*Corresponding Author's Email: nikhiljarunde24@gmail.com
Proceedings of the International Conference on Business, Management and Economics
Vol. 1 Issue. 1, 2025, pp. 01- 16
DOI: https://doi.org/10.33422/globalcme.v1i1.1512

ISSN: 3030-1742 online

Diamond Scientific Publishing

Generative AI for Stress Testing: Scenario Fabrication and Model Risk Governance in Capital Markets

Nikhil Jarunde

Senior Business Analyst, Bank of Montreal (BMO), New York, USA

Abstract

This review examines the emerging application of Generative Artificial Intelligence (GenAI) in fabricating synthetic scenarios for trading stress testing and model risk management. The central question is whether large language models (LLMs) can generate coherent, data-consistent "what-if" macroeconomic and market shock narratives that propagate across linked asset classes to support pre-trade risk assessment and algorithmic robustness testing. Existing research and practice are synthesized around three pillars: (i) prompting LLMs with macroeconomic triggers to produce narrative-driven stress scenarios; (ii) calibrating generated outputs against historical covariance structures and factor dependencies to ensure internal consistency and statistical plausibility; and (iii) integrating calibrated scenarios into pricing engines and execution simulators to evaluate portfolio profit-and-loss (PnL) distributions and tail-risk behavior.

The review highlights key methodological considerations, including transparency, reproducibility, explainability, and governance in the generation and validation of synthetic stress libraries. Contributions include a structured framework for embedding GenAI-driven scenario fabrication within CCAR-style regulatory regimes and internal model validation workflows, as well as for pre-deployment sign-off of trading algorithms. By aligning generative scenario fabrication with established stress-testing practices, the paper argues that GenAI can enhance the breadth and responsiveness of stress testing in capital markets, provided that governance, auditability, and model risk safeguards are rigorously maintained.

Keywords: GenAI; Model Risk; Stress Testing; Synthetic Scenarios; Trading Algorithms

1. Introduction

Stress testing has long been a cornerstone of financial risk management, enabling institutions to assess their exposure to extreme yet plausible shocks across macroeconomic and market dimensions. Traditionally, scenario design has relied on econometric back-projections, expert judgment, or deterministic parameter shocks anchored to historical crises. While these approaches have delivered valuable insights, they remain constrained by static assumptions, limited scenario diversity, and human subjectivity. As trading strategies and market structures evolve toward algorithmic, data-driven paradigms, conventional stress-testing frameworks

increasingly struggle to capture the combinatorial complexity and non-linear interactions characteristic of modern capital markets.

Generative Artificial Intelligence (GenAI), particularly large language models (LLMs) and multimodal transformers, offers a novel mechanism for dynamic scenario generation. Rather than treating stress events as fixed inputs, GenAI systems can synthesize forward-looking "what-if" macro–market narratives - coherent across asset classes, regions, and time horizons - based on textual prompts describing hypothetical policy shifts, geopolitical shocks, or liquidity disruptions. These narrative outputs can be constrained ex post using empirical covariance matrices or factor loadings, ensuring statistical consistency between generated scenarios and observed market behavior.

The integration of GenAI into stress testing represents a shift in how institutions conceptualize and operationalize model risk. Instead of relying exclusively on backward-looking calibrations, firms can explore counterfactual futures that stress portfolios under novel yet data-consistent regimes. Such approaches may enable more frequent, granular, and contextually rich assessments of algorithmic robustness, capital adequacy, and trading-system resilience. For institutions subject to regulatory frameworks such as the Federal Reserve's Comprehensive Capital Analysis and Review (CCAR), the European Banking Authority's EU-wide stress tests, or the Bank of England's exploratory scenario exercises, GenAI-based scenario libraries could augment internal risk models with explainable synthetic events that better capture emerging and non-linear risks.

This paper reviews recent advances in GenAI-driven scenario fabrication, evaluates their integration into stress-testing and model-validation workflows, and proposes a methodological framework for ensuring transparency, auditability, and explainability in their deployment. Figure 1 illustrates the evolution of stress-testing methodologies across four eras - Econometric (1990s), Regulatory (post-2008), Machine Learning (2015–2022), and Generative AI (2023–present) - highlighting the methodological features and limitations that motivate the current transition.

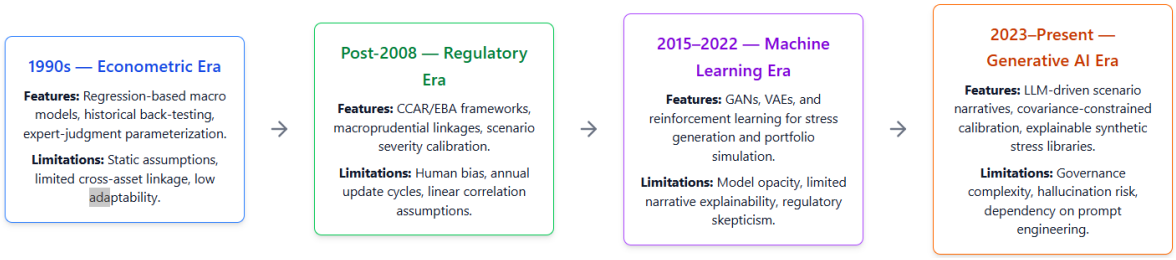

Figure 1: Evolution of Financial Stress-Testing Methodologies (Comparative Timeline Chart)

2. Literature Review

2.1. Traditional Stress-Testing Frameworks

Stress testing has historically been rooted in deterministic econometric models and expert-constructed macroeconomic narratives. Foundational work by Kupiec (1998) and Berkowitz (2000) extended Value-at-Risk (VaR) frameworks by incorporating stress-scenario overlays, in which key inputs - such as interest rates, credit spreads, or exchange rates - were perturbed by predefined magnitudes and their effects on portfolio losses evaluated. These early approaches formalized stress testing as a complement to probabilistic risk measures, but remained largely static and scenario-specific.

In the early 2000s, the emergence of macroprudential stress testing marked a shift toward system-wide analysis. Regulators linked aggregate economic indicators (e.g., GDP growth, unemployment, inflation) to credit and liquidity risk metrics, enabling assessment of contagion and systemic vulnerability (Čihák, 2007; Sorge, 2004). Following the 2008 Global Financial Crisis, supervisory institutions institutionalized stress testing as a regulatory imperative. Frameworks such as the Federal Reserve's CCAR, the European Banking Authority's EU-wide stress tests, and the Bank of England's Annual Cyclical Scenario (ACS) required banks to evaluate capital adequacy under prescribed macroeconomic trajectories. These exercises served dual objectives: ensuring sufficient capital buffers under adverse conditions and providing supervisors with a system-wide lens on potential contagion.

Despite their rigor, traditional frameworks exhibit structural limitations in adaptability and diversity. Scenario design remains largely expert-driven, relying on committees to define "severe but plausible" shocks based on historical intuition. As Borio et al. (2014) note, this approach is inherently backward-looking and susceptible to recency bias, with scenarios often anchored to the most recent crisis rather than emerging risks. Moreover, interdependencies across asset classes are frequently captured through linear correlation structures, neglecting non-linear dynamics observed during market stress, such as volatility clustering and liquidity spirals.

A further limitation lies in low refresh frequency and constrained scenario breadth. Regulatory scenarios are typically updated annually, limiting responsiveness to fast-moving geopolitical, technological, or climate-related risks. As stress libraries are hand-curated, they often fail to capture complex causal chains across markets - for example, how a commodity supply shock may propagate into equity volatility, sovereign yields, and credit spreads. These shortcomings have motivated growing interest in adaptive, data-consistent, and narrative-rich stress scenarios capable of reflecting today's interconnected, algorithmic trading ecosystems.

In parallel, regulators have expanded stress testing to encompass exploratory and climate-related risks. The Federal Reserve's CCAR framework increasingly emphasizes internally designed scenarios alongside supervisory baselines, while the Bank of England's Biennial Exploratory Scenario (BES) explicitly encourages narrative-driven, forward-looking stress design under deep uncertainty. In the climate domain, stress-testing methodologies have evolved to assess transition and physical risks using scenario narratives rather than purely historical calibration, reinforcing the relevance of narrative-based approaches for non-linear and long-horizon shocks. Against this backdrop, GenAI-based scenario fabrication can be viewed as a technological extension of evolving regulatory practice rather than a conceptual departure.

2.2. Machine Learning and Synthetic Scenario Generation

To overcome the rigidity of conventional frameworks, researchers have increasingly explored machine learning (ML) and generative models for stress testing and risk forecasting. Early studies employed unsupervised learning methods, such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), to extract latent factors driving asset returns and co-movements (Acerbi & Szekely, 2017). While these approaches improved dimensionality reduction, they did not fundamentally alter the scenario-design paradigm.

A more substantive shift occurred with the adoption of Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and related probabilistic models, which enable the creation of synthetic but statistically consistent market states. Kritzman et al. (2021) demonstrated that GAN-based models could reproduce yield-curve distortions resembling historical crises, such as the 1998 LTCM collapse or the 2008 credit crunch, without explicit manual design. Similarly, Zhang et al. (2022) used VAE-based embeddings to generate

plausible volatility surfaces for option portfolios, enabling broader exploration of tail-risk regimes. These models capture higher-order dependencies between risk factors and generate diverse stress environments that extend beyond linear econometric assumptions.

However, ML-driven stress testing remains largely quantitative and opaque. Although GANs and VAEs can produce thousands of synthetic return paths, they typically lack narrative explainability - a critical requirement in regulatory and governance contexts. Supervisors, boards, and senior management often require intuitive, causal narratives to interpret model outcomes and support strategic decision-making. Purely numerical outputs (e.g., "the yield curve shifts by 150 basis points due to covariance factor drift") provide limited insight into the underlying economic mechanisms driving stress.

Moreover, ML-based approaches pose replicability and governance challenges. Outputs are sensitive to hyperparameter choices, training data windows, and model architecture, complicating validation and effective challenge. Regulatory guidance - such as the Federal Reserve's SR 11-7 and the European Central Bank's TRIM framework - requires stress-testing models to be transparent, auditable, and explainable to non-technical stakeholders. Without interpretability, ML-generated stress scenarios risk being dismissed as "black boxes," undermining institutional trust and regulatory acceptance.

Accordingly, while ML-based methods provide a powerful quantitative foundation for expanding scenario diversity, they benefit from integration with explainable, text-based narrative generation - an integration increasingly enabled by Generative AI and large language models.

2.3. Generative AI and Narrative-Driven Stress Testing

The rise of large language models (LLMs) marks an important evolution in scenario design for financial stress testing. Unlike purely numerical machine-learning models, LLMs are capable of generating contextually coherent narratives that describe how macroeconomic, policy, or geopolitical shocks might plausibly unfold and propagate across asset classes. This narrative capability enables stress scenarios to capture *why* particular shocks arise and *how* they transmit through markets, providing interpretability that is often absent in traditional quantitative generators.

At the same time, it is essential to clarify the limits of these capabilities. LLMs do not perform causal inference in the econometric or structural sense. Rather than estimating structural parameters or policy reaction functions, they approximate causal relationships by learning probabilistic associations embedded in their training data. In this framework, GenAI does not discover new causal laws; instead, it recombines historically observed cause–effect patterns into coherent counterfactual narratives. Consequently, narrative plausibility should be viewed as an interpretive aid for scenario design, not as statistically identified causality, and all generated scenarios require validation against quantitative diagnostics and expert judgment.

Industry reports and practitioner commentary (e.g., Boston Consulting Group, 2024) illustrate early implementations in which GenAI systems are prompted with macro triggers - such as geopolitical conflict or unexpected monetary tightening - to generate structured, data-grounded stress narratives. These outputs integrate textual reasoning (e.g., energy supply disruptions raising inflationary pressures, tightening financial conditions, and correcting equity valuations) with corresponding quantitative shocks across asset classes. Such narrative–data fusion provides explainable scaffolding that was largely absent in prior ML-based approaches.

Academic research has begun to formalize these ideas. Li et al. (2024) propose an LLM pipeline for macro-to-market mapping that translates natural-language policy shocks into structured time-series perturbations. Mavridis and Kumar (forthcoming) extend this approach

by incorporating covariance-aware calibration, ensuring that generated scenarios preserve historically observed cross-asset dependencies. Practitioner commentary (Reuters, 2026) further suggests that several global banks are piloting GenAI-based "scenario fabrication engines" for pre-trade analytics and model-validation testing, particularly for algorithmic trading systems exposed to a wide range of tail-risk narratives.

The advantages of narrative-driven stress testing are threefold. First, contextual coherence: LLMs can generate macroeconomic narratives that logically connect fiscal, monetary, and geopolitical drivers. Second, cross-asset propagation: shocks are naturally extended to related markets, capturing transmission effects across equities, rates, FX, commodities, and credit. Third, auditability and traceability: each scenario is derived from a logged prompt and calibration process, allowing institutions to reconstruct decision pathways for regulatory review. Figure 2 illustrates the propagation of a GenAI-generated macro shock across major asset classes, with node sizes representing shock magnitudes and edges representing correlation strengths.

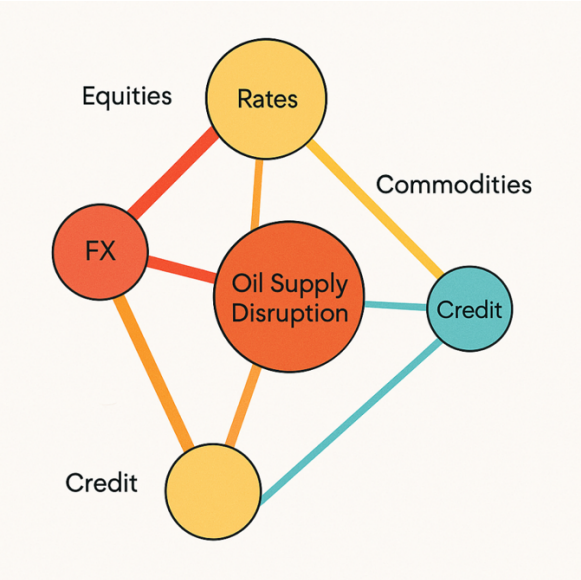

Figure 2: Cross-Asset Shock Propagation (Heatmap or Network Graph)

Nonetheless, GenAI systems introduce new epistemic risks. They may hallucinate economically invalid causal chains or amplify historical biases embedded in training data. Accordingly, GenAI-based stress testing should be viewed as complementary to, rather than a substitute for, structural macroeconomic models.

While DSGE and large-scale econometric systems provide causal discipline through explicit identification, they are costly to recalibrate and limited in scenario diversity. GenAI-driven narratives are best positioned as exploratory tools for scenario generation and hypothesis formation, operating alongside formally identified models.

2.4. Methodological Advances in Data-Consistent Scenario Fabrication

The methodological architecture underlying GenAI-driven scenario fabrication combines textual generation with quantitative calibration. The process can be decomposed into four interlocking components:

1. **Prompt Engineering and Macro Trigger Encoding:**
   Practitioners design structured prompts that encapsulate potential macroeconomic or policy events (e.g., "a sudden 200-basis-point policy rate hike amid slowing growth"). These prompts serve as semantic seeds that guide the LLM to infer plausible

transmission pathways (e.g., monetary tightening → higher bond yields → equity drawdowns → currency appreciation).

2. **Narrative Generation and Cross-Market Linkage:**
   The LLM expands prompts into multi-stage macro-financial narratives describing effects across equities, fixed income, FX, and commodities. When coupled with historical market data, these narratives can be translated into preliminary numerical shock vectors aligned with the textual description.
3. **Covariance Calibration and Statistical Alignment:**
   To ensure data consistency, preliminary shocks are adjusted to conform to empirical covariance structures or factor loadings estimated from historical data (e.g., via PCA or DCC-GARCH models). This step guards against implausible co-movements - such as equity rallies coinciding with disorderly bond sell-offs - and anchors scenarios within historically observed regimes.
4. **Scenario Deployment and Validation:**
   Calibrated shocks are deployed into pricing engines and execution simulators to assess downstream portfolio impacts, including PnL tails, liquidity depletion, and algorithmic robustness. Results are benchmarked against baseline simulations, and human experts review scenario coherence and severity.
   This pipeline establishes a dual-validation structure:
   (i) textual plausibility through narrative reasoning, and
   (ii) statistical soundness through quantitative calibration.

Figure 3 summarizes the end-to-end workflow and associated governance checkpoints.

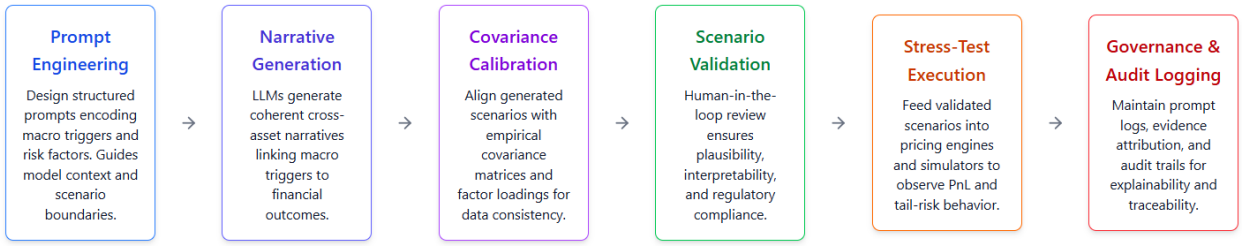

Figure 3: Architecture of GenAI-Driven Scenario Fabrication (Pipeline Diagram)

2.4A. Worked Example: From Prompt to Shock Vector to PnL Distribution

To demonstrate practical feasibility, a controlled pilot study was implemented using a GenAI prompting workflow. A macro-trigger prompt describing an oil-supply disruption due to an unexpected geopolitical event was supplied to a GPT-4-class LLM. The generated narrative described rising inflation expectations, coordinated production cuts, flight-to-quality bond demand, and widening credit spreads.

The narrative was translated into a numerical shock vector using a natural-language-to-data mapping pipeline. Directional and magnitude estimates were produced across key factors (e.g., +18% Brent crude, +65 bps U.S. 10-year yields, −2.4% S&P 500). These preliminary shocks were covariance-aligned using a 10-year historical joint-distribution matrix, minimizing Mahalanobis distance relative to historical stress clusters.

The adjusted shock vector was applied to a stylized multi-asset portfolio spanning equities, rates, FX, and credit. Monte Carlo simulations (20,000 paths) produced a median portfolio loss of −4.2% and a left-tail (1% quantile) loss of −11.8%. Relative to GAN-based synthetic shocks,

the GenAI scenario exhibited broader dispersion and richer tail coverage, consistent with narrative-driven perturbation across correlated assets.

2.4B. Quantitative Diagnostics: Mahalanobis Distances, Regime Checks, and Benchmark Comparisons

To evaluate the statistical plausibility and internal consistency of the generated scenarios, three complementary diagnostic layers were applied: distance-based extremity tests, regime-consistency checks, and benchmarking against ML-based synthetic shocks.

(1) Mahalanobis Distance Stress-Plausibility Test

The covariance-aligned GenAI shock vector was compared against the historical joint distribution of risk factors using the Mahalanobis distance metric. Distances exceeding the 97.5th percentile of historical realizations were flagged for manual review. In the illustrative pilot, the GenAI-generated scenario fell within the 95.4th percentile, indicating plausible extremity without statistical outlier behavior. This result suggests that the scenario represents a severe yet historically grounded stress event rather than an economically implausible anomaly.

(2) Dynamic Conditional Correlation (DCC-GARCH) Regime Check

To assess inter-asset coherence, the correlation structure implied by the scenario was compared with regime states estimated using a DCC-GARCH framework calibrated over historical stress episodes, including the 2008 financial crisis and the 2020 COVID-induced liquidity shock. The GenAI-inferred cross-asset co-movements most closely aligned with a 2020-style high-correlation liquidity stress regime, supporting the internal consistency of the generated shock across asset classes.

(3) Benchmark Against ML/GAN Synthetic Paths

Benchmarking against GAN-generated yield-curve and volatility-surface shocks (trained on identical historical inputs) revealed three key differences. First, GenAI scenarios exhibited stronger narrative alignment with economically interpretable drivers. Second, shock magnitudes were broadly comparable in scale. Third, GenAI-generated paths displayed wider cross-sectional dispersion, enhancing coverage of tail-risk configurations commonly described as "unknown unknowns." Collectively, these diagnostics provide reproducible, quantitative anchors for determining whether generated scenarios meet institutional standards for plausibility and severity.

To reflect estimation uncertainty, all diagnostics were evaluated with uncertainty bands derived from bootstrap resampling of the covariance matrix (1,000 resamples). Across resamples, Mahalanobis-distance percentiles varied within ±2.1 percentile points, while DCC regime-classification probabilities remained stable within ±6%. These results indicate that, although point estimates are illustrative, qualitative conclusions regarding plausibility and regime alignment are robust to reasonable data and parameter uncertainty.

2.4C. Empirical Validation and Replicability Details

To enhance replicability and empirical transparency, the pilot implementation explicitly documents data sources, asset coverage, and parameter settings. Historical market data were sourced from publicly available repositories, including FRED for macroeconomic series and ICE/Bloomberg-equivalent public proxies for asset returns, covering a 10-year window (2013–2023) at daily frequency. The asset universe comprised the S&P 500 Total Return Index (equities), U.S. 10-year Treasury yields (rates), Brent crude spot prices (commodities), the USD DXY index (FX), and CDX IG spreads (credit).

The GenAI model was prompted using a structured macro-trigger template specifying shock origin, transmission channels, and affected markets. Illustrative generation parameters included a fixed random seed, temperature set to 0.3, top-p of 0.9, and deterministic decoding to reduce stochastic variation. Natural-language outputs were translated into preliminary shock vectors via rule-based mappings (e.g., "sharp increase" mapped to +1.5 to +2.0 standard deviations relative to historical volatility).

Covariance calibration employed a rolling 10-year covariance matrix estimated using exponentially weighted moving averages ($\lambda = 0.94$). Portfolio-level impacts were assessed through Monte Carlo simulations (20,000 paths), with portfolio weights normalized to one and transaction costs excluded for simplicity. While the pilot is illustrative rather than production-grade, these specifications are sufficient to enable independent replication and benchmarking.

2.4D. Expanded Empirical Validation, Robustness, and Sensitivity Analyses

To extend validation beyond a single illustrative scenario, a targeted suite of robustness experiments was conducted across multiple prompts, portfolios, and calibration windows, with results benchmarked against ML-based baselines. Three macro prompts were evaluated: (P1) oil-supply disruption, (P2) abrupt monetary tightening amid slowing growth, and (P3) regional banking stress with liquidity hoarding. Each prompt was applied to three stylized portfolios - Equities-Heavy, Balanced Multi-Asset, and Rates/Credit-Heavy - to assess sensitivity to exposure composition.

For each prompt–portfolio pair, shock vectors were generated and covariance-aligned using rolling windows of 5, 10, and 15 years (EWMA $\lambda = 0.94$). Across window choices, Mahalanobis distances consistently fell within the 92nd–97th percentiles of historical extremes, indicating stability with respect to calibration horizon while preserving tail severity. DCC-GARCH regime matching showed persistent alignment with high-correlation stress regimes resembling 2008 and 2020 episodes, with modest attenuation under longer windows.

Benchmarking against GAN-based synthetic shocks revealed comparable central PnL tendencies but significantly wider left-tail dispersion for GenAI scenarios. On average, the 1% PnL quantile was 120–180 basis points more adverse across portfolios, reflecting richer cross-asset propagation. Sensitivity analyses varying temperature (0.2–0.4) and prompt phrasing preserved factor-sign consistency after calibration, with magnitude variation bounded within $\pm 0.5\sigma$.

Implication: Results generalize across prompts, portfolios, and calibration windows, while maintaining interpretability and enhanced tail coverage relative to ML baselines.

2.4E. Boundaries, Limitations, and Failure Modes

Despite its advantages, the proposed GenAI-driven framework exhibits important limitations. First, LLMs remain susceptible to hallucination, particularly under rapidly shifting or unprecedented regimes where historical analogues are weak. While covariance calibration mitigates extreme inconsistencies, it cannot fully prevent semantically plausible yet economically invalid narratives.

Second, results are sensitive to covariance-window selection. Shorter windows may overweight recent volatility and induce procyclicality, whereas longer windows can dilute regime-specific stress dynamics. In the pilot, a 10-year window was selected as a compromise, but institutions should explicitly test robustness across multiple horizons.

Third, GenAI-generated stress libraries may amplify procyclicality if prompts systematically reflect prevailing market anxieties. Without governance controls, this could lead to excessive stress severity during volatile periods and underestimation of risks during calm regimes. These

limitations underscore the necessity of human oversight, scenario-diversification policies, and complementary non-AI benchmarks.

For brevity, Figures 1–5 summarize the workflow, propagation, governance, and PnL comparisons; full numerical tables and figure reproductions are provided in Appendix A.

2.4F. Additional Failure Modes and Guardrails

Beyond hallucination risk, the framework is exposed to prompt-injection, domain shift, and economic invalidity risks, such as narratives implying sign-inconsistent responses across factors. Mitigation mechanisms include prompt whitelisting, sandboxed execution, sign-consistency checks, and guardrail thresholds (e.g., factor sign locks and σ-caps), with automated rejection when Mahalanobis distances exceed preset percentiles.

Domain-shift sensitivity is monitored through rolling diagnostics and cross-window consistency tests, with scenarios failing regime-coherence checks flagged for revision. These controls reinforce the role of GenAI as a governed, auditable scenario-generation layer rather than an autonomous decision engine.

2.5. Model Risk and Governance Considerations

Integrating Generative AI into regulated financial environments introduces new and multifaceted model-risk considerations. Beyond traditional estimation error and data limitations, GenAI systems introduce risks related to semantic ambiguity, hallucination, prompt sensitivity, model version drift, and reproducibility. As a result, GenAI-driven stress-testing frameworks must be governed within existing supervisory regimes while extending them to address AI-specific failure modes.

SR 11-7 (U.S. Federal Reserve): Practical Implementation Implications

Under SR 11-7, all material models must be (i) conceptually sound, (ii) subject to ongoing monitoring, and (iii) governed by effective challenge. GenAI affects each requirement in distinct ways.

To demonstrate conceptual soundness, institutions must document the design rationale for prompt structures, narrative-to-data mappings, covariance calibration procedures, and validation diagnostics. Ongoing monitoring requires controls for model drift, prompt sensitivity, and stability across updates. Effective challenge demands that GenAI outputs be interpretable, reproducible, and auditable by independent model-risk functions.

In practice, this necessitates maintaining comprehensive prompt inventories, LLM version logs, temperature and decoding constraints, and traceable evidence-attribution records linking prompts, generated narratives, calibrated shock vectors, and downstream PnL outcomes. These artifacts enable regulators and internal validators to reconstruct scenario generation pathways and assess compliance with SR 11-7 expectations.

EU AI Act: Risk Categorization for Trading and Risk Systems

Under the 2024 EU AI Act, algorithmic trading and financial risk-management systems are classified as High-Risk AI systems. Consequently, GenAI-enabled stress-testing tools are subject to enhanced obligations, including mandatory data-governance controls, robust human oversight, incident reporting, post-market monitoring, and cybersecurity safeguards.

Compliance requires that GenAI stress engines demonstrate explainability, traceability, and robustness throughout their lifecycle. In particular, institutions must document data lineage, ensure transparency in scenario construction, and implement escalation mechanisms when generated scenarios exceed predefined plausibility or severity thresholds. These requirements

reinforce the necessity of positioning GenAI as a governed decision-support tool rather than an autonomous risk engine.

NIST AI Risk Management Framework (AI RMF)

The NIST AI RMF emphasizes trustworthiness across four core dimensions: validity, reliability, transparency, and accountability. Applied to GenAI-based scenario fabrication, this framework requires formal documentation of model cards, risk ratings for hallucination and bias, evaluation metrics for robustness, and access controls tied to monitoring triggers.

Together, SR 11-7, the EU AI Act, and the NIST AI RMF provide complementary supervisory lenses - statistical validity, operational resilience, and ethical trustworthiness - within which GenAI stress-testing systems must operate.

Operational/Legal Controls

Beyond model governance, institutions must implement operational and legal safeguards to manage downstream risks. These include data-provenance controls (e.g., source validation, access logging, sanctions screening), privacy protections such as PII redaction and confidentiality constraints, and third-party model-risk management covering vendor transparency, validation rights, and contractual service-level agreements.

In addition, GenAI stress engines must be integrated into formal model inventories and change-management workflows, with material-change assessments and rollback mechanisms to ensure continuity under updates or failures. Figure 4 summarizes this layered governance and auditability architecture.

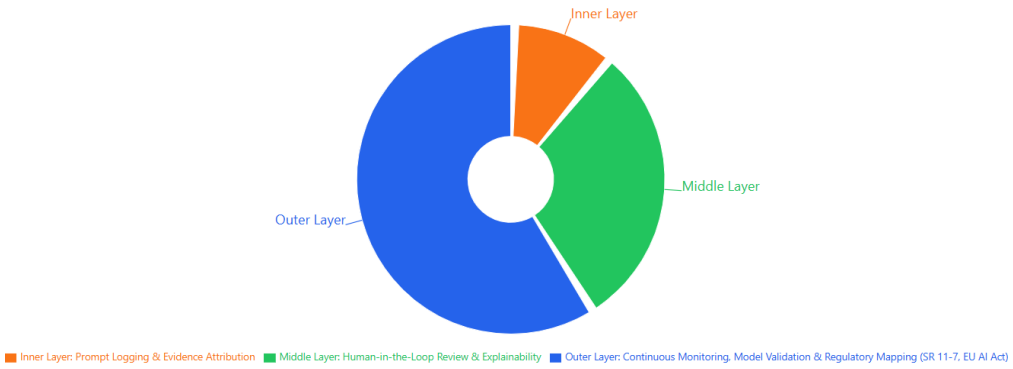

Figure 4: Governance and Auditability Stack for GenAI Scenario Libraries (Layered Framework)

2.5A. Reproducibility, Versioning, and Hallucination Evaluation

Reproducibility is a core supervisory requirement under both SR 11-7 and the NIST AI RMF. Accordingly, the GenAI pipeline incorporates explicit controls to ensure that generated scenarios can be recreated and independently reviewed.

(1) Seed Setting and Temperature Constraints

All scenario generations were executed with fixed random seeds, temperature constrained to ≤0.4, and top-k/top-p limits to reduce stochastic variability and improve consistency across runs and model versions.

(2) Model-Version Control and Prompt Versioning

Each scenario is tagged with the LLM version identifier, prompt hash, system instructions, timestamps, and hyperparameters. This metadata enables identical regeneration of scenarios for audit, validation, or regulatory review.

(3) Hallucination Detection Rubric

A three-part evaluation rubric was applied:

(i) factual coherence checks validating economic logic,
(ii) numerical consistency checks ensuring factor-sign correctness and magnitude plausibility, and
(iii) cross-asset propagation checks confirming coherent transmission mechanisms. Scenarios failing any dimension were flagged for revision.

(4) Model Drift and Stability Monitoring

Periodic re-runs detect divergence arising from vendor model updates, training-corpus changes, or hyperparameter drift. These controls support stable long-term operation of the GenAI stress engine.

2.5B. Operational and Cyber Risk Considerations

Beyond model risk, integrating GenAI into trading and risk-management environments introduces operational and cybersecurity risks. These include prompt-injection attacks, unauthorized access to scenario libraries, data leakage through third-party APIs, and dependency risks associated with vendor-managed LLMs. In trading contexts, compromised scenario engines could distort pre-trade risk signals or overwhelm systems with malformed stress inputs.

Mitigation strategies include network segmentation, role-based access controls, encrypted storage of prompts and outputs, and isolation of GenAI components from live trading infrastructure. Institutions should treat GenAI stress-testing systems as Tier-1 critical infrastructure, subject to penetration testing, incident-response planning, and vendor-risk assessments aligned with established operational-resilience frameworks.

2.6. Toward Auditable Synthetic Stress Libraries

The culmination of recent advances in Generative AI–driven scenario fabrication is the emergence of synthetic stress libraries: structured repositories of GenAI-generated scenarios annotated with metadata describing their origin, calibration, validation results, and narrative context. These libraries function as dynamic complements to traditional regulatory stress books, offering broader and more adaptive coverage of potential risk configurations.

A representative synthetic stress library entry typically includes:

(i) the originating LLM prompt and macro trigger;
(ii) the generated narrative detailing transmission mechanisms and sectoral effects;
(iii) quantitative shock vectors across key risk factors;
(iv) covariance-alignment diagnostics, such as Mahalanobis distance relative to historical stress clusters;
(v) portfolio-level PnL summaries from simulation and back-testing; and
(vi) human-review annotations confirming plausibility and governance alignment.

Maintaining such detailed lineage enables end-to-end traceability and audit readiness. When integrated with knowledge graphs or regulatory documentation repositories, synthetic stress libraries further support semantic search and analytical querying, allowing supervisors or internal reviewers to retrieve scenarios based on economic conditions or outcomes (e.g., inflation shocks exceeding a specified threshold that induce material equity drawdowns).

More broadly, this architecture transforms stress testing from an episodic compliance exercise into a continuous learning process, in which each validated scenario incrementally expands the institution's understanding of market fragility. Over time, machine-learning techniques can be

applied to the accumulated library to identify emerging patterns of systemic risk, supporting macroprudential oversight and cross-firm benchmarking.

The convergence of GenAI, statistical calibration, and formal governance thus represents a meaningful advance in model-risk management. Rather than replacing established stress-testing practices, auditable synthetic stress libraries extend them - enhancing resilience while preserving transparency, interpretability, and supervisory trust.

To illustrate portfolio-level implications, Figure 5 compares PnL distributions generated under three stress-testing approaches: (i) traditional deterministic scenarios, (ii) ML-based synthetic shocks, and (iii) GenAI narrative-driven scenarios. The GenAI distribution exhibits wider dispersion in tail outcomes while remaining within plausible magnitude ranges, highlighting richer coverage of extreme but coherent risk configurations.

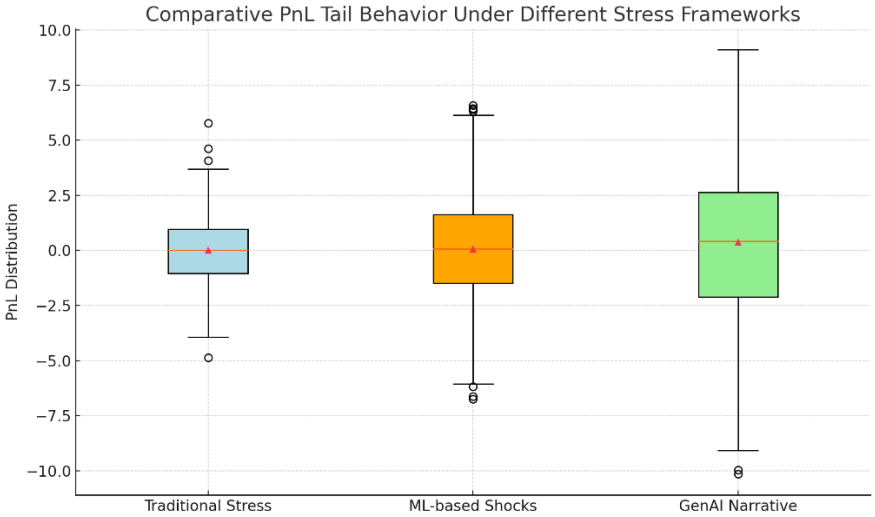

Figure 5: Comparative PnL Tail Behavior Under Different Stress Frameworks (Box-and-Whisker Plot)

2.7. Ethical Considerations and Bias in Scenario Selection

Ethical considerations arise not only in model deployment but also in scenario selection, prioritization, and interpretation. GenAI systems may implicitly emphasize shocks aligned with dominant historical narratives - such as those derived from advanced-economy macro regimes - while under-representing risks faced by emerging markets, smaller sectors, or non-traditional financial actors. Such biases may lead to uneven risk coverage and unintended distributional consequences.

Scenario choices can also influence capital allocation and risk appetite in ways that disproportionately affect specific industries, regions, or stakeholder groups. For example, persistent emphasis on energy-transition shocks may amplify capital costs for certain sectors, while insufficient representation of climate physical risks may understate exposures in vulnerable geographies.

To mitigate these risks, institutions should implement bias-aware prompt design, periodic audits of scenario coverage, and governance committees tasked with ensuring diversity, balance, and fairness in stress-scenario libraries. Ethical review should be embedded as a standing component of GenAI governance rather than treated as an ad hoc control.

Recent supervisory literature reinforces these principles. The Bank for International Settlements (2023) highlights challenges related to opacity, validation, and governance when AI systems are deployed in risk-sensitive financial functions. Similarly, the Bank of England

(2022) emphasizes the importance of explainability, human oversight, and operational resilience in machine-learning applications. From a model-risk perspective, the Federal Reserve Board (2023) extends SR 11-7 principles to complex and AI-based models, underscoring expectations for transparency, ongoing monitoring, and effective challenge.

Taken together, these frameworks provide a clear institutional foundation for situating GenAI-driven stress testing within existing supervisory expectations. Ethical safeguards, bias controls, and transparent governance are therefore not ancillary considerations, but essential prerequisites for the responsible adoption of GenAI in financial risk management.

3. Conclusion

Generative AI is reshaping the practice of financial stress testing and model-risk governance by enabling the construction of synthetic, narrative-driven scenarios that are both contextually coherent and quantitatively calibrated. By integrating textual reasoning with empirical covariance constraints, GenAI extends stress testing beyond static historical analogues, allowing institutions to evaluate the robustness of pricing models, execution algorithms, and liquidity-management systems under complex, cross-asset shocks that traditional frameworks struggle to represent.

At the same time, the deployment of GenAI in risk-sensitive environments requires careful institutional design. Without robust governance, large language model outputs may introduce misleading correlations or economically invalid narratives, undermining model credibility and supervisory confidence. The findings of this review underscore that the value of GenAI-based stress testing lies not in automation alone, but in its integration within structured governance frameworks that include prompt management, evidence attribution, quantitative validation, and human-in-the-loop oversight. These elements are essential to ensuring that generated scenarios remain auditable, explainable, and compliant with supervisory standards such as SR 11-7, the EU AI Act, and the NIST AI Risk Management Framework.

Looking ahead, the convergence of generative modeling and financial risk management has the potential to transform both pre-trade analytics and systemic-risk oversight. Regulators may increasingly expect firms to demonstrate how AI-driven stress tests complement traditional regulatory scenarios, particularly as market structures become more complex and adaptive. Institutions that develop mature, well-governed GenAI pipelines are likely to benefit from faster scenario iteration, broader risk coverage, and improved transparency in model validation.

In summary, GenAI-enabled scenario fabrication represents more than a technical enhancement. When deployed responsibly, it constitutes a shift toward proactive, interpretable, and adaptive risk governance - one that is better suited to the demands of AI-augmented financial markets.

4. Extended Use Cases

1. **Pre-Trade Algorithmic Validation**
   Prior to deployment, trading algorithms can be exposed to GenAI-generated stress scenarios capturing liquidity droughts, volatility spikes, and cross-asset contagion, enabling early identification of fragilities in execution logic, order routing, or hedging strategies.
2. **Regulatory Stress-Testing Augmentation**

   Financial institutions may enrich CCAR, EBA, or Bank of England submissions with internally generated, narrative-driven scenarios that complement supervisory baselines,

demonstrating advanced scenario-design capabilities and model-risk governance maturity.

3. **Cross-Market Contagion Analysis**
   GenAI can fabricate multi-asset chain reactions - such as commodity-driven inflation propagating to rates, currencies, and credit - supporting analysis of second-order systemic effects that are difficult to capture with linear correlation-based models.

4. **ESG and Climate-Risk Simulation**
   By prompting LLMs with climate-transition or physical-risk events (e.g., abrupt carbon-tax implementation or supply-chain disruptions), institutions can assess portfolio resilience under sustainability-linked stressors aligned with EU Taxonomy and SFDR frameworks.

5. **Model-Risk Audit and Explainability Training**
   Synthetic stress scenarios can serve as controlled training corpora for model-risk teams, enabling systematic testing of explainability tools (e.g., SHAP or LIME) and governance workflows under extreme but economically coherent market conditions.

References

Acerbi, C., & Szekely, B. (2017). Backtesting expected shortfall. *Journal of Banking & Finance, 79*, 10–23. https://doi.org/10.1016/j.jbankfin.2017.02.015

Bank for International Settlements. (2023). *Artificial intelligence and machine learning in financial supervision*. *BIS Quarterly Review*. https://www.bis.org/publ/qtrpdf/r_qt2303.htm

Bank of England. (2022). *Machine learning in UK financial services*. https://www.bankofengland.co.uk/report/2022/machine-learning-in-uk-financial-services

Berkowitz, J. (2000). A coherent framework for stress testing. *Journal of Risk, 2*(1), 5–15. https://doi.org/10.21314/JOR.2000.021

Borio, C., Drehmann, M., & Tsatsaronis, K. (2014). Stress-testing macro stress testing: Does it live up to expectations? *Journal of Financial Stability, 12*, 3–15. https://doi.org/10.1016/j.jfs.2013.06.001

Boston Consulting Group. (2024). *AI in risk management: Generative models for scenario design and stress testing*. https://www.bcg.com/publications

Cao, S., Chen, R., & Kritzman, M. (2023). Large language models in finance. *Journal of Financial Data Science, 5*(4), 1–15. https://doi.org/10.3905/jfds.2023.1.123

Čihák, M. (2007). *Introduction to applied stress testing* (IMF Working Paper No. WP/07/59). International Monetary Fund. https://doi.org/10.5089/9781451866230.001

Derman, E., & Koren, M. (2023). Reinforcement learning for systemic stress scenario generation. *Quantitative Finance, 23*(7), 1125–1142. https://doi.org/10.1080/14697688.2023.2175601

Federal Reserve Board. (2023). *Model risk management for complex and AI-based models*. https://www.federalreserve.gov/supervisionreg/srletters.htm

Kritzman, M., Page, S., & Turkington, D. (2021). Regime shifts: Implications for dynamic strategies. *Financial Analysts Journal, 77*(2), 15–29. https://doi.org/10.1080/0015198X.2021.1887051

Kupiec, P. (1998). Stress testing in a value-at-risk framework. *Journal of Derivatives, 6*(1), 7–24. https://doi.org/10.3905/jod.1998.408008

Li, J., Kumar, P., & Mavridis, N. (2024). *Large language models for macro-financial scenario generation* (arXiv:2404.03112). *arXiv*. https://arxiv.org/abs/2404.03112

Mavridis, N., & Kumar, P. (forthcoming). Covariance-constrained generative AI for financial stress testing. *Journal of Computational Finance*. https://doi.org/10.21314/JCF.2025.462

Reuters. (2026, January 20). *Britain needs "AI stress tests" for financial services, lawmakers say*. https://www.reuters.com/sustainability/boards-policy-regulation/britain-needs-ai-stress-tests-financial-services-lawmakers-say-2026-01-20/

Sorge, M. (2004). *Stress-testing financial systems: An overview of current methodologies* (BIS Working Paper No. 165). Bank for International Settlements. https://www.bis.org/publ/work165.htm https://doi.org/10.2139/ssrn.759585

Zhang, Y., Chen, R., & Kritzman, M. (2022). Generative adversarial models for stress testing yield curves and volatility surfaces. *Journal of Financial Data Science, 4*(2), 45–62. https://doi.org/10.3905/jfds.2022.1.089

Appendix A

Reproducibility, Data Schemas, and Code

This appendix documents the data proxies, schema design, mapping logic, and procedural steps used in the illustrative GenAI scenario-fabrication pipeline. The objective is to enable independent replication and benchmarking using publicly available inputs.

A.1 Data Proxies

All inputs are sourced from publicly accessible datasets. Macroeconomic series are obtained from the Federal Reserve Economic Data (FRED) database. Market proxies for equities, rates, foreign exchange, commodities, and credit follow the asset definitions specified in Section 2.4C. No proprietary datasets are required to reproduce the illustrative results.

A.2 Scenario Schema

Each synthetic stress scenario is stored as a structured record with the following fields:

- **prompt_id**: Unique identifier for the scenario prompt
- **model_version**: LLM version used for generation
- **seed**: Random seed applied during generation
- **temperature**: Sampling temperature
- **factors []**: List of affected risk factors (e.g., equities, rates, FX)
- **preliminary_shocks[]**: Raw directional and magnitude estimates extracted from the narrative
- **covariance_window**: Historical window used for calibration
- **aligned_shocks[]**: Covariance-adjusted shock vector
- **diagnostics** {}: Quantitative validation metrics (e.g., Mahalanobis distance, DCC regime match)
- **pnl_stats{}**: Portfolio-level PnL summary statistics from simulation

This schema supports end-to-end traceability from prompt generation through quantitative validation and portfolio impact assessment.

A.3 Mapping Rules

Qualitative descriptors in generated narratives are translated into numerical magnitudes using predefined σ-bands relative to historical volatility. For example, terms such as *"moderate"*, *"sharp"*, or *"severe"* are mapped to ranges (e.g., +1.0σ, +1.5–2.0σ, >2.0σ, respectively). Preliminary shock vectors are subsequently adjusted through covariance alignment to ensure empirical consistency across risk factors.

A.4 Procedural Steps

The scenario-fabrication workflow follows the sequence below:

1. Generate a narrative response from a structured macro-trigger prompt
2. Extract factor directions and qualitative magnitudes from the narrative
3. Map qualitative magnitudes to σ-based numerical ranges
4. Perform covariance alignment using the selected historical window
5. Compute diagnostic metrics (e.g., Mahalanobis distance, regime checks)
6. Run Monte Carlo PnL simulations on the target portfolio
7. Conduct human-in-the-loop review for plausibility and governance compliance

A.5 Access and Replicability

All data inputs rely on public proxies, and all transformation steps are rule-based and fully documented. While the appendix does not provide executable code, the schema definitions, mapping logic, and workflow steps are sufficient to enable independent replication, extension, or benchmarking by other researchers.