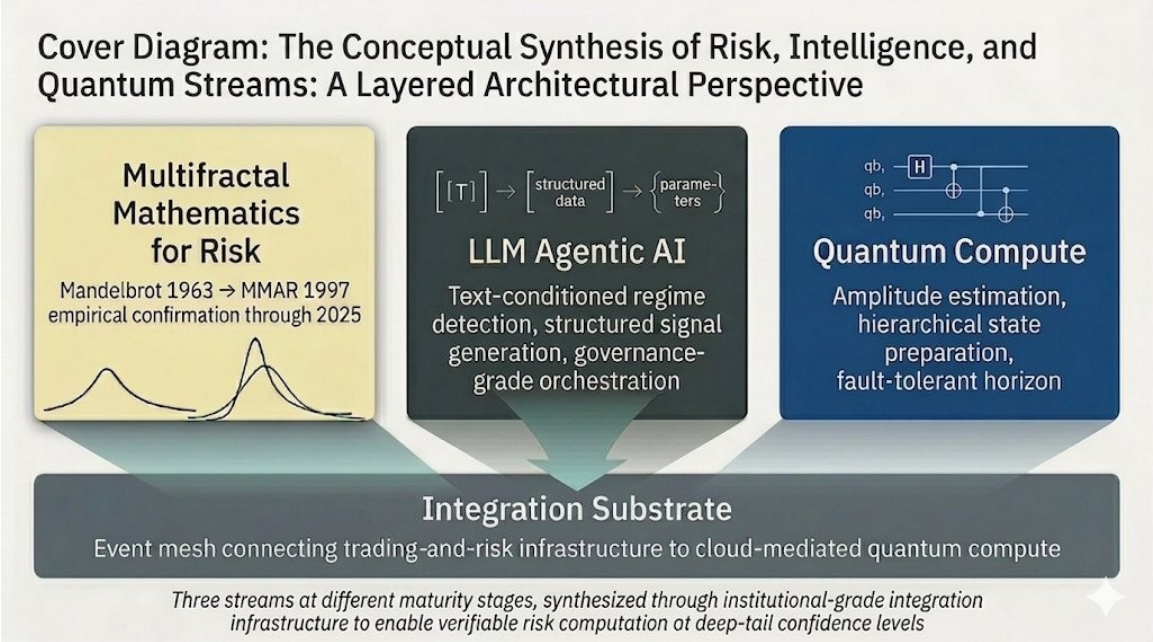
Quantum-Augmented Risk Management for Capital Markets

An Architectural Perspective — Executive Summary

Multifractal Tail Risk, LLM Orchestration, and the Compute Stack of the Next Decade



Anantha Padmanabhan

Capital Markets AI · capmarkets-ai.com

May 2026

This is the executive summary of a companion detailed write-up. It is self-contained: it states the thesis, the architecture, and the research agenda in compressed form. The full architectural treatment, literature engagement, and open-question detail are available in the companion paper, "Quantum-Augmented Risk Management for Capital Markets: An Architectural Perspective." Like the full paper, this is a research-agenda document: it proposes an architectural framework and identifies the open research questions the framework raises; it does not present validated results.

The thesis

Capital markets risk management has a structural mismatch at its core. The models that determine regulatory capital, stress tests, and tail-risk preparation overwhelmingly assume Gaussian or near-Gaussian return distributions, while six decades of empirical work — beginning with Mandelbrot (1963) — show that financial returns are fat-tailed, long-memory, and regime-dependent. The mismatch matters most precisely where risk matters most: at the deep tail, the confidence levels at which capital adequacy and systemic resilience are actually decided.

This write-up proposes an architecture that addresses the mismatch by synthesizing three intellectual streams standing at different maturity stages. Two are production-ready today. Multifractal mathematics, mature since the late 1990s, provides a generative model whose empirical fit to financial returns is strongest exactly at the deep tail. LLM agentic AI, production-deployable since 2024–2025, provides a bounded mechanism for reading the qualitative textual information — central bank communications, policy shifts, auction results — that classical regime-detection methods cannot process. The third stream, quantum compute, is research-stage; the architecture positions it at a backend-agnostic compute layer to absorb its potential advantage if and when commercial viability arrives, while operating fully on the first two streams today.

The framework is therefore best understood not as three co-equal components assembled in parallel, but as a working two-stream classical architecture, deployable now, deliberately structured to intercept a research-stage third stream at the point where that stream would add value. The contribution is architectural synthesis and research-direction articulation, not numerical demonstration — and its firmer half (the classical two-stream architecture and the integration infrastructure that makes it deployable) stands independent of how the quantum questions resolve.

The three streams and their asymmetric maturity

The streams differ not only in maturity but in architectural role, and conflating the two has been a persistent source of confusion in the quantum-finance literature.

Multifractal mathematics is the foundation. The Multifractal Model of Asset Returns (Mandelbrot, Fisher, and Calvet 1997) builds a loss distribution through a recursive multiplicative cascade. Its empirical contribution is concentrated at the deep tail: at Value at Risk 95%, Gaussian and multifractal models often agree; at Value at Risk 99% and beyond, the multifractal model captures probability mass that Gaussian models systematically underestimate. This is the empirical reason the foundation matters — the tail confidence levels where models actually drive capital decisions are exactly where the Gaussian-versus-multifractal distinction becomes consequential. The choice of a multifractal cascade over rough-volatility models — which lead on volatility-surface fitting and option pricing — is deliberate: on deep-tail VaR and Expected Shortfall the cascade family (in its forecasting-grade Markov-

switching form) remains competitive to dominant, the roughness-versus-multifractal question is itself unresolved, and the cascade's scale-invariant structure is precisely what the quantum question turns on. It is used here as the cleanest representative of that structure, not as a claim to be the last word.

LLM agentic AI is the adaptation layer, deliberately bounded. Classical regime-detection methods — hidden Markov models, regime-switching econometrics — operate well on quantitative inputs and are not replaced. The LLM fills the gap they cannot: reading unstructured text and producing a categorical regime signal drawn from a small, fixed, auditable set of classifications. A hardcoded classical policy rule, not the LLM, then maps that category to model parameters. This separation is deliberate. It confines the probabilistic, hallucination-susceptible step to a discrete, inspectable classification, while the quantitative parameterization downstream remains deterministic and hand-recomputable — a governance property essential for any model-risk-managed pipeline. The text-to-categorical step itself is well-trodden rather than novel: an active 2024–2026 literature extracts structured signal from central-bank communication and validates it against realized rates and volatility (Ahrens et al. 2025; Silva, Moriya, and Veyrune 2025; Collodel 2025). What that literature does not do — and the white space this architecture occupies — is compose such a signal with a forward, regime-conditional multifractal tail-risk engine.

Quantum compute is a future backend, honestly research-stage. Goldman Sachs publicly scaled back its quantum effort in April 2026 after concluding that practical portfolio-optimization applications required resources far beyond current hardware — on the order of eight million logical qubits against the fewer than one hundred available today. That is best read as a fault-tolerant resource threshold, and the lesson generalizes: any quantum advantage this architecture would rely on is asymptotic — a claim about scaling in target precision, not about today's machines — realizable only past a threshold, of unresolved location, where the better scaling overtakes error-correction overhead. Comparing today's quantum to today's classical hardware answers the wrong question. Herbert (2021) proved that standard state-preparation methods eliminate the theoretical quadratic speedup of quantum amplitude estimation for the very distribution classes finance routinely assumes — the "loading bottleneck." Those classes are log-concave; the multifractal cascade is not, which is why the loading question is open for this architecture rather than closed by Herbert's result. The architecture does not claim near-term quantum advantage; its quantum value, where any exists, would lie at the sampling layer rather than the loading layer, and the write-up is explicit that this remains genuinely open.

The architecture in brief

The architecture comprises seven layers. A governance layer composes audit artifacts. An LLM orchestration layer performs the bounded text-to-categorical-signal adaptation. An integration substrate — an event mesh — connects on-premises trading and risk infrastructure to cloud-mediated compute. A risk engine constructs the multifractal cascade and produces regime-conditional loss distributions. A backend-agnostic tail-sampling solver routes computation to classical (GPU) or quantum backends based on tail depth and structural compatibility — today every route resolves to the classical backend, and the architecture is designed so that future

quantum availability requires no change to the layers above the solver. Below the solver sit the compute manager and the cloud-resident hardware primitives.

The design intent is that the production-ready streams operate today, and the quantum layer can be absorbed at a single, well-defined point — the tail-sampling solver — when it matures, without re-architecting the system around it.

The distinctive contribution: the integration substrate

The published quantum-finance literature systematically elides the operational layer that connects financial-institution infrastructure to quantum compute. Financial institutions will not own quantum hardware — the economics, the operational expertise, and the hardware roadmap volatility all argue against it; they will access it through cloud services. That fact has architectural consequences the literature largely ignores, and addressing them is the write-up's most concrete contribution.

The integration substrate must connect three substantively different compute environments: on-premises trading systems with sub-millisecond latency budgets, mixed on-premises and cloud risk systems, and queue-managed cloud-resident quantum compute with latencies measured in seconds. Three design implications follow. Latency is asymmetric, so a classical fast path serves pre-trade and intraday risk while a decoupled, asynchronous, batch-scheduled path serves the quantum-augmented analytics — there is no near-real-time quantum routing, and the architecture does not claim one. Data residency is handled by on-premises pre-aggregation: raw trade data is reduced to risk-factor exposures before any transmission to cloud compute, keeping the most sensitive data within the institution's regulatory boundary. And quantum execution cost, which scales with circuit shots, makes routing an economic decision taken on a batch schedule rather than per-trade.

This is the layer that distinguishes a deployable architecture from a research demonstration, and it is where the bulk of real engineering, regulatory engagement, and institutional risk management actually happens.

The research agenda, compressed

The write-up identifies a structured set of open questions. Three are decisive.

Does the multifractal cascade admit a genuine quantum advantage, and if so, where? This resolves into a loading question (can the self-similar cascade structure be prepared efficiently on a quantum state?) and a sampling question (even if the distribution is classically representable, do the specific deep-tail expectations resist efficient classical contraction?). The honest current expectation is that financial cascades likely fall on the classically simulable side at the loading layer, which makes the sampling-layer question the decisive one. The advantage the architecture would capture therefore exists only in the conjunction — a cascade cheap to prepare yet carrying a deep-tail functional expensive to contract classically — and whether multifractal cascades occupy that narrow region is the open question. Notably, the experiment that would settle this is itself classical — a numerical tensor-network simulation requiring no quantum hardware to begin.

How is verifiability preserved under quantum shot noise? Quantum execution cannot be reproduced to the decimal the way a fixed-seed classical Monte Carlo can; hardware calibration drifts between runs. The audit-trail unit must shift from exact reproducibility to statistical-distribution-level reproducibility under documented execution conditions — and whether model-risk frameworks such as SR 11-7 and TRIM will accept that shift is itself unresolved.

What is the regulatory path? Model-risk acceptance of quantum-augmented models will likely lag technical readiness by years. The write-up treats the demonstration sequence that would persuade supervisors as an open question rather than a solved problem.

Framing, access, and what comes next

The architecture is agnostic across gate-model quantum backends — trapped-ion, superconducting, neutral-atom, photonic, silicon-spin — and the research questions it identifies apply across vendors; this gate-model agnosticism does not extend to quantum annealing, which is a different computational paradigm rather than an alternative backend. It does not propose a commercialization roadmap, does not claim current quantum advantage, and does not relitigate the settled empirical case for fat-tailed returns.

One downstream consequence is worth surfacing. Because the architecture reaches quantum compute through cloud-mediated access rather than on-premises hardware, and because the integration substrate is the equalizer, institutional-grade tail-risk modeling capability need not remain confined to tier-one technology firms. As commercial viability emerges, the framework is in principle accessible to a far broader range of institutions — mid-tier banks, regional broker-dealers, pension funds, sovereign wealth funds, insurance reserving operations, and central clearing counterparties. This is not the primary contribution, but it is a genuine consequence of the design.

The full architectural treatment — the layer-by-layer detail, the literature engagement, the complete set of open questions, and the honest accounting of what is demonstrated versus what remains open — is available in the companion detailed write-up. The author welcomes engagement, critique, and collaboration on the open questions it raises.

Acknowledgments

This document and its companion were prepared with assistance from Claude (Anthropic) as a thinking partner across outline development, structural framing, and prose composition, and underwent cross-model adversarial review by Gemini (Google DeepMind) and Grok (xAI). The author retains responsibility for intellectual content, citation choices, and architectural positions.

PoC Results — Empirical Update on the Loading-Layer Question

This section is appended to the executive summary on the basis of empirical work completed after the body of this document was finalized. It reports the headline finding of the proof-of-concept measurement identified above — the classical numerical tensor-network experiment described in the research-agenda section — and updates the architectural posture stated earlier in this document. The full result, with all methodology, scope, and limitations, is reported as Part 2 of the companion combined paper.

The question, restated. The body of this executive summary articulates the architecture's central open question: does the multifractal cascade admit a quantum loading advantage? The honest current expectation stated above was that financial cascades likely fall on the classically simulable side of the boundary at the loading layer — making the sampling layer the decisive location for any quantum advantage. The classical numerical experiment that would settle this expectation was identified, in compressed form, in the research-agenda section.

The experiment. A direct numerical measurement of the bond dimension of the matrix product state representation of the multiplicative cascade was conducted, varying cascade depth K from 4 to 20 and the intermittency parameter across both lognormal multipliers (six values of λ^2) and conservative-binomial multipliers (six values of p), with 30 realizations per cell. The experiment is classical throughout — it uses the singular value decomposition, not quantum hardware — while the question it addresses is about quantum advantage.

The result. Bond dimension saturates the maximum possible value 2^(K/2) at every nonzero intermittency tested, at every cascade depth up to K = 20, for both multiplier laws, at every accuracy tolerance $\varepsilon \in \{10^{-2}, 10^{-3}, 10^{-6}, 10^{-10}\}$. The exponential growth rate matches the saturation rate to within 1–4% across all eleven nonzero-intermittency cells. The result holds at every bipartition cut, not just the centre — a deductive consequence of which is that the balanced binary tree tensor network with natural qubit ordering also saturates. The result is independent of cascade construction: conservative-renormalized and canonical non-renormalized cascades agree within 1.3% at $\varepsilon = 10^{-3}$.

What this updates in the architecture. The earlier expectation — that the cascade was likely classically simulable at the loading layer, making sampling the only decisive location for quantum advantage — is not supported by the measurement under the matrix product state ansatz and, by deductive extension, the balanced binary tree tensor network with natural qubit ordering. The loading-layer question therefore reopens as a candidate location for quantum advantage. The architecture now carries two open quantum-advantage questions rather than one: at the loading layer, whether a tree tensor network respecting the cascade's generative tree (rather than the dyadic-scale ansatz tested) compresses more efficiently, and whether polynomial-depth quantum preparation circuits exist; at the sampling layer, the question previously identified. Neither is settled; both are now empirically constrained rather than merely conjectured. The two-stream classical architecture and the integration substrate that makes it deployable today remain unchanged and operationally complete.

Scope and what is not claimed. The measurement scopes its claims precisely: matrix product state and balanced-natural-ordering tree-tensor-network representations only (not all classical representations); the cascade measure μ (not the full MMAR price process $X(t) = B_H(\theta(t))$); the canonical lognormal and conservative-binomial constructions (not all multifractal models); cascade depths $K \leq 20$ (the asymptotic claim is deferred to a tensor-cross-interpolation extension); and the L2 truncation norm (alternative norms untested). Establishing exponential classical representation cost is a necessary but not sufficient condition for quantum advantage at this distribution class: the measurement removes one classical escape route without establishing a working quantum alternative. The companion paper (Part 2) provides the full scope and limitations discussion.