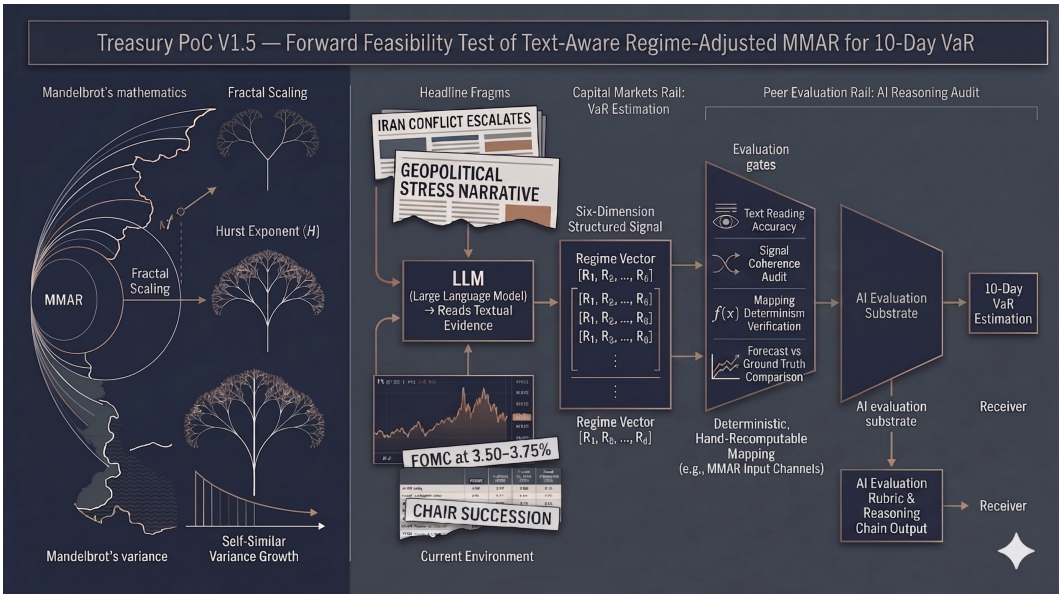
Treasury PoC V1.5 - Detailed Findings

Contents

Treasury PoC V1.5

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)



Treasury PoC V1.5 — Forward Feasibility Test of Text-Aware Regime-Adjusted MMAR for 10-Day VaR

Companion documents: Open Methodological Questions (separate); Executive Summary (separate); corpus_manifest.json; regime_signal.json; sensitivity tables; engine_results.json; baselines.json. 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.*
>
> ***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.*

Treasury PoC V1.5 — Detailed Findings

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) Discipline: Bar 2 forward (single LLM, single pass, frozen corpus, no internet during scoring) Position: $30M synthetic UST book ($10M each in 2Y, 5Y, 10Y)

All numeric values in this document are V1 illustrative. The PoC tests whether the architecture works as designed; it does not claim alpha and does not claim the coefficients are calibrated.

---

1. Context and Thesis

The Risk Intelligence framework (SSRN 6584378; SSRN 6615841) proposes that an LLM can read textual regime evidence — FOMC statements, minutes, central bank speeches, market commentary — and emit a structured signal that a deterministic MMAR engine then converts into a regime-adjusted Hurst exponent and accompanying volatility / tail multipliers. The framing is deliberate: the LLM does not compute numbers. It reads, classifies, and emits a structured signal. The numbers come from a deterministic mapping that any reader can recompute by hand.

V1.5 tests the **forward feasibility** of this architecture — that is, whether the regime-scoring step can be executed under Bar 2 discipline (single LLM, single pass, no internet during scoring, frozen corpus) and produce engine inputs that meaningfully differ from a regime-blind baseline. V1 (the published framework paper) proposed the architecture in retrospective form. V1.5 runs it forward, on a real position over a real two-week window, with the score frozen before the window opens.

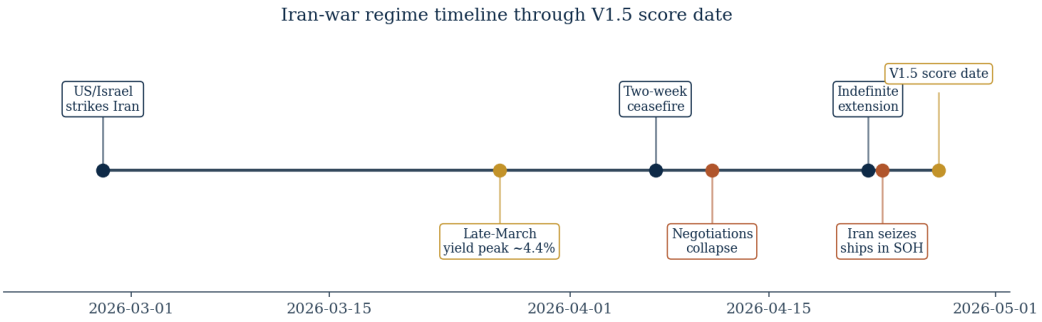
The hypothesis under test is narrow:

> *When the LLM reads the corpus and emits a structured regime signal, the deterministic mapping channels that signal into engine outputs that are materially different from both the Gaussian baseline and the text-blind static MMAR. Whether those different numbers turn out to be more accurate than the baselines is a separate question, addressable only with backtest data and ultimately requiring multi-window evaluation.*

What this PoC can demonstrate by design: that the architecture runs end-to-end, that the deterministic mapping arithmetic is auditable, and that the magnitude of the text-aware lift is bounded and replicable.

What this PoC cannot demonstrate: that the regime score is calibrated, that the engine output is more accurate than competitors, or that the approach generalizes to other asset classes or windows. Those claims require V2 with multiple windows and rigorous coefficient calibration.

The active regime under test is the Iran-war-driven multi-shock environment of April 2026. The figure below shows the timeline of the regime through the V1.5 score date.



Iran-war regime timeline through the V1.5 score date

---

2. Position and Design

The book is a synthetic $30 million long UST position, allocated as $10M notional each across the 2Y, 5Y, and 10Y constant-maturity points. Modified durations were derived from semiannual par-bond pricing at the April 24, 2026 close yields (3.78% / 3.92% / 4.31%). 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 window is the morning of Monday April 27, 2026, with corpus materials covering through Friday April 24 close plus weekend and Monday morning news. 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 discipline means three things:

1. **Single LLM, single pass.** The score is produced once against the frozen corpus. There is no iteration, no committee scoring, no fallback to a second model.
2. **No internet during scoring.** The LLM scores using only the materials in the frozen corpus. Anything not in the corpus cannot influence the score.
3. **Frozen corpus.** The corpus manifest is locked before scoring begins and cannot be modified to "rescue" any dimension after the score is produced. This is the property that makes the score

replicable: a different reader scoring the same corpus under the same rule should land within ±0.5 anchor units of the original.

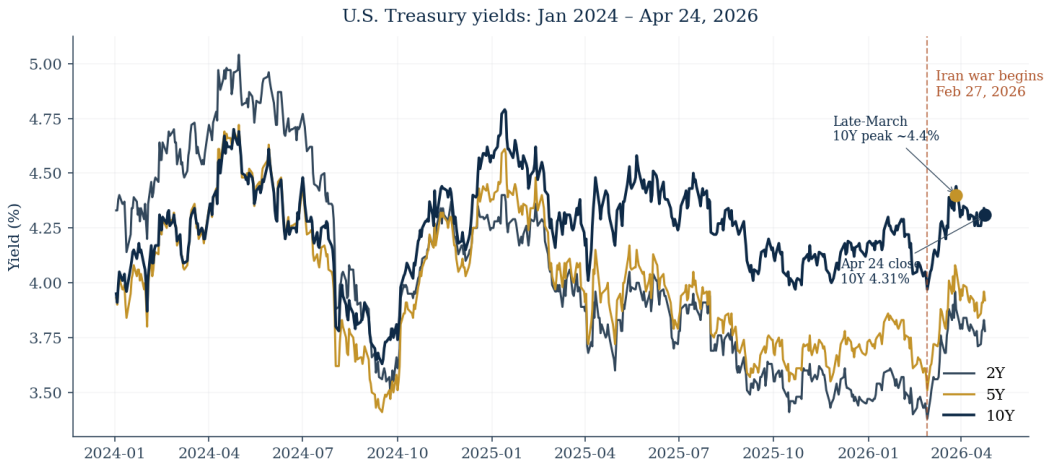
V1.5 is *not* Bar 3. Bar 3 would require multi-LLM ensemble scoring with disagreement adjudication, plus production-grade governance over corpus assembly. V1.5 is the feasibility bar that comes before Bar 3.

---

3. Baseline Computation

The historical baselines were computed from the FRED-equivalent daily Treasury yields for DGS2, DGS5, and DGS10 over January 2, 2024 through April 24, 2026 (578 trading days). FRED endpoints were not directly accessible from the execution environment, so US Treasury daily yield CSVs were used; the data series matches FRED to within rounding.

The yield path provides factual context for the regime under test.



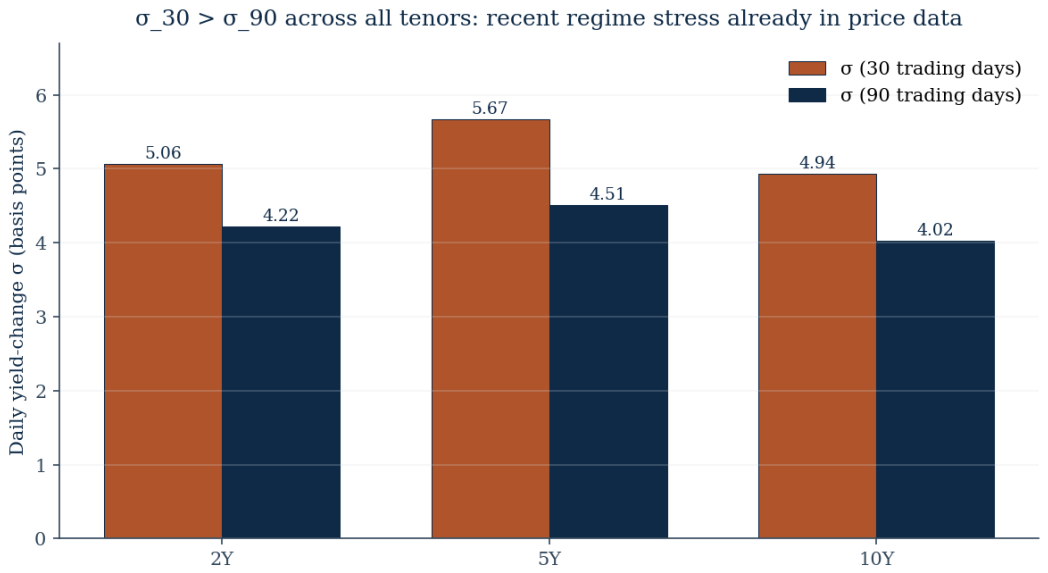
U.S. Treasury yields, January 2024 – April 24, 2026

Daily yield changes (in basis points) were computed via first-differencing. From these:

Daily volatility under two windows:

Tenor	σ (30d) bp	σ (90d) bp
DGS2	5.06	4.22
DGS5	5.67	4.51
DGS10	4.94	4.02

The σ_30 > σ_90 ordering across all three tenors is itself a regime-stress indicator — the recent 30 trading days carry the war's first 8 weeks of volatility, and the longer 90-day window dilutes them with the calmer pre-war January–February data. Any engine that uses σ_30 inherits some of the regime adjustment automatically; this is worth being explicit about, because a careless reader could mistake the σ_30 lift for the text-aware regime channel.



σ_30 vs σ_90 across tenors

Hurst exponent via rescaled-range (R/S) analysis:

Tenor	H (12m)	H (24m)
DGS2	0.489	0.568
DGS5	0.520	0.576
DGS10	0.541	0.582

The 12-month window puts H near 0.5 (random-walk-like), while the 24-month window shows clear persistence (H ≈ 0.58). The 24-month H captures the regime-switching pattern from the cuts cycle (Sept–Dec 2025) into the current hold (Jan–April 2026), which the 12-month window cannot see because it sits inside a single regime. This is why the headline engine runs use the 24-month H — it captures multi-regime persistence, which is the structural feature the framework is meant to detect.

Empirical correlation matrix of daily yield changes over the last 90 trading days:

	DGS2	DGS5	DGS10
DGS2	1.000	0.911	0.803
DGS5	0.911	1.000	0.946
DGS10	0.803	0.946	1.000

Adjacent tenors are tightly co-moving; the 2Y-10Y correlation is the lowest of the three (still 0.80). This curve-segment structure means parallel-shift VaR aggregation will overstate risk slightly at the 2Y-10Y interface; the empirical correlations capture this.

---

4. Source Corpus and the Coverage Decision

The frozen corpus has nine source clusters plus the computed baseline file. They are:

ID	Source	Date	Weight class
S01	FOMC statement, March 18	Mar 18, 2026	highest authority
S02	SEP + Powell press conference	Mar 18, 2026	highest authority
S03	FOMC minutes (released April 8)	Mar 17–18, 2026	highest authority
S04	Powell at Harvard EC10b lecture	Mar 30, 2026	highest authority
S05	Macro/geopolitical synthesis	through Apr 26, 2026	context
S06	Waller Auburn + Williams remarks	Apr 16–17, 2026	highest authority
S07	IMF Spring Meetings	Apr 13–18, 2026	context
S08	Market action / yield trajectory	through Apr 24, 2026	context
S09	Voting members: Bowman / Cook / Miran / Jefferson primary text	Jan 30 – Apr 7, 2026	highest authority, bounded
BASELINE	Computed σ, Hurst, DV01	Apr 24, 2026 close	engine input

The corpus was initially frozen with eight items (S01–S08) and a reference-only treatment of the four other voting members. After GATE 1 review, S09 was added with primary text for Bowman, Cook, Miran, and Jefferson, the manifest was reset, and the score was produced against the expanded ten-item corpus. The reason to document this expansion explicitly is that under Bar 2 discipline, who chose to add a source — and on what grounds — is part of the replicability story.

What S09 added and what it did not change: it confirmed the dovish wing exists as a real signal (Miran overt; Bowman pre-war stance with three cuts in her December SEP), confirmed Jefferson reinforces the Powell-Williams central tendency, and added Cook's stance-silent March 26 financial-stability speech as a soft continuity data point. It did not displace the central tendency — Powell, Williams, and Waller continue to dominate by recency × authority × topic relevance.

What the coverage explicitly does not include and why:

- *No fresh Powell scripted monetary-policy speech between March 30 and April 26.* This is consistent with the FOMC's informal pre-meeting blackout convention and is treated as an absence of new signal, not as a missing data point.
- *No fresh Bowman monetary-policy speech post-Iran-war.* Her March 31 (small business) and March 3 (liquidity) speeches are regulatory, not monetary. Her January 30 stance is the most recent; we treat it as stale on monetary policy and bound its weight by recency.
- *No primary text for Logan, Hammack, Paulson.* These voting members did not give material monetary-policy speeches in the window. Their absence is bounded by Powell-Williams-Waller dominance; if any had dissented at the March meeting, the dissent record (S01) would have caught it.
- *No real-time intraday April 27 data.* The corpus closes at Friday April 24 with weekend and Monday morning context as of the freeze timestamp (12:15 PM ET, April 27).

The replicability bar this corpus sets: a different reader applying the same conflict-resolution rule against the same frozen corpus should produce a composite severity score within ±0.5 anchor units of 7.0 — that is, between 6.5 and 7.5. Anything wider than that and the scoring methodology is failing the Bar 2 forward criterion of being a *single-pass, single-LLM* operation that other readers can sense-check.

---

5. Bar 2 Forward Scoring

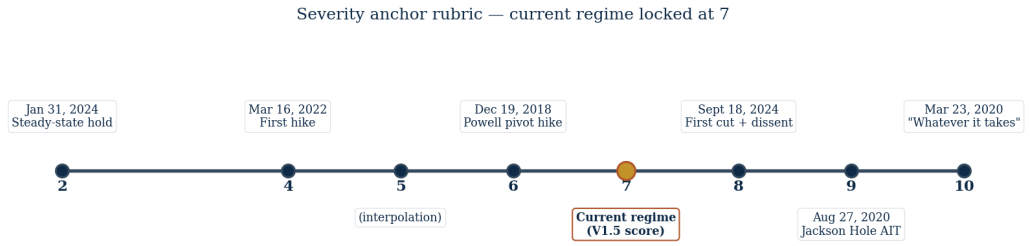
5.1 Methodology

The score has two structural elements: a six-anchor severity rubric and a six-dimension regime signal.

The **anchor severity rubric** locks current and historical regimes to specific dates so that scoring against it is a comparison act, not an absolute-magnitude estimate. The locked anchors are:

Score	Anchor event
10	March 23, 2020 — pandemic backstop, “whatever it takes,” market dysfunction
9	August 27, 2020 — Jackson Hole AIT framework reset
8	September 18, 2024 — first cut of cycle, Bowman dissent
7	Multi-shock dual-mandate stress with internal FOMC dispersion spanning cut/hold/hike-option, anchored long-run inflation expectations, orderly markets, and concurrent leadership-transition signaling. Distinguished from anchor 8 by multi-shock cumulation and institutional/leadership overhang; distinguished from anchor 9 by the absence of a regime-level framework reset.
6	December 19, 2018 — Powell pivot hike
5	(interpolated)
4	March 16, 2022 — first hike of tightening cycle
2	January 31, 2024 — steady-state hold

The figure below is the rubric visualized; the score (severity 7) is one anchor in a comparative scale, not an absolute-magnitude reading. Adjacent anchors carry structurally different regime characters; severity alone does not determine engine output (see Section 8).



Severity anchor rubric

Anchors 5 and 7 are interpolation points by design. Anchor 7 was originally written as a generic interpolation and was rewritten *during corpus expansion at GATE 1, before the V1.5 score was produced.* The rewrite was triggered by the S09 addition surfacing two distinctive features that warranted explicit naming: the dispersion-spanning-three-directions feature, and the leadership-

transition overhang. Both are operative differentiators from the 6 and 8 anchors and must appear in the rubric so that the engine grades adjacency cleanly. Importantly, the rewrite preceded the score and was not a post-score adjustment of the rubric to fit the result.

The **conflict-resolution rule** for divergent voices in the corpus is: voting > non-voting; Chair > governors; scripted > unscripted; recent > earlier. Genuine residual divergence after this weighting increases the dispersion-across-sources sub-score.

The **six dimensions** of the regime signal capture the structural inputs the deterministic mapping requires:

#	Dimension	What it does for the engine
D1	Severity score	Anchored against the rubric; sets baseline ΔH magnitude
D2	Persistence	Transient vs structural; sets sign-and-distance of ΔH from 0.5
D3	Policy direction bias	Conditional drift sign for asymmetric VaR
D4	Inflation-expectations anchoring	Term-premium scaling via vol multiplier
D5	Dispersion across sources	Uncertainty premium via dispersion factor
D6	Market functioning	Tail multiplier (GFSR-style stress signal)

Each dimension is scored from the corpus with at most one verbatim quote of fifteen words or fewer per source per dimension, supporting the rubric call.

5.2 Per-Dimension Walkthrough

D1 Severity = 7.0. The current regime materially matches the locked 7-anchor description. Multi-shock cumulation: explicit. Internal FOMC dispersion spanning cut (Miran), hold (Powell-Williams-Waller-Jefferson), and hike-option (per minutes' "some" voices): explicit. Long-run inflation expectations anchored: confirmed by IMF and Powell. Orderly markets and adequate liquidity: confirmed by IMF GFSR. Concurrent leadership-transition signaling: Powell term ending May 15; Warsh confirmation blocked by Tillis pending DOJ probe outcome; Cook contested at SCOTUS. None of the features that would push severity into the 8-band are present (no active policy turn, no expectations de-anchoring, no market dysfunction). None of the features that would pull severity into the 6-band are present either (this is multi-shock, not isolated-policy-mistake territory).

D2 Persistence = 0.65 (ambiguous, leaning structural). The shock has both transient elements (war could resolve quickly, per Waller's swift-resolution scenario) and structural elements (Powell explicitly invoked the five-year shock cumulation: tariff plus pandemic plus energy; Waller framed the sequence-

of-shocks as keeping inflation elevated for “quite some time”; the labor-supply structural shift from immigration changes is multi-year; succession overhang is at minimum 12-month structural). The structural elements dominate but do not entirely displace the possibility of swift war resolution.

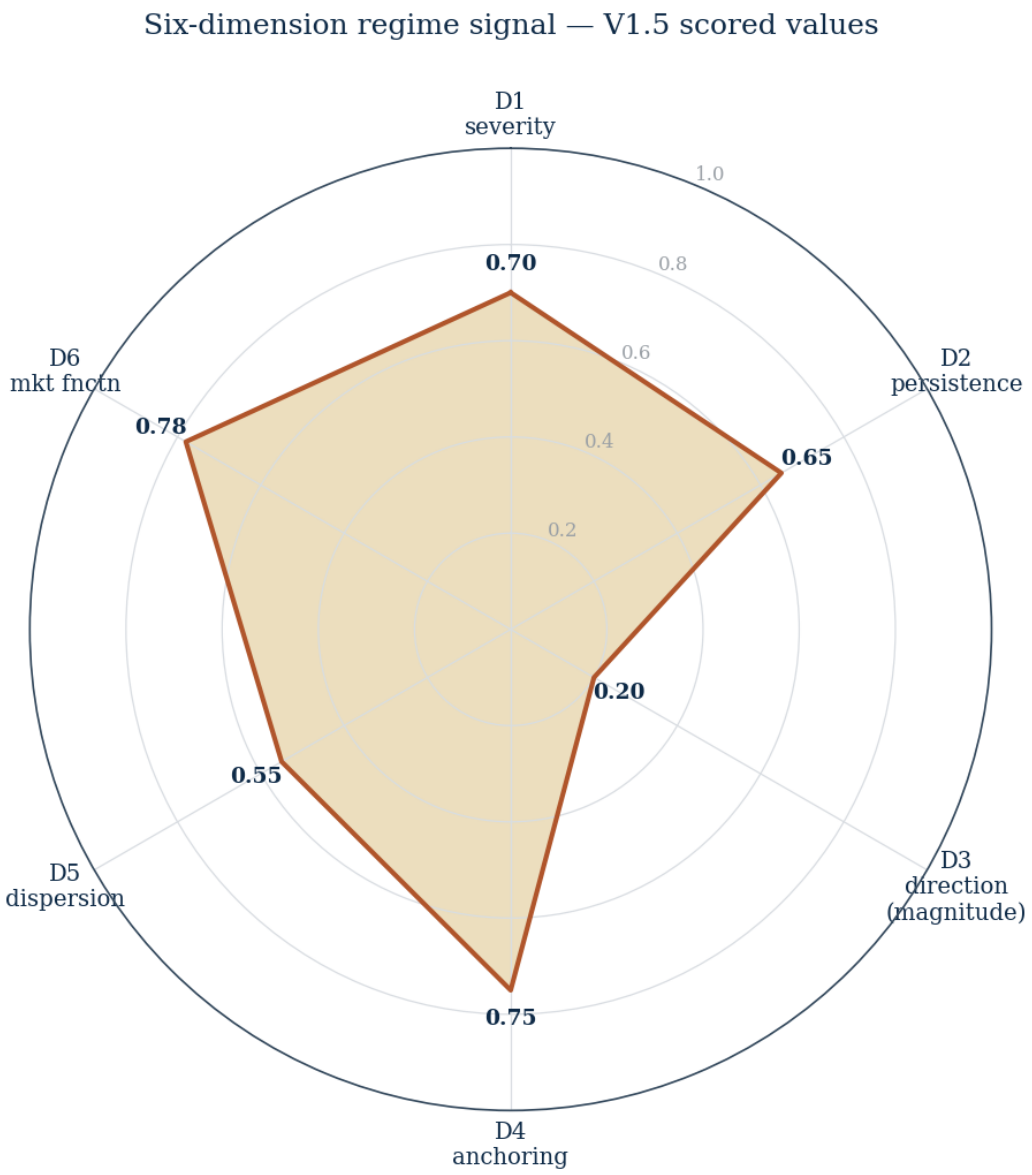
D3 Policy direction bias = −0.20 (hold, with soft conditional cut bias). Central tendency strongly favors hold across post-war voting members. Powell: “in a good place to wait and see.” Williams: “well above 3% inflation, no time for forward guidance.” Waller (the dovish dissenter at January): now cautious on cuts and would maintain rate if inflation risks dominate. Soft cut bias from Miran (lone March dissenter for 25 bp cut) and pre-war Bowman (three cuts in December SEP, stale). Hike option flagged in minutes by “some” voices but lacks a named voting champion. Markets price 26–43% chance of a single cut by year-end (range reflects ceasefire-news whipsaw).

D4 Inflation expectations anchoring = 0.75 (anchored, with short-term softening). IMF cross-country read: long-term expectations remain well-anchored. Short-term expectations have moved up in the U.S., per Georgieva. Powell repeatedly invokes anchoring concern as the operative risk that would force action. SEP raised 2026 PCE to 2.7% (from 2.5%); Waller forecasts March PCE at 3.5%; March CPI was 3.3% YoY (highest since May 2024). The anchoring is intact but increasingly contested.

D5 Dispersion across sources = 0.55 (moderate-to-high). Direction-of-next-move dispersion spans all three directions (cut / hold / hike-option) but with strong central tendency on hold. Severity-of-shock dispersion is uniformly low; virtually all voices acknowledge significant uncertainty. After conflict-resolution weighting, effective dispersion is moderate: the dovish flag (Miran) and hawkish-tail flag (minutes “some” voices) are both real and recent, but neither displaces the central tendency. This is the locked 7-anchor description’s defining feature, reproduced at the score level.

D6 Market functioning = 0.78 (orderly, with localized stress at the margins). GFSR characterizes markets as orderly. No margin-call or forced-deleveraging stress. Equities at record highs (S&P 500). Localized stress: late-March 2y/5y/7y auctions weak (primary dealer absorption 25% on the 2Y vs 11% historical average; 1.8 bp tail). Partial repair on April 22 with the 20Y “slightly above average demand.” $\sigma_{30} > \sigma_{90}$ across all tenors confirms recent stress in price data, but it is bounded.

The radar visualizes the scored dimensions in their six-axis structure.



Six-dimension regime signal — V1.5 scored values

5.3 Composite Assessment

The composite severity anchor is **7.0**, with confidence band ±0.5. The full set of dimensional values flowing into the deterministic mapping:

```
D1 severity = 7.0
D2 persistence = 0.65 (numeric)
D3 direction = −0.20 (numeric)
D4 anchoring = 0.75
```

```
D5 dispersion = 0.55
D6 market_functioning = 0.78
```

These produce, via the V1 illustrative deterministic mapping:

```
severity_norm = 0.70
persistence_signed = 0.30
dispersion_factor = 1.55
ΔH = (severity_norm − 0.5) × persistence_signed × (1 + dispersion) × α
   = 0.20 × 0.30 × 1.55 × 0.25
   = 0.0233 (under cap |ΔH| ≤ 0.10)

vol_multiplier = 1 + (1 − anchoring) × β = 1 + 0.25 × 0.20 = 1.050
tail_multiplier = 1 + (1 − market_functioning) × γ = 1 + 0.22 × 0.20 = 1.044
```

These three numbers — ΔH = +0.0233, vol_mult = 1.050, tail_mult = 1.044 — are what the engine consumes. Every other element of the score either feeds into these three, or is preserved for the documentation audit trail.

---

6. Why Discrete Anchored Scoring? The APGAR / Glasgow Coma Analogy

The framework converts qualitative regime evidence into a discrete anchored score on a 2–10 ordinal scale. This is structurally similar to how the APGAR score (newborn assessment) and the Glasgow Coma Scale (neurological assessment) work in medicine. The structural similarity is not coincidence; it is the answer to a specific class of measurement problem.

In all three cases, the underlying state being measured is multi-dimensional and qualitatively assessed, with no single continuous measurement that captures it cleanly. APGAR scores newborns on heart rate, breathing, muscle tone, reflex response, and color, summed to a 0–10 integer. Glasgow Coma scores patients on eye-opening, verbal response, and motor response, summed to a 3–15 integer. Both scales have remained in clinical use for sixty-plus years because they capture three properties that continuous measurement cannot:

1. **Anchored interpretability.** Each integer point on the scale has a referent state any trained clinician recognizes. APGAR 7 is not “0.7 of full health” — it is a specific clinical picture. The Risk Intelligence severity scale is the same: anchor 8 is “September 2024 first cut with Bowman dissent,” not “80% of maximum severity.”

2. **Replicability across observers.** Two clinicians scoring the same patient with APGAR will land within one integer of each other almost always. The reproducibility property is what makes the score usable in clinical practice. Bar 2 forward discipline pursues the same property at the regime-scoring level: a different reader scoring the same corpus should land within ±0.5 anchor units.
3. **Forced disagreement on dimensions, agreement on summary.** If two clinicians disagree on a Glasgow component (motor response, say), they can adjudicate the disagreement at the dimension level rather than relitigating the entire patient assessment. The six-dimension structure of the regime signal does the same: dispersion in D5 captures genuine residual disagreement after conflict-resolution, and the per-dimension verbatim excerpts make adjudication concrete rather than abstract.

These three properties — anchoring, reproducibility, dimensional adjudicability — are why discrete anchored scoring is appropriate for a regime that is multi-dimensional, qualitatively assessed, and required to be replicable across observers. The alternative (continuous regression on hand-engineered features) loses all three properties at the cost of false precision.

---

7. Why Not Just Sentiment Analysis? Five Structural Differences

The most common skeptical question is: “Isn’t this just sentiment analysis with extra steps?” It isn’t, and the differences are structural rather than rhetorical.

1. **Sentiment analysis maps text to a scalar (positive / negative). The Nexus signal maps text to a six-dimensional structured object.** The dimensional structure exists because the deterministic mapping requires structural inputs the engine can use; a scalar sentiment score cannot drive a Hurst adjustment, a vol multiplier, and a tail multiplier independently. The engine literally cannot consume sentiment.
2. **Sentiment analysis pools voices uniformly. The Nexus signal applies a documented conflict-resolution rule (voting > non-voting; Chair > governors; scripted > unscripted; recent > earlier).** Pool-uniformly approaches treat a junior staff economist’s blog post and the Chair’s prepared press conference as comparable inputs. The conflict-resolution rule treats them as differently weighted, with the weighting documented and replicable.
3. **Sentiment analysis is anchored to the model’s training corpus. The Nexus score is anchored to a fixed historical rubric that survives training-corpus drift.** A sentiment model trained in 2022 will score 2026 text against 2022 baselines silently. The Nexus rubric is anchored to specific dated regimes (March 2020, August 2020, September 2024, December 2018, etc.) that any reader can look up. The anchor does not drift with the model’s training cutoff.

4. **Sentiment analysis produces a number that goes directly into a downstream model. The Nexus signal flows through a deterministic mapping any reader can recompute by hand.** The deterministic mapping is the audit trail. A risk officer can write the score, the formulas, and the resulting engine inputs on a whiteboard and walk through them with auditors. A sentiment-driven model cannot be audited at this level because the mapping from sentiment-to-engine-input is opaque.
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 framework treats genuine residual disagreement as information that widens the engine's uncertainty band rather than as noise to be averaged away.

The first difference is the structural one — sentiment analysis cannot drive the engine because the engine needs structured inputs. The other four are governance properties that matter for institutional adoption (auditability, replicability, anchoring, uncertainty quantification).

7B. Why Not Just Use Prices? The Efficient-Market Objection

A second skeptical question, distinct from the sentiment-analysis question of Section 7 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, and any architecture using both is double-counting at best.* This is the Efficient-Market Hypothesis applied to LLM-emitted regime signals. The objection is serious. V1.5 does not dismiss it, does not claim to defeat it, and does not assert that the framework provides forecasting edge that survives semi-strong-form market efficiency at long horizons. What V1.5 *can* defend is the architectural claim that the framework's value persists *given* significant EMH effects — and it commits to a falsifiable empirical test of the residual informational claim in V2.

1. **The framework is structurally a regime-confirmation framework, not a regime-detection framework, and acknowledges that status explicitly.** Markets routinely lead official communication; pricing the next FOMC move ahead of the FOMC is the modal pattern in any normal regime. The V1.5 corpus is composed of voting Fed members and official documentation precisely because auditability requires fixed sources with editorial provenance. The framework's signal is, by design, slightly lagging — it confirms what the official apparatus believes about the regime, in a structured form that price action alone does not provide. For institutional users defending capital adequacy positions to regulators, "the bond market told us" is a weaker defense than "the FOMC said X and our framework processed it through documented methodology Y." Auditability has institutional value beyond predictive accuracy.

2. **The framework's plausible informational contribution is in the consolidation and structuring of dispersed-source narratives, not in scooping prices on individual events.** Markets price individual headlines efficiently; they integrate diverse multi-source narratives more slowly and inconsistently. When five voting members make differently-emphasized statements over two weeks, the structured Nexus signal aggregates the dimensional content (severity, persistence, dispersion) faster and more replicably than a price-based estimate of the same regime characteristics. Whether this consolidation actually produces a 10-day VaR forecast superior to price-only baselines is a hypothesis, not a demonstrated property. V2 multi-window calibration is designed to test it directly.

3. **The deterministic mapping is auditable in a way that price-only models cannot be.** A risk officer can write the regime score, the formulas, and the resulting engine inputs on a whiteboard and walk through them with auditors. A price-only Hurst-driven VaR cannot be audited at this level — the price series is a single observable but the *interpretation* of recent price action (regime-aware vs. transient noise) is precisely what the LLM signal makes explicit. This is governance value rather than forecasting value, and it persists whether or not the framework adds incremental information beyond prices.

4. **The framework's contribution is hypothesized to concentrate during regime transitions, when both price signals are noisiest and structured information has the most value.** This is the architecture's hypothesized area of maximum marginal value relative to price-only baselines, but it is presented here as architectural intent grounded in MMAR theory rather than as empirically demonstrated property. MMAR exists because returns are non-homogeneously distributed across time — calm periods are statistically uninteresting, while regime transitions concentrate the volatility and tail behavior that drive risk outcomes. The LLM-integrated MMAR extends this thesis to the textual signal: the regime channel's contribution is hypothesized to be most pronounced when narratives are dispersed and price signals haven't yet integrated the multi-source picture (March 2020, December 2018, August 2022 are the historical examples that motivate the hypothesis). Whether transition periods are *uniquely* the place where the regime channel adds value beyond price-derived signals is the empirical question that V2 multi-window calibration with deliberate transition-window over-sampling (Q14) is designed to test. The current response is defensible as architectural intent; it is not yet defended as empirically demonstrated.

5. **The empirical test is pre-registered and falsifiable.** V2 will run the framework on multiple historical 10-day windows with cross-LLM scoring and compare engine output to realized P&L. If the regime channel adds no information beyond price-derived Hurst (the EMH-redundancy null hypothesis), multi-window backtesting will reveal that — the E3 (Nexus-adjusted) calibration will not improve on E2 (static MMAR) systematically. If the regime channel does add information, the calibration improvement will be measurable. Methodological Question Q14 (added in this revision) formalizes this falsification test explicitly.

6. **Why doesn't the market already do this?** The most direct EMH-flavored objection is that professional market participants — discretionary macro traders, primary dealer desks, sell-side rates strategists, high-frequency algorithmic systems — already parse the identical central-bank corpus in real time. If that human and algorithmic parsing already incorporates the narrative content into prices, the LLM layer is redundant replication rather than incremental information. A related objection from the Grossman-Stiglitz tradition holds that even if the framework did extract a residual signal, deploying it at scale would arbitrage that signal away, returning the system to EMH-equivalence. Both objections are serious and V1.5 does not dispute either: the framework's value does not require that it extract information unavailable to competitive market participants, nor that the residual signal survive scale deployment. What V1.5 does claim is that the framework's value persists in dimensions that informal market parsing does not provide — bounded structured output that maps to engine parameters arithmetically, methodology that other practitioners can replicate without access to proprietary trading infrastructure, audit trail that satisfies institutional governance requirements regardless of whether the underlying regime characterization was already in prices. Nexus is designed for institutional risk management, not alpha generation. Its goal is not to outperform competitive market parsing on informational content; it is to ensure the risk-reporting layer is as structurally complex as the market it monitors, with the audit trail that regulated institutions require. This positioning is robust to both the latency/replication objection and the Grossman-Stiglitz arbitraging argument: the value V1.5 claims survives both.

The single-window May 11 result, by itself, cannot adjudicate this objection. A single observation cannot distinguish between "the framework added information" and "the framework happened to produce a number close to realized loss in this window." The objection is genuinely undecided pending V2 evidence. What V1.5 demonstrates is that the architecture runs end-to-end, produces materially different VaR outputs from price-only baselines, and structures the question precisely enough that V2 can answer it empirically rather than rhetorically.

---

8. Sensitivity Analyses

Two sensitivity tests were run to characterize the engine's behavior around the score: a cap saturation sweep over severity, and a severity adjacency comparison under both fixed and anchor-character-consistent dimensional profiles.

8.1 Cap Saturation

Holding D2–D6 at their scored values, sweeping D1 from 5.0 through 10.0:

- ΔH varies linearly from 0.0 (at severity 5.0) to 0.058 (at severity 10.0).
- The cap ($|\Delta H| \leq 0.10$) does not bind anywhere in the [5, 10] range under current D2/D5 values.

- The cap would only bind under simultaneous extremes (severity ≈ 10, persistence ≈ 1.0, dispersion ≈ 1.0), which the locked 7-anchor description structurally precludes.

In the ±0.5 confidence band on the actual score (i.e., [6.5, 7.5]), ΔH ranges from 0.0174 to 0.0291 — a 67% relative spread but absolutely small (about 1.2 bp on Hurst). The engine output's sensitivity to scoring uncertainty is therefore modest and bounded.

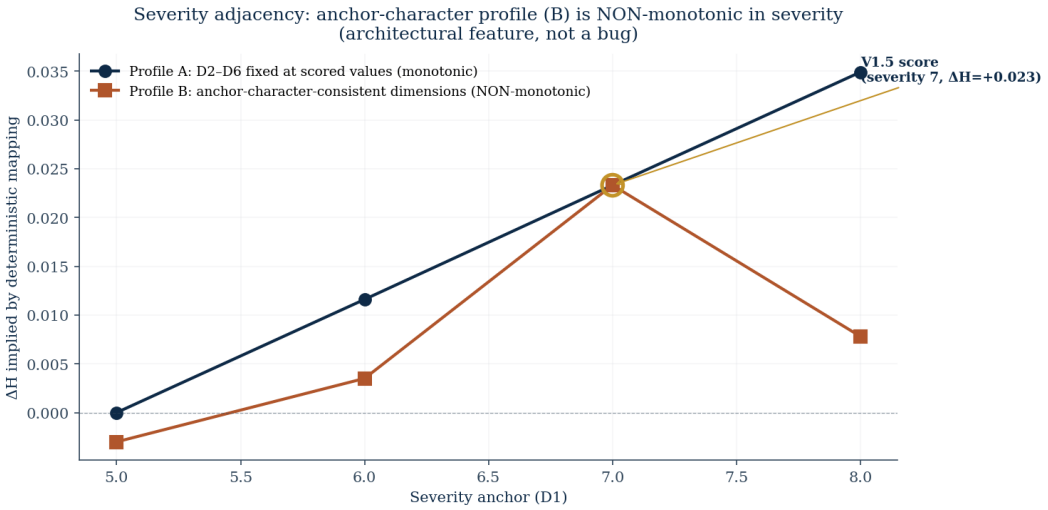
8.2 Severity Adjacency, and the Non-Monotonicity Finding

Under fixed dimensional profile (D2–D6 held constant), the engine produces a smoothly increasing ΔH with severity. This is the expected behavior.

Under **anchor-character-consistent dimensional profiles**, the engine's behavior becomes more interesting. Each severity anchor implies a specific dimensional structure:

Anchor	D2 persistence	D3 direction	D4 anchoring	D5 dispersion	D6 mkt funct	ΔH implied
5.0 (calm-to-stress)	0.45	0.00	0.85	0.20	0.90	−0.003
6.0 (Dec 2018 pivot)	0.55	+0.30	0.85	0.40	0.65	+0.004
7.0 (current regime)	0.65	−0.20	0.75	0.55	0.78	+0.023
8.0 (Sept 2024 first cut)	0.55	−0.50	0.85	0.30	0.92	+0.008

The figure below shows ΔH versus severity under both profiles. The dip at anchor 8 in the orange line is a feature of the architecture, not a bug; the prose that follows explains why it appears and why it is the central architectural claim being tested.



Severity adjacency: Profile A (fixed dimensions) is monotonic; Profile B (anchor-character-consistent) is not

Note that under anchor-consistent dimensions, ΔH is *not* monotonic in severity. The current regime (anchor 7) actually produces a *higher* ΔH than the Sept 2024 first cut (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 (the cycle had just turned), lower dispersion (the dissent was bounded), more orderly markets (no parallel war). The Hurst-adjustment channel responds to regime *structure* — to the dimensional sub-scores — 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: that the regime signal is a structured object whose dimensional components carry independent information that severity headlines cannot capture. A purely severity-headline scorer would mis-rank the 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 the engine output, because the *dimensional structure* dominates. Conversely, two scorers who agree on severity 7 but disagree on persistence will produce materially different engine outputs. This relocates the 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.

A note on the dimensional profiles in the table above. The dimensional values assigned to historical anchors (Sept 2024 first cut, Dec 2018 pivot hike, etc.) are described as "anchor-character-consistent" and are plausible reconstructions, but they are not independently derived from contemporaneous text using the same six-dimension methodology applied to V1.5. A skeptical reader can fairly object that if these profiles were chosen post-hoc to produce dimensional structures consistent with the architectural

claim, the non-monotonicity finding is constructed rather than discovered. The framework acknowledges this and names a falsifying test: rescore each historical anchor against its contemporaneous corpus using the same methodology and conflict-resolution rule, and compare the derived profile to the values in the table above. If derived profiles match within ±0.10 per dimension, the architectural claim survives empirical test. If they materially differ, the V1.5 claim must be revised. This test is V2 priority 1 (Open Methodological Questions, Q11) — the single most consequential robustness check the framework faces.

---

9. Three-Engine Results

The three engines are:

1. **Engine 1 — Gaussian VaR.** Standard parametric VaR with sqrt-t scaling. Uses σ daily and DV01. No regime adjustment whatsoever. This is the baseline a risk-blind reader would produce.
2. **Engine 2 — Static MMAR.** Replaces sqrt-t scaling with t^H scaling, where H is the historical Hurst exponent. Captures self-similar variance growth that exceeds Gaussian when H > 0.5. Uses no text input — the regime is read entirely from price history.
3. **Engine 3 — Nexus-adjusted MMAR.** Uses the regime-adjusted Hurst (H_base + ΔH_regime), the vol multiplier on σ, and the tail multiplier on the quantile. This is the architecture under test.

9.1 Headline Results

For the σ_90d / H_24m run (long baseline, captures multi-regime persistence):

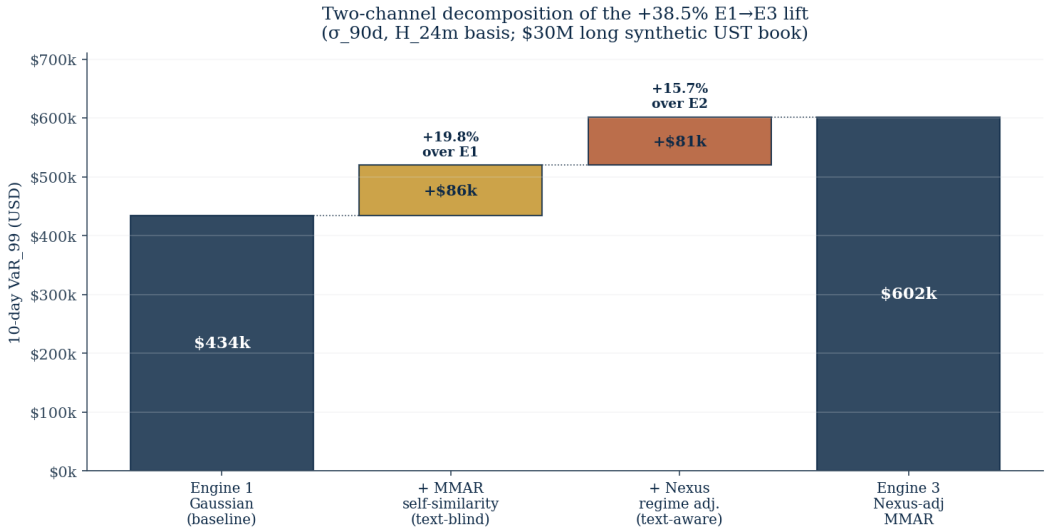
Engine	VaR 95%	VaR 99%	Lift (99%)
E1 Gaussian	$307,162	$434,424	(baseline)
E2 Static MMAR	$367,842	$520,246	+19.8% vs E1
E3 Nexus-adj MMAR	$425,452	$601,725	+15.7% vs E2; +38.5% vs E1

For the σ_30d / H_24m run (recent stress weighting):

Engine	VaR 95%	VaR 99%
E1 Gaussian	$379,071	$536,128
E2 Static MMAR	$453,965	$642,052
E3 Nexus-adj MMAR	$525,064	$742,608

9.2 The Two-Channel Decomposition

The figure below shows the +38.5% E1→E3 lift as two stacked architectural channels. Reader caution: the +38.5% combined lift includes the +19.8% pure MMAR contribution that is text-blind. Only the +15.7% increment from E2 to E3 is the text-aware Nexus channel. The decomposition is the headline finding because it isolates the architecture-under-test from the multifractal property of the price series itself.



Two-channel decomposition of the +38.5% E1→E3 lift

The 38.5% lift from E1 to E3 splits cleanly into two architectural channels:

Channel 1: Pure MMAR self-similarity (E1 → E2, +19.8%). Comes from t^H scaling instead of sqrt(t) scaling. With H_24m around 0.58, 10^0.58 ≈ 3.80 versus 10^0.50 ≈ 3.16, which is a 20% widening of the variance window before any text gets read. This channel is text-blind. It would be present even if the LLM reading step were skipped entirely.

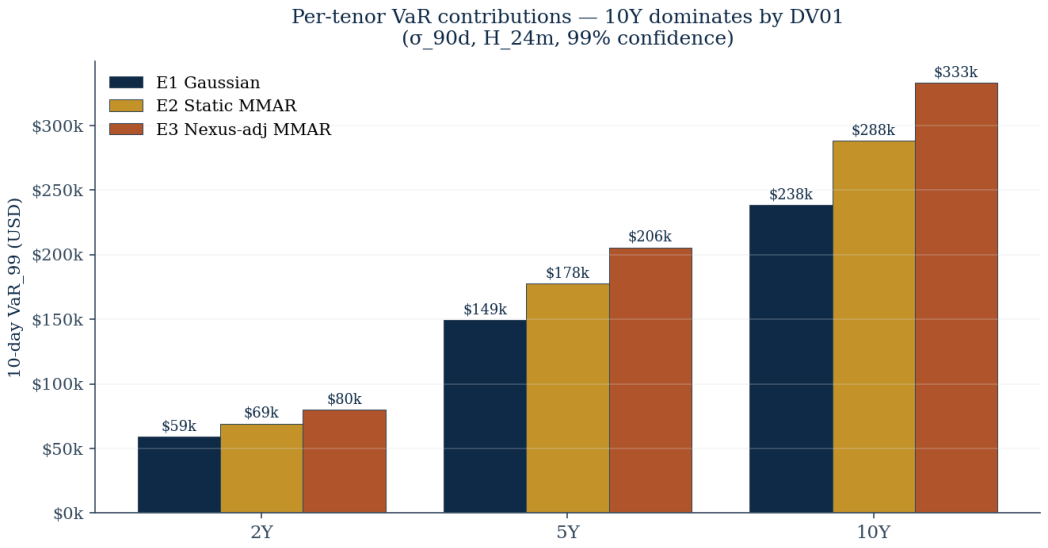
Channel 2: Text-aware regime adjustment (E2 → E3, +15.7%). Comes from three multiplicative components: ΔH = +0.0233 contributes about +5.5% to t^H; vol_mult = 1.050; tail_mult = 1.044. Multiplied together: 1.055 × 1.050 × 1.044 ≈ 1.157, matching the observed +15.7% to two decimals. The deterministic mapping arithmetic checks.

The ability to decompose the lift cleanly is itself a property of the architecture. A monolithic risk model that produced "VaR went up by 38%" could not say which channel contributed how much. This decomposition is what makes the architecture auditable.

9.3 Per-Tenor Breakdown

For the σ_90d / H_24m run at 99%:

Tenor	E1	E2	E3	E3 − E1	E3 − E2
DGS2	$59,221	$69,259	$80,106	+$20,885	+$10,847
DGS5	$149,192	$177,724	$205,559	+$56,367	+$27,835
DGS10	$238,442	$287,994	$333,099	+$94,657	+$45,105



Per-tenor VaR contributions — 10Y dominates by DV01

The 10Y dominates because of DV01 dominance ($8,055/bp versus $4,501 and $1,909). The 5Y has the highest σ but a smaller DV01. The 2Y is the smallest VaR contributor in absolute terms.

9.4 Position Sizing Sanity Check

The E3 VaR_99 of $601,725 on the $30M book is 2.0% of notional over a 10 trading day horizon. Under the stress-weighted σ_30 baseline, E3 VaR_99 is $742,608, or 2.5% of notional. These figures sit in a plausible range for a 10-day VaR on a long UST book in a stressed regime; nothing in the engine output breaks face validity.

9.5 What the σ_30 vs σ_90 Spread Means

Engine 1 alone produces a 23% higher VaR using σ_30 versus σ_90 ($536,128 versus $434,424). This gap is real but should not be interpreted as evidence of the text-aware channel. It is the recent-stress-weighted volatility baseline; a regime-blind reader using σ_30 would inherit some of the regime adjustment automatically, simply because the σ_30 window includes the war's first 8 weeks.

The full text-aware story is: the *gap* between σ_30 and σ_90 is one piece of the regime signal (price-data evidence of recent stress), and the text-aware Nexus adjustment is another piece (regime-narrative evidence of structural shock). Both are needed for a complete picture. A V1 reader could mistake either piece in isolation for the whole story. The empirical test for whether the two pieces double-count — quantifying overlap between σ_30 stress weighting and the text-aware vol_mult / tail_mult / dispersion-factor channels — is V2 work (Open Methodological Questions, Q13).

---

10. Reasoning Capability Decomposition

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.

Question 1 — Did the LLM read the text correctly? This is a comprehension property. It can be measured by checking whether the verbatim excerpts the LLM extracted are actually present in the corpus and whether they support the dimensional claims they are attached to. This is replicable because the corpus is frozen and the excerpts are direct quotes.

Question 2 — Did the LLM produce a coherent structured signal? This is an internal-consistency property. It can be measured by checking whether the dimensional sub-scores cohere — for example, a regime scored as "expectations de-anchoring" (D4 low) should also show "policy direction toward hike" (D3 positive) under most plausible scenarios. Internal contradictions across dimensions are an automatic failure.

Question 3 — Did the deterministic mapping correctly translate the signal into engine inputs? This is an arithmetic property. It can be measured by recomputing the mapping on a calculator. The architecture's commitment to a deterministic, hand-recomputable mapping is what makes this question *trivially* answerable. A reader can verify the mapping in five minutes with a pencil.

Question 4 — Did the engine forecast align with the realized outcome? This is a calibration property. It can be measured only after the project window closes, by comparing realized P&L to the VaR thresholds at each engine's confidence level. With a single window, this is a single-point check; with many windows, it becomes a calibration distribution.

The four questions are independent in principle: an LLM could pass questions 1, 2, and 3 and fail question 4, or vice versa. Pass-fail patterns across the four questions diagnose which part of the architecture is responsible for any failure.

This decomposition also serves as a bridge to the broader training-substrate thesis. If LLMs are to become trusted components in financial risk infrastructure, the “did it reason well” question has to be addressable by methods that resemble the audit methods institutional risk officers use today. The four-question decomposition provides exactly those methods. A system that fails question 1 has a comprehension problem; a system that fails question 2 has a coherence problem; a system that fails question 3 has an implementation problem; a system that fails question 4 has a calibration problem. Each failure is repaired by different work.

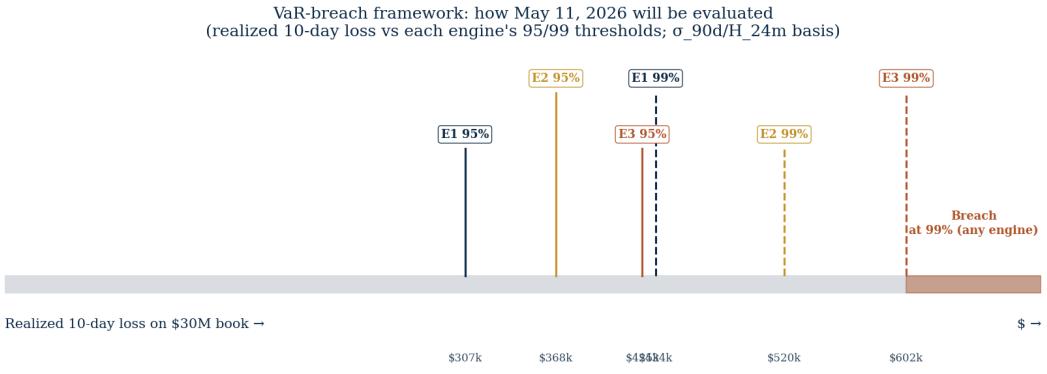
This is also why the Bar 2 forward discipline matters more than retrospective evaluation. A retrospective study can confound questions 1–3 with question 4 because the LLM may have seen the realized outcome during training. Forward discipline isolates question 4 from the others by placing the score before any realized data exists.

---

11. Forward Backtest Framework

The 10-trading-day project window closes on Monday May 11, 2026. On that date, the realized 10-day P&L on the synthetic book can be computed from realized yield changes April 28 through May 11 multiplied by per-tenor DV01s. This realized number is then compared to each engine’s VaR_95 and VaR_99 thresholds.

The figure below shows the framework structure. Reader caution: the diagram does not pre-judge any engine. It places the six VaR thresholds (E1/E2/E3 × 95%/99%) on a single axis so that, on May 11, the realized loss can be located on the same axis and any breaches identified directly. We do not yet know which engine’s number is right; we know what we will measure.



VaR-breach framework: the six thresholds against which May 11 realized loss will be compared

The single-observation interpretation:

- If realized loss > VaR_99 from any engine, that engine produced a "breach" at the 99% level.
- If realized loss > VaR_95 from any engine, that engine produced a breach at the 95% level.
- If realized loss < all VaR_95 thresholds, no engine was breached and the test is uninformative on which engine is best.

A single observation cannot determine which engine is best calibrated. The expected breach rate at the 99% level is 1% — meaning, in expectation, one breach per hundred 10-day windows. 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 was so far from realized that it failed face-validity (e.g., a VaR_95 breach an order of magnitude smaller than realized loss, or a VaR_99 figure ten times larger than realized 10-day vol).

The framework should therefore be run on multiple non-overlapping windows in V2. A 12-window evaluation (one year of monthly windows) gives roughly 12 observations at the 95% level, of which 0–1 expected breaches, and 12 observations at the 99%, of which 0–1 expected breaches. Even at this scale, the test is power-limited; meaningful calibration requires 100+ windows.

The current single-window forward test is a feasibility test, not a calibration test. It demonstrates that the architecture produces interpretable outputs that can be checked against reality; it does not, and cannot, demonstrate that the outputs are correctly calibrated.

---

12. Limitations and V1 Disclosures

The following are explicit limitations of V1.5 that any reader should treat as binding caveats.

The coefficients are illustrative. $\alpha = 0.25$ (Hurst-adjustment scaling), $\beta = 0.20$ (vol-multiplier sensitivity to anchoring), $\gamma = 0.20$ (tail-multiplier sensitivity to market functioning) are chosen to produce engine outputs in a plausible range, not to be empirically calibrated. Empirical calibration is V2 work; it requires multi-window historical scoring and would meaningfully change the magnitude of the lift.

ΔH is term-structure-agnostic. The same regime adjustment is applied uniformly across 2Y, 5Y, and 10Y. Any economically meaningful regime should affect the curve heterogeneously — short-end persistence is driven by Fed expectations, long-end by term premium and structural factors. V2 should condition ΔH on tenor.

The MMAR implementation uses Gaussian quantiles. True MMAR specifies a multifractal distribution with fat tails. V1.5 retains Gaussian quantile multipliers (1.645 at 95%, 2.326 at 99%) and applies regime adjustments to the σ scaling and tail multiplier. A V2 enhancement would use a multifractal copula and tenor-specific scaling factors; the engine output magnitude would shift, though the relative E1-E2-E3 ranking should be preserved.

The corpus is bounded by what was publicly available by April 27, 2026. Internal Fed communications, OIS-implied path detail, and primary-dealer survey results were not in the corpus. A production system would have access to the FedWatch data, primary-dealer surveys, and possibly internal positioning data; the score would be richer.

The conflict-resolution rule is rule-based, not learned. Voting > non-voting; Chair > governors; scripted > unscripted; recent > earlier. These weights are imposed by design and are not empirically validated against historical regime accuracy. A learned weighting scheme might outperform; it would lose the auditability property the rule was designed to preserve.

Single LLM, single pass. Bar 2 forward by definition. Bar 3 with multi-LLM ensemble is the production target; V1.5 is the feasibility step before Bar 3.

No alpha claim. V1.5 does not claim its forecasts are more accurate than the 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 the baseline — that claim the engine results support.

---

13. Literature and Novelty

Two recent works in the LLM-and-monetary-policy literature inform this PoC and should be cited explicitly. 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 (GDP, inflation, policy rates, sector exposures), 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. Specifically, *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

(Soleimani's λ mixing calm and crisis covariance matrices; Nexus's anchored severity rubric driving ΔH). The deterministic-mapping-with-auditability pattern is therefore converging industry practice as of late 2025, not a unique architectural commitment of the Nexus framework.

The structural differences between Soleimani and Nexus are real but specific. Soleimani is a *counterfactual scenario generator* — it produces forward synthetic shocks for stress testing across G7 economies. Nexus is a *current-state regime scorer* — it reads contemporaneous text and emits a structured signal characterizing the prevailing regime. Soleimani uses linear PCA factor channels and polynomial factor channels for shock translation; Nexus uses multifractal Hurst-exponent adjustment. Soleimani has no anchored ordinal severity rubric tied to dated historical regimes; Nexus does. Soleimani has no conflict-resolution rule across speakers; Nexus does. The use cases are complementary rather than competing.

The novelty claims of the Nexus framework, as concretized in V1.5, are revised against this updated reading of the literature:

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 the question of how to parameterize regime structure for the multifractal scaling exponent; the LLM-emitted six-dimension structured signal is one answer to this parameterization question. This remains the genuine architectural novelty of the framework.

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 a 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 λ 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 decomposition (text reading, signal coherence, deterministic mapping, forecast vs ground truth) provides a substrate for evaluating LLM reasoning quality in financial risk applications using methods that resemble institutional audit practices. The decomposition is not an end-to-end audit of the LLM's reasoning — Q3 (deterministic mapping) is calculator-trivial on the framework's own formulas, and 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 thesis is that the four questions provide more diagnostic granularity than “did the LLM reason well?” — not that the LLM is rendered fully auditable.

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 executive summary’s institutional-audience discussion and in V2 work; V1.5 does not exercise it because the score (7.0) does not cross the breaker threshold.

14. Events Inside the Project Window (Forward Application)

The V1.5 score was frozen on Monday April 27, 2026 at 12:15 PM ET against a corpus closing Friday April 24. Bar 2 forward discipline holds the score fixed for the duration of the project window — April 28 through May 11, 2026 — regardless of what news arrives during that window. The score was not, and will not be, updated.

This section documents two events that broke inside the project window and applies the framework methodology *as if* the events had been in the original corpus. The purpose is twofold. First, transparency: a reader should know what the score-time information set excluded. Second, framework demonstration: applying the same six-dimension methodology and the same conflict-resolution rule to inside-window events shows how the architecture would have responded if the information had been available, without breaking forward discipline. The May 11 backtest will compare realized 10-day P&L against the V1.5 thresholds (the actual test of the architecture's calibration), not against the hypothetical inside-window-aware thresholds developed below.

14.1 Powell continuation as Federal Reserve Governor (signaled April 28, 2026)

Source events. The U.S. Attorney for the District of Columbia announced on Friday April 25 that the Justice Department was closing its investigation of Powell and handing remaining matters to the Fed's Inspector General. The Senate Banking Committee scheduled the Warsh confirmation vote for April 29. Powell himself, in a March press conference referenced widely on April 28, said he had *"no intention of leaving the board until the investigation is well and truly over, with transparency and finality."* EY chief economist Gregory Daco wrote in a client note that Powell remaining on the board would *"help preserve institutional continuity, anchor the existing communication approach, and provide a stabilizing counterweight during the transition."* Warsh's communication agenda — characterized as wanting to *"recalibrate how the Fed operates"* with *"regime change"*, a new inflation framework, smaller balance sheet, less forward guidance, and possibly no post-meeting press conferences — sits in tension with Powell-as-continuity.

Six-dimension assessment if this had been in the V1.5 corpus.

Dim	V1.5 score	If-in-corpus	Δ	Rationale
D1 severity	7.0	7.5	+0.5	Twin-power Chair/ Governor configuration is structurally novel; pushes toward 8-band but does not cross (no policy turn, no expectations de-anchoring)
D2 persistence	0.65	0.75	+0.10	Powell-on-Board through January 2028 extends regime overhang from May 2026 horizon to a 21-month minimum
D3 direction	−0.20	−0.15	+0.05	Marginal softening of hold bias; Warsh “regime change” rhetoric introduces tail-risk option for hike or cut depending on framework interpretation
D4 anchoring	0.75	0.70	−0.05	Warsh’s proposed “new inflation framework” raises anchoring concern at the margin; partially offset by Powell continuity
D5 dispersion	0.55	0.70	+0.15	Two centers of communication authority (Powell-stable, Warsh-change) is the

Dim	V1.5 score	If-in-corpus	Δ	Rationale
				largest dispersion increment of any V1.5 dimensional shift
D6 mkt functioning	0.78	0.78	0	Markets read Powell continuity as stabilizing; no immediate dysfunction signal

Engine effect under deterministic mapping. Updated severity_norm = 0.75; persistence_signed = 0.50; dispersion_factor = 1.70. Updated ΔH = 0.25 × 0.50 × 1.70 × 0.25 = 0.0531 (vs V1.5 ΔH = 0.0233). Updated vol_mult = 1.060 (vs 1.050). Tail_mult unchanged at 1.044. Channel-2 lift if Powell continuation alone in corpus: 1.124 × 1.060 × 1.044 ≈ 1.243, vs V1.5's 1.157 — approximately +24% text-aware lift instead of +15.7%.

The architecture is responsive to the structural feature (twin-power configuration) primarily through the dispersion and persistence dimensions, which feed into ΔH multiplicatively. The Powell event is the more architecturally consequential of the two events covered in this section.

14.2 UAE departure from OPEC and OPEC+ (announced April 28, 2026, effective May 1)

Source events. The UAE announced on April 28 that it would leave OPEC and OPEC+ effective May 1, 2026, ending sixty years of membership. UAE Energy Minister Suhail Al Mazrouei said the move reflected a *"policy-driven evolution aligned with long-term market fundamentals."* WTI crude broke above $100 per barrel for the first time since April 10, reaching $102 in early April 28 trading; Brent rose to ~$113. Iran negotiations stumbled — Trump publicly rejected the latest Iranian proposal on April 27 — and Strait of Hormuz transit had collapsed to roughly six ships per day from the pre-war 130. Rystad's Jorge Leon characterized the UAE exit as having *"near-term effects muted given ongoing disruptions in the Strait of Hormuz; longer-term implication is a structurally weaker OPEC."*

Six-dimension assessment if this had been in the V1.5 corpus.

Dim	V1.5 score	If-in-corpus	Δ	Rationale
D1 severity	7.0	7.0	0	Energy shock is incremental, not regime-defining; oil at $102 was substantially priced in
D2 persistence	0.65	0.65	0	OPEC governance shift is structural for energy markets, but UST regime impact is via inflation channel which is pricing-cycle
D3 direction	−0.20	−0.20	0	No direct policy bias change for the FOMC
D4 anchoring	0.75	0.70	−0.05	Higher oil → 5y5y forward inflation expectations more contested; this is the channel UAE/ OPEC actually moves
D5 dispersion	0.55	0.55	0	No new voices on the FOMC; external commentary not voting members
D6 mkt functioning	0.78	0.75	−0.03	Oil-driven volatility imports into Treasury via inflation expectations channel; modest market-functioning hit

Engine effect. ΔH unchanged (D2 and D5 unchanged). Vol_mult = 1.060 (vs 1.050) — same as the Powell scenario, different mechanism. Tail_mult = 1.050 (vs 1.044). Channel-2 lift if UAE/OPEC alone in corpus: 1.055 × 1.060 × 1.050 ≈ 1.174, vs V1.5's 1.157 — approximately +1.5 percentage points of additional channel-2 lift.

14.3 Combined inside-window scenario

If both events had been in the V1.5 corpus, the dimensional profile would have been D1 = 7.5, D2 = 0.75, D3 = −0.15, D4 = 0.65, D5 = 0.70, D6 = 0.75. The deterministic mapping produces ΔH = 0.0531, vol_mult = 1.070, tail_mult = 1.050. Channel-2 lift would have been 1.124 × 1.070 × 1.050 ≈ 1.262 — versus V1.5's 1.157, which is approximately +9.1 percentage points of additional channel-2 lift.

E3 VaR_99 under the inside-window-aware profile would have been approximately $655,000, compared to V1.5's $602,000 — a difference of roughly $53,000 on the $30M book.

The forward backtest framework does not change. The May 11 evaluation compares realized 10-day P&L against the V1.5 thresholds, not against the hypothetical inside-window-aware thresholds in this section. Bar 2 forward discipline requires that the score not be moved during the project window. This section is documentation of inside-window events and a transparent illustration of framework responsiveness — not a revised score.

The inside-window comparison is itself a framework demonstration. A V2 deployment that *did* update scores intra-window — a sliding-window or event-driven scoring regime — would converge toward the inside-window-aware figures over the course of April 28–May 11. V1.5 holds the score fixed by design to make the forward-discipline test clean. V2 work should examine whether intra-window updating improves backtest performance (Open Methodological Questions, Q7 multi-window framework).

Additional inside-window events may be added to this section as they occur; the May 11 backtest update will close the section.

14.4. 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).

Tenor	April 28, 2026 close	May 11, 2026 close	Change
DGS2	3.84%	3.95%	+11.0 bp
DGS5	3.97%	4.07%	+10.0 bp
DGS10	4.36%	4.42%	+6.0 bp

All three tenors moved in the same direction over the window. The curve shifted upward modestly with short-end concentration, producing a small bear-steepening. The 2Y move dominated in basis-point terms, the 10Y move dominated in DV01-weighted P&L terms.

Per-tenor realized P&L on the $30M synthetic UST book ($10M each tenor, long positions).

Tenor	Δy (bp)	DV01 per $10M	P&L
2Y	+11.0	$1,909/bp	−$20,999
5Y	+10.0	$4,501/bp	−$45,010
10Y	+6.0	$8,055/bp	−$48,330
Total			**−$114,339**

The realized 10-day loss for the $30M book is **$114,339**.

Breach matrix vs. published thresholds.

Engine	VaR_95 threshold	Breach?	VaR_99 threshold	Breach?
E1 Gaussian	$307,162	No	$434,424	No
E2 Static MMAR	$367,842	No	$520,246	No
E3 Nexus-adjusted MMAR	$425,452	No	$601,725	No

No engine breached at either confidence level. The realized loss represents 37% of the lowest published threshold (E1 VaR_95) and 19% of the highest (E3 VaR_99). Headroom to the headline E3 VaR_99 threshold is $487,386, approximately 4.3× the realized loss.

Comparison to Section 14.3 Window Watch hypothetical.

Section 14.3 documented that an inside-window-aware re-score incorporating the Powell continuation signal (signaled April 28) and UAE/OPEC dispersion (announced April 28, effective May 1) would have produced an E3 VaR_99 of approximately $655,000 versus the published $601,725. The realized loss of $114,339 falls far below both the published threshold and the Window Watch hypothetical. In

this particular window, the cost of Bar 2 forward discipline (declining to re-score on inside-window evidence) is undetectable in P&L terms — the published thresholds held with substantial margin regardless of whether re-scoring would have raised them.

What this outcome demonstrates and does not demonstrate.

The architecture ran end-to-end against realized market data. The corpus assembled April 24, the score frozen April 27, the engine outputs computed April 28, and the published thresholds produced ahead of the window all held against realized P&L over the subsequent ten trading days. This is face validity of the V1.5 publication: the document committed to specific numbers ahead of time, and those numbers were not breached by realized loss. 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 99% threshold is, by construction, breached approximately 1% of the time; observing no breach in a single window is statistically uninformative about long-run calibration. The 95% threshold is breached approximately 5% of the time; even there, a single no-breach observation tells us little. The outcome is consistent with the architectural claim of Section 7B that the framework's value persists under approximate market efficiency — face validity in a calm window is what V1.5 promised, and what multi-window V2 calibration must extend to (or refute).

What the result does **not** demonstrate, and what V1.5 has been careful throughout not to claim:

- It does not demonstrate that the V1 illustrative coefficients (α=0.25, β=0.20, γ=0.20) are well-calibrated. That is V2 Q1 calibration work, requiring multi-window historical sampling against realized P&L across calm and transition regimes.
- It does not demonstrate that the regime channel adds information beyond price-derived Hurst. That is V2 Q14 falsification work, pre-registered with publication commitment for the falsifying outcome (E3 = E2 in multi-window calibration).
- It does not demonstrate that the framework would have outperformed simpler alternatives. A trivial baseline of "use last month's realized volatility scaled to 10 days" might have produced similar face validity in this window. Distinguishing the framework from such baselines requires V2 work, not single-window evidence.

What V1.5 demonstrates, after the realized outcome lands, is that the architecture runs end-to-end with full transparency, that the published numbers held against realized loss in this window, and that the framework is now an empirical artifact ready for the empirical pressure V2 will apply. The window closes here; the calibration question opens.

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 that V2 must run. Section 7B, also added in this revision, engages the Efficient-Market objection at the architectural level and commits the framework to publishing the falsifying outcome if Q14's null hypothesis (regime channel informationally redundant with price-derived Hurst) is not rejected. The architecture stands as a feasibility demonstration; V2 converts it into a calibrated framework with bounded empirical claims, or surfaces the falsifying evidence transparently and revises positioning accordingly.

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, term-structure-conditional ΔH — split the Hurst adjustment by curve segment so that front-end (Fed-expectations-driven) and long-end (term-premium-driven) responses to a regime score differ. These three V2 deliverables, paired with the cross-LLM robustness check, are the path from feasibility-demonstration to institutional pilot. The non-monotonicity finding (Section 8.2) 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, and the single-window forward test demonstrates feasibility before any calibration claim. 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 (Section 13) is narrowed to reflect that auditable deterministic-mapping-with-LLM-pipelines is converging industry practice as of late 2025 — most directly evidenced by Soleimani’s “LLM-Generated Counterfactual Stress Scenarios” (arxiv 2512.07867, November 26, 2025), which independently commits to “transparent and fully auditable” structured outputs and deterministic post-processing. The novelty claims of the Nexus framework are revised to foreground what remains genuinely distinctive: the multifractal pairing, the anchored ordinal severity rubric, and the conflict-resolution rule. The deterministic-mapping commitment alone is no longer claimed as novel.
2. Section 8.2 (severity adjacency and the non-monotonicity finding) adds a falsification-test caveat acknowledging that the dimensional profiles assigned to historical anchors are plausible reconstructions rather than independently derived. The empirical test that would falsify the architectural claim — rescoring each historical anchor against its contemporaneous corpus using the same six-dimension methodology — is named as the top V2 priority (Open Methodological Questions, Q11).
3. Three new methodological questions are added to the Open Methodological Questions appendix: Q11 (independent derivation of historical anchor dimensional profiles), Q12 (Hurst exponent window robustness), and Q13 (channel orthogonality and double-counting ablation). The V2 priority ranking is reshuffled; Q11 is now priority 1, ahead of coefficient calibration.

Section 14 (Events Inside the Project Window) is added as forward-application content — applying the framework methodology to two events that broke inside the project window (Powell continuation as Federal Reserve Governor; UAE departure from OPEC) without updating the V1.5 score. Bar 2 forward discipline holds the score fixed; Section 14 documents what the score-time information set excluded and demonstrates framework responsiveness without breaking discipline.

We thank Gemini and Grok (independently prompted as adversarial reviewers) for surfacing the issues this revision addresses. The architecture stands; the framing is more honest about boundaries.

Addendum — May 12, 2026. This revision incorporates three additions made after the original V1.5 launch and before public closure of the project window:

1. Section 7B (“Why Not Just Use Prices? The Efficient-Market Objection”) is added immediately after Section 7. It engages the structural objection that price series already incorporate the textual information the framework reads, defending V1.5’s value on auditability, structured-narrative consolidation, transition-period responsiveness, and replicability grounds, and committing the framework to falsifiable empirical testing of the residual informational claim in V2.

2. Q14 (“Does the regime channel add information beyond the price-derived Hurst signal?”) is added to the Open Methodological Questions appendix. It formalizes the EMH-redundancy falsification test that V2 multi-window calibration must run, pre-registering three outcomes with explicit publication commitment for the falsifying outcome (Outcome 3: regime channel informationally redundant; framework’s forecasting-edge claim refuted; positioning revises to auditability-only). The V2 priority ranking is updated to include Q14 as priority 3.

3. Section 14.4 (“Realized Outcome — May 11, 2026 Close”) is added as the closing entry of Section 14. It documents the realized 10-day P&L of the $30M synthetic UST book over the April 28 → May 11 project window, the breach matrix against published thresholds, and the comparison to the Window Watch hypothetical scenarios discussed in Section 14.3. The realized loss of $114,339 fell below all six published VaR thresholds. The closing entry is explicit about what a single-window no-breach observation does and does not demonstrate: face validity of the V1.5 publication is established; calibration validation and EMH-redundancy adjudication are V2 work.

Additions 1 and 2 are textual scaffolding; they do not modify the framework architecture, the engine, the rubric, or any V1.5 numerical result. Addition 3 reports a single realized data point; it does not adjust any V1.5 commitment made before the window opened. All three additions preserve V1.5’s discipline of distinguishing what the architecture demonstrates (feasibility, auditability, replicability under Bar 2 forward discipline) from what V2 must demonstrate empirically (calibration validity, channel orthogonality, EMH-residual information content, cross-LLM agreement).

Prior to public deployment, the three additions and the consolidated document were submitted to Gemini and Grok for a second-round adversarial review. Their feedback converged on three substantive areas that this final pre-deployment revision incorporates: (a) Section 14.4 was tightened with an explicit statistical-power disclaimer and an explicit process-rigor affirmation, separating operational feasibility (demonstrated) from statistical validation (V2 work), and adding a bridging sentence to Section 7B; (b) Section 7B Response 4 (transition-period concentration) was reframed as an architectural hypothesis to be V2-tested rather than as a claimed empirical property, and a new Response 6 was added covering the latency/replication objection and the Grossman-Stiglitz information paradox (both pointing to the framework’s positioning as institutional risk management infrastructure rather than alpha generation); and (c) Q14 test design was tightened with Expected

Shortfall (ES, the quantile loss function at 99%) committed as the primary evaluation metric with VaR breach counts secondary, three named statistical tests pre-specified (Diebold-Mariano, encompassing, likelihood-ratio), a pre-specified stratified-sampling window protocol, and a quantitative falsification threshold (rejection at $p < 0.05$ with effect size > 5%). All Gemini/Grok feedback-driven changes remain textual; no architectural changes, coefficient revisions, or V1.5 numerical commitments are modified.

End of Detailed Findings. Companion documents: corpus_manifest.json, regime_signal.json, sensitivity_table_cap_saturation.json, sensitivity_table_severity_adjacency.json, engine_results.json, baselines.json. Code: 01_baselines.py, 02_engines.py, 03_charts.py.