Treasury PoC V1.5 - Executive Summary

Contents

Treasury PoC V1.5 — Executive Summary

Forward Feasibility Test of Text-Aware Regime-Adjusted MMAR for 10-Day VaR

Score date: April 27, 2026. Project window: April 28 – May 11, 2026 (10 trading days). Position: $30M synthetic UST book ($10M each in 2Y, 5Y, 10Y). Discipline: Bar 2 forward (single LLM, single pass, frozen corpus).

Companion document: Detailed Findings (full document with sensitivity tables, per-dimension walkthrough, methodology defenses, three-engine arithmetic, and Open Methodological Questions appendix). Code: 01_baselines.py, 02_engines.py, 03_charts.py.

At a Glance

The PoC. Treasury PoC V1.5 runs the Risk Intelligence framework forward on a real position over a real two-week window. A $30M synthetic UST book, allocated $10M each across 2Y, 5Y, and 10Y, is held through a 10-trading-day window opening Tuesday April 28, 2026 — a window that contains a live FOMC meeting on its first day. The score is frozen on the morning of Monday April 27 against a corpus closing at Friday April 24. Bar 2 forward discipline applies: single LLM, single pass, no internet during scoring, frozen corpus. All numeric coefficients are V1 illustrative; the property under test is whether the architecture runs end-to-end, not whether the coefficients are calibrated.

The Methodology. The framework reads textual regime evidence — FOMC statements, minutes, central-bank speeches, IMF assessments, market commentary — and emits a structured six-dimension regime signal. A six-anchor severity rubric tied to dated historical regimes (March 2020, August 2020, September 2024, December 2018, March 2022, January 2024) makes the score replicable across observers. A documented conflict-resolution rule (voting > non-voting; Chair > governors; scripted > unscripted; recent > earlier) handles divergent voices. The signal flows through a deterministic, hand-recomputable mapping into three engine inputs: a Hurst-exponent adjustment (ΔH), a volatility multiplier, and a tail multiplier. Three engines compare: Engine 1 (Gaussian baseline), Engine 2 (static MMAR with historical Hurst), Engine 3 (Nexus-adjusted MMAR with text-aware regime adjustment). V1.5 is the feasibility bar (Bar 2 forward); the production target (Bar 3) requires multi-LLM ensemble scoring and cross-LLM robustness checks that are V2 work.

The Conclusion.

For Quant Teams. The +38.5% E1→E3 lift in 10-day VaR_99 decomposes cleanly into +19.8% pure MMAR self-similarity (text-blind) and +15.7% text-aware regime channel. Channel decomposition is auditable; the deterministic-mapping arithmetic checks to two decimals. Severity-adjacency analysis

shows non-monotonicity under anchor-character-consistent dimensions — a feature of the architecture, not a bug, demonstrating that dimensional structure dominates headline severity in driving engine output. Forward backtest framework lands May 11, 2026; single-window evaluation is feasibility, not calibration. **The May 11 backtest closed with realized 10-day loss of $114,339 — no engine breached at either 95% or 99% confidence level. This is Outcome A (modal expected outcome in a calm regime), establishing face validity of the V1.5 publication while providing essentially zero statistical power on long-term calibration accuracy — V2 multi-window calibration is the test that adjudicates calibration.**

For AI Teams. Bar 2 forward discipline isolates the calibration question from the comprehension question by fixing the score before any realized data exists. The four-question reasoning decomposition (text reading / signal coherence / deterministic mapping / forecast vs ground truth) gives an auditable evaluation substrate where each question is addressable by a different methodology. Single-LLM single-pass replicability bar is ±0.5 anchor units; cross-LLM robustness is V2 priority. The architecture's commitment to a deterministic, hand-recomputable mapping is a counter to "LLMs are black boxes": the audit trail is complete from the structured signal forward to engine inputs, and any reader can reproduce the engine input from the score in five minutes with a calculator. The signal-emission step itself remains bounded by Bar 2 forward discipline and the replicability bar — the LLM's qualitative judgment is constrained, not eliminated.

Position, Scoring, and Three-Engine Comparison

Position and Window

A synthetic $30 million long UST position, allocated $10M each across 2Y, 5Y, and 10Y constant-maturity points. DV01 per $10MM notional: 2Y = $1,909; 5Y = $4,501; 10Y = $8,055. Total book DV01 for a parallel curve shift = $14,465 per basis point. The score is produced on Monday April 27, 2026; the project window is the ten trading days from Tuesday April 28 through Monday May 11. The April 28–29 FOMC meeting falls inside the project window — the V1.5 forecast is therefore through-and-beyond a live FOMC decision, not a calmer interval.

Bar 2 Forward Scoring

The score has two structural elements: a six-anchor severity rubric and a six-dimension regime signal. The rubric locks current and historical regimes to specific dates so that scoring is a comparison act, not an absolute-magnitude estimate. The six dimensions are: D1 severity (anchored to rubric), D2

persistence (transient vs structural), D3 policy direction bias, D4 inflation-expectations anchoring, D5 dispersion across sources, and D6 market functioning. A documented conflict-resolution rule (voting > non-voting; Chair > governors; scripted > unscripted; recent > earlier) handles divergent voices.

Scored values: D1 = 7.0; D2 = 0.65; D3 = −0.20; D4 = 0.75; D5 = 0.55; D6 = 0.78. Confidence band ±0.5 on D1.

The deterministic mapping converts these into three engine inputs:

```
ΔH = (severity_norm − 0.5) × persistence_signed × (1 + dispersion) × α
   = 0.20 × 0.30 × 1.55 × 0.25 = 0.0233 (under cap 0.10)
vol_multiplier = 1 + (1 − anchoring) × β = 1.050
tail_multiplier = 1 + (1 − market_funct) × γ = 1.044
```

V1 illustrative coefficients α / β / γ = 0.25 / 0.20 / 0.20. The mapping is hand-recomputable in five minutes with a calculator.

Three-Engine Headline Results (σ_90d / H_24m)

Engine	VaR 95%	VaR 99%	Lift (99%)
E1 Gaussian	$307,162	$434,424	(baseline)
E2 Static MMAR	$367,842	$520,246	+19.8% vs E1
E3 Nexus-adj MMAR	$425,452	$601,725	+15.7% vs E2; +38.5% vs E1

The +38.5% E1→E3 lift decomposes cleanly into two architectural channels. Channel 1 (E1→E2, +19.8%) is pure MMAR self-similarity — text-blind, present even if the LLM step were skipped. With H_24m ≈ 0.58, t^H scaling widens the variance window roughly 20% beyond sqrt-t. Channel 2 (E2→E3, +15.7%) is the text-aware Nexus channel: ΔH contributes ≈+5.5% to t^H, vol_mult contributes 1.050, tail_mult contributes 1.044 — multiplied: 1.157, matching the observed lift to two decimals. The deterministic-mapping arithmetic checks. The decomposition itself is the headline finding, because it isolates the architecture-under-test from the multifractal property of the price series.

E3 VaR_99 of $601,725 on the $30M book is 2.0% of notional over 10 trading days; under stress-weighted σ_30 the figure is 2.5%. Both sit in a plausible range for a long UST book in a stressed regime.

Sentiment vs Structured Scoring

The first response to any LLM-driven regime score is to ask whether it differs from sentiment analysis. Five differences distinguish the Nexus scorer from sentiment:

1. **Sentiment analysis is unstructured by construction** — it produces a scalar (typically [−1, +1] or [0, 1]) and the engine cannot consume it without an additional, post-hoc bridge that is precisely what the deterministic mapping commits to avoiding.
2. **Sentiment analysis discards dispersion** — the Nexus D5 dispersion sub-score is explicit and is used by the engine; sentiment analysis averages over disagreement.
3. **Sentiment analysis is observer-dependent** — without an anchored rubric, sentiment scores drift; the Nexus anchored ordinal rubric ties scoring to dated historical regimes.
4. **Sentiment analysis has no conflict-resolution rule** — voting members and non-voting commentary are treated identically. The Nexus conflict-resolution rule (voting > non-voting; Chair > governors; scripted > unscripted; recent > earlier) makes the score replicable.
5. **Sentiment analysis is silent about its own uncertainty** — the Nexus score reports a confidence band (±0.5 anchor units) and a dispersion sub-score (D5) explicitly.

The first difference is structural — sentiment analysis cannot drive the engine because the engine needs structured inputs. The other four are governance properties that matter for institutional adoption.

Why Not Just Use Prices? (EMH Objection — Summary)

A second skeptical question, distinct from the sentiment-analysis question and arriving from both the trading and institutional sides of the audience, is structural: *if markets already incorporate the textual information the framework reads, the LLM signal is redundant with information already in the price series.* This is the Efficient-Market Hypothesis applied to LLM-emitted regime signals. V1.5 does not claim forecasting edge that survives semi-strong-form market efficiency. The framework's value is defended on six architectural grounds: (1) the framework is structurally a regime-confirmation framework, not detection; (2) the framework's plausible informational contribution is in consolidation and structuring of dispersed-source narratives, not in scooping prices on individual events; (3) the deterministic mapping is auditable in a way price-only models cannot be; (4) the framework's contribution is hypothesized to concentrate during regime transitions, a hypothesis V2 will test rather than a claim V1.5 demonstrates; (5) the empirical test of EMH-redundancy is pre-registered and falsifiable in V2 Q14; (6) Nexus is designed for institutional risk management, not alpha generation —

its value persists under both the latency/replication objection (markets already parse the corpus) and the Grossman-Stiglitz arbitraging argument. *See Detailed Findings Section 7B for the full five-response defense.*

Severity Adjacency: The Non-Monotonicity Finding

Under fixed dimensional profile (D2–D6 held constant), ΔH increases monotonically with severity — expected. Under anchor-character-consistent dimensional profiles, ΔH is not monotonic: the current regime (anchor 7) produces a higher ΔH than September 2024 (anchor 8), even though anchor 8 is structurally rated as more severe. The reason is that severity 8 was a structurally calmer regime — lower persistence, lower dispersion, more orderly markets, no parallel war. The Hurst-adjustment channel responds to dimensional structure, not to headline severity alone. This is a feature of the architecture, not a bug — it is in fact the central architectural claim being tested. A purely severity-headline scorer would mis-rank these regimes; the Nexus structured scorer ranks them by structural depth. The implication for risk officers: a 7-vs-8 disagreement among scorers is not necessarily a meaningful difference for engine output, because the dimensional structure dominates. This relocates substantive disagreement from "what number is severity?" to "is the regime persistent or transient?" — which is a question the dimensional structure forces the scorer to confront.

The dimensional profiles assigned to historical anchors are plausible reconstructions, not independently derived. The falsifying test for the architectural claim — rescoring each historical anchor against its contemporaneous corpus using the same methodology — is V2 priority 1 (Open Methodological Questions, Q11). If derived profiles diverge from the table values, the V1.5 architectural claim must be revised.

Forward Backtest Framework

The 10-trading-day project window closes Monday May 11, 2026. Realized 10-day P&L is computed from realized yield changes April 28 through May 11 multiplied by per-tenor DV01s, then compared to each engine's VaR_95 and VaR_99 thresholds. A single observation cannot determine which engine is best calibrated — expected breach rate at the 99% level is 1%, meaning one breach per hundred 10-day windows in expectation. Concluding from a single window that "Engine X was best calibrated" is statistically meaningless. What a single window can show is whether any engine produced a number that fails face validity. Multi-window calibration (12+ monthly non-overlapping windows; 100+ for meaningful power) is V2 work. V1.5 is a feasibility test, not a calibration test.

Reasoning Capability Decomposition (For AI Teams)

A persistent challenge in evaluating LLM-driven systems is that "did the LLM reason well?" is too coarse to be measurable. The Nexus architecture turns this question into four independently measurable sub-questions, each addressable by a different evaluation methodology.

- **Q1 — Did the LLM read the text correctly?** A comprehension property, measurable against the frozen corpus by checking whether the verbatim excerpts the LLM extracted are actually present and support the dimensional claims they are attached to.
- **Q2 — Did the LLM produce a coherent structured signal?** An internal-consistency property, measurable by checking whether the dimensional sub-scores cohere (e.g., a regime scored as expectations-de-anchoring should also show direction-toward-hike under most plausible scenarios).
- **Q3 — Did the deterministic mapping correctly translate the signal into engine inputs?** An arithmetic property, trivially measurable on a calculator because the mapping is hand-recomputable.
- **Q4 — Did the engine forecast align with the realized outcome?** A calibration property, measurable only after the project window closes, requiring multi-window observation to be statistically meaningful.

The four questions are independent in principle: an LLM can pass Q1, Q2, Q3 and fail Q4, or vice versa. Pass-fail patterns across the four diagnose which part of the architecture caused any failure — comprehension, coherence, implementation, or calibration. Each failure is repaired by different work. This is also why Bar 2 forward discipline matters: retrospective evaluation can confound Q1–Q3 with Q4 because the LLM may have seen the realized outcome during training. Forward discipline isolates Q4 by placing the score before any realized data exists.

Literature and Nexus Novelty Claims

Two recent works in the LLM-and-monetary-policy literature inform this PoC. Fernandez-Fuertes (October 2025, SSRN) studies LLM reading of central-bank communication for monetary-policy-shock identification, using LLM scoring as a regressor in a structural VAR. Soleimani (arxiv 2512.07867, "LLM-Generated Counterfactual Stress Scenarios for Portfolio Risk Simulation via Hybrid Prompt-RAG Pipeline," November 26, 2025) proposes a "transparent and fully auditable" LLM pipeline for macro-financial stress testing, with structured JSON outputs, deterministic run modes, hash-verified artifact manifests, and factor-based mapping to portfolio VaR/ES.

Soleimani's framework shares more with the Nexus architecture than the V1 framework paper acknowledged: both commit to deterministic post-processing of LLM output, both commit to auditability and reproducibility as design principles, and both use a regime-severity construct. The deterministic-mapping-with-auditability pattern is therefore converging industry practice as of late 2025, not a unique architectural commitment of Nexus. The structural differences are real but specific: Soleimani is a forward counterfactual scenario generator using PCA and polynomial factor channels for shock translation; Nexus is a current-state regime scorer using multifractal Hurst-exponent adjustment. The use cases are complementary rather than competing.

The novelty claims of the Nexus framework, revised against this updated literature reading:

1. **MMAR + LLM pairing.** No prior work, to our knowledge, channels LLM-derived regime signal into a multifractal volatility model. Forty years of multifractal finance literature has wrestled with how to parameterize regime structure for the multifractal scaling exponent; the LLM-emitted six-dimension structured signal is one answer. This remains the genuine architectural novelty.
2. **Auditable mapping coupled with multifractal substrate.** The deterministic, bounded mapping formalism on its own is converging industry practice (Soleimani 2025; concurrent works). What distinguishes Nexus is the pairing of the auditable mapping with the multifractal volatility substrate and with the anchored ordinal scoring. The combination is novel; the deterministic-mapping commitment alone is not.
3. **Anchored severity scoring with conflict-resolution rule.** Ordinal scoring against fixed historical anchors, combined with the documented voting/Chair/scripted/recent rule, makes the score replicable across observers — the property Bar 2 forward discipline depends on. Soleimani uses a regime-severity construct but not anchored to dated historical regimes; Fernandez-Fuertes uses LLM scoring but not against an ordinal rubric. The anchored-rubric-with-conflict-resolution pattern is novel to Nexus.
4. **AI training-signal thesis (qualified).** The four-question reasoning decomposition provides more diagnostic granularity than "did the LLM reason well?" — but it is not an end-to-end audit of the LLM. Q3 (deterministic mapping) is calculator-trivial on the framework's own formulas; the LLM's contribution to Q1 and Q2 (text reading, signal coherence) is bounded by Bar 2 forward discipline and the ±0.5 replicability bar but is not eliminated.

The Fractal Circuit Breaker — a higher-severity construct that engages when the regime score crosses a threshold and triggers structural engine-mode changes rather than continuous parameter adjustments — is a separate Nexus novelty claim addressed in the framework paper and in V2 work; V1.5 does not exercise it because the score (7.0) does not cross the breaker threshold.

Limitations (V1 Disclosures)

The coefficients α, β, γ are V1 illustrative; empirical calibration is V2 work. ΔH is term-structure-agnostic in V1; tenor-conditional ΔH is a V2 priority. The MMAR implementation uses Gaussian quantiles for the V1 run; multifractal copula and tenor-specific scaling factors are V2 enhancements. The corpus is bounded by what was publicly available by April 27, 2026. The conflict-resolution rule is rule-based, not learned. Single LLM, single pass — Bar 2 forward by definition; Bar 3 multi-LLM ensemble is the production target. **No alpha claim.** V1.5 does not claim its forecasts are more accurate than baselines; the single-window forward test cannot establish accuracy. The architectural claim is that the regime signal channel exists, is auditable, and produces materially different outputs from baseline — that claim the engine results support.

Events Inside the Project Window (Forward Application)

Two events broke inside the project window after the score was frozen on April 27. The V1.5 score was not, and will not be, updated — Bar 2 forward discipline holds the score fixed for the duration of the window. This section briefly documents the events and applies the framework methodology as if they had been in the original corpus, for transparency and framework demonstration.

Event 1 — Powell continuation as Federal Reserve Governor (signaled April 28). The DOJ closed its Powell investigation on April 25, handing matters to the Fed Inspector General. The Senate Banking Committee scheduled the Warsh confirmation vote for April 29. Powell signaled openness to remaining as Governor through 2028; market commentary framed this as “stabilizing counterweight” amid Warsh’s “regime change” agenda. Six-dimension impact: D1 +0.5, D2 +0.10, D5 +0.15 (twin-power configuration is the largest dispersion increment), D3 +0.05, D4 −0.05, D6 unchanged. If the event had been in the corpus, ΔH would have risen from 0.0233 to 0.0531; vol_mult to 1.060; channel-2 lift would have been ≈+24% rather than V1.5’s +15.7%.

Event 2 — UAE departure from OPEC and OPEC+ (announced April 28, effective May 1). The UAE ended sixty years of OPEC membership; WTI crude broke $100/bbl. Six-dimension impact: D4 −0.05, D6 −0.03, all other dimensions unchanged (no FOMC voting impact, structural OPEC shift but pricing-cycle for UST regime). If alone in corpus, channel-2 lift would have risen from +15.7% to ≈+17.4% — a smaller increment than the Powell event.

Combined inside-window scenario: if both events had been in the V1.5 corpus, E3 VaR_99 would have been ≈$655,000, vs V1.5’s $602,000 — a difference of roughly $53,000 on the $30M book. The May 11 backtest framework does not change: realized 10-day P&L will be compared against V1.5 thresholds, not the inside-window-aware figures here. See Detailed Findings Section 14 for the full six-dimension assessment with verbatim source attributions.

Realized Outcome — May 11, 2026 Close

The project window opened April 28, 2026 and closed May 11, 2026. The V1.5 regime score, frozen on April 27 against the corpus closing April 24, was not updated during the window. Bar 2 forward discipline held.

Realized yield changes (FRED DGS2, DGS5, DGS10; H.15 Selected Interest Rates). DGS2 moved from 3.84% to 3.95% (+11.0 bp); DGS5 from 3.97% to 4.07% (+10.0 bp); DGS10 from 4.36% to 4.42% (+6.0 bp). All three tenors moved in the same direction over the window — a small bear-steepening upward shift.

Realized 10-day P&L on the $30M synthetic UST book ($10M each tenor, long positions): −$114,339. Per-tenor contribution: 2Y −$20,999; 5Y −$45,010; 10Y −$48,330.

Breach matrix: No engine breached at either 95% or 99% confidence level. The realized loss represents 37% of the lowest published threshold (E1 VaR_95 = $307,162) and 19% of the highest (E3 VaR_99 = $601,725). Headroom to the headline E3 VaR_99 threshold is $487,386, approximately 4.3× the realized loss. The realized loss of $114,339 also falls far below the inside-window-aware hypothetical (≈$655,000 if Powell continuation and UAE/OPEC events had been in the corpus). In this window, the cost of Bar 2 forward discipline is undetectable in P&L terms.

What this outcome demonstrates and does not demonstrate. The framework executed cleanly end-to-end under Bar 2 forward constraints with zero post-hoc adjustments, fulfilling the narrow feasibility objective of V1.5. This single-window no-breach result provides evidence only of operational feasibility, process discipline, and adherence to pre-published thresholds; it supplies essentially zero statistical power on long-term calibration accuracy or predictive content, both of which remain V2 questions. This is the modal expected outcome under any reasonable VaR methodology in a calm regime. A single no-breach observation cannot distinguish between “well-calibrated framework,” “conservatively-calibrated framework,” and “framework that adds no information beyond baseline” — all three are consistent with this outcome. The outcome is consistent with the architectural claim of Section 7B that the framework’s value persists under approximate market efficiency: while Section 7B establishes the theoretical necessity of a narrative-aware risk layer for institutional risk management, the realized outcome demonstrates the first successful forward-run of that layer, providing the face validity required to move into the large-sample empirical calibration of V2.

Forward note. V2 multi-window calibration is the next empirical step. Q14 (added to the methodological appendix in this revision) formalizes the EMH-redundancy falsification test with Expected Shortfall as primary metric, named statistical tests (Diebold-Mariano, encompassing, likelihood-ratio), and a pre-specified quantitative falsification threshold. See Detailed Findings Section 14.4 and Methodological Questions Q14 for the complete analysis.

Looking Forward

V1.5 establishes feasibility. What comes next is calibration, robustness, and the broader question of what the architecture means beyond financial risk.

For Quant Teams: From Feasibility to Institutional Pilot. Three V2 priorities convert the V1.5 demonstration into a calibrated framework. First, coefficient calibration — replace the V1 illustrative values for α, β, γ with values derived from a defined procedure (historical-window minimum-deviation, Bayesian update on incoming windows, or cross-asset-class consistency). Second, multi-window forward testing — run the framework on a year of monthly non-overlapping windows so that breach-rate calibration becomes empirically grounded rather than asserted. Third, EMH-redundancy falsification (Q14) — pre-registered test of whether the regime channel adds information beyond price-derived Hurst, with Expected Shortfall as primary metric and a quantitative falsification threshold. These three V2 deliverables, paired with the cross-LLM robustness check, are the path from feasibility-demonstration to institutional pilot. The non-monotonicity finding is the architectural claim the V2 work will further validate or invalidate; the empirical test is dimensional-vs-severity scorer agreement on multiple corpora.

For AI Teams: Evaluation Substrate Beyond Financial Risk. The four-question reasoning decomposition (text reading / signal coherence / deterministic mapping / forecast vs ground truth) is generalizable beyond financial regimes. Any domain where an LLM emits a structured signal that drives downstream computation can use the same decomposition: the text-reading question is corpus-anchored; the coherence question is structure-internal; the mapping question is arithmetic; the forecast question requires forward observation. Bar 2 forward discipline — score before any realized data exists — is a template for evaluating LLM reasoning where calibration evidence requires the passage of time. The deterministic-mapping audit-trail property is a direct counter to the “LLMs are black boxes” criticism that blocks institutional adoption: in this architecture, the mapping is hand-recomputable in five minutes with a calculator, and the LLM’s contribution is bounded to producing the structured signal that feeds the mapping. The boundary between “LLM judgment” and “deterministic computation” is the institutional risk officer’s primary requirement; the architecture honors it.

The conclusion of V1.5 is narrow and the implication is broad. The narrow conclusion: the architecture works end-to-end, the +38.5% E1→E3 lift decomposes auditably into two channels, the single-window forward test demonstrates feasibility before any calibration claim, and the realized May 11 outcome held against published thresholds. The broad implication: LLMs as components in financial risk infrastructure are ready for institutional pilot, provided the architecture commits to the audit-trail property V1.5 demonstrates. The pairing of Mandelbrot’s multifractal volatility model with LLM-emitted regime signal closes a forty-year-old parameterization gap in MMAR while providing AI

evaluation methodology with a financial-risk substrate. Forty years of multifractal finance got stuck on "how do you parameterize the regime?" The LLM-emitted structured signal is one answer. V2 will tell whether it is the right answer.

Revision History

This document represents Treasury PoC V1.5 as launched. Prior to public release, the document underwent adversarial review by two independent LLM evaluators (Gemini and Grok), prompted as skeptical institutional reviewers. Their feedback surfaced three substantive issues which this document incorporates: (1) the literature and novelty section is narrowed to acknowledge that auditable deterministic-mapping-with-LLM-pipelines is converging industry practice — most directly evidenced by Soleimani (arxiv 2512.07867, November 2025); the deterministic-mapping commitment alone is no longer claimed as novel. (2) The non-monotonicity finding adds a falsification-test caveat acknowledging that historical anchor dimensional profiles are plausible reconstructions rather than independently derived. (3) Three new methodological questions (Q11 independent derivation; Q12 Hurst window robustness; Q13 channel orthogonality ablation) are added to the appendix; the V2 priority ranking is reshuffled. The Events Inside the Project Window section is added as forward-application content for transparency. We thank Gemini and Grok (independently prompted as adversarial reviewers) for surfacing the issues this revision addresses.

Addendum — May 12, 2026. This final revision incorporates three additions made after the original V1.5 launch and before public closure of the project window: (a) the "Why Not Just Use Prices?" EMH defense (summarized above; see Detailed Findings Section 7B for the full six-response treatment); (b) Q14 to the Open Methodological Questions appendix, formalizing the EMH-redundancy falsification test that V2 multi-window calibration must run; and (c) the Realized Outcome section above, documenting the May 11 backtest closure with the realized loss of $114,339 falling below all six published VaR thresholds (Outcome A — modal expected outcome).

Prior to public deployment, the consolidated document was submitted to Gemini and Grok for a second-round adversarial review. Their feedback converged on three substantive areas which this final pre-deployment revision incorporates: (i) the Realized Outcome section was tightened with an explicit statistical-power disclaimer, an explicit process-rigor affirmation, and a bridging sentence to the EMH defense; (ii) the EMH defense Response 4 (transition-period concentration) was reframed as an architectural hypothesis to be V2-tested rather than a claimed empirical property, and a new Response 6 was added covering the latency/replication objection and the Grossman-Stiglitz information paradox; (iii) Q14 test design was tightened with Expected Shortfall (the quantile loss function at 99%) committed as the primary evaluation metric, three named statistical tests pre-specified (Diebold-Mariano, encompassing, likelihood-ratio), a stratified-sampling window protocol, and a quantitative

falsification threshold. All feedback-driven changes remain textual; no architectural changes, coefficient revisions, or V1.5 numerical commitments are modified. We again thank Gemini and Grok for the second-round adversarial discipline.

End of Executive Summary. For full methodology, per-dimension walkthrough with verbatim Fed quotes, sensitivity tables, complete engine arithmetic, the full Section 14 inside-window analysis with verbatim source attributions, the full Section 7B EMH defense, the full Section 14.4 realized-outcome analysis, and the full Open Methodological Questions appendix, see the companion Detailed Findings document.