Treasury PoC V1.5 - Open Methodological Questions

Contents

Treasury PoC V1.5 — Open Methodological Questions

Companion document to Detailed Findings Score date: April 27, 2026

This document collects the methodological questions that remain genuinely open after V1.5 — questions where the V1.5 evidence is either silent, ambiguous, or insufficient to settle them. Each is a target for V2 work or for adversarial review before institutional adoption.

The intent is not defensive disclosure. It is to make explicit what V1.5 has *not* answered, so a reader does not infer answers V1.5 cannot support.

Q1 — How are the deterministic mapping coefficients calibrated?

V1.5 uses $\alpha = 0.25$ (Hurst-adjustment scaling), $\beta = 0.20$ (vol-multiplier sensitivity to inflation expectations anchoring), and $\gamma = 0.20$ (tail-multiplier sensitivity to market functioning). These are illustrative — chosen to produce engine outputs in a plausible range, not derived from a calibration procedure.

The open question is: what calibration procedure should produce these coefficients? Three candidate approaches:

- **Historical-window minimum-deviation.** Run the architecture retrospectively on N historical regimes with known realized 10-day P&L distributions; choose coefficients that minimize average breach-rate deviation from theoretical (1% at 99% level). This is the classical risk-model calibration approach and inherits its weaknesses: it tunes to the historical sample, may overfit to specific regimes, and assumes the regime distribution is stable.
- **Coefficient-as-prior-and-Bayesian-update.** Treat the V1 coefficients as priors and update them as new windows are scored and back-tested. This is robust to non-stationarity but slow to converge, and requires a principled prior (which is itself a calibration problem).
- **Cross-asset-class consistency.** Calibrate coefficients on Treasury data, then check that the same coefficients produce sensible outputs on credit, FX, or equity data. This forces the calibration to be structural rather than asset-specific. The risk is that no single coefficient set works across asset classes, in which case the architecture needs asset-class-specific calibration.

V2 should pick one and commit. Without a chosen procedure, the coefficients remain illustrative and the framework cannot graduate to production.

Q2 — When does Powell-Williams-Waller dominance break?

The conflict-resolution rule in V1.5 effectively concentrates scoring weight on three voting members: Chair, FOMC Vice Chair (NY Fed), and the dovish dissenter (Waller). Other voting members are bounded: their absence from recent scripted speeches, or speech topics that do not address monetary policy, prevents them from influencing the score.

This works as long as the Chair-Vice-dissenter triangle is informationally dominant. It would break in three identifiable scenarios:

- **A non-voting alternate publishes a regime-defining piece.** If a non-voting Reserve Bank president publishes an analysis that subsequently moves the consensus, the conflict-resolution rule under-weights it because of the voting > non-voting priority. The V1 rule cannot capture "this non-voter said something the voting members are about to repeat."
- **The Chair is silent and a junior voter leads.** During Powell's pre-FOMC blackout, the dovish dissenter's voice has the highest scripted weight by recency. If Miran (or his successor) becomes the *only* recent scripted voice, the rule would push toward a dovish-biased composite even when the broader committee remains hold-biased. V1.5 happens to be in this neighborhood (Powell silent post-March-30).
- **Succession transition.** If Powell leaves and Warsh assumes Chair, the rule must transfer Chair-weight to Warsh immediately. If Warsh has not yet given scripted Chair-level remarks, the most-recent-scripted-Chair-voice is stale or absent. The V1 rule does not handle "the Chair just changed and has not yet spoken at Chair level."

V2 should consider weighting decay (the most recent scripted voice from each member, with weight decaying as time passes), explicit handling of blackout periods, and a transition-protocol for chair handoffs.

Q3 — Why is ΔH term-structure-agnostic?

V1.5 applies the same ΔH to 2Y, 5Y, and 10Y. This is wrong in any economically meaningful sense: front-end Treasuries respond primarily to Fed-funds-path expectations, while long-end Treasuries respond to term premium, fiscal conditions, and structural inflation expectations.

Why was V1.5 implemented this way? Because the regime signal is a single scalar pair (severity + dispersion + persistence + …), and the deterministic mapping does not yet resolve which dimension affects which segment of the curve. A more sophisticated mapping would split:

- D3 (policy direction bias) primarily into the front-end ΔH
- D4 (anchoring) primarily into the long-end ΔH
- D2 (persistence) into both
- D6 (market functioning) into a curve-segment-specific tail multiplier

The open question is whether such a split would empirically improve the engine outputs. It might also introduce coefficient-explosion (now there are α_2Y, α_5Y, α_{10}Y, β_2Y, β_5Y, β_{10}Y, etc.). The V2 trade-off is more dimensionality vs more identifiable structure.

A reasonable V2 first step: split into front-end (≤2Y) and long-end (≥10Y) only, with the belly inheriting an interpolation. This gives one extra coefficient pair without explosion.

Q4 — Is dispersion measurement well-formed?

D5 (dispersion across sources) is currently an LLM-emitted scalar in [0, 1]. The conflict-resolution rule maps the corpus voices to weighted positions, and dispersion captures the residual disagreement after weighting.

Two open questions:

- **Is the residual a meaningful regime feature or a noise artifact?** If genuine residual dispersion correlates with future realized volatility, dispersion is a regime feature and belongs in the deterministic mapping. If it is essentially noise from the conflict-resolution rule, it is double-counting and should be removed. The right test is: does dispersion (D5) carry incremental information beyond severity (D1) and persistence (D2)? V1.5 cannot answer this with a single window.
- **Should dispersion drive vol or tail rather than ΔH?** V1.5 uses dispersion to multiplicatively boost ΔH (via the dispersion_factor = 1 + dispersion). An alternative architecture would have dispersion drive the tail multiplier (high disagreement → wider tails) without affecting the central scaling. Both choices are defensible; V1 picked one without empirical evidence for the choice.

V2 should run an ablation: compare engine outputs with dispersion driving ΔH (current), dispersion driving tail multiplier (alternative), and dispersion removed entirely (baseline). Differences across these three would tell us whether the dispersion channel is information or noise.

Q5 — The non-monotonicity finding: feature or bug?

The severity adjacency table shows that under anchor-character-consistent dimensional profiles, ΔH is *not* monotonic in severity: anchor 7 produces higher ΔH than anchor 8.

This is presented in the detailed findings as a feature — the engine responds to dimensional structure, not headline severity. This presentation is correct, but it carries a non-trivial implication that should be tested.

The implication is: severity is being used as a *labeling* device, not as a *driver* of engine output. A purely severity-headline scorer would mis-rank regimes. But a scorer who agrees on the dimensional structure but disagrees on the severity label would produce *the same* engine output. This is fine if dimensions are observable independent of severity. It is concerning if scorers tend to score dimensions to be consistent with their severity prior — in which case the severity label is the carrier signal and the dimensions are post-hoc justification.

The empirical test for this: does scorer-on-scorer dimensional agreement *exceed* scorer-on-scorer severity agreement? If so, dimensions are independently observable. If not, severity is the carrier and dimensions are confound.

V1.5 cannot answer this because it has only one scorer. V2 should run the same corpus through two or three different LLM scorers (or different prompt variations on the same LLM) and measure dimensional agreement vs severity agreement. If dimensional agreement is materially higher, the architecture's claim is supported.

Q6 — Does the architecture produce different outputs across LLMs?

Bar 2 forward discipline is "single LLM, single pass." V1.5 was scored on a single LLM. This raises the question: would a different LLM (different model family, different training cutoff, different prompt formulation) produce a meaningfully different score?

The Bar 2 forward design is silent on this. Bar 3 (the production target) explicitly addresses it via multi-LLM ensemble. But V1.5 cannot say whether the score is robust to LLM choice.

A V2 robustness check: score the same corpus with three or more LLMs (e.g., Claude family, GPT family, open-weights model). Measure cross-LLM dispersion in (a) severity, (b) dimensional sub-scores, (c) verbatim excerpts chosen, (d) engine outputs after deterministic mapping. If cross-LLM dispersion exceeds the ±0.5 single-LLM confidence band by a meaningful margin, single-LLM scoring is inadequate even at Bar 2.

The honest framing for V1.5: we have set a self-consistent replicability bar (±0.5 on the same LLM with the same corpus and rule). We have not validated that this bar holds across LLMs.

Q7 — Multi-window forward test design

A single-window test is feasibility-only. Calibration requires N >> 1 windows. Three design choices for the multi-window test:

- **Non-overlapping monthly windows.** Score on the first business day of each month, project 10 trading days forward, backtest realized P&L. Twelve windows per year. Modest power; clear separation between observations.
- **Rolling weekly windows.** Score every Monday, project 10 trading days. Roughly 50 windows per year. More power but observations are not independent (overlapping projection windows correlate).
- **Event-triggered windows.** Score only around FOMC meetings, CPI releases, employment reports. Eight FOMC meetings per year plus 12 CPI plus 12 employment ≈ 30 windows. High signal-to-noise but small sample.

The right answer depends on the question being asked. For calibration of breach rates, monthly non-overlapping is the cleanest. For event-response measurement (does the architecture handle major announcements correctly?), event-triggered is right. For year-round risk-officer use, rolling weekly is what production demands.

V2 should commit to all three on different time scales: rolling weekly for risk-officer use, monthly non-overlapping for calibration, event-triggered for event-response evaluation.

Q8 — When should the Fractal Circuit Breaker engage?

The Nexus framework includes the Fractal Circuit Breaker concept: a structural mode-change in the engine that triggers when severity crosses a threshold (proposed: severity ≥ 9 with persistence ≥ 0.85 and market functioning ≤ 0.40). Below the threshold, the engine continuously parameter-adjusts. Above, it engages dysfunction-mode tail multipliers, corner-solution VaR (e.g., switch to historical max-drawdown over the past N years), and disables the deterministic-mapping continuity assumption.

V1.5 does not exercise the breaker — the scored regime (7.0, persistence 0.65, market functioning 0.78) is well below the proposed threshold. The open question is what the breaker should do mechanically when it engages, and whether the threshold should be the proposed (9, 0.85, 0.40) or different.

The historical anchor for the breaker is March 23, 2020 (anchor 10 in the rubric). At that moment, continuous risk-model parameter adjustment was inadequate; only structural mode-change captured the regime. The breaker is meant to formalize this.

V2 should specify the breaker behavior in detail (which engine internals change, what the corner-solution VaR is, how it transitions back to continuous mode) and should validate the threshold against historical March 2020 data and any other anchor-10-class events.

Q9 — Comparison against sentiment-based approaches

The Detailed Findings argue (Section 6.2) that the Nexus signal differs structurally from sentiment analysis. This argument is presented qualitatively. It has not been validated quantitatively.

The proper comparison: run a sentiment-based scoring approach on the same corpus, compute its implied engine inputs, and compare engine outputs. If sentiment-based scoring produces outputs that are functionally equivalent (correlation > 0.9 with the Nexus engine output across multiple windows), the structural difference is real but operationally moot. If sentiment-based outputs differ materially, the structural difference is operationally consequential.

V1.5 has not run this comparison. V2 should.

A reasonable comparator is: a scalar sentiment score applied multiplicatively to baseline σ, with sentiment score derived from a sentiment-classifier on the same corpus. Compare its 10-day VaR to E3 across multiple windows. If V2 finds sentiment is “good enough,” that is itself a meaningful finding — it would suggest the Nexus framework’s added complexity is not justified for this asset class and time horizon. The framework’s design claim is that the complexity *is* justified; the empirical test is V2 work.

Q10 — How are inflation expectations measured?

D4 (inflation-expectations anchoring) is the most consequential dimension for the vol multiplier. V1.5 uses LLM judgment of corpus excerpts (Powell statements, IMF assessments, market-implied measures cited in news commentary) to land on 0.75.

The open question is: should D4 be replaced by a direct market measure?

- **Five-year five-year forward breakeven** is the standard market-implied measure of long-run inflation expectations. It is observable daily. It would remove LLM judgment from this dimension entirely.

- **University of Michigan survey expectations** is the standard household-survey measure. Less timely (monthly) but captures a different population.
- **TIPS-implied breakevens at multiple horizons.** Captures the term structure of expectations.

The case for replacing D4 with a market measure: it removes a degree of LLM judgment and uses a measure with known statistical properties.

The case against: LLM-emitted D4 captures verbal-commitment evidence (Powell saying "we will do what it takes") that market measures cannot. These are different kinds of anchoring evidence.

V2 should consider a hybrid: D4 = (market measure) × (LLM verbal-commitment multiplier), with both inputs preserved in the audit trail. This keeps the auditability while reducing LLM-judgment dependence.

Q11 — Independent derivation of historical anchor dimensional profiles

The severity-adjacency analysis in Detailed Findings Section 8.2 reports that under "anchor-character-consistent" dimensional profiles, ΔH is non-monotonic in severity — the central architectural claim that dimensional structure dominates headline severity in driving engine output. The dimensional profiles for historical anchors (Sept 2024 first cut: D2 0.55, D3 −0.50, D4 0.85, D5 0.30, D6 0.92; Dec 2018 pivot hike: D2 0.55, D3 +0.30, D4 0.85, D5 0.40, D6 0.65; etc.) are described as "anchor-character-consistent" and are plausible, but they are not independently derived from contemporaneous text using the same six-dimension methodology applied to V1.5.

This raises a legitimate adversarial concern: if the historical profiles were chosen to produce dimensional structures consistent with the architectural claim being demonstrated, the non-monotonicity finding is post-hoc, not structural.

The falsifying test is direct. For each historical anchor in the sensitivity table, assemble a contemporaneous corpus matching the V1.5 corpus assembly principles (FOMC statement, SEP and Chair press conference if applicable, FOMC minutes when released, voting-member speeches in the relevant window, IMF/OECD assessments, market-functioning evidence). Score each anchor against its corpus using the same six-dimension methodology and the same conflict-resolution rule. Compare the resulting dimensional profiles to the values currently in Detailed Findings Table 8.2.

Three possible outcomes:

1. *Derived profiles match table values within ±0.10 per dimension.* The architectural claim survives empirical test. The non-monotonicity finding is structural rather than constructed.

2. *Derived profiles materially differ but the rank-ordering of ΔH across anchors is preserved.* The specific values in the V1.5 table need updating, but the central claim (dimensional structure dominates severity headline) survives.
3. *Derived profiles produce a monotonic ΔH-vs-severity relationship.* The architectural claim is falsified. Section 8.2 must be rewritten as "the framework's response to historical anchors is monotonic; the non-monotonicity in V1 illustrative profiles was a numerical artifact, not a structural feature."

V2 must run this test before any institutional pilot. It is the single most consequential robustness check the framework faces.

Q12 — Hurst exponent window robustness

The headline V1.5 engine runs use H_24m (rescaled-range Hurst on a 24-month window). The document defends this choice on the grounds that 24 months captures multi-regime persistence (cuts cycle into hold) that 12 months cannot. The defense is reasonable but is not stress-tested against alternative window lengths.

A skeptical reviewer can fairly object that picking a window which delivers "structural persistence" is data snooping — the choice may reflect the analyst's narrative preference rather than an inherent property of the data.

V2 should run H at five window lengths: 12, 18, 24, 36, and 60 months. For each, recompute E2 and E3 and report the resulting VaR_95 and VaR_99 figures. The robustness check has three pass conditions:

1. *E1→E2 channel-1 lift is monotonic in H.* Expected if H itself increases monotonically with window. A monotonic relationship means the engine response is structural rather than artifact.
2. *E1→E2 lift is bounded and stable across windows.* If a 12-month H produces +5% lift while a 60-month H produces +40% lift, the headline +19.8% in V1.5 is window-dependent and the document overstates magnitude precision.
3. *E2→E3 channel-2 lift is approximately stable across windows.* The text-aware channel should be roughly invariant to historical-window choice because it scales the Hurst delta, not the base. If channel-2 lift varies materially with window, the channel and base are not orthogonal.

V2 should report all three conditions and either confirm robustness or document the magnitude sensitivity.

Q13 — Channel orthogonality and double-counting ablation

The two-channel decomposition (E1→E2 pure MMAR self-similarity at +19.8%; E2→E3 text-aware regime channel at +15.7%) is presented in Detailed Findings Section 9.2 as cleanly separable. The document acknowledges in Section 9.5 that σ_30 already inherits regime stress from price action. A skeptical reviewer reasonably objects that the text-aware channel may double-count information already present in σ_30, leading to overestimation when both are used.

The orthogonality test is an ablation. Run E3 four ways:

1. *Full E3 with σ_90:* the V1.5 baseline at $601,725 VaR_99.
2. *E3 with σ_30 substituted for σ_90:* if the text-aware channel is orthogonal, this should produce roughly E3_σ30 = E3_σ90 × (σ_30/σ_90), preserving channel proportions. If the text-aware channel double-counts σ_30 stress, the substitution will produce a smaller-than-proportional lift.
3. *E3 with vol_mult disabled (set to 1.0):* isolates the ΔH and tail_mult contributions. Compare to V1.5 1.157 channel-2 lift. The arithmetic difference attributable to vol_mult was 1.050 in V1.5; if it disappears cleanly under disablement, vol_mult is orthogonal to base σ.
4. *E3 with tail_mult disabled (set to 1.0):* isolates ΔH and vol_mult. Tail_mult was 1.044 in V1.5; same orthogonality test.

A fifth run with dispersion factor disabled (set to 1.0, neutralizing D5's effect on ΔH) tests whether the dispersion sub-score carries incremental information beyond the severity-and-persistence channel.

The purpose is not to discredit any channel; it is to *quantify* channel overlap. A finding that "vol_mult contributes 4 percentage points beyond what σ_30 already captures" is more credible than "vol_mult contributes 5%." V2 should produce per-channel orthogonality metrics so that institutional reviewers can audit channel attribution rather than accept the headline decomposition on faith.

Q14 — Does the regime channel add information beyond the price-derived Hurst signal?

Section 7B (added in this revision) addresses the Efficient-Market objection at the architectural level: the framework's value is defended on auditability, structured-narrative consolidation, transition-period responsiveness, and replicability grounds, not on a claim of forecasting edge that survives semi-strong-form market efficiency. The empirical question — whether the LLM-emitted regime channel adds information not already present in the price-derived Hurst exponent — is left explicitly unresolved at the V1.5 stage. This methodological question formalizes the falsification test that V2 must run.

The hypothesis to be tested, stated precisely: under a multi-window calibration design, does E3 (Nexus-adjusted) produce a tighter empirical loss distribution against realized P&L than E2 (static MMAR) does? E2 carries the price-derived Hurst signal alone. E3 adds the text-aware regime channel on top. If the regime channel is informationally redundant with what the price-derived Hurst already captures, E3 should not systematically outperform E2 in calibration. If it does add information, the calibration improvement should be measurable.

The test design has three components.

First, window selection. V2 must run on at least 60 historical 10-day windows. Window selection follows a pre-specified stratified-sampling protocol: windows are stratified by regime anchor (anchored against the V1.5 severity rubric: severity ≥6 = transition; severity 3-5 = elevated; severity ≤2 = calm) and by FOMC cycle phase (pre-meeting, meeting-inside, post-meeting). Approximately one-third of windows are drawn from the transition stratum, one-third elevated, one-third calm. Pure random sampling would under-weight transitions because transitions are rare; the stratified protocol pre-commits to the sample composition before any backtesting begins, eliminating the data-snooping concern about post-hoc window selection.

Second, evaluation metric. The **primary metric is Expected Shortfall (ES) — the quantile loss function at 99% confidence, also known as Conditional VaR.** ES is a proper scoring rule that distinguishes between models even when neither breaches the threshold, addressing the statistical power concern that VaR breach counts alone are underpowered at 60+ windows (where only ~0.6 breaches per engine are expected under correct 99% calibration). The **secondary metric is VaR breach count** (the traditional measure) plus the empirical breach rate vs. nominal level. The primary E3-vs-E2 comparison is on ES; the secondary comparison is on breach counts and rates. Both must show E3 advantage for the regime channel to be considered informational; ES advantage alone is the credible primary signal.

Third, statistical test specification. Three named tests, pre-specified, run in sequence:

1. **Diebold-Mariano forecast accuracy test** comparing E3 vs E2 ES values across the test windows. Null hypothesis: E3 = E2 in forecast accuracy. Rejection at $p < 0.05$ with effect size > 5% improvement supports the architectural claim.
2. **Encompassing test** for E2 forecasts encompassing E3 forecasts (and vice versa). If E2 encompasses E3 (E3 adds no information not already in E2), the regime channel is redundant.
3. **Likelihood-ratio test on VaR backtest scores** as a robustness check on the breach-count secondary metric, comparing E3 vs E2 against realized loss distributions.

Fourth, quantitative equivalence threshold. The null hypothesis "E3 = E2 in calibration" is *not* rejected — and Outcome 3 (regime channel informationally redundant) is concluded — when the Diebold-Mariano test fails to reject the null ($p \geq 0.05$) AND the ES improvement effect size is below 5%. This quantitative threshold pre-commits to a falsification criterion before the test begins, eliminating wiggle room in the interpretation of "not systematically better."

Fifth, null-hypothesis discipline. The test is structured as falsification, not validation. The null is "E3 = E2 in calibration." The framework's claim is rejected if E3 does not improve on E2 systematically by the criteria above. A failure to reject the null does not merely defer the architectural claim — it actively weakens it.

Three possible outcomes:

1. *E3 systematically outperforms E2 on ES (primary metric) with effect size > 5%, Diebold-Mariano $p < 0.05$, and the advantage concentrated in transition-period windows.* The architectural thesis is empirically supported. The regime channel adds information beyond price-derived Hurst, and the contribution is most pronounced in the regime where the framework is designed to be most useful.
2. *E3 outperforms E2 on ES with effect size > 5% and Diebold-Mariano $p < 0.05$, but the advantage is approximately uniform across calm and transition windows.* The regime channel is informational but does not behave as architecturally hypothesized. Section 7B's transition-period framing must be revised; the framework retains forecasting value but loses one of its architectural claims.
3. *Diebold-Mariano fails to reject the null at $p \geq 0.05$, OR ES improvement effect size is below 5%, OR the encompassing test shows E2 encompasses E3.* The EMH-redundancy null is not rejected. The regime channel is, on this evidence, informationally redundant with the price-derived Hurst signal. The framework's institutional value (auditability, replicability, structured-narrative consolidation) persists; its forecasting-edge claim does not. V2 must publish this result honestly and the framework's positioning must be revised accordingly.

The third outcome is the falsifying outcome. The framework commits to publishing it if it occurs. This commitment, combined with the pre-specified quantitative falsification threshold (rejection at $p < 0.05$ with effect size > 5%), is what makes Q14 a genuine empirical test rather than a rhetorical hedge.

Summary of V2 priorities

The questions above are not equally pressing. A V2 priority ranking based on what most threatens V1.5's claims if left unresolved:

1. **Q11 — independent derivation of historical anchor dimensional profiles.** This is the falsifying test for the central architectural claim. If Q11 fails, no other V2 work compensates.

2. **Q1 — coefficient calibration.** Without this, the framework cannot graduate to production.
3. **Q14 — does the regime channel add information beyond price-derived Hurst?** The EMH-redundancy falsification test. If Q14 fails (E3 does not outperform E2 systematically), the framework's forecasting-edge claim is empirically refuted and positioning must revise to auditability-only.
4. **Q12 — Hurst window robustness.** Tests whether the headline magnitudes are window-dependent; addresses the data-snooping concern directly.
5. **Q13 — channel orthogonality ablation.** Quantifies double-counting between σ_30 stress and text-aware channels; the basis for credible channel attribution.
6. **Q5 — non-monotonicity sensitivity to scorer agreement.** This validates whether dimensional structure is an independently observable feature.
7. **Q6 — cross-LLM robustness.** This validates the Bar 2 forward discipline.
8. **Q7 — multi-window forward test.** This converts feasibility evidence into calibration evidence.
9. **Q3 — term-structure dependence in ΔH.** This addresses the most obvious V1 simplification.
10. **Q9 — comparison against sentiment.** This addresses the most predictable adversarial challenge.
11. **Q4 — dispersion measurement formalism.** This is internal to the architecture.
12. **Q2 — Powell-Williams-Waller dominance.** This is a robustness concern; less urgent in calm periods.
13. **Q10 — inflation expectations measurement.** This is a refinement, not a foundational issue.
14. **Q8 — Fractal Circuit Breaker design.** This matters only when the breaker engages, which is rare by construction.

A V2 work package addressing items 1–7 would convert V1.5 from a feasibility demonstration into a calibrated, validated framework ready for institutional pilot.

End of Open Methodological Questions.