance statistic: the observed Sharpe ratio is discounted by the expected maximum attainable under the realized number of effectively independent trials and the higher-moment structure of the returns. A configuration clears Stage 5 only on the deflated statistic, under fixed (not co-optimized) risk constraints, against a binding five-criteria charter gate—the bar designed to reject artifacts of search rather than evidence of edge.

7. Risk as a Structural Constraint

Risk is enforced preventively at the sizing layer rather than reactively at a kill-switch, because an overnight or weekend gap can render a reactive stop too late. Let P_t be the monotonically ratcheting all-time equity high (never reset), Q_t current equity, D the hard drawdown ceiling, ℓ_u the worst-case adverse loss per unit (protective stop plus a per-instrument gap reserve estimated from the instrument's own history), and n the position size. Sizing solves

$$n = \min\left(n_{pre},\ \left\lfloor \left(Q_t - P_t \cdot (1 - D)\right) / \ell_u \right\rfloor \right),$$

blocking entry when $Q_t \leq P_t \cdot (1-D)$, where n_{pre} is the pre-risk size from the specification's own sizing model. The construction guarantees that even a worst-case-in-data gap-through keeps realized drawdown, measured from the true peak, below D—by sizing, not by reaction. D is a structural constant, implemented identically in S, C, and E; it is never a calibrated dimension. A reactive intrabar control remains as a backstop. The design target is capital efficiency for small accounts: the budget is utilized, not merely respected.

8. Stages 6–7: Independent Re-Derivation and Governance

Sign-off. Packaging produces a content-hashed dossier (configuration parameters, model hashes, data-snapshot hash,

compliance heat map). A configuration is marked ready only after an independent re-derivation on held-out partitions reproduces the result within tolerance, and a separate validator signs off on train=serve integrity—both under credentials distinct from the producer's. Promotion is gated on a signed certificate (GREEN and hash-matched), never on a claimed pass.

Governance. Seven chartered agents own the stages (Calibrator, Trainer, Validator, Sentinel, Researcher, Auditor, Adjudicator). Producer/verifier independence is enforced by per-agent key isolation, making the separation cryptographic and externally observable rather than procedural. The oversight role assigns lanes and adjudicates; it does not hand-run lanes or override gates. Binding rules are mechanized as fail-closed gates rather than restated, because a control that depends on recall is not a control.

9. Verdicts and Residual Risk

The pipeline emits three verdicts and never asserts optimality. **GREEN (residual risk: …)**—cornerstone clean, charter gate met on the deflated statistic, residual risks enumerated. **AMBER (conditional on: …)**—ships only after named conditions clear. **RED (blocked by: …)**—structural defect, unverifiable claim, compliance flag, or unresolved fail. Every GREEN carries an explicit residual-risk statement—the failure modes that survive a passing grade—and the live evidence required to keep the grade valid.

10. Limitations

Intellectual honesty requires stating the boundary of the claims. **Non-stationarity and regime change:** purged CPCV and a deflated statistic bound in-sample selection bias; they do not immunize against structural breaks in the data-generating process after deployment—out-of-sample here means out-of-sample within the certified history. **Gap-reserve estimation risk:** the preventive sizing bounds drawdown against the worst gap observed in the instrument's history; a future gap exceeding the historical envelope is a tail the construction does not cover. **Model and data risk:** train=serve and health gates constrain the HAI overlay; they do not certify that the learned relationships are causal or durable. **Execution, leverage, liquidity, and venue/regulatory risk** are not eliminated by calibration integrity and remain in every residual-risk statement.

The pipeline's contribution is precise and bounded: it makes specification–implementation divergence and selection-biased validation structurally hard to ship, and it makes every shipped result independently reproducible. It does not—and does not claim to—predict the future.
