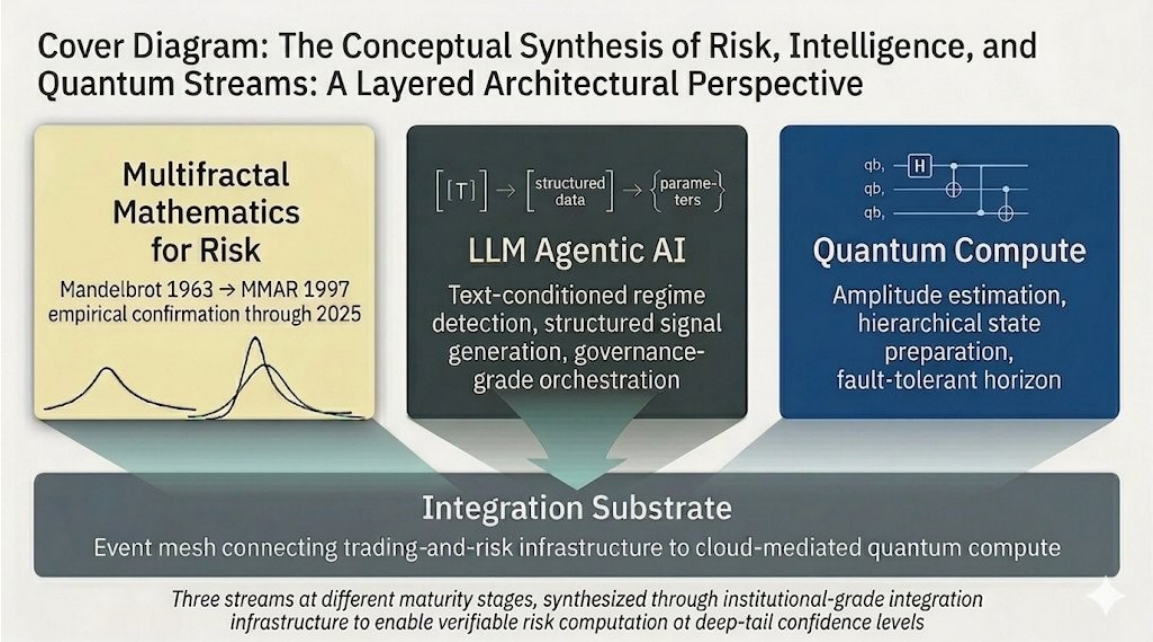
Quantum-Augmented Risk Management for Capital Markets: An Architectural Perspective and Empirical Loading-Layer Measurement

Anantha Padmanabhan *Founder and Principal Researcher, Capital Markets AI* capmarkets-ai.com *June 2026*



This document combines two complementary works on quantum-augmented risk management for capital markets, presented together as a single citable artifact.

Part 1: An Architectural Perspective. A three-stream architectural framework synthesizing multifractal mathematics, LLM agentic AI, and a backend-agnostic compute layer designed to absorb quantum hardware advantages if and when they mature. Two streams are production-ready today; the third is positioned for future absorption through an integration substrate connecting on-premises trading infrastructure to cloud-mediated compute.

Part 2: Empirical Loading-Layer Measurement. The empirical companion. A direct numerical measurement of whether the multifractal cascade — the foundation of Part 1's risk-modeling layer — is efficiently representable as a matrix product state. This measurement was identified in Part 1 §8 Q1 as the experiment that would settle one of the architecture's central open questions; Part 2 reports its execution and result.

Combined Abstract

This document combines an architectural framework for quantum-augmented capital-markets risk management with an empirical measurement that updates one of the framework's central open questions. **Part 1** articulates a three-stream synthesis — multifractal mathematics (production-ready), LLM agentic AI (production-ready), and quantum compute (research-stage, structurally positioned for future absorption) — integrated through an event-mesh substrate connecting on-premises trading infrastructure to cloud-mediated compute. **Part 2** reports a direct numerical measurement of whether the multifractal cascade underlying Part 1's risk-modeling layer is efficiently representable as a matrix product state. The measurement was specified in Part 1 §8 Q1 as the experiment that would settle the loading-layer question; its result — that bond dimension saturates the maximum possible value 2^(K/2) at every nonzero intermittency tested up to K = 20, for both lognormal and conservative-binomial multiplier laws, and by deductive extension also for the balanced binary tree tensor network with natural qubit ordering — reopens the loading-layer question as a candidate location for quantum advantage. The architecture now carries two open quantum-advantage questions rather than one. The two-stream classical architecture is deployable today regardless of how either quantum question ultimately resolves.

Table of Contents

Part 1 — An Architectural Perspective

Abstract

This paper articulates a research agenda. It proposes an architectural framework and identifies the open research questions the framework raises; it does not present validated results.

This paper proposes a seven-layer architecture for capital markets risk computation that operates today on two production-ready streams — multifractal mathematics, mature since the late 1990s, and LLM agentic AI, production-deployable since 2024–2025 — while positioning a third, research-stage stream (quantum compute, with commercial pilots emerging) at a backend-agnostic tail-sampling solver to absorb its potential advantage if and when commercial viability arrives. The architecture advances a conditional technical hypothesis — that the multifractal cascade structure may admit efficient hierarchical quantum state preparation, offering a possible escape from the loading bottleneck (Herbert 2021) for fat-tailed distributions — but treats it as open research whose value is contingent on the questions in Section 8, not as a load-bearing claim; the paper argues that any durable quantum advantage for this architecture would lie at the sampling layer rather than the loading layer, and that even this is genuinely open. The paper's firmer contribution is the integration substrate connecting on-premises trading infrastructure to cloud-mediated quantum compute — the operational layer that the published quantum-finance literature systematically elides — and the articulation of a structured research agenda, including the specific numerical experiments that would resolve its central open question. The architecture is agnostic across gate-model quantum backends and the research questions it identifies apply across vendors.

---

1. Why the question matters now

This paper addresses how to synthesize two production-ready streams — multifractal mathematics and LLM agentic AI — with a third, research-stage stream, quantum compute, through institutional-grade integration infrastructure. The three streams stand at different maturity stages, and that asymmetry is central to the architecture rather than incidental to it: the system operates fully on the first two today and is structured to absorb the third if and when it matures.

Multifractal mathematics for risk modeling has been mature since the late 1990s. Mandelbrot's 1963 work on cotton prices established fat-tailed return distributions empirically; the Multifractal Model of Asset Returns (Mandelbrot, Fisher, and Calvet 1997) provided a rigorous cascade-based generative framework; six decades of subsequent empirical work have confirmed fat tails across asset classes, time scales, and market regimes (Plerou and Stanley 1999, Cont 2001, Bouchaud and Potters 2003, Calvet and Fisher 2008). This stream is practitioner-deployable today.

LLM agentic AI reached production deployability in 2024–2025. Text-conditioned regime detection, structured signal generation, and governance-grade orchestration are operational at financial institutions today, with documented audit-trail discipline and cross-model adversarial review practices supporting institutional use. This stream is practitioner-deployable today.

Quantum compute remains research-stage. Goldman Sachs publicly scaled back its quantum effort in April 2026 after internal analysis concluded that practical applications for portfolio optimization required computational timelines on the order of millions of years on current hardware (Bloomberg 2026). Herbert (2021) proved that standard Grover-Rudolph state preparation eliminates the theoretical quadratic speedup of quantum amplitude estimation when applied to the log-concave distributions that finance applications routinely assume. Pistoia et al. (2021), writing from JPMorgan's quantum research team, observe that no end-to-end application of quantum machine learning with exponential speedup over its classical counterpart has been discovered. This stream is not practitioner-deployable today for institutional-scale risk computation.

The architectural question is therefore not whether quantum compute is currently delivering institutional-scale advantage — it is not — but how to synthesize a risk-computation architecture that operates today on production-ready streams (multifractal mathematics, LLM orchestration) while positioning the quantum layer at the right architectural point to absorb its potential advantage when commercial viability arrives. This paper proposes such a synthesis: a seven-layer architecture in which LLM agentic AI provides text-to-structured-signal adaptation, a multifractal cascade engine produces regime-conditional loss distributions, and a backend-agnostic tail-sampling solver routes computation to classical or quantum backends based on tail depth and structural compatibility. The paper articulates structural reasons that the Multifractal Model of Asset Returns may be a natural target for hierarchical quantum state preparation if the underlying state-preparation question can be solved, identifies what is currently demonstrated

versus what remains open research, and engages with the institutional infrastructure realities that determine whether such a synthesis is operationally deployable.

1.1 Scope: portfolio risk, not derivative pricing

The framework here is concerned with portfolio-level risk management — Value at Risk, Expected Shortfall, regime-conditional tail estimation, regulatory capital — not with derivative pricing. The two activities answer different epistemic questions. Derivative pricing is naturally addressed through relative-value replication off liquid market instruments, where vol-surface-trading desks already operate close to this paradigm even when academic literature describes Black-Scholes-Merton closed-form pricing. Risk management requires probability distributions over outcomes for underlying risk factors, where distribution shape (Gaussian, fat-tailed, multifractal) materially changes the answer.

The derivative-pricing line in quantum finance (Stamatopoulos et al. 2020, Chakrabarti et al. 2021) is cited here for its algorithmic contributions — amplitude estimation as a primitive, end-to-end resource analysis under fault-tolerant assumptions — rather than for its foundational framework. The architecture proposed here engages with risk distributions directly, sitting upstream of the Black-Scholes-Merton-tradition debate.

---

2. The compute stack today: CPU → GPU → Quantum

The honest comparison between classical and quantum compute for capital markets risk management begins by acknowledging what GPU acceleration actually delivers today. Goldman Sachs reports approximately a 25% improvement in real-time data processing from GPU integration into trading systems. JPMorgan Chase reports approximately a 30% reduction in calculation time for risk computations. Academic benchmarking on portfolio Value at Risk with 4,000 risk factors shows up to 169-times speedup of GPU over CPU (Pagès, Wilbertz, and Wilkens 2018). These are production deployments at tier-one financial institutions and represent the mature commercial reality of accelerated risk computation in 2026.

But GPU acceleration, however large its constant factor, preserves the convergence regime of classical Monte Carlo. Sampling at scale on parallel hardware reduces wall-clock time proportional to silicon count, but the underlying $1/\sqrt{N}$ convergence law that governs Monte Carlo error estimation does not change. At Value at Risk 95%, this matters little: a few thousand paths suffice. At Value at Risk 99.99% or Expected Shortfall 99.5%, where multifractal tail mass concentrates, the number of paths required for tight confidence intervals grows past 10^8 even with variance reduction, and parallel scale-out continues to scale linearly with path budget while convergence continues to scale with $1/\sqrt{N}$.

Quantum amplitude estimation, where it works, changes the convergence regime itself. The asymptotic convergence is $1/N$ rather than $1/\sqrt{N}$ (Brassard et al. 2002), translating to a quadratic reduction in the number of samples required to achieve a given precision. This is not a constant-factor improvement, however large; it is a different scaling law. The practical realization of this advantage is constrained by the loading-bottleneck question discussed in Section 3 and by hardware fault-tolerance horizons that remain multi-year, but the underlying mathematical structure is genuinely distinct.

This is the central honest framing: GPU shifts the cost of feasibility downward; quantum, where it works, shifts the boundary of feasibility outward. Four specific qualitative differences support this framing.

The convergence law itself. Classical Monte Carlo's $1/\sqrt{N}$ error scaling and quantum amplitude estimation's $1/N$ error scaling produce fundamentally different cost curves at deep-tail confidence levels. At 1% precision target, classical Monte Carlo requires approximately 10^4 paths; quantum amplitude estimation requires approximately 10^2. At 0.01% precision target, classical requires approximately 10^8; quantum requires approximately 10^4. The gap widens monotonically as the tail deepens.

Rare-event sampling structure. Classical Monte Carlo of rare events requires importance sampling — biasing the proposal distribution toward where the tail mass is presumed to lie, then reweighting samples. This requires knowing where the tail is. For Gaussian models the answer is straightforward. For fat-tailed multifractal regime-dependent processes, the tail location is precisely what one is trying to discover; using importance sampling for that task is structurally

backwards. Quantum amplitude amplification biases sampling toward rare events structurally rather than through prior knowledge.

Memoryful processes. Aghamohammadi, Crutchfield, and Mahoney (2018) demonstrated that quantum memory advantage for rare-event sampling on classical memoryful stochastic processes ranges from polynomial to exponential, with the advantage diverging in certain regimes where classical memory grows without bound while quantum memory reaches a finite limit. Long-memory financial processes such as fractionally integrated GARCH and multifractal random walks (Bacry, Delour, and Muzy 2001) share the memoryful, broadly stationary character to which this result applies. The MMAR cascade is a less clean fit, and the distinction is worth being precise about: the Aghamohammadi–Crutchfield–Mahoney result concerns stationary, causal processes with a one-dimensional temporal memory structure (the ε-machine / hidden-Markov family), whereas an MMAR cascade is a scale-indexed multiplicative object whose dependence runs across scales rather than along a single causal timeline. The result is therefore suggestive motivation for the broader long-memory class, not a formal guarantee that transfers to scale-indexed cascades — a representation-versus-computation distinction Section 4.1 takes up again.

Where the quantum candidates are narrow and specific. The quantum-advantage candidates are not “non-separable workloads” in general. Modern GPU clusters handle high-dimensional, non-separable workloads — Quasi-Monte Carlo with Sobol sequences, copula transformations as massively parallel tensor-core operations — with great efficiency, and on those they are not merely adequate but dominant. The narrow region where a structural quantum candidate exists is the intersection of three conditions: deep-tail confidence levels, fat-tailed long-memory processes, and high-precision requirements. At that intersection two properties become relevant that constant-factor acceleration does not address — amplitude estimation’s 1/N convergence, where the binding constraint is precision at the deep tail rather than throughput, and the quantum memory advantage for rare-event sampling on memoryful processes. Outside that intersection, classical compute dominates regardless of dimensionality or separability.

These differences together support the qualitative-versus-quantitative framing. GPU and quantum are not different points on the same compute axis; they sit on different axes. Where the precision requirement is modest, GPU dominates and quantum cannot compete, regardless of dimensionality or separability. Where the workload involves deep tails at high precision on memoryful, fat-tailed processes, the asymptotic argument tilts toward quantum — subject to the practical constraints addressed in the next section.

---

3. Where quantum advantage matters in risk management

Quantum compute is currently positioned for narrow but meaningful application classes in risk management, with substantial limitations from current hardware that determine the deployment horizon. This section is explicit about both.

In-scope use cases with structural fit between application and quantum primitive:

- Deep-tail Value at Risk and Expected Shortfall on fat-tailed distributions, via quantum amplitude estimation (Woerner and Egger 2019)
- High-dimensional portfolio optimization with cardinality and integer constraints, via QAOA on appropriately structured Hamiltonians
- Correlated default modeling and Clayton-family copula sampling, via quantum sampling
- Reverse stress testing — combinatorial search for parameter combinations that breach a target loss threshold
- Path-dependent options under non-Gaussian path measures, via quantum walks (De Backer et al. 2024, 2025)

Out-of-scope use cases where quantum offers no structural advantage:

- Static FRTB Standardized Approach calculations and BCBS standardized capital methodologies, which are closed-form or near-closed-form
- Basic parametric Value at Risk under Gaussian assumptions
- Accounting-flavored risk reporting and most regulatory submissions where the binding constraint is data quality, not compute throughput
- Most low-dimensional pricing tasks already adequately handled by analytical or near-analytical methods

The loading bottleneck and NISQ-era reality. Herbert (2021), publishing as a researcher at Cambridge Quantum (now Quantinuum), proved a result that has reshaped the quantum-finance landscape. The standard Grover-Rudolph state preparation method, applied to the log-concave distributions that finance applications routinely assume (Gaussian, exponential, log-normal), does not deliver the quadratic speedup of quantum amplitude estimation. The classical integration cost embedded in the state preparation step cancels the quantum advantage in the sampling step. This is not a hardware limitation that better quantum hardware will fix; it is a structural result about how loading-then-sampling composes for that class of distributions.

It is worth being precise about the scope of this no-go, because the precision is what keeps the door open. Herbert's result is established specifically for log-concave distributions — Gaussian, exponential, log-normal and their relatives — prepared via Grover-Rudolph. The multifractal cascade is not log-concave; its defining feature is exactly the fat-tailed, non-log-concave mass on which the foundation of this architecture rests. The no-go therefore does not transfer to it by assumption: whether an analogous obstruction applies to cascade-structured fat-tailed distributions is an open question, not a closed one. This is not a claim that the cascade escapes the bottleneck — Section 4.2 is candid that it most likely does not, falling on the classical-simulability side rather than the quantum-advantage side — but the narrower and more important point that the cascade sits outside the regime where the no-go has been proven, so

the loading question for multifractal distributions must be settled on its own terms rather than inherited from the log-concave case.

The field has responded with several workarounds: variational state preparation via quantum generative adversarial networks (Zoufal, Lucchi, and Woerner 2019), matrix product state preparation for low-entanglement distributions (Iaconis, Johri, and Altman 2024), quantum signal processing approaches with improved gate-count scaling (Stamatopoulos and Zeng 2024), and direct generation approaches that sidestep loading by producing the distribution through quantum-walk dynamics (De Backer et al. 2024). The quantum-walk approach warrants a specific caveat for this architecture: it generates distributions dynamically through quantum evolution rather than loading a pre-specified distribution, which means its output state does not compose directly with the canonical amplitude-estimation circuit this architecture uses at L2. Using quantum-walk-generated distributions as an input to amplitude estimation would require a state-transformation step that is itself unstudied and potentially expensive — the naive route, re-extracting a loadable representation from the dynamically evolved state, could erode or erase the very speedup the approach is meant to deliver; the composition is an open question, not a solved path. None of these workarounds has cleanly resolved the loading question for the generic distribution case. Section 4 of this paper proposes a structural hypothesis specifically for multifractal cascade distributions.

Current quantum hardware adds additional constraints. Two-qubit gate error rates need to drop below approximately 10^{-6} for consistent quantum advantage in amplitude-estimation-based computations at realistic problem sizes (threshold analyses developed in Stamatopoulos et al. 2020 and subsequent literature for amplitude-estimation primitives generally), and current hardware does not deliver this. Chakrabarti et al. (2021) provide the canonical end-to-end resource estimate for amplitude-estimation-based quantum computation under fault-tolerant assumptions; their threshold analysis places practical quantum advantage at problem sizes and hardware qualities significantly beyond NISQ-era capabilities.

The Goldman scale-back as field-reality anchor. Goldman Sachs publicly scaled back its quantum effort in April 2026. Their internal analysis concluded that practical investment applications were unattainable on current hardware: for the specific portfolio-optimization problem they investigated, researchers estimated a requirement on the order of eight million logical qubits, against the fewer than one hundred logical qubits available on today's machines (Bloomberg 2026). That figure is best read not as a statement about runtime but about a fault-tolerant *resource threshold* — logical qubits already fold in the error-correction overhead — and it concerns portfolio optimization, a different algorithm from the tail-sampling use this architecture contemplates. The general lesson nonetheless transfers, and bears stating plainly: any quantum advantage of the kind this architecture would rely on is *asymptotic* — amplitude estimation's quadratic improvement over classical Monte Carlo is a claim about scaling in target precision, not about current hardware — and realizing it requires crossing a fault-tolerant resource threshold at which the better scaling overtakes the constant-factor overhead of error correction and magic-state distillation. Where that threshold sits for deep-tail risk functionals is unresolved, and on present roadmaps it is not met. This is consistent with the Herbert critique and the Chakrabarti threshold analysis, and it is precisely why this paper treats quantum as a

research-stage backend: the open question is the threshold's location, not the advantage's existence in principle. Comparing today's quantum hardware to today's classical compute therefore answers the wrong question.

The architectural implication is straightforward. A capital markets risk architecture proposed in 2026 cannot reasonably claim near-term commercial advantage from quantum compute. What it can do — and what this paper proposes — is position the quantum layer at the right architectural point so that the system operates today on classical compute and is structurally prepared to absorb quantum advantage when commercial viability arrives. The architectural value is in the positioning, not in current deployment.

---

4. Where multifractal mathematics fits

A structural observation about the field motivates the foundation this section develops. The quantum-finance literature exhibits a tooling-induced bias toward log-concave and Gaussian-derivative distributions — not because practitioners believe asset returns are log-concave (the empirical case for fat tails has been settled since Mandelbrot 1963, and fat-tailed frameworks from Lévy processes to rough volatility are well established in classical practice) but because the log-concave regime is where quantum amplitude-estimation speedups are provable and state-preparation methods (notably Grover-Rudolph) are best developed. The tools selected the distributions, rather than the data selecting the tools. One consequence is that the empirically better-supported multifractal models remain comparatively under-explored *specifically in the quantum context* — not because which fat-tailed model is operationally best is a settled question (it is a legitimate live debate) but because the quantum-finance subfield has had structural reasons to look elsewhere. The empirical foundation the architecture itself builds on, by contrast, is straightforward and long-settled, and it is where this section begins.

Multifractal mathematics provides the mathematical foundation for the architecture's risk computation: a fat-tailed, long-memory, regime-dependent generative framework with empirical fit to financial returns across asset classes and time scales. The Multifractal Model of Asset Returns (Mandelbrot, Fisher, and Calvet 1997) builds the loss distribution via a recursive multiplicative cascade. At each cascade level, a random multiplier from a fixed distribution (often log-normal, log-Poisson, or binomial) modifies the inherited measure on each sub-interval. The cumulative effect across cascade levels produces a distribution with multifractal scaling properties: moments scale as power laws of aggregation level with non-linear exponents, fat tails persist across time scales, and volatility clustering emerges naturally from the multiplicative structure.

A word on the choice of MMAR specifically, since a natural objection is that the multifractal model has been superseded. The objection conflates two tasks. For describing volatility *dynamics* — fitting the implied-volatility surface, pricing options, forecasting realized volatility — rough volatility (Gatheral, Jaisson, and Rosenbaum 2018) has become the leading framework of the past decade, and on that task the objection is fair. But the task here is deep-tail loss estimation, Value at Risk and Expected Shortfall at the confidence levels that drive capital adequacy, and on *that* task the multifractal cascade family remains competitive to dominant: across the asset classes studied, Markov-switching multifractal models — the estimable, forecasting-grade reformulation of the cascade idea (Calvet and Fisher 2008) — repeatedly rank among the models that cannot be outperformed in superior-predictive-ability tests of VaR and ES, while rough-volatility methods are largely absent from the tail-risk forecasting literature. Moreover, the apparent rivalry between roughness and multifractal long memory is itself unresolved: the two are near-observationally-equivalent descriptions of the same volatility persistence, and which the data truly supports remains an open econometric question. The cascade is therefore not a superseded choice but a live one — and it is the right one *for this paper* for a further, specific reason: its recursive multiplicative, scale-invariant structure is

precisely the property the quantum representability question of Section 4.1 turns on. MMAR is used here as the cleanest representative of that cascade structure; the Markov-switching reformulation shares the same structure and is the member to reach for when forecasting performance, rather than structural clarity, is the priority.

MMAR's empirical contribution is concentrated at the deep tail. At Value at Risk 95%, Gaussian and multifractal models often produce similar numbers; at Value at Risk 99% and beyond, particularly at Expected Shortfall 99.5% and Value at Risk 99.99%, the multifractal model captures probability mass that Gaussian models systematically underestimate. This is the empirical reason a multifractal foundation matters for risk computation: the tail confidence levels at which models actually matter for capital adequacy, reverse stress testing, and tail-event preparation are precisely where the distinction between Gaussian and multifractal becomes consequential.

This empirical asymmetry establishes the first structural alignment with quantum compute. Quantum amplitude estimation's advantage over classical Monte Carlo sharpens with tail depth. The combination of a model whose contribution is concentrated at the deep tail and a compute primitive whose advantage sharpens with tail depth is a natural pairing — though, as discussed in Section 3, this pairing requires solving the loading bottleneck before it can be operationally realized.

4.1 Two structural reasons MMAR aligns with quantum compute

The MMAR–quantum coupling rests on two structural reasons, but before turning to them it is worth stating the axis on which the coupling actually rests, because it is easy to misidentify. The wager of this architecture is that the multifractal cascade's hierarchical, scale-invariant structure — not any notion of discreteness — is the property that may align with efficient quantum representation. The relevant contrast is not continuous versus discrete: quantum computation is itself built on continuous Hilbert space and continuous amplitudes, so a discreteness argument would be misplaced. The contrast is scale-invariant and hierarchical versus smooth and single-scale. Hierarchical, self-similar states are precisely the class that tensor-network methods can sometimes represent with low bond dimension, and that representability is the bridge between classical simulability and quantum state preparation. The two structural reasons below are distinguished by whether they apply specifically to MMAR or to a broader class of processes that MMAR happens to belong to.

First: general quantum advantages for fat-tailed long-memory processes — properties MMAR possesses. Two general quantum-advantage results apply to the class of processes MMAR belongs to. Quantum amplitude estimation provides quadratic convergence improvement over classical Monte Carlo for any sufficiently structured distribution, with the advantage most pronounced at deep tails. Aghamohammadi, Crutchfield, and Mahoney (2018) demonstrated polynomial-to-exponential quantum memory advantage for rare-event sampling on classical memoryful stochastic processes, with the advantage diverging in certain regimes. A caution travels with this result: a memory or representation advantage — fewer qubits needed to *encode* the process — is not the same as a query or runtime advantage in *computing* a deep-tail expectation over it, and only the latter is what the architecture ultimately needs. The two are

adjacent but distinct, and the gap between them is precisely the sampling-layer question taken up in 4.2. MMAR's fat-tailed and long-memory properties place it within the class of processes these results apply to, but the results are not MMAR-specific. The first structural alignment is real but general.

Second: MMAR-specific hypothesis — cascade structure may admit efficient hierarchical quantum state preparation. The scale-invariant multiplicative cascade structure of MMAR exhibits hierarchical recursion that is structurally similar to the recursion exploited by matrix product state preparation, Tucker tensor decomposition, and other hierarchical amplitude protocols (Iaconis, Johri, and Altman 2024; Ran 2020; Mori, Mitarai, and Fujii 2024). The hypothesis is that the multiplier distribution and cascade depth might be loaded onto a quantum state via a circuit whose depth scales with cascade levels rather than with the discretization resolution of the final distribution. If this construction works, it would constitute a structural escape from the Herbert (2021) loading bottleneck for the specific distribution class most relevant to fat-tailed risk modeling. But escaping the loading bottleneck is necessary, not sufficient — and the distinction is the crux of the whole quantum case. As 4.2 develops, a cascade cheap enough to prepare is for that very reason likely cheap enough to *simulate* classically, so the loading result on its own does not establish any advantage; it may even argue against one. The operative hypothesis of the architecture is therefore not the loading claim alone but the conjunction of two claims: that the cascade is efficiently preparable (the loading claim made here) *and* that the deep-tail functional computed over it resists efficient classical contraction even when the underlying state has low bond dimension (the sampling claim of 4.2). Only that pairing — loading-tractable yet sampling-advantaged — leaves a durable role for quantum hardware; either claim without the other collapses the case.

This two-part conditional hypothesis — loading-tractable and sampling-advantaged together — is the central open *question* the paper poses, not its central *claim*. The firmer contribution, developed in Sections 6 and 7, is the deployable two-stream classical architecture and the integration substrate that makes it operational; the quantum coupling is the research bet layered on top of that, and it is open research, not a demonstrated result. The technical questions include whether cascade-level correlations keep bond dimension bounded as cascade depth grows, whether the resulting quantum state represents the distribution-over-cascade-outcomes rather than a single realization, whether the construction yields a normalized and well-defined state, and how the prepared state composes with downstream amplitude estimation. None of these has been worked out in the published literature. Section 8 lists these as the central open research questions of the architecture.

Note: The numerical experiment proposed in this section — measuring how the bond dimension of the matrix-product-state representation of the cascade scales with cascade depth — has now been executed. The empirical result is reported as Part 2 of this combined document, and updates the conjecture stated in §4.1 and §4.2 of efficient hierarchical preparation along the dyadic-scale chain. See the postscript to §4.2 below for the substantive empirical update; see Part 2 for full methodology, scope, and limitations.

4.2 The classical tensor-network objection

A natural objection to the cascade-loading hypothesis deserves substantive engagement. If MMAR cascade structure admits efficient hierarchical state preparation via matrix product state methods, the same hierarchical structure may also admit efficient classical simulation via tensor networks on GPU clusters. If true, this would eliminate the quantum advantage: any computation possible on the quantum-prepared state would also be possible on the classically-simulated state, and the GPU pathway would presumably be cheaper and more accessible than quantum hardware for the foreseeable future.

The technical question that determines whether this objection bites concerns the bond dimension of the matrix product state representation of MMAR cascade outputs. States with bounded bond dimension are classically tractable; states whose bond dimension grows without bound require quantum resources to represent at scale. The boundary between these regimes is well-studied in the tensor-network literature (Schuch et al. 2008, Verstraete, Murg, and Cirac 2008). MMAR cascades have not been characterized within this framework in the published quantum-finance literature, and an honest assessment must concede the likely answer: financial distributions, even multifractal ones, do not exhibit physical non-local entanglement. The "entanglement" in a tensor-network representation of a financial distribution is a mathematical artifact encoding temporal correlation and cross-asset dependency, not a physical quantum resource. Structured financial cascades therefore probably fall on the low-entanglement, area-law side of the boundary — the classically simulable side. A back-of-the-envelope expectation, pending actual numerical tensor-network simulation of MMAR cascades, is that the cascade-loading hypothesis is more likely to reduce to classical tractability than to demonstrate quantum necessity at the loading layer. (A terminological caution is warranted here: the multiplicative *return* cascade discussed throughout this paper is unrelated to the interbank *network-contagion* cascade studied in the quantum-systemic-risk literature (Aboussalah, Chi, and Lee 2023), which propagates losses across a graph of counterparties rather than across the scales of a single return process — the two share a word, not a mechanism.)

This concession matters less than it first appears, because the architecture's quantum value does not rest on the loading layer. The decisive point is that quantum advantage, where it exists, lives at the *sampling* layer, not the *loading* layer. Even if the MMAR distribution can be prepared or compressed classically via tensor networks, the question of computing a deep-tail expectation over that distribution to high precision is a separate problem, and it is the problem amplitude estimation addresses with its 1/N convergence.

This sampling-layer defense must itself be stated honestly, because it is not unconditional. If a distribution sits on the low-entanglement, area-law side of the boundary, classical tensor-network contraction can in many cases evaluate expectations over the state directly, without a path-dependent Monte Carlo sampler — which means the same low bond dimension that makes loading classically tractable can also make sampling classically tractable. The sampling-layer advantage therefore does not survive automatically; it survives only where the specific tail functional being computed resists efficient classical contraction even when the underlying state has low bond dimension. Whether the deep-tail expectations relevant to capital adequacy fall into that residual class — distributions cheap to represent but expensive to integrate at the tail

— is itself an open question, not a settled result. The honest position is that the quantum-advantage case for this architecture is genuinely open at both layers, and that the cascade-loading hypothesis (Section 4.1, second pairing) and the sampling-layer functional question (Section 8, Q1) are the two technical questions whose resolution determines whether the architecture's quantum component delivers advantage at all. The architecture is structured so that the production-ready classical streams stand on their own regardless of how these questions resolve; the quantum layer is a conditional enhancement whose value is contingent on them, not a settled pillar.

Postscript — empirical update from Part 2. The architectural posture stated above — that financial cascades "probably fall on the low-entanglement, area-law side of the boundary" and that "the cascade-loading hypothesis is more likely to reduce to classical tractability than to demonstrate quantum necessity at the loading layer" — was a back-of-the-envelope expectation pending the direct numerical measurement identified in §8 Q1. Part 2 of this combined document reports that measurement. The empirical finding, for the matrix product state ansatz and (by deductive extension) the balanced binary tree tensor network with natural qubit ordering, is that bond dimension saturates the maximum possible value 2^(K/2) at every nonzero intermittency, every cascade depth tested up to K = 20, for both lognormal and conservative-binomial multiplier laws, at every L2 truncation accuracy $\varepsilon \in \{10^{-2}, 10^{-3}, 10^{-6}, 10^{-10}\}$. The cascade does *not* fall on the area-law side under these representations.

The loading-layer question therefore reopens as a candidate location for quantum advantage, alongside the sampling-layer question discussed above. The architecture now carries two open quantum-advantage questions rather than one. The earlier framing of "loading is probably cheap classically" is not supported under the standard tensor-train ansatz at the cascade depths and tolerances tested. Two questions remain open: whether a tree tensor network respecting the cascade's *generative* tree (rather than the dyadic-scale ansatz tested in Part 2) compresses more efficiently, and whether a polynomial-depth quantum preparation circuit exists for the cascade. The careful scoping of Part 2 (necessary but not sufficient for quantum advantage; MPS and balanced-natural-ordering TTN only; K ≤ 20 brute-force; alternative tree topologies, higher K via tensor cross interpolation, and alternative truncation norms deferred to future work) is preserved; readers seeking full methodology, scope, and limitations should consult Part 2 directly.

4.3 Relationship to prior work

The multifractal-plus-LLM framework that this architecture extends has been developed in two prior working papers covering market, credit, and liquidity risk and institutional risk intelligence respectively (Padmanabhan 2026a, 2026b), and demonstrated as a classical software baseline in the Capital Markets AI Treasury PoC V1.5 (Padmanabhan 2026c). Those works do not address quantum compute. The contribution of the present paper is the extension of that classical framework to engage with the quantum compute layer as a potential L2 backend — articulating the structural reasons for the coupling and the open research questions it raises. The prior work is referenced here as a parallel software baseline, not as validation of the quantum-side claims, which remain open.

These prior works do not propose quantum extensions. The current paper extends the architectural framework to engage with the quantum compute layer as a potential L2 backend, articulating the structural reasons for the coupling and the open research questions it raises.

---

5. The LLM orchestration layer

The LLM orchestration layer in this architecture performs a specific and bounded function: text-to-structured-signal adaptation. Given a corpus of textual inputs that classical regime-detection methods cannot directly read — FOMC meeting minutes, Treasury auction statements, regulatory communications, internal policy documents, monetary policy speeches — the LLM produces a categorical regime signal drawn from a small, fixed set of classifications. That category, not the LLM, is then mapped to a parametric MMAR family by a hardcoded classical rule.

This framing matters and is deliberate. Classical regime-detection methods — hidden Markov models, regime-switching econometrics, statistical clustering on realized volatility and skew — are well-developed and widely used in production risk systems. They are not replaced by the LLM. They operate on numerical and quantitative inputs and produce reliable, verifiable, deterministic regime classifications. The LLM's contribution is specifically for the textual input domain where classical methods do not operate: extracting structured signal from qualitative monetary policy language, central bank communications, geopolitical risk narratives, and similar inputs that carry information classical methods cannot directly process. The LLM provides input-adaptation capability that complements classical regime detection rather than replacing it.

5.1 Why a third layer matters

The multifractal risk engine can operate on quantitative inputs alone, using classical regime-detection methods for parameterization. The LLM orchestration layer adds a specific capability on top of that working baseline: it incorporates qualitative textual information — central bank communications, regulatory policy shifts, monetary policy language — that classical regime-detection methods cannot read. Configuration of model parameters from qualitative inputs, audit-trail composition for regulatory acceptance, and supervisory narrative generation all benefit from an orchestration layer between the application surface where institutional users operate and the compute surface where the engine executes. LLM agentic AI has reached production deployability for this orchestration function in 2024–2025 (Pistoia et al. 2021). The architectural question is what specific role the LLM plays and where it fits — and the answer, developed below, is bounded: it is an input adapter for the textual domain, not a co-equal computational necessity.

5.2 Prior work: LLM-derived signals from central-bank text

The bounded, text-to-categorical role assigned to the LLM here should be read against an active and fast-growing body of work on extracting structured signal from central-bank communication, which has matured rapidly across 2024–2026 and which establishes the textual-adaptation step as well-trodden rather than novel. Ahrens et al. (2025), in the *Journal of Econometrics*, map Federal Reserve speech text to implied macroeconomic-forecast revisions and use them to explain realized volatility and tail risk in Treasury futures and equities through a continuous-time jump-diffusion model. Silva, Moriya, and Veyrune (2025), in IMF Working Paper 25/109, classify some twenty-one million sentences drawn from 169 central banks along four

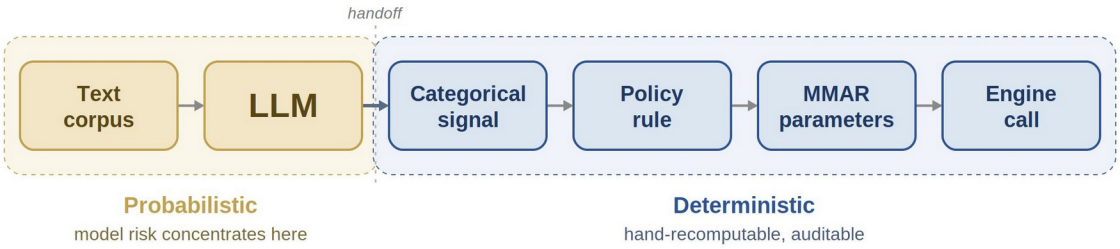
categorical axes — topic, communication stance, audience, and sentiment — and show that forward-looking sentiment predicts future policy and market-based rates while forward-looking risk communication predicts future market volatility. Collodel (2025), at the Central Bank of Malta, has large language models simulate a cross-section of heterogeneous synthetic traders reading ECB press-conference transcripts, using the dispersion of their interpretations to forecast OIS-rate volatility across tenors. The same neighborhood includes work extracting probabilistic policy expectations from Federal Reserve communications (Fernández-Fuertes 2025) and LLM-generated counterfactual macro-stress scenarios feeding Value-at-Risk and Expected-Shortfall assessment (Soleimani 2025). Taken together, these establish that systematic extraction of forward-looking signal from central-bank text — increasingly via LLM and transformer methods — is an active, validated research program with demonstrated linkage to realized market outcomes.

Two of these works bear directly on the design choices made here, and the architecture stands on them rather than apart from them. The IMF study's rationale for a fine-tuned categorical classifier over a general-purpose generative model is explicitly reproducibility, auditability, and cost — commercial LLM outputs being, in their phrasing, non-reproducible and non-transferable across runs — which is the same governance argument that motivates the bounded, deterministic-downstream design developed below; that work is therefore adopted here as a methodological template for the categorical-classification step, not merely cited as adjacent. Collodel's design runs in the opposite direction, and is instructive precisely for that reason: it treats interpretive dispersion across many synthetic readers as the signal, whereas the architecture here deliberately collapses textual interpretation to a single bounded, auditable category. The contrast sharpens the governance trade-off — dispersion is informative but unbounded and difficult to audit, while a fixed categorical output is narrower but fully inspectable — and explains why a model-risk-managed pipeline favors the latter.

What none of this literature does — and what none of it set out to do — is construct a forward, regime-conditional tail-risk engine. These studies extract a text signal and validate it against realized volatility or rates; the signal is the endpoint, not an input to a downstream loss-distribution model. None is multifractal, and none engages a quantum compute layer. Collodel is explicit on the point, naming the use of its synthetic measure as a direct input to a risk model as a downstream extension it does not pursue. The white space this architecture occupies is therefore not the text-to-signal step — which this paper treats as established prior art that it builds on, not as a contribution it claims — but the composition of that bounded signal with a multifractal tail-risk engine, the step to which the remainder of this section now turns.

5.3 Architectural pattern

The orchestration pattern is:



The LLM never touches numerical compute and never emits numerical parameters. Single LLM, single pass, frozen corpus for governance-grade runs. The LLM's output is confined to a small, fixed set of categorical classifications (for example, a discrete regime tier with a confidence flag) — a deliberately narrow, auditable output space. It is important to be precise about where variability enters: the LLM's interpretation of the textual corpus into one of these categories is the probabilistic step — the same corpus could in principle yield a different category — and this is where model risk concentrates. Because the output space is a small discrete set, that risk is bounded and inspectable: a reviewer can enumerate the categories, audit the classification, and challenge it. The mapping from the chosen category to MMAR cascade parameters is then fully deterministic, executed by a published, hand-recomputable classical policy rule rather than by the LLM. This separation is deliberate: it isolates the probabilistic, hallucination-susceptible component (categorical text interpretation) from the quantitative parameterization (a hardcoded rule), so the model-risk surface is confined to a discrete, auditable classification step rather than diffusing through the numerical pipeline. The cascade then produces the regime-conditional loss distribution that flows into the tail-sampling solver at L2.

This pattern is operationally tractable today on classical compute; it has been implemented as a classical software baseline and run forward under pre-commitment, with the LLM signal locked before any realized data in the window exists (Padmanabhan 2026c).

5.4 Governance properties

The orchestration pattern carries governance properties that make it institutionally usable:

- **Bar 2 forward discipline** (developed in Treasury PoC V1.5): the LLM produces its regime signal before any realized data within the forward window exists. Score locked, signal frozen, no parameter adjustment during the evaluation window. This is the discipline that distinguishes forward prediction from post-hoc explanation.
- **Four-question reasoning decomposition** (developed in Treasury PoC V1.5): each LLM run is decomposed into four explicit reasoning questions covering text reading,

signal coherence, deterministic mapping, and forecast versus ground truth. Each question is audited separately.

- **Auditable hand-recomputable mapping**: the deterministic mapping from structured signal to MMAR parameters is published and verifiable. A reviewer can reproduce the parameter set from the signal without re-running the LLM.
- **Cross-model adversarial review** across Anthropic Claude, Google DeepMind Gemini, and xAI Grok (applied in Treasury PoC V1.5 and both SSRN papers): each result is reviewed by independent LLM systems with different training distributions and architectures, surfacing disagreements as candidate errors for human review.

These governance properties address the legitimate concern that introducing an LLM into a model-risk-management pipeline introduces hallucination risk, non-determinism, and audit complexity. The properties above mitigate each risk: Bar 2 discipline addresses retrospective rationalization; four-question decomposition addresses opaque reasoning; hand-recomputable mapping addresses non-determinism downstream of the LLM; adversarial review addresses isolated-model error.

5.5 Quantum-side implications: shot noise as an audit problem

When the L2 tail-sampling solver routes computation to a quantum backend, the result is probabilistic with shot-noise confidence intervals. Classical Monte Carlo is also probabilistic, but the analogy must not be overstated, because quantum execution is materially harder to reproduce retrospectively. A classical Monte Carlo run with a fixed random seed can be re-executed months later on any conforming processor and reproduce the identical Value at Risk to the last decimal. Cloud-mediated quantum execution cannot make this guarantee: quantum hardware calibration drifts, sometimes hour to hour, and the gate error rates, relaxation times, and error-mitigation stacks that shape the shot-noise distribution change between executions. Even with an identical circuit specification and shot budget, a re-execution weeks later — or on a different vendor's hardware — can produce a statistically distinguishable tail estimate. The "calibration snapshot" that documents execution conditions is a historical record of a machine state that no longer exists, not a reproducibility guarantee.

This is a genuine governance escalation, not a solved problem, and the architecture should be honest about it. The audit-trail unit cannot be "Value at Risk exactly reproducible under any execution." It must instead be "Value at Risk plus confidence interval, reproducible at the level of statistical distribution under documented execution conditions, with the calibration snapshot archived in tamper-evident form." Whether model-risk committees and supervisors under SR 11-7 and TRIM will accept distribution-level rather than exact reproducibility is itself an open question (Section 8, Q2). The architecture does not claim to have resolved it; it identifies the relaxation that quantum compute requires and flags the regulatory-acceptance question as unsettled. Acknowledging this honestly is more defensible than presenting quantum execution as classically-equivalent in its auditability, which it is not.

5.6 LLM as text-to-structured-signal adapter

The role of the LLM in this architecture is precisely the role classical regime-detection methods cannot fill: reading unstructured textual inputs and producing structured outputs. The LLM is not

a replacement for hidden Markov models or regime-switching econometrics; those methods continue to operate on quantitative inputs. The LLM is the adapter for the textual input domain. Without it, the architecture cannot incorporate signal from qualitative inputs (Federal Reserve communications, central bank guidance, regulatory policy shifts, geopolitical risk narratives) that materially affect market behavior and that classical methods cannot directly process.

This framing is more bounded than positioning the LLM as a “regime-conditional distribution selector.” The LLM does not select the distribution family; the architectural specification does that. The LLM does not determine the parametric values directly; the deterministic mapping does that. The LLM produces a structured signal from textual input. Other layers handle the consequences.

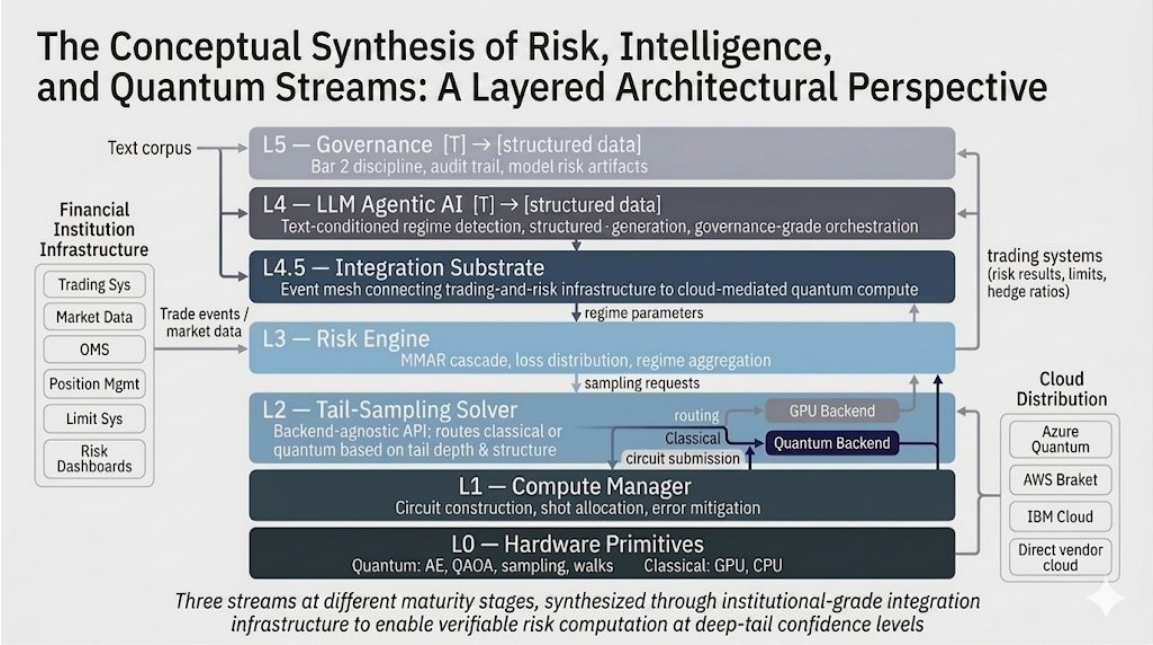
5.7 Prior demonstration

The orchestration pattern described here has been implemented as a classical software baseline and run forward under pre-commitment: a synthetic US Treasury book scored against a single frozen corpus before a ten-trading-day window, with the regime signal and all VaR thresholds locked before any realized data existed (Padmanabhan 2026c). The baseline exercises the cascade's self-similar *scaling* channel (a Hurst-exponent adjustment) together with the text-aware regime adjustment; it does not yet instantiate a full multifractal fat-tailed quantile, which is deferred work. The text-aware channel produced tail estimates materially and auditably different from both a Gaussian baseline and a text-blind MMAR baseline, and the pre-committed thresholds held against the realized outcome. Whether those different estimates are also more accurate than price-only baselines is a separate, multi-window question, pre-registered there as a falsifiable test rather than claimed here. The architectural extension this paper proposes is the routing of the downstream tail-sampling computation to a quantum backend at L2 if and when commercial viability arrives. The orchestration layer above L2 runs on commodity classical CPU hardware today — no GPU or quantum compute is involved; the substitution at the L2 backend is the open architectural question this paper engages with.

---

6. Layered architecture sketch

The layered model described in this section is a proposed design, not a deployed system; the present-tense description that follows characterizes how the architecture is intended to operate rather than asserting a built artifact. With that framing understood, the description is stated directly for clarity. The architecture comprises seven layers, six numbered (L0 through L5) plus an intermediate integration substrate layer (L4.5) that connects the application-layer stack to the financial institution's trading and risk infrastructure.



The layers, top to bottom:

- **L5 — Governance.** Bar 2 forward discipline, audit trail composition, model risk artifacts for SR 11-7, TRIM, MAS, FCA compliance. Composes audit artifacts from all layers below.
- **L4 — LLM Orchestration.** Text-to-structured-signal adaptation. Reads FOMC minutes, auction results, regulatory communications. Produces a categorical regime signal (for example, a discrete regime-tier classification with confidence) drawn from a small, fixed, auditable set of categories. A hardcoded classical policy rule — not the LLM — maps that category to numerical MMAR parameters, so the LLM never emits numerical parameters directly.
- **L4.5 — Integration Substrate.** Event mesh connecting on-premises trading and risk infrastructure (Trading Sys, Market Data, OMS, Position Mgmt, Limit Sys, Risk Dashboards) to the cloud-mediated compute layers below. High-throughput, low-latency, schema-governed, bidirectional. Carries trade events, market data, position state, risk results, and limit feedback.
- **L3 — Risk Engine.** MMAR cascade construction, loss distribution generation, regime-conditional aggregation. Invokes L2 for tail-sampling estimates.

- **L2 — Tail-Sampling Solver.** Backend-agnostic API. Routes computation to classical (GPU) or quantum backend based on tail depth and structural compatibility. Returns Value at Risk or Expected Shortfall estimate with confidence interval.
- **L1 — Compute Manager.** Circuit construction, shot allocation, error mitigation (ZNE, PEC, dynamical decoupling), classical post-processing. SDK layer: Qiskit, TKET, Braket SDK, Cirq depending on backend.
- **L0 — Hardware Primitives.** Cloud-resident; no on-premises quantum hardware assumed. Quantum primitives (amplitude estimation, QAOA, sampling, walks); classical primitives (GPU, CPU). Cloud distribution via Azure Quantum, AWS Braket, IBM Cloud, direct vendor cloud services.

Backend-selection logic at L2. The routing decision is deterministic and based on the structural properties of the requested computation:

- Shallow tail (Value at Risk 95% or below), separable, embarrassingly parallel → GPU backend, available today
- Deep tail (Value at Risk 99% or beyond, Expected Shortfall 99.5% or beyond), memoryful, regime-conditional, high-dimensional, or non-separable → quantum backend target, contingent on commercial viability

The LLM at L4 does not influence backend selection. The architectural commitment is that backend routing is a technical decision at the compute layer, not a governance decision at the orchestration layer. Today, all routing decisions resolve to the GPU backend because the quantum backend is not yet commercially viable. The architecture is designed so that future quantum-backend availability requires no changes to L3, L4, L4.5, or L5 — only the routing rule at L2 expands.

---

7. Integration architecture: connecting quantum compute to trading and risk infrastructure

The published quantum-finance literature on capital markets applications systematically elides the integration architecture that connects financial institution trading and risk systems to the quantum compute resource. Egger et al. (2020), Stamatopoulos et al. (2020), Woerner and Egger (2019), and Chakrabarti et al. (2021) all assume the quantum compute resource is reachable from the application layer, without engaging with the operational reality of how trade data, market data, position state, and risk results actually flow between on-premises trading systems and cloud-resident quantum compute. This section addresses that gap directly.

7.1 Financial institutions will not own quantum hardware

The economic and operational logic against on-premises quantum hardware at financial institutions is straightforward. Current systems cost upwards of $10M in hardware capex plus substantial infrastructure for cryogenics, calibration, and environmental control. Operational expertise in trapped-ion physics, superconducting circuit calibration, error correction code maintenance, and quantum-classical interface engineering does not align with the skill profiles financial institutions hire for. Hardware roadmaps evolve faster than capex amortization cycles. The major potential quantum-finance users — JPMorgan Chase, Goldman Sachs, Barclays, BBVA, BNP Paribas, HSBC, and others — universally access quantum compute through cloud services or research partnerships, not through on-premises ownership.

The dominant access model is hyperscaler-mediated. Microsoft Azure Quantum provides the broadest QPU lineup, including Quantinuum H-series, IonQ, Pasqal, and Atom Computing hardware. AWS Braket provides IonQ, Rigetti, QuEra, and IQM access. IBM Cloud distributes IBM's own quantum hardware. Direct vendor cloud (Quantinuum Quantum Cloud Services, IonQ's direct offering, others) provides hardware-specific access. The architectural implication is that L0 hardware primitives are reached via cloud-mediated access, not via local execution. The integration substrate at L4.5 must therefore handle the boundary between on-premises trading-and-risk infrastructure and cloud-resident compute layers.

A note on what this backend-agnosticism does and does not claim. The architecture is agnostic across gate-model quantum backends — trapped-ion, superconducting, neutral-atom, photonic, and silicon-spin — because all of them expose the amplitude-estimation and sampling primitives the L2 solver targets; the integration substrate is written to the primitive, not to the hardware. That hedge is the rational response to a genuinely unsettled hardware roadmap, and recent federal funding makes the point concretely. The U.S. Department of Commerce's CHIPS quantum awards (announced May 21, 2026 — $2.013 billion in planned incentives across nine companies, with the government taking a minority, non-controlling equity stake in each) deliberately spread support across every major modality rather than betting on one: superconducting (IBM, the anchor recipient at $1 billion, and Rigetti), trapped-ion (Quantinuum, $100 million earmarked for scaling fault-tolerant trapped-ion systems), neutral-atom (Atom Computing, Infleqtion), photonic (PsiQuantum), silicon-spin (Diraq), and even annealing (D-

Wave). The earlier DARPA Quantum Benchmarking Initiative (Stage B, November 2025) likewise advanced eleven firms across competing modalities. No approach has yet separated decisively, which is precisely why an architecture committing to one today would be premature. Two caveats bound the claim. First, agnosticism at the interface level is not equal performance across hardware: the same circuit runs with materially different fidelity, depth, and cost on different machines, and backend selection at L2 must account for that. Second, the agnosticism is specific to the gate model. It does not extend to quantum annealing — the D-Wave paradigm, itself among the CHIPS recipients — which solves optimization problems by adiabatic evolution rather than executing the amplitude-estimation circuits this architecture's tail-sampling layer relies on; annealing is a different computational paradigm, not an alternative backend for this design.

7.2 The integration boundary

The integration substrate connects three substantively different compute environments:

- **Trading systems:** on-premises, colocated near exchanges and electronic communication networks, sub-millisecond latency for order management and execution paths, FPGA-accelerated pricing on high-frequency desks. Latency budgets in the hundreds of microseconds for pre-trade risk checks.
- **Risk management systems:** mixed environment. Real-time and intraday risk computation often runs on on-premises grid compute (Calypso, internal risk engines, vendor risk platforms). Overnight Value at Risk, regulatory capital, stress testing often runs on cloud-based batch compute. Latency budgets range from seconds (intraday) to hours (overnight).
- **Cloud-resident quantum compute:** queue-managed, with execution latency dominated by cloud round-trip plus circuit execution plus shot budget plus error mitigation. Round-trip latencies measured in seconds for the network leg, plus quantum execution time. Shot costs are per-circuit and non-trivial.

These three environments have different security boundaries (on-premises trading systems carry different regulatory and competitive constraints than cloud-resident compute), different regulatory compliance requirements (data residency under SR 11-7 model risk management (Federal Reserve and OCC 2011), TRIM internal model approvals (ECB 2021), MAS guidance for Singapore-resident institutions, FCA expectations for UK institutions), and different cost structures (on-premises grid compute is amortized capex; cloud compute is per-execution opex; quantum compute is shot-budget opex). Connecting them is not an API integration. It is a data-plane plus control-plane architectural design.

7.3 Event mesh as the integration substrate

A high-throughput, low-latency event mesh is the natural design pattern for the L4.5 integration substrate. Candidate technologies include Solace PubSub+ (purpose-built for capital markets event distribution; Solace 2023), Apache Kafka with latency tuning (the partitioned-log design originating in Kreps, Narkhede, and Rao 2011), Aeron (an open-source low-latency transport used at high-frequency trading firms for sub-microsecond messaging; Real Logic 2023), and other event-streaming platforms. The specific vendor choice is less important than the architectural requirements the substrate must satisfy. One clarification preempts a natural

objection: the sub-microsecond capability these transports offer is dictated by the classical real-time risk plane, not by the quantum path. As Section 7.4 develops, the quantum-augmented computation consumes from the mesh only on a batch schedule; the substrate is justified by the classical workload on its own, and the quantum path simply reuses the same backbone as a low-frequency, asynchronous consumer.

Requirements:

- **High-throughput streaming** for trade events, market data, and position updates without backpressure on the upstream trading systems
- **Low-latency synchronous paths** for the request-response interactions where risk must respond to trading decisions in near-real-time (pre-trade limit checks, intraday Value at Risk recomputation)
- **Asynchronous streaming** for the batch-oriented analytics where risk results flow back without blocking trading
- **Guaranteed-delivery semantics** for regulatory-audit-trail paths where every trade event must reach risk infrastructure exactly once with reproducible ordering
- **Schema-governed contracts** between producers (trading systems) and consumers (risk systems), versioned and verifiable
- **Bidirectional flow** because risk results have to flow back into trading systems as limit-usage updates, capital allocation changes, and hedge-ratio recommendations

7.4 Three design implications

Three operational realities shape the integration substrate design:

Asymmetric latency profiles. Real-time pre-trade risk requires sub-millisecond round-trip from trade event to risk decision. Quantum compute, even at fault-tolerant scale, will not deliver sub-millisecond response for the foreseeable future — quantum circuit execution, shot budgets, error mitigation, and cloud round-trip add up to latencies measured in seconds at minimum. The architectural response is separation of concerns: classical fast path for pre-trade risk (running on existing on-premises grid compute and accelerated GPU infrastructure), quantum-augmented path for intraday and overnight analytics where seconds-to-minutes latency is acceptable. The event mesh routes trade events into both paths, and risk results from the quantum-augmented path return to update intraday and overnight limit allocations rather than per-trade decisions.

Data residency and security boundaries. Trade data is among the most sensitive data financial institutions hold. Direct routing of raw trade data through cloud-resident quantum compute crosses regulatory boundaries that exist for legitimate reasons under SR 11-7, TRIM, MAS, and equivalent frameworks. The architectural response is on-premises pre-aggregation. Raw trade data is aggregated to risk-factor exposures (delta, gamma, vega, key-rate duration, default-probability-weighted exposure by counterparty) before transmission to the cloud compute layer. The cloud layer operates on risk-factor exposures, not on individual trades. This reduces the regulatory surface area and limits the consequences of any cloud-side security event. The event mesh enforces this boundary: trade events flow to on-premises pre-aggregators; only aggregated risk-factor states cross to the cloud quantum path.

Quantum compute cost as a routing constraint, and the asynchronous decoupling it requires. Per-circuit execution cost on quantum cloud services is non-trivial today and will remain so for several years. The cost structure differs fundamentally from cloud classical compute (per-second VM time) — quantum cost scales with shots executed per circuit, and shot counts for production-grade Value at Risk computations are large. The architectural response is intelligent batching and request aggregation. It is important to be explicit that this batching places the quantum-augmented path entirely outside the event mesh's low-latency streaming loop: the quantum path is asynchronous and decoupled, triggered as low-frequency batch operations (for example hourly or end-of-day) to optimize shot-budget economics. There is no near-real-time quantum-augmented routing, and the architecture does not claim one. The event mesh's low-latency synchronous paths (Section 7.3) serve the classical pre-trade and intraday-refresh cadences exclusively; the quantum path consumes aggregated risk-factor states from the mesh on a batch schedule and returns results that update limit allocations and capital figures asynchronously. This separation eliminates any pipeline-stall tension: the streaming substrate is never asked to hold back exposure updates to accumulate a quantum batch, because the quantum batch is assembled on its own clock from the durable state the mesh already maintains, not from the synchronous path. The L2 backend-selection logic uses these

batched aggregations to determine when quantum routing is economically justified versus when classical Monte Carlo on the fast path is the appropriate path.

7.5 Why this layer is load-bearing

The integration substrate is the layer that distinguishes a deployable architecture from a research demonstration. Without it, the quantum compute layer is operationally unreachable from the trading and risk infrastructure that exists at financial institutions, and even the classical two-stream architecture cannot connect to live trading and risk systems. This layer is where the bulk of the engineering work, regulatory engagement, and institutional risk management actually happens — and where the published quantum-finance literature, written largely from the technology side, has its largest blind spot. The architectural contribution of this paper is not just “use quantum compute for tail risk”; it is the recognition that the integration substrate is the precondition for any of the streams to deliver value inside a real institution.

---

8. Open questions: a research agenda

The architecture proposed here rests on a number of open research questions. Each is identified explicitly rather than assumed resolved. The most consequential are listed first.

Q1 — Quantum advantage for multifractal tail computation: open at both the loading and sampling layers. Does the MMAR cascade admit a genuine quantum advantage, and if so, where? This resolves into two linked sub-problems. The *loading* question: does the scale-invariant multiplicative cascade structure admit an efficient hierarchical quantum state preparation circuit, analogous to matrix product state preparation of low-entanglement distributions — and on which side of the classical-tensor-network-simulability boundary do MMAR cascades fall? The *sampling* question: even where the distribution has low bond dimension and is therefore classically loadable, do the specific deep-tail expectations relevant to capital adequacy resist efficient classical tensor-network contraction, leaving a residual advantage for amplitude estimation? Sub-questions: (a) does cascade-level correlation keep bond dimension bounded as cascade depth grows; (b) does the construction yield a normalized, well-defined quantum state representing the distribution over cascade outcomes rather than a single realization; (c) for low-bond-dimension cascade states, are there tail functionals whose classical contraction cost is nonetheless prohibitive while amplitude estimation remains efficient; (d) does the prepared state compose with downstream amplitude estimation to deliver end-to-end advantage. This is the most distinctive open question raised by the architecture. The honest current expectation (Section 4.2) is that financial cascades likely fall on the classically simulable side at the loading layer, which makes sub-question (c) — the sampling-layer functional question — the decisive one for whether the quantum component delivers advantage at all.

This question is empirically approachable, and specifying the experiment is itself part of the contribution. The concrete experiment is to express the level-to-level multiplicative cascade map as a parameterized quantum circuit — formally, a recursive production-rule construction in which a single generative operation is reapplied across cascade levels rather than enumerated per outcome — prepare the MMAR cascade state for increasing cascade depth d, and measure how circuit depth, qubit count, and the bond dimension of the matrix-product representation scale with d. Two outcomes bound the result. If these resources stay bounded as d grows, the cascade is efficiently preparable — but the same bounded bond dimension likely renders it classically simulable, pushing the question to whether the specific deep-tail functional resists efficient classical contraction even for a low-bond-dimension state (sub-question c). If the resources grow without bound, loading is quantumly hard in the generic Herbert sense and the cascade's self-similarity confers no preparation advantage. The region in which a durable quantum advantage survives is therefore narrow and precisely specifiable: a cascade state cheap to prepare yet carrying a tail functional expensive to contract classically. Characterizing whether MMAR cascades occupy that region is a well-posed numerical tensor-network experiment that requires no quantum hardware to begin — the experiment that would decide the architecture's central conditional hypothesis is itself classical.

Empirical update — Q1 sub-question (a). The classical numerical tensor-network experiment described above as the empirical approach to Q1 has been executed and is reported as Part 2 of this combined document. The headline result for the matrix product state ansatz and the balanced binary tree tensor network with natural qubit ordering: bond dimension saturates the maximum possible value 2^(K/2) at every nonzero intermittency tested, K ≤ 20, both multiplier laws. The hypothesized "classically simulable at the loading layer" outcome is not supported by the present measurement under these ansatzes for the cases tested. Sub-question (a) is therefore answered negative under MPS / balanced-natural-ordering TTN; sub-questions (b), (c), and (d) remain open. See Part 2 for full result, scope, and limitations.

Q2 — Verifiability under shot noise. How does Bar 2 forward discipline extend to probabilistic quantum compute? What is the reproducibility unit when the quantum backend's output is a Value at Risk estimate plus confidence interval rather than a deterministic value? How does the model-risk-management audit trail compose when results include circuit specification, shot budget, hardware identifier, and calibration snapshot?

Q3 — Backend-routing governance. Should the L4 orchestration layer ever influence the L2 backend-selection decision, or is that an anti-pattern? The architecture proposed here commits to deterministic technical routing at L2 with no L4 influence; the question is whether this commitment holds under all operational scenarios or whether edge cases require L4 awareness of routing.

Q4 — Error-mitigation labor. Quantum error mitigation today requires expert tuning of zero-noise extrapolation parameters, probabilistic error cancellation circuits, and dynamical decoupling sequences. Does the orchestration layer compress this work into machine-tractable processes, or does it require vendor-side automation? The answer determines whether the architecture is operationally accessible at a 3-year horizon or only at a 7-year horizon.

Q5 — Formal characterization of the LLM's output family. What is the precise specification of the parametric distribution family that the LLM orchestration layer produces? The dimensional reduction from arbitrary empirical distributions to a structured parametric family is what makes loading tractable, and the formal characterization of this family is part of what determines the architecture's mathematical scope.

Q6 — Integration substrate latency-throughput specification. What are the appropriate latency-throughput service-level agreements for the L4.5 integration substrate across different risk-pipeline cadences (pre-trade, intraday refresh, end-of-day, regulatory submission)? How does the SLA structure differ between cloud-resident quantum compute and on-premises classical compute, and how does the event mesh handle this heterogeneity?

Q7 — Data-residency and security envelope. What is the regulatory-compliant data flow specification for cloud-mediated quantum risk computation under SR 11-7, TRIM, MAS, FCA, and equivalent frameworks? Specifically, can risk-factor-exposure summaries be routed to cloud quantum compute without triggering regulatory data-residency restrictions, and what is the audit-trail composition that satisfies supervisory expectations?

Q8 — Regulatory acceptance lag. Model risk management acceptance timelines for quantum-augmented models likely lag technical-readiness by 3 to 5 years. What does early validation work look like? What demonstration sequence persuades regulators that quantum-backed Value at Risk estimates are auditable and defensible?

Q9 — Realistic deployment timeline. Expert estimates of when quantum hardware delivers commercial advantage for the specific application classes addressed here vary widely. Three-year horizons are aggressive; seven-year horizons are defensible; fifteen-year horizons capture the broader uncertainty band. The architecture's value depends on its preparation for an outcome whose timing is genuinely uncertain.

---

9. Closing

The architecture proposed in this paper operates today on two production-ready streams — multifractal mathematics and LLM agentic AI — synthesized through an institutional-grade integration substrate into a coherent framework for portfolio-level risk management at deep-tail confidence levels. It is built so that a third stream, quantum compute, can be absorbed at the L2 backend if and when it reaches commercial viability, without requiring its presence today and without re-architecting the layers above it. The framework is therefore best understood not as three co-equal components assembled in parallel, but as a working two-stream classical architecture deliberately structured to intercept a research-stage third stream at the point where that stream would add value.

The originality claim is correspondingly specific, and should not be conflated with the technical hypothesis it surrounds. A literature search found no prior work composing multifractal cascade modeling, bounded LLM regime adaptation, and quantum tail-sampling into a single risk architecture; that three-way synthesis is novel as a matter of literature composition, and the claim holds regardless of how the separate, deliberately open technical question — where, or whether, quantum delivers advantage at the sampling layer — is ultimately resolved. The architecture is deployable on its two classical streams today, and absorbs the third if and when the technical question resolves favorably.

This is a research-agenda paper, not a research-results paper. It articulates an architectural framework, identifies open questions, and provides intellectual context for engagement with the field. Empirical validation of the specific claims — particularly the cascade-state-preparation tractability question, the loading-boundary efficiency question, and the integration-substrate latency specification — remains future work. The contribution intended is architectural synthesis and research-direction articulation, not numerical demonstration.

The framework is also explicit about what it does not do. It does not propose a commercialization roadmap. It does not claim quantum compute is currently delivering institutional-scale advantage. It does not endorse the Black-Scholes-Merton foundational framework that underlies much of the published quantum-finance derivative-pricing literature, though it cites that literature for its algorithmic contributions. It does not relitigate the empirical case for fat-tailed return distributions, which has been settled since Mandelbrot (1963) and reinforced for sixty years.

Broadening access beyond tier-one firms

One downstream consequence of the architectural design is worth surfacing explicitly. The cloud-mediated access model for quantum compute, combined with the integration substrate that connects on-premises trading infrastructure to cloud-resident compute resources, means that institutional-grade tail risk modeling capability is not structurally limited to firms with on-premises quantum hardware, internal quantum research teams, or tier-one technology budgets. The integration substrate is the equalizer. As commercial viability emerges, the architecture is in principle accessible to a broader range of institutions than the tier-one technology firms that

currently dominate published quantum-finance work — including mid-tier banks, regional broker-dealers, large pension funds, sovereign wealth funds, insurance reserving operations, and central clearing counterparties. This accessibility implication is not the primary contribution of this paper, but it is worth surfacing as a downstream consequence of the cloud-mediated, integration-substrate-based design choice. Tail risk modeling sophistication has historically been concentrated at institutions whose scale gave them access to specialized compute and quantitative talent; the architectural framework proposed here is structured so that, when quantum commercial viability arrives, the sophistication can flow more broadly through the institutional landscape.

Audience and future directions

The paper is intended for senior practitioners and researchers in capital markets, AI, and quantum computing. A further line of inquiry, beyond this paper's scope, concerns whether the multifractal substrate's scale-invariant structure carries implications for the resilience of the broader financial system — whether systems whose risk is modeled and managed through scale-invariant rather than single-scale representations distribute and absorb shocks differently — which the author intends to explore in future work.

The author welcomes engagement, critique, and collaboration on the open questions raised in Section 8.

Part 2 — Empirical Loading-Layer Measurement

Abstract

A quantum-amplitude-estimation speedup over classical Monte Carlo for capital-markets risk computation requires that the underlying distribution can be loaded onto the quantum device cheaply — a result first sharpened by Herbert (2021) for log-concave distributions. We measure, directly, whether the multiplicative cascade measure underlying the Mandelbrot–Calvet–Fisher Multifractal Model of Asset Returns (MMAR) is efficiently representable as a matrix product state (MPS), equivalently a quantics tensor train (QTT) indexed by dyadic scale. We find that the bond dimension at the central bipartition saturates the maximum possible value 2^(K/2) at every cascade depth K from 4 to 20, at every nonzero intermittency tested, for both lognormal and conservative-binomial multiplier laws and at four accuracy targets $\varepsilon \in \{10^{-2}, 10^{-3}, 10^{-6}, 10^{-10}\}$. The exponential growth rate matches the saturation rate to within 1–4% across all 11 nonzero-intermittency cells. The result is independent of cascade construction (conservative renormalized vs. canonical non-renormalized), and the saturation holds at every bipartition cut, not just the centre — a deductive consequence of which is that a balanced binary tree tensor network with natural qubit ordering also saturates. The cascade is therefore not efficiently representable in two of the standard classical tensor-network ansatzes. This is a necessary but not sufficient condition for a quantum loading advantage at this distribution class: it removes one classical escape route without establishing a working quantum alternative. The loading-layer question for multifractal distributions is reopened rather than resolved.

---

1. Introduction

Quantum amplitude estimation offers a theoretical quadratic speedup over classical Monte Carlo. Herbert (2021) showed that this speedup is erased for log-concave probability distributions, because Grover–Rudolph state preparation requires resources comparable to the classical computation it sought to accelerate. The lesson — that state preparation can be the binding constraint on end-to-end quantum advantage — extends naturally to any distribution class, with the answer depending on whether the relevant state can be prepared efficiently.

The Mandelbrot–Calvet–Fisher Multifractal Model of Asset Returns (Mandelbrot, Fisher, and Calvet 1997; Calvet and Fisher 2008) is a canonical model for fat-tailed financial returns with empirically observed multifractal scaling. Its central object is a multiplicative cascade measure μ on a binary tree of dyadic scales, composed with fractional Brownian motion to produce the price process. The cascade is, by construction, neither log-concave nor smooth: it sits outside the distribution classes for which efficient quantum state-preparation methods have been established (Carrera-Vazquez and Woerner 2020 list only log-concave, machine-learned, and smooth-via-piecewise-polynomial-MPS as known efficient classes).

In a companion architectural paper (Padmanabhan 2026), we argued that the loading layer for multifractal distributions is open: classical tensor-network methods may or may not yield polynomial bond dimension for the cascade, and only direct measurement can settle the question. The present paper reports that measurement.

The contribution is empirical. We construct the cascade explicitly, encode the resulting probability measure as the amplitude state of a K-qubit system in the quantics scale ordering, and measure the bond dimension at every bipartition cut as a function of cascade depth K and multiplier intermittency. We make no claims about the full price process X(t) = B_H(θ(t)) — only about the cascade core that carries the multifractal structure. We make no claims about quantum hardware behaviour — the experiment is classical throughout. We make no claim of quantum advantage — we constrain one necessary condition for it. The result is a measurement, not a theorem.

To state the position one more time, because the framing is easy to misread on a quick skim: the paper reports a *constraint* on a classical escape route, not progress toward quantum advantage. Establishing that one canonical classical representation is exponentially expensive is a necessary condition for any quantum loading advantage at this distribution class; it is not a sufficient condition, and we do not propose, design, or analyze a quantum preparation circuit. The question of whether a polynomial-depth quantum loading method exists for the multifractal cascade is the natural quantum-side counterpart to the present measurement (§6.4) and remains open.

2. Background

The multiplicative cascade. A binary multiplicative cascade is constructed iteratively on a binary tree of depth K. Starting from total mass 1 at the root, each parent at level $\ell < K$ distributes its mass to two children using a pair of multipliers (M_L, M_R) drawn from a chosen distribution. After K levels, the cascade yields 2^K leaf masses {m_i} summing to 1

(conservative construction) or to a random total ≈ 1 (canonical construction). The intermittency parameter — λ^2 (the variance of log M) for the lognormal cascade, p (the smaller child mass) for the conservative binomial — controls the multifractal strength of the resulting measure. $\lambda^2 = 0$ (lognormal) or p = 0.5 (binomial) yields a deterministic uniform distribution.

Quantics tensor trains. A quantics tensor train (QTT; Khoromskij 2011) encodes a function on a 2^K grid as a tensor train with one tensor per dyadic scale. In this representation, the bond dimension at the edge between tensor ℓ and tensor ℓ+1 is the Schmidt rank of the bipartition between the coarsest ℓ scales and the remaining K − ℓ scales. For multiplicative-cascade-generated measures, this bond dimension is a direct quantitative measurement of cross-scale entanglement — the multiplicative-cross-scale dependence structure that defines multifractality. (The term *bond dimension* originates in tensor-network literature; in the present setting it is the size of the auxiliary index linking two tensors in the chain. It has no relationship to fixed-income instruments.)

For a normalized amplitude state $|\psi\rangle$ with amplitudes $a_i = \sqrt{m_i}$ on the 2^K-dimensional computational basis, the bond dimension at bipartition cut k is the rank of the reshaped 2^k × 2^(K−k) matrix. The maximum possible value at cut k is min(2^k, 2^(K−k)) = 2^min(k, K−k), maximised at k = K/2 with value 2^(K/2). A state is *efficiently representable* as an MPS / QTT if its bond dimension grows at most polynomially in K; the contrasting regime — exponential, saturating the maximum — is incompressible.

State preparation in quantum finance. Carrera-Vazquez and Woerner (2020) classify known efficient state-preparation routes for quantum amplitude estimation into three families: log-concave distributions (Grover–Rudolph), machine-learned distributions (e.g., qGAN), and smooth functions approximated by piecewise polynomials with associated MPS representations. Multiplicative cascade measures fall in none of these. Whether they admit any efficient classical tensor-network representation is the open question; bond-dimension measurement provides the direct test.

3. Methods

Cascade construction. We implement two constructions:

- *Conservative lognormal:* at each parent, two independent lognormal multipliers M_L, M_R with E[M] = 1/2 and Var(log M) = λ^2 are drawn and *renormalized* so M_L' = M_L/(M_L+M_R) and M_R' = M_R/(M_L+M_R). Mass is conserved exactly at each step; the marginal distribution of M_L' deviates slightly from pure lognormal for large λ^2.
- *Conservative binomial:* at each parent, the children receive masses (p, 1−p) with the assignment randomized; p ∈ (0, 1/2] is the smaller mass.

We also implement a *canonical lognormal* construction (independent lognormal multipliers, no per-step renormalization; mass conserved only in expectation) for use in the construction-artifact check (§4.3).

All random draws use a deterministic SHA-256-based seed protocol over the tuple (K, multiplier law, intermittency value, realization index), enabling exact reproducibility across runs and identical-realization comparisons between methods.

Amplitude state and bond-dimension measurement. For each realization, the leaf masses {m_i} are converted to a normalized amplitude state a_i = √m_i. At every bipartition cut k ∈ {1, ..., K−1}, the amplitude vector is reshaped into a 2^k × 2^(K−k) matrix and its singular value decomposition is computed. For each cut we report two bond-dimension measures:

- *Machine-precision rank*: count of singular values above the numpy default `matrix_rank` tolerance σ_max · max(shape) · ε_machine, scaling correctly with matrix size.
- *Bond dimension at accuracy ε*: the minimum χ such that the L2 truncation error from keeping only the χ largest singular values is at most ε. We report χ at four values of ε spanning practical accuracy targets: 10^{-2}, 10^{-3}, 10^{-6}, and 10^{-10}.

The von Neumann entanglement entropy is also recorded at every cut as a continuous companion to the integer-valued bond dimension.

Sanity-check discipline. Before any non-trivial measurement is reported, the cascade-construction code is verified against three gates: mass conservation (every realization sums to 1 within machine precision); deterministic-limit behaviour (intermittency = 0 yields uniform leaf masses, hence bond dimension 1 at every cut and entropy 0); and reproducibility (identical seeds produce identical cascades). The deterministic-limit gate is treated as load-bearing — halt-on-failure with no warn-and-continue — because it is the strongest single check that the construction code is correct. The gate passes uniformly for K = 4..20 in both laws.

Sweep parameters. The principal sweep uses cascade depths K ∈ {4, 8, 12, 16, 20}, R = 30 realizations per cell, both multiplier laws, and intermittency grids λ² ∈ {0.01, 0.05, 0.10, 0.15, 0.30, 0.50} for lognormal and p ∈ {0.50, 0.45, 0.40, 0.35, 0.30, 0.20} for binomial. K = 20 is the upper bound at which brute-force SVD on the 2^K amplitude vector remains tractable in memory; the extension to higher K via Tensor Cross Interpolation (TCI) on the QTT (Núñez Fernández et al. 2022) is a natural follow-up not undertaken in this paper.

Code availability. All cascade-construction, measurement, sweep, and analysis code is implemented in Python (numpy, scipy, matplotlib) and is available from the author on request. The deterministic seed protocol ensures that any reported numerical result can be regenerated by a reader who runs the same (K, law, intermittency, realization-index) tuple.

4. Results

4.1 Bond-dimension scaling across intermittency

For each (K, law, intermittency) cell, we report the mean centre-cut bond dimension over the R = 30 realizations. The headline result is the scaling of this quantity with K at fixed intermittency. Figure 1 shows the saturation behaviour and its dependence on intermittency at K = 20.

[Figure 1: compressibility ratio bd/max_possible at K=20, as a function of intermittency, for lognormal (left) and binomial (right) cascades. Curves are shown at four accuracy targets ε ∈ {10^{-2}, 10^{-3}, 10^{-6}, 10^{-10}}. At deterministic (λ^2=0 or p=0.5), bd=1 (ratio ≈ 0); at any nonzero intermittency, the cascade reaches ≥ 93% of full saturation at every tight tolerance, with the loosest tolerance ε=10^{-2} dropping only modestly to 70–85%.]

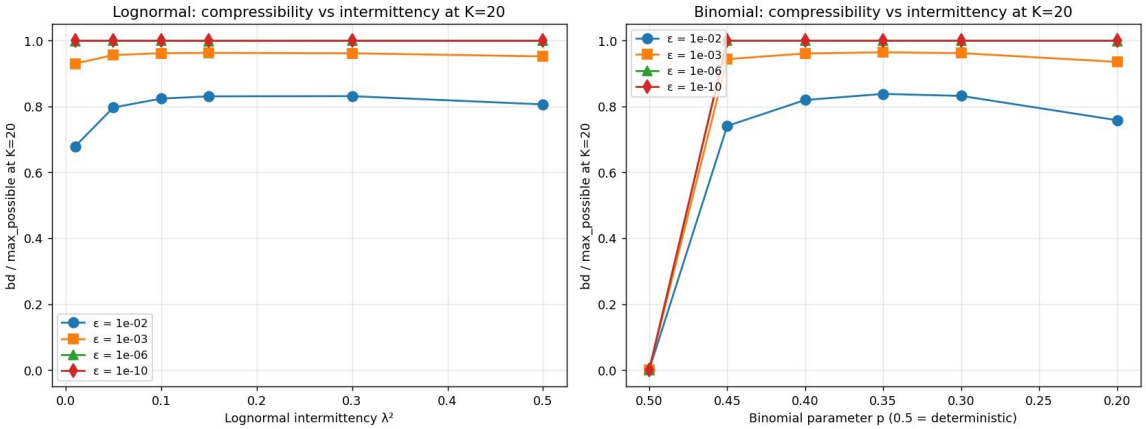


Table 1 reports the scaling fit γ such that bd ~ exp(γK), compared to the saturation rate γ_sat = log(2)/2 ≈ 0.347 (which corresponds to bd = 2^(K/2)). Across all 11 nonzero-intermittency cells, γ / γ_sat ∈ [0.991, 1.037]: the cascade scales as 2^(K/2), the maximum possible, with no detectable deviation.

Law / intermittency	bd @ K=4	K=8	K=12	K=16	K=20	γ_fit	γ/γ_sat
Lognormal λ² = 0.01	3.9	15.2	59.7	238.6	952.7	0.343	0.991
Lognormal λ² = 0.05	3.9	15.7	61.6	245.3	979.1	0.345	0.994
Lognormal λ² = 0.10	4.0	15.7	62.0	246.8	985.0	0.345	0.994
Lognormal λ² = 0.15	4.0	15.7	62.2	247.3	986.0	0.344	0.993
Lognormal λ² = 0.30	4.0	15.8	62.3	247.4	984.8	0.344	0.994
Lognormal λ² = 0.50	4.0	15.8	62.2	246.1	975.1	0.344	0.991
Binomial p = 0.45	3.2	15.3	60.7	241.9	966.5	0.354	1.023
Binomial p = 0.40	3.5	15.7	62.1	246.5	984.1	0.351	1.013
Binomial p = 0.35	3.4	15.7	62.3	247.8	987.7	0.353	1.017
Binomial p = 0.30	3.1	15.8	62.2	247.6	985.2	0.357	1.030
Binomial p = 0.20	2.8	15.8	62.0	243.5	958.0	0.360	1.037

Table 1. Centre-cut bond dimension at ε = 10^{-3} across K and intermittency. The maximum possible bond dimension at K = 20 is 2^10 = 1024. γ_sat = log(2)/2 ≈ 0.347. All nonzero-intermittency cells are within 4% of full saturation.

The saturation is essentially deterministic across the 30 realizations per cell. The empirical standard deviation of the centre-cut bond dimension is below 1 unit at K ≤ 12, below 2.2 units at K = 16, and below 8.4 units at K = 20 across all nonzero-intermittency cells; the worst-case coefficient of variation across the entire grid is 0.85% (at K = 20, lognormal λ² = 0.50). No realization deviates from its cell mean by more than 2.5% at any tolerance. The saturation finding is therefore not driven by outlier realizations or by a small subset of realizations dominating the cell mean — every individual cascade in every cell saturates.

A striking feature is the discontinuity at the deterministic boundary. At p = 0.5 (binomial) or λ² = 0 (lognormal), bond dimension is exactly 1 at every cut for all K. At any infinitesimally nonzero randomness — p = 0.45 (a 10% deviation from deterministic), λ² = 0.01 (1% log-variance) — bond dimension at K = 20 already reaches 93% or more of the maximum. The transition from polynomial to exponential bond dimension occurs at intermittency = 0+, not at any finite threshold.

The result is uniform across the four accuracy targets at tight tolerances ($\varepsilon \le 10^{-6}$ gives ≥ 99% saturation across all nonzero-intermittency cells), with only modest loosening at $\varepsilon = 10^{-2}$ (where the centre-cut ratio drops to 80–93%). The cascade is not significantly more compressible at practical loading accuracies than at machine precision.

4.2 Per-cut profile: extension to balanced tree tensor networks

The centre-cut bond dimension reported in §4.1 measures one bipartition. A balanced binary tree tensor network (TTN) with natural qubit ordering — the most common alternative to MPS for tree-structured data — measures bond dimension at a *subset* of MPS cuts: at the root, the bipartition between leaves with first bit 0 and first bit 1 (equivalent to MPS cut at k = 1); at level 2, bipartitions at k = 2 and k = K − 2; and so on. Consequently:

If MPS bond dimension is saturated at every cut, balanced TTN with natural qubit ordering also saturates at every internal edge. This follows deductively from the subset relation; no separate TTN measurement is required.

We verify the antecedent by measuring MPS bond dimension at every cut $k \in \{1, ..., 15\}$ for K = 16, at the cells (lognormal, $\lambda^2 = 0.30$) and (binomial, p = 0.40). The minimum saturation ratio across all cuts is 96.6% (lognormal) and 96.3% (binomial) at $\varepsilon = 10^{-3}$, with 100% saturation at every cut at $\varepsilon \le 10^{-6}$. Every cut except the centre is at 98–100% saturated even at $\varepsilon = 10^{-2}$. The antecedent holds; the consequence — that balanced TTN with natural ordering also saturates — follows.

Custom (non-balanced or non-natural-ordering) TTN topologies are not addressed by this argument and remain an explicit acknowledged limitation (§5).

4.3 Construction-artifact check

The conservative cascade construction enforces per-step mass conservation by renormalizing independent lognormal multipliers. A natural concern is whether the renormalization itself, rather than the underlying multiplicative structure, drives the saturation. We test this by repeating the K = 16, $\lambda^2 = 0.30$ lognormal measurement using the *canonical* construction (independent lognormal multipliers without per-step renormalization; mass conserved only in expectation; the amplitude state is normalized globally at the end), R = 20 realizations.

Tolerance ε	Conservative bd	Canonical bd	Difference	Conservative ratio	Canonical ratio
10^{-2}	215.9	205.2	−10.8	84.3%	80.1%
10^{-3}	247.3	244.1	−3.2	96.6%	95.3%
10^{-6}	255.9	255.9	0.0	100.0%	100.0%
10^{-10}	256.0	256.0	0.0	100.0%	100.0%

Table 2. Centre-cut bond dimension at K = 16, lognormal, $\lambda^2 = 0.30$ for both cascade constructions. Maximum possible bd at K = 16 centre cut is 256. Constructions agree within 1.3% at $\varepsilon = 10^{-3}$ and are identical at $\varepsilon \le 10^{-6}$. The canonical construction shows slightly higher compressibility at the loosest tolerance, with corresponding mean entanglement entropy 2.19

nats (canonical) versus 1.55 nats (conservative), reflecting that per-step renormalization modestly concentrates the singular spectrum without altering its rank.

The saturation is not a renormalization artifact. The two standard cascade constructions produce essentially identical bond-dimension behaviour; the modest difference at the loosest tolerance is consistent with the canonical cascade's slightly broader singular spectrum and does not change the qualitative conclusion.

4.4 Entanglement entropy: rank-full but information-light

A subtlety worth recording: while bond dimension is saturated uniformly across nonzero intermittency, the von Neumann entanglement entropy varies smoothly with intermittency. At K = 20 lognormal: 0.15 nats at $\lambda^2 = 0.01$, growing to 2.72 nats at $\lambda^2 = 0.50$. At K = 20 binomial: 0.28 nats at p = 0.45, growing to 3.54 nats at p = 0.20. The maximum possible entropy at the K = 20 centre cut is log(2^10) = 6.93 nats; the cascade entropy is at most ≈ 50% of this maximum even at high intermittency.

The cascade is therefore *rank-full but information-light*: its singular-value spectrum contains a long tail of small but above-tolerance singular values whose count saturates the maximum rank, while most of the L2 weight is concentrated on a few dominant modes. Intermittency controls how much weight the tail carries, but not whether the tail exists. This pattern is itself a structural observation about the cascade's representation — the singular spectrum may bear a multifractal signature analogous to the cascade itself — but a detailed characterisation is deferred to future work.

A natural reader concern arises here: if the cascade is rank-full but information-light, could a variational MPS algorithm with adaptive bond-dimension growth, or an approximation scheme using non-linear singular-value truncation, exploit the spectral concentration to achieve polynomial representation? Our multi-tolerance measurement provides a partial answer. At $\varepsilon = 10^{-2}$ — a 1% L2 truncation, looser than any practical quantum-loading accuracy target would require — the centre-cut bond dimension at K = 20 still ranges from 80% to 93% of the maximum across all nonzero-intermittency cells (Figure 1, left and right panels at the $\varepsilon = 10^{-2}$ curve). The information-light spectrum does not collapse to polynomial bond dimension within the L2 norm at any of the tested tolerances. Whether *non-L2* truncation criteria (e.g., total variation, Kullback–Leibler divergence, or sup-norm) could escape the saturation regime is an open question whose answer would itself constitute a separate measurement; we explicitly do not claim that all approximate compression schemes are excluded by the present work, only that L2-norm truncation at practical accuracies is not a viable escape.

5. Scope and Limitations

The result above is scoped precisely. We state what it does and does not establish, in five points, ranked by the severity of the limitation.

(i) MPS and balanced-natural-ordering TTN only, not all classical representations. The measurement establishes exponential bond-dimension scaling for the matrix-product-state ansatz and (by the deductive argument of §4.2) for the balanced binary TTN with natural qubit

ordering. Custom TTN topologies, non-natural qubit orderings, hierarchical Tucker decompositions, neural-network state representations, sparse-grid methods, and other classical schemes have not been tested. A classical-representation efficiency result for one of these alternatives is not excluded by what we report.

(ii) The cascade measure, not the full MMAR price process. We measure the multiplicative cascade μ. The MMAR price process is $X(t) = B_H(\theta(t))$, the composition of a fractional Brownian motion with a multifractal trading time derived from μ. The fBm composition is monofractal and Gaussian, and is assumed — but not measured — not to alter the cascade's representational cost. The companion fBm-composition measurement is a natural follow-up.

(iii) One distribution family, not the broader multifractal class. Multifractality is a property exhibited by many distribution families: the multiplicative cascade we measure, multifractal random walks (Bacry, Muzy, and Delour 2001), continuous-time multifractal models, α-stable distributions, and others. Our finding extends to the canonical lognormal and conservative-binomial cascade constructions. Generalization to other multifractal models requires their own measurement.

(iv) Cascade depth $K \leq 20$ and L2 truncation norm. Brute-force SVD becomes infeasible above $K \approx 24$ (memory). The scaling fit through $K = 20$ is excellent and consistent ($\gamma_{fit} / \gamma_{sat} \in [0.99, 1.04]$), but the strictly asymptotic claim warrants extension via tensor cross interpolation (Núñez Fernández et al. 2022) to $K \approx 40$, deferred to future work. Truncation accuracy is measured in L2 throughout; alternative norms (total variation, Kullback–Leibler divergence, sup-norm) may yield qualitatively different scalings and have not been tested.

(v) Necessary but not sufficient for quantum advantage. Exponential classical representation cost is a necessary condition for any quantum loading advantage at this distribution class. It is not sufficient. A quantum loading method might also have exponential cost — the question of whether the multifractal cascade admits a polynomial-depth quantum preparation is open and is not addressed by the present work.

6. Discussion

6.1 What the result reopens

The companion architectural paper (Padmanabhan 2026) hypothesized that the cascade was likely classically efficiently representable, and consequently that the open quantum question for capital-markets risk consolidated at the *sampling* layer rather than the loading layer. The present measurement reopens the loading-layer question for the specific case of MPS and balanced-natural-ordering TTN representations of the cascade measure.

The honest reading is: the loading layer is now empirically constrained rather than merely conjectured. Whether the constraint is resolved by a smarter classical representation (a custom TTN topology, a neural-network ansatz, a method not yet identified) or by a polynomial-depth quantum preparation method (which would establish genuine quantum advantage), is open. Two future measurements — one classical, one quantum — together would close the question. The present paper closes one classical sub-question and identifies the remaining structure of the problem.

6.2 On the choice of object measured

A natural objection: the measurement targets the cascade measure μ rather than the full price process X(t). We address this objection in three parts.

First, *scope honesty*: we claim only loading-layer representability of the cascade, not of the price process. The conclusions are bounded accordingly.

Second, *the cascade is the locus of the relevant property*: the multifractal structure that defines the MMAR's fat tails — the property motivating quantum methods in the first place — lives in the cascade. The fBm composition is a smooth, monofractal, Gaussian factor whose representational cost is known to be benign (Sano et al. 2026 for the smooth-function case generally; the QTT representation of fBm-related Gaussian objects has been studied separately, e.g., Carrera-Vazquez and Woerner 2020). Probing the cascade alone is therefore probing the multifractal source, not an arbitrary tractable simplification.

Third, *direction of bias*: adding the fBm composition and real-market microstructure features (jumps, regime shifts, intraday effects) would, if anything, *increase* representational cost. Our negative result on the bare cascade is a conservative lower bound on the difficulty of the full process. If the cascade is already not efficiently representable, the price process composed with it is unlikely to become more representable. The objection therefore strengthens rather than weakens the result.

6.3 Connection to the loading-layer literature

The result complements two recent strands of the tensor-network state-preparation literature. Sano et al. (2026) derive rigorous asymptotic expansions for the bond-dimension decay of MPS representations of *smooth* functions, showing very slow growth. Guzman, Tiunov, and Aolita (2026) demonstrate that *generated* fractal noise fields, including 3D synthetic turbulence, can be represented directly in tensor-train format with logarithmic complexity. The present work fills a gap between these: a *measured* result on stochastic multifractal cascades — neither smooth

nor generated, but sampled from the standard generative tree — and the finding is qualitatively distinct from both. The cascade's roughness places it outside Sano et al.'s smooth regime; its stochasticity places it outside Guzman et al.'s deterministic-fractal-generation regime. The bond dimension does not behave like either.

The QTT–renormalization-group connection (Tang et al. 2025) suggests that the bond dimension in the quantics representation may be analytically related to the count of rescaled couplings generated at each coarse-graining step. For a multiplicative cascade, this connection is illuminating in a way worth making explicit. The cascade's correlations are built *top-down* through the generative tree: each scale's randomness is conditionally independent given the parent multipliers at coarser scales, with long-range cross-scale dependencies arising from the multiplicative path-product structure. The QTT representation, by contrast, indexes the state *linearly* by dyadic scale — the K-qubit MPS chain is a sequential unrolling of the tree. The mismatch between the tree-generative structure of the data and the chain-sequential structure of the ansatz is the precise structural reason bond dimension saturates: long-range cross-scale correlations that the cascade builds at one tree level must be propagated across many MPS bonds when the tree is unrolled, and the rank required to carry that information accumulates exponentially with K.

This observation is a feature, not a bug, of the chosen ansatz. It tells us that *any* representation that does not respect the cascade's natural tree topology — including the balanced binary TTN with natural qubit ordering (§4.2), which is itself a tree but a different tree from the cascade's generative tree — should face the same saturation. The corresponding *tree-respecting* representation that matches the cascade's generative structure remains an open empirical question (§6.4). Deriving an analytical prediction for cascade bond-dimension scaling from the QTT-RG framework, and comparing it to the saturation measurement reported here, is an open theoretical question.

6.4 Future work

The measurement reported here constrains the loading-layer question but does not close it. Four open directions follow naturally from the present work, in rough descending order of priority:

- **Tree tensor networks that respect the cascade's generative structure.** The cascade is generated by a binary tree; our deductive TTN argument (§4.2) addresses only the balanced binary TTN with natural qubit ordering, which respects the *dyadic-scale* tree of the QTT indexing but not the *generative* tree of the cascade itself. A TTN ansatz with topology and contraction order matching the cascade's generative tree may exhibit qualitatively different bond-dimension scaling. This is the single most important open empirical question raised by the present measurement; if such a representation compresses the cascade efficiently, the loading-layer question for the multifractal cascade returns to the classical side.
- **Cascade depth beyond K = 20.** Tensor cross interpolation (Núñez Fernández et al. 2022) extends the QTT construction to higher K without forming the full 2^K amplitude vector. A target of K ≈ 40 would test the asymptotic scaling claim against finite-size artifacts that the brute-force K = 4..20 range does not strictly rule out.

- **The full MMAR price process.** The fBm composition $X(t) = B_H(\theta(t))$ has not been measured directly. The conservative direction-of-bias argument (§6.2) suggests adding the composition can only increase representational cost, but direct measurement remains the principled test.
- **A constructive quantum loading proposal.** This paper establishes that classical MPS / balanced-natural-ordering TTN preparation of the cascade state is exponentially expensive. It does not propose, design, or analyze a polynomial-depth quantum circuit that would prepare this state efficiently. Whether such a circuit exists is the natural quantum-side counterpart to the present measurement; constructing a candidate algorithm — a quantum tensor cross interpolation, a variational state preparation, or a circuit exploiting the cascade's generative tree directly — would, if successful, convert the present negative classical constraint into a positive quantum claim. We do not undertake this construction here, but we identify it as the natural next step for any work that seeks to demonstrate the quantum-side counterpart to the loading-layer hardness reported above.

Broader generalizations — alternative multiplier laws including log-Poisson and truncated forms, continuous-time multifractal models (Bacry, Muzy, and Delour 2001), and alternative truncation norms (total variation, KL divergence, sup-norm) — are tracked separately as longer-horizon work.

Acknowledgments

This combined document was prepared with assistance from **Claude (Anthropic)** as a thinking partner across outline development, structural framing, prose composition, and the empirical measurement work reported in Part 2 (cascade construction, bond-dimension measurement implementation, sweep design, plotting, and analysis). The architectural perspective in Part 1 and the empirical results in Part 2 underwent cross-model adversarial review by **Gemini (Google DeepMind)** and **Grok (xAI)** in separate review rounds; substantive reviewer feedback was incorporated into both parts. The author retains responsibility for intellectual content, citation choices, architectural positions, empirical methodology, and any errors. The reviewers were given the documents as standalone artifacts and asked for unsparing critique; their engagement materially improved both parts and the seams between them.

References

Aboussalah, A. M., Chi, C., and Lee, C.-G. (2023). “Quantum computing reduces systemic risk in financial networks.” *Scientific Reports*, 13, Article 3990.

Aghamohammadi, C., Crutchfield, J. P., and Mahoney, J. R. (2018). “Extreme Quantum Memory Advantage for Rare-Event Sampling.” *Physical Review X*, 8(1), 011025.

Ahrens, M., Erdemlioglu, D., McMahon, M., Neely, C. J., and Yang, X. (2025). “Mind Your Language: Market Responses to Central Bank Speeches.” *Journal of Econometrics*, 249, 105921.

Bacry, E., Delour, J., and Muzy, J.-F. (2001). “Multifractal Random Walk.” *Physical Review E*, 64(2), 026103.

Bloomberg (2026, April 26). “Goldman Sachs Scales Back Quantum Computing Effort.” Bloomberg News.

Bouchaud, J.-P., and Potters, M. (2003). *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management.* Cambridge University Press.

Brassard, G., Høyer, P., Mosca, M., and Tapp, A. (2002). “Quantum Amplitude Amplification and Estimation.” *Contemporary Mathematics*, 305, 53–74.

Calvet, L. E., and Fisher, A. J. (2008). *Multifractal Volatility: Theory, Forecasting, and Pricing.* Academic Press.

Carrera-Vazquez, A., and Woerner, S. (2020). "Survey of quantum algorithms for finance: state preparation methods." *IBM Research Zurich Technical Report*.

Chakrabarti, S., Krishnakumar, R., Mazzola, G., Stamatopoulos, N., Woerner, S., and Zeng, W. J. (2021). “A Threshold for Quantum Advantage in Derivative Pricing.” *Quantum*, 5, 463.

Collodel, U. (2025). “Interpreting the Interpreter: Can We Model Post-ECB Conference Volatility with LLM Agents?” arXiv:2508.13635 [econ.GN]. Central Bank of Malta.

Cont, R. (2001). “Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues.” *Quantitative Finance*, 1(2), 223–236.

De Backer, B., Rocha, J. C., Ryckebusch, J., and Schoors, K. (2024). “On the Potential of Quantum Walks for Modeling Financial Return Distributions.” arXiv:2403.19502.

De Backer, B., et al. (2025). “Characterizing Asymmetric and Bimodal Long-Term Financial Return Distributions through Quantum Walks.” arXiv:2505.13019.

Egger, D. J., Gambella, C., Marecek, J., McFaddin, S., Mevissen, M., Raymond, R., Simonetto, A., Woerner, S., and Yndurain, E. (2020). “Quantum Computing for Finance: State-of-the-Art and Future Prospects.” *IEEE Transactions on Quantum Engineering*, 1, 1–24.

European Central Bank (2021). “ECB Guide to Internal Models.” (Targeted Review of Internal Models — TRIM project.) European Central Bank Banking Supervision.

Federal Reserve and Office of the Comptroller of the Currency (2011). “Supervisory Guidance on Model Risk Management.” Board of Governors of the Federal Reserve System SR Letter 11-7 / OCC Bulletin 2011-12.

Fernández-Fuertes, R. (2025). “Monetary Policy Shocks: A New Hope — Large Language Models and Central Bank Communication.” SSRN Working Paper 5669212; BAFFI CAREFIN Working Paper 25257, Bocconi University.

Financial Conduct Authority (2023). Model risk management and internal model expectations, FCA Handbook (PRA/FCA supervisory materials). United Kingdom.

Gatheral, J., Jaisson, T., and Rosenbaum, M. (2018). “Volatility Is Rough.” *Quantitative Finance*, 18(6), 933–949.

Guzman, J., Tiunov, A., and Aolita, L. (2026). "Local Interpolation via Low-Rank Tensor Trains." arXiv:2601.03885.

Herbert, S. (2021). “No quantum speedup with Grover-Rudolph state preparation for quantum Monte Carlo integration.” *Physical Review E*, 103, 063302. arXiv:2101.02240.

Hull, I., Sattath, O., Diamanti, E., and Wendin, G. (2024). *Quantum Technology for Economists.* Springer (Contributions to Economics series). Working paper version: Sveriges Riksbank Working Paper 398; arXiv:2012.04473.

Iaconis, J., Johri, S., and Altman, E. (2024). “Quantum State Preparation of Normal Distributions Using Matrix Product States.” *npj Quantum Information*, 10, Article 15.

Khoromskij, B. N. (2011). "O(d log N)-quantics approximation of N-d tensors in high-dimensional numerical modeling." *Constructive Approximation*, 34(2), 257–280.

Kreps, J., Narkhede, N., and Rao, J. (2011). “Kafka: A Distributed Messaging System for Log Processing.” *Proceedings of the NetDB Workshop on Networking Meets Databases.*

Mandelbrot, B. B. (1963). “The Variation of Certain Speculative Prices.” *Journal of Business*, 36(4), 394–419.

Mandelbrot, B. B., Fisher, A., and Calvet, L. (1997). “A Multifractal Model of Asset Returns.” Cowles Foundation Discussion Paper 1164.

Monetary Authority of Singapore (2013, rev. subsequent). “Guidelines on Risk Management Practices — Market Risk and Internal Models.” MAS supervisory guidance. Singapore.

Mori, T., Mitarai, K., and Fujii, K. (2024). “Efficient State Preparation for Multivariate Monte Carlo Simulation.” arXiv:2409.07336.

Núñez Fernández, Y., Jeannin, M., Dumitrescu, P. T., Kloss, T., Kim, J., Shinaoka, H., von Delft, J., Waintal, X., and Parcollet, O. (2022). "Learning Feynman diagrams with tensor trains." *Physical Review X*, 12(4), 041018.

Padmanabhan, A. (2026a). “Risk Intelligence: A New Era for Capital Markets.” SSRN Working Paper 6584378.

Padmanabhan, A. (2026b). “Risk Intelligence: A New Era for Institutional Finance.” SSRN Working Paper 6615841.

Padmanabhan, A. (2026c). “Treasury PoC V1.5: Forward Feasibility Test of Text-Aware Regime-Adjusted MMAR for 10-Day VaR.” Capital Markets AI working documentation, May 2026.

Pagès, G., Wilbertz, B., and Wilkens, S. (2018). “GPU-Accelerated Quasi-Monte Carlo Methods for Risk Management.” Various conference proceedings and technical reports on GPU acceleration of VaR computation.

Pistoia, M., Wang, S. M., Pham, M., Gunnels, J., Lin, X.-S., and Yusue, C. (2021). “Quantum Machine Learning for Finance.” arXiv:2109.04298.

Plerou, V., Gopikrishnan, P., Amaral, L. A. N., Meyer, M., and Stanley, H. E. (1999). “Scaling of the Distribution of Price Fluctuations of Individual Companies.” *Physical Review E*, 60(6), 6519–6529.

Ran, S.-J. (2020). “Encoding of Matrix Product States into Quantum Circuits of One- and Two-Qubit Gates.” *Physical Review A*, 101, 032310.

Real Logic (2023). “Aeron: Efficient Reliable UDP Unicast, Multicast, and IPC Message Transport.” Open-source project documentation, github.com/real-logic/aeron.

Sano, M. et al. (2026). "Entanglement scaling in matrix product state representation of smooth functions and their shallow quantum circuit approximations." *Physical Review Research*, 8(2).

Schuch, N., Wolf, M. M., Verstraete, F., and Cirac, J. I. (2008). “Entropy Scaling and Simulability by Matrix Product States.” *Physical Review Letters*, 100, 030504.

Silva, T. C., Moriya, K., and Veyrune, R. (2025). “From Text to Quantified Insights: A Large-Scale LLM Analysis of Central Bank Communication.” IMF Working Paper 25/109. International Monetary Fund.

Solace (2023). “PubSub+ Platform for Capital Markets: Event-Driven Architecture for Trading and Risk.” Solace technical documentation and capital-markets reference materials.

Soleimani, M. (2025). “LLM-Generated Counterfactual Stress Scenarios for Portfolio Risk Simulation via Hybrid Prompt-RAG Pipeline.” arXiv:2512.07867 [q-fin.RM].

Stamatopoulos, N., Egger, D. J., Sun, Y., Zoufal, C., Iten, R., Shen, N., and Woerner, S. (2020). “Option Pricing Using Quantum Computers.” *Quantum*, 4, 291.

Stamatopoulos, N., and Zeng, W. J. (2024). “Derivative Pricing Using Quantum Signal Processing.” *Quantum*, 8.

Tang, H. et al. (2025). "Entanglement across scales: Quantics tensor trains as a natural framework for renormalization." arXiv:2507.19069.

Verstraete, F., Murg, V., and Cirac, J. I. (2008). “Matrix Product States, Projected Entangled Pair States, and Variational Renormalization Group Methods for Quantum Spin Systems.” *Advances in Physics*, 57(2), 143–224.

Woerner, S., and Egger, D. J. (2019). “Quantum Risk Analysis.” *npj Quantum Information*, 5, Article 15.

Zoufal, C., Lucchi, A., and Woerner, S. (2019). “Quantum Generative Adversarial Networks for Learning and Loading Random Distributions.” *npj Quantum Information*, 5, Article 103.

Appendix: Acronyms and terminology

Acronyms organized by domain. Each entry: term, expansion, brief definition.

Quantum compute

- **AE** — Amplitude Estimation. Quantum algorithm providing quadratic convergence improvement (1/N vs classical 1/√N) for expectation values of structured distributions.
- **Bond dimension** — Parameter in matrix product state representation that bounds the entanglement structure a state can represent. States with bounded bond dimension are classically tractable.
- **MPS** — Matrix Product State. Hierarchical representation of quantum states that exploits limited entanglement structure for compact encoding and efficient classical simulation when bond dimension is bounded.
- **NISQ** — Noisy Intermediate-Scale Quantum. Current era of quantum hardware: limited qubit counts, significant noise rates, not yet fault-tolerant.
- **PEC** — Probabilistic Error Cancellation. Quantum error mitigation technique requiring expert-tuned parameters per circuit family.
- **QAE** — see AE.
- **QAOA** — Quantum Approximate Optimization Algorithm. Quantum algorithm for combinatorial optimization problems including portfolio selection under cardinality and integer constraints.
- **QPU** — Quantum Processing Unit.
- **Tucker decomposition** — Hierarchical tensor decomposition method related to matrix product states, used in classical and quantum state preparation.
- **ZNE** — Zero-Noise Extrapolation. Quantum error mitigation technique requiring expert-tuned parameters per circuit family.
- **QTT** — Quantics Tensor Train. Tensor-train representation indexing a function on a 2^K grid by dyadic scale; equivalent to a matrix product state on K qubits for the amplitude state. Bond dimension in this representation is the Schmidt rank of bipartitions between coarse and fine scales.
- **TTN** — Tree Tensor Network. Tensor-network representation with tree topology rather than chain topology; bond dimension at each internal edge is the Schmidt rank of the bipartition induced by removing that edge.
- **TCI** — Tensor Cross Interpolation. Method for constructing tensor-train representations of functions sampled at strategically chosen points, avoiding the need to form the full exponentially-large amplitude vector.
- **SVD** — Singular Value Decomposition. Standard matrix factorization underlying the measurement of bond dimension; singular value count above a tolerance ε measures the rank of the bipartition matrix.
- **fBm** — Fractional Brownian motion. Gaussian, monofractal process indexed by the Hurst exponent H; composes with the multifractal cascade to produce the MMAR price process X(t) = B_H(θ(t)).

AI and orchestration

- **LLM** — Large Language Model. Class of AI systems that produce structured outputs from textual inputs.
- **RAG** — Retrieval-Augmented Generation. LLM technique that grounds outputs in a retrievable corpus of source documents.

Risk frameworks and metrics

- **CVaR** — Conditional Value at Risk. Same as Expected Shortfall.
- **ES** — Expected Shortfall. Expected loss conditional on the loss exceeding a Value at Risk threshold.
- **FRTB** — Fundamental Review of the Trading Book. BCBS regulatory framework for market risk capital, in effect at most global banks.
- **FRTB IMA** — FRTB Internal Models Approach. The internal-model alternative to standardized capital calculation under FRTB.
- **SA-CCR** — Standardized Approach for Counterparty Credit Risk. BCBS framework for counterparty exposure measurement.
- **SIMM** — Standard Initial Margin Model. ISDA framework for non-cleared derivatives initial margin.
- **VaR** — Value at Risk. Loss quantile at a specified confidence level (e.g., Value at Risk 99% = loss exceeded with 1% probability over the horizon).

Multifractal mathematics

- **Cascade** — Recursive multiplicative construction producing a multifractal measure with scale-invariant generative structure.
- **Hurst exponent** — Parameter characterizing long-memory in stochastic processes; values above 0.5 indicate positive long-memory.
- **MMAR** — Multifractal Model of Asset Returns (Mandelbrot, Fisher, Calvet 1997). Cascade-based generative model for fat-tailed long-memory financial returns.
- **MRW** — Multifractal Random Walk (Bacry, Delour, Muzy 2001). Continuous-parameter multifractal model.

Regulatory and supervisory

- **BCBS** — Basel Committee on Banking Supervision. Issues international banking regulatory standards.
- **BCBS 248** — Specific BCBS standard for intraday liquidity monitoring.
- **FCA** — Financial Conduct Authority. UK financial regulator.
- **ICAAP** — Internal Capital Adequacy Assessment Process. Required capital adequacy framework under prudential regulation.
- **ILAAP** — Internal Liquidity Adequacy Assessment Process. Required liquidity adequacy framework.
- **MAS** — Monetary Authority of Singapore.
- **SR 11-7** — U.S. Federal Reserve Supervisory Guidance on Model Risk Management.
- **TRIM** — Targeted Review of Internal Models. European Central Bank framework for internal model approval.

Capital markets

- **BSM** — Black-Scholes-Merton. Standard derivative-pricing framework assuming log-normal underlying asset distributions.
- **CCP** — Central Counterparty. Clearing entity for derivative contracts.
- **FOMC** — Federal Open Market Committee. The U.S. Federal Reserve body that sets monetary policy.
- **G-SIB** — Global Systemically Important Bank. BCBS designation for the largest internationally active banks.
- **OMS** — Order Management System. Trading-system component handling order lifecycle.
- **SIFI** — Systemically Important Financial Institution.
- **T-bill** — U.S. Treasury Bill. Short-maturity government debt instrument.
- **UST** — U.S. Treasury.
- **XVA** — Collective acronym for valuation adjustments (CVA, DVA, FVA, KVA, MVA) applied to derivative valuations.

Architecture-specific terminology

- **Bar 2 forward discipline** — Methodological discipline (Capital Markets AI Treasury PoC V1.5) requiring score lock before any realized data exists in the forward evaluation window.
- **Event mesh** — Messaging infrastructure providing high-throughput, low-latency, schema-governed event distribution across heterogeneous systems.
- **L0 through L5** — Layer numbering in the architecture proposed here. L4.5 designates the integration substrate between application and compute layers.